Unsupervised multilingual models of speech representation, an approach inspired by cognitive science

Maureen de Seyssel

PSL
UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à l'École Normale Supérieure

Apprentissage non supervisé de modèles multilingues de représentation de la parole, une approche inspirée des sciences cognitives

Unsupervised multilingual models of speech representation, an approach inspired by cognitive science

Soutenue par
Maureen de SEYSSEL
Le 27 novembre 2023

École doctorale n°158
Cerveau, cognition, comportement (3C)

Spécialité
Sciences cognitives

Composition du jury :

Martine ADDA-DECKER Directrice de Recherche, Sorbonne Nouvelle	*Présidente*
Odette SCHARENBORG Assoc. Professor, Technische Universiteit Delft	*Rapporteure*
Hao TANG Lecturer, University of Edinburgh	*Rapporteur*
Casey LEW-WILLIAMS Professor, Princeton University	*Examinateur*
Maria GIAVAZZI Maîtresse de Conférence, ENS	*Examinatrice*
Hervé BREDIN Chargé de Recherches, IRIT	*Examinateur*
Emmanuel DUPOUX Directeur d'Études, EHESS, ENS, INRIA	*Directeur de thèse*
Guillaume WISNIEWSKI Maître de Conférence, Université Paris Cité	*Invité (co-encadrant)*
Camille DUTREY Ministère des Armées	*Invitée*

1794
ENS | PSL

Multilingual and unsupervised models of speech representation, an approach inspired by cognitive science

Abstract

Speech, serving as a key input in the early language acquisition process, carries different types of information. This includes linguistic information - denoting the inherent meaning of the communicated message - and indexical information - which is tied to the speaker's identity (including the identity of the language spoken). In this thesis, we are interested in how infants process these two types of information. We explore how the specificities of an infant's language environment, particularly exposure to multiple and diverse languages, shape their speech perception abilities. We also question whether the representation of indexical information, and language(s) in particular, can influence linguistic learning. Adopting a computational modelling approach, we model infant speech perception for indexical and linguistic information, leveraging recent advancements in machine learning and speech processing. Consequently, our contributions have significant implications for both cognitive science and speech processing.

Throughout this thesis, we, in turn, model indexical speech perception and linguistic speech perception (which involves simulating early language acquisition) from raw speech input, with varying input patterns and conditions, with a particular focus on multilingual speech input. This modelling became feasible due to our development of suitable frameworks and measures for appropriate learning simulations of speech perception and early language acquisition. Our work allows us to underscore the advantages of computational modelling in speech perception in infants, providing guidelines for such an approach. These simulations also enable us to shed light on some hypotheses behind infants' speech processing by serving as proofs of concept. We found that statistical learning mechanisms were enough to simulate early language acquisition in monolingual infants. However, although we found some linguistic learning with the same mechanisms when given bilingual input, we could not replicate bilingual infants' patterns, potentially suggesting that these statistical mechanisms are not sufficient in their language learning process. We also discuss how our work has implications in the speech processing field, discussing the effect of language distance and negative interference.

Keywords: unsupervised speech processing; multilingual; cognitive science; psycholinguistics; machine learning

Apprentissage non supervisé de modèles multilingues de représentation de la parole, une approche inspirée des sciences cognitives

Résumé

La parole, qui est essentielle à l'acquisition du langage, véhicule différents types d'informations. Parmi elles, les informations linguistiques (propres au sens du message communiqué) et indexicales (liées à l'identité du locuteur, dont la langue parlée). Dans cette thèse, nous nous intéressons à la manière dont les nourrissons traitent ces deux types d'informations. Nous explorons de quelle façon les spécificités de l'environnement linguistique d'un nourrisson, en particulier l'exposition à des langues multiples et diverses, façonnent leur perception de la parole. Nous nous demandons également comment la façon dont les informations indexicales sont représentées influence l'apprentissage linguistique. En adoptant une approche de modélisation computationnelle, nous modélisons la représentation des informations indexicales et linguistiques lors de la perception de la parole chez le nourrisson, en tirant parti des avancées récentes en matière d'apprentissage automatique et de traitement de la parole. Par conséquent, nos contributions ont des implications significatives à la fois pour les sciences cognitives et pour le traitement de la parole.

Tout au long de cette thèse, nous modélisons tour à tour la perception indexicale de la parole et la perception linguistique de la parole (qui implique la simulation de l'acquisition du langage) à partir de parole comme seule donnée d'entrée, sous différentes conditions, et en mettant particulièrement l'accent sur l'entrée de parole multilingue. Cette modélisation est passée par le développement de structures et de mesures appropriées pour des simulations d'apprentissage linguistique. Notre travail nous permet de souligner les avantages de la modélisation informatique dans la perception de la parole et l'apprentissage du langage chez le bébé, en fournissant des lignes directrices pour une telle approche. Ces simulations nous permettent également d'éclairer certaines hypothèses sur le traitement de la parole chez les nourrissons en servant de preuves de concept. Nous avons constaté que les mécanismes d'apprentissage statistique étaient suffisants pour simuler l'acquisition précoce du langage chez les nourrissons monolingues. Cependant, bien que nous ayons constaté un apprentissage linguistique avec les mêmes mécanismes avec des données d'entrées bilingues, nous n'avons pas pu reproduire les tendances observées chez les nourrissons bilingues. Ceci pourrait suggérer que ces mécanismes statistiques ne sont pas suffisants dans leur processus d'apprentissage du langage. Pour finir, nous examinons également les implications de notre travail dans le domaine du traitement de la parole, en discutant de l'effet de la distance linguistique et de l'interférence négative.

Mots-clés: traitement automatique non supervisé de la parole; multilingue; sciences cognitives; psycholinguistique; apprentissage machine

Résumé substantiel

La parole est une modalité fondamentale à l'apprentissage du langage chez les bébés dans les communautés entendantes. Cependant, ce signal, en plus de véhiculer des informations *linguistiques* (se rapportant au message partagé), contient également des informations *indexicales* (c'est à dire liée à l'identité du locuteur, et donc indépendante du sens du message). Cela crée la nécessité pour les nourrissons de distinguer l'information linguistique de l'information indexicale, une nécessité qui devient encore plus complexe dans les environnements bilingues où les nourrissons sont exposés à deux langues distinctes (laquelle identité étant considérée comme information indexicale). Comment les bébés parviennent-ils, le cas échéant, à démêler les différents types d'informations ? Quelles sont les interactions existantes entre ces différents types d'informations? Ces questions sont tout aussi pertinentes dans le domaine du traitement de la parole, où les modèles formés sur la parole peuvent également bénéficier de l'exploration de ces éventuelles interactions entre informations linguistiques et indexicales. C'est dans ce contexte de recherche que se situe notre travail dans cette thèse.

Notre recherche se focalise sur l'influence de l'environnement parlé du nourrisson sur sa perception de la parole durant ses premières années. Plus précisément, nous nous intéressons à la manière dont différentes informations interagissent, tant *au sein* de ces informations (par exemple l'influence de la langue parlée sur la capacité à traiter l'identité du locuteur ou les interactions existantes entre différents niveaux linguistiques) qu'*à travers* ces informations (comme l'effet de l'exposition à plusieurs langues sur l'apprentissage linguistique).

Nous adoptons ici une approche de modélisation computationnelle basée sur l'approche de « reverse-engineering » de Dupoux (2018), consistant à mieux comprendre des effets cognitifs observés en les reproduisant à l'aide de modèles informatiques. Avec une telle approche, nous pouvons surmonter certaines des difficultés présentes dans la recherche comportementale en ayant un meilleur contrôle des variables, et permettant de tester certaines hypothèses autrement difficiles à évaluer. La discrimination linguistique signifie-t-elle nécessairement une séparation des langues ? Pouvons-nous répliquer certains effets liés à la perception de la parole sur une multiplicité de langues ? Y a-t-il un rôle de l'exposition à la langue dans la perception bilingue des informations linguistiques ? Comment la quantité de données d'entrée affecte-t-elle le développement de la perception d'information linguistique dans la parole ? Enfin, grâce à cette approche de modélisation informatique, notre travail, bien que fondé sur les principes et les motivations des sciences cognitives, tire parti de certaines des dernières avancées en matière de traitement de la parole, en particulier dans le domaine de l'apprentissage non supervisé. Par conséquent, nous estimons que nos avancées contribuent non seulement aux sciences cognitives, mais qu'elles ont également des implications significatives pour le domaine du traitement de la parole.

Tout au long de notre thèse, nous nous appuyons sur les recommandations de Dupoux (2018) pour la modélisation des systèmes de perception et acquisition du langage et de la parole chez les enfants. C'est à dire que nous considérons un modèle comme "apte" à modéliser ces procédés s'il (1) utilise comme donnée d'entrée de la

parole (et non une approximation tels du texte), (2) n'a *pas besoin d'étiquettes* et donc de supervision, et (3) peut reproduire des *performances humaines* sur une variété de tâches de perception de la parole. Nous faisons attention à utiliser dans nos travaux des modèles de traitement automatique de la parole qui remplissent au mieux ces critères.

Les recherches exposées dans cette thèse résultent d'une combinaison d'articles que nous avons publiés au fil de nos années de doctorat et de travaux originaux. Ces recherches sont structurées autour de trois chapitres, que nous allons résumer ici.

Chapitre 2: Modéliser les informations indexicales dans la perception de la parole

Dans le premier chapitre, nous nous intéressons à la perception de la parole au niveau indexical, en mettant l'accent sur les informations d'identité de langue et d'identité du locuteur. Nous utilisons des modèles d'i-vecteurs (Dehak et al., 2011), initialement proposés en traitement automatique de la langue pour des tâches d'identifications de locuteurs, en tant que modèle de perception de la parole chez le nouveau-né. Ces modèles fournissant une représentation globale des énoncés parlés, nous les considérons comme d'excellents modèles pour la perception d'informations indexicales. Ce chapitre se structure autour de trois sections.

Dans la première section, nous nous inspirons de l'approche de Carbajal et al. (2016) pour modéliser la capacité des nouveau-nés à discriminer les langues, à partir de données d'entrée monolingues (une seule langue) et bilingues (deux langues). Nous démontrons également que la capacité d'un modèle à *discriminer* deux langues (c'est-à-dire à distinguer une langue d'une autre) ne se traduit pas nécessairement par une capacité à *séparer* les langues (c'est-à-dire à regrouper de manière non supervisée les énoncés présentés en différents clusters de langue sans connaître le nombre de langues présentes). Ces résultats remettent en question l'idée selon laquelle les enfants bilingues catégorisent nécessairement les langues de leur environnement au cours de leur développement.

Dans la seconde section, nous modélisons l'effet de familiarisation de la langue (Language Familiarisation Effect, LFE) - qui consiste en la difficulté à discriminer des locuteurs parlant une langue qui n'est pas notre langue maternelle - en nous appuyant sur les travaux de Thorburn et al. (2019). Nous approfondissons leurs travaux initiaux en démontrant que nous pouvons (1) obtenir des résultats graduels et comparables (ce qui n'est pas possible dans les études comportementales) et (2) reproduire cet effet sur un large éventail de paires de langues, attestant de l'universalité de l'effet. Nous présentons également des travaux, en se basant sur cette modélisation, qui atteste de l'importance de la *stabilité* dans la modélisation computationnels de processus cognitifs.

Enfin, dans ces deux premières sections, nous observons un effet de *distance linguistique* sur les performances, avec des paires de langues plus distantes conduisant à de meilleurs scores de discrimination linguistique et à des LFE plus significatifs. C'est suite à cette observation que nous proposons, dans la troisième section, d'utiliser ces modèles d'i-vecteurs comme outil de calcul automatique de distance

acoustique entre les langues. En développant un tel outil, nous ouvrons ainsi la voie à l'utilisation de telles méthodes pour des applications telles que l'annotation de corpus de parole, mais aussi à l'utilisation dans des systèmes automatiques de traitement de la parole.

Non seulement ce chapitre nous a permis de soutenir le modèle d'i-vecteurs comme un bon modèle de perception indexicale de la parole chez les enfants, mais il nous a également permis de mettre en évidence différents concepts méthodologiques importants dans la modélisation computationnelle des processus cognitifs, lesquels sont particulièrement pertinents lorsqu'ils portent sur l'utilisation de plusieurs langues. Nous pouvons citer l'évaluation symétrique des langues, la gradualité, la stabilité, et l'utilisation de modèles cognitifs comme outils.

Chapitre 3: Elaboration d'un système développemental de modélisation d'acquisition du langage

Dans le deuxième chapitre de nos travaux, nous réorientons notre attention des informations indexicales vers les informations linguistiques, qui se rapportent à la substance du message véhiculé par la parole. La capacité à percevoir des informations linguistiques se développant au cours de la petite enfance, nous nous focalisons spécifiquement sur ce processus d'acquisition du langage. Nous nous posons la question de savoir si un processus complexe, impliquant plusieurs niveaux linguistiques, peut être représenté à travers une seule simulation. Dans ce contexte, nous faisons appel à des modèles du traitement automatique de la parole qui s'appuient sur des techniques d'apprentissage non supervisé (Self-Supervised Learning, SSL). Ces modèles capturent l'information à l'échelle de courtes fenêtres du signal, et sont ainsi plus aptes à modéliser l'information à un niveau segmental et suprasegmental, qui englobe l'étendue des informations linguistiques. En particulier, nous combinons un modèle acoustique (qui est composé d'un modèle CPC (Oord et al., 2018) et d'un algorithme de clustering) avec un modèle du langage (LSTM, Hochreiter and Schmidhuber, 1997).

Dans la première section de ce chapitre, nous présentons une structure de simulation de l'apprentissage du langage qui est à la fois *développementale* (elle permet la génération de courbes de développement) et *interlinguistique* (elle autorise la comparaison des conditions de perception linguistique au niveau natif et non natif). Cette simulation repose sur un modèle de l'environnement (les données d'entrée sont issues de discours en français et en anglais provenant de livres audio), un modèle de l'apprenant (la combinaison du modèle acoustique et du modèle de langage décrit précédemment, que nous nommons STELA pour STatistical learning of Early Language Acquisition) et un modèle des évaluations. À ce stade, nous introduisons deux types d'évaluations : phonétique (la tâche machine ABX Schatz et al., 2013 servant comme tâche de discrimination phonétique) et lexicale (sWuggy, une tâche de reconnaissance des mots). Grâce à cette simulation, nous démontrons que nous pouvons reproduire l'apprentissage graduel et simultané aux niveaux phonétique et lexical, tel qu'observé chez les nourrissons. Notre simulation se basant uniquement sur des mécanismes d'apprentissage statistique, nous soutenons également que l'hypothèse de l'apprentissage statistique (ou "statistical learning",

voir Saffran and Kirkham, 2018) est suffisante pour initier l'acquisition du langage chez les bébés. Enfin, nous montrons que cet apprentissage ne nécessite pas de catégories linguistiques préétablies, ni au niveau phonétique (comme déjà défendu par Schatz et al., 2021) ni au niveau lexical.

Dans la deuxième section du chapitre, nous étendons nos travaux en proposant une tâche d'évaluation, ProsAudit, au niveau prosodique, et plus particulièrement au niveau de la prosodie structurelle (relative à la façon dont la prosodie contribue à l'organisation du discours en délimitant les structures linguistiques telles que les mots ou les phrases). Nous proposons deux sous-tâches permettant d'évaluer cette connaissance prosodique dans les modèles de parole SSL à différents niveaux (protosyntaxe et lexicale). Nous avons également évalué des humains sur ces mêmes tâches, permettant des comparaisons directes entre performances humaines et machines. Enfin, nous montrons que la simulation d'apprentissage du langage exposée précédemment permet de générer des courbes développementales tout aussi graduelles et simultanées que leur homologues phonétiques et lexicales. Ces courbes s'avèrent être en accord avec les résultats psycholinguistiques existants. Il convient de souligner que, bien que les tâches phonétiques et lexicales soient disponibles en anglais et en français, cette tâche n'est, pour l'instant, disponible qu'en anglais.

Nous clôturons ce chapitre par une discussion approfondie sur le rôle et l'avenir des simulations d'apprentissage des langues telles que celles présentées dans ce chapitre. Entre autres, nous proposons une série de critères qui, selon nous, devraient guider les futures simulations d'apprentissage, à savoir qu'elles devraient viser à être causales, quantitatives, réalistes et englober plusieurs niveaux linguistiques

Chapitre 4: Modélisation de l'acquisition du langage chez les enfants bilingues

Dans ce dernier chapitre, nous adaptons la structure de simulation d'apprentissage linguistique présentée ci-dessus à la modélisation d'acquisition du langage chez les enfants bilingues, en modifiant le modèle de l'environnement de façon à donner commes entrées au modèle de l'apprenant deux langues au lieu d'une seule au préalable (nous avons désormais des modèles entraînés sur de l'anglais et du français simultanément). Cela nous permet donc des comparaisons directes entre l'apprentissage monolingue et l'apprentissage bilingue tel que modélisé par notre simulation.

Dans un premier temps, nous nous intéressons à la représentation des différents types d'information, indexicale et linguistique, dans des modèles monolingues et bilingues. Nous nous intéressons particulièrement à la capacité de nos modèles, originellement proposés comme modèle de perception linguistique de la parole, à discriminer des langues comme le faisaient les modèles i-vecteurs, proposés comme modèles de la perception indexicale. Nous trouvons que les modèles bilingues (entraînés sur deux langues) parviennent à discriminer les langues, mais que les modèles monolingues (entraînés sur une seule langue) ne le parviennent pas. Cependant, la capacité des modèles bilingues à discriminer les langues semble

aussi dépendre de la quantité de données d'apprentissage, les modèles entraînés sur peu de données ne parvenant pas non plus à une telle discrimination. Ces résultats, qui diffèrent des résultats en psycholinguistique, suggèrent que (1) des modèles SSL multilingues de la parole sont capable d'apprendre une discrimination de leurs langues d'entrapinement et (2) qu'ils ne sont pas forcément adaptés à une simulation de la perception indexicale de la parole, et qu'une combinaison avec des modèles plus globaux comme les i-vecteurs seraient nécessaires pour cette application.

Nous nous penchons ensuite sur la comparaison des courbes développementales produites par les modèles dans les conditions monolingues et bilingues, et ce au niveau phonétique, lexical et prosodique. Bien que les résultats varient en fonction des différents niveaux et représentations étudiées, nous trouvons systématiquement un coût à l'exposition à des données bilingues, coût qui n'est pas observé chez les enfants dans les études psycholinguistiques. Nous trouvons également qu'il y a un effet important de la proportion d'exposition à chacune des langues, effet qui n'est pas nécessairement linéaire. Nous concluons de ces analyses que le modèle STELA, proposé initialement comme modèle de l'acquisition du langage chez les enfants monolingues, ne permet pas de simuler fidèlement l'acquisition du langage chez les enfants bilingues. Nous avançons l'hypothèse que cette situation pourrait résulter du besoin de mécanismes supplémentaires pour un apprentissage bilingue, lorsque les mécanismes d'apprentissage statistiques suffisent dans le cadre d'un apprentissage monolingue.

De plus, nos résultats ont des implications dans le domaine du traitement automatique de la parole, car mettant en exergue l'importance d'interférences négatives lors d'entraînements multilingues, et ce particulièrement lorsque les modèles sont entraînés sur relativement peu de données.

Conclusion

En conclusion, nos recherches apportent une contribution significative aux domaines de la psycholinguistique et de la perception de la parole. Elles offrent une meilleure compréhension du traitement de la parole chez le nourrisson, à la fois au niveau indexical et linguistique et leurs éventuelles dépendances. Elles mettent également en lumière les défis potentiels et les opportunités dans l'entraînement multilingue (que ce soit par le rôle de la distance entre langues ou les conséquences en interférences négatives) et la représentation de la parole des informations indexicales et linguistiques. Notre travail sert de passerelle cruciale entre ces deux domaines, exploitant les atouts de chacun pour renforcer l'autre. Les connaissances acquises grâce à cette recherche soulignent non seulement la valeur de la modélisation computationnelle dans l'étude de la psycholinguistique, mais démontrent également comment les théories psycholinguistiques peuvent inspirer et informer la conception de modèles de traitement de la parole plus efficaces.

Pour faire avancer notre travail, une direction prometteuse pourrait résider dans une meilleure simulation de la perception de la parole bilingue. En intégrant à la fois les modèles globaux (*indexicaux*) et (supra-)segmentaux (*linguistiques*) de la perception de la parole dans un cadre plus global, et en incorporant des mécanismes

d'apprentissage supplémentaires démontrés par les nourrissons bilingues, nous espérons refléter plus précisément la trajectoire de développement de l'acquisition du langage bilingue. De tels progrès pourraient non seulement conduire à une meilleure compréhension de l'acquisition du langage chez les enfants bilingues, mais pourraient aussi offrir une solution potentielle au coût dû à l'exposition multilingue observé dans notre recherche. Ceci pourrait, à son tour, avoir des implications significatives pour le domaine du traitement de la parole multilingue.

Acknowledgments / Remerciements

Here comes the time for acknowledgements, arguably the part I dreaded writing the most. Not that it is uncomfortable, all the contrary, but more that I fear unintentionally overlooking some people - as I probably will - among all the marvellous persons who have supported me these last years and made this work possible. Although the rest of this thesis is in English, I will be switching languages depending on whom I am acknowledging.

First of all, I would like to express my gratitude to all members of my jury who have accepted to evaluate this thesis: Hao Tang and Odette Scharenborg who are carrying the difficult work of being rapporteurs, Martine Adda-Decker, Casey Lew-Williams, Maria Giavazzi, Hervé Bredin and Camille Dutrey. I am sincerely appreciative of the time and effort you have committed to reviewing this thesis, and although I have not received it yet, I am certain your constructive feedback will be very valuable to my work, present and future. Merci Maria et Hervé d'avoir, en plus de faire partie de mon jury, accepter de me suivre tout au long de ces trois années en tant que membres du comité de suivi (et surtout merci d'avoir jonglé avec les demandes de suivi de dernière minute!).

Bien sûr, je ne pourrais pas avoir effectué le travail de ces dernières années sans le soutien et la supervision de mes deux directeurs de thèse, Guillaume et Emmanuel. Guillaume, merci pour ton suivi ces dernières années, tes retours minutieux sur chaque papier et chaque écrit. Merci également de t'être pris au jeu et intéressé à mon sujet dès la première discussion que nous avons eu ensemble. J'espère que l'approche plus "parole" de mes travaux t'aura d'autant plus donné envie de continuer sur cette lancée, et, qui sait, peut-être aussi un petit goût pour les sciences cognitives! Emmanuel, je ne saurais même pas par où commencer. Merci de m'avoir donné ma chance lorsque je t'ai contacté de nulle part en une fin d'été 2019. Merci de m'avoir soutenu dans la recherche de financements, et de m'avoir permis de commencer cette thèse. Merci de m'avoir laissé tant de liberté dans ma recherche. Merci pour toutes nos discussions. Enfin, plus que tout, merci de m'avoir inculqué plus profondément un amour de la recherche, méthodologie et rigueur scientifique. Je ne saurais te dire combien je te suis reconnaissante.

Speaking of people who have motivated me to do research, there are two persons particularly that I owe a lot, and without whom I probably would not have continued down this path. Ansgar, c'est finalement grâce à toi que tout a commencé, toi qui, peut-être sans t'en rendre compte, m'a donné le goût pour l'approche computationnelle au sein des sciences cognitives. Je me rappelerai toujours la première fois où tu m'as demandé de faire tourner un *script* via le *terminal* - du charabia pour moi à l'époque! Qu'il est long le chemin parcouru. Comment, sans toi, aurais-je décidé de m'intéresser au traitement automatique du langage et de la parole? And that's where you come into play, Peter. Thank you for believing in me when I came to you ready to start a research project which implied deep machine learning methods, although acknowledging I probably knew much less than most other students at the time. Thank you for supporting me

throughout this research project and after. Ansgar and Peter, I am confident I would not have been equipped with the necessary tools to write this thesis without your support, past and present.

Voici venu le tour de l'équipe CoML. En fait, je ne saurais même pas imaginer à quoi aurait pû ressembler ces quatre dernières années sans la présence et le soutien de chacun d'entre vous, à travers vents, marées et covid. Tout d'abord, Marvin. Evidemment, tu sais déjà ce que je pense de toi, tu es un des chercheurs les plus brillants que je connaisse, et comme si ce n'était pas assez, également l'une des personnes les plus oufs que je connaisse. Bon déjà, soyons clair, il est évident que cette thèse ne serait pas la même sans toi, et je n'aurais pas pu rêver meilleur collaborateur au long de ce travail ! Evidemment, toutes nos discussions vont me manquer. Juliette, bah toi tu me manques déjà au labo, alors finalement c'est pas plus mal que cette thèse finisse. Merci aussi pour nos discussions, scientifiques ou non (surtout non), et j'espère bien que nos petits dîners vont continuer ! Rachid, soyons clair, finalement c'est grâce à toi que je me suis retrouvée ici, lorsque tu as acceptée de me rencontrer lorsque je cherchais à rebound avec les sciences cognitives, et que tu m'as fait l'apologie du labo. Tu as bien fait ! Salah, ok cela fait quelques années que tu as quitté l'équipe, mais t'as été sans aucun doute celui qui m'a le plus motivée toute cette première année, la vie au bureau aurait été différente sans toi ! Mathieu, Manel, Marianne - que serait le bureau 22 sans vous ? Bon c'est vrai que je n'ai pas été la plus présente ces derniers mois, mais que du bonheur de savoir vous y retrouver lors de mes passages à Paris ! Xuan-Nga, Catherine, Sabrina: heureusement que vous étiez là, non seulement parce que vous êtes les personnes les plus efficaces qu'on aurait pu avoir ici, mais aussi parce que vous êtes beaucoup trop cools. Et tous les autres, présents ou passés, évidemment on est tellement nombreux que je vais en oublier à tous les coups, mais juste merci de faire de l'équipe ce qu'elle est, Robin, Tu-Anh, Rahma, Paul, William, Hadrien, Maxime, Mathieu B, Julien, Nick, Julien, Mitja et tant d'autres ! Finally, Andrea and Arthur, thank you both for having been great interns. I hope you are so proud of your work!

Il y en a d'autres que, bien que déjà partis depuis (plus ou moins) belle lurette lorsque je suis arrivée au labo, je ne peux pas oublier: Bogdan, Ewan,Abdellah et Thomas, c'est un peu comme si vous faisiez toujours parti des murs ! Merci pour vos discussions et collaborations ces dernières années! De même, Sharon et Alex (C), merci pour tous vos retours et discussions si pertinentes. Enfin, merci Alex (de C) et Justine de m'avoir donné la chance d'enseigner, c'est une expérience qui va clairement rester gravée de manière positive !

I would like to recognise those friends I have encountered throughout my academic journey. The memories we have created and the bonds of friendship we have forged have not only enriched my life personally but also profoundly shaped the researcher I have become. Sofia, what would London life have been without you? Kai (see this time I could not forget you!), Anna, Jason, Simon, Enno: would I have even been able to make it through the masters without you? Let's plan

an Edinburgh reunion soon. Bérengère: let's be honest, why do you think I got interested in bilingualism? I could not have imagined a better PhD role model!

Vient le moment du paragraphe le plus compliqué à écrire, du fait d'à quel point vous êtes si nombreux, amis d'enfance ou plus récents, à m'avoir soutenu pendant ces années. Je ne sais pas si je vais pouvoir tous vous citer, et je vais très probablement vous oublier, mais à chacun d'entre vous, merci, merci d'avoir supporté mes sautes d'humeur et mes coups de stress. Merci aussi pour tous les fous rires, voyages, week-ends, et nombreuses discussions plus ou moins philosophiques. Merci tout simplement de votre présence et votre soutien constant. Vous avoir à mes côtés est une chance inestimable. Clémence, Mahaut, Tiphaine, Marguerite, Félicité, Héloïse, Pierre-Louis, Laurence, Philippine, Guillaume, Valentin, Paul, Xavier, Camille, Laure, Eléonore, Gaëlle, Marie-Astrid (les deux!), Victor, Constance, Laura, Marie, Nina, Juliette (R et I), Aurore, Inès, Sophie, Julien, et tous les autres, merci! Merci également à Laure, Coralie, Germain, Morgan, Elise, Olivier, Louise et Mathieu d'avoir rendu sans le savoir mon arrivée à Nice si facile !

Evidemment, je ne peux finir cette série de remerciements sans exprimer du fond du cœur ma gratitude à toute ma famille.

Théophile, tu sais que je peine à m'exprimer, mais sache que tu es juste l'un des meilleurs frères qui soit, et je suis tellement contente de t'avoir dans ma vie et j'espère bien réussir à te convaincre de déménager au soleil dans les prochaines années. Lorraine, clairement tu es juste formidable, et je suis tellement fière de tout ce que tu as fait ces quatre dernières années. Brieuc, tu es juste devenu beaucoup trop grand beaucoup trop vite, et tu m'impressionnes tellement. Il parait que l'on se ressemble - alors, à quand la thèse ? Marin, petit marinos, plus si petit que ça ! Ta gentillesse, ton amour du social et des autres ne sont que des exemples pour nous tous. Merci, merci, merci, merci à tous les quatre pour votre soutien, et juste d'être vous, vous qui finalement aurez (presque) tous fini vos études avant votre grande sœur, où va le monde. Merci à vous mes grand-parents, Mam, Dad (qui de là-haut est si fier je le sais), Bon-Papa et Bonne-Maman. Vous êtes des soutiens inébranlables, et c'est si agréable de se savoir entourée comme cela. Merci également Sophie et Edouard, vous qui avez finalement suivi de près cette thèse, et que je considère désormais comme faisant partie de la famille.

Enfin, merci Papa et Maman, pour tout votre amour et soutien toutes ces années, pendant et avant la thèse. Je ne pourrais pas l'avoir fait sans vous.

Je ne peux évidemment finir sans te remercier, Maxime. Tu le sais, tu as été le meilleur soutien dont j'aurais pu rêver ces trois dernières années. Et quelle joie de savoir que tu continueras à l'être pour toutes les années à venir. Merci.

List of author's publications

* denotes equal contribution as first author

Included in the body of the thesis

- **de Seyssel, M.** & Dupoux, E. (2020). Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors. In *Proceedings for the Annual Meeting of the Cognitive Science Society 2020.*
- **de Seyssel, M.**, Wisniewski, G., & Dupoux, E. (2022). Is the Language Familiarity Effect gradual ? A computational modelling approach. In *Proceedings for the Annual Meeting of the Cognitive Science Society 2022.*
- **de Seyssel, M.**, Wisniewski, G., Dupoux, E., & Ludusan, B. (2022). Investigating the usefulness of i-vectors for automatic language characterization. In *Proc. Speech Prosody 2022, 460-464.* doi:10.21437/SpeechProsody.2022-94
- **de Seyssel, M.**, Lavechin, M., Adi, Y., Dupoux, E., & Wisniewski, G. (2022). Probing phoneme, language and speaker information in unsupervised speech representations. In *Proc. Interspeech 2022, 1402-1406.* doi:10.21437/Interspeech.2022-373
- Lavechin, M.*, **de Seyssel, M.***, Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., & Dupoux, E. (2022). *Can statistical learning bootstrap early language acquisition? A modeling investigation.* [under review]. ArXiv. doi:10.31234/osf.io/rx94d
- **de Seyssel, M.**, Lavechin, M., Titeux, H., Santos, A., Virlet, G., Thomas, A., Wisniewski, G., Ludusan, B., & Dupoux, E. (2023). ProsAudit, a prosodic benchmark for self-supervised speech models. In *Proc. Interspeech 2023.*
- **de Seyssel, M.***, Lavechin, M.*, & Dupoux, E. (2023). Realistic and broad-scope learning simulations: first results and challenges. *Journal of Child Language*, 1-24. doi:10.1017/S0305000923000272

Included in the appendices

- Nguyen, T. A.*, **de Seyssel, M.***, Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., & Dupoux, E. (2020). The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *Neurips Workshop on Self-Supervised Learning for Speech and Audio Processing.*

Not included in the manuscript

- Endress, A. & **de Seyssel, M.** (2022). *The specificity of sequential Statistical Learning: Statistical Learning accumulates predictive information from unstructured input but is dissociable from (declarative) memory.* [under review]. PsyArXiv. doi:10.31234/osf.io/u9z4a

- Lavechin, M., **de Seyssel, M.**, Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E. & Cristia, A. (2023). *Statistical learning models of early phonetic acquisition struggle with child-centered audio data.* [under review]. PsyArXiv. doi:10.31234/osf.io/hav58

- Dunbar, E., Bernard, M., Hamilakis, N., Nguyen, T. A., **de Seyssel, M.**, Rozé, P., Rivière, M., Kharitonov, E., and Dupoux, E. (2021). The Zero Resource Speech Challenge 2021: Spoken Language Modelling. In *Proc. Interspeech 2021,1574–1578.*

- Lavechin, M., **de Seyssel, M.**, Gautheron, L., Dupoux, E., & Cristia, A. (2021). Reverse-engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics, 8*, 389-407. doi:10.1146/annurev-linguistics-031120-122120

- Nguyen, T.A., **de Seyssel, M.**, Algayres, R., Roze, P., Dunbar, E., Dupoux, E. (2020). *Are word boundaries useful for unsupervised language learning?* [Unpublished technical report]. ArXiv. doi:10.48550/arXiv.2210.02956

Contents

Acronyms

BLiMP Benchmark of Linguistic Minimal Pairs

CPC Contrastive Predictive Coding

DNNs Deep Neural Networks

ERPs Event-related Potentials

GMM Gaussian Mixture Models

IPA International Phonetic Alphabet

LDA Linear Discriminant Analysis

LFE Language Familiarity Effect

LSTM Long Short-Term Memory model

MFCC Mel-Frequency Cepstral Coefficients

OPOL One Parent One Language

RNN Recurrent Neural Network

sBLiMP Synthesised Benchmark of Linguistic Minimal Pairs

SSL Self-Supervised Learning

STELA STatistical Learning of Early Language Acquisition

TP Transitional Probability

VAD Voice Activity Detection

WALS World Atlas of Language Structures

List of Tables

List of Figures

Chapter 1

General Introduction

Speech is the foundational input in the language acquisition process, serving as the primary source of language knowledge for infants in hearing communities. Indeed, speech carries primarily linguistic information, which is information directly related to the message conveyed, including elements of sounds, which in turn form words and sentences. However, there is also a lot more information carried by the speech signal which does not directly relate to its meaning, including indexical information (which pertains to the speaker, such as their identity and the language they speak) and paralinguistic information (relating to the manner the message is delivered and includes information such as emotions or speaking rate). All of these different types of information coexist, creating the need for infants to discern them from one another. In fact, this process becomes even more complex in bilingual environments where infants are exposed to two distinct languages. Indeed, not only are there interactions between these different types of information, but also the frontiers between the different levels are not necessarily clear-cut, and some characteristics considered indexical in one language could be linguistic in another. An example of such is the creakiness feature, which, while in languages like English, is considered indexical information, can also have a phonemic status and therefore be classified as linguistic information in other languages such as Jalapa Mazatec (Silverman et al., 1995). In such cases, not only must indexical and linguistic information be separated, but the separation is also dependent on the language spoken, making it pivotal in the language acquisition process. Similarly, in the field of speech processing, the disentanglement of linguistic and indexical information from raw speech input is a compelling research question. *How are different types of information extracted, and how do these representations translate into abstract language knowledge? Can models effectively differentiate between these information types?* It is within this multifaceted context that we situate our work in this dissertation.

Bridging these considerations, our research pivots towards the understanding of how an infant's language environment, especially exposure to multiple and diverse languages, shapes their speech perception abilities during their first years of life. *Is there an effect of native language on the perception of speaker identity? If an infant constructs a representation of the signal that captures indexical information, how will they eventually abstract these representations to learn the linguistic information?*

Given our interest in the impact of multilingual input, we dedicate a significant part of our research to the question of bilingual speech environments and their potential impact on speech perception and language acquisition. *How do environmental factors such as speech input patterns and language distance affect indexical speech perception? Is there a cost to bilingual compared to monolingual input in the language development process? Can a bilingual-born infant learn linguistic information without first resorting to language separation?*

As a specificity to our work and to delve deeper into these intricate processes and their possible interactions, we adopt a computational modelling approach based on Dupoux (2018) 's *reverse-engineering approach* to modelling speech perception. This approach, which endeavours to gain an understanding of observed cognitive effects by replicating them with computational models, will be further defined in this introduction (§1.5). With such an approach, we can overcome some of the difficulties present in behavioural research by having better control over variables that are otherwise difficult to differentiate, allowing us to answer additional questions. *Does language discrimination necessarily mean language separation? Can we replicate the modelled effects of speech perception on a multiplicity of languages? Is there a role of language exposure in the bilingual perception of linguistic information? How does the quantity of input affect early language acquisition?* Finally, as a result of this computational modelling approach, our work, although grounded in the principles and motivations of cognitive science, leverages some of the latest advances in speech processing, especially in machine learning and unsupervised learning. Consequently, we find our advancements not only contribute to cognitive science but also carry significant implications for the field of speech processing.

In this general introduction, we first elaborate on the notion of speech perception (§1.1) and further define the specific indexical and linguistic aspects of interest in this work. We then present an overview of some psycholinguistics findings regarding infant's speech perception, both at the indexical (§ 1.2) and linguistic (§ 1.3) levels. Still presenting psycholinguistics results, we also delve into the specific case of infant speech perception in bilingual environments (Section 1.4). In all three sections, we will focus our interest on psycholinguistic findings that we will model in this thesis (literature on relevant speech processing models and findings will be presented within the different chapters). We then present the computational modelling approach we adopt in this thesis and discuss how such an approach can be helpful in the context of our work (§ 1.5). Finally, we summarise our upcoming motivations and contributions, along with an outline of the different chapters that make up this thesis (§ 1.6).

1.1 Levels of speech perception

When speaking of speech perception, people primarily think of the content of the message perceived and understood from the processed speech (e.g. phonemes, word-form recognition). This type of information is predominantly embedded within the

speech signal but represents only one level of speech perception. However, there are other levels to consider. To establish consistent terminology in our thesis, we group them into three primary levels: the *linguistic*, *indexical*, and *paralinguistic* levels. Each type of information also operates at different granularities on the speech segment, that is, spanning different durations within the speech signal.

We provide details on these levels below, also outlined in Table 1.1 and Figure 1.1.

Level	**Focus**	**Definition**	**Examples of information**
Linguistic	What?	Content or meaning of speech	Phonemes, words, structural prosody
Paralinguistic	How?	How speech is delivered and emotional content	Tone, emotions, non-verbal cues
Indexical	Who?	Identity of the speaker	Gender, language, speaker characteristics

Table 1.1: Overview of the different levels of speech perception

Linguistic This first level encompasses all aspects related to the *meaning* or content of the speech. Within this level, information can occur at different granularities: the *segmental* granularity, which spans a duration of approximately 5 to 20ms and includes linguistic information such as phonemes and words, and the *suprasegmental* granularity, which can last up to approximately 1 second and includes, for example, structural prosody (intonation and rhythm), which can give us information about the meaning of the sentence (e.g. where stress or pauses are located in an utterance). We can identify linguistic information by answering the *what is being said?* question. Finally, when referring to language acquisition, we usually mean the developmental learning of this linguistic-level information that enables us to transform sounds into meaningful information.

Indexical This second level focuses on the *identity* of the speaker, including their gender and spoken language, and is therefore referred to as "indexical" in psycholinguistics literature. This time, the meaning of the speech content is set aside. Information at the indexical level is processed at a *global* granularity, spanning an entire utterance. In other words, indexical information refers to the *who is speaking?* question.

Paralinguistic This final level concerns the *manner* speech is delivered and includes emotional content, voice manner, styles, and non-verbal cues such as

laughter. Paralinguistic information concerns the *how is it being said?* question. As for linguistic information, paralinguistic information is at the *segmental* and *suprasegmental* granularities. Paralinguistic information is not a focus of this thesis.

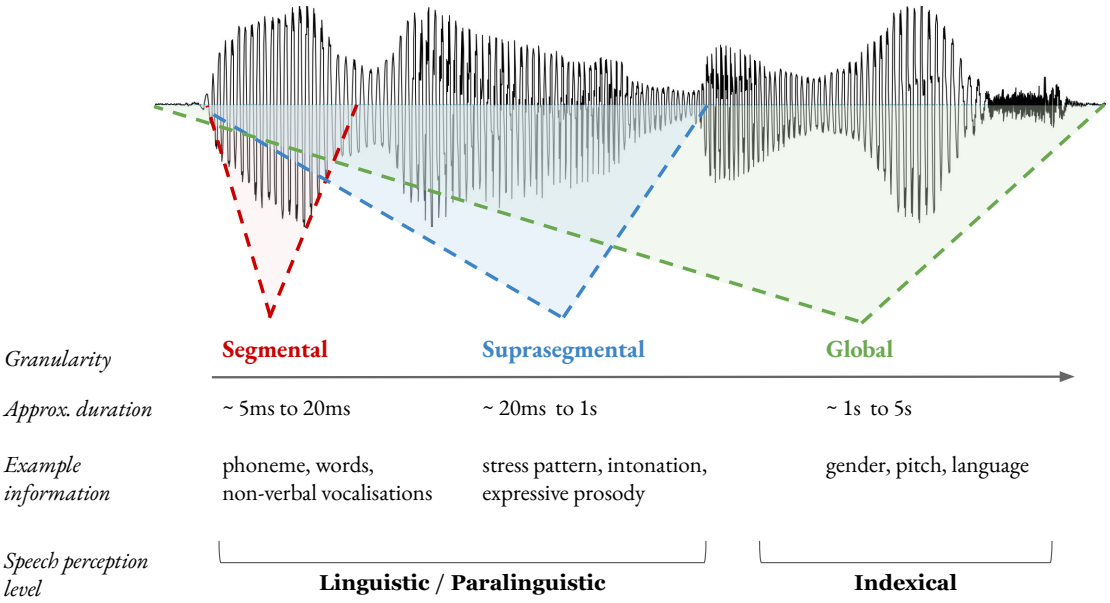


Figure 1.1: Illustration of the distinct levels of focus granularity observed across different levels of speech perception

1.2 Early development of indexical representation

We now delve into how indexical information (information pertaining to the speaker's identity) is processed by infants. We constrain our focus on language and speaker identity, as these are the two aspects of indexical information discussed in the thesis.

1.2.1 Representation of language(s)

Language discrimination is one of the most well-studied facts in infant language speech perception and one which we particularly focus on in the present thesis. This relates to the ability to *differentiate speech in one language from that in another language* when presented with the two. In a seminal experiment, Mehler et al. (1988) showed that four-day-old infants exhibited different sucking-rate patterns in a high-amplitude sucking procedure when presented with their native language (French) and a non-native language (Russian). After being exposed to speech from their native language during a habituation phase, infants who were then presented with non-native speech exhibited an increase in sucking rate compared to those presented with native speech again, suggesting a surge of interest and, hence, some language discrimination ability. This effect held even when the

speech was low-pass filtered, retaining only prosodic (rhythmic and stress-related) information. Similarly, Moon et al. (1993) found evidence of preference for the infants' native language using an English-Spanish language pair and a preference procedure, ultimately supporting the same conclusion of language discrimination ability.

Subsequent research investigated further the role of rhythm in such language discrimination ability. Nazzi et al. (1998) found that newborns showed language discrimination capacities between low-pass filtered speech from two non-native languages if the two languages belonged to different rhythmic classes (that is, differing in their rhythmic pattern, e.g. English and Japanese), but not if they belonged to the same rhythmic class (e.g. English and Dutch). Follow-up studies on other language pairs replicated these findings (Ramus et al., 2000; Ramus, 2002b).

Interestingly, as infants age, their ability to discriminate language does not rely as much on rhythm anymore but on their familiarity with one of the languages. Studies of infants in their first semester have shown that they find it more difficult to discriminate between two unfamiliar languages when they were able to do so earlier in age, and this is even when the two languages are of different rhythmic classes (Bahrick and Pickens, 1988; Bosch and Sebastián-Gallés, 1997; Nazzi et al., 2000). These results suggest a gradual tuning over time to the features of the infant's native language. However, these results should be approached cautiously as, in a recent meta-analysis, Gasparini et al. (2021) were not able to confirm this effect of age on language discrimination based on rhythm class[1]. Therefore, there is still a need for more research on this subject.

Finally, Ramus (2002a) suggested that speaker identity can interact with language perception, potentially making it more difficult for infants to disentangle speaker and language identity from one another.

1.2.2 Representation of speaker(s)

Extensive research in the past literature has shown that newborns exhibit a preference for their mother's voice, suggesting that they have the ability to discriminate it from other voices (Mills and Melhuish, 1974; Mehler et al., 1978; DeCasper and Fifer, 1980). This preference is even present before birth, as suggested by fetus-based studies, with the heart rate increasing when hearing the mother's voice compared to other voices (e.g. Hepper et al., 1993; Kisilevsky et al., 2009). This suggests that the preference for the mother's voice is due to the familiarity of the newborn with it, which they have been exposed to during prenatal development. Although it may appear that the preference is a mother voice only prerogative, infants also show speaker discrimination abilities when they are familiar with at least one of the voices, such as their father's voice (Decasper and Prescott, 1984; Ward and Cooper, 1999), and this even when the speakers are of the same gender. Furthermore, newborns (Floccia et al., 2000), and young infants (Miller et al., 1982; Miller, 1983) also show the ability of gender-based voice discrimination

[1]However, they did find an effect of age on native preference, with an increasing preference over time for a non-native language only if the language is in the same rhythmic class as their native language.

even when the voices are unfamiliar. Finally, although it seems more complex, this ability also extends to unfamiliar, within-speaker voices, with infants able to discriminate between - but not recognise - unfamiliar same-gender voices (see Fecher et al., 2019 for a discussion on the topic). Interestingly, the ability to detect speaker changes seems to vary depending on the language spoken by the speakers, with better discrimination observed when the speakers are native in the infant's language (Johnson et al., 2011; Fecher and Johnson, 2018). This effect, also called the Language Familiarity Effect (LFE), was initially found in adults (Goggin et al., 1991; Johnson et al., 2018), and suggests a strong interweaving between speaker and language perception (we will present this effect in more depth in Chapter 2, §2.3).

Finally, evidence suggests that infants progressively learn to abstract from talker identity during linguistic learning. At six months old, infants recognise different realisations of the same vowel as belonging to the same category, even across gender (Kuhl, 1979). Furthermore, electrophysiological studies indicate that newborns' brains can account for speaker variations in their ability to extract phonetic categories (Dehaene-Lambertz and Pena, 2001). Infants aged 7 to 9 months struggle to recognise words as identical when spoken by voices of different genders, but they manage to do so by ten and a half months (Houston and Jusczyk, 2000), suggesting that speaker identity can indeed affect the processing of lexical learning.

Most of the work presented in these sections provides evidence of speaker and language discrimination and preference. However, we must determine to what extent infants use this discrimination ability in their downstream processing. We discussed evidence that speaker variability could sometimes affect linguistic learning, but *what does this teach us about how they represent speech according to this information? Is there one indexical piece of information that is more important than another? Can (and do) infants necessarily separate speech based on this indexical information (in other words, if they do not know how many languages or speakers they are presented with, will they automatically separate the speech along these dimensions)?* These are some of the questions we will attempt to address in this thesis.

1.3 Early development of linguistic representation

Another essential aspect of speech perception, and perhaps the one most focused on, is that of linguistic information. In other words, how do infants and young children perceive the linguistic components that ultimately help them understand the meaning of the perceived speech? Linguistic perception is intrinsically linked with language development or acquisition, as the perception gradually changes along with the learning process. Therefore, the perception of linguistic components involves the ability to discern phonemes (the sounds of a language), morphemes and words (combinations of phonemes that form meaningful units), syntax (the structure of such units), and semantics (the meaning behind each component). Prosody, also referred to as the music of speech (Wennerstrom, 2001), is an additional level that

can carry information about the speech's content but interplays multiple aspects of linguistic learning.

These different levels of language acquisition are intrinsically interconnected, with learning at some levels aiding other levels, not always linearly, leading to what has been termed linguistic bootstrapping (Höhle, 2009). We will discuss this interplay of levels in linguistic learning in greater detail in Chapter 3; for now, we will provide an overview of the main acquisition milestones for infants in early language acquisition on the levels we focus on in this thesis (principally phonetic, lexical and prosodic levels).

1.3.1 Phonetic representation

Phonetic perception refers to the processing of speech sounds and the capacity to distinguish these diverse speech sounds, or *phonemes*, that constitute a language. Young infants have the ability to discriminate between phonemes of most languages, including those to which they have never been exposed to (Streeter, 1976; Eimas et al., 1971; Aslin et al., 1998), and even when these contrasts are remarkably subtle (Sundara et al., 2018). Yet, adults find it very difficult to discriminate between phonetic units which are not used phonemically in their native language - in other words, those that do not distinguish meaning in their language (Best et al., 2001; Iverson et al., 2003; Miyawaki et al., 1975; Zhang et al., 2005; Guion et al., 2000). The most famous evidence of such difficulty, and the focus of most of these studies, is the challenge for Japanese speakers to discriminate between the two English consonants [r] and [l], a task which English speakers perform easily (Goto, 1971; Miyawaki et al., 1975).

When does this native language sound specificity appear in infants? Actually, this attunement refines during the first year of life, with ten months of age being the commonly accepted threshold at which infants are no longer able to discriminate non-native phonemic contrasts (Kuhl et al. (1992); Polka and Werker (1994); Kuhl et al. (2006); Werker and Tees (1984); see Maurer and Werker (2014) for a review). Moreover, the decline in vowel discrimination seems to start slightly before that for consonants (Polka and Werker, 1994). While most of these studies have used behavioural designs, these attunement patterns have also been replicated in electrophysiological studies using Event-related Potentials (ERPs) (Rivera-Gaxiola et al., 2005; Minagawa-Kawai et al., 2007).

The primary explanation for infants' attunement to phonemes of their native language is based on the *distributional learning* hypothesis. This hypothesis posits that infants learn the frequency distributions of sounds they are exposed to, specifically those of their native language (Anderson et al., 2003; Yoshida et al., 2010; Maye et al., 2002; Werker et al., 2007). According to this hypothesis, language exposure is the key predictor of the attunement effect and relies on the infant's capacity for *statistical learning*.

While language exposure remains the main factor affecting phonetic learning in infants, it is worth noting that other factors can also influence speech perception, including sensorimotor information (at six months of age, Bruderer et al. (2015)) and attention (Jusczyk et al., 1990).

1.3.2 Lexical representation

Higher up in the hierarchy of linguistic units are words, which are combinations of sounds within a language that convey specific meanings even when used independently of other sounds[2].

The collection of words known by a person constitutes their *lexicon* or *vocabulary*. To build their lexicon and begin recognising words as independent elements of their language, infants must master two additional linguistic aspects: *phonotactics* and *word segmentation*. Phonotactics refers to the rules that govern sound patterns in a language. These rules are primarily language-specific, though some combinations are forbidden in all languages due to the physical impossibility of realisation. For instance, while words cannot start with the consonant cluster [tl] in French and English, this pattern is legal in Hebrew (Hallé and Best, 2007). As with phonetic perception, mastery of the phonotactics of one's native language seems to occur between 6 and 9 months old (Friederici and Wessels, 1993; Jusczyk et al., 1994), relying once again on distributional cues and language exposure.

Furthermore, phonotactic knowledge can provide helpful clues in word segmentation, which is the ability to separate words from the continuous stream of speech. Indeed, some phonological patterns are constrained to specific positions within a word (Mattys and Jusczyk, 2001). Of course, word segmentation is crucial to building a lexicon, and infants as young as 7.5 months old can succeed at the task (Jusczyk and Aslin, 1995). Besides phonotactics, prosodic cues can also help in the word segmentation process (see §1.3.3). However, the most compelling proposition for how infants manage to segment words and, in turn, build their internal lexicon is the *statistical learning hypothesis*. This hypothesis is based on Transitional Probabilities (TPs), which refer to the probabilities between adjacent syllables in the speech stream. The idea is that a TP between two syllables of the same word will be higher than between two syllables of different words. For instance, the probability between the syllables [mə] and [mi] in the phrase "mummy reads" should be higher than the one between [mi] and [ɹiː]. A series of experiments using artificial languages have demonstrated that infants as young as eight months old can utilise such cues to segment words from the speech they hear (Saffran et al., 1996; Pelucchi et al., 2009). Since then, the statistical learning hypothesis has expanded to encompass more aspects of language acquisition than merely word segmentation, including all mechanisms that rely on the distributional cues discussed earlier (see Saffran and Kirkham, 2018 for a review).

Word segmentation is then just a tiny step away from the learning process of one's language's lexicon. When do infants start recognising words as part of their native language? Infants as young as 4.5 months old prefer to listen to their own names compared to phonotactically matched non-words (Mandel et al., 1995), and both 6 and 7.5 months old prefer to hear stories containing words they were familiarised with earlier in the experiment (Bortfeld et al., 2005; Jusczyk and Aslin,

[2]Although we focus on words in this section, we should note that they are not the smallest meaningful units of a language; they are typically composed of morphemes. While morphemes can contain meaning, they must be combined with other morphemes to form words and be used independently.

1995). Moreover, evidence suggests that infants have already started building their lexicon by eight months, preferring passages containing words they were familiarised with more than two weeks before the test (Jusczyk and Hohne, 1997).

However, as we addressed earlier, this lexicon may contain very fine-grained representations of stored "words" rather than general exemplars, with infants at a certain age not recognising words when spoken by a speaker of a different gender (Houston and Jusczyk, 2000). This is the case with a change of pitch at 7.5 months old (Singh et al., 2008) (but not nine months old) and of speech affect for 7.5 months old (Singh et al., 2004) (but not 10.5 months). Therefore, infants start recognising words from their language very early, but this recognition, in the first months, also depends on non-phonetic variabilities of the acoustic input. Infants start to build proper prototypes around ten months old (Jusczyk and Luce, 2002; Werker and Yeung, 2005), with more familiar (i.e. more frequent) words being recognised first (see Carbajal et al., 2021 for a meta-analysis).

Finally, while other cues have been shown to help word-form recognition and understanding in infants, such as mapping the word with its object or meaning, this drifts away from our focus, and we will not delve into the topic here (but see Samuelson and McMurray, 2017 for a review).

1.3.3 Prosodic representation

We now turn to the case of *prosody* as an aspect of linguistic information. As alluded to earlier, prosody refers to the acoustic aspects of speech that differ from the specific sounds (or phonemes) but rather pertain to the musicality of the speech stream (Wennerstrom, 2001), including rhythm, stress, and intonation. An essential aspect of prosody relates to the perception of *prosodic constituent structures*, or prosodic units, that is, the perception of multiple words grouped in the speech stream due to prosodic cues (e.g. pauses, changes in pitch or duration) (Nespor and Vogel, 2012). These prosodic units can be defined in relation to their prosodic *boundaries* (the edges of the constituent) and their prosodic *prominence* (the head of the constituent), both of which have been shown to correlate with acoustic properties (see Wagner and Watson, 2010 for a review on the topic). In fact, these prosodic constituents have also been shown to correlate well with linguistic units such as words and syntactic structures (Steedman, 1991; Cole, 2015). We refer to this aspect of prosody as 'structural prosody', which is the aspect we focus on in our thesis, in contrast with other types of prosody linked, for example, to affect or focus.

1.3.4 Other linguistic representations

Of course, other aspects of linguistics are required to master a language, aspects which often come later in the language acquisition timeline due to their higher level in terms of linguistic components. Although they are not the main focus of this thesis, we want to briefly give an overview of two of them: semantics and syntax.

Semantics Semantics refers to the knowledge of the *meanings* behind different words, phrases, and sentences. Naturally, this implies that semantic learning occurs concurrently and interdependently with other aspects of language learning, particularly lexical and syntactic learning. However, our focus in this paragraph is on semantic learning at the word level. The issue of how infants know that a specific word they have identified in the speech stream corresponds to a particular referent and thus infer its meaning is known as the *mapping problem* (Quine and Van, 1960). Multiple mechanisms have been proposed to help solve this problem (see Samuelson and McMurray, 2017, but also Wojcik et al., 2022 for why we should shift away from an exclusive focus on mapping in semantic learning). Focusing on the speech input alone, both phonological and distributional cues have been found helpful in semantic learning (Lany and Saffran, 2011). Regarding the acquisition timeline, first evidence of such semantic learning appears in the first year of life, with 6-month-old infants beginning to acquire some meaning for familiar words (Tincoff and Jusczyk, 1999), and 8-month-olds being able to learn new meanings from new objects, provided they are given cross-modal cues (Gogate et al., 2001).

Syntax Moving up from the word level, infants must learn how words are combined in a syntactically correct manner, i.e. they must learn the rules of their language at the sentence level (e.g. word order, phrases combination). This also encompasses learning the correct syntactic and part-of-speech categories of different words (e.g. function vs content words, verbs, nouns). Newborns and 6-month-old English infants have been found able to differentiate function words (e.g. "and", "the", "or") from content words (e.g. "cat", "play", "mummy") (Shi et al., 1999; Shi and Werker, 2001). However, this distinction seems to be due to acoustic and phonological differences (with content words being more salient than function words) rather than a proper understanding of the syntactic categories (Shi and Werker, 2003). Knowledge of function words has also been found to help 11-month-old infants infer word-form recognition of content words (Hallé et al., 2008). Infants seem to infer content words' categories by the end of the second year using different mechanisms, including bootstrapping using function words (Kedar et al., 2006; Shi and Melançon, 2010) and prosodic knowledge (Christophe et al., 2008). Regarding grammatical regularities, Gomez and Gerken (1999) found that one-year-old infants could extract and generalise grammatical rules from an artificial language, and Marcus et al. (1999) extended this to seven-month-old infants with simple algebraic rules.

To conclude, evidence of language learning starts appearing in the first year of life and in a somewhat *simultaneous* manner for the different levels. In fact, the different levels of linguistic knowledge *interact* to aid general learning. Also, while many other cues have proved to be helpful, the use of distributional cues and hence *statistical learning mechanisms* from the sole speech input have been proposed as a potential learning mechanism of language acquisition at these multiple levels (Romberg and Saffran, 2010; Erickson and Thiessen, 2015). We will delve further into this in Chapter 3. We now shift our focus to speech perception in the specific case of dual-language (bilingual) input.

1.4 Early indexical and linguistic development with bilingual input

1.4.1 Defining bilingualism

Researchers have proposed numerous definitions to define bilingualism, ranging from the native proficiency of two languages (Bloomfield, 1933) to the approximate knowledge of another language in relation to the native tongue (Macnamara, 1967). In the context of infancy, bilingualism can be classified into two categories: simultaneous bilinguals (where the infant learns both languages at the same time from birth as they are exposed to a multilingual speech environment) and sequential bilinguals (where the infant is first exposed to their native language during their early development, and then later exposed to a second language either at an early age or later in life). This thesis focuses on the first type of bilinguals: infants who have always been exposed to two languages and therefore commence learning both *simultaneously*.

Even within this group, many factors can differentiate one simultaneous bilingual from another, and we will lay out here some of them. First of all, there is the *strategy* used by the parents, purposely or not, that is, the pattern of how the different languages are spoken at home. In the One Parent One Language (OPOL) strategy, each parent speaks a different language, leading to a direct correlation between the speaker's identity and the spoken language. In contrast, in the opposite "mixed" setup, both parents can speak both languages interchangeably, with even potential instances of intra-utterance code-switching (the mix of two languages within a single utterance) (Kremin et al., 2022). Although this disambiguation is primarily focused on parents and, therefore, input pattern at home (with infants hearing both languages at home), it can also be extended to the case where one language is spoken at home and another one in the outside world, which would resemble an OPOL pattern (different speakers always speak only one language each). Up to now, there has not been strong evidence of either one or the other strategy being preferable for bilingual language acquisition (De Houwer, 2007). Another aspect of variability in the speech input pattern is the *proportion of speech input* for each language infants are exposed to: some infants could be exposed to both languages in a balanced pattern (50%-50%), or one could be more present in their environment (e.g. 75%-25%). Differences in proportion can affect language acquisition of the two languages, with perfectly balanced exposure having the best chance at good language learning in both languages (Thordardottir, 2011; Poulin-Dubois et al., 2013; Place and Hoff, 2011). Moreover, the effect of this input is also dependent on whether the speakers are native or not in the language (Place and Hoff, 2016). In fact, studies involving bilingual homes' naturalistic recordings unveiled considerable variability in the different patterns of input speech bilingual infants are exposed to (Orena et al., 2020). Finally, the *distance* between the languages the infant is exposed to can also affect speech perception and language acquisition, with the direction of the effect differing dependent on the level studied (Sundara and Scutellaro, 2011; Floccia et al., 2018).

Let us put asides all these variabilities within simultaneous bilinguals. Why would their speech perception and processing differ from that of monolinguals? In fact, bilingual infants are faced with multiple challenges unknown to their monolingual counterparts: not only do they need to disentangle and differentiate the two languages they are exposed to, and this potentially without any non-speech cue indicating the language they hear, but they are also provided with less input in each of their native languages than a monolingual who is only exposed to a single language. How do they deal with such challenges? Do their developmental processes follow those of monolingual infants?

1.4.2 Indexical representations

In Section 1.2, we gave a brief overview of the speech perception of monolingual infants on two particular types of indexical information: language and speaker. How does this differ for infants raised in bilingual environments?

The speech perception pattern at the indexical level seems entirely comparable to that of monolingual infants, with bilingual newborns able to discriminate their two languages from birth. However, they show no preference for one over the other (when monolingual newborns show a preference for their native language) (Byers-Heinlein et al., 2010).

Additionally, we already discussed that monolingual infants could discriminate between two rhythmically similar languages provided they are familiar with one. Two-month-old bilingual infants can also do so, even though they are supposedly familiar with both languages, as shown by the study on Spanish and Catalan led by Bosch and Sebastián-Gallés (2001). Similar results were found by Molnar et al. (2014a) with 3.5 months old bilinguals able to discriminate their familiar Basque-Spanish languages, whilst the task proved to be more difficult (but possible) for monolingual infants. Some researchers took the ability to discriminate between their native languages as evidence that infants perceive and process the two languages separately, although this has been argued since (see Byers-Heinlein (2014)). In fact, it seems that this ability for bilingual infants to differentiate their languages gradually improves over time and is dependent on linguistic learning (Byers-Heinlein, 2014). There is also evidence that 20 months old can monitor the language they are attending to (Byers-Heinlein et al., 2017), suggesting that there is indeed some growing separate language categories in older infants.

Interestingly, regarding the processing of speaker information, there seems to be a difference between bilingual and monolingual infants. While nine months old monolingual infants find it more difficult to discriminate talkers of an unfamiliar language than their native language, this difficulty does not exist in bilingual infants (Fecher and Johnson, 2019, 2022). This observation suggests a heightened sensitivity and improved processing of indexical information in bilingual infants (Fecher and Johnson, 2022).

1.4.3 Linguistic representations

Regarding the language acquisition process, simultaneous bilingual infants seem to follow overall the same trajectories and milestones as their monolingual counterparts.

We saw earlier that monolingual newborns could discriminate between phonemic contrasts of most languages but gradually tune to those of their native language, reinforcing their native contrasts but losing the ability to discriminate between contrasts of their non-native language. This tuning, or perceptual narrowing, is also present in bilingual infants. No difference has been found in the trajectory of most phonemic contrasts between monolingual and bilingual infants (Burns et al., 2007; Sundara et al., 2008). Yet, a U-shaped pattern has been found on some specific phonemic contrasts for bilingual infants, with 4 and 12 months old Spanish-Catalan bilingual discriminating contrasts of their native languages but eight months old failing to do so (Bosch and Sebastián-Gallés, 1997; Sebastián-Gallés and Bosch, 2009). However, later studies suggested that this U-shaped pattern was potentially due to the higher flexibility of infants to accept phonemic variations but that they did follow the same linear perceptual narrowing pattern when tested on another paradigm (Albareda-Castellot et al., 2011). Therefore, there does not seem to be any specific cost or delay to the phonetic perception of speech in bilingual infants compared to monolingual infants.

In terms of word segmentation, most studies also show a similar developmental pattern between monolingual and bilingual infants, with young bilingual and monolingual 7 to 10 months old infants preferring to listen to read passages containing words they were familiarised with before from their native language(s)(Bosch et al., 2013). This was the case even under a dual-language setup (the two languages being mixed during familiarisation), although there seems to be an important effect of the proportion of language exposure (Hoff, 2018; Orena and Polka, 2019). This effect of language exposure proportion was also found in phonotactics knowledge. Monolingual and bilingual infants exhibited a comparable ability to distinguish between legal and illegal word endings in their native language; however, this parity was only observed provided the language was dominant in the bilingual infants' environment (Sebastián-Gallés and Bosch, 2002). Moreover, in an artificial language experiment, Kovács and Mehler (2009) showed a potential bilingual advantage in that 12 months old bilingual infants were able to compute and generalise regularities from two distinct artificial grammar speech inputs when monolingual infants were not able to do so, suggesting a more flexible learning mechanism for bilingual infants. Finally, regarding proper word-form recognition abilities without prior familiarisation, monolingual and bilingual infants also seem to share a similar learning timeline, with 11-months-old bilingual infants able to recognise words from their native languages (English and Welsh) (Vihman et al., 2007).

As was already hinted by the language discrimination and word segmentation results presented above, it seems that bilingual infants also follow the same trajectory as monolingual infants in terms of prosody, with equal (if not better) capability to discriminate different prosodic patterns present in their native language (Bijeljac-Babic et al., 2012; Abboub et al., 2015), and comparable development

of prosodic biases (Bijeljac-Babic et al., 2016). They also show similar use of prosody to bootstrap higher levels of linguistic information, with seven months old bilinguals able to use prosodic cues from their languages to solve relevant syntactic word order (Gervain and Werker, 2013).

Hence, at least during the initial months of language development, the learning trajectories between monolingual and bilingual infants appear remarkably similar, with no noticeable delay due to exposure to multiple languages. This effect seems to persist in the higher stages of language acquisition, with similar paths in vocabulary and grammar measures when considering the total vocabulary size of bilingual infants (that is, calculating the total number of words understood across the two languages for bilingual infants) (Hoff et al., 2012). Research into the impact of dual language input on speech perception and linguistic development is still a relatively novel field, and much more work is required to distinguish the mechanisms utilised in each condition. However, the numerous confounding factors we discussed make this an extremely challenging task.

For more complete recent overviews on the topic of early speech perception and language comprehension in bilingual infants, see Sebastian-Galles and Santolin (2020); Höhle et al. (2020); Grosjean and Byers-Heinlein (2018); Byers-Heinlein et al. (2017).

1.5 Computational modelling in the study of speech perception

Computational modelling lies at the heart of cognitive science, providing the possibility to simulate specific cognitive processes through the development of specific algorithms which can recreate the outcomes observed in real life. This method enhances our understanding of the underlying mechanisms within the human mind by concentrating on the algorithmic level of cognitive science analyses, as illustrated by Marr (1983). With this approach, we do not consider the physical realisations of how these processes are implemented in the brain, i.e., their neurological realisation. Instead, our attention is on the algorithms that can transform a given input into an observed output, yielding valuable insights into the cognitive processes at work.

Evidently, the study of language processing as a cognitive process has given rise to numerous computational models, which have, in turn, contributed to novel theories on the topic. We will introduce some of these models in the following chapters. For now, we will discuss how such an approach can help in the general study of language processing, and we will outline some key characteristics we believe are essential for conducting effective research with computational modelling. Specifically, we will rely on the notion of *reverse-engineering*; that is, in the context of cognitive science, the fact of depending on observable inputs and outcomes in an attempt to model and thus understand the internal mechanisms that give rise to said outcomes (Schierwagen, 2012). By working backwards from the observable data, researchers can develop computational models that simulate these cognitive processes and ultimately improve our understanding of human cognition.

Dupoux (2018) has proposed to use this term of reverse-engineering in the context of developmental cognition, and particularly language acquisition. Stemming from the observation that modern-day advances in machine learning are now capable of emulating human capabilities, such as language understanding and generation (Hirschberg and Manning, 2015) or speech recognition (Wang et al., 2019), they suggest harnessing these advances to simulate the infant's language development process.

One of the core propositions of their proposal is the use of *realistic input* to avoid any assumptions in the model that could bias the results. They warn that by using some already processed input, such as phonemes, in simulating language development, assumptions are made regarding the necessity and existence of such input in the speech perception process. However, these assumptions can be debatable, with, for example, the very existence of phonemes which has been questioned (Schatz et al., 2021; McMurray, 2022). Moreover, the sole use of such assumptions can also lead to a so-called bootstrapping problem, with, still in the case of phonemes, phonemes being required for learning words and words for learning phonemes. In that case, which input should we use to model learning of one or the other? Henceforth, the best solution to overcome such potentially biased assumptions is to begin with the most unprocessed input available, in this case, *raw speech*. This is where recent advancements in machine learning and technology can play a significant role, as most recent models are capable of processing such raw input. Of course, raw input is far more complex than components like phonemes in textual form, and often, certain approximations are still made, yet one should aim for the least processing of input possible.

There are different levels of how realistic such data can be, and using raw (or minimally processed) speech in speech and language processing models is already a significant advancement. Yet, in a paper that we co-authored, Lavechin et al. (2022) insisted on the need for even more ecological data in models of infants' language acquisition and processing. They advocated for using long-form recordings as input to such models, that is, recordings directly extracted from an infant's environment with the help of a child-worn microphone. In the work presented in this thesis, we do not go to such extent in input data realism, as many side issues can arise compared to using cleaner speech data (Lavechin et al., 2023), and stick to raw or minimally processed cleanly recorded speech. However, it is important to remember that this is something we should aim for in the future.

Input is not the only observable that needs to be as realistic as possible for computational models to be helpful in a cognitive science reverse-engineering approach. In fact, Dupoux (2018) also advocated for the generation of *realistic outcome measures* achieved through psycholinguistically-inspired tasks. Indeed, by evaluating a model on tasks similar to those given to humans, we can obtain measures that are more easily comparable to human results. One of the best examples of such a psycholinguistically-inspired task for computational models in the study of speech perception is the ABX task, proposed initially as a phoneme discrimination task by Schatz et al. (2013), which we use extensively in the present thesis and which description is provided in Appendix A.

Finally, while we have focused up until now on the use of computational models as a means to gain better insights into the processes underlying human cognition, we would like to pause for a moment to consider another application of such models, as emphasised by the distinction made by Fourtassi (2023) between simulations as a *model* versus as a *tool*. The idea behind the latter is that we can use technological advances to assist us in cognitive science research by automating certain aspects, often in the realm of annotation. For instance, phoneme recognition models can be valuable in annotating large corpora (Adams et al., 2018), as is the case for classifiers of speech type (Lavechin et al., 2020). In Chapter 2, we will present one such "model as a tool" study, and therefore, we believe it is essential to reflect on this distinction.

To conclude, in line with the guidelines laid out above, a good simulation of the indexical and linguistic aspects of speech perception, therefore, requires models that (1) are based on *raw speech*, (2) do *not need labels* or supervision, and (3) can reproduce a *human level of performance* on a variety of speech perception tasks. Moreover, depending on the type of information we want to model, they should learn to represent speech with different levels of granularity: some at a global level, making them good candidates for indexical speech perception modelling and at a more fine-grained level, making them fit for linguistic perception modelling. Recent advances in the speech-processing field have made such models possible, which we will present in the relevant chapters of this thesis.

1.6 General outline and contributions

In this thesis, we merge the disciplines of cognitive science and speech processing, adopting a reverse-engineering approach advocated by Dupoux (2018). We enhance this approach by integrating the most recent advances in machine learning, unveiling and addressing critical questions regarding the interactions between different speech perception processes. Our work offers meaningful contributions to both fields, which we will now present along with the general outline of this thesis.

In Chapter 2, we aim to simulate speech perception at the *indexical level*, focusing on language and speaker identity. We model indexical speech perception through speaker and language discrimination tasks using i-vector models and compare different input patterns, such as language and speaker. We also develop a tool for the automatic computation of acoustic language distance. We investigate how environmental factors such as speech input patterns and language distance can affect indexical speech perception. From this, we conclude that speaker and language information are intrinsically related in global models of speech perception. This work also allows us to lay out some of the core concepts in language-related modelling and showcase how such an approach can help us bring new knowledge otherwise near impossible to test with behavioural work.

In Chapter 3, we shift our focus to *linguistic information*. We design a developmental framework for modelling early language acquisition based on "STELA", a combination of an acoustic self-supervised speech model and a language model.

This developmental framework allows for the generation of models' "developmental curves" for native and non-native speech input. To measure the linguistic knowledge in the different models, we make use of the existing phonetic ABX task but also create novel zero-shot metrics and benchmarks (at the lexical, prosodic, semantic and syntactic levels). Our results indicate that we can simulate parallel and gradual learning at the phonetic, lexical, and prosodic levels using this developmental framework. We further show that learning can arise without sharp linguistic categories and that statistical learning mechanisms are enough to explain these developmental patterns. Finally, we propose criteria and guidelines for learning simulations in modelling language acquisition.

Lastly, in Chapter 4, we delve into how language, as an indexical factor, influences linguistic development, specifically examining the impact of *exposure to multiple languages* on early language acquisition. This expands upon the work from the previous chapters. We carry out an analysis of both indexical and linguistic information in monolingual and bilingual STELA models, in addition to studying language discrimination patterns within these models. We compare monolingual and bilingual models' linguistic developmental curves across the phonetic, lexical, and prosodic levels. Our findings suggest that while bilingual models can distinguish between languages, monolingual models cannot, and this discrimination ability in bilingual models is contingent on the size of the input data. Furthermore, we identify a cost associated with bilingual input on linguistic learning at various levels and show that this cost also depends on the proportion of language exposure.

Chapter 2

Modelling indexical information in speech perception

2.1 Introduction

In this chapter, we attempt to model speech perception of indexical information processing, specifically language and speaker identity. We are also particularly interested in the interaction of such two types of information in speech perception. Before providing an overview of the sections that make up this chapter, we will review recent work proposed in such a type of modelling in the speech processing community.

2.1.1 Modelling indexical speech perception

As we already discussed in the Introduction (§1.1 and 1.2), indexical information, encompassing speaker and language identity, is predominantly accessible at a global granularity, meaning it can be retrieved using holistic information from an entire utterance altogether. Therefore, modelling such information calls for a model which captures the information at the same global granularity, i.e. over an entire utterance.

I-vector models (Dehak et al., 2011) do just that: they capture global information over a complete utterance, incorporating spectral and temporal cues. These models yield i-vectors, which are fixed-length vector representations of entire spoken utterances. These representations measure the degree to which an utterance acoustically differs from a reference distribution of speech used in training the underlying model. Coming from the speech processing field, these models, which are built on Gaussian Mixture Models (GMM), were first proposed as a way to capture speaker identity (which is what the "*i*" in *i*-vectors stands for), and were primarily used for speaker normalisation in speech recognition applications (Dehak et al., 2011). They were also found to capture language-specific information, leading to their use in language identification tasks (Martinez et al., 2012, 2011).

Beyond these speech processing applications, i-vector models have been recently proposed as a model of speech perception of the infant, shown to replicate cognitive effects such as the language discrimination effect (Carbajal et al., 2016; Carbajal,

2018) and the speaker-related Language Familiarity Effect (LFE) (Thorburn et al., 2019). Following the cognitive computational modelling guidelines laid out in Chapter 1 (Section 1.5), they make a good model of speech perception in that they can be trained on speech input only, without the need for further supervision. Additionally, they only need minimal amounts of data to provide good representations. However, it should be noted that minimal preprocessing of the speech signal must be performed, as the i-vector model contains a signal transformation step which transforms the raw speech into a set of acoustic features, usually Mel-Frequency Cepstral Coefficients (MFCC). These features are specifically designed to capture the most relevant aspects of the speech signal while reducing the complexity and dimensionality of the data. Most importantly, they are inspired by the human auditory perception system (Mermelstein, 1976), supporting further the notion that i-vector models can serve as plausible models of speech perception, despite not being trained on raw speech. Finally, although in speech processing applications, training the i-vector model usually includes a supervised step, the Linear Discriminant Analysis (LDA), which requires speaker or language labels, this step is not necessary and is not used in speech perception modelling studies (Carbajal et al., 2016; Carbajal, 2018; Thorburn et al., 2019). We will provide in-depth information on the training and inference steps of this i-vector model in the following sections of the chapter.

Other speech processing models have been proposed to capture information at a global granularity, usually in speaker or language identification applications. X-vectors (Snyder et al., 2018) are extensions of i-vectors, differing in that they are based on Deep Neural Networks (DNNs), and necessitate some speaker or language discrimination step in their training phase, which necessarily require speaker/language labels and therefore make them unfit for modelling early speech perception. Moreover, because they are trained on DNNs, they necessarily require much more input data than the i-vector models. Finally, other neural-network-based speech models have been shown to capture some language and speaker information. However, their segmental and suprasegmental primary focus (rather than global) renders them beyond the scope of this discussion. Instead, we will explore these models in the context of linguistic modelling in Chapters 3 and 4.

2.1.2 Outline of Chapter 2

In this chapter, we will use the abovementioned i-vector model to model psycholinguistic findings related to indexical speech perception. Our investigation will encompass both language and speaker identity, examining factors such as language distance, different input patterns, and the universality of the effect. We will also leverage some of our findings to propose a novel use of the i-vector model, this time "as a tool" following the distinction introduced by Fourtassi (2023).

In Section 2.2, we will focus on infants' language discrimination ability, a key aspect of speech processing that was discussed in Chapter 1. We will replicate results from Carbajal et al. (2016) in modelling this ability. Additionally, we will look into the effect of different bilingual speech input patterns and ask whether

this discrimination necessarily entails unsupervised clustering to separate speech based on languages.

Section 2.3 of this thesis revolves around the Language Familiarity Effect (LFE), a cognitive process that highlights the significant interaction between language and speaker information in speech perception, already briefly introduced in Chapter 1. By modelling this effect, we aim to highlight key contributions of computational modelling in contrast to behavioural work, such as the ability to yield gradual measures. We will also explore how language distance can influence this process.

Finally, Section 2.4 builds upon the concept of language distance introduced in the previous sections. Here, we will demonstrate how the i-vector model can serve as a tool for automatically calculating language distance at the acoustic level, providing a more objective and precise measure of language similarity.

2.2 Modelling bilingual language discrimination and separation using I-vectors

In this first section, we focus on the well-known effect of language discrimination in newborns, which refers to the ability of newborns to differentiate between their native language and a foreign language, as documented by (Mehler et al., 1988). This discrimination effect has also been observed in newborns exposed to two native languages simultaneously during pregnancy (simultaneous bilinguals) (Byers-Heinlein et al., 2010). As we have previously discussed in Chapter 1 (§1.2), this effect holds great significance as it provides initial evidence of infants processing indexical information.

I-vectors models were proposed as a model of the language discrimination effect first by Carbajal et al. (2016). However, while Carbajal et al. (2016) were able to reproduce the language discrimination effect for "monolingual" models (trained on one language), they found that training the model on what they referred to as a "mixed" condition (following the OPOL strategy discussed in Chapter 1, i.e. two different languages spoken by different speakers) yielded lower discrimination scores than the monolingual condition, going against results from psycholinguistics on bilingual language discrimination where no difference was observed between monolingual and bilingual infants. Can we replicate such results using a different dataset? What factors in bilingual input can affect language discrimination? Another question we ask is whether language discrimination necessarily entails language separation (if a model can differentiate two languages, does it mean the representations will automatically be clustered along these two languages?). We try to answer these questions in the work presented in this section.

The work carried out in this section is presented as a paper:

de Seyssel, M. & Dupoux, E. (2020). Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors. In *Proceedings for the Annual Meeting of the Cognitive Science Society 2020*

2.2.1 Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors

Paper summary

As briefly mentioned earlier, the present study has a dual focus. Firstly, we aim to model language discrimination by replicating previous findings from both modelling work and psycholinguistics. Secondly, we ask whether language separation, which we perceive as the capacity to automatically separate speech input into distinct language clusters even without prior knowledge of the number of languages involved, is necessarily consequential to language discrimination.

Language discrimination. To begin with, we are interested in modelling the language discrimination results in both monolingual and bilingual setups and uncovering some of the factors in input patterns that influence such a process. For this purpose, we replicated the experiments from Carbajal et al. (2016) in modelling language discrimination using i-vectors with monolingual and mixed conditions. Language discrimination was measured using an adapted version of the machine ABX task (Schatz et al., 2013, see Appendix A). We used another dataset than the one used in Carbajal et al. (2016), the EMIME dataset (Wester, 2010), which was explicitly proposed as a bilingual dataset with clean recordings of speech in two language pairs: English-Finnish and English-German. With this dataset, we were also able to add a third condition to the monolingual and mixed conditions proposed by Carbajal et al. (2016): a "bilingual" condition, in which the model is trained on utterances from two different languages spoken by the *same* speakers (contrary to the mixed condition where speakers were different for each language). This condition, along with the "mixed" condition, both emulate language learning in a bilingual environment but with two different strategies, a One Parent One Language (OPOL) environment in which the speakers are directly associated with different languages, and a non-OPOL environment in which all speakers can speak the two languages. Finally, we analysed the role of speaker information in this effect. For this purpose, we also introduced a novel technique, anti-LDA, aiming at reducing the information specific to the speaker.

Our results showed that the i-vector model is able to reproduce the language discrimination effect on all three conditions (monolingual, mixed, and bilingual), corroborating with findings from psycholinguistics and supporting the efficacy of the model as a good global model of speech perception. Furthermore, we found an advantage for the mixed condition when the model is given enhanced speaker information. Although this "enhanced speaker condition" could be plausible cognitively (a separate mechanism may help the infant to separate speakers, including visual/sensory information), we cannot directly put these results in perspective with psycholinguistics results as the latter does not allow for graduality in their setup but can only give evidence of presence or absence of the effect. Finally, we found an effect of language distance, with smaller language discrimination scores when the models were tested within the "close" language pair (English-German) than the "distant" language pair (English-Finnish), and this within all three conditions.

Language separation. As a second focus, we seek to answer whether language discrimination necessarily implies language separation in the context of bilingual speech perception. To make this distinction, we define language separation as the ability to differentiate languages without prior knowledge of the number of languages present. Put simply, can unsupervised clustering of speech representations lead to language clusters, and what implications can such separation (or lack of) have for infants' language perception?

Specifically, we looked at whether models trained on bilingual and mixed conditions showed signs of language separation by applying an unsupervised hierarchical clustering algorithm on the same utterances representations the models were trained on, with increasing target numbers of clusters. We measured the language purity of the retrieved clusters, i.e., the extent to which the cluster of utterances contained only utterances from a single language. The experiment showed that only the model with the mixed condition could separate languages, and this was because the clusters separated speakers directly correlated with languages in this condition. All in all, this leads us to conclude that while showing language discrimination abilities, the models do not intrinsically separate languages. This study highlights this differentiation between language discrimination and separation and asks about its consequences on language development research. For the first time, we were able to show that one did not necessarily entail the other. This distinction is crucial for understanding how infants acquire and process multiple languages. Contrary to existing assumptions, our research challenges the notion that language acquisition processes are inherently separate for each language, potentially impacting our understanding of bilingual infants' language learning process.

Symmetrical testing. Finally, we would like to stop a moment and highlight a concept used in this study that we refer to as *symmetrical testing*, a counterbalancing design we believe to be fundamental in language-based computational experiments. When testing a pair of languages, we aim to counterbalance the conditions as much as possible to ensure fairness and unbiased analysis of the overall effect examined (e.g., the effect of nativeness). We define symmetrical testing as the design of testing both languages in both directions (e.g., native speakers of language A tested on language B and native speakers of language B tested on language A). Further analysis can then investigate potential asymmetries between the languages. Importantly, this design prevents us from assuming any language effect that may be attributed to differences between the training datasets or evaluated tasks. Although such a design is wished for in behavioural experiments, it is rarely applied because of the difficulties in gathering the relevant populations. Therefore, a computational modelling approach allows us to address this limitation and achieve symmetrical testing, provided enough data in both languages is available.

Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors

Maureen de Seyssel (maureen.deseyssel@gmail.com)
Emmanuel Dupoux (emmanuel.dupoux@gmail.com)
Laboratoire de Sciences Cognitives et Psycholinguistique, ENS-PSL/EHESS/CNRS/INRIA
Paris, France

Abstract

The language discrimination process in infants has been successfully modeled using i-vector based systems, with results replicating several experimental findings. Still, recent work found intriguing results regarding the difference between monolingual and mixed-language exposure on language discrimination tasks. We use two carefully designed datasets, with an additional "bilingual" condition on the i-vector model of language discrimination. Our results do not show any difference in the ability of discriminating languages between the three backgrounds, although we do replicate past observations that distant languages (English-Finnish) are easier to discriminate than close languages (English-German). We do, however, find a strong effect of background when testing for the ability of the learner to automatically sort sentences in language clusters: bilingual background being generally harder than mixed background (one speaker one language). Other analyses reveal that clustering is dominated by speakers information rather than by languages.

Keywords: language discrimination; language diarization; i-vectors; bilingualism; speaker information

Introduction

Bilingualism is a widespread phenomenon, with the majority of children being born in a bilingual environment. It also appears that being raised bilingual does not result in any particular delay in the language acquisition milestones of children compared to the monolingual peers (Oller, Eilers, Urbano, & Cobo-Lewis, 1997; Vihman, Thierry, Lum, Keren-Portnoy, & Martin, 2007; Petitto et al., 2001), nor to any confusion between the different languages (Petitto & Holowka, 2002; Byers-Heinlein & Lew-Williams, 2013). In fact, infants from both monolingual and bilingual environments seem to be able to discriminate between distant languages from birth (Byers-Heinlein, Burns, & Werker, 2010; Mehler et al., 1988), and rhythmically similar languages as young as 5 months old (Nazzi, Jusczyk, & Johnson, 2000; Bosch & Sebastián-Gallés, 1997). How do they do it? What kind of computational system can achieve language discrimination from the raw signal only? Are there pairs of languages or language backgrounds which would make such discrimination easier or harder? One way of addressing these questions is to use automatic language discrimination techniques as a model of how infants process and discriminate languages.

Related work

I-vectors (Dehak, Torres-Carrasquillo, Reynolds, & Dehak, 2011) are fixed-length vector representations of entire utterances which characterize how much an utterance deviates acoustically from a background distribution of speech used to train the system. These representations are typically used for speaker identification and discrimination (Dehak et al., 2011) but can also represent languages (Martinez, Burget, Ferrer, & Scheffer, 2012; Martinez, Plchot, Burget, Glembek, & Matějka, 2011).

I-vectors based systems have been shown to reproduce key findings in language discrimination experiments: the ability to detect a change in language within a bilingual speaker (language discrimination) (Carbajal, Dawud, Thiollière, & Dupoux, 2016), the distance effect between different language pairs, with close languages being harder to discriminate than more distant languages (Carbajal, 2018), and the ability to discriminate based on prosody (Martinez, Lleida, Ortega, & Miguel, 2013; Carbajal, 2018). However, they also resulted in an intriguing prediction that has not so far been verified experimentally. Notably, Carbajal et al. (2016) found that learners exposed to a mixture of languages have more difficulties to discriminate languages than learners exposed to monolingual backgrounds. These results are counter-intuitive: one would think that having a mixed background should help discrimination not hinder it. They also have potentially important empirical and practical implications. Indeed, if true, they would reveal an undocumented discrimination deficit for infants in a bilingual or mixed background. This is why we wanted to replicate them with more controlled stimuli. Indeed, the initial study used English and Xitsonga recordings from completely different datasets, raising the possibility that results might come from recording-specific properties rather than the language characteristics.

Present work

The mixed background deficit effect found by Carbajal et al. (2016), if true, is important both for theoretical and practical reasons. The current study is devoted to reproducing the original effect, test its robustness, and to more fully understand how language background may affect a learner's ability to discriminate languages.

The first aim of the study is to reproduce the original experiments using more controlled and ecological stimuli. First, to discard potential acoustic artifacts, all recordings used in the experiment were from the same corpus. Second, we used a better counterbalancing design allowing the different con-

ditions to be perfectly comparable, all containing the exact same recordings. Third, the datasets are also more ecological, containing a smaller number of speakers ($N = 12$), simulating an infant's exposure to speech better than the original study containing an implausible number of speakers ($N = 168$).

We also introduce three novelties to explore the robustness of the results. First, we compare two language pairs, one being closely-related (English and German) and the other one being more distant (English and Finnish). Besides enhancing the generalizability of the results, this also allows us to test whether close language pairs are more difficult to discriminate than distant language pairs. Second, along the monolingual and mixed conditions, we introduce a new "bilingual" background condition, with speech from the same speakers speaking in both languages. This new condition simulates an environment in which the infant is exposed to bilingual speech from the same persons (e.g. parents switching constantly between language A and language B). Recent theories in psychology support the idea that such a fully bilingual environment can harm the children's linguistic development and therefore suggests that parents should follow the "One Parent, One Language" (or OPOL) strategy (Genesee, 1989). We are therefore able to investigate whether, in modeling language discrimination, a mixed environment (OPOL) and a fully bilingual environment result in any processing differences. Finally, we analyze the effect of speaker information on language discrimination. This was partly done in Carbajal et al. (2016) by applying a Linear Discriminant Analysis (LDA) to the i-vectors to select a new representation that increases the separation between speaker. Here we add a method which, by taking the orthogonal complement of this LDA representation, does the opposite, i.e. normalizes the representation across speakers.

Finally, to more fully understand how language background could affect discrimination, we test language discrimination in two different ways. The first one is based on psycholinguistic experiments run in infants in the laboratory. In such experiments, infants are presented with sentences from a single bilingual speaker speaking one of their languages, and the reaction of the infant to an unpredictable change in language is measured (through behavioral proxies such as looking time or non-nutritive sucking). Children are said to discriminate the two languages if there is a statistical difference between the set of children who had a switch of language and those who did not. As in (Carbajal et al., 2016), we model this task with a machine-ABX discrimination metric (Schatz et al., 2013). We argue, however, that contrary to the standard interpretation of the discrimination paradigm, a statistical difference between groups is not fully ecological. It does not necessarily indicate that infants can sort out individual utterances from their environment according to their language. In practice, infants are not confronted with a single speaker, but with multiple ones, the decision has to be made sentence by sentence (sometimes words by words in the presence of code switching), and the number of languages that they speak is unknown. This second problem can be defined as a **language diarization task**, which we model as a clustering problem. More precisely, we apply a clustering algorithm to the modeled acoustic space of the different training backgrounds, and look at the extent to which the formed clusters correlate with language labels.

Methods

Materials

We used the EMIME bilingual corpus (Wester, 2010). It is a read speech corpus containing bilingual speech (utterances from two languages recorded by the same speaker) with a 16kHz sampling rate. It was split into two datasets, one with English and Finnish speech, and the other with English and German speech. In each subset, the speakers are bilingual, although English is always their second language. For each language, each speaker reads on average twice the same set of 145 sentences, leading to some sentence repetitions in the train set.

We designed three conditions for each dataset: a *monolingual* one composed of speech from a single language; a *mixed* condition in which the two languages are represented but with each person speaking only one of the two languages; and a *bilingual* condition, containing speech from both languages, uttered by the same speakers. To ensure all conditions are fully comparable, we further split the training sets into subsets. Each subset was used independently, and results were then averaged within the conditions. This way, within each dataset (English-Finnish and English-German), each averaged condition contains the *same speech utterances*. A summary of the different training conditions is presented in Table 1. The average utterance duration is of 4.44 seconds in the English-Finnish dataset and 4.52 seconds in the English-German dataset. The total duration of each training set was therefore between 4h23 and 4h37. Additionally, a test set was created for each dataset, using bilingual speech from the highest-rated accent male and female for each language (2 speakers per set). Each test set is composed of 200 utterances (100 per language).

Pipeline

The following section describes the methodology behind the different steps carried out in the experiment. The whole workflow is applied independently to each training set. Unless stated otherwise, the open-source tool Kaldi (Povey et al., 2011) was used for the different stages of the process.

Feature Extraction Mel frequency cepstral coefficients (MFCCs) features (Mermelstein, 1976) were extracted for all train and test sets, with 13 coefficients (including energy). They were calculated on 25ms speech frames, using 10ms shift. These features, widely popular in speech processing, are based on human perception and are therefore adequate for modeling cognitive processes of speech. Shifted-delta coefficients (SDC) are also calculated. They capture long-distance

Table 1: Summary of train datasets

Dataset	Background	N speakers (N males)	N utterances
English-Finnish	Mono	12 (6)	6910
	English	*6 (3)*	*3480*
	Finnish	*6 (3)*	*3430*
	Bilingual	12 (6)	6910
	subset 1	*6 (3)*	*3454*
	subset 2	*6 (3)*	*3456*
	Mixed	12 (6)	6910
	subset 1	*6 (3)*	*3480*
	subset 2	*6 (3)*	*3430*
English-German	Mono	12 (6)	6960
	English	*6 (3)*	*3480*
	German	*6 (3)*	*3480*
	Bilingual	12 (6)	6960
	subset 1	*6 (3)*	*3504*
	subset 2	*6 (3)*	*3456*
	Mixed	12 (6)	6960
	subset 1	*6 (3)*	*3480*
	subset 2	*6 (3)*	*3480*

information from the neighboring frames, adding some dynamic information to the speech structure.

I-vectors model Following the I-vector model (Dehak et al., 2011), a Gaussian Mixture Model (GMM) is first trained over all speech features of the train set, resulting in a large probabilistic representation of the acoustic space called Universal Background Model (UBM). It can be defined by a supervector m containing the means of all gaussian components. Using factor analysis, the components of highest variability are then projected into a low-dimensional space, the Total Variability space, which is defined by a Total Variability matrix T. An utterance μ can then be defined as $\mu = m + Tv$. The variable v can be used as a fixed dimension representation of μ, and is typically referred to as an i-vector. This process is depicted in Figure 1. We extracted i-vectors for utterances of both the test and train sets. We used a GMM with 128 Gaussians, and dimensionality of 150 for the i-vectors, as these parameters seemed to yield satisfactory results in small datasets (Carbajal et al., 2016).

LDA and Orthogonal Complement Two additional steps were also optionally performed, in an attempt to investigate the effect of speaker information on language discrimination. These supervised methods, applied on the i-vectors, use the speaker labels from the train set to either enhance or diminish the speaker information. They assume that the child is able to identify speakers on an independent basis, and uses this information to either amplify speaker separation or decrease it. To increase speaker information, Linear Discriminant Analysis (LDA) based on the speaker labels is computed on the i-vectors from the train set to estimate a transformation matrix which maximizes the distance between speakers. I-vectors from the train and test sets are then transformed using this matrix, resulting in i-vectors of dimension 11 (*N_speakers* - 1). The opposite stance was also taken by calculating the orthogonal complement of the LDA subspace and then using it to transform the i-vectors. This allowed us to retrieve all the information from the initial i-vector space excluding the information which is in the LDA. By doing this, we remove the information which is used to maximize the distance between speakers, normalizing all speaker information. For clarity reasons, we refer to this extra step as "anti-LDA". The orthogonal complement was calculated using the *scipy* Python package (Virtanen et al., 2019). In cognitive terms, this process would amount to the ability of a child to identify the cues which are speaker-specific, and then removing them from the language identification processes.

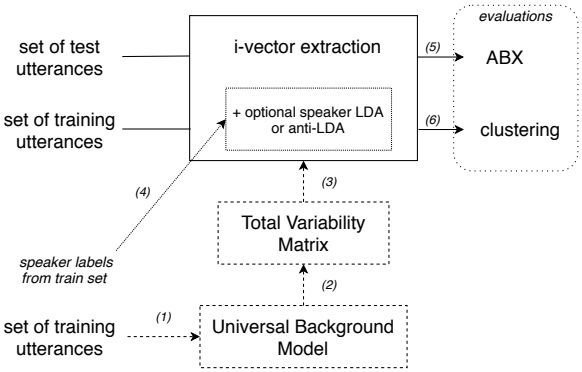


Figure 1: The different stages of the experimental pipeline. In a first training phase, indicated by dotted lines, we construct an i-vector extractor in three steps (1,2,3), followed by an optional step enhancing or reducing the effect of talker variability (4). In the evaluation phrase, indicated by plain lines, we either run a machine equivalent of a discrimination task on novel sentences (5), or cluster the training utterances (6).

Evaluation Methods

Two evaluation methods were implemented, each focusing on one of the language discrimination and language diarization processes.

ABX Scores Language discrimination experiments in psycholinguistics often consist in a first familiarization phase during which the child is exposed to speech from a language A, and an evaluation phase during which the child is presented with two sentences uttered by a new speaker, one of the sentence being from the same language A, and the second sentence being from a novel language B. If the infant can discriminate between the two languages, there should there-

fore be a surprise effect when language B is presented. Although this method is often used as a proxy to assess if children automatically differentiate languages, it is strictly a way to evaluate if children are able to discriminate between two languages, and we therefore restrict our discussion of such results to this particular set-up.

We use the machine ABX paradigm (Schatz et al., 2013) to simulate such a language discrimination experiment. This is done by computing, over the whole set of test i-vectors, multiple triplets of items A, B and X; A and X being i-vectors of utterances sharing the same language and B being an i-vector from an utterance of a different language. For each triplet, the cosine distances of A to X and B to X are then computed. If the distance between A and X is smaller than the distance between B and X, a score of 1.0 is attributed to this triplet, otherwise the score is 0.0. The average of scores across all triplets is then computed, yielding an average ABX score. Perfect discrimination would therefore yield an ABX score of 1.0 (or 100%), as the distance of same-language utterances would always be smaller than the distance of utterances from different languages. To compare our results to psycholinguistics experiment, we compute the triplets within speaker, that is all three items A, B and X will always share the same speaker.

Clustering As a proxy for evaluating whether children cluster multilingual speech from their environment into languages, we apply a clustering algorithm with K clusters to the i-vectors from the multilingual train sets (bilingual and mixed), and evaluate the purity of the formed clusters. If languages are perfectly clustered in the acoustic space, we would expect a purity score of 1.0 when $K = 2$. K-means algorithm was ran 20 times for each K, in the range of $K = 2$ to $K = 20$, yielding an average and standard deviation of the purity scores for each K. We also extended this method to calculate the purity scores on speaker clusters, with $K = 12$ (i.e. the accurate number of speakers). This method was applied to the raw i-vectors from the train sets, as well as the LDA and anti-LDA transformed i-vectors.

Results

ABX scores / Language discrimination

Within speakers ABX scores were computed on the raw, LDA and anti-LDA test i-vectors for each train condition. Results for each dataset are presented in Table 2. Scores in both datasets suggest that the i-vectors successfully allow discrimination between the two languages in all conditions and datasets (no discrimination would yield chance level scores at 50%). As expected, scores in the English-Finnish (different language family) dataset are significantly higher than those in the English-German (same language family) dataset.

There does not seem to be any significant difference with the raw i-vectors between the bilingual, mixed and monolingual conditions, suggesting that the input type in the background's composition does not have an effect on language discrimination of unknown speech. Removing speaker information from the test i-vectors (using the anti-LDA transformation matrix estimated on the train i-vectors) very slightly lowers the discrimination scores in all conditions. In both datasets, however, enhancing speaker information with LDA leads to an increase in ABX scores in the mixed condition, which can be explained by the additional use of speaker information in the language discrimination task, each speaker only corresponding to a single language. It does not yield a stable pattern for the monolingual and bilingual conditions, deteriorating the scores in the English-Finnish dataset but leading to a slight increase in the English-German dataset.

Table 2: Summary of ABX results (in % correct) in both datasets for the different training backgrounds, on the standards, LDA (+LDA) and anti-LDA (-LDA) i-vectors. The scores are calculated within speaker.

		ABX scores		
Dataset	Background	standard	+ LDA	- LDA
English-Finnish	Bilingual	75.1	66.0	74.4
	subset 1	*73.1*	*67.0*	*72.3*
	subset 2	*77.1*	*65.0*	*76.5*
	Mixed	75.5	88.7	73.2
	subset 1	*76.4*	*91.1*	*74.1*
	subset 2	*74.6*	*86.2*	*72.2*
	Mono	73.7	68.0	72.6
	English	*71.8*	*68.9*	*70.4*
	Finnish	*75.5*	*67.0*	*74.8*
English-German	Bilingual	63.3	65.3	62.8
	subset 1	*62.5*	*61.8*	*61.9*
	subset 2	*64.0*	*68.8*	*63.7*
	Mixed	64.2	72.5	62.6
	subset 1	*63.4*	*77.1*	*61.7*
	subset 2	*64.9*	*67.8*	*63.4*
	Mono	63.6	64.9	62.9
	English	*63.1*	*63.3*	*62.5*
	German	*64.1*	*66.4*	*63.2*

Clustering / Language diarization

Kmeans clustering with K clusters (from $K = 2$ to $K = 20$) was applied to the train i-vectors in each mixed and bilingual conditions. Results are presented in Figure 2. The purity score for $K = 2$ is close to 0 in all conditions with, for the raw i-vectors, an average of 0.090 (S = .083) in the mixed condition and of 0.003 (S = .008) in the bilingual condition. This suggests that the acoustic space is not clustered primarily by language.

As presented in Figure 2, with the raw i-vectors, the larger the number of clusters, the larger the difference between the mixed and bilingual conditions, with clusters in the mixed condition getting significantly higher language homogeneity scores. In the mixed condition, language identity is fully correlated with speaker identity, whereas there is absolutely no such correlation in the bilingual condition, each speaker having utterances in both languages. It is therefore probable that

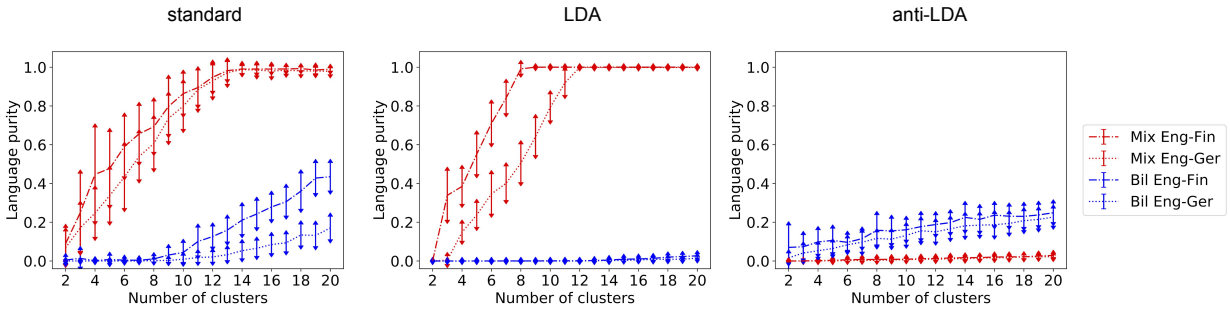


Figure 2: Average language purity as a function of the number of clusters for the different condition, with the standard, LDA and anti-LDA i-vectors. Clustering was done over 20 trials using k-means clustering.

Table 3: Average language purity (in %) using $K = 2$ to $K = 20$ clusters in the different training conditions, with the standard (raw) i-vectors, LDA i-vectors (+lda) and anti-lda i-vectors (-lda).

	English-Finnish			English-German		
Background	raw	+ lda	- lda	raw	+ lda	- lda
Mixed	77.0	83.2	1.3	71.6	68.3	1.0
Bilingual	14.6	0.6	17.0	4.3	0.3	13.6

the acoustic space is clustered primarily by speakers, explaining the highest language purity scores in the mixed condition. Moreover, when the number of clusters is equal to the number of speakers ($K = 12$), clusters in the mixed conditions start reaching perfect purity, while bilingual condition purity scores only start increasing.

The speaker-based cluster hypothesis seems to be confirmed by the results with the LDA and anti-LDA i-vectors. Enhancing speaker information with the LDA favors the mixed condition at the detriment of the bilingual condition, whereas removing this speaker information by taking the LDA's orthogonal complement prevents any language clusters to be formed in the mixed condition, but allows the i-vectors in the bilingual condition to form clusters with language purity scores > 0 when $K < 12$.

It is also worth noting that, in all conditions, the clusters in the English-Finnish dataset have higher purity scores than those in the English-German dataset, suggesting that the language information present in the distant language pair's acoustic space is more discriminatory than those in the close language pair.

We calculated the speaker purity scores for $K = 12$ (the total number of speaker per set). As expected, anti-LDA i-vectors do not cluster speakers at all (M = .011, SD = .003), whereas the LDA i-vectors reach a nearly perfect speaker purity (M = .999, SD = .001). Raw i-vectors also yield very high speaker purity scores (M = .940, SD = .019), suggesting that the standard i-vectors already hold a lot of speaker-specific information.

Discussion

Our experiments successfully replicate the major key findings from previous language discrimination studies, with our model being able to discriminate between languages even with very small exposure. We also found that close language pairs were harder to discriminate than distant ones. However, unlike Carbajal et al. (2016), we found no difference in the standard system between the monolingual and mixed conditions. These results, however, corroborate experimental findings on bilingual children (Byers-Heinlein et al., 2010; Bosch & Sebastián-Gallés, 1997). Although more careful investigation would be required, it is strongly possible that the model in the original study primarily captured recording-specific differences rather than language-specific ones, as the two languages come from distinct datasets. There was also no difference in our raw system with the additional bilingual condition. This would suggest that being exposed to a multilingual environment in which each speaker speaks multiple languages does not hinder the language discrimination process compared to an OPOL-like environment. Although this does not necessarily extend to further processes of language acquisition, such results emphasize the importance of quantitative evidence in supporting psycholinguistics claims.

We found that manipulating the significance of speaker information led to small modulations in language discrimination. Enhanced speaker information slightly improved discrimination in the mixed condition, sometimes to the detriment of the other conditions. Removing this information on the other hand only led to a common very small decrease in discrimination. Such speaker information manipulations assume, in terms of cognition, that infants are able to infer the identity of the speakers in their environments from external modalities (e.g. visual cues). External cognitive processes would then either automatically diminish or enhance speaker-related information when processing speech.

Both theories are also as equally plausible as the standard system, and experimental findings support both ideas: infants are able to recognize speech from their mother (Mehler, Bertoncini, Barriere, & Jassik-Gerschenfeld, 1978) but also fail at strangers voice discrimination tasks when prosody is disturbed (Johnson, Westrek, Nazzi, & Cutler, 2011). Because all three models (raw and with speaker modulation) are reasonable, it would be imprudent to conclude that there are any differences between any of the three exposure conditions.

Findings that bilingual infants are able to discriminate languages (Byers-Heinlein et al., 2010; Genesee, 1989) are not sufficient evidence to assume that they necessarily cluster speech from their multilingual environments into distinct languages. For both mixed and bilingual conditions, and even when manipulating speaker information, the i-vectors used to represent the acoustic space never clustered into two language clusters. This suggests that, even when the number of languages is known, sorting utterances in homogeneous languages clusters is extremely hard. If the number of clusters is increased to the number of speakers, language purity scores increase but only in the mixed condition, corresponding to the intuition of some parents to adopt the OPOL strategy. This indicates that speaker information is not only more salient than the language one, but also that both are intertwined in a way which makes it hard to get them disentangled, even by amplifying or decreasing this speaker information. Nevertheless, it does not mean that these results should be taken as an argument for the OPOL strategy, as there is still no evidence that language separation is a necessary prior to later steps of language acquisition for bilingual children. Hence the underlying question: are children really able to do language diarization? If not, what consequences can it have on language acquisition in bilingual environments?

Another point worth considering in future research is the question of accented speech in bilingual environments. As mentioned previously, the dataset used in the present experiments is composed of non-native bilingual speakers, sometimes leading to the presence of slightly accented English speech. This does not discredit the cognitive inferences made from our results in that even in a family where both parents are native of the two languages, they will often still display accented speech in one language (Major, 1992). However, it would be interesting to replicate the experiments with a corpus solely composed of recordings from native bilinguals, not only to confirm the present results but also to get more insights on the effect of different input types and degrees of accented speech on language discrimination and language diarization.

Acknowledgments

This work was funded in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute), the CIFAR LMB program, and a grant from Facebook AI Research (Research Grant).

References

Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*(1), 33–69.

Byers-Heinlein, K., Burns, T. C., & Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychological science*, *21*(3), 343–348.

Byers-Heinlein, K., & Lew-Williams, C. (2013). Bilingualism in the early years: What the science says. *LEARNing landscapes*, *7*(1), 95.

Carbajal, M. J. (2018). *Separation and acquisition of two languages in early childhood : A multidisciplinary approach*. Doctoral dissertation, Université de recherche Paris Sciences et Lettres.

Carbajal, M. J., Dawud, A., Thiollière, R., & Dupoux, E. (2016). The "language filter" hypothesis: A feasibility study of language separation in infancy using unsupervised clustering of i-vectors. In *2016 joint ieee international conference on development and learning and epigenetic robotics (icdl-epirob)* (pp. 195–201).

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association.*

Genesee, F. (1989). Early bilingual development: One language or two? *Journal of child language*, *16*(1), 161–179.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011.

Major, R. C. (1992). Losing english as a first language. *The Modern Language Journal*, *76*(2), 190–208.

Martinez, D., Burget, L., Ferrer, L., & Scheffer, N. (2012). ivector-based prosodic system for language identification. In *2012 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 4861–4864).

Martinez, D., Lleida, E., Ortega, A., & Miguel, A. (2013). Prosodic features and formant modeling for an ivector-based language recognition system. In *2013 ieee international conference on acoustics, speech and signal processing* (pp. 6847–6851).

Martinez, D., Plchot, O., Burget, L., Glembek, O., & Matějka, P. (2011). Language recognition in ivectors space. In *Twelfth annual conference of the international speech communication association.*

Mehler, J., Bertoncini, J., Barriere, M., & Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception*, *7*(5), 491–497.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143–178.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, *116*, 374–388.

Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by english-learning 5-month-olds: Effects

of rhythm and familiarity. *Journal of Memory and Language*, *43*(1), 1–19.

Oller, D. K., Eilers, R. E., Urbano, R., & Cobo-Lewis, A. B. (1997). Development of precursors to speech in infants exposed to two languages. *Journal of child language*, *24*(2), 407–425.

Petitto, L. A., & Holowka, S. (2002). Evaluating attributions of delay and confusion in young bilinguals: Special insights from infants acquiring a signed and a spoken language. *Sign Language Studies*, *3*(1), 4–33.

Petitto, L. A., Katerelos, M., Levy, B. G., Gauna, K., Tétreault, K., & Ferraro, V. (2001). Bilingual signed and spoken language acquisition from birth: Implications for the mechanisms underlying early bilingual language acquisition. *Journal of child language*, *28*(2), 453–496.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... Vesely, K. (2011, December). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline..

Vihman, M. M., Thierry, G., Lum, J., Keren-Portnoy, T., & Martin, P. (2007). Onset of word form recognition in english, welsh, and english–welsh bilingual infants. *Applied Psycholinguistics*, *28*(3), 475–493.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... Contributors, S. . . (2019, Jul). SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python. *arXiv e-prints*, arXiv:1907.10121.

Wester, M. (2010). *The emime bilingual database* (Tech. Rep.). The University of Edinburgh.

2.3 How entangled are speaker and language information? The case of LFE

To what extent are speaker and language information intertwined in speech perception? How is this interplay reflected in models of speech perception? Results from the previous section suggest that models of speech perception have a hard time differentiating language information from speaker information, with speaker information appearing to be the most prominent, as evidenced by the separation occurring at the speaker level rather than the language level. These results align with research in psycholinguistics, which has uncovered a cognitive effect closely linked to this interweaving between language and speaker information in speech perception, known as the Language Familiarity Effect (LFE). Indeed, as presented in Chapter 1, the LFE relates to the difficulty to differentiate speakers who are speaking a language unknown to the listener and has been found to occur in both adults and infants (Goggin et al., 1991; Johnson et al., 2011; Fecher and Johnson, 2018).

Recently, and following Carbajal et al. (2016) 's approach to using i-vector models as a model of speech perception, Thorburn et al. (2019) proposed to use the same model as a way to model the LFE, using English and Japanese speech. In this chapter, we build upon the work of Thorburn et al. (2019) by exploring the extent to which the i-vector model can capture the cognitive processes underlying the LFE. In §2.3.1, we investigate whether the LFE can be modelled across multiple languages, providing evidence for its universality as a cognitive effect. We also tackle the impact of language distance in this effect and its implications for speech perception. Then, in §2.3.2, we delve deeper into the model and address questions about the stability and reproducibility of our results. We consider both the training data and the model's intrinsic properties in this second analysis.

The first part of this work is presented in the form of a paper:

de Seyssel, M., Wisniewski, G., and Dupoux, E. (2022). Is the Language Familiarity Effect gradual? A computational modelling approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44

2.3.1 Is the Language Familiarity Effect gradual? A computational modelling approach

Paper summary

As introduced above, in this paper, we focus on the question of graduality in the Language Familiarity Effect. Specifically, we aim to address one of the limitations of psycholinguistic designs that evaluate language effects, such as the LFE, which often yield only binary scores and therefore do not allow for comparing effect sizes across differing conditions. Even when effect sizes are analysed, comparability between behavioural studies is often hindered by uncontrollable design variables, such as different languages and populations (Levi, 2019). We argue that computational

models can help solve this issue, as they can (1) entirely control the conditions to ensure only the tested variable differs and (2) yield gradual measures rather than binary scores. This directly relates to the concept of symmetrical testing introduced in the previous section, as we could independently compare language pairs and input patterns.

In this paper, we not only replicated the results of Thorburn et al. (2019) on new speech data and languages, supporting the i-vector model's effectiveness as a model of the LFE and holistic speech perception, but we also demonstrated the model's ability to yield gradual measures of the effect. To assess the LFE, and similarly to the approach we took in the previous section, we used the ABX machine task (Schatz et al., 2013), focusing, this time, on speaker information, and ensured we followed symmetrical testing, which was also done in Thorburn et al. (2019). We first confirmed the model's ability to provide gradual measures by replicating literature findings that the LFE decreases when one language is accented in the other (Goggin et al., 1991). For this purpose, we used the same dataset as in Section 2.2. In the second part of the paper, we replicated this effect for 36 language pairs using a subset of the CommonVoice dataset (Ardila et al., 2019), which provides read sentences in multiple languages. Our results suggest that the effect may be universal, with positive LFE scores on average. Additionally, our findings suggest an effect of language distance, as we observed in the language discrimination effect discussed in the previous chapter: language pairs belonging to the same language family had lower LFE scores than those that did not.

Is the Language Familiarity Effect gradual ? A computational modelling approach

Maureen de Seyssel[1,2] **(maureen.deseyssel@gmail.com)**
Guillaume Wisniewski[2] **(guillaume.wisniewski@u-paris.fr)**
Emmanuel Dupoux[1] **(emmanuel.dupoux@gmail.com)**
[1] Cognitive Machine Learning (ENS–CNRS–EHESS–INRIA–PSL Research University)
[2] Université Paris Cité, CNRS, Laboratoire de Linguistique Formelle
Paris, France

Abstract

According to the Language Familiarity Effect (LFE), people are better at discriminating between speakers of their native language. Although this cognitive effect was largely studied in the literature, experiments have only been conducted on a limited number of language pairs and their results only show the presence of the effect without yielding a gradual measure that may vary across language pairs. In this work, we show that the computational model of LFE introduced by Thorburn, Feldman, and Schatz (2019) can address these two limitations. In a first experiment, we attest to this model's capacity to obtain a gradual measure of the LFE by replicating behavioural findings on native and accented speech. In a second experiment, we evaluate LFE on a large number of language pairs, including many which have never been tested on humans. We show that the effect is replicated across a wide array of languages, providing further evidence of its universality. Building on the gradual measure of LFE, we also show that languages belonging to the same family yield smaller scores, supporting the idea of an effect of language distance on LFE.
Keywords: language familiarity effect ; computational modelling ; i-vectors

Introduction

The Language Familiarity Effect (LFE) is a cognitive effect observed in language processing, according to which people are better at discriminating speakers who speak in their native language, compared to speakers of another unfamiliar language (Goggin, Thompson, Strube, & Simental, 1991; Johnson, Bruggeman, & Cutler, 2018). Two views are commonly proposed to explain the LFE (T. K. Perrachione, 2018). According to the Phonetic Familiarity hypothesis, the lack of familiarity with the foreign language's lower levels of linguistic characteristics (rhythm, phonetics, acoustics) is enough to explain the effect. For proponents of the Linguistic Processing hypothesis, on the other hand, the effect is in great part explained by the lack of understanding (due to knowledge of lexicon and syntax). However, even in this second view, the role of low-level linguistic features is accepted (Bregman & Creel, 2014; T. Perrachione, Dougherty, McLaughlin, & Lember, 2015).

Methodological issues Although numerous experimental studies run in humans (henceforth behavioural studies) found evidence of the effect, the lack of systematicity makes it hard to compare the results directly (Levi, 2019). First, the evaluation tasks used to assess the presence of LFE differ from one study to the next, ranging from identification tasks (voice line-up) to discrimination tasks (AX task). Critically, a same language pair evaluated on different tasks can yield opposite results regarding the presence of LFE (Levi, 2019). Another source of variability comes from the initial testing conditions. Two setups are principally used, the "1 Group 2 Languages" (or 1G2L) and the "2 Groups 1 Language" (or 2G1L). In the first, most common condition, participants are all native speakers of the same language and are evaluated on their ability to discriminate between speakers in both their native language and a second unfamiliar language. In the 2G1L condition, two groups of participants, native speakers of languages A and B, are tested on only the same language A.

One more issue raised from behavioural studies is the restricted number of language pairs tested. Although this effect has been found over multiple language pairs, leading to qualifying the effect of universal (see Levi (2019); T. K. Perrachione (2018) for reviews), it turns out that only a small number of languages was tested. For instance, only a handful of studies test a language pair that does not contain English (Köster, Schiller, et al., 1997; Johnson, Westrek, Nazzi, & Cutler, 2011; Perea et al., 2014). In order to get more robust evidence of the universality of the effect, a wider array of languages must be tested.

LFE as a gradual effect Because of how the LFE has been evaluated behaviourally, it has mainly been presented as either present or absent. Very few attempts have been made at looking at the effect gradually: T. K. Perrachione (2018) computed effect sizes in LFE experiments, but they are hardly comparable due to differences in setup. Having a systematic gradual measure would allow deeper analyses of specific conditions. Hence, we could directly compare different language pairs or different atypical populations on the LFE. Some studies looked into the role of language distance in LFE. For example, (Levi, 2019) showed, in an extensive literature review, that language pairs both from the same and different rhythmic classes could yield an LFE. However, this does not allow a gradual ranking of language pairs. A few studies directly tested multiple languages with the same population, permitting ranked comparisons, assuming that phonologically similar languages yield better performance in speaker identification. However, these studies never tested more than three languages at a time, and conflicting results were found. Köster et al. (1997) and Zarate, Tian, Woods, and Poeppel (2015) results confirmed this assumption (testing Chinese, English and

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

Spanish on German adult listeners and English, German and Mandarin on English adult listeners, respectively). However, no difference between phonologically similar (English and Dutch) and dissimilar (English and Mandarin) were found in infants by (Johnson et al., 2011), thus a need for further studies on the question. Hence, there is a need for a way to test and rank in a systematic manner a large number of language pairs, varying in language similarity.

Additionally, having a gradual measure of the LFE can help analyse finer granularity than that of language differences. An existing example is the case of accented speech. Indeed, some studies found that, for a language pair A-B, if the test stimuli in language A are spoken by native speakers of language B, and therefore accented in B, the LFE can be reduced (Goggin et al., 1991), and even totally cancelled (Goldstein, Knight, Bailis, & Conover, 1981). This suggests that the LFE can be modulated by how heavily accented the speech is and, to a further extent, that acoustically similar dialects should give rise to smaller LFE, corroborating the idea that language distance plays a role in this cognitive effect. These results also show that the effect is gradual, emphasising the need for a gradual measure.

I-vectors as a model of LFE Recently, Thorburn et al. (2019) were able to computationally model the LFE using i-vectors (Dehak et al., 2010), an unsupervised algorithm that allows to compute a representation of whole speech utterances. Computational modelling of LFE can help circumvent some of the methodological problems presented earlier, and we believe it can help compute a systematic, gradual and comparable measure of the effect.

I-vectors models, typically used for speaker-identification applications in speech processing, consist in training a Gaussian Mixture Model on speech features of the train sets utterances to define a new representation of the acoustic space. Then, projecting the components of highest variability onto a lower-dimensional space, we can create a new representation of speech (the i-vector) for all utterances from the train set. By extension, we can predict representations for novel utterances based on this operation. Furthermore, because computed at the utterance level rather than at a finer frame level, we capture the acoustic information that is representative of the utterance as a whole, such as speaker or language information. The lack of time-dependencies in the representation means that only low-level features of linguistics (rhythm, some phonology) are captured. Because of that, and the fact that training such models only necessitates a small amount of input data, the approach has mainly been proposed in models of infants' speech perception. Still, we believe i-vectors can equally model some aspects of adult speech perception that do not require access to higher levels of linguistics.

I-vectors were first proposed in the context of speech perception by Carbajal, Dawud, Thiollière, and Dupoux (2016) as a model of language discrimination. Still modelling language discrimination processes, de Seyssel and Dupoux (2020) showed that they also capture speaker information, even without relying on any supervised components usually present in speech processing applications of i-vectors. Because the LFE depends on both language and speaker information, the i-vector model has the necessary attributes to model it, and this is indeed what showed Thorburn et al. (2019). In their paper, the authors focused on the English-Japanese pair. They showed that the scores from a speaker discrimination task carried out on i-vector representations extracted on both languages were significantly better when the i-vectors were extracted using a model trained on the same language than on another unfamiliar language, effectively replicating the effect found in humans.

Contributions

One underlying contribution of this paper is a replication of the computational approach of Thorburn et al. (2019) on new speech stimuli and language pairs. This reinforces the validity of the i-vector approach to model the LFE. Most importantly, and as the main contribution, we inquire about the capacity of the approach to yield a gradual, comparable measure.

In a first experiment, we look into *reproducing the human findings* according to which accented speech minimises the LFE compared to native speech. Precisely, we replicated an experiment from (Wester, 2012) testing LFE on two language pairs that are always accented in one of the languages. This first experiment also allows us to directly *compare* a close language pair to a distant one. We expect three primary outcomes : a replication of the LFE on the native condition; an LFE is smaller or non-existent in the accented condition compared to the native condition; an LFE is smaller in the close language distance pair than in the distant one. Such results would corroborate with the idea of a gradual effect of the LFE, which we could measure using the i-vector approach, allowing for systematic comparisons.

Findings from the first experiment then lead us to generalise the method to *additional language pairs*. In the second experiment, we evaluate the LFE on 36 language pairs, with many that have never been tested in humans. We can then systematically (1) test and (2) compare the LFE on a large number of languages pairs in a way that would be impossible behaviourally. We expect the LFE to replicate on most of these pairs and to find an effect of language distance.

General methods

The methods presented here are common to both experiments. We use as an example a setup in which we want to evaluate the LFE on a language pair A, B. For each language, we have a set of speech utterances, split between a train and a test set. The former is used to train the models, and the latter is the evaluation stimuli used to test the presence of LFE.

Training pipeline

We first extract Mel Frequency Cepstral Coefficients (MFCCs) (Mermelstein, 1976) for all utterances (train and test), with 13 coefficients including energy, along with double

delta coefficients. We also include pitch information through computation of the fundamental frequency, as it is thought to be relevant in language discrimination (Lin & Wang, 2005).

We then train two i-vector models using the MFCCs from the train sets, one on language A (model A) and one on language B (model B), following the approach first proposed in Dehak et al. (2010). The only difference with the original i-vector approach is that we do not carry out a Linear Discriminant Analysis (LDA), originally aiming at maximising the distance between speakers and/or language (Kanagasundaram, Vogt, Dean, Sridharan, & Mason, 2011; Dehak, Torres-Carrasquillo, Reynolds, & Dehak, 2011). Indeed, as in previous studies using i-vectors as models of speech perception (Carbajal et al., 2016; de Seyssel & Dupoux, 2020; Thorburn et al., 2019), we ensure that the pipeline is unsupervised and therefore better suited to cognitive models. Finally, we extract i-vector representations from the two test sets on both models. That is, we extract i-vectors using model A on tests sets A and B, and similarly for model B. This leaves us with four sets of i-vectors: language A trained on A, language A trained on B, language B trained on A and language B trained on B.

The models are trained with 128 (2,048) Gaussians and i-vectors of dimension 150 (400) in Experiment 1 (Experiment 2). The difference in parameters between the two experiments is explained by the larger number of speakers in the training sets of Experiment 2. Feature extraction, models training and i-vectors extraction were conducted using the Kaldi toolkit (Povey et al., 2011).

Evaluation

Following Thorburn et al. (2019), we first use a machine ABX task (Schatz et al., 2013) to evaluate the capacity of a model to discriminate speakers. In this setup, we create triplets of three utterances from the same language: a, b and x, with a and x being pronounced by the same speaker and b by a different speaker. If the Euclidean distance between the representations (i.e. i-vectors) of utterances a and x is larger than the distance between the representation of b and x, we consider that the model did not manage to discriminate between the speakers, and we count an error for this specific triplet. The ABX error score is the error rate estimated over all possible triplets in the test set.

This framework can be extended to evaluate the LFE by comparing the capacity of a model to discriminate between speakers in a 'familiar' condition, in which the representation is learnt and tested on the same language, to its capacity to discriminate between speakers in an 'unfamiliar' condition, in which test utterances are in a different language than the ones used to train the model. More precisely, we define the LFE score as follows: for a language pair (A,B), we compute the ABX error rates for all four conditions (Ts stands for *test* and Tr for *training*): $\mathrm{Ts}(A)_{\mathrm{Tr}(A)}$, $\mathrm{Ts}(A)_{\mathrm{Tr}(B)}$, $\mathrm{Ts}(B)_{\mathrm{Tr}(B)}$ and $Ts(B)_{Tr(A)}$, where $Ts(A)_{Tr(B)}$ corresponds to the evaluation of the ABX error rate on the language A when the representation has been trained on language B. We then average the scores in the 'familiar' condition ($\mathrm{Ts}(A)_{\mathrm{Tr}(A)}$ and $\mathrm{Ts}(B)_{\mathrm{Tr}(B)}$, the test and train sets being matched in language), and the scores in the 'unfamiliar' condition ($\mathrm{Ts}(A)_{\mathrm{Tr}(B)}$ and $\mathrm{Ts}(B)_{\mathrm{Tr}(A)}$ in which train and test languages are different). The LFE score is defined as the relative percentage increase from the 'familiar' to the 'unfamiliar' condition:

$$\mathrm{LFE} = \frac{S_{\mathrm{diff}} - S_{\mathrm{same}}}{S_{\mathrm{same}}} \tag{1}$$

where:

$$S_{\mathrm{same}} = \frac{\mathrm{Ts}(A)_{\mathrm{Tr}(A)} + \mathrm{Ts}(B)_{\mathrm{Tr}(B)}}{2} \tag{2}$$

$$S_{\mathrm{diff}} = \frac{\mathrm{Ts}(A)_{\mathrm{Tr}(B)} + \mathrm{Ts}(B)_{\mathrm{Tr}(A)}}{2} \tag{3}$$

We use a Two-Sample Fisher-Pitman Permutation Test with Monte-Carlo sampling to test whether this effect is significant. The score is significant if discrimination scores in the S_{same} and S_{same} groups are significantly different. A positive significant LFE score reflects an effect of language familiarity, with a higher ABX error rate in the non-familiar condition than in the familiar condition.

Because we are looking at the LFE on the language pair symmetrically, that is analogous to a '2 groups 2 languages' (or '2G2L') approach, (two groups of participants, native in two different languages, are tested on both languages). Hence, we are controlling for any biases due to a specific training set yielding better speaker discrimination performance, and thus singling out the actual LFE process. This is a more robust evaluation setup than what is commonly done in behavioural work, where LFE is looked into from the perspective of a single language only.

Experiment 1: LFE and accented speech

First, we focus on two language pairs, English-Finnish and English-German. For each of these pairs, we compare a "native" setup, where all tested speakers are native in the languages, and an "accented" setup, with English utterances being spoken by non-native speakers, hence Finnish accented or German-accented.

Materials

We retrieved audiobooks in English, German and Finnish from the LibriVox project[1] using the `Libri-Light` tools (Kahn et al., 2020), and used a Voice Activity Detection model (Lavechin, Bousbib, Bredin, Dupoux, & Cristia, 2020) to segment speech. We then created for each language a 10 hours training set, balanced equally between 10 speakers.

The test sets were built from the EMIME bilingual corpus (Wester, 2010), which contains English, German and Finnish read speech uttered by native speakers, as well as English spoken by German and Finnish speakers, and is therefore accented. We built five different test sets: native Finnish, native German, native English, Finnish-accented English and

[1] https://librivox.org

Table 1: Summary of test sets in Experiment 1.

Language	Accent type	N speakers (N male)	Mean (SD) utt dur (in s)
English	native	12 (6)	3.21 *(1.04)*
	Finnish	12 (6)	4.37 *(1.48)*
	German	12 (6)	4.56 *(1.52)*
Finnish	native	12 (6)	4.6 *(1.29)*
German	native	12 (6)	4.6 *(1.32)*

German-accented English. Each test set is balanced equally between 12 speakers and has an average duration of 25 min (348 utterances). See Table 1 for more information.

Results

We calculated the LFE score on four language pairs following the procedure presented in the General Methods: native English and native German; native English and native Finnish; German-accented English and native German; Finnish-accented English and native Finnish. We refer to the two first pairs as *native* and the two last as *accented*.

Table 2: LFE scores on the native and accented conditions for both language pairs in Experiment 1. Significance was estimated using a Two tailed Paired Fisher-Pitman Permutation Test with Monte-Carlo sampling (*: p<.05)

Language Pair	LFE (%) native	LFE (%) accented
English - Finnish	**+19.21***	-8.62
English - German	**+10.77 ***	-1.1

The first thing to notice from Table 2 is that both language pairs yield a *significant* LFE score in the native condition (the familiar models yield better discrimination scores than the unfamiliar ones, p<.05 in both pairs), giving further support to the i-vector approach as a good model of LFE. Moreover, the LFE score is higher in the English-Finnish pair than in the English-German pair, suggesting that the distance between languages could modulate the LFE.

In the accented condition, there is no longer a significant difference between the familiar and unfamiliar models' scores on language discrimination, and this on both language pairs. Hence, whilst the LFE scores indicate the effect was present in the native conditions, it is no longer the case in the accented condition, that is, when one of the two languages is uttered with an accent from the other language. These results, which replicate the behavioural findings from Wester (2012) as well as previous studies on accents, suggest that we can use the i-vector models to obtain a gradual measure of the LFE.

Experiment 2: Testing LFE on many language pairs

Results from the first experiment not only validate further the i-vector approach as a good model of the LFE, but they also suggest that the resulting measure is gradual and thus comparable. Furthermore, they suggest that there might be an effect of language distance on LFE.

In this second experiment, we generalise the experiment to many language pairs, including pairs that have not been tested on humans. This allows us to 1) verify the universality of the effect, 2) make use of the gradual measure to compare pairs with varying language distances.

Materials

We used stimuli from the `CommonVoice 6.1` (CV) corpus (Ardila et al., 2019), which gathers read speech from a large number of languages. We selected nine languages (those for which we had enough data) and generated, for each of them, training sets of 15 hours split between 60 speakers and test sets of 30 minutes split between 20 speakers (see Table 3 for the complete list). The high number of speakers is closer to the setup proposed by Thorburn et al. (2019) than in the first experiment and ensures more variability in the training set, leading to a more robust model.

Table 3: Summary of languages in Experiment 2. Train sets have an average duration of 15 hours (60 speakers) and test sets have an average total duration of 30mn (20 speakers).

Language	**ISO**	**Avg utt dur (s)**	**Family**
Catalan	cat	5.10 *(SD=1.82)*	indo-european
Welsh	cy	4.52 *(SD=1.67)*	indo-european
German	deu	4.42 *(SD=1.50)*	indo-european
English	eng	4.81 *(SD=1.74)*	indo-european
Farsi	fas	3.80 *(SD=1.42)*	indo-european
French	fra	4.78 *(SD=1.51)*	indo-european
Italian	ita	5.35 *(SD=1.73)*	indo-european
Kabyle	kab	3.38 *(SD=1.23)*	afro-asiatic
Kinyarwanda	kin	5.14*(SD=1.80)*	niger-congo

Results

Models were trained on the nine languages, and evaluation was run on all possible language pairs, yielding 36 LFE scores.

Speaker ABX scores averaged across all pairs are presented in Figure 1, and detailed LFE scores are available in Table 4. Speaker discrimination scores are overall significantly higher in the 'familiar' condition than in the 'unfamiliar' one, with a mean LFE of 13.78 (significance was calculated using a 95% confidence interval with bootstrapping on languages, with 10,000 permutations, CI = [2.71–18.55]). These results corroborate the idea of a universal LFE that can be expected on language pairs that were not tested on humans.

However, the difference is not systematically significant in every pair, with one language pair (German-Welsh) yielding a significant inverse LFE.

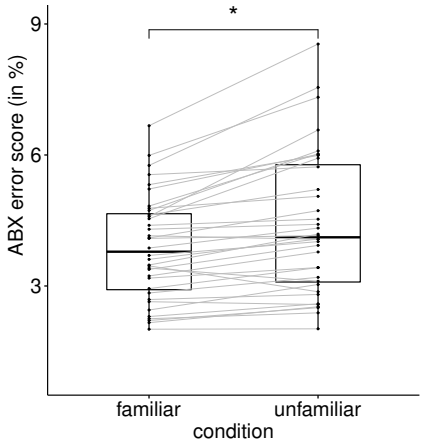


Figure 1: Speaker ABX error scores averaged across the 36 language pairs in the CV dataset. LFE score = 13.78. The asterisks on top illustrate the significance level (*, 95% CI).

We then divided the language pairs into two groups: the 'same family' and the 'different family', based on whether the languages in the pair belong or not to the same language family (as defined by the WALS typology (Dryer & Haspelmath, 2013), see Table 3). As shown in Figure 2, the LFE scores from 'same family' pairs (M=21.46, SD=9.62), N=15 are significantly lower than the 'different family' pairs (M=6.13, SD=9.47, N=21) (significance was tested using a 99% confidence interval with bootstrapping on language within family with 10,000 permutations, CI = [7.28,29.07]).

Discussion

In the first experiment, we successfully replicated results from Thorburn et al. (2019) by showing that the i-vector approach yields a significantly positive LFE score on two new language pairs (native condition). Moreover, we further validated this model by replicating another behaviour observed in humans, that is, the fact that the LFE can be diminished or cancelled with accented speech (Goldstein et al., 1981; Thompson, 1987). Specifically, the stimuli we use in our 'accented' conditions come from the same dataset as in (Wester, 2012), in which they also found no significant effect of LFE in humans, neither in the English-German nor in the English-Finnish pair.

Differences in the LFE between accented and native speech support the idea of the LFE as a gradual effect that can be modulated by languages variations. Current psycholinguistic setups make such changes hard to capture in humans, but our results suggest that with the i-vector approach, we can capture and grade such changes. Being able to capture this granularity allows one to investigate the role of different variables

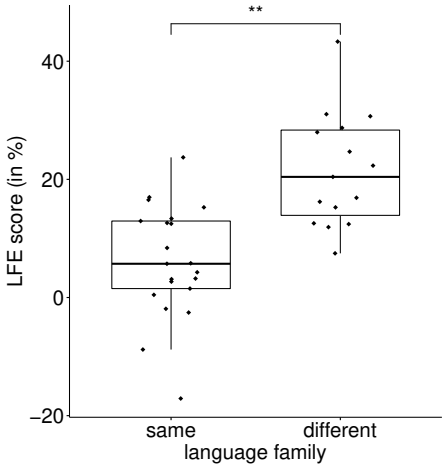


Figure 2: LFE scores averaged across the 'same family' and 'different family' conditions. The asterisks on top illustrate the significance level (**, 99% CI).

on the LFE, with the most obvious one being that of language distance. As discussed earlier, the role of language distance is hard to analyse in behavioural experiments. However, results from Experiment 1 do suggest an effect, with the close distance language pair (English-German) yielding a lower LFE than the distant language pair (English-Finnish).

In the second experiment, we tested the model on a larger number of language pairs, of which many have never been tested on humans. We could compute comparable LFE scores for each of the 36 language pairs. We note that the LFE was present overall (across all pairs together), validating the approach again. However, not all language pairs yielded a significant effect. Multiple things could cause this: these specific pairs might not actually yield an LFE in humans, or the LFE might be too small in humans, and the model is not sensitive enough to capture it. Regardless, we should replicate the experiment behaviourally, especially if the pair had not been tested on humans before. Finally, there might also be specific biases in the stimuli resulting in an absence of LFE. For example, two pairs, Welsh-English and Welsh-German, actually yielded a negative LFE score: the unfamiliar condition yielded better discrimination scores than the familiar one. However, it is likely that a large part of the Welsh utterances in the CV corpus was pronounced by English native speakers, which, in light of the previous results on accented speech and LFE, could partially explain the lack of LFE.

Interesting results arose when the language pairs were divided into two groups: those in which both languages of the pair belong to the same family and those in which they do not. There was a significant difference in LFE scores between the two groups, with the same family language pairs yielding much lower LFE scores than the different family pairs. This corroborates with results from Experiment 1, suggesting that

Table 4: LFE scores for all possible CommonVoice language pairs. Two tailed Paired Fisher-Pitman Permutation Test with Monte-Carlo sampling with Bonferroni correction (*: p<.05; **: p<.005)

	ca	**cy**	**de**	**en**	**fa**	**fr**	**it**	**kab**
cy	8.39							
de	12.49	*-17.10***						
en	5.80	-8.81	0.45					
fa	**16.53***	1.51	4.27	-2.55				
fr	**23.72****	12.95	**13.36****	**16.97***	**12.63****			
it	3.11	3.21	-1.91	**15.27****	5.71	2.68		
kab	**20.42****	12.42	**16.88****	7.48	**11.92****	**16.21****	**12.55****	
rw	**24.69****	**22.32****	**43.32****	**30.68****	**15.26****	**28.71****	**27.97****	**31.04****

closer languages lead to a smaller LFE. The possibility for the LFE to be affected by language distance raises many interesting points regarding the cognitive processes behind this phenomenon. Commonalities and differences between languages can occur at various linguistic levels, but the i-vector approach only focuses on low-level cues (mainly phonology, prosody and phonotactics). Yet, it suggests an effect of language distance, supporting further the idea that the LFE is largely due to the familiarity with the phonetics and phonology of the foreign language, as proposed by the *phonetic familiarity* hypothesis (Fleming, Giordano, Caldara, & Belin, 2014; Orena, Theodore, & Polka, 2015). Still, the distance between two languages at higher linguistic levels may also enhance the phenomenon.

To conclude, although it is not guaranteed that the language distance effect suggested by the model is equally present in humans, our results give us strong incentives to investigate this. This should be done in a systematic setup allowing for direct comparison of language pairs, potentially by designing a wide-scale online speaker discrimination study in many languages. Still, it is yet unclear whether we can obtain a fine enough gradual measure in humans.

LFE score stability One of the central issues in computational modelling is the impact of data on the models. Here, we consider that the i-vector-based LFE score is stable in that it is not affected by changes in train or test sets. This is why we can confidently compare two conditions (languages, setup, number of speakers, recording condition) as long as all other factors are controlled for. However, we have not tested whether the results are prone to variations based on the train and evaluation stimuli, for example by running the same experiments on a new train or test sampled from the same original dataset. Only if the results are stable can we fully validate the approach. This stability aspect also raises the question of how representative of a language the training sets are. Indeed, while humans have had years of being exposed to their native language, which allows them to build an internal language prototype, we only train the models on a few hours of data in the current approach. Despite being a considerable advantage in data collection, it also increases the probability for the model's prototype to be biased. In the second experiment, we purposely used a high number of speakers and diversity in the recording setup, and we recommend any further work to follow this lead.

Finally, we would like to address the fact that the i-vector model was initially proposed as a model of infant perception (Carbajal et al., 2016; de Seyssel & Dupoux, 2020), and used as such in the scope of the LFE to support the evidence that the effect only requires low-level linguistic knowledge and is present in infants (Thorburn et al., 2019). Here, we focused on the LFE in general, without restriction to a specific age group. Indeed, although the present model only requires knowledge of the acoustics of the language, it still reproduces behavioural results found in adults, and we can only assume that adding higher up knowledge that makes use of language's understanding will only reinforce the effect found here. Therefore, even if the model could be completed by adding such features, the present approach can still be seen as a model of LFE in both infants and adults and has the advantage of only requiring very little data.

Conclusion

To conclude, our results further validate the i-vector approach as a good model of the LFE by replicating Thorburn et al. (2019) on novel languages and replicating human experiments on accented speech. These results on accents also suggest that the effect can be modulated, hence gradual. The i-vector model allows computation of gradual LFE scores, meaning that we can then directly compare different conditions. We showed further evidence of the universality of the effect by evaluating it on a large number of pairs systematically, in a way that can only be done computationally. We also found an effect of language distance, with larger LFE yielded when the two languages are dissimilar. These results should be replicated with humans in a setup allowing systematic evaluation, and attention should be given to the design of such an experiment. Finally, a more thorough analysis of the stability of the model depending on the training set could be done to ensure that the scores are fully stable and that different data would not give different scores, skewing the comparisons.

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., ... Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning. *Cognition*, *130*(1), 85–95.

Carbajal, M. J., Dawud, A., Thiollière, R., & Dupoux, E. (2016). The "language filter" hypothesis: A feasibility study of language separation in infancy using unsupervised clustering of i-vectors. In *2016 joint ieee international conference on development and learning and epigenetic robotics (icdl-epirob)* (pp. 195–201).

Dehak, N., Dehak, R., Glass, J. R., Reynolds, D. A., Kenny, P., et al. (2010). Cosine similarity scoring without score normalization techniques. In *Odyssey* (p. 15).

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., & Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association.*

de Seyssel, M., & Dupoux, E. (2020). Does bilingual input hurt? a simulation of language discrimination and clustering using i-vectors. In *Cogsci 2020-42nd annual virtual meeting of the cognitive science society.*

Dryer, M. S., & Haspelmath, M. (Eds.). (2013). *Wals online*. Leipzig: Max Planck Institute for Evolutionary Anthropology. Retrieved from `https://wals.info/`

Fleming, D., Giordano, B. L., Caldara, R., & Belin, P. (2014, September). A language-familiarity effect for speaker discrimination without comprehension. *Proceedings of the National Academy of Sciences*, *111*(38), 13795–13798. Retrieved 2021-01-29, from `https://www.pnas.org/content/111/38/13795` (Publisher: National Academy of Sciences Section: Social Sciences) doi: 10.1073/pnas.1401383111

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & cognition*, *19*(5), 448–458.

Goldstein, A. G., Knight, P., Bailis, K., & Conover, J. (1981). Recognition memory for accented and unaccented voices. *Bulletin of the Psychonomic Society*, *17*(5), 217–220.

Johnson, E. K., Bruggeman, L., & Cutler, A. (2018). Abstraction and the (Misnamed) Language Familiarity Effect. *Cognitive Science*, *42*(2), 633–645.

Johnson, E. K., Westrek, E., Nazzi, T., & Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, *14*(5), 1002–1011.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., ... others (2020). Libri-light: A benchmark for asr with limited or no supervision. In *Icassp 2020-2020 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 7669–7673).

Kanagasundaram, A., Vogt, R., Dean, D., Sridharan, S., & Mason, M. (2011). I-vector based speaker recognition on short utterances. In *Proceedings of the 12th annual conference of the international speech communication association* (pp. 2341–2344).

Köster, O., Schiller, N. O., et al. (1997). Different influences of the native language of a listener on speaker recognition. *Forensic Linguistics*, *4*, 18–28.

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., & Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656*.

Levi, S. V. (2019). Methodological considerations for interpreting the language familiarity effect in talker processing. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(2), e1483.

Lin, C.-Y., & Wang, H.-C. (2005). Language identification using pitch contour information. In *Proceedings.(icassp'05). ieee international conference on acoustics, speech, and signal processing, 2005.* (Vol. 1, pp. I–601).

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, *116*, 374–388.

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition*, *143*, 36–40.

Perea, M., Jiménez, M., Suárez-Coalla, P., Fernández, N., Viña, C., & Cuetos, F. (2014). Ability for voice recognition is a marker for dyslexia in children. *Experimental Psychology*.

Perrachione, T., Dougherty, S., McLaughlin, D., & Lember, R. (2015). The effects of speech perception and speech comprehension on talker identification. In *Icphs.*

Perrachione, T. K. (2018, December). Recognizing Speakers Across Languages. In S. Frühholz & P. Belin (Eds.), *The Oxford Handbook of Voice Perception* (pp. 514–538). Oxford University Press. doi: 10.1093/oxfordhb/9780198743187.013.23

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... others (2011). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding.*

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline..

Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*, *1*(2), 121–131.

Thorburn, C. A., Feldman, N. H., & Schatz, T. (2019). A quantitative model of the language familiarity effect in infancy. In *Proceedings of the conference on cognitive computational neuroscience.*

Wester, M. (2010). *The emime bilingual database* (Tech. Rep.). The University of Edinburgh.

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781–790.

Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific reports*, *5*(1), 1–9.

2.3.2 Investigating the stability of the i-vector model on modelling speech perception through LFE

In research in general, and cognitive science in particular, robustness, which we define as the reliability and stability of research findings across different conditions, samples, and methodologies, is a crucial concept that requires careful attention in experimental design to ensure that effects are significant and replicable. In order to achieve such stability, experimental work typically relies on stringent significance testing and numerous replications on many participants since a participant can only be tested on n numbers of stimuli at test time. Computational modelling, on the other hand, allows for extensive testing on a single model since machines are tireless (Lavechin et al., 2022) (i.e. the same model can be presented with as many evaluation measures as necessary). Because of such, it is usually assumed that robustness is found through this extensive testing. However, to validate the robustness of an effect, it is also critical to assess the stability of the model itself, that is, the ability of a model to return consistent results over time, under varying conditions, or when different inputs are provided. This encompasses the model's ability to make accurate predictions not only on the data it was trained on but also on unseen data or data that might slightly deviate from the training set. In machine learning, this is evaluated mainly using two approaches: model rerunning in the case of non-deterministic algorithms (using different random initialisation) and resampling techniques such as bootstrapping or cross-validation.

Despite being relatively well studied in machine learning, this question of model stability is often left aside in cognitive computational modelling (Lee et al., 2019). However, when possible, such techniques allow direct insights into the stability of the algorithm. In the case of resampling, by generating multiple training sets all sampled from the same source, we obtain multiple models with the same mechanisms trained on similar but different data, and all tested on the same dataset. This allows computation of variance statistics on the model and thus to quantify the variations of the observed measurements if the experiments were repeated.

This subsection examines the question of model stability from two perspectives, taking support on the LFE. First, we investigate whether LFE results on a specific language pair (French and English) can be reproduced across different datasets. In other words, we aim to determine how sensitive the model is to data specificities and whether similar results can be obtained when the dataset changes. We also use resampling to assess the model's intrinsic stability. Second, we explore the role of the model's hyper-parameters in the reproducibility of the results and investigate how dependent the results are on hyper-parameters.

2.3.2.1 Materials

The CommonVoice and LibriVox datasets We use two open-source projects to create our train and test sets, the LibriVox project (Kearns, 2014)[1] and Mozilla's

[1] https://librivox.org

CommonVoice project (Ardila et al., 2019)[2]. Regarding CommonVoice (Ardila et al., 2019), we used the English and French train and test sets introduced in Subsection 2.3.1, which consist of 15 hours training sets split between 60 speakers and 30-minute test sets with 20 speakers. LibriVox is an open-source project that collects recordings of free-of-rights audiobooks in multiple languages. Using the Libri-Light tools developed by (Kahn et al., 2020), we gathered audiobooks in English and French, which we segmented into utterances using an open-source Voice Activity Detection (VAD) model (Lavechin et al., 2020). We then created English and French training sets of 15 hours with ten speakers and a test set of ten different speakers with a total duration of 30 minutes.

Resampled data For the resampling analyses, we resampled new train sets for English and French in both the CommonVoice and LibriVox datasets. To do so, we sampled new data from the original sources, following the same constraints presented above to create as many train sets as possible. We did not resample new test sets. We generated three additional English train sets and two French train sets from the LibriVox corpus, as well as six and five additional English and French train sets, respectively, from CommonVoice.

2.3.2.2 Methods

The methods in training the i-vectors models and evaluating the LFE score on these models are identical to those presented in §2.3.1. However, in this subsection, we consider two pipelines which differed in terms of the hyper-parameters used; these two pipelines will allow us to compare the role of hyper-parameters in model stability. The first, LD (Low-Dimension), pipeline uses the same number of Gaussian mixtures (128) and i-vector dimension (150) as previous works on modelling language discrimination (Carbajal et al., 2016; de Seyssel and Dupoux, 2020). Because these hyper-parameters allowed modelling of language discrimination, it was thought sensible to assume that the same hyper-parameters would successfully model the Language Familiarity Effect. The hyper-parameters used in the second, HD (High-Dimension), pipeline (2,048 Gaussian mixtures and i-vectors of dimension 400) are the ones used in Thorburn et al. (2019) and the default values used in the Kaldi toolkit (Povey et al., 2011) that have proven to be successful in different speaker recognition tasks.

2.3.2.3 Experiment and Results

In order to test whether the i-vector models proposed in Thorburn et al. (2019) yield results which can be trusted in terms of stability, we propose an approach based on data resampling. We used the resampled train sets from the English and French original corpora as presented in §2.3.2.1, considering both the CommonVoice and LibriVox datasets. We chose these languages as the ones which allowed us to generate the larger number of train sets possible, and this in both the LibriVox CommonVoice corpora. We estimated the LFE on each possible combination of

[2]https://voice.mozilla.org

English and French train sets following the method introduced in §2.3.1, keeping the test sets constant (we only have one English and one French test set). This was replicated both in the LD and HD setups. This means that we were able to vary the models for the English-French pair in three ways: across datasets (LibriVox and CommonVoice), across sampled data within the same dataset (resampling method), and across hyper-parameters (LD and HD setups).

Train	LibriVox		CommonVoice	
	LD	HD	LD	HD
French	10.10 *(13.00)*	4.91 *(44.21)*	30.75 *(8.63)*	28.96 *(25.81)*
English	1.95 *(10.91)*	32.06 *(47.07)*	-2.13 *(9.78)*	10.47 *(17.20)*
Balanced (overall)	6.03 *(5.29)*	18.48 *(12.34)*	14.31 *(6.80)*	10.85 *(5.70)*

Table 2.1: Mean and Standard Deviation (italicised) of LFE scores obtained from all possible resampled combinations for the different train conditions with the LibriVox and CommonVoice datasets.

We can observe in Table 2.1 and Figure 2.1 that, for both CommonVoice and LibriVox, all "balanced" LFE scores (that is averaged over both languages in a symmetrical testing design) are on average positive, suggesting that i-vectors are better at discriminating speakers from the language they were trained on in both the HD and LD setup, confirming the findings of Thorburn et al. (2019) that they are good models of LFE.

We also report, in Table 2.1, results for each train set: instead of counterbalancing the LFE score by considering simultaneously models trained on language A and language B as did Thorburn et al. (2019), and as presented in §2.3.2.2, we only consider a single model trained on language A. In this case, we only have $S_{\text{same}} = Ts(A)_{Tr(A)}$, and $S_{\text{diff}} = Ts(B)_{Tr(A)}$. When analysing such scores, we can see one instance (the LD condition on CommonVoice) in which the model trained on English did not yield better discrimination scores on the English set than the French set and, therefore, is not showing a Language Familiarity Effect. This could suggest both variance in the models and asymmetry within the language pair.

We can see that the variance of these raw, not counterbalanced, scores is high. This suggests that training multiple models on datasets generated from the *same* source and language will be relatively unstable, as it will yield speaker discrimination scores and, by extension, LFE scores with a high variance. However, when looking at the balanced LFE scores, that is, with each model being counterbalanced by its opposite language counterpart, the results seem relatively stable, with a smaller variance and the LFE mainly going in the same direction. It should still be noted that, as shown in Figure 2.1, even if the average balanced LFE scores suggest an effect of Language familiarity, there are still English-French train combinations in which the LFE score is negative.

Another interesting observation is the difference yielded by the choice of setup. Whilst the high-dimension setup yields higher variance with the LibriVox, the opposite effect is found with the CommonVoice, with overall less stability in the

low-dimension setup. As will be discussed further in the related discussion, this difference is likely due to the difference in the number of speakers in the two datasets, with higher dimension models being more suited to datasets with higher numbers of speakers.

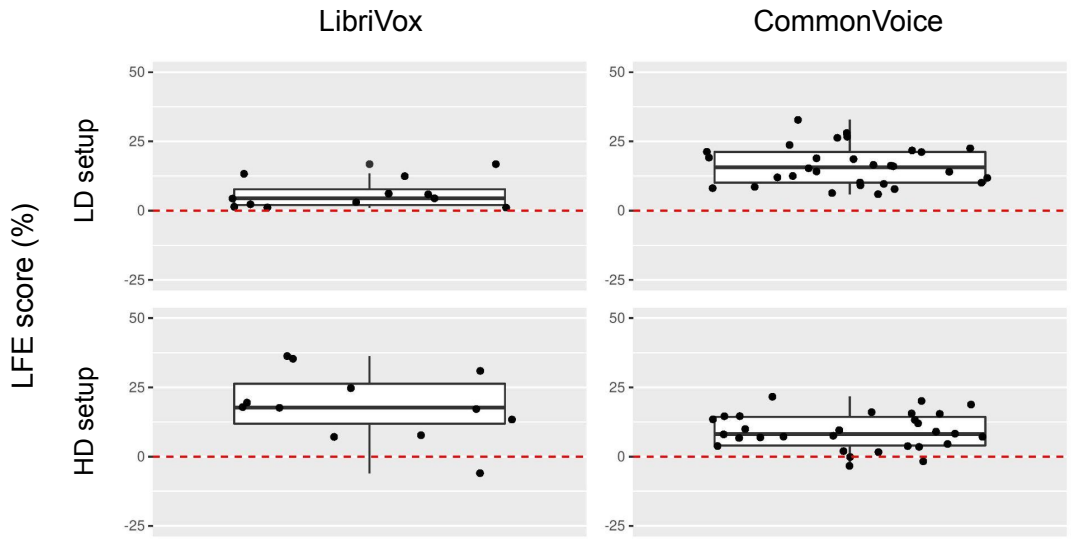


Figure 2.1: Comparison of LFE scores on the English-French language pair, using resampled train sets from the CommonVoice and LibriVox corpora. The scores are computed using both low and high-dimension models and are averaged over the English and French test sets.

2.3.2.4 Discussion of Subsection 2.3.2

As in Subsection 2.3.1, we could reproduce the Language Familiarity Effect using i-vector models on the English-French language pair, with better speaker discrimination when the model is tested on speakers from the same language it was trained on. This replication was the case for both datasets and in both the LD and HD setups. Nevertheless, our experiments revealed some lack of stability within the models when the languages were analysed individually (i.e. not counterbalanced), with high variance in LFE scores. This emphasises the importance of symmetrical testing models in this computational model of LFE and in computational modelling in general. Indeed, it allows for controlling the case where participants from one language generally show higher speaker discrimination scores, regardless of the languages they are tested on. Although probably not as likely in behavioural experiments, this is a primordial control in speech modelling when looking at language differences, where unwanted differences in the training sets can lead to better speaker discrimination scores in one language over the other.

Moreover, the fact that some instances of train and test set combinations yielded negative LFE scores, even if, on average, the effect was positive, emphasises the importance of investigating a model's stability in computational modelling.

Indeed, our observation should lead us to be cautious in interpreting results based on a single training set. Gorman and Bedrick (2019) have recently shown that considering a single train-test split when comparing NLP models can result in erroneous conclusions. However, to our knowledge, this is the first time a similar conclusion has been drawn for cognitive computational modelling. Unfortunately, considering multiple splits is not always possible, especially when the models require data with particular constraints. This is, for example, the case in the experiments reported in the following chapters, and interpretation should be led with caution in such cases. Naturally, further work could focus on optimising the models to increase their stability without the need for generating additional splits.

Stability and hyperparameters' choice Another source of variance revealed by our experiments was the choice of hyper-parameters in training the model and, more particularly, how it seems to correlate with the training data characteristics. Indeed, while the high-dimension setup leads to higher variance rates in LibriVox compared to its low-dimension counterpart, this effect is less present, if not inverted, with CommonVoice. The fact that train sets from CommonVoice count more speakers than those from LibriVox (60 vs 10) could explain these results, with more hyper-parameters being needed to capture the larger diversity of speakers in the former. Fewer hyper-parameters are helpful in the case of LibriVox, and the "extra" dimensions in the HD setup could capture non-relevant acoustic information, potentially resulting in cases of over-fitting, which can lead to more variance at test time. Indeed, the number of hyper-parameters in the HD setup is based on Thorburn et al. (2019) setup, which follows the recommendations for state-of-the-art speaker discrimination models trained on a high number of speakers.

Although we could not reproduce our stability experiments on languages other than the English-French pair due to the lack of enough data in other languages, confirming the trends on other languages and data sources would be necessary. Moreover, our data suggest that the hyper-parameter choice should be fitted to the training data's characteristics, particularly with the number of speakers in the data. Further work on making this choice without requiring large amounts of data to compute variance would be of great interest.

Stability and cognition Beyond these parametrical issues, we believe it is worth approaching this instability aspect in terms of cognition. What if the variance found in our experiments also reflects the variance, and possibly lack of stability, of the Language Familiarity Effect itself? Indeed, although this effect has been largely replicated and proven robust in adults, it has only been looked into very recently in infants. Therefore, it is plausible that the LFE is less robust than it is in adults. As the i-vector approach is a model of infant speech processing, one could argue that the variance found in our models reflects the fragility of the effect in infants, attesting further to the model's validity.

These findings underscore the importance of replication and stability testing in computational modelling, shedding light on a less-explored aspect of this field. While these results may cast doubt on previous findings, they do not necessarily

undermine existing cognitive hypotheses. Instead, these outcomes underscore the need for our models to achieve greater stability before they can be deemed reliable computational representations of cognitive processes. However, this does not necessarily mean that the pipeline itself is flawed. Integrating an optimisation step based on the training set may be necessary to improve the models' generalisation. A possibility would be to choose the hyper-parameters that would bring the smaller possible amount of variance.

2.3.3 Conclusion

In this section, we focused on using i-vectors as a model of speech perception to model the Language Familiarity Effect. We first presented a paper where we (1) demonstrated the ability of the i-vector model to obtain a gradual measure of the LFE and (2) replicated the effect on a wide range of language pairs, with indications that language distance may play a role in the size of the effect. Beyond these experimental results, the paper highlights the advantages of computational modelling in the study of cognitive effects in general and speech perception specifically. It allows for experimentation under various conditions (in this case, different languages) in a completely controlled setup, which is not possible in experimental studies due to a lack of symmetrical testing in cross-linguistic experiments, difficulties in finding native speakers of a wide array of languages with similar experience, and other limitations. In conclusion, this paper has shown that computational models have the potential to significantly enhance our understanding of cognitive processes by allowing for better control over experimental setups.

In the latter part of this section, we delved into examining the robustness of i-vector models in representing the LFE. Despite effectively encapsulating the LFE, these models demonstrated vulnerability to instability stemming from alterations in the training data and hyper-parameters. These findings, which do not negate the preceding conclusions, highlight the significance of exercising prudence while interpreting outcomes derived from these models. Further work is needed to improve the stability of these models, but the potential benefits of using computational models in cognitive research make this a promising avenue for future investigation.

To conclude, our work in this section has allowed us to define better guidelines and concepts proper to the cognitive modelling approach, which we will attempt to follow throughout the remainder of the thesis. Moreover, the work further supports the i-vector model as a good model of speech perception of indexical information by replicating behavioural findings of the LFE and, by extension, mimicking the speaker-language entanglement present in speech perception. It also highlights further the role of acoustic language distance in such perception, which we will leverage in the next section of this chapter.

2.4 Using I-vectors to compute language similarity

In Sections 2.2 and 2.3, we showed that a global model such as the i-vector model, typically used in applications like speaker and language identification and

normalisation, can emulate cognitive phenomena, indicating its potential as a model for human speech perception. Additionally, we tackled the topic of language distance in both chapters, highlighting its impact on cognitive process modelling, with more acoustically distant languages being easier to discriminate and yielding higher language familiarity effects. In this section, we aim to investigate if we can use this effect to develop an automatic acoustic method for measuring acoustic language distance (i.e. which can be quantised from direct properties relative to the signal, such as phonetics or prosody). In contrast to the previous sections, in this work, we use the i-vector model as a tool rather than as a model of speech perception.

This work is presented in the form of a paper:

de Seyssel, M., Wisniewski, G., Dupoux, E., and Ludusan, B. (2022). Investigating the usefulness of i-vectors for automatic language characterisation. In *Proc. Speech Prosody 2022, 460-464*. doi:10.21437/SpeechProsody.2022-94

2.4.1 Investigating the usefulness of i-vectors for automatic language characterization

Paper summary

Determining language distance can be challenging, as it involves various language properties such as phonology, grammar, lexicon, and prosody. One can envision a lot of different typologies, contingent on the language level and characteristics considered. Numerous language typologies are available; resources like the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) document them. In this work, our focus is primarily on *acoustic language distance*. The approach we present here allows us to measure language distance from an acoustic standpoint. Yet, as the paper below demonstrates, the measurement encompasses various linguistic properties, including phonetic, phonological, and syntactic properties.

To propose a method for computing acoustic, i-vector-based language distance, we trained an i-vector model on multiple languages. We then calculated the distance between the centroids of these different languages as a measure of acoustic distance. This method requires only a few utterances in each language, allowing us to overcome the low-resource problem typically encountered in large multilingual models.

Firstly, we visually demonstrated the correlation between the i-vector centroids and different language families, suggesting that the distance calculated is meaningful. Additionally, we discovered a correlation between the acoustic distance we computed and the distance based on syntactic properties calculated by experts. This finding is surprising, but previous research has suggested a correlation between word order and prosody in some languages (Nespor and Vogel, 2012). Leveraging this discovery, we propose a method for predicting a specific word order feature, suggesting that this automatic method could be used to classify languages automatically.

The work presented in this paper is promising as an annotation tool for linguists and typologists. However, it can also be used in speech processing, with applications where data selection is based on acoustic language distance, as we will further discuss in Chapter 5.

Investigating the usefulness of i-vectors for automatic language characterization

Maureen de Seyssel[1,2], *Guillaume Wisniewski*[2], *Emmanuel Dupoux* [1], *Bogdan Ludusan*[3]

[1]Cognitive Machine Learning (ENS–CNRS–EHESS–INRIA–PSL Research University), France
[2]Université de Paris, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France
[3]Phonetics Workgroup, Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Germany

maureen.deseyssel@gmail.com, bogdan.ludusan@uni-bielefeld.de

Abstract

Work done in recent years has shown the usefulness of using automatic methods for the study of linguistic typology. However, the majority of proposed approaches come from natural language processing and require expert knowledge to predict typological information for new languages. An alternative would be to use speech-based methods that do not need extensive linguistic annotations, but considerably less work has been done in this direction. The current study aims to reduce this gap, by investigating a promising speech representation, i-vectors, which by capturing suprasegmental features of language, can be used for the automatic characterization of languages. Employing data from 24 languages, covering several linguistic families, we computed the i-vectors corresponding to each sentence and we represented the languages by their centroid i-vector. Analyzing the distance between the language centroids and phonological, inventory and syntactic distances between the same languages, we observed a significant correlation between the i-vector distance and the syntactic distance. Then, we explored in more detailed a number of syntactic features and we proposed a method for predicting the value of the most promising feature, based on the i-vector information. The obtained results, an 87% classification accuracy, are encouraging and we envision to extend this method further.

Index Terms: i-vector, language typology, suprasegmental information, prosody, syntax

1. Introduction

Languages differ on a variety of levels, and studying these variations is fundamental in understanding how language is structured [1, 2]. A lot of effort has been put in defining features to classify languages at multiple levels: from phonology [3, 4], morphology and syntax [5] up to semantics [6]. These characterizations of languages are done by expert linguists and have been collected in several typological databases such as WALS [7], PHOIBLE [8] or SSWL [9]. However, this documentation is not complete, as the majority of languages spoken in the world today still lack description in terms of numerous typological features, thus making a general classification of languages difficult to achieve.

Automatic methods of language typology may help with this problem by characterising specific aspects of languages either based on annotated linguistic features [10] or directly from a corpus [11, 12]. Such methods can, in turn, learn to predict missing features [13], or can be used in downstream language-processing models [14]. Although the bulk of the methods that have been proposed in the literature are coming from natural language processing (NLP; see [15] for an overview), there is also some work done towards speech-based language characterization. Those studies include analyses which focus on prosody, either by performing comparative analyses of dialects [16] and languages [17] using suprasegmental information, by employing long-term information for the syntactic description of languages [18], or even by attempting automatic, signal-based, prosodic typology [19].

The results of these studies provide evidence that signal-based approaches, especially those based on prosodic information, may be developed to help automatic typology in different linguistic areas. We investigate here a promising speech representation which captures long-term information, i-vectors, with the goal of aiding the automatic characterization of languages. The i-vectors, features which are able to represent entire utterances of speech into fixed-dimension representations, have been shown to capture speaker-specific characteristics, being initially successfully used in speaker identification applications [20]. Subsequent studies established that these features may capture also language specific characteristics, when language labels are explicitly given, for the task of language identification [21, 22]. Moreover, when the input features used for computing the i-vectors contain prosodic information such as pitch or intensity, this latter type of information is reflected in the composition of the i-vectors [23]. More recent work employing i-vectors for a comparative analysis of dialects [24], has shown that the i-vector distances between the four investigated Latvian dialects correlated with their geographic position (and presumably with the inter-dialect distances, although no objective evaluation was performed to attest this).

We propose to investigate here whether acoustic distance between language, based on i-vectors, can be used to predict various typological distances between languages (Section 3). For this we employ a large set of languages belonging to several linguistic families and we evaluate the method by means of objective distances based on expert linguistic features. In the second part of the study, we explore an approach based on pairwise language distances to directly predict specific features of the given languages (Section 4).

2. General methods

2.1. Materials

We used languages from CommonVoice 6.1 to generate our dataset. This collaborative corpus, an initiative supported by the Mozilla Foundation[1], consists of recordings of people reading

[1]http://voice.mozilla.org/

10.21437/SpeechProsody.2022-94

prompts in various languages and environments. 60 languages were available in the original dataset. We selected utterances from 24 languages[2] to create a balanced dataset, with a total duration of one hour per language, equally split between 60 speakers. A high number of speakers was chosen in order to have a high within-language variability. Preliminary experiments with larger training sizes show that one hour was sufficient for our purposes, while allowing us to employ more languages. The average number of utterances per language set was 761, with an average utterance duration of 4.62 seconds. No significant variance in the number of utterances was observed between languages.

2.2. Training pipeline

We first extracted Mel frequency cepstral coefficient features (MFCCs) [25] for all utterances in the 24 languages, with 13 coefficients including energy (related to intensity), along with double-delta coefficients and pitch information. We then used these features to train a standard i-vector system on all languages, using the Kaldi toolkit [26], with 2,046 Gaussians and i-vectors of dimension 400. In order to maximise the distance between languages, a transformation matrix based on a Linear Discriminant Analysis (LDA) was also computed, and applied to the i-vectors.

Next, we generated i-vector representations for all utterances from our dataset and we calculated the mean i-vector for each language, averaging over all i-vector representations of the language. We call these vectors "centroids".

Finally, we determined the distance between a pair of languages by computing the Euclidean distance between the centroids of those two languages (previous work suggesting that Euclidean distance works best with language i-vectors [24, 27, 28]). Distances were computed for all possible 276 pairs yielded by the languages in our dataset.

3. Experiment 1 : Estimating language distances using i-vectors

In this first experiment, we are comparing the i-vector distances to expert-annotated linguistic distances.

3.1. Linguistic distances

We retrieved the inventory, phonological and syntactic language vectors from the URIEL database [10], having a size of 28, 158 and 103, respectively. These vectors contain various featural information belonging to these three linguistic areas (inventory, phonology and syntax), and were gathered from different typological databases such as WALS and PHOIBLE. To avoid using sparse vectors, missing features were predicted following the method proposed in [13]. We also concatenated, for each language, the vectors corresponding to the three different categories of information into one, which we refer to as the "general" linguistic vector. Distances between languages were then derived for each of the 276 language pairs using the cosine distance, following [10] and [13], for each of the four vectors (three linguistic areas, one general vector).

[2] Arabic, Catalan, Czech, Dutch, Welsh, German, English, Spanish, Basque, Persian, French, Frisian (Netherlands), Italian, Kabyle, Polish, Portuguese, Russian, Kinyarwanda, Swedish, Tamil, Turkish, Tatar, Ukrainian and Mandarin (Mainland China)

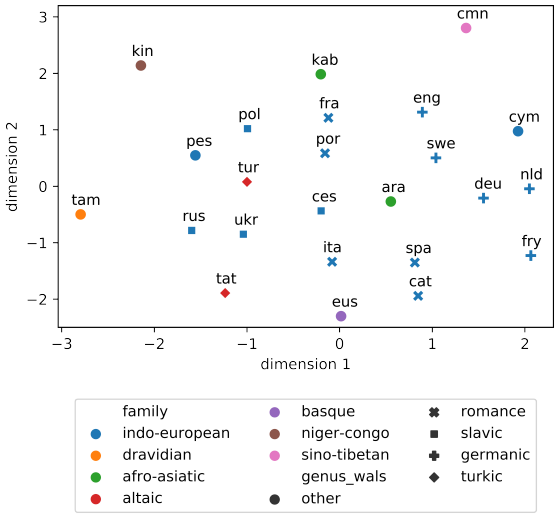


Figure 1: *Visualisation of the centroid language vectors using multidimensional scaling. Colors indicate the language family and shape the language genus as documented in the WALS. Genuses related to only one language are grouped into the "other" category.*

3.2. Results

We applied multidimensional scaling to the centroid i-vectors to visualise how the different languages were scattered around the acoustic space. As shown in Figure 1, the most distinctive languages of our set such as Mandarin, Tamil and Kinyarwanda are separate from the other languages. Similarly, languages sharing common roots seem to located in the same places, as is the case for Russian, Ukrainian, Czech and Polish, or for English, Swedish, German and Dutch. This qualitative analysis seems to corroborate the fact that the information present in the i-vectors may capture some sort of language distance.

We computed the Pearson correlation coefficient between the i-vector distance scores and the general linguistic vectors distance scores for all language pairs, as well as its 95% confidence interval using bootstrapping with 9,999 samples. We then compared it to the correlation obtained when randomly pairing i-vector distances and linguistic distances (for 9,999 times, by resampling with replacement). The correlation is significant when the boundaries of the two confidence intervals do not overlap. As reported in Table 1, the i-vector distances and the general linguistic distances were positively correlated, further supporting the fact that i-vectors encode language information.

We continued our analysis by calculating the correlation between i-vector distances and each of the three categorical linguistic distances, in order to gain more knowledge regarding which type of information is captured by the i-vectors. As presented in Table 1, there was no significant correlation neither between the phonology distances and the i-vector distances nor between the inventory distances and the i-vector distances. The syntactic distances however, yielded a significant positive correlation with the i-vector distance scores. Because the datapoints in our correlations correspond to language pairs and are therefore not totally independent from each other, in addition to computing the random permutations, we also re-ran the analysis on the syntax and i-vector distances correlation removing each

Table 1: *Correlation scores between the i-vector distances and each of the general, phonology, inventory and syntax distances. The median Pearson R value is reported over the bootstrapped alternative hypothesis along with its 95% confidence interval (* indicates significance). CI for the random permutation is also reported.*

	Pearson R	95% CI	Random perm. 95% CI
general	0.52*	[0.44, 0.59]	[-0.29, 0.34]
phonology	0.34	[0.23, 0.44]	[-0.29, 0.36]
inventory	0.22	[0.12, 0.32]	[-0.28, 0.35]
syntax	0.55*	[0.47, 0.62]	[-0.28, 0.33]

time one of the languages (so 23 language pairs). A significant correlation was obtained every time, suggesting that the initial results are robust.

4. Experiment 2: Predicting syntactic features from speech representations

We have seen in Experiment 1 that the distances between the i-vectors centroids correlate best (among the distances investigated here) with the syntactic distance between language pairs. In this experiment we would like to determine which are the syntactic features most correlated with the i-vector distances. Moreover, we conduct a preliminary investigation into using the information given by the i-vector distances to predict values for languages which have not been yet described.

Based on evidence from prosodic phonology [29], showing that prosodic information (the placement of prosodic prominence within phonological phrases) is correlated with the relative order of heads and complements in a language, and from speech processing revealing that long-term information (the shape of the amplitude modulation spectrum) may discriminate between head-complement and complement-head languages [18], we focused our analysis on word order features.

4.1. Methods

We chose those word order features from the WALS for which our languages were represented only by two classes and an optional third class, for "mixed" or "no dominant order". One of the two classes was then coded with the value 1, while the other class with the value 0. In case the optional mixed/no dominant order class existed, it was coded with the value 0.5. We then kept only those features which had at least three instances of languages for each of the two classes (0 or 1). Languages for which their feature value was not recorded in the WALS, were marked with a question sign (see Table 2) and were not used to compute the correlation. The following six features were employed in this experiment:

- 83A: Order of Object and Verb
- 85A: Order of Adposition and Noun Phrase
- 86A: Order of Genitive and Noun
- 87A: Order of Adjective and Noun
- 90A: Order of Relative Clause and Noun
- 93A: Pos. of Interrogative Phrases in Content Questions

Table 2: *The WALS syntactic features employed in this experiment. We used features for which the considered languages had only two distinct values (coded by 0/1) and, optionally, a mixed/no dominant order (coded by 0.5). Features missing a value are coded by a question mark in the table and not used in the correlation computation. The last column shows the prediction of the feature 90A using the proposed approach.*

Lang.	**WALS features**						**Pred.**
	83A	**85A**	**86A**	**87A**	**90A**	**93A**	**90A**
ara	0	0	0	0	1	0.5	1
cat	0	0	0	0	1	?	1
ces	0	0	0.5	1	1	0	1
cmn	0	0.5	1	1	0	0	0
cym	0	0	0	0	1	1	1
deu	0.5	0	0	1	1	1	1
eng	0	0	0.5	1	1	1	1
eus	1	1	1	0	0	0	0.5
fas	1	0	0	0	1	0	1
fra	0	0	0	0	1	1	1
fry	0.5	0	0.5	1	1	1	1
ita	0	0	0	0	1	?	1
kab	0	0	0	0	1	?	1
kin	0	0	?	0	?	0	0
nld	0.5	0	0	1	1	?	1
pol	0	0	0	1	1	1	1
por	0	0	0	0	1	?	1
rus	0	0	0	1	1	1	1
spa	0	0	0	0	1	1	1
swe	0	0	1	1	1	1	1
tam	1	1	1	1	0	0	0
tat	1	1	1	1	0	?	0.5
tur	1	1	1	1	0	0	1
ukr	0	0	?	1	1	?	1

We then determined, for each feature, the distance (*feat_dist*) between all pairs of languages which had values for that particular features, by computing the absolute difference between the value of the two classes. For example, if one class had the value 0 and the other one value 1, the absolute difference between them was equal to 1. Thus, languages belonging to a class always had a difference equal to 0 to the other languages from the same class, a distance of 1 to the instances of the other class and a distance of 0.5 to the mixed/no dominant class elements. The pairwise *feat_dist* for all language pairs was subsequently correlated to the distance between the centroid of the i-vectors of the same pairs of languages. The R software [30] was used to compute the Pearson r correlation coefficient and to test the significance of the correlation.

Finally, we employed the most promising feature (the one having the highest correlation to the i-vector distance) to predict the values of languages, both of those that are described and of those for which no value is given in the WALS. For the languages which had values, we proceeded as follows: we replaced the original value of the language by either 0, 1 or 0.5 and we recomputed *feat_dist* and its correlation to the i-vectors. We then considered as the predicted class the one which gave the highest correlation among the three. Also for the languages without values in the WALS, an identical procedure was applied (the only difference is that we actually consider all the pairs which contain that particular languages, as they were initially not included due to not having a value for that feature).

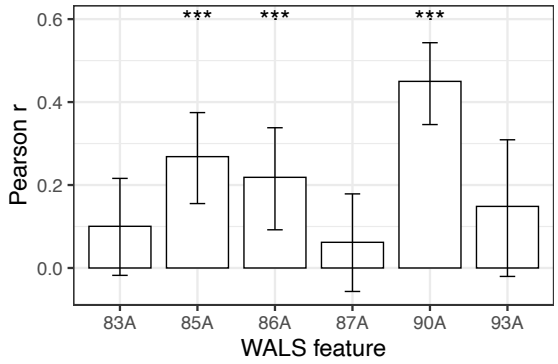


Figure 2: *Obtained correlations between the i-vector distances and the feat_dist corresponding to that particular WALS feature. The error bars represent confidence intervals. The asterisks on top illustrate the significance level (***, $p < 0.001$).*

4.2. Results

The correlation results between the pairwise distances *feat_dist* and the i-vector ones are illustrated in Figure 2. We observe positive low to medium correlations for all the investigated features, with a maximum value of 0.45 obtained for the feature 90A. Three of the feature distances, 85A, 86A and 90A reached significant correlations with the i-vector distances.

We then used the best feature, 90A, to predict the values of the languages that had values for this feature in the WALS, as well as for Kinyarwanda, the only language from our set of 24 languages which was missing a value for that feature. These results are presented in the last column of Table 2. Comparing the predictions with the values given in the WALS we see that most languages are correctly predicted. The three languages which were not correct, belonged to the class 0, with two of them being classified as mixed/no dominant order. The proposed approach predicted Kinyarwanda to be a *Relative Clause - Noun* language (class 0), similar to the prediction made by [13].

5. Discussion and conclusions

Using an i-vector model of language identification, and relying on the average representation of each language in our train set, we were able to compute distances between language pairs. We found that these languages correlated with the general distance from [13], based on the concatenation of multiple expert-annotated linguistic features, at different levels. These results extend those of [24], by showing that i-vectors encode relevant information to discriminate also between languages. Moreover, we evaluated our distances against expert-derived observations, thus providing robust evidence for the suitability of using i-vectors and showing their appropriateness for methods for automatic language characterisation.

One of the main advantage of this approach is that only relatively small amount of speech per language is required (here we used one hour, but we could probably reduce it further). However, it is important to have sufficient within-language variability in the training set languages (e.g., by increasing the number of different speakers or recording conditions). An alternative would be to first train a model on a fixed number of languages which have enough data, and use it to compute representation vectors of novel languages with less data. Assuming we have an adequate amount of data and diversity among the languages in the training set, it might be possible to compute distances to new languages with only a few utterances. Finally, whilst not reported here, we also found a significant, although slightly weaker, correlation when no Linear Discriminant Analysis was applied, suggesting that the model can capture language characteristics even in a totally unsupervised fashion.

Having found that the i-vector distances correlated with general expert-derived linguistic features, we analysed further whether there were particular levels of linguistics that correlated with this new distance. We looked at three different levels: inventory, phonology and syntax, and found that the syntax-derived distances yielded a significant correlation with the i-vector distances. The fact that the i-vector distances do not correlate with neither inventory nor phonology was surprising but not unexpected, as previous attempts to take into consideration phoneme information using i-vectors were done by modeling phoneme information in a supervised fashion [31, 32]. Further analyses will also be required to determine whether the structure of the employed corpus might have had an effect on these results. Finally, the fact that syntactic distances correlated with i-vector distances can be explained by the links between prosody and syntax (e.g. [29]), the former type of information being likely captured by our model.

In order to better investigate which syntactic distances might relate to those captured by i-vector distances, we tested six word-order features from the WALS. We observed that three features significantly correlated with our i-vector distances, with two of them capturing phrase-level word order characteristics. These results are in line with the prosodic phonology theory [29], stating that prosody information may help determine word order, as well as with the findings of previous speech processing studies (e.g. [18]). Finally, we found that our approach was able to correctly predict the 90A feature for 20 out of the 23 languages for which we had this information, and that the value it predicted for the only language missing this information (Kinyarwanda) was the same as the one predicted by the method in [13]. These preliminary results are promising in that they suggest that i-vectors could potentially be used in prediction of missing linguistic features.

We can see multiple applications to using i-vector models for language characterization. First, their output could be employed in downstream speech processing tasks, in the same way as text-based models are used in downstream NLP tasks, for example to select which languages to pretrain models from, in the case of under-resourced speech recognition. Secondly, the preliminary results we obtained on feature prediction are encouraging in that such models can bring additional knowledge to be used in predicting some features, particularly syntactic. Future work could focus on using these representations along with supervised or unsupervised learning paradigms, rather than with correlations, for determining specific language features.

6. Acknowledgements

This work was performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011012315). MS work was partly funded by l'Agence de l'Innovation de Défense.

7. References

[1] J. H. Greenberg, "The nature and uses of linguistic typologies," *International Journal of American Linguistics*, vol. 23, no. 2, pp. 68–77, 1957.

[2] B. Comrie, "Linguistic typology," *Annual Review of Anthropology*, vol. 17, no. 1, pp. 145–159, 1988.

[3] R. Jakobson, *Child language, aphasia and phonological universals*. De Gruyter Mouton, 2014.

[4] L. M. Hyman, "Where's phonology in typology?" *UC Berkeley PhonLab Annual Report, 2(2)*, 2007.

[5] B. Comrie, *Language universals and linguistic typology: Syntax and morphology*. University of Chicago press, 1989.

[6] N. Evans *et al.*, *Semantic typology*. Oxford University Press, 2010.

[7] M. S. Dryer and M. Haspelmath, "The world atlas of language structures online (max planck institute for evolutionary anthropology, leipzig)," *Available at wals. info. Accessed October*, vol. 9, p. 2014, 2013.

[8] S. Moran and D. McCloy, Eds., *PHOIBLE 2.0*. Jena: Max Planck Institute for the Science of Human History, 2019. [Online]. Available: https://phoible.org/

[9] C. Collins and R. Kayne, "Syntactic structures of the world's languages," *New York: New York University*, 2009.

[10] P. Littell, D. R. Mortensen, K. Lin, K. Kairis, C. Turner, and L. Levin, "Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 2017, pp. 8–14.

[11] B. Snyder and R. Barzilay, "Unsupervised multilingual learning for morphological segmentation," in *Proceedings of acl-08: hlt*, 2008, pp. 737–745.

[12] S. B. Cohen and N. A. Smith, "Shared logistic normal distributions for soft parameter tying in unsupervised grammar induction," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2009, pp. 74–82.

[13] C. Malaviya, G. Neubig, and P. Littell, "Learning language representations for typology prediction," *arXiv preprint arXiv:1707.09569*, 2017.

[14] H. O'Horan, Y. Berzak, I. Vulić, R. Reichart, and A. Korhonen, "Survey on the use of typological information in natural language processing," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 2016, pp. 1297–1308.

[15] E. M. Ponti, H. O'horan, Y. Berzak, I. Vulić, R. Reichart, T. Poibeau, E. Shutova, and A. Korhonen, "Modeling language variation and universals: A survey on typological linguistics for natural language processing," *Computational Linguistics*, vol. 45, no. 3, pp. 559–601, 2019.

[16] A. Suni, M. Włodarczak, M. Vainio, and J. Šimko, "Comparative Analysis of Prosodic Characteristics Using WaveNet Embeddings," in *Proc. Interspeech 2019*, 2019, pp. 2538–2542.

[17] J. Šimko, A. Suni, K. Hiovain, and M. Vainio, "Comparing Languages Using Hierarchical Prosodic Analysis," in *Proc. Interspeech 2017*, 2017, pp. 1213–1217.

[18] L. Varnet, M. C. Ortiz-Barajas, R. G. Erra, J. Gervain, and C. Lorenzi, "A cross-linguistic study of speech modulation spectra," *The Journal of the Acoustical Society of America*, vol. 142, no. 4, pp. 1976–1989, 2017.

[19] U. Reichel, K. Mády, and S. Benus, "Acoustic profiles for prosodic headedness and constituency," in *Proc. Speech Prosody 2018*, 2018, pp. 699–703.

[20] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[21] D. Martinez, L. Burget, L. Ferrer, and N. Scheffer, "ivector-based prosodic system for language identification," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4861–4864.

[22] D. Martinez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language recognition in ivectors space," in *Twelfth annual conference of the international speech communication association*, 2011.

[23] D. Martinez, E. Lleida, A. Ortega, and A. Miguel, "Prosodic features and formant modeling for an ivector-based language recognition system," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 6847–6851.

[24] A. A. Bērziņš, "Usage of i-vectors for automated determination of a similarity level between languages," *Proceedings of the Institute for System Programming of the RAS (Proceedings of ISP RAS)*, vol. 31, no. 5, pp. 153–164, 2019.

[25] P. Mermelstein, "Distance measures for speech recognition, psychological and instrumental," *Pattern recognition and artificial intelligence*, vol. 116, pp. 374–388, 1976.

[26] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[27] H. Behravan, T. Kinnunen, and V. Hautamäki, "Out-of-set i-vector selection for open-set language identification." in *Odyssey*, vol. 2016, 2016, pp. 303–310.

[28] E. San Segundo, A. Tsanas, and P. Gómez-Vilda, "Euclidean distances as measures of speaker similarity including identical twin pairs: a forensic investigation using source and filter voice characteristics," *Forensic Science International*, vol. 270, pp. 25–38, 2017.

[29] M. Nespor and I. Vogel, *Prosodic phonology*. De Gruyter Mouton, 2012.

[30] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2020. [Online]. Available: https://www.R-project.org/

[31] J. Franco-Pedroso and J. Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition," *Speech Communication*, vol. 76, pp. 61–81, 2016.

[32] L. F. D'Haro Enríquez, O. Glembek, O. Plchot, P. Matějka, M. Sou?far, R. de Córdoba Herralde, and J. Černocký, "Phonotactic language recognition using i-vectors and phoneme posteriogram counts," in *InterSpeech 2012 - 13th Annual Conference of the International Speech Communication Association*, 2012, pp. 1–4. [Online]. Available: http://oa.upm.es/20403/

2.5 Discussion and conclusion of Chapter 2

In the first chapter of this thesis, we aimed to shed light on the representation of speaker and language information in the speech perception process and how these types of information interact by comparing them with relevant findings in psycholinguistics. To investigate this question, we focused on an unsupervised model of speech representation, which has already proven itself as a model of speech perception: the i-vector model. This i-vector model has the specificities to represent speech at the utterance level, capturing mainly global information. As a result, it is ideally suited for the modelling of indexical information in speech perception.

In the three sections that comprise this first chapter, we have further shown the ability of the i-vector model as a model of speech perception, replicating, on a larger scale than what was previously done, the cognitive effects of language discrimination and Language Familiarity Effect. Overall, the results re-assessed the possibility of producing speaker and language discrimination with only speech as input and further proved the entanglement between these two features in speech processing models. This was shown, of course, by the presence of the LFE, but also the significant effect of speaker information on language discrimination results in bilingual settings. Finally, we found an effect of language distance in the different effects examined, which, in turn, led us to develop a method to use this model as a tool to generate some language distance measures based on acoustic information only.

2.5.1 Methodological concepts in cognitive modelling

We have also addressed a range of methodological concepts in the previous sections when adopting a cognitive modelling approach, allowing us to uncover some aspects of speech perception that may not have been highlighted otherwise.

Experimental design: variable control Experimental design is a critical aspect of research and is particularly taken care of in empirical work. However, it is often overlooked in computational modelling. Despite this, using computational models allows for near-complete *control over experimental design*, enabling testing that would be otherwise very challenging, if not impossible. Indeed, as demonstrated in Sections 2.2 and 2.3, computational modelling allows us to have complete control over many, if not all, variables, resulting in more meaningful comparisons between different conditions. The primary example is that of languages: when wanting to compare the effect of different languages on the same process, the task is tedious in behavioural studies, as not only is it challenging to collect native speakers from different languages, but different cultures often will have other cognitive differences which might undermine the comparison. In computational modelling, only the input language/speech can be changed, leading to a much better comparison. This, in turn, allows us to test effects such as *universality*, as we did in Section 2.3.

Experimental design: counterbalancing *Counterbalancing* is another crucial concept related to experimental design that empirical studies recognise as essential, much like variable control. Specifically, we previously introduced the concept of *"symmetrical testing"* when dealing with experiments relying on more than one language. Symmetrical testing is equivalent to counterbalancing of languages, which involves testing both languages in both directions to eliminate potential confounding effects due to differences in input, training sets, or evaluation measures between the languages. By ensuring symmetry, researchers can obtain more reliable and accurate results when the experiment involves multiple languages, whether studying nativeness or comparing language pairs.

Stability The concept of *stability* is inherently linked with the one of replication, as it refers to the consistency of results across multiple replications of an experiment. Therefore, researchers often aim to achieve stability in their experimental designs to ensure the results are reliable and can be replicated in future studies. The importance of stability has been highlighted in empirical studies when unveiling the so-called replication crisis in psychology (Collaboration, 2015; Maxwell et al., 2015). However, the importance of stability and robustness has yet to receive equal attention in the computational modelling community (Lee et al., 2019). We raised this issue in Section 2.3. In computational modelling, there is often a mistaken assumption that models are inherently stable when, in reality, they rely on various variables, including model parameters and training data which can be prone to variations. Therefore, it is crucial to acknowledge this limitation and replicate experiments when necessary to ensure the validity and reliability of the results.

Graduality Our research has demonstrated that computational models can help us generate *gradual measures*, facilitating comparisons of specific conditions. This is particularly advantageous as often in behavioural work, and especially in infant studies, the measures yielded are dichotomous. Even when effect sizes exist, the lack of standardised experimental protocols can hinder meaningful comparisons. Computational modelling provides a more precise and nuanced way of evaluating cognitive processes, enabling researchers to explore complex phenomena comprehensively. Overall, combining experimental work and computational modelling can enhance the study of cognitive processes and provide researchers with a complete understanding of complex cognitive phenomena.

Computational modelling as a tool While slightly divergent from the previously introduced concepts and contributions, we would like to highlight the potential of computational modelling as a tool. This approach was first introduced by Fourtassi (2023) (see Chapter 1, §1.5), and we have demonstrated its practicality using the i-vector model previously used in modelling speech perception. Specifically, we obtained measures of language distance using this model. While these results provide insight into the potential of computational models, one should exercise caution to avoid circularity issues when using them as tools and models concurrently. For example, using the same model to calculate language distance

and then using it again to determine the effect of language distance on an effect produced by the model can introduce biases.

2.5.2 Limitations of the i-vector model as a model of speech perception

Despite proving itself to be a good model of speech perception when it comes to modelling effects related to language and speaker information, we want to address some of the limitations of the i-vector model as such. First, we presented it as an unsupervised model of speech representation, relying only on speech input. However, this is somewhat of an approximation. Indeed, at least in the form we presented in this chapter, the model takes as input speech input which has already been processed into Mel-Frequency Cepstral Coefficients, although, as we presented earlier, they are based on human perception (Mermelstein, 1976) and should therefore be adequate for our purposes[3]. Nevertheless, we input some features that have already undergone transformations when one should aim to use raw features to avoid any approximation, as advocated in Dupoux (2018). Second, as previously mentioned, the i-vector model mainly captures global information as it operates at the utterance level. While this is adequate for investigating processes that focus solely on global information, it falls short as a comprehensive model for speech perception. It cannot capture more detailed information at a more fine-grained level, which makes it unsuitable for studying fine-grained aspects of language, such as phonetic or lexical representation.

2.5.3 Towards a framework of language acquisition

In the following chapter, we propose a framework for language acquisition modelling that can capture these more fine-grained aspects of language while enabling a comprehensive study of the language acquisition process by modelling developmental curves of the learning process.

After establishing such a framework, we will be able to look into a question introduced in this chapter: Can language learning happen in a bilingual setup without prior language separation? Indeed, in Section 2.2, we showed that language discrimination did not necessarily entail language separation, questioning the hypothesis that language acquisition in a bilingual environment involves separate processes for each language. In Chapter 4, we will therefore use the framework for language acquisition modelling to analyse bilingual language acquisition.

[3]Other popular models of speech representations, presented as unsupervised, make use of such features as input, such as HuBERT (Hsu et al., 2021).

Chapter 3

Building a developmental modelling framework of early language acquisition

3.1 Introduction

In this chapter, we shift focus from indexical to linguistic information, which relates to the meaningful content inherent in speech. Because the ability to capture linguistic information is acquired through infancy, we are particularly interested in this language acquisition process and whether such a multilayered process can be modelled in a single approach. In other words, can we construct a framework that models early language acquisition solely from speech and at multiple linguistic levels altogether?

3.1.1 Self-supervised learning speech models

In the previous chapter, we focused on modelling speech perception at the indexical level, which by nature has a global granularity, and therefore requires extracting information over a whole utterance. For this purpose, i-vector models were the most suitable models. Modelling linguistic aspects of speech perception requires using models able to capture speech information at a more fine-grained level, as linguistic information spans segmental and supra-segmental granularities (Chapter 1, §1.1). Moreover, following the cognitive computational guidelines laid out in Chapter 1 (§1.5), we want these models to be fully unsupervised, learn from raw speech and reach human-like performance in speech processing. Very recent advances in deep learning and speech processing have led to a new category of such models, trained on speech, which rely on what are called Self-Supervised Learning (SSL) algorithms (see Mohamed et al., 2022 for a review)[1].

Self-Supervised Learning algorithms were initially proposed to build latent representations of image and text input, allowing, for example, to build generative

[1]Although i-vector models presented in the previous chapter could be considered as SSL models, we follow here the approach taken by Mohamed et al., 2022 and refer to SSL models those which make use of deep learning algorithms.

text models such as the now famous GPT models (Radford et al., 2019; Brown et al., 2020). The adaptation of such techniques to speech is more recent. Indeed, it requires not only taking a much more complex and noisy input than images or text, but the input is also continuous rather than delimited by boundaries (as is the case for words with text input). To overcome this problem, most SSL speech models learn their representations at the *frame* level: the speech input is divided into windows (often of a few milliseconds), which the models learn a representation of. Finally, proposed initially to compensate for the lack of enough labelled training data in speech processing tasks, SSL speech models are usually used as pre-training, leading to constrained speech representations. These pre-training models can then be fine-tuned on downstream applications with much less labelled data that would have been initially necessary.

Mohamed et al. (2022) proposed to classify SSL speech models into three categories, depending on the nature of the main algorithm used in learning the representations: the *generative* models, the *predictive* models and the *contrastive* models. In the first category are models that learn to directly reconstruct the data from the signal, including Audio Word2vec (Chung et al., 2016), VQ-VAE (Van Den Oord et al., 2017), and Speech2vec (Chung and Glass, 2018). Predictive models such as HuBert (Hsu et al., 2021) or WavLM (Chen et al., 2022) are given an input utterance containing masks and learns to output a probability distribution for these masked segments. This output distribution is presented over a set of discrete vocabulary, which itself has to be learnt prior to that (e.g. using clustering algorithms on pre-processed speech features such as MFCCs. Finally, for contrastive models, the training task is to select for reconstruction the correct sample out of other incorrect samples, for example, in the future (e.g. CPC, Oord et al., 2018) or from masked segments of the speech (e.g. Wav2Vec 2.0, Schneider et al., 2019).

In order to reach a good enough quality of representations for reaching good performances in downstream tasks, these SSL models require hundreds if not thousands of hours of speech input, ideally with clean acoustics qualities. For example, the LibriSpeech corpus (Panayotov et al., 2015) has been extensively used in the SSL speech community, comprising hundreds of hours of English audiobooks.

CPC as a Linguistic Speech Perception and Language Acquisition Model? The previously mentioned CPC model relies on Contrastive Predictive Coding. This technique, initially proposed by Oord et al. (2018), entails predicting the future from past input. The objective is to accurately select the succeeding input from a collection of incorrect inputs sampled from the same dataset. In speech, the input corresponds to fixed-size speech frames, typically ten milliseconds. Rivière et al. (2020) later proposed an enhanced version of the original CPC model for speech input, which optimised the quality of speech representations by adjusting the architecture and specific implementation details. Henceforth, when we refer to the CPC model, we imply the speech CPC modified model as first introduced in Rivière et al. (2020).

So, why is the emphasis here on this particular model? The CPC model makes a good candidate for computational modelling of the speech perception process

for multiple reasons. First, unlike several other SSL speech models previously discussed, it possesses a certain degree of cognitive plausibility. Indeed, predictive coding in speech processing has been detected in the brain (Caucheteux et al., 2023) and during the initial stages of language development (Ylinen et al., 2017). Similarly to other SSL models, the CPC model is based on statistical mechanisms, which, as outlined in Chapter 1, are suggested as a crucial mechanism in infants' speech perception and language learning. Additionally, during the training phase of the CPC model, the prediction is executed not only on the following speech frame but also up to a certain number n of succeeding speech segments. This enables the model to learn *short-term dependencies* from the input, representing speech at both *segmental* and potentially *supra-segmental* granularities - the granularities at which linguistic information operates. Lastly, the model has demonstrated evidence of linguistic knowledge at phonetic, lexical, and even semantic and syntactic levels when trained on sufficient data, particularly when combined with other downstream unsupervised models (Nguyen*, de Seyssel* et al., 2020; Dunbar et al., 2021). We will delve deeper and expand upon this subject in the following sections of this chapter.

Probing and Evaluation of SSL speech models Although initially used in pre-training, SSL models have also been found to capture linguistic information and perform relatively well on various linguistic tasks. How can this be evaluated? We can test a model's linguistic knowledge in two primary ways. The first technique comes from the study of text-based models and is called *probing* (Alain and Bengio, 2016). It consists of checking for the presence of a specific information/feature by training a classifier on the output representations of a test dataset based on the desired feature and then calculating the accuracy on the same classifier for a held-out test set. For example, one might train a probe to predict the speaker's gender of a speech segment based on its representation in a language model. If the probe can perform the task well, it suggests that the representations learned by the language model contain information about the gender of the speaker. Although this technique allows us to unveil what information was captured by the models, it still requires a supervised component and therefore is not a good outcome task in cognitive computational modelling as we defined it in Chapter 1 (§1.5). This can be remediated using *zero-shot* (or zero-resource) metrics, tasks the model needs to have relevant enough (often linguistic) knowledge to solve. These tasks are often inspired by psycholinguistic tasks of speech perception and language acquisition. An example of such a task is the ABX phoneme discriminability task (Schatz et al., 2013, see Appendix A). The Zero-Resource challenge series (Versteegh et al., 2015; Dunbar et al., 2017, 2019, 2020, 2021) specialises in using these zero-shot metrics to evaluate linguistic knowledge of speech models. In the challenge's last iteration, we developed lexical, semantic and syntax metrics of the sort (Nguyen*, de Seyssel* et al., 2020), which we will describe later in this chapter.

3.1.2 Outline of Chapter 3

Section 3.2 presents a novel developmental and cross-linguistic framework for modelling early language acquisition. Building upon the CPC model, we introduce STatistical Learning of Early Language Acquisition (STELA), a learning simulation made of unsupervised models trained directly from raw speech, that we refer to as the *STELA learner*. As this learner is based only on statistical mechanisms, it can also give us insights into whether the statistical learning hypothesis is a viable one in the context of early language acquisition. This is a question we are interested in in this section, focusing primarily on phonetic and lexical learning.

In Section 3.3, we propose an additional measure of linguistic learning in SSL models of speech, this time at the *prosodic* level. Besides turning it into a prosodic benchmark, we also integrate this additional prosodic level into the developmental framework presented in the previous section.

Finally, in Section 3.4, we provide an overview and discussion on how learning simulations such as STELA can be helpful in the general study of early language acquisition.

Work presented in Sections 3.2 and 3.4 was carried out in collaboration with fellow PhD student Marvin Lavechin, resulting in two papers with a joint co-first authorship.

3.2 STELA: Developing a language acquisition framework

Can a single model learn from raw speech input only and simulate multiple levels of early language acquisition? Are mechanisms based on statistical learning sufficient for such modelling? What type of linguistic information can be learnt? These are some of the questions we attempt to answer in this first section, in which we present a developmental framework for modelling early language acquisition, focusing primarily on the phonetic and lexical linguistic levels.

The work presented here is in the form of a paper, which is still under submission at the time of writing. The study was done in collaboration with fellow PhD student Marvin Lavechin, resulting in a shared first authorship:

Lavechin, M.*, **de Seyssel, M.***, Titeux, H., Bredin, H., Wisniewski, G., Cristia, A. & Dupoux, E. (2022) *Can statistical learning bootstrap early language acquisition? A modeling investigation.* [under review]. ArXiv. doi:10.31234/osf.io/rx94d

3.2.1 Can statistical learning bootstrap early language acquisition? A modeling investigation

Paper summary

Language Acquisition Framework In this paper, we propose a novel framework for modelling language acquisition at (1) multiple linguistic levels and (2) using realistic input and outcomes (both of these concepts will be further discussed in Section 3.4). We present *STatistical Learning of Early Language Acquisition (STELA)*, which, as its name suggests, is a framework for modelling early language acquisition using statistical learning mechanisms exclusively. Our framework consists of three separate components: an *environment* model, a *learner* model, and an *outcome* model. The environment model refers to the input given to the learner model and consists of clean raw speech available in English or French. The learner model is designed to replicate the learning mechanisms of an infant. To achieve this, we use the baseline model we developed for the ZeroSpeech Challenge 2021 (Nguyen*, de Seyssel* et al., 2020). Finally, the outcome model measures phonetic and lexical learning using zero-shot metrics, which are adapted English and French versions of the ABX and sWuggy metrics we initially developed for the same ZeroSpeech Challenge 2021.

To model the developmental aspect of language acquisition, we set up a *developmental design* by training multiple models with varying training sizes ranging from 50 to 3,200 hours, which allows us to create "developmental curves". To ensure comparability, we split the large 3,200-hour training set into smaller subsets to have the same total amount of data at each stage of training. In addition, and following the symmetrical testing introduced in Chapter 2, we designed our framework to be *cross-linguistic*. We trained models in both English and French and tested them on evaluations in both languages. This design allows us to create native and non-native conditions for evaluation.

Statistical Learning and Early Language Acquisition The first and primary question we ask in this paper is whether statistical learning mechanisms are sufficient to bootstrap early language acquisition. Specifically, we investigate whether the statistical learning hypothesis can account for the gradual and parallel learning in infants at the phonetic and lexical levels. Using the developmental framework we introduce in this paper, we demonstrate that our model, based solely on statistical mechanisms, reproduces this gradual and parallel learning at the phonetic and lexical levels. We find that the amount of training data directly correlates with improved phonetic and lexical scores, with the learning curve starting as early as 50 hours. Importantly, we observed that complete knowledge of one level is not a prerequisite for learning the other. We also find a correlation between the two levels, with models which yield high scores in the phonetic evaluation also yielding high scores in the lexical evaluation.

Linguistic Categories and Early Language Acquisition In a second step, we question the necessity of defined linguistic categories in early language acquisition,

building upon the work of Schatz et al. (2021). Using another model based on statistical mechanisms, they found that phonetic learning could occur without the models producing precise phonetic representations. Here, we replicate this result at the phonetic level by demonstrating that although the model learns representations that approximate linguistic categories, they are not phonemes but rather more fine-grained representations of these phonemes. We also examine the lexical level and similarly find that the models do not contain proper word representations, although they exhibit similar lexical features.

To conclude, not only does this research have important implications for our understanding of the mechanisms underlying early language acquisition, as will be further discussed in Section 3.4, it also provides us with a flexible framework of development which can be adapted to test varieties of other research questions, modifying the input data (environment model), the algorithms (learner model) or the type of evaluation (outcome measures). In the next section, we propose a novel outcome measure complementing the phonetic and lexical evaluation measures presented here, focusing on prosodic knowledge.

Note: Additional information on the generation of the CommonVoice-based evaluation sets used in the paper is available in Appendix B.

Can statistical learning bootstrap early language acquisition? A modeling investigation

Marvin Lavechin[a,b,c,1,2], **Maureen de Seyssel**[a,b,d,1,2], **Hadrien Titeux**[a,b], **Hervé Bredin**[e], **Guillaume Wisniewski**[d], **Alejandrina Cristia**[a], **and Emmanuel Dupoux**[a,b,c]

[a]Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Etudes cognitives, ENS, EHESS, CNRS, PSL University, Paris, France; [b]Cognitive Machine Learning Team, INRIA, Paris, France; [c]Meta AI Research, Paris, France; [d]Laboratoire de linguistique formelle, Université de Paris Cité, CNRS, Paris, France; [e]IRIT, Université de Toulouse, CNRS, Toulouse, France

Before they even produce their first word, infants become attuned to the phonetic properties of their native language, recognize the auditory form of an increasing number of words, and develop a rudimentary knowledge of grammatical categories. What kind of learning mechanism could produce such a puzzling pattern of gradual and overlapping improvement at different linguistic levels ? In-laboratory experiments have shown that young infants are exquisitely sensitive to fine-grained statistical regularities of their language input, leading researchers to propose that *"statistical learning"* could provide such a mechanism. Yet, statistical learning abilities have only been demonstrated in infants with simple artificial languages and remain controversial as a cornerstone for early language bootstrapping. Two questions remain lingering: could statistical learning work at all when fed with the full complexity and variability of natural language? Could it account for overlapping learning at multiple levels? Here, we introduce STELA, a computational model that simulates how infants might bootstrap into language from raw audio signals using statistical learning principles. STELA is built from machine learning algorithms that predict future representations of speech based on past ones. When fed with increasing quantities of raw continuous speech from multiple speakers in French and English (no preprocessing nor human annotation), STELA reproduces the observed pattern of gradual and overlapping specialization to the "native" language across levels: it improves in discriminating sounds, recognizing the auditory form of words, and organizing sounds and words along linguistic dimensions. STELA provides a proof of feasibility that statistical learning from raw speech is sufficient to bootstrap early language acquisition at the sound and word levels. Subsequent analyses indicate that this process occurs without the use of linguistic categories at these levels.

language acquisition | artificial intelligence | self-supervised learning | statistical learning | predictive learning

Infants master critical aspects of the language(s) spoken around them well before they produce their first word. Between 6 and 12 months, infants' discrimination of native sounds shows an improvement, while those of non-native sounds shows a decline (1–4). Not only do infants learn to discover sounds of their native language, they also start learning words very early on. Evidence for word learning starts as early as 4 months, where infants have been shown to recognize their own names (5). At 6-7 months, infants recognize the auditory form of frequent words (6, 7), show a preference for content over function words (8), and segment words from fluent speech (9). For their first birthday, a typically-developing American English infant comprehends around 80 words (10). Evidence suggests, therefore, a scenario of early language acquisition where learning sounds and words develop concurrently. However, it has proved devilishly difficult to understand how infants break into the intricate system that human language is. In other words, it remains unclear how infants manage to bootstrap phonetic and lexical learning from sensory information only.

One mechanism that has been proposed to explain language acquisition is *statistical learning* (11): learning from the statistical regularities of the speech input, i.e. frequency, distribution, variability, transitional probabilities, etc. Concerning phonetic acquisition, in-laboratory experiments suggest that infants use distributional information to discriminate between sounds (12–14). Regarding word learning, in a seminal experiment, Saffran et al. (15) used an artificial grammar to show that infants can track transitional probabilities across syllables to identify word boundaries. Since then, statistical learning has been studied in countless experiments across ages, domains, and species (16). Although there is a consensus among researchers that infants are sensitive to statistical regularities of their speech input, the extent to which statistical learning can explain language acquisition is at the heart of heated debates (17, 18).

One of the most prominent criticisms of the statistical learning hypothesis is that infants are embodied in a much more diverse and complex environment than what is typically present in laboratory experiments. In particular, many experiments use synthetic stimuli and artificial languages, which has the undeniable advantage of isolating the contribution of individual variables, but makes it hard to generalize to real-life language input. Indeed, two critical aspects of natural language are missing in artificial languages used in experiments. First, language is highly variable. However, in word segmentation experiments, it is common to employ artificial languages where every word shares the same length; when more variability in length is introduced, infant's ability to use transitional probabilities to segment words is severely diminished (17). Similarly, sound discrimination experiments use prototypical sounds and cherry-picked contrasts that fail to account for the large variability found in natural languages (19). Second, language is hierarchically organized into linguistic levels. In artificial languages, variability is typically frozen from all levels except the one under study. For instance, in phonetic learning experiments the language introduces phonetic variations (usually along a single dimension) but is made

Author contributions: M.L., M.S, H.B., G.W., A.C. and E.D. designed research; M.L. and M.S. performed research; H. T. created the lexical evaluation set; M.L. and M.S. analyzed data; M.L., M.S, and E.D. wrote the paper with contributions from H.B., G.W., and A.C.

The authors declare no competing interest.

[1]M.L and M.S. contributed equally to this work. Authorship order was decided by a coin flip.

[2]Address for correspondence: marvinlavechin@gmail.com or maureen.deseyssel@gmail.com

only of two monosyllabic utterances. Vice versa, in lexical learning experiments, the language contains more syllables and long "utterances", but syllables are identical copies with no phonetic variability or coarticulation effects. Even though infants have been shown to use statistical learning mechanisms in these simplified languages, could similar mechanisms work when faced with the complexity and variability of real languages and reproduce some of the observed developmental patterns?

In face of the lack of ecological validity of laboratory experiments, one possible answer consists of building computational models of language acquisition, adopting the reverse-engineering approach (20). After all, if language acquisition occurs through statistical learning, algorithms should be able to reproduce behavioral patterns observed in infants when fed with similar input. Unfortunately, the development of language learning algorithms addressing the full complexity and variability of language from raw speech input is not an easy enterprise (see (21) for a review). This is why early attempts at simulating language acquisition through computer models had also to resort to simplifying assumption and/or focus only on one aspect of language at a time. For instance, early statistical models of phonetic learning (22) did not use real continuous speech input but synthetic data generated from average formants measured in isolated syllables. Similarly, statistical models of word learning worked not from real speech, but from phonetic transcription of this input by adults who have already learned the language(23), thereby implicitely assuming that phonetic learning is completed before word learning can take place. These algorithms are useful in advancing our understanding of language acquisition as they provide proofs of learnability under certain hypotheses. However, to the extent that their simplifying assumptions are not met in real life, they do not allow to assess whether statistical learning can really address the full complexity and variability of language, from lower-level sound units to higher-level word units.

Recent advances in machine learning have provided some hope that some of these roadblocks can be lifted. For instance, Schatw et al. (24) proposed a phonetic learning model that, for the first time, learns from raw speech. They showed that a representation learning algorithm based on mixtures of Gaussian applied English or Japanese recordings could reproduce patterns of phonetic attunement as found in infants. Hitczenko et al. (25) showed that, even though language-specific statistical patterns are often obscured by the variability in running speech, such variability can be reduced by taking into account a larger window of analysis incorporating local phonetic context. Both studies constitute substantial evidence in favor of the feasibility of statistical learning hypothesis for early phonetic development. However, both studies are still only addressing one linguistic level in isolation. Would learning algorithms as applied to raw speech result in sufficiently abstract representations to sustain learning at other levels? This question is not a trivial one, given that Schatz et al. (24) found that their model was unable to converge to interpretable phonemic or even phonetic categories. Is it possible to learn words or syntax on top of such non-linguistic representations? In other words, is statistical learning restricted, in practice, to patterns of attunements to the phonology of the native language? Or can higher levels of language acquisition be reached through statistical learning?

Here, we introduce STELA (STatistical Learning of Early Language Acquisition), a learning simulation addressing for the first time the joint learning of phonetic and lexical information from raw speech. Building on recent advances in speech processing and unsupervised representation learning (26, 27), we show in Experiment 1 that a neural network trained to predict the near-future from raw speech signal, and tested with psycholinguistically-inspired discrimination and preference tasks can reproduce gradual and simultaneous learning at both the phonetic and lexical levels. At the phonetic level, the network is increasingly better at discriminating native than non-native sounds, reproducing the so-called perceptual narrowing effect documented in infants (2). At the lexical level, the network reproduces patterns of preference for real words over pseudo-words, i.e. non-existent but plausible words (9). This constitutes the first demonstration that statistical learning is *sufficient* to bootstrap early phonetic and word learning in a simultaneous fashion. In Experiment 2, we investigate whether the learned representations correspond to interpretable linguistic categories. We show that, as the quantity of speech received by the network increases, phonetic and grammatical categories become more linearly separable in the learned representations. However, the learned acoustic representations remain shorter and more variable than phonetic categories (24). A similar phenomenon occurs at the lexical level: the network does not explicitly represent words or word boundaries. Thus, in addition to providing a proof of feasibility to statistical learning potentially explaining multilevel language learning, our STELA simulation further suggests a new hypothesis, i.e., that linguistic categories are not needed to account for patterns of early language development. In the General Discussion, we discuss the consequences of these findings for theories of early language acquisition.

Approach

STELA follows the reverse engineering approach described in (20) whereby a full computational simulation of language acquisition addresses three components of the learning situation: the environment, the learner, and the outcome measures (see Figure 1). Here, we give only a high-level sketch of these components described in more details in the Methods Section.

As in (24, 28), we take the *environment* of the infant to be composed of raw speech input. Here, we extract 3,200 hours of speech from French and English audiobooks, which corresponds to the upper limit of what infants could hear during the first three years of their life (29). For each training language (English or French), we built training sets by randomly splitting the whole set of audio segments into mutually exclusive training sets of 50 hours. These 50-hours training sets were then merged two by two to build the 100-hours training sets. This procedure was repeated until convergence, which left us with 64, 32, 16, 8, 4, 2, and 1 training sets of 50h, 100h, 200h, 400h, 800h, 1,600h and 3,200h of speech.

We simulate the *learner* by using the winning entry of the ZeroSpeech 2021 international challenge on unsupervised representation learning (26). It consists of two components. The *Acoustic Model* takes as input raw audio and outputs a discrete unit every 10ms slice of time. The *Language Model* takes the discretized version of the audio as input and outputs a prediction for the next units, similarly to text-based language

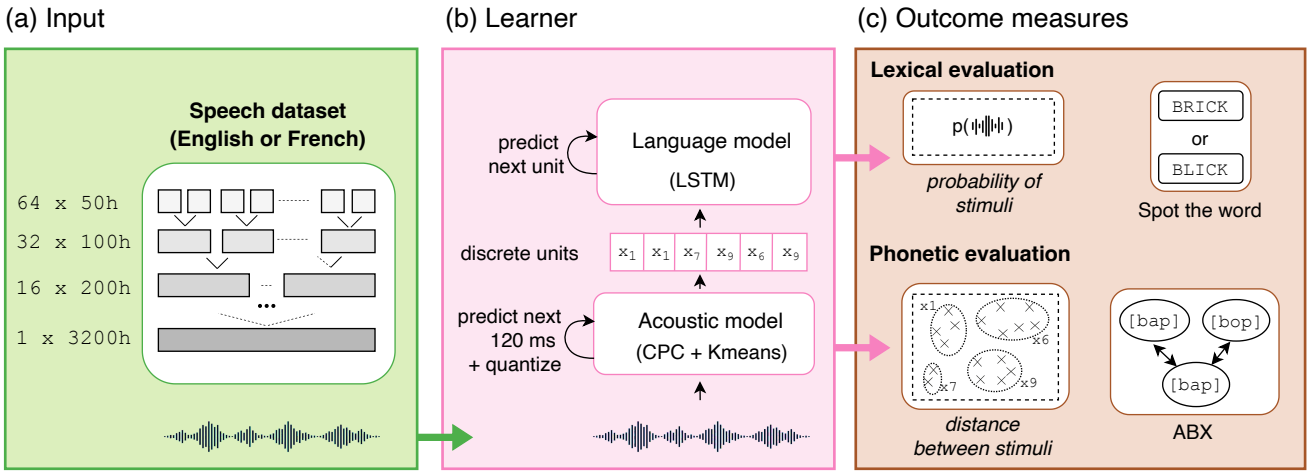


Fig. 1. Overall setup for the training and testing of STELA. a. The audio environment of learners of different 'ages' are modeled using audiobooks segmented and aggregated in increasingly larger sets matched for number hours and of speakers across two languages (See Table 1). b. The learner is composed of an 'Acoustic Model', first trained with predictive coding and followed by a K-means algorithm returning discrete units and a LSTM 'Language Model' trained to predict future units based on past units. c. Outcome measures are obtained by modeling an ABX sound discrimination task at the (discretized) output of the Acoustic Level, and an auditory lexical preference task (Spot-the-Word) by using the ability of the Language Model to compute the probability of stimuli.

models except that the latter are trained on words. The two components are trained by minimizing self-supervised objective functions on the same chunk of data. In other words, the model learns from the raw speech only, without any human annotation intervening in the loop. It thus obeys a critical constraint for modeling infant language development. Children are never explicitly given linguistic knowledge, so neither should computational models. Once trained, a model constitutes a simulation of an infant exposed to a particular language for a given amount of exposure.

We measure our learners' language *outcomes* at two linguistic levels: the phonetic level (sounds) and the lexical level (words), drawing inspiration from psycholinguistic studies (see Section A.3). At the phonetic level, we simulate an ABX auditory discrimination task using phonetic contrasts, e.g. /ɪ/ versus /ɛ/as in "bit" versus "bet". At the lexical level, we simulate a spot-the-word task: the model is asked to identify which of two audio stimuli (e.g., "brick" and "blick") is a word (the former), and which is a pseudo-word (the latter). For each trained model and each target language, we obtain a phonetic and a lexical score, such that 100% and 50% indicate perfect and chance-level accuracy, respectively. We compute the average phonetic and lexical scores in the native condition (the English model evaluated on English, and the French model evaluated on French) and the non-native condition (English model on French, French model on English). Contrary to humans, machines can be presented with thousands of trials for a given stimulus type (words or phonetic contrasts), allowing us to extract robust measures of learning outcomes.

The comparison between native and non-native scores allows us to identify what our model has learned due to exposure to its native language (as opposed to exposure to another language). In other words, the non-native model acts as a control for the native model. By assessing our models' language capabilities as a function of the quantity of speech they have been exposed to, we draw developmental trajectories and ask whether or not learning outcomes exhibited by our model share similarities with infant language development. Finally, we supplement these two tasks with in-depth analysis of the representations learned by the system.

1. Experiment 1 : Can statistical learning bootstrap both phonetic and lexical learning?

The objective of our first experiment is to investigate whether our model demonstrates phonetic and lexical learning outcomes and whether such learning occurs gradually and concurrently, similar to how it does in infants, as aligned with the primary question presented in the introduction.

A. Material and Methods. In this section, we provide a more comprehensive description of the model's implementation, including details on the input data, learner design and outcome measures.

A.1. Training sets. We used 10,000 hours of English audiobooks from the Librivox platform (30) and 10,000 hours of French audiobooks from litteratureaudio (31). We constructed 64 twin chunks of 50 hours of speech (3200 hours total) made of entire book chapters in each language, such that the number of speakers would be as matched as possible across the two languages. To achieve this, we applied a stochastic sampling algorithm that matches across English and French: 1) the cumulated duration, 2) the number of speakers per chunk of 50h, and 3) the number of chunks per speaker. We then randomly aggregated the 64 chunks of 50 hours two by two to obtain 32 chunks of 100, until we obtained one large 3200h chunk in each language. Table 1 provides further statistics that demonstrate the matching between the English and French training sets.

Table 1. Statistics for the French and English training sets varying in quantity of speech. Average number of speakers per training set, average quantity of speech for the least talkative and the most talkative speaker per training set.

Training	French			English		
sets	N	min (h)	max (h)	N	min (h)	max (h)
64x50h	9.7	0.33	16.96	9.7	0.75	15.84
32x100h	17.0	0.19	24.11	17.3	0.55	20.81
16x200h	28.7	0.14	35.61	29.6	0.41	29.90
8x400h	46.9	0.06	58.45	49.1	0.32	45.22
4x800h	73.7	0.05	94.84	74.7	0.23	75.43
2x1600h	107.0	0.04	187.89	108.5	0.19	133.75
1x3200h	147.0	0.01	334.17	147.0	0.17	267.50

A.2. Learner design. We describe below our proposed model learning speech representations from the raw waveform (26). The learner is composed of two components: 1) the Acoustic Model that learns discretized representations of the the raw waveform, and 2) the Language Model that takes the discretized representation as input and returns a probability distribution over the set of discrete units.

The acoustic model. It consists of a Contrastive Predictive Coding (CPC) algorithm (27, 32). The key idea behind CPC is to predict the near future of a sequence given its past context (see Appendix for more details). The learner is given an example that is drawn from the near future up to 120 ms (called positive example), and multiple examples that are not drawn from the near future (called negative examples). Given the past context of a sequence, the learner is asked to maximize the categorical cross-entropy of classifying the positive sample correctly (see Appendix 1.1 for more details). The continuous context-dependent representations output by CPC are then fed to a simple K-means clustering algorithm that returns a discrete representation of the audio.

The language model. The language model takes as input the discrete representation of the audio file returned by the acoustic model. It is trained to predict the next discrete unit via a cross-entropy loss function (see Appendix 1). At test time, the model is simply used to produce a probability of a stimulus $S = q_1, q_2, ..., q_T$ by applying the following formula:

$$P(q_1, ..., q_T) = -\frac{1}{T}\sum_{t=1}^{T} \log p(q_t \mid q_1, ..., q_{t-1})$$

Based on this probability, it becomes possible to simulate a preference task between two stimuli. A stimulus A is preferred over B if its probability, as estimated by the language model, is higher.

A.3. Outcomes measures. **Phonetic evaluation: the machine ABX sound discrimination task**

General principle. The ABX sound discrimination task was first proposed by (33) to offer a way to evaluate models' phonetic discrimination capabilities in a setup comparable to how humans are evaluated. The task consists of generating a wide range of triplets of sounds in the format A, B and X, with A and X corresponding to different variations of the same triphone ('bop') and B to another triphone where the central phone changes ('bap'). Distances between A and X, and B and X are then computed using Dynamic Time Warping (DTW) based on a frame-to-frame cosine distance. A score of 1 is given if $d(A, X) < d(B, X)$, otherwise the score is 0. An average score is finally computed over all possible triplets.

The ABX task can be used on any type of speech representation, and has already proven robust with the CPC+K-means architecture presented here (26). In this paper, we use the discrete representations output from the K-means algorithm to compute the ABX score.

Materials. The triplets are generated over carefully tailored English and French speech test sets, which are subsets of the CommonVoice dataset (34). These test sets, already presented in (35) and (28), consist of 10 hours of read speech balanced between 24 speakers (12 males and 12 females). All utterances from the English and French test sets are tagged as "US accent" and "France accent" respectively in the original CommonVoice dataset. The phone-level alignment was obtained by aligning the audio stream with its transcript using Kaldi recipes (36), eventually allowing us to generate triplets for the ABX task. The ensuing phonetic inventory in International Phonetic Alphabet (IPA) standard for both languages is shown in Table S1.

Lexical evaluation: the spot-the-word task

General principle. The evaluation of lexical knowledge in a recurrent neural network was first proposed in (37) using the spot-the-word task. It consists of presenting the network with a minimal pair of word and non-word (e.g., 'brick' versus 'blick') and evaluating whether the probability given by the network to the word is higher or lower than the probability given to the non-word, yielding an accuracy score, which was averaged across all of the pairs in the test set.

Materials. The pairs are constructed using the Wuggy toolbox (38), which generates lists of nonwords matched for syllabic and phonotactic structure with a given list of words. To build our test set, we first selected the list of words present in our environments and constructed for each word a set of associated non-words using Wuggy and pronunciation dictionaries for French and English (39). We then reduced this list to a single non-word by applying a filter maximizing the frequency of unigrams and bigrams of phonemes between the words and the non words. We then synthesized the words and non-words using the Google text-to-speech API (40) in 4 voices (2 males, 2 females) in each language.

The resulting list of word/non-word pairs was further sorted into frequency bands by intersecting them with the different environments. The highest frequency band was constructed by selecting the words that appeared at least once in each of the 64 50-hours environments. The second highest frequency band was made of words that appeared at least once in each of the 32 100h environments and that were not in the preceding list, and so-forth until we had the corresponding 7 frequency bands. In Figure 2, we only displayed the results of the highest frequency bands. The results for each frequency band can be found in Supplementary Figure S5.

B. Results and discussion. Panels (a) and (b) of Figure 2 show the scores obtained on the phonetic and lexical tasks, for the native and the non-native learners, as a function of input quantity. Results indicate that native models trained on 3,200 hours of speech succeed in discriminating sounds (81.64% phonetic score) and, to a lesser extent, recognize the auditory form of words (62.98% lexical score).

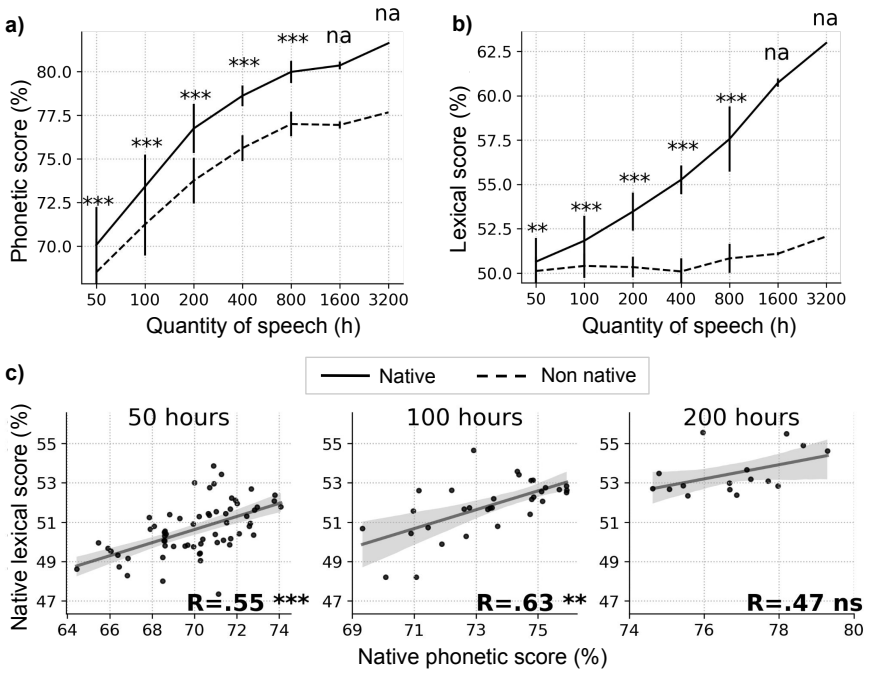


Fig. 2. Gradual and parallel learning across the phonetic and the lexical levels. a) Phonetic score, in terms of ABX accuracy, obtained by the discrete representations for native and non-native input. b) Lexical score, in terms of accuracy on the spot-the-word task, on the high frequency words for native and non-native input. For a) and b), two-way ANOVAs with factors nativeness and training language were carried out for each quantity of speech. Significance scores indicate whether the native models are better than the non-native ones. c) Correlation between the phonetic and lexical scores obtained across individual native models trained for 50h, 100h and 200h in English and French. R is the Pearson correlation coefficient. Significance levels: na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

The developmental aspect of STELA also allows us to assess the evolution of the learning trajectories. In the native condition, both phonetic and lexical scores increase gradually as a function of quantity of speech. Phonetic and lexical scores obtained by the native model are systematically higher than those obtained by the non-native model. This difference increases with input quantity, reaching a relative difference of 5% for the phonetic score, and 18.97% for the lexical score between our native and non-native learners trained on 3,200 hours of speech. Using two-way ANOVAs with factors nativeness and training language, we found that native scores were significantly higher than non-native scores for as little as 50 hours of speech ($F(1, 252) = 18.95$, $p<.001$ for the phonetic score, $F(1, 252) = 15.81$, $p<.0001$ for the lexical score). Significance tests on 1,600 and 3,200 hours of speech are not available due to the low number of models. To summarize, the proposed algorithm learns key aspects of its native language at both the phonetic and the lexical levels in a gradual and simultaneous fashion, consistently with what has been observed in young infants (7, 41–43).

The phonetic score obtained by the non-native model improves with input quantity (as previously noticed in (24)). This developmental pattern might seem to run counter to experiments that report a loss in non-native sound discrimination in infants (2). However, our setup differs from the usual infants experiments, as we systematically average performance over all possible phonetic contrasts in the present study (see Supplementary Table S1 for the list of evaluated phonemes, and Supplementary Section 4 for similar comments on the non-native lexical score). In infant studies, the non-native sound discrimination loss was documented only for a small number of carefully selected phonetic contrasts which are known to be difficult for the non-native language tested (such as the "r" versus "l" as in /rock/ versus /lock/ in Japanese infants). Besides, we know that many non-native contrasts map onto native ones (44), which would explain the phonetic learning even in the non-native condition. For instance, interdental fricatives can map from one language to the other (/s/ and /θ/ in English map to /s/ and /z/ in French). The increase in phonetic score by the non-native model is an interesting observation that could be tested in infants. On the other hand, this non-native learning is not observed in the lexical task. This was expected as, contrary to the phonetic task, there is no overlap between auditory word forms in the two languages.

Further evidence for lexical learning in the native condition is provided by an additional analysis (Supplementary Section 7) showing that the higher the frequency of the evaluated words, the higher the lexical score obtained by the native model. This frequency effect has been widely documented in young infants and has been argued to be an important requirement for any successful account of language acquisition (45). Investigation of a large-scale study of human reaction times in auditory lexical decision (deciding whether a word exists) revealed that word probabilities computed by the native model correlate with linguistic factors shown to influence human lexical decision times (such as the duration, the frequency and the number of phonological neighbors of the word; see Supplementary Section 8). All in all, we found evidence for learning at the phonetic

and lexical levels using an algorithm exclusively based on statistical learning.

Two models exposed to the same quantity of speech can perform differently on the phonetic and lexical tasks. This is due to: 1) the training set itself that may constitute a more or less adequate language experience; and 2) the randomness in the weights' initialization and in the way data is presented, which may advantage or disadvantage the model. With this in mind, we can attempt to characterize the relationship between phonetic and lexical outcomes obtained by our models. Panel (c) of Figure 2 shows significant positive correlations between the scores obtained on the phonetic and lexical tasks, respectively, across models trained on 50h, 100h, or 200h of speech (there were fewer than 8 models trained on larger quantities of speech, not enough to compute meaningful correlations). This result indicates that models that are better at discriminating native sounds are also better at solving the spot-the-word task. This is compatible with infant studies suggesting a positive correlation between native discrimination and vocabulary size at 11 months (46, 47). Similarly, multiple longitudinal studies show that early sound discrimination capabilities predict later language development (48–50). Further work could assess specifically whether there exists a positive correlation between native discrimination and auditory word form recognition.

All the analyses presented in this section can also be found separately across the English and the French model in Supplementary Sections 4 and 5.

2. Experiment 2 : Are linguistic categories required?

In the previous Experiment, we have shown that our models improve in both lexical and phonetic tasks, more so for native than non-native tests, which parallels findings with human infants. In an attempt to better understand the nature of the learned representations, we dedicate the current section to a deeper analysis of how similar these representations are to linguistic categories.

A. Methods. Additional analyses are carried out to compare the model's representations to linguistic categories. Linguistic categories are analyzed at two levels: at the acoustic model for the phonetic categories (phone class/sonority, place of articulation, and voicing) and at the language model for the lexical categories (broad *function vs. content word* differentiation, and content words' part of speech). For these analyses, the same English and French test sets as presented in Section A.3 are used, consisting in speech-to-phones and speech-to-words alignments.

For each category, a qualitative and quantitative analysis is run. The qualitative analysis consists of a 2D visualization of the output speech representations from the model trained on most data (3200h), colored in terms of their linguistic category. We extracted the output acoustic (language) model representation of every test sentence in the language the 3200h model was trained on (the representations are therefore context-dependent). We then extracted the representation for every phone (word) and used a mean-pooling function to obtain a fixed-dimension representation for each of these phones (words). We applied a t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to reduce the representations to 2 dimensions on a subset of the data (N=6,000). Finally, we plotted the resulting dimensions and color-coded the data points based on their target characteristic category (phone class / place of articulation / voicing / part of speech).

The second, quantitative, analysis focuses on the emergence of linguistic categories as a function of input quantity. This allows an understanding of whether the models' speech representations become closer to the linguistic categories when trained on more data. To do so, we split the test set into sub-training and sub-test sets. The sub-training set contains models' representations of all phonemes minus one phoneme. The sub-test set contains models' representations of the final phoneme (the same is done at the lexical level with representations of 50 word types per category chosen out of the 100 most frequent word types - the other 50 being used as a development set - see below). A logistic regression model is then trained for all sub-training set representations, using the desired linguistic categories as targets, before being tested on the sub-test representations. This process is done iteratively with all possible phonemes (word type) being part of the test set, using Leave-One-Out cross-validation. This allows us to retrieve an average classification error for the model on the specified information. This is done on all models of all different training sizes, allowing us to draw developmental curves of these classification errors. The chance classification error and error calculated on raw MFCCs were also computed. Finally, we check the significance of the developmental curve's slope (correlation between classification score and quantity of input) using Spearman's rank correlation.

For the phonetic analyses, representations were extracted from the last hidden layer of the CPC model (these are the same representations used for the ABX task). We use all phones for the 'sonority' analyses, however we only keep consonants for the 'place of articulation' and 'voicing' categories, as vowels are not relevant here*. Regarding the lexical analyses, we chose the hidden layer yielding the best classification error scores on the 3200 hours model, using a development set also formed of 50 word types per category sampled out of the 100 most frequent word types per category. The best hidden layer was the third (last) for both the English and French models (logistic regression scores on all layers are available in Supplementary Table S2).

B. Results and discussion.

B.1. The emergence of phonetic categories. Although our model works with 10-ms frames, the Acoustic Model often assigns the same discrete unit to multiple successive frames. Do these duplicated discrete units share commonalities with phonemes, in terms of duration and perplexity? Our analysis reveals both that these units are much shorter than actual phonemes (see Figure 1), and that a same unit can encode multiple phonemes (see Supplementary Section 10). These conclusions mirror results with a different acoustic model found in (24). We also found that this pattern does not change with the amount of training data. If anything, the learned units become shorter as the amount of data increases (top graphs of panels (b) and (d) of Supplementary Figure S8). At the same time, we observe an opposite trend for the number of units associated to each phoneme: the unit-to-phone perplexity decreases with input quantity. This indicates that the more speech the model receives, the more fine-grained the learned discrete units are.

*We also discard approximants from the place and voicing analyses, as well as the English h and the French ʁ, as they are alone with their place of articulation label.

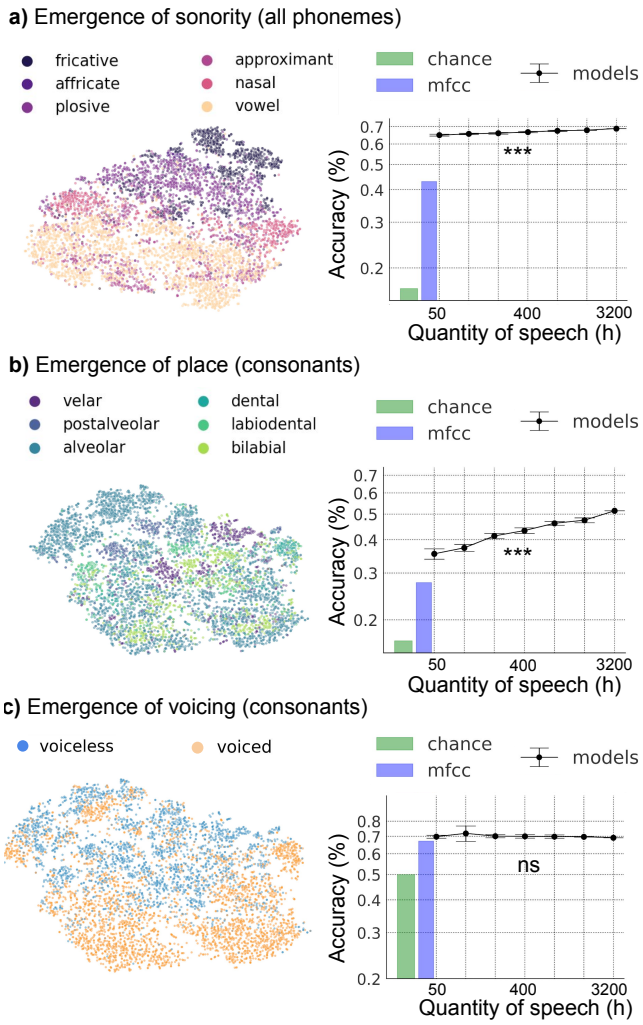


Fig. 3. Emergence of latent linguistic categories at the phonetic level for the English models. Left: tSNEs of the continuous representations of the acoustic model (last layer) pooled within phones in a test set, according to sonority (a), place (b) and voicing (c) for the 3200h English model. Right: developmental curves from a leave-one-phoneme-type-out classification errors as a function of input quantity (taking all 256 dimensions into account). Chance level and MFCCs performances are also given. The asterisks indicate a significant correlation of classification error and input quantity. na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

Although linguistic categories are not hard-coded and therefore do not exist *per se* in our model, could the learned representations encode important linguistic information? Using a t-distributed stochastic neighbor embedding (t-SNE) method on the representations learned by the English acoustic model trained on 3,200h of speech, Figure 3 shows that the learned acoustic representations encode multiple phonetic features. Phone representations are organized along a continuum spanning from sounds that are very sonorous (vowels) to not sonorous (fricatives) (panel (a)). Similarly, consonant representations are clustered by place of articulation (place where the constriction and obstruction of air occur when producing the consonant), and by voicing (whether or not produced with vocal cord vibration) (panels (b) and (c)).

The projection of high-dimensional representations in 2D spaces results in an important loss of information and constitutes only a qualitative analysis of the learned representations. Therefore, we use logistic regressions as probes to analyze quantitatively the information encoded within the models (51, 52). We train a linear classifier on top of the continuous acoustic features to measure the extent to which previously studied phonetic features (sonority, place of articulation, and voicing) are present in the learned representations (Figure 4). We compare the classification scores of our probes with those obtained both by a random linear classifier (representing chance level, in green) and by one trained with mel-frequency cepstral coefficients (MFCCs, representing acoustic representations, in blue). Results indicate that sonority, place of articulation, and voicing are encoded in the learned representations even by the model trained on the smallest quantity of input (50h) of English speech, since all scores are better than both the random baseline in green and the acoustic representations in blue. Classification errors on sonority and place of articulation improve with data quantity, showing a positive effect of exposure. This does not hold for voicing, for which the linear classifier obtains a high classification score. Equivalent analyses on the French model and further details can be found in Supplementary Section 6.

Results presented in this section show that, although the learned representations are too fine-grained to correspond to phonetic categories as defined by linguists, they nonetheless contain information that encodes critical phonetic features. In addition, our study found that for two of the three phonetic features we examined, such a perceptual organization emerges in a gradual fashion, with a positive effect of input quantity.

B.2. The emergence of lexical and grammatical categories. Next, we look at whether lexical and grammatical information is present in the representations learned by the language models. We follow the same procedure as above and analyze word representations in a qualitative way using t-SNE and in a quantitative way using linear classifiers. In particular, we probe two dimensions: 1) the distinction between function and content words; and 2) part-of-speech categories among content words (nouns, verbs, adjectives).

Experimental studies suggest that infants know at least some of the function words of their native language around one year of age (53), and that they use this information to infer part-of-speech categories among content words in their second year of life (54, 55). Mainstream theories like prosodic bootstrapping hold that both the distinction between function and content words, and the part-of-speech categories among content words are crucial cues in early language acquisition, particularly in lexical segmentation and syntactic parsing (56).

A 2D t-SNE projection of the word-level representations learned by the language model does not reveal a clear separation between function and content words (left of panel (a), Fig. 4), although some regions of the space seem specific to each grammatical class. The same conclusion can be drawn when coloring content word representations according to their part of speech categories (left of panel (b), Fig. 4).

However, it is not because t-SNE does not exhibit a clear separation between linguistic categories that the information is not present in the learned representations (as mentioned in the previous section, t-SNE leads to a loss of information). As a matter of fact, linear probing on the learned representations suggests that linguistic information is indeed present. Specifically, it is possible both to classify whether a word is a

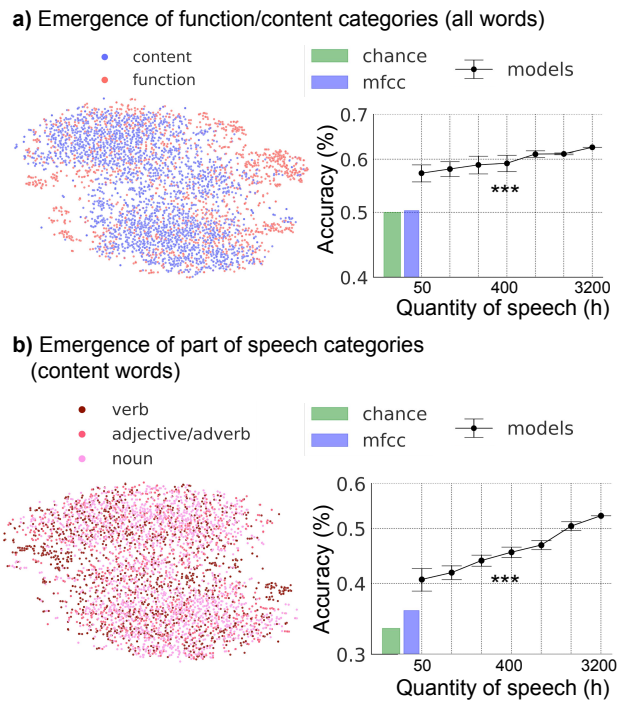


Fig. 4. Emergence of latent linguistic categories at the word level. Left: tSNEs of the continuous representations of the language model (last layer) pooled over words according to (a) function/content distinction and (b) part of speech for the 3200h English model. Right: corresponding developmental curves of leave-one-word-out classification error as a function of input quantity (taking all dimensions into account). Chance level and MFCCs performances are also given. The asterisks indicate a significant correlation of classification error and input quantity. na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

function word or a content word (right of panel (a), Figure 4) and to classify the part-of-speech categories of content words (right of panel (b), Figure 4). Linear classifiers trained on the learned representations of the language model are indeed better than chance and better than the MFCC baseline on both classification tasks. Importantly, accuracy increases when the representations are learned on a larger quantity of data, showing that categories become more linearly separable as the quantity of speech increases, showing a positive effect of exposure (p<.0001).

All in all, results in this section suggest that the learned representations are somewhat structured by word categories. This organization emerges in a gradual fashion, with a positive effect of input quantity.

3. General Discussion

Whether statistical learning can account for infant early language acquisition, despite plethora of experimental infant studies on the topic (see (16) for a review), still remains an open question (17, 57, 58). Besides studies carried out directly on infants, some computational models showed the feasibility of language acquisition in statistical learning, but these models either only focused on a single aspect of language acquisition (phone discrimination (24); word learning, (59)), or they made strong assumptions on the input data (using processed signal or text), making their plausibility as a model of infants questionable. Recent studies (24, 25) provided strong evidence in favor of the statistical learning hypothesis for early phonetic learning. Can statistical learning account for higher-level aspects of early language learning?

STELA constitutes the first proof of feasibility of statistical learning to account for early language acquisition at the phonetic and lexical level. We further showed that phonetic and lexical learning was possible without linguistic categories. More generally, we proposed the first developmental psycholinguistic analysis of a state-of-the-art machine learning model. In this section, we reflect on STELA key findings and limitations.

Statistical learning is enough to bootstrap language learning In our STELA simulation, we have shown for the first time that a self-supervised learning model built within the scope of the statistical learner hypothesis can reproduce developmental patterns of gradual learning at both the phonetic and lexical levels when provided with untranscribed, raw, clean speech data. The model works by implementing a neuro-cognitively motivated statistical learning mechanism (predictive coding) within a linguistically interpretable division of levels (discrete acoustic vs. high-level abstract), trained on raw audio data. We found that linguistic knowledge at these two levels (phonetic and lexical) emerges gradually and in parallel, as attested by psycholinguistic-inspired tests and analyses. Such results constitute strong proof of feasibility for the statistical learning hypothesis, suggesting that such computations are sufficient to bootstrap phonetic and lexical knowledge when provided with raw, clean speech.

However, there are several limits that need to be addressed before claiming that statistical learning alone can bootstrap the entire linguistic system. First, we only analyzed two linguistic levels: phonetic and lexical, the latter being restricted to word forms. Bootstrapping language would require to show the other linguistic levels that have been documented as emerging in young children (prosody, syntax, semantics) also emerge thanks to the same mechanisms. Specific tests inspired by infant psycholinguistics probing these levels would need to be developed and applied to the model[†]. Second, we used as input audiobooks, which are much less noisy than the audio available to infants. It is possible that additional mechanisms besides statistical learning are needed to cope with such variability (28).

Finally, infants are much more than simple statistical learners, and previous studies have found that cross-modal learning and social interactions play a significant role in infants' language acquisition (58, 64–66). We want to point out that our study does not question this, and that these other types of input could well be critical in the development process. Instead, our proof of feasibility shows that relying only on a statistical learning mechanism to start bootstrap language is possible.

Phonetic and lexical learning without linguistic categories The seminal work carried out by Schatz et al. (24) suggested that statistical learning can be used to reproduce developmental patterns in phonetic learning without creating phoneme-like units, therefore questioning the presence of such categories in infants (see also (67)). Analyses carried out on the representations learned by our model point in the same direction:

[†] Work in spoken language modeling (60–63) suggests that these levels can emerge from statistic mechanisms applied to the raw speech, but such models typically require much more input data than is available to infants and it remains to be seen that they can reproduce plausible developmental curves.

the learned units do not correspond to the usual phonetic categories. The discrete level itself (acoustic units) is not linguistically interpretable and does not tend to become more phoneme-like with more input data, but rather to correspond to more fine-grained sub-phonetic units, mirroring Schatz' results with a different model.

Examining the lexical level for the first time, we surprisingly found a similar pattern: the learned representations do not directly map to word-level categories such as part-of-speech. Two main lessons can be drawn from these results: 1) sub-phonetic units are sufficient to learn higher-level aspects of language; and 2) word categories are not required to recognize the auditory form of words. Thus, our work questions the need for any linguistic categories, and not only phonetic ones, in the early stages of language acquisition. Similarly, even though we found evidence for the emergence of lexical and grammatical information, this information does not seem to be grounded into a segmentation of the input into word-like chunks (see Supplementary Section 9).

Gradual and parallel learning in STELA Within the STELA framework, we introduced a carefully designed developmental setup, which allows us to compute the effect of quantity on phonetic and lexical learning, to generate their respective developmental curves, and to compare them to experimental results.

At a qualitative level, the developmental curves show a gradual and parallel increase in both phonetic discrimination and lexical preference. How can we account for such a pattern? Our algorithm works by minimizing quantities called *loss functions*. We use three such functions that are minimized jointly. The acoustic model minimizes the prediction errors over continuous acoustic representations (predictive coding) and then discretize them using a compactness score (discretization). The language model minimizes the prediction error over the discrete units. The two prediction errors are minimized by the stochastic gradient descent algorithm and the compactness score by a variant of the expectation-maximization algorithm. The gradual and parallel aspect of the results is due to the fact that these three loss functions are optimized to lower values as more data is presented (see Supplementary Figure S1).

Does statistical learning actually bootstrap early language acquisition? The core demonstration of our work consists in showing that a statistical learning mechanism can exploit the information present in the raw speech signal and reproduce patterns of early stages of language acquisition, such as measured by our psycholinguistically-inspired evaluation tasks. This shows that infants could rely on statistical learning mechanisms to bootstrap language acquisition, but it does not show that they necessarily do.

In other words, while STELA is valuable in providing a proof of feasibility of the statistical learning mechanism in early language learning, it cannot at present be considered a fully fledged model of the infant because of several limitations. One limitation of the current implementation of STELA is that while it provides a series of cross-sectional predictions (by simulating infants of different ages), it does not allow for longitudinal studies: models are trained anew for every quantity of input, and led to convergence every time. Implementing a longitudinal framework would require larger datasets, with, ideally, each training set representing a single child's input, and a modification of the learning algorithm to yield incremental results for each increasing amount of input.

Regarding the model of the learning outcomes, although heavily inspired by psycholinguistic experiments, it does not directly simulate the experiments as they are run in a laboratory setting: preferential looking, high-amplitude sucking, etc. These procedures have been designed to explore processes at different stages of the infant's speech perceptual development and aim at eliciting specific behavioral responses from the infant. In this regard, the machine evaluation tasks are far simpler and directly interpretable in terms of: 1) distance between sound representations for the phonetic evaluation; 2) prediction error of words and pseudo-words for the lexical evaluation. The next challenge will likely consist in allowing better comparison between infants' language learning outcomes and those obtained via our in-silico simulations, i.e., moving beyond qualitative comparison. The noise inherent to infants' behavioral responses might prevent us from doing that in the near future, but a promising approach might consist in comparing learning outcomes obtained by the machine against large-scale cumulative empirical infant data.

Finally, the model of the environment adopted in the present study used relatively well-articulated speech without background noise. As shown in (28), infants have the additional task of separating speech from noise, which is not taken into account in the present simulation. Once these limitations are addressed, it may be possible to more directly compare the predictions of STELA with actual infant's outcomes, and validate or invalidate it as a possible model for early language acquisition.

4. Conclusion

Overall, this proof of feasibility shows that self-supervised learning models are good *a priori* candidates to help us understand trajectories in infant language development. Machine learning solves deep puzzles in cognitive development and provides quantitative models that make numerical predictions as a function of the amount of input data. While this proof of feasibility shows that phonetic and lexical bootstrapping is possible using only statistical learning mechanisms, there remain many challenges, including going further towards ecological audio data and benchmarking against actual infant experimental results. Even more challenging will be the issue of closing the gap between computational models and the actual cognitive learning processes used by infants: To what extent do infants actually make use of statistical learning mechanisms during language acquisition? And what is the place of other mechanisms (social learning, intrinsic motivation) in the developmental pathway?

STELA offers the potential to simulate the entire language acquisition process in the early years of life using a fully implemented model that operates on real audio input. This could generate a wealth of quantitative predictions that can be compared to data on infants. By open sourcing the model, we hope to inspire a shift towards a more quantitative approach in infant research.

ACKNOWLEDGMENTS. We are especially grateful to Sharon Peperkamp for enlightening discussions and proofreading sessions. We are grateful to LAAC, and CoML members for helpful discussion. All errors remain our own. A.C. gratefully acknowledges financial and institutional support from Agence Nationale de la

Recherche (ANR-16-DATA-0004 ACLEW, ANR-17-EURE-0017); the J. S. Mc-Donnell Foundation (Understanding Human Cognition Scholar Award); and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (ExELang, Grant agreement No. 101001095). E.D., in his academic role (EHESS), acknowledges funding from Agence Nationale de la Recherche (ANR-19-P3IA-0001 PRAIRIE 3IA Institute), a grant from CIFAR (Learning in Machines and Brains), and the HPC resources from GENCI-IDRIS (Grant 2020-AD011011829). M.S. acknowledges PhD funding from Agence de l'Innovation de Défense.

1. RE Eilers, WR Wilson, JM Moore, Developmental changes in speech discrimination in infants. *J. Speech Hear. Res.* **20**, 766–780 (1977).
2. PK Kuhl, et al., Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Dev. science* **9**, F13–F21 (2006).
3. FM Tsao, HM Liu, PK Kuhl, Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *The J. Acoust. Soc. Am.* **120**, 2285–2294 (2006).
4. Y Sato, Y Sogabe, R Mazuka, Discrimination of phonemic vowel length by japanese infants. *Dev. Psychol.* **46**, 106 (2010).
5. DR Mandel, PW Jusczyk, DB Pisoni, Infants' recognition of the sound patterns of their own names. *Psychol. Sci.* **6**, 314–317 (1995).
6. PW Jusczyk, EA Hohne, Infants' memory for spoken words. *Science* **277**, 1984–1986 (1997).
7. MJ Carbajal, S Peperkamp, S Tsuji, A meta-analysis of infants' word-form recognition. *Infancy* **26**, 369–387 (2021).
8. R Shi, JF Werker, Six-month-old infants' preference for lexical words. *Psychol. Sci.* **12**, 70–75 (2001).
9. PW Jusczyk, RN Aslin, Infants' detection of the sound patterns of words in fluent speech. *Cogn. psychology* **29**, 1–23 (1995).
10. MC Frank, M Braginsky, D Yurovsky, VA Marchman, Wordbank: An open repository for developmental vocabulary data. *J. child language* **44**, 677–694 (2017).
11. AR Romberg, JR Saffran, Statistical learning and language acquisition. *Wiley Interdiscip. Rev. Cogn. Sci.* **1**, 906–914 (2010).
12. J Maye, JF Werker, L Gerken, Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition* **82**, B101–B111 (2002).
13. J Maye, DJ Weiss, RN Aslin, Statistical phonetic learning in infants: Facilitation and feature generalization. *Dev. science* **11**, 122–134 (2008).
14. S Tsuji, A Cristia, Perceptual attunement in vowels: A meta-analysis. *Dev. psychobiology* **56**, 179–191 (2014).
15. JR Saffran, RN Aslin, EL Newport, Statistical learning by 8-month-old infants. *Science* **274**, 1926–1928 (1996).
16. JR Saffran, NZ Kirkham, Infant statistical learning. *Annu. review psychology* **69**, 181 (2018).
17. EK Johnson, MD Tyler, Testing the limits of statistical learning for word segmentation. *Dev. science* **13**, 339–345 (2010).
18. J Lidz, A Gagliardi, How nature meets nurture: Universal grammar and statistical learning. *Annu. Rev. Linguist.* **1**, 333–353 (2015).
19. D Swingley, Contributions of infant word learning to language development. *Philos. Transactions Royal Soc. B: Biol. Sci.* **364**, 3617–3632 (2009).
20. E Dupoux, Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition* **173**, 43–59 (2018).
21. O Räsänen, Computational modeling of phonetic and lexical learning in early language acquisition: Existing models and future directions. *Speech Commun.* **54**, 975–997 (2012).
22. GK Vallabha, JL McClelland, F Pons, JF Werker, S Amano, Unsupervised learning of vowel categories from infant-directed speech. *Proc. Natl. Acad. Sci.* **104**, 13273–13278 (2007).
23. S Goldwater, TL Griffiths, M Johnson, A bayesian framework for word segmentation: Exploring the effects of context. *Cognition* **112**, 21–54 (2009).
24. T Schatz, NH Feldman, S Goldwater, XN Cao, E Dupoux, Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proc. Natl. Acad. Sci.* **118**, e2001844118 (2021).
25. K Hitczenko, NH Feldman, Naturalistic speech supports distributional learning across contexts. *Proc. Natl. Acad. Sci.* **119**, e2123230119 (2022).
26. TA Nguyen, et al., The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint arXiv:2011.11588* (2020).
27. Avd Oord, Y Li, O Vinyals, Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
28. M Lavechin, et al., Statistical learning models of early phonetic acquisition struggle with child-centered audio data. *PsyArXiv* (2022).
29. A Cristia, A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations (https://osf.io/c86ew/?view_only=f9af0cf7d2574234a8517c38151e4210) (2019).
30. J Kearns, Librivox: Free public domain audiobooks in *Reference Reviews*. (Emerald Group Publishing Limited), (2014).
31. A Brunault, C Pitton, Litterature audio (2007).
32. M Riviere, A Joulin, PE Mazaré, E Dupoux, Unsupervised pretraining transfers well across languages in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (IEEE), pp. 7414–7418 (2020).
33. T Schatz, et al., Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline in *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*. pp. 1–5 (2013).
34. R Ardila, et al., Common voice: A massively-multilingual speech corpus in *Language Resources and Evaluation Conference (LREC)*. (2020).
35. M de Seyssel, M Lavechin, Y Adi, E Dupoux, G Wisniewski, Probing phoneme, language and speaker information in unsupervised speech representations. *ArXiv* **abs/2203.16193** (2022).
36. D Povey, et al., The kaldi speech recognition toolkit in *Automatic Speech Recognition and Understanding (ASRU) workshop*. (IEEE Signal Processing Society), (2011).
37. G Le Godais, T Linzen, E Dupoux, Comparing character-level neural language models using a lexical decision task in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. pp. 125–130 (2017).
38. E Keuleers, M Brysbaert, Wuggy: A multilingual pseudoword generator. *Behav. research methods* **42**, 627–633 (2010).
39. RL Weide, The cmu pronouncing dictionary. *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict* (1998).
40. Avd Oord, et al., Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016).
41. PK Kuhl, et al., Phonetic learning as a pathway to language: new data and native language magnet theory expanded (nlm-e). *Philos. Transactions Royal Soc. B: Biol. Sci.* **363**, 979–1000 (2008).
42. E Bergelson, D Swingley, At 6–9 months, human infants know the meanings of many common nouns. *Proc. Natl. Acad. Sci.* **109**, 3253–3258 (2012).
43. M Sundara, L Polka, F Genesee, Language-experience facilitates discrimination of/d- /in monolingual and bilingual acquisition of english. *Cognition* **100**, 369–388 (2006).
44. CT Best, , et al., The emergence of native-language phonological influences in infants: A perceptual assimilation model. *The development speech perception: The transition from speech sounds to spoken words* **167**, 233–277 (1994).
45. B Ambridge, E Kidd, CF Rowland, AL Theakston, The ubiquity of frequency effects in first language acquisition. *J. child language* **42**, 239–273 (2015).
46. B Conboy, M Rivera-Gaxiola, L Klarman, E Aksoylu, PK Kuhl, Associations between native and nonnative speech sound discrimination and language development at the end of the first year in *Supplement to the proceedings of the 29th Boston University conference on language development*. (2005).
47. BT Conboy, JA Sommerville, PK Kuhl, Cognitive control factors in speech perception at 11 months. *Dev. psychology* **44**, 1505 (2008).
48. FM Tsao, HM Liu, PK Kuhl, Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child development* **75**, 1067–1084 (2004).
49. PK Kuhl, BT Conboy, D Padden, T Nelson, J Pruitt, Early speech perception and later language development: Implications for the" critical period". *Lang. learning development* **1**, 237–264 (2005).
50. TC Zhao, O Boorom, PK Kuhl, R Gordon, Infants' neural speech discrimination predicts individual differences in grammar ability at 6 years of age and their risk of developing speech-language disorders. *Dev. Cogn. Neurosci.* **48**, 100949 (2021).
51. G Alain, Y Bengio, Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644* (2016).
52. Y Belinkov, Probing classifiers: Promises, shortcomings, and advances. *Comput. Linguist.* **48**, 207–219 (2022).
53. R Shi, Perception of function words in preverbal infants in *10th International Congress for the Study of Child Language, Berlin, Germany*. (2005).
54. C Fisher, SL Klingler, Hj Song, What does syntax say about space? 2-year-olds use sentence structure to learn new prepositions. *Cognition* **101**, B19–B29 (2006).
55. S Bernal, J Lidz, S Millotte, A Christophe, Syntax constrains the acquisition of verb meaning. *Lang. learning development* **3**, 325–341 (2007).
56. A Christophe, S Millotte, S Bernal, J Lidz, Bootstrapping lexical and syntactic acquisition. *Lang. speech* **51**, 61–75 (2008).
57. S Peperkamp, R Le Calvez, JP Nadal, E Dupoux, The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition* **101**, B31–B41 (2006).
58. A Seidl, R Tincoff, C Baker, A Cristia, Why the body comes first: Effects of experimenter touch on infants' word finding. *Dev. science* **18**, 155–164 (2015).
59. B Jones, M Johnson, MC Frank, Learning words and their meanings from unsegmented child-directed speech in *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics*. pp. 501–509 (2010).
60. K Lakhotia, et al., On generative spoken language modeling from raw audio. *Transactions Assoc. for Comput. Linguist.* **9**, 1336–1354 (2021).
61. E Kharitonov, et al., Text-free prosody-aware generative spoken language modeling in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 8666–8681 (2022).
62. TA Nguyen, et al., Generative spoken dialogue language modeling. *arXiv preprint arXiv:2203.16502* (2022).
63. Z Borsos, et al., Audiolm: a language modeling approach to audio generation. *arXiv e-prints* pp. arXiv–2209 (2022).
64. R Abu-Zhaya, A Seidl, R Tincoff, A Cristia, Building a multimodal lexicon: Lessons from infants' learning of body part words in *Proc. GLU 2017 International Workshop on Grounding Language Understanding*. pp. 18–21 (2017).
65. C Yu, DH Ballard, RN Aslin, The role of embodied intention in early lexical acquisition. *Cogn. science* **29**, 961–1005 (2005).
66. K Nelson, *Young minds in social worlds: Experience, meaning, and memory*. (Harvard University Press), (2007).
67. NH Feldman, S Goldwater, E Dupoux, T Schatz, Do infants really learn phonetic categories? *Open Mind* **5**, 113–131 (2022).

Supplementary material

1. Proposed model

In this section, we described the proposed model which corresponds to the low-budget baseline architecture from the zero resource challenge 2021 (1).

A. Acoustic model.

A.1. Training objective. As originally proposed in (2), we used a contrastive loss which forces the latent space to retain information that is useful to predict future samples. Precisely, the input sequence of observations x_t is mapped to a sequence of latent representations through an encoder g_{enc}, such that $z_t = g_{enc}(x_t)$. Then, all $z_{\leq t}$ are aggregated with an auto-regressive model that produces a context-dependent latent representation $c_t = g_{ar}(z_{\leq t})$. Given the past context c_t, a predictor g_{pred} is asked to predict future representations z_{t+k} for $k \in \{1, ..., K\}$. Given a set $X = \{x_1, ..., x_n\}$ of N random samples containing one positive sample from the true positive distribution $p(x_{t+k} \mid c_t)$ and $N-1$ negative samples from the proposal negative distribution $p(x_{t+k})$, we optimize the categorical cross-entropy loss of classifying the positive sample correctly:

$$\mathcal{L} = -\frac{1}{K}\sum_{k=1}^{K} \log\left[\frac{\exp\left(g_{pred}(c_t)^T z_{t+k}\right)}{\sum_{x_j \in X}\exp\left(g_{pred}(c_t)^T z_j\right)}\right]$$

On top of the context-dependent representations c_t, we train a simple K-means algorithm to minimize the within-cluster sum of squares:

$$\mathcal{L} = \sum_{k=1}^{K}\sum_{c_{t,i} \in S_i} d(c_{t,i}, \mu_i)$$

where K is the number of clusters, S_i is the set of points belonging to the i^{th} cluster for $i \in [1..K]$, μ_i is the centroid of points in S_i, d is a distance function defined on the context-dependent representations c_t.

A.2. Implementation details. As proposed in (3), the encoder g_{enc} consists of a 5-layer convolutional neural network with kernel sizes $[10, 8, 4, 4, 4]$ and strides $[5, 4, 2, 2, 2]$ that returns a 256-dimensional vector every 10 milliseconds. The auto-regressive model g_{ar} is a 2-layer long short-term memory network of dimension 256. The model is asked to predict up to $K = 12$ time steps in the future (which is equivalent to 120 ms). The predictor g_{pred} is a single multi-head transformer layer with $K = 12$ heads, each predicting at time step $k \in \{1, ..., 12\}$. Negative samples are drawn from sequences that are temporally close to the sequence the positive sample are drawn from. More precisely, creating a batch consists of selecting 64 successive sequences in the case of the domain-general learner (or 64 successive sequences that have been pronounced by the same speaker for the domain-specific learner). For a current sequence seq_i, negative samples are taken from all other sequences seq_j, with $j \neq i$. All models have been trained on 8 GPUs with batches of 64 sequences, and each sequence has a duration of 1.28 seconds. All models are trained until convergence, and the best epoch is selected according to validation loss (5% of the original training set).

The K-means algorithm was trained with $K = 50$ using a euclidean distance function. All K-means were trained online with 200 sequences of 0.64 seconds using 1 GPU. All models are trained until convergence. At inference time, the input 10ms-frame is assigned the cluster label whose centroid is the closest.

B. Language model: LSTM.

B.1. Training objective. We train a language model on the discretized version of the audio files returned by the Acoustic Model. The Language Model is trained to predict the next unit of a sequence given its past context via a cross-entropy loss function:

$$\mathcal{L} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K} y_{t,k}\log(\hat{y}_{t,k})$$

where T is the length of the input sequence, K is the number of clusters, $y_{t,k}$ is the real cluster at time t, and $\hat{y}_{t,k}$ is the predicted probability at time-step t for cluster k.

B.2. Implementation details. The language model is a 3-layer LSTM with an embedding layer of size 200, hidden layers of size 1024 and a feed-forward output layer of size 200. We used the implementation proposed in (4).

2. Analysis: the training objectives computed on the evaluation set

A. Experimental protocol. We compute the training objectives of the Acoustic model and the Language model on the set of audio files used in the ABX discrimination task. Note that these audio files have never been seen during training.

B. Results. Figure S1 shows the 3 training objectives averaged across the native models (English evaluated on English, French evaluated on French) for the Acoustic Model and the Language Model. Results indicate that the higher the quantity of speech in the training set, the lower the test losses. This indicates a positive effect of exposure on the training objectives. This result is achieved via gradient descent.

3. Analysis: Detailed phonetic and lexical scores

Detailed phonetic and lexical scores are presented in Figure S2. Here, the scores are presented separately for each of the English and French test set.

When tested on English, both phonetic and lexical results follow the trends discussed in the main paper. When tested on French, however, the phonetic scores yielded by the English (non-native) models are closer to the scores yielded by the French (native) models. For the lexical task, the French (native) models do yield higher scores than the English (non-native) models, but the difference between the two curves is lesser than the one observed on the English test set.

Such results could indicate a potential asymmetry between languages. Although it is difficult to provide precise evidence, the fact that the English model tested on the French lexical task (bottom right graph) gets progressively above chance with input quantity might be explained by the high number of cognates and loanwords in English.

Yet, one should stay cautious about such comparisons. As mentioned in the Results Section, such patterns can also be the effect of the training set themselves. For example, (5) found that the presence of non-speech in such models can deteriorate their quality, and it is possible that such differences exist between the French and English training sets (with one being noisier than the other). This is why we recommend focusing on results aggregated symmetrically over the native and non-native conditions, as presented in the results. Still, further work could focus on potential asymmetries between languages.

4. Analysis: Phonetic scores predict lexical scores

A. Experimental protocol. For this analysis, we consider models trained on 50h, 100h, or 200h of English or French speech. We evaluate their phonetic score using the ABX discrimination task, and their lexical score using the spot-the-work task, both described in the Methods Section. Both scores are evaluated in the native condition, i.e. on the training language. Lexical scores are computed either on: 1) words belonging to the 64th frequency band (high frequency words); 2) words belonging to the 1st frequency band (low frequency words) or 3) as the average accuracy across all frequency bands.

B. Results. Figure S3 shows the correlation between the phonetic score and the lexical score obtained by individual models for different training set sizes (column-wise) and for a lexical score computed on different frequency bands (row-wise). Results indicate that, in general, models that are less accurate at discriminating native sounds, are less good in the spot-the-word task. This effect seems more important on high frequency words as shown by the 50h English model that exhibits a Pearson's R correlation coefficient of .52 ($p<.0001$) on high frequency words, .36 ($p<.05$) across frequency bands, and .12 (non-significant) on low frequency words. While the 100-hours and the 200-hours English models seem to exhibit a similar pattern, models trained on French speech show more constant correlation scores across the different frequency bands.

5. The emergence of latent linguistic structure

A. Layer-wise LOO classification scores for the lexical analyses. Leave-One-Out classification scores for the function vs. content (FC) and part-of-speech (POS) categories were computed on a development set for all hidden layers of the 3200h English and French models (see Methods). Results are presented in Table 2. Layer 3 yielded the best scores overall for both the English and French models, and was subsequently chosen to carry out the lexical probing analyses.

B. Results on the French models. In the Results Section, we presented qualitative and quantitative analyses of the emergence of phonetic and lexical categories in the English models. The same analyses on the French models are presented in Figure S4. Experimental methods are the same as described in the Methods).

As for the English model, qualitative analyses carried out on the French 3200h model suggest that this model clearly encodes information about sonority, place and voicing, with the categories being visually well separated (panel a). Moreover, all of these three types of information get progressively better encoded with more training data (panel b). Interestingly, contrary to results on the English models, even the voicing information present in the models gets significantly better.

Regarding the emergence of the lexical and proto-syntactic categories, the patterns are the same as for the English models. No clear categories of function vs content and Part of Speech (POS) can be qualitatively distinguished from the t-SNE(s) on the 3200h French model (panel c). Yet, probing analyses carried out on all models show that this categorisation can be better learnt with models trained on more data, suggesting that this information gets gradually better encoded (panel d). The main simplifying assumption regarding the word segmentation problem in this work is that utterances are represented as strings of phonemes. Any computational model comes with its set of simplifying assumptions, which is fine. However, the authors should discuss this in more detail. In particular, the assumption mentioned above is problematic for two reasons. First, this assumes that children can assign a single phoneme to each phone they hear in an error-free manner. However, evidence suggests that children segment some words way before their perception have reached that of an adult

6. Analysis: the frequency effect

We evaluate the Language model using the spot-the-word task described in the Methods. The lexical score obtained by the native model is displayed in the diagonal of Figure S5 (panels (a) and (d)). The anti-diagonal shows the lexical score obtained by the non-native model (panels (c) and (d)). The number of trials per frequency band is presented in Table 3. In the native condition, results indicate that the higher the quantity of speech, the higher the lexical score, showing a positive effect of exposure. We only observe a slight increase in the non-native condition, which suggests that the non-native model is mostly unable to solve the lexical task. The positive effect of exposure in the native condition seems more important on high-frequency words than low-frequency words (native curves are steeper as the frequency increases).

7. Analysis: the emergence of lexical factors

A. Dataset. We use the Massive Auditory Lexical Decision (MALD) dataset (6) that contains reaction times of human participants on the auditory lexical decision task. In this psycholinguistic task the participant hears an audio stimuli and has to classify it as either a word or a nonword. The MALD contains reaction times for 26,793 words and 9592 nonwords. This sums up in reaction times for 227,179 auditory lexical decisions from 231 unique monolingual English listeners. In addition to reaction times, each stimuli is annotated for various lexical descriptors: the duration of the stimuli, the frequency of the stimuli, the number of phonological neighbors, the phone index of the phonological uniqueness point of the stimuli within the CMU-A dictionary (7), the mean phone-level Levenshtein distance of the item from all entriers within the CMU-A, etc. A detailed description of all descriptors can be found in (6). All data on nonwords were discarded and only words were included in the present analysis.

B. Experimental protocol. We compute the probability of each word of the MALD dataset with the Language Model, and look at which lexical factors are significant predictors of this probability. We do so using a nested linear regression analysis.

We first start with the predictor that leads to the highest R^2. Then, we add the second predictor that increases the R^2 in the most significant way (i.e., the selection criterion is the p-value such as computed by a likelihood-ratio test). We do until the addition of predictor does not yield a significant increase in R^2.

We run the same analysis with human reaction times and then compare the lexical factors for both target: pseudo-probabilities returned by the Language Model, and human reaction times.

C. Results. Panel (a) of Figure S6 shows the various descriptors (duration, frequency, phonological features, part-of-speech categories, etc.) that are: 1) significant predictors of human reaction times (in green); 2) significant predictors of the Language Model probability (in red); 3) both 1) and 2) (intersection of green and red surfaces); or 4) not significant for both human reactions times and pseudo-probabilities (in white).

Results indicate that the duration and the frequency of the word are significant predictors of both the human reaction time and the Language Model probability. PhonND which indicates the number of phonological neighbors (defined as one phone edit away) for the word within the CMU-A dictionary, and PhonUP which indicates the phone index of the phonological uniqueness point of the item within the CMU-A are also significant predictors of both the human reaction time and the Language Model probability. Significant predictors of the Language Model probability capture 31% of its variance, while significant predictors of the human reaction time capture 26% of its variance.

Panel (b) of Figure S6 shows the R^2 obtained by the nested linear regression models as a function of quantity of speech in the training set. The blue curve corresponds to a linear model containing only Duration as a predictor, the orange curve both Duration and PhonLev (the mean phone-level Levenshtein distance to all entries within the CMU-A dictionary), etc. Results indicate that the higher the quantity of speech in the training set, the higher the R^2 obtained by the different nested models. In other words, as the Language Model receives more speech, the abovementioned linguistic factors become more predictive of the probability.

8. Analysis: the emergence of word boundaries

In this analysis, we look into whether the emerging grammatical structure learned by our model is grounded on some notion of words or morphemes as a cohesive sequence of phonemes. In a seminal paper, Elman (8) presented a language model trained on letters and discovered that the probability assigned at each time step gradually increases inside words and sharply decreases between words. This important result suggests that the language model trained on letters implicitly performs word segmentation, although the model is not provided with breaks. In this section, we perform a similar analysis, with, contrary to Elman, our language model that is trained from the raw acoustic input.

A. Experimental Protocol. The analysis shows the probability assigned by the language model as the sentence unfolds over time. We consider either words or sentences from the Common Voice audio files that have also been used in the ABX discrimination task, and that have never been seen during training.

B. Results. Figure S7 present behaviors of the Language Model probability as the audio unfolds over time. The model considered was trained on 3200 hours of English speech.

Panel (a) shows probability curves as a function of length rank (1st rank contains shortest words, 10th rank contains longest words). Probability curves are linearly interpolated so that each word belonging to the same length rank share the same target length (median length for this rank). Results show a length effect, with the probability increasing as the word unfolds over time. Sharp decreases in the beginning and the end of words can be noticed which seems to indicate that the Language Model has a harder time predicting the next token on word boundaries.

Panel (b) shows a similar analysis on sentences. Sentences are sorted depending on their number of words, linearly interpolated so that sentences with the same number of words share the same target duration (median duration)), and averaged. Results show a sharp increase in the probability at the beginning of sentences, then a slight decrease as the sentence unfolds over time. A sharp increase can be noticed at the end of sentences, which indicates that the Language Model is better at predicting the next token at the end of sentences than at the beginning/middle.

Panel (c) shows the probability for the sentence "*I can see a smiling face in the clouds*" (in grey) and the probability estimated by averaging N=500 words of the same size (in red). Results indicate a noisy behavior when considering a single stimuli, despite having applying a moving average of 10 frames (100 ms). However, when considering the average profile of the probability, we notice that the probability slightly increases inside words, and sharply decreases between words.

We draw on the same conclusion than Elman: the language model seems sensitive to word boundaries, but only when averaging across hundreds of inputs. The acoustic variability infants are facing bring a much more difficult problem: normalizing the input across the various acoustic dimensions (speaker, speech rate, etc.)

9. Analysis: the learned units

A. Experimental protocol. Here, we compare the discrete units learned by the K-means algorithm to phones as recognised by an Automatic Speech Recognition (ASR) algorithm. To compute the ASR phones, we use the MLS speech corpus (9), which is an aggregation of read speech taken from the LibriVox project (10). For each language, we select 100h of speech data. We first train a phone bigram language model on each training set using the SRILM toolkit (11). We then train for each language a hybrid GMM-DNN phone recogniser based on a time-delay neural network architecture (12), adapting the s5 librispeech recipe from the Kaldi speech recognition toolkit (13). Finally, we infer ASR phones for our English and French Common Voice test sets using the English and French newly trained phone recognizers respectively*.

A.1. Analyses. We can now compare how K-means units and ASR phones compare to the gold phones from the test set. For each model, we compute $p2u$ (phone-to-unit), the perplexity of gold phones given the ASR phones or the K-means units distribution, and $u2p$ (unit-to-phone), the perplexity of ASR

*The phone accuracy yielded by the phone recognisers on the English and French test sets of 24.7 and 24.6% respectively.

Supplementary references

1. TA Nguyen, et al., The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. *arXiv preprint arXiv:2011.11588* (2020).
2. Avd Oord, Y Li, O Vinyals, Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748* (2018).
3. E Kharitonov, et al., Data augmenting contrastive learning of speech representations in the time domain in *Spoken Language Technology Workshop (SLT)*. (2021).
4. M Ott, et al., fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038* (2019).
5. M Lavechin, et al., Statistical learning models of early phonetic acquisition struggle with child-centered audio data. *PsyArXiv* (2022).
6. BV Tucker, et al., The massive auditory lexical decision (mald) database. *Behav. research methods* **51**, 1187–1204 (2019).
7. RL Weide, The cmu pronouncing dictionary. *URL: http://www. speech. cs. cmu. edu/cgibin/cmudict* (1998).
8. JL Elman, Finding structure in time. *Cogn. science* **14**, 179–211 (1990).
9. V Pratap, Q Xu, A Sriram, G Synnaeve, R Collobert, Mls: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411* (2020).
10. J Kearns, Librivox: Free public domain audiobooks in *Reference Reviews*. (Emerald Group Publishing Limited), (2014).
11. A Stolcke, Srilm-an extensible language modeling toolkit in *Seventh international conference on spoken language processing*. (2002).
12. V Peddinti, D Povey, S Khudanpur, A time delay neural network architecture for efficient modeling of long temporal contexts in *Sixteenth annual conference of the international speech communication association*. (2015).
13. D Povey, et al., The kaldi speech recognition toolkit in *Automatic Speech Recognition and Understanding (ASRU) workshop*. (IEEE Signal Processing Society), (2011).
14. T Schatz, NH Feldman, S Goldwater, XN Cao, E Dupoux, Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proc. Natl. Acad. Sci.* **118**, e2001844118 (2021).

Table 1. Evaluated phonetic inventory in Metropolitan French and American English in International Phonetic Alphabet (IPA) standard.

Manner of articulation	Metropolitan French	American English
	Consonants	
Stops:	p,b,t,d,k,g	p,b,t,d,k,g
Nasals:	m,n,ɲ	m,n,ŋ
Fricatives:	f,v,s,z,ʃ,ʒ,ʁ	f,v,θ,ð,s,z,ʃ,ʒ,h
Approximants:	j,w,l	j,ɹ,w,l
Affricates:	ʤ	ʤ,ʧ
	Vowels	
Oral	i,y,e,ø,œ,ɛ,a,ə,ɔ,o,u	i,ɪ,ɛ,æ,ɚ,ʌ,ɐ,u,ʊ,ɔ,ɑ
Nasal:	ɑ̃, ɛ̃, œ̃,ɔ̃	
Diphtongs:		aɪ,ɔɪ,aʊ,eɪ,oʊ

Table 2. Leave-One-Out Classification Scores (CS). Scores are computed on the English and French 3200h models using the dev sets for the function vs content (FC) and part-of-speech (POS) categories classification tasks. Best average classification scores are indicated in bold.

Language	Hidden layer	FC CS	POS CS	Average CS
English	1	57.07	46.15	51.61
English	2	58.26	50.57	54.41
English	3	60.42	55.89	**58.16**
French	1	61.37	39.21	50.29
French	2	62.91	47.41	55.16
French	3	66.34	45.43	**55.89**

Table 3. Number of trials in the spot-the-word task. The numbers have to be divided by 4 (number of synthesised voices) to get the number of word/nonword pairs.

	Frequency band						
test language	**1st (rare)**	**2nd**	**4th**	**8th**	**16th**	**32th**	**64th (frequent)**
English	70,136	60,664	49,324	40,204	28,132	17,544	15,108
French	51,956	53,700	42,944	32,032	23,168	16,336	12,976

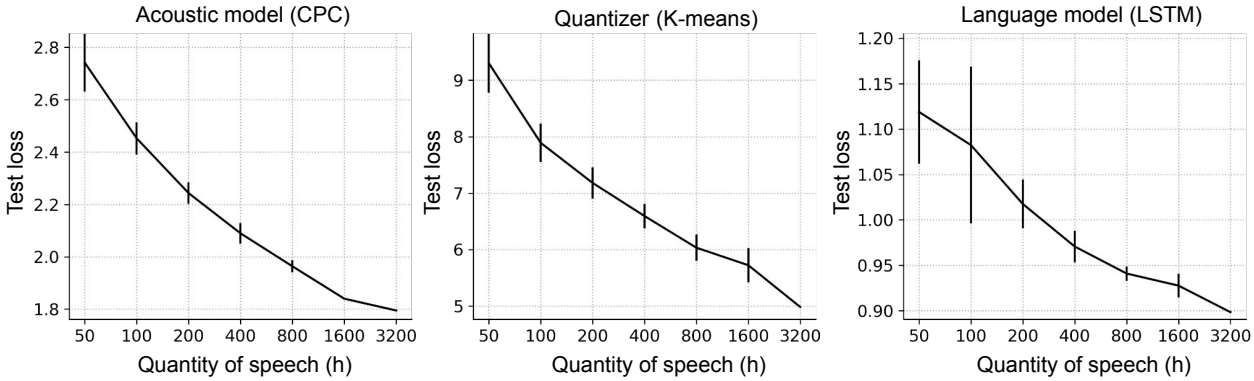


Fig. S1. Graduality and parallelism of the training objectives. The three losses, computed on the test set, for the 2 components of our model: the acoustic model minimizes the cross-entropy of classifying the positive sample correctly (contrastive predictive coding); and the within-cluster sum of squares (K-means); the language model minimizes the cross-entropy of predicting the next token correctly.

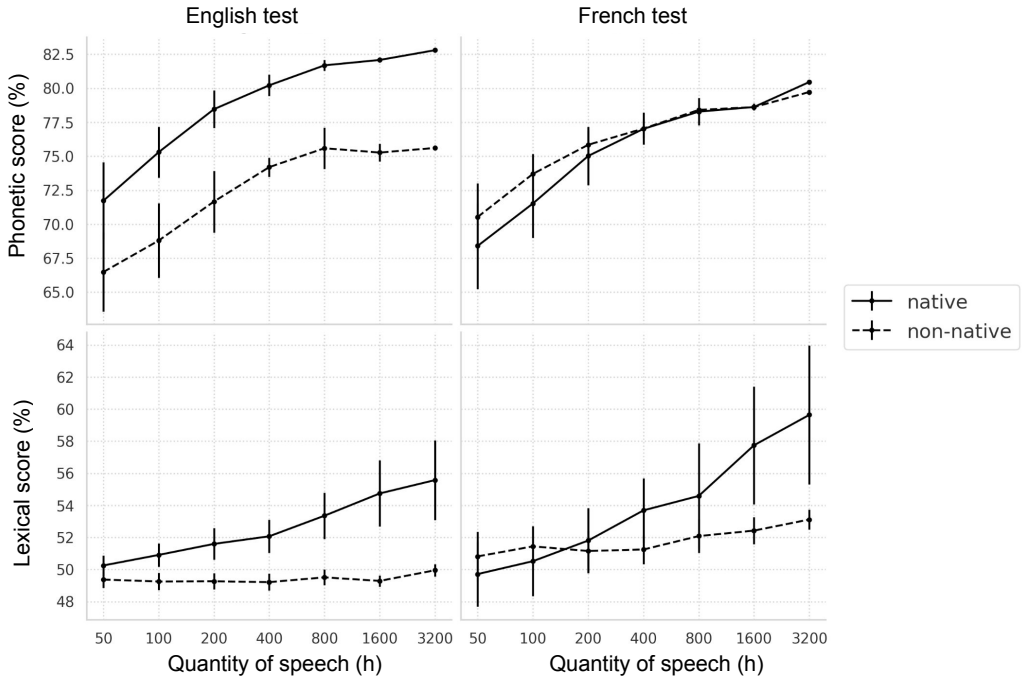


Fig. S2. Phonetic and lexical scores per training and test languages. Phonetic and lexical scores are presented on both English and French test sets, separately for each trained language. Phonetic scores are presented on the top row and lexical scores on the bottom row. On the left column, we show scores calculated on the English test set, and on the right column, scores calculated on the French test set. For the lexical scores, scores are first averaged over each frequency band then per training size. Error bars for the phonetic and lexical scores correspond to the standard deviation between the averaged scores for all models of a same training size and language.

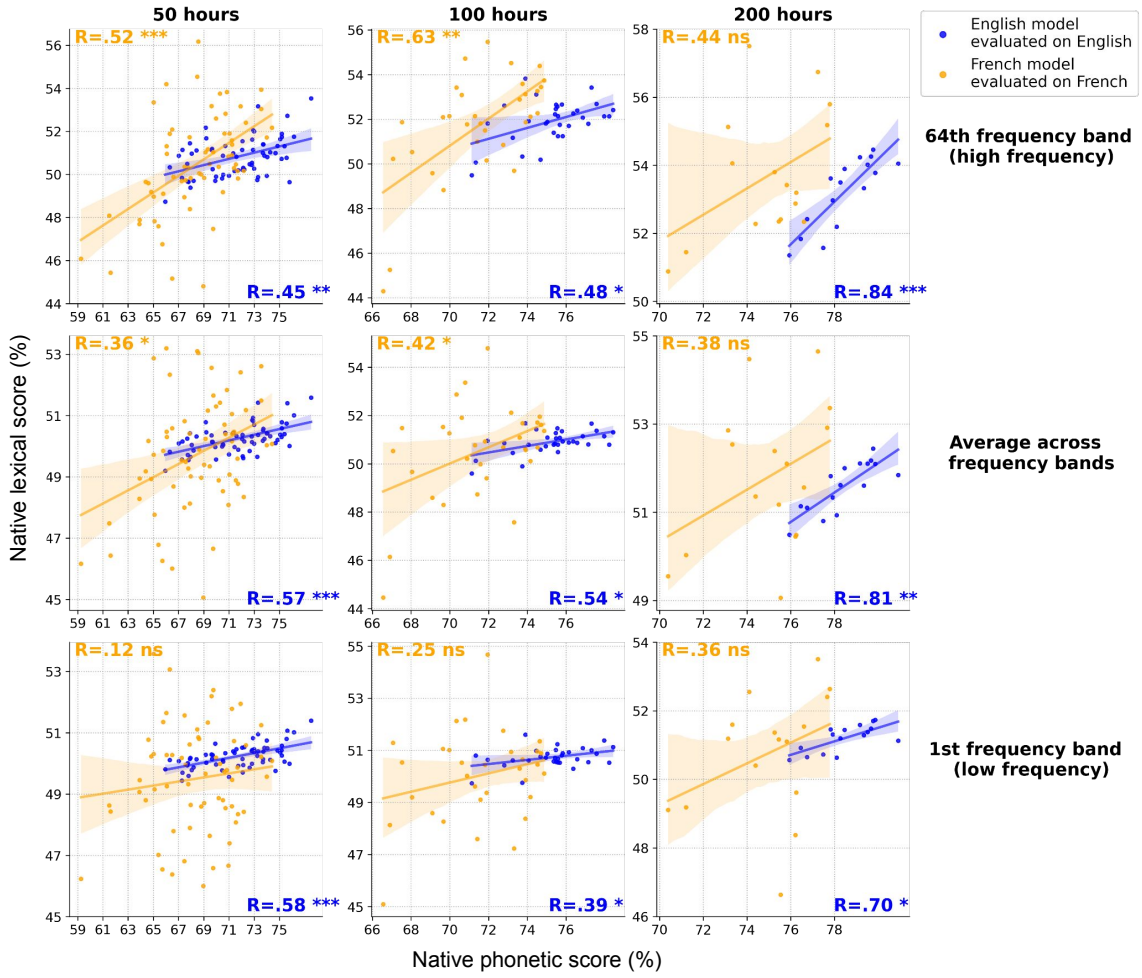


Fig. S3. Phonetic scores predict lexical scores. Correlation between the phonetic and lexical scores across English (in blue) and French (in orange) models trained on 50h (first column), 100h (second column) and 200h (third column) of speech. The lexical score is evaluated on the high frequency words (first row), the average across frequency bands (second row), or low frequency words (third row). R is the Pearson correlation coefficient. Significance levels: na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

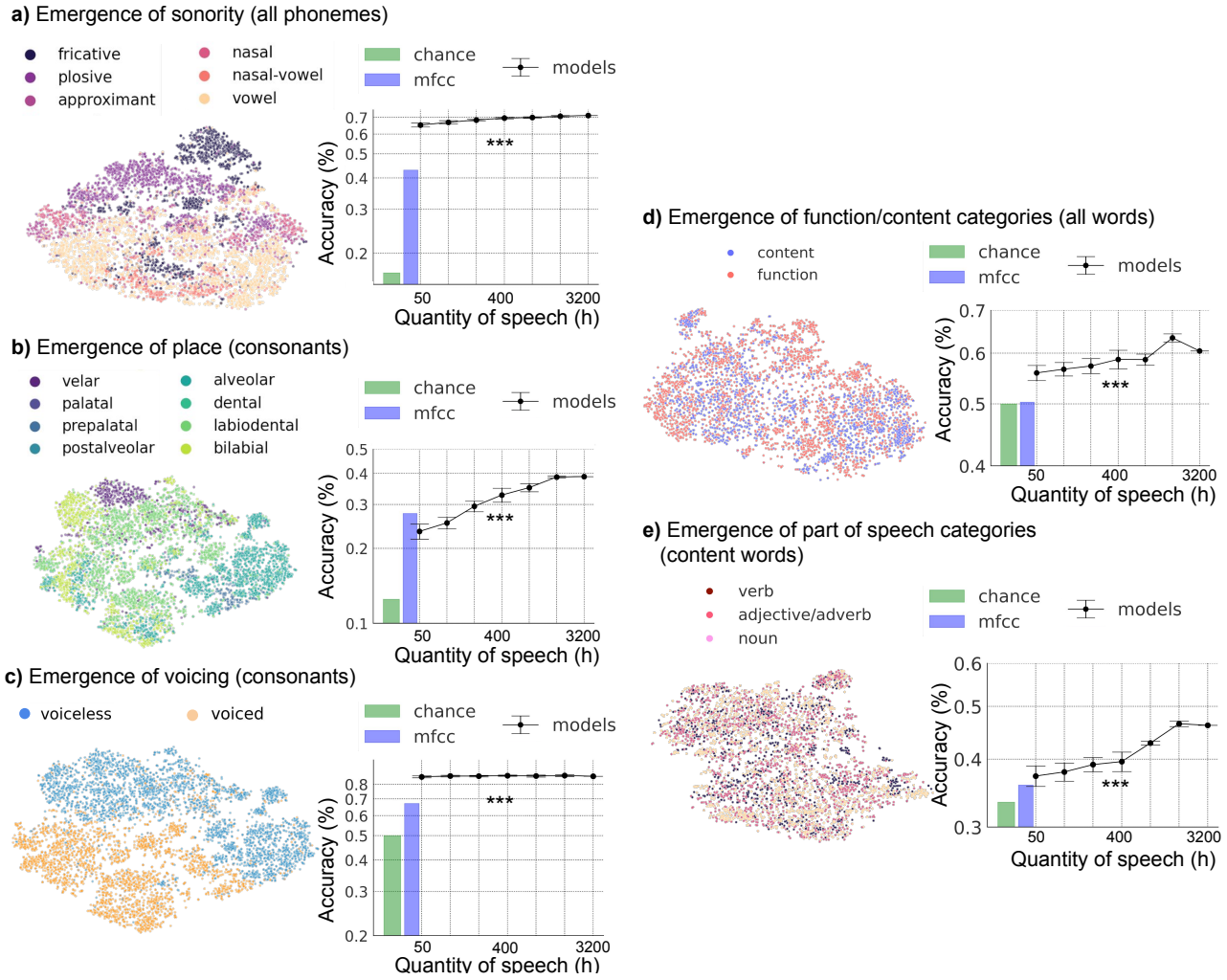


Fig. S4. Emergence of latent linguistic structures at the phonetic level for the French models. Left: tSNEs of the continuous representations of the acoustic model (last layer) pooled within phonetic tokens in a test set, according to sonority (a), place (b) and voicing (c) for the 3200h English model. with their corresponding developmental curves of leave-one-phoneme-type-out classification errors as a function of input quantity (taking all 256 dimensions into account). Chance level and MFCCs performances are also given. Right: tSNEs of the continuous representations of the language model (last layer) pooled over words tokens according to (d) function/content distinction and (e) part of speech for the 3200h English model with their corresponding developmental curves of leave-one-word-out classification error as a function of input quantity (taking all 1024 dimensions into account). Chance level and MFCCs performances are also given. The asterisks indicate a significant correlation of classification error and input quantity. na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

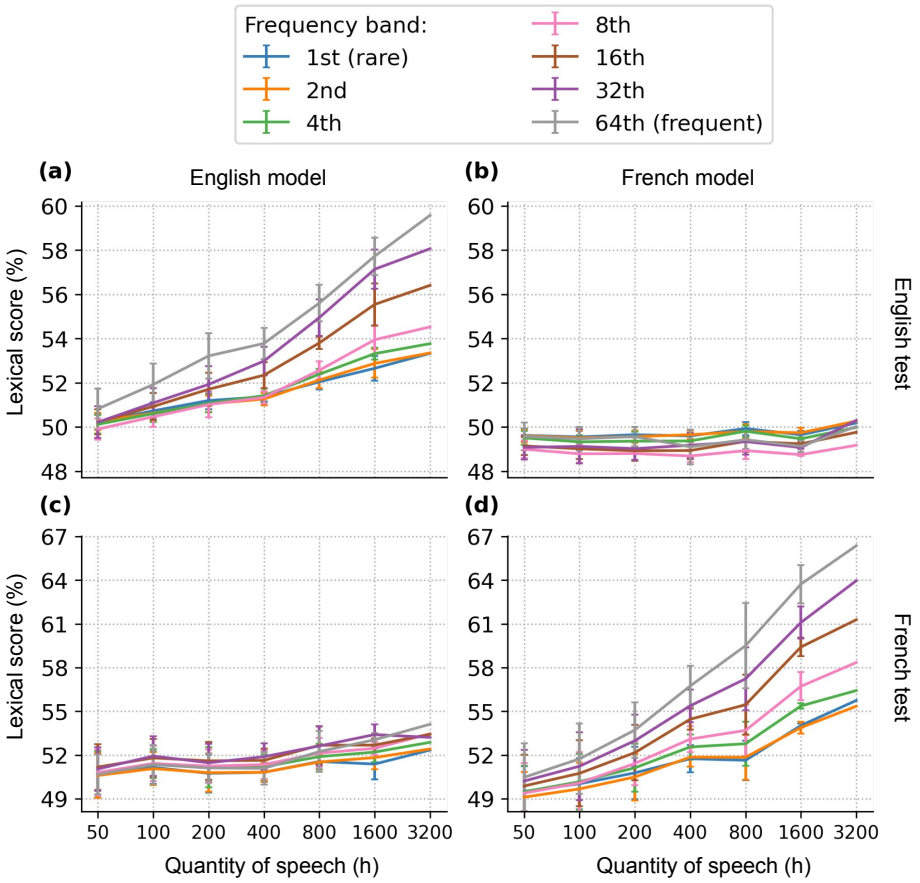


Fig. S5. The frequency effect on the lexical task. Effects of input frequency and input quantity on phonetic and lexical tasks. Left: Right: Lexical scores obtained by English (first column) and French (second column) models on the English (first row) and the French (second row) lexical test. Panels (a) and (d) on the diagonal show lexical scores obtained by native models. Panels (b) and (c) on the anti-diagonal show lexical scores obtained by non-native models. Scores are given as functions quantity of speech available in the training set, and class of frequency of evaluated words. Words in the 64th class of frequency are present at least one time in the 50-hours training sets, two times in the 100 hours, four times in the 200 hours. Words belonging to the 32th class of frequency are present at least one time in the 100 hours training sets, 2 times in the 200 hours, etc. Error bars represent standard errors computed across mutually exclusive training sets whose number depends on the quantity of data available. The last data point along the x-axis is computed on a single learner (trained with all available data), then the number of learners doubles with each step along the x-axis, as the quantity of audio is divided by two.

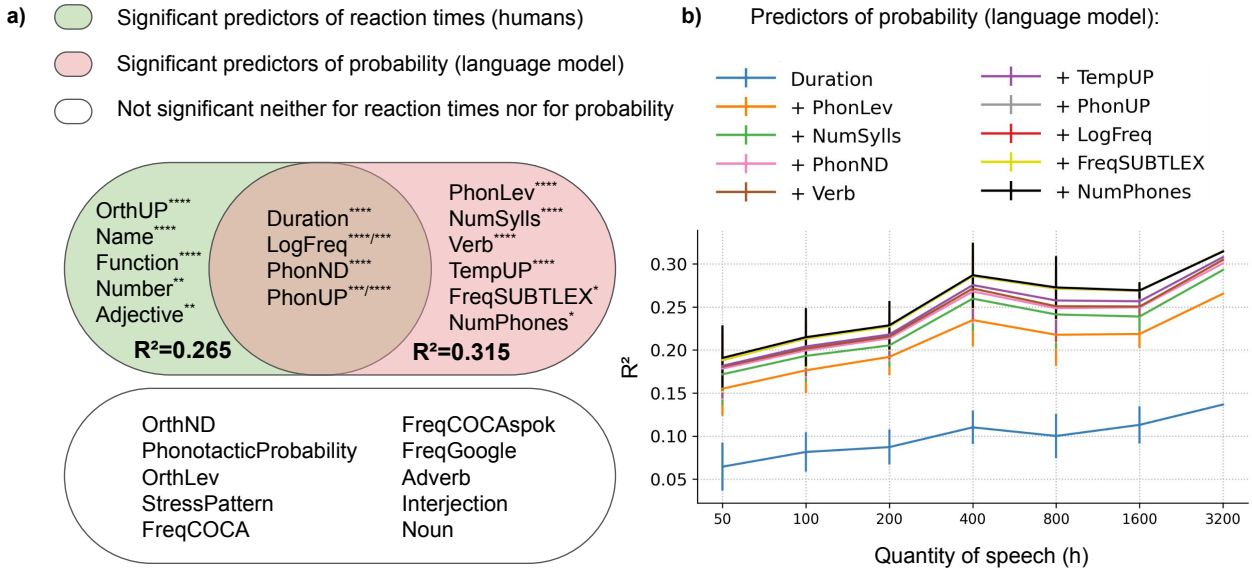


Fig. S6. Emergence of lexical factors in the English Language Model. (a) Nested multiple regression results on models trained with 3200 of English (red) or human reaction times (green) from the MALD dataset with various descriptors (6). Using a linear regression model, we start from the descriptor that leads to the largest R^2 and keep adding descriptors until the addition does not yield a significant increase in R². RT corresponds to log(RT) and pseudo-prob to mean(log(prob)). (b) Developmental curve of the R² of the nested models as a function of input quantity. Significance levels: * $p < 5 \times 10^{-2}$, ** $p < 10^{-3}$, *** $p < 10^{-4}$, **** $p < 10^{-5}$

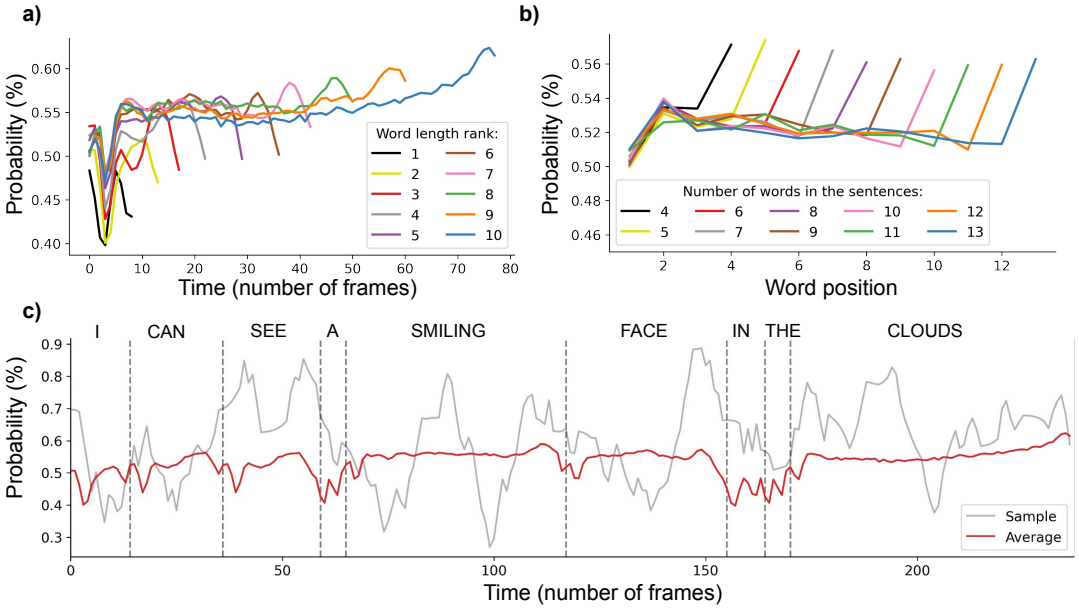


Fig. S7. Emergence of word boundaries in the Language Model. Probability, such as computed by the Language Model, as a function of time (10ms frames) in a sample sentence for a model trained on 3200h. In red, the average pseudo probability estimated from words of the same size (N=500). The pseudo-probabilities are smoothed by a moving average of 100ms (10 frames).

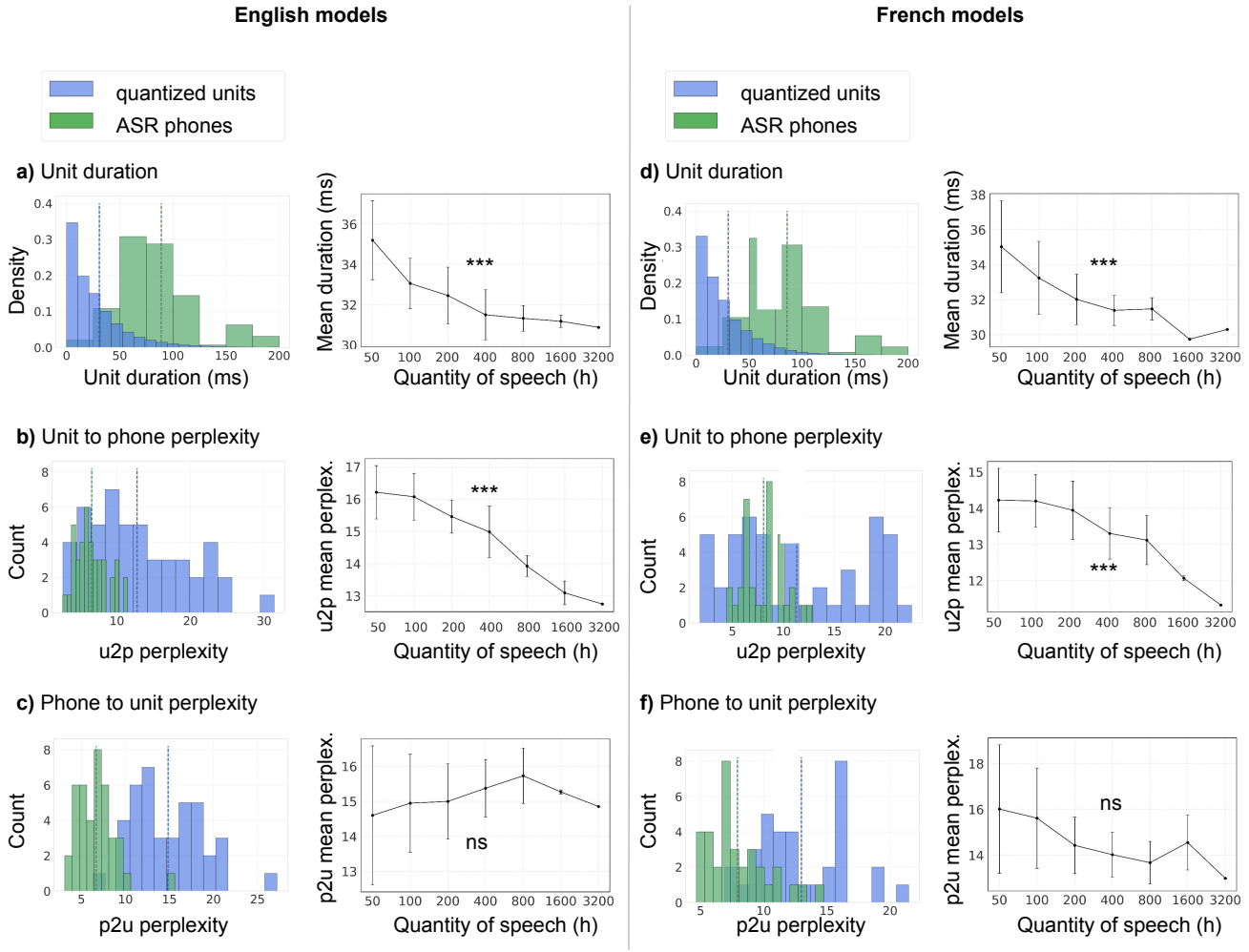


Fig. S8. Analysis of discovered pseudo-phonetic units.. (a),(b),(c) *left*: duration and perplexity histograms of the discrete units relative to phones and vice versa compared to an ASR baseline for the 3200h English model. (d),(e),(f) left: the same for French. (a), (b), (c) *right*: corresponding developmental curves of the median duration and perplexities for the English models. (d),(e),(f) right: the same for French. na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001

3.3 Beyond phonetic and lexical acquisition: evaluating prosodic knowledge in models

Phonetic and lexical learning are amongst the most impressive improvements made by infants in early language acquisition. However, another linguistic level at which they also make tremendous improvements is prosody: the rhythmic and melodic aspects of speech. Moreover, as we have seen in the Introduction (§1.3), knowledge of the prosody of their language can help bootstrap the learning of other linguistic information, such as at the lexical and syntactic levels.

In this section, we propose a new zero-shot metric, *ProsAudit*, for evaluating structural prosodic information captured by SSL speech models, that is, information about the prosodic constituents (Chapter 1,§1.3.3). We then use this metric as an outcome measure in the developmental framework presented in the previous chapter in order to simulate prosodic learning.

This research results from a collaboration with two interns who worked on the topic when in the team, Andrea Santos Revilla and Arthur Thomas. It is presented in the form of a paper:

de Seyssel, M., Lavechin, M., Titeux, H., Thomas, A., Virlet, G., Santos Revilla, A., Wisniewski, G., Ludusan, B. & Dupoux, E. (2023) ProsAudit, a prosodic benchmark for self-supervised speech models. In *Proc. Interspeech 2023.*

3.3.1 ProsAudit, a prosodic benchmark for self-supervised speech models

Paper summary

In the paper, we propose a new metric, evaluation set and benchmark, ProsAudit, which evaluates the presence of English *structural prosody* in unsupervised speech representation models. By structural prosody, we mean prosodic cues that help organise speech by indicating the boundaries between words, phrases, and sentences. As for the metrics introduced in the Zero-Resource Challenges (Nguyen*, de Seyssel* et al., 2020 in particular), this ProsAudit metric is zero-shot and inspired by psycholinguistics experiments as the one carried out by Ludusan et al. (2021). We also integrate it into the Speech Modelling track of the Zero-Resource 2021 Challenge (Dunbar et al., 2021).

In a nutshell, the task operates on the same general principles as the sWuggy task. It consists in presenting thousands of pairs of English utterances, each pair comprising one utterance with natural prosody and one with unnatural prosody. We then evaluate the model's probability of generating each utterance. If the model's probability for the natural utterance is higher than for the unnatural one, the pair receives a score of 1; otherwise, it receives a score of 0. By calculating the average score for all pairs, we obtain an accuracy metric. Unlike sWuggy, where the speech was synthesised, the utterances used in ProsAudit are taken

from natural speech, in which we inserted artificial silences at either natural or unnatural locations.

The ProsAudit benchmark consists of two subtasks: a *lexical* subtask and a *protosyntax* subtask. In the lexical subtask, the natural boundary is placed between words, while the unnatural boundary is placed within words. In the protosyntax subtask, the natural boundary is placed at a strong prosodic break, and the unnatural boundary is placed at a weak prosodic break, always between words. While only prosodic knowledge is required for the protosyntax subtask, lexical knowledge is necessary for the lexical subtask, as the "natural" boundaries are placed in locations where prosodic boundaries would not normally be present, such as within clitic groups.

We first evaluated a series of SSL speech models on the ProsAudit benchmark, including the STELA models presented in Section 3.2. We found that all evaluated models performed above chance on both subtasks. Additionally, we conducted a human evaluation on a subset of the two subtasks and found that humans also performed above chance in both subtasks.

In the second phase of our study, we investigated the impact of size and nativeness on prosody by conducting the ProsAudit evaluation on both the English and French STELA models and generating "developmental curves", as we did for the phonetic ABX and sWuggy metrics in Section 3.2. As for the phonetic and lexical evaluations, our analysis revealed that size significantly affected the native (English) models, with increasing scores on both the protosyntax and lexical ProsAudit subtasks with larger training sizes. Interestingly, even the 50-hour models exhibited above-chance effects, indicating the presence of prosodic-related linguistic information even in the small models. Additionally, we observed a nativeness effect on both subtasks, with the non-native (French) models consistently yielding lower scores than the native (English) ones. Nonetheless, the non-native models still produced above-chance results, indicating a cross-linguistic influence of prosody, as had already been reported in psycholinguistics (Endress and Hauser, 2010). Finally, while the non-native models' protosyntax subtask performance improved with size, the scores on the lexical subtask remained relatively constant, consistent with our findings on the sWuggy lexical task in Section 3.2 (more training data does not lead to better lexical scores for the non-native models). This reinforces the notion that the lexical ProsAudit subtask is inherently linked to lexical knowledge.

ProsAudit, a prosodic benchmark for self-supervised speech models

Maureen de Seyssel[1,2], *Marvin Lavechin*[1,5], *Hadrien Titeux*[1], *Arthur Thomas*[†], *Gwendal Virlet*[4†], *Andrea Santos Revilla*[†], *Guillaume Wisniewski*[2], *Bogdan Ludusan*[3], *Emmanuel Dupoux*[1,5]

[1] Cognitive Machine Learning, ENS–CNRS–EHESS–INRIA–PSL Research University, France
[2] Université Paris Cité, CNRS, Laboratoire de linguistique formelle, Paris, France
[3] Faculty of Linguistics and Literary Studies & CITEC, Bielefeld University, Germany
[4] PEGASE, INRAE, Institut Agro, Saint-Gilles, France [5] Meta AI Research, France

maureen.deseyssel@gmail.com

Abstract

We present ProsAudit, a benchmark in English to assess structural prosodic knowledge in self-supervised learning (SSL) speech models. It consists of two subtasks, their corresponding metrics, and an evaluation dataset. In the protosyntax task, the model must correctly identify strong versus weak prosodic boundaries. In the lexical task, the model needs to correctly distinguish between pauses inserted between words and within words. We also provide human evaluation scores on this benchmark. We evaluated a series of SSL models and found that they were all able to perform above chance on both tasks, even when evaluated on an unseen language. However, non-native models performed significantly worse than native ones on the lexical task, highlighting the importance of lexical knowledge in this task. We also found a clear effect of size with models trained on more data performing better in the two subtasks.

Index Terms: prosody, speech representation, self-supervised learning, human evaluation

1. Introduction

In recent years, self-supervised learning (SSL) speech models such as Wav2Vec [1], CPC [2, 3], HuBert [4] have made groundbreaking advancements, while removing the need for labeled data as they use information extracted from the input raw audio itself. Multiple benchmarks and metrics have been since developed to test (and exhibit) the linguistic knowledge of such models at different levels. For instance, the Zero-Resource speech challenge [5, 6, 7] offers zero-shot evaluation metrics at the phonetic, lexical, syntactic and semantic levels and SUPERB [8] allows downstream evaluation on speech processing tasks, including paralinguistics and speaker-related tasks.

An important aspect of language that has received little to no attention in SSL models is prosody. As a result, we still know little about SSL encoding of structural prosodic knowledge. Yet, through its key components: rhythm, stress, and intonation, prosody plays a significant role in language processing [9, 10], interfacing with other linguistic levels (e.g., lexical, syntactic levels), while also carrying paralinguistic information (e.g., emotion) (for a literature review, see [9] and [10]). Recently, [11] proposed to explicitly implement prosodic knowledge into SSL models by forcing prediction of pitch and duration within the models, but only human evaluation on downstream tasks was used as an indirect cure of prosodic encoding in the models. Besides, [12] proposed a benchmark to test pragmatics aspects of prosody of the models using downstream tasks. Yet, there is currently no zero-shot metrics which allows systematic evaluation of prosody in such models.

In this paper, we fill this gap by proposing an evaluation benchmark, ProsAudit, which assesses SSL speech models' ability to learn prosodic information at the structural level. By such, we refer to how prosody contributes to the organisation of speech by marking the boundaries between words, phrases, and sentences. Our benchmark, in English, consists of two subtasks. The protosyntax task tests the model's ability to identify strong versus weak prosodic boundaries (e.g., see [13] for an evaluation of human performance). The lexical task tests the model's ability to distinguish between pauses inserted between and within words. Being a proxy for detecting word boundary versus word internal, this task requires some lexical knowledge. Crucially, we also provide results on these two tasks carried out on human evaluation. We bring our benchmark to the Spoken Modelling track of the Zero-Resource Speech Challenge [7] and present some baseline results, which will be integrated into the leaderboard. Finally, we conduct further analysis on factors like input quantity and nativeness, putting the models in perspective with prosodic learning in humans.

2. Methods

2.1. ProsAudit benchmark

We created prosodic benchmarks in English, composed of two tasks which each focus on different aspects of structural prosody: the lexical task and the protosyntax task. These tasks are designed to evaluate the models' understanding of prosody by presenting them with pairs of stimuli that only differ in the placement of a pause.

2.1.1. Material and data preprocessing

We used the Boston University Radio News Corpus (BU) [14] dataset to create the evaluation set, as it includes word and phone-level transcriptions along with prosodic hierarchy annotations based on the American English ToBI system [15]. The dataset is a collection of professionally-read news stories.

We selected segments based on the criteria outlined in [13]. To qualify, a segment had to start and end with an intonation phrase (IP) boundary (ToBI level 4) and contain one internal prosodic boundary, which could be an IP boundary or an intermediate boundary (ToBI level 3). Additionally, the qualifying segments had to meet specific criteria: a minimum and maximum duration of 2 and 5 seconds; the internal prosodic boundary could not be between the two first or two last syllables of the utterances, and a pause should be annotated wherever a prosodic IP was annotated[1]. The next step consisted in automatically deleting all existent annotated pauses from the stimuli, applying

[†] Work performed while the authors were employed at CoML

[1] This is not systematic as IP can be present without the speaker marking a pause

Table 1: *Examples of stimuli in the English benchmark, for the ProtoSyntax and Lexical tasks. Numbers in subscript correspond to prosodic break tiers in ToBI format.*

task	condition	stimuli
protosyntax	natural	$_4$*She*$_1$ *went*$_1$ *to*$_1$ *jail*$_4$ <***PAUSE***> *for*$_1$ *what*$_1$ *appeared*$_1$ *to*$_1$ *be*$_1$ *a*$_1$ *murder*$_4$
protosyntax	unnatural	$_4$*She*$_1$ *went*$_1$ *to*$_1$ *jail*$_4$ *for*$_1$ *what*$_1$ *appeared*$_1$ *to*$_1$ <***PAUSE***> *be*$_1$ *a*$_1$ *murder*$_4$
lexical	natural	$_4$*She*$_1$ *went*$_1$ <***PAUSE***> *to*$_1$ *jail*$_4$ *for*$_1$ *what*$_1$ *appeared*$_1$ *to*$_1$ *be*$_1$ *a*$_1$ *murder*$_4$
lexical	unnatural	$_4$*She*$_1$ *went*$_1$ *to*$_1$ *jail*$_4$ *for*$_1$ *what*$_1$ *a-*<***PAUSE***> *-ppeared*$_1$ *to*$_1$ *be*$_1$ *a*$_1$ *murder*$_4$

some crossfading (10ms on each side of the pause) to prevent abrupt cut-off or jump in the audio.

2.1.2. Creating the protosyntax and lexical tasks

In the protosyntax task, one stimulus has a pause placed at a "natural" location : a prosodic phrase boundary (ToBI levels 3 or 4). In contrast, the other stimulus has the pause placed at an "unnatural" location where there are no higher level prosodic breaks (ToBI levels 1 or 2). This task aims to assess the models' understanding of structural prosody at the sentence level. In the lexical task, the prosodic boundary is present at a word boundary in the natural condition (levels 1 or 2) and within a word in the unnatural condition. Because the differences between these two conditions are less marked prosodically, lexical knowledge should be required on top of prosodic knowledge to perform well in this task. Examples of both tasks are presented Table 1.

While there is only one possibility for where the pause can be in the natural condition, multiple stimuli can be created for each pair in the unnatural condition. Therefore, we included constraints on the position of the break from the start and end of the audio regarding the number of syllables and seconds. We sampled the final stimuli by implementing similarity losses for these constraints, aiming for the smallest divergence between the distributions of the natural and unnatural conditions. Once the stimuli pairs and break positions were defined, we inserted a 400ms pause with crossfading at the chosen position, resulting in a total of 5,234 pairs in the protosyntax task and 5,178 pairs in the lexical task. Since the pause is artificially inserted in both unnatural and natural conditions, the duration of the two stimuli is the same in both conditions, with the only difference being the location of the pause within the stimuli. Finally, we randomly sampled about 10% of the pairs to create a dev set, resulting in 2,355 *(262)* pairs in the protosyntax task for the test *(dev)* set and 2,330 *(259)* pairs in the lexical task for the test *(dev)* set.

2.1.3. Metric

We employed, in both subtasks, the same metric as the sWuggy and sBlimp metrics proposed in [6] : we compute, for a pair of stimuli, the probability of the model evaluated (or pseudo-probability) to generate these stimuli. If the probability of the "natural" stimuli is higher than the unnatural stimuli, we give a score of 1 for this pair, otherwise a score of 0. The final score corresponds to the average score for all dev or test pairs in the given subtask (see [6] for more details).

2.2. Baselines

We evaluated several self-supervised learning models of speech on the protosyntax and lexical tasks, as well as a human baseline.

2.2.1. GSLM and pGSLM baselines

We evaluated our models against the prosody-aware models presented in [11], which we refer to as pGSLM models. These models are an extension of the "GSLM" models, presented in an earlier study [16], and that we also evaluated. All of these models are trained on a "clean" 6k hours sub-sample of the Libri-Light dataset [17], which is made of English audiobooks. The GSLM models have three main components: an acoustic model (HuBert), a quantizer, and a language model (transformer). Here, we present two versions of these GSLM models, the standard version and the "deduplicated" version, where the units output by the quantizer are deduplicated. The pGSLM models build upon the GSLM models by adding tasks for the language model, which predicts the fundamental frequency (f0) and duration of the following frames, in addition to the frame's content. Because the duration is predicted, the information is also removed from the deduplicated units.

The f0 and duration prediction tasks should enable the model to learn the rhythm and intonation of the speech, which are important cues for understanding the meaning of speech. Hence, this modification allows the pGSLM models to incorporate prosodic information into their predictions. By incorporating this information, the pGSLM models may be able to produce more natural-sounding speech and make more accurate predictions. In this paper, we consider four versions of these pGSLM models, trained on continuous (cont.) or discrete (disc.) input, and with or without a prosodic shift in the prediction (see [11] for more details). However, we only compute the pseudo-probability based on the original LM token prediction, similarly to what is done in [11].

2.2.2. STELA baselines

The tasks were evaluated on English and French models from [18], which we refer to as STELA models. By also using models trained in an unseen language (French), we can test the effect of nativeness on structural prosody. These models, trained on English and French audiobooks, use a Contrastive Predictive Coding objective [2, 3] for the first acoustic model, which generates continuous representations. These representations are then quantized using k-means, and a language model (LSTM) is trained on the resulting units. Additionally, these models are trained on varying amounts of data, allowing for the creation of training size curves[2]. To ensure a fair comparison with other models, a "deduplicated" version of the 3,200 hour English model was also trained.

2.2.3. Human evaluation

Finally, we ran a human evaluation by presenting the same protosyntax and lexical tasks in an online experiment. We used Mechanical Turk to recruit 389 participants for 790 sessions, which sessions were compensated $1.5 USD each.In a session, the participant was presented with a series of 7 pairs from the protosyntax and 7 pairs from the lexical task in a random order, along with two examples at the beginning of the session,

[2]To ensure comparable results, models at all train sizes are overall trained on the same amount of data, with more models being trained on the smallest train sizes. See [18] for more details.

Table 2: *ProsAudit accuracy scores (%) for the different models.*

model	dataset	protosyntax dev	protosyntax test	lexical dev	lexical test
STELA	3200h audiob.	**72.5**	**74.9**	68.7	68.3
STELA deduplicated	3200h audiob.	58.0	58.5	48.7	46.7
GSLM	Librilight 6k	58.8	58.1	53.3	54.1
GSLM deduplicated	Librilight 6k	67.2	66.5	73.8	70.5
pGSLM - cont.	Librilight 6k	65.7	66.8	73.8	71.5
pGSLM - cont. + shift	Librilight 6k	69.1	66.8	**74.9**	71.9
pGSLM - disc.	Librilight 6k	67.6	64.8	74.5	71.1
pGSLM - disc. + shift	Librilight 6k	69.1	65.9	72.6	**72.9**

and four additional 4 controls[3] specifically easier in order to filter out participants who were not paying attention to the task. For each pair, a webpage was presented to them, asking them to listen to two audios and decide which of the two audio was the most natural (the condition order was randomised), as well as how sure they were of their decision (slightly, moderately, strongly). We also included two examples at the beginning of the test, as well as 4 controls in order to filter out participants who were not doing the task correctly. Participants were paid $1.5 USD per session. For the analyses, we discarded all non-native English participants, as well as all sessions where the participant did not correctly pass at least 5 out of the 6 examples + controls, resulting in a total of 255 participants (515 sessions). Finally, we only included in our analysis stimuli pairs that were listened to at least 5 times. We refer to these final stimuli as the "human subset", independent from the dev and test sets, which is composed of 521 pairs for the protosyntax task and one of 510 pairs for the lexical task.

3. Results

3.1. Benchmark

The benchmark (with both the dev and test sets) is available as part of a new evaluation metric for Track 4 of the Zero-Resource Speech challenge[4], and a new leaderboard is also setup.

Scores on the protosyntax and lexical prosodic tasks evaluated on English models are presented in Table 2. First, all models perform above chance in both the protosyntax and lexical tasks, suggesting that all of these models have some prosodic knowledge about the structure of sentences and words.

An interesting observation is that while the GSLM and pGSLM models perform better on the lexical task, suggesting a better understanding of word boundaries, the STELA models actually perform better on the protosyntax task, indicating a stronger sensitivity to the prosodic structure of sentences. This is surprising given that the pGSLM models are specifically trained to also predict the duration and pitch of the next frames, in addition to the content of the frame. However, it must be reminded here that the pseudo-probability for the pGSLM model is only computed at the token level. A potential improvement for future work would be considering the duration and pitch losses when computing the pseudo-probability.

There is little difference in performance between the four pGSLM models, although the continuous pGSLM with prosodic shift performs slightly better overall. More surprisingly, the GSLM model with deduplicated units performs similarly to the pGSLM models, even though the deduplication process removes the duration information, which is an essential aspect of prosodic information (while the GSLM models are also trained on deduplicated units, the duration information is reinjected as a separate loss). This suggests that the lexical and grammatical information present in the units may be sufficient for the model to perform well on the two tasks, even without the prosodic information. It is possible that the model learned to extract the relevant information from the units even with the duration information removed.

Finally, while deduplicating the units in the GSLM model greatly helps in both tasks, the opposite happens when deduplicating the units with the STELA model, which yields much poorer results. While surprising at first, this could be caused by multiple factors that vary between the two model types: the quantity of data (there is nearly double the amount of data in the GSLM models) and the architecture of the language model (an LSTM in the case of the STELA model and a transformer in the GSLM ones). It would be interesting to study further the impact of deduplication on other language evaluation tasks for these two types of models.

Table 3: *ProsAudit accuracy scores (%) for the human evaluation and best performing STELA and GSLM models, on the test and human subset.*

	protosyntax test	protosyntax human set	lexical test	lexical human set
humans native	-	80.50	-	60.38
STELA	74.86	73.32	68.28	71.18
pGSLM	65.94	64.88	72.88	73.33

3.2. Further analyses

To better understand the presence of structural prosody in SSL models, we conducted a series of analyses using our ProsAudit benchmark.

Human/Machines comparison. Table 3 shows the scores of the human cohort on the lexical and protosyntax tasks, compared to the best-scoring STELA and (p)GSLM models on the same human subset[5]. Humans perform better than chance on both tasks. The protosyntax results (80.5%) are slightly lower than those presented in [13] (93.2%), where participants were presented with a Japanese version of the task. However, this can be explained by the fact that the stimuli in [13] were much more curated and manually selected than the ones in ProsAudit, which compensates this by having much more stimuli. Still, in the protosyntax task, humans perform better than the models, acting as a topline. On the lexical task, both models score higher than humans, who were less confident in their ratings than in the protosyntax task. This may be due to the wording of the experiment, asking the participants to choose the most natural sentence without any cues to focus on the lexical aspect. Moreover, in this task, both conditions sound pretty unnatural, as the "correct" stimuli have a pause placed at a word boundary with no prosodic break. This hypothesis is supported by additional analyses indicating lower participant confidence in the lexical task compared to the protosyntax task.

Effect of size. Results averaged over all STELA models from a same size of the training dataset, both in English and French, are presented in Figure 1. Focusing solely on the English models (solid line), we can note multiple things. First, even models trained on a small amount of data (50 hours) achieve above-chance performance in the prosodic tasks, indicating that struc-

[3]For a control pair, the pause is inserted at an intonation phrase boundary (natural) and within a word (unnatural).

[4]https://download.zerospeech.com/datasets/prosaudit-dataset.zip

[5]Computation of the correlation between human and machine scores for the different items was not possible due to the high variability and insufficient number of responses per items in the human condition.

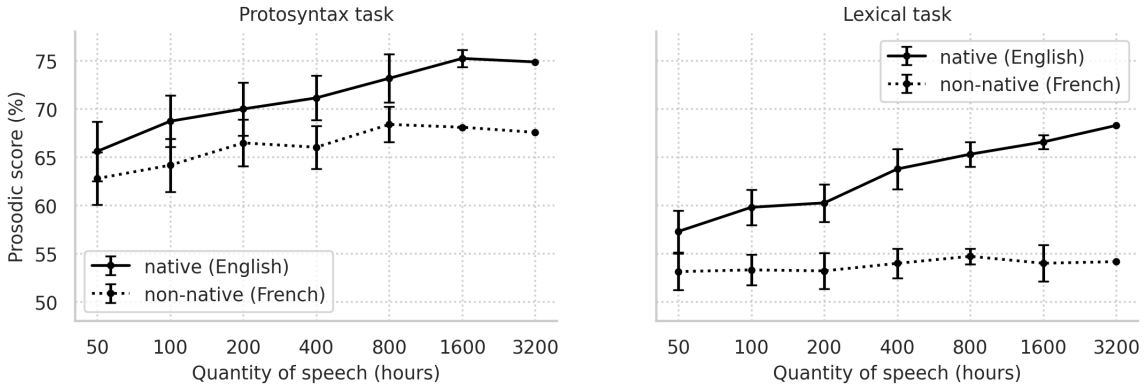


Figure 1: *ProsAudit accuracy scores (%) for the Protosyntax and Lexical subtasks, on the STELA models, w.r.t. the training size.*

tural prosodic knowledge can be acquired from limited data. Additionally, the results show a clear trend of improvement in performance as the size of the training dataset increases, particularly for the lexical task, which exhibits a logarithmic growth. However, in the case of the protosyntax task, the study observes a plateau in the improvement of performance between 1,600 and 3,200 hours, potentially indicating a ceiling effect, meaning that beyond this point, increasing the size of the dataset does not have a significant impact on the model's performance.

Effect of nativeness. We analysed the performance of models trained on French data when evaluated on English prosodic tasks, as depicted by the dotted line in Figure 1. When considering only models trained on the largest amount of data, it is found that there is a clear native advantage for English models, achieving better performance than their French counterparts in both tasks. This difference is more pronounced in the lexical task, where a strong lexical knowledge of English is a key factor for success. A comparison of the overall developmental curves reveals that while the performance of French models improves in the protosyntax task with increasing training data, this is not the case for the lexical task. The lack of improvement for French models in this second task results in a widening performance gap between native and non-native models. These findings, although unsurprising, indicate that non-native models cannot acquire the necessary lexical knowledge to excel in this task solely through additional training data.

4. Discussion & Conclusion

We introduced ProsAudit, a zero-shot benchmark for measuring the English prosodic knowledge of speech SSL models, made of two subtasks. The protosyntax task assesses the model's understanding of structural prosody at the syntactic level. All the models we evaluated performed well above chance in this task, indicating that this knowledge is embedded in speech SSL models. Besides, models trained on another language perform relatively well in this task, which is in line with findings in humans [13]. This suggests that some prosodic knowledge can be universal [19]. However, the gap between native and non-native models increases with more data, suggesting that some prosodic cues are language-specific, indicating that the models are in line with findings from the psycholinguistics literature [20].

The lexical task requires both hierarchical prosodic knowledge at the sentence level and some lexical knowledge, as prosody is not always enough to differentiate between word boundary and within word breaks. The GSLM models (with and without prosody) performed better on this task than the protosyntax one. This could be because they have some strong lexical knowledge of English (these models perform relatively well on lexical metrics, see [16, 7]), levelling up with their knowledge of prosody. Conversely, the STELA models do not perform as well on this task as the protosyntax one, although they still score way above chance, suggesting their lexical knowledge is not strong enough to surpass the protosyntax task. Unsurprisingly, there is a much larger native effect in this task, with non-native models performing only slightly above chance, regardless of the size of the training set. In contrast, the native models' performance is strongly correlated with data size. Further research could examine how much lexical knowledge in a model correlates with their performance on the ProsAudit lexical task.

We also found that the pGSLM models, despite supposedly encoding prosody-specific features and yielding better mean opinion scores on downstream tasks [11], score only slightly better on the protosyntax and lexical tasks than their vanilla GSLM counterparts. Two things could explain these results. First, we only compute the pseudo-probability taking into account the loss at the token level (discarding losses at the pitch and duration levels), and finding news ways to generate this pseudo-probability by taking into account these two components could increase the models' scores. Second, the benchmark we propose here only evaluates specific aspects of prosody at the structural level, while other aspects of prosody as emotion, not evaluated in this benchmark, could be more related to the metrics presented in [11].

To conclude, we propose for the first time a new zero-shot benchmark for evaluating structural prosodic knowledge in speech models, along with a human evaluation topline. We hope that it will inspire further research to enhance the prosodic capabilities of SSL models. This is crucial, as research has demonstrated that models with better prosodic understanding lead to more advanced generative speech models (as seen in [11]). Additionally, the benchmark is now incorporated into Track 4 of the Zero-Resource Speech challenge [6], providing a platform for continuous improvement and comparison of models. In the future, we aim to develop this benchmark in languages other than English, and to expand the benchmark to subtypes targeting other aspects of prosody.

Acknowledgments. MS's work was partly funded by l'Agence de l'Innovation de Défense and performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011012315). ED in his EHESS role was supported in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute) and a grant from CIFAR (Learning in Machines and Brains).

5. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[3] M. Riviere, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7414–7418.

[4] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[5] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 323–330.

[6] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," in *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*, 2020.

[7] E. Dunbar, M. Bernard, N. Hamilakis, T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, and E. Dupoux, "The Zero Resource Speech Challenge 2021: Spoken Language Modelling," in *Proc. Interspeech 2021*, 2021, pp. 1574–1578.

[8] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[9] A. Cutler, D. Dahan, and W. Van Donselaar, "Prosody in the comprehension of spoken language: A literature review," *Language and speech*, vol. 40, no. 2, pp. 141–201, 1997.

[10] D. Dahan, "Prosody and language comprehension," *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 6, no. 5, pp. 441–452, 2015.

[11] E. Kharitonov, A. Lee, A. Polyak, Y. Adi, J. Copet, K. Lakhotia, T. A. Nguyen, M. Riviere, A. Mohamed, E. Dupoux *et al.*, "Text-free prosody-aware generative spoken language modeling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8666–8681.

[12] G.-T. Lin, C.-L. Feng, W.-P. Huang, Y. Tseng, T.-H. Lin, C.-A. Li, H.-y. Lee, and N. G. Ward, "On the utility of self-supervised models for prosody-related tasks," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 1104–1111.

[13] B. Ludusan, M. Morii, Y. Minagawa, and E. Dupoux, "The effect of different information sources on prosodic boundary perception," *JASA Express Letters*, vol. 1, no. 11, p. 115203, 2021.

[14] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.

[15] K. E. Silverman, M. E. Beckman, J. F. Pitrelli, M. Ostendorf, C. W. Wightman, P. Price, J. B. Pierrehumbert, and J. Hirschberg, "Tobi: A standard for labeling english prosody." in *ICSLP*, vol. 2, 1992, pp. 867–870.

[16] K. Lakhotia, E. Kharitonov, W.-N. Hsu, Y. Adi, A. Polyak, B. Bolte, T.-A. Nguyen, J. Copet, A. Baevski, A. Mohamed *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[17] J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.-E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen *et al.*, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7669–7673.

[18] M. Lavechin, M. de Seyssel, H. Titeux, H. Bredin, G. Wisniewski, A. Cristia, and E. Dupoux, "Can statistical learning bootstrap early language acquisition? a modeling investigation," *PsyArXiv preprint PsyArXiv:rx94d*, 2022.

[19] A. D. Endress and M. D. Hauser, "Word segmentation with universal prosodic cues," *Cognitive psychology*, vol. 61, no. 2, pp. 177–199, 2010.

[20] B. Höhle, R. Bijeljac-Babic, B. Herold, J. Weissenborn, and T. Nazzi, "Language specific prosodic preferences during the first half year of life: Evidence from german and french infants," *Infant Behavior and Development*, vol. 32, no. 3, pp. 262–274, 2009.

3.4 Modelling language learning in cognitive science

In the first two sections of this chapter, we proposed a framework for modelling early language acquisition and reported first results and developmental curves on our model at different linguistic levels (phonetic, lexical and prosodic). In this section, we tackle the implications that learning simulations (models of the learner) such as STELA can have for research in cognitive science and in early language acquisition in particular.

This section is presented under the form of a paper (as it follows the work published in section 3.2, this work was also done in close collaboration with fellow PhD student Marvin Lavechin, with whom the first-authorship of the present paper is shared):

de Seyssel, M.*, Lavechin, M.* & Dupoux, E. (2023). Simulating early language acquisition: first results and challenges. *Journal of Child Language*, 1-24. doi:10.1017/S0305000923000272

3.4.1 Realistic and broad-scope learning simulations: first results and challenges

Paper summary

Starting from the observation that language acquisition research is currently facing a theory crisis due to the shortcomings of existing theories (Fried, 2020; Muthukrishna and Henrich, 2019; McPhetres et al., 2021), we first propose a set of four criteria that a unifying theory should meet to overcome these limitations. Specifically, an effective theoretical framework should be:

- *Causal*: It should clearly specify the learning mechanism involved in language acquisition.
- *Quantitative*: It should generate numerical outcomes that can be compared to human performance data.
- *Realistic*: It should provide realistic specifications of the environment and outcome measures.
- *Broad-scope*: It should encompass multiple levels of language acquisition, including phonetics, syntax, semantics, and pragmatics.

By meeting these criteria, a unifying theory can provide a comprehensive and cohesive explanation of how humans learn language and may help to resolve the current theoretical crisis in language acquisition research.

In a second step, we discuss the potential of learning simulations to fulfil the four criteria for a unifying theory of language acquisition. While most learning

simulations already satisfy the causal and quantitative criteria, we emphasise the importance of building *realistic* and *broad-scope* simulations to fully address the complexities of language acquisition.

Building on the STELA learning simulation introduced in Section 3.2, we focus on one of the main use cases of learning simulations in language acquisition research: their use as a proof of concept. The objective of a proof of concept is to demonstrate that a particular mechanism or input can produce outcomes similar to those found in humans, providing evidence that the mechanism or input is a viable candidate for further research. Specifically, we show how we can use learning simulations as a proof of concept for addressing three enduring questions in language acquisition. Building upon our work presented in Section 3.2, we demonstrate that these simulations can provide insights into the statistical learning hypothesis in early language acquisition, as well as the hypothesis that learning can occur without linguistic categories. Lastly, we focus on the question of ecological audio and its impact on the plausibility of statistical learning as the only mechanism involved in language acquisition.

Finally, we provide guidelines, limitations and directions for future work on using such learning simulations in language acquisition research by focusing, in turn, on the environment, the learner and the outcome measures. In conclusion, this paper strongly grows on Dupoux (2018)'s reverse-engineering approach, providing concrete illustrations of the benefits of using learning simulations in cognitive science.

Journal of Child Language (2023), 1–24
doi:10.1017/S0305000923000272

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Realistic and broad-scope learning simulations: first results and challenges

Maureen de SEYSSEL[1,2,†], Marvin LAVECHIN[1,3,†] and Emmanuel DUPOUX[1,3]

[1]Laboratoire de Sciences Cognitives et de Psycholinguistique, Département d'Études Cognitives, ENS, EHESS, CNRS, PSL University, Paris, France
[2]Laboratoire de Linguistique Formelle, Université Paris Cité, CNRS, Paris, France
[3]Meta AI Research, Paris, France
Corresponding authors: Maureen de Seyssel and Marvin Lavechin; Emails: maureen.deseyssel@gmail.com; marvinlavechin@gmail.com

(Received 04 October 2022; revised 24 April 2023; accepted 04 April 2023)

Abstract
There is a current 'theory crisis' in language acquisition research, resulting from fragmentation both at the level of the approaches and the linguistic level studied. We identify a need for integrative approaches that go beyond these limitations, and propose to analyse the strengths and weaknesses of current theoretical approaches of language acquisition. In particular, we advocate that language learning simulations, if they integrate realistic input and multiple levels of language, have the potential to contribute significantly to our understanding of language acquisition. We then review recent results obtained through such language learning simulations. Finally, we propose some guidelines for the community to build better simulations.

Keywords: Language acquisition; computational modelling; phonetic learning; word learning; phonetic categories

What is needed and why?

Theory in crisis

The field of language acquisition is prolific, with an extensive range of high-quality research published every year. However, there has been surprisingly slow progress in solving some long-standing controversies regarding the basic mechanisms that underlie language acquisition. For instance, do infants learn language primarily from extracting statistics over speech inputs (Romberg & Saffran, 2010; Saffran & Kirkham, 2018), from examining cross-situational correlations over multisensory inputs (Smith & Yu, 2008; Suanda, Mugwanya & Namy, 2014; Yu & Smith, 2017; Zhang, Chen & Yu, 2019), or by relying on social interactions and feedback (Tomasello, 2003; Tsuji, Cristia & Dupoux, 2021; Yu & Ballard, 2007)? Do they learn by leveraging discrete linguistic categories or

†M.S and M.L. contributed equally to this work. Authorship order was decided by a coin flip.

continuous sensory representations (Kuhl et al., 2008; McMurray, 2021)? Do they rely on language-specific or domain-general learning mechanisms (Elman, Bates & Johnson, 1996; Karmiloff-Smith, 1994; Pinker, 1994)? Such a lack of headway may be due in part to the 'replication crisis': the experimental study of human cognition in general and infant cognition, in particular, is inherently noisy and difficult (Frank, Bergelson, Bergmann, Cristia, Floccia, Gervain, Hamlin, Hannon, Kline & Levelt, 2017), slowing down cumulative progress. Here, we explore the possibility that there is, in addition, a 'theory crisis'. To say it bluntly, perhaps, current theories have shortcomings that prevent us from even finding the right experimental setup to make progress on basic questions about learning mechanisms.

Several papers have already been devoted to the theory crisis in psychology in general; psychological theories have been claimed to be mere statistical model fitting (Fried, 2020), too descriptive or fragmented (Muthukrishna & Henrich, 2019), or to not contribute in cumulative theory building (McPhetres et al., 2021). In developmental psychology, Kachergis, Marchman, and Frank (2021) called for a 'standard model' that would allow integration of results in a cumulative fashion. In this paper, we explore the possibility proposed in Dupoux (2018) that recent advances in machine learning could help address the theory crisis through systems that realistically simulate how infants learn language in their natural environment. Such learning simulations are computer models that would ideally learn from similar inputs as the ones available to infants (raw sensory data), and reproduce the broad spectrum of outcome measures as obtained in laboratory experiments or corpus studies. To the extent that these new computer models are powerful enough to address the complexity and variability of data available to infants during language development, they could help us make progress in some of the aforementioned controversies. At best, such learning simulations can provide proof of principle that a given hypothesis (e.g., the statistical learning hypothesis) can account for learning outcomes as observed in infants. In addition, they can help us go beyond said long-standing controversies by providing new insights into the learning process and a wealth of associated quantitative predictions.

In this paper, we first discuss how these new types of learning simulations are complementary to more familiar theoretical approaches in cognitive development and argue that they provide one step towards the needed cumulative integrative theories or standard models. We then present STELA, a recent learning simulation implementing the hypothesis that infants are statistical learners, and show how it provides insights into some long-standing controversies.

Varieties of theories in language acquisition

The theoretical landscape of language development is vast and complex. Even if one focuses on early language development, there are wild varieties of theoretical approaches that differ not only in scope (the range of phenomena they cover) but also in style (verbal, statistical, formal, computational). Here, far from making a comprehensive survey of these approaches, we attempt to classify them into types and sort them along dimensions that outline their respective strengths and weaknesses with regard to addressing basic questions/controversies about learning mechanisms. Familiar types are verbal frameworks (among others: The competition model: MacWhinney & MacWhinney, 1987; WRAPSA: Jusczyk, 1993; Usage-based theory: Tomasello, 2005; NML-e: Kuhl et al., 2008; PRIMIR: Werker & Curtin, 2005), which weave a narrative around a large body of experimental research using verbally defined concepts, sometimes complemented by

box-and-arrow schemas (e.g., the ScALA framework from Tsuji et al., 2021). Correlational approaches (e.g., Fernald, Marchman & Weisleder, 2013; Hart & Risley, 1995; Swingley & Humphrey, 2018) aim to identify the main variables that predict language development outcomes through statistical models. Formal models (e.g., Jain, Osherson, Royer & Sharma, 1999; Tesar & Smolensky, 2000) and computational models (e.g., Brent, 1997) aim to study how algorithms can learn language through mathematical proofs or empirical study of the learning outcomes. All theoretical approaches of early language development recognise that infants receive inputs from their environment, and have a learning mechanism, which produces a linguistic competence that can be accessed through outcome measures. The differences between these theoretical approaches lie in the simplifying assumptions and degree of specifications they make about inputs, learning mechanisms and outcome measures. We distinguish four dimensions or axes to sort these theoretical approaches: Causal versus Correlational, Quantitative versus Qualitative, Realistic versus Abstract, and Broad Scope versus Narrow Scope.

Causal/Correlational

A theory is causal when it provides a specification/implementation of the learning mechanism underlying language acquisition; it is correlational when it focuses on the input/outcome relationship without specifying a learning mechanism. A correlational model can outline the important factors that drive learning and therefore provide insights into the development of learning mechanisms. However, only a causal model can provide proof of principle that a postulated learning mechanism is sufficient to reproduce a developmental outcome given an input. As a result, to the extent that they can be effectively implemented, causal models are better positioned to resolve disagreements about learning mechanisms than correlational models.

Quantitative/Qualitative

A theory is quantitative if it can produce numerical outcomes that can be compared to human performance. It is qualitative when it produces predictions about the possible presence of a significant effect without a numerical prediction about its strength. Qualitative models are useful to inspire novel experimental paradigms, and provide insights about learning mechanisms, but are hard to refute and difficult to compare to one another. Quantitative theories make very precise predictions and can be compared to one another by computing the degree of fit of their predictions against some observed outcome. As a result, they are better positioned to solve disagreements about learning mechanisms than qualitative theories.

Realistic/Abstract

A theory is realistic when its model of the environment is as close as possible to the actual sensory/motor environment of the child. It is abstract when the environment is specified through synthetic data, or human/categorical annotations of observed environments (e.g., textual transcriptions). Abstract theories are useful because they enable a high degree of control and interpretability and provide insights into what type of input information can yield particular outcomes. However, they cannot prove that their conclusions apply to real-world data as perceived by infants and are therefore not very informative when it comes to solving long-standing controversies. Realistic theories, in

contrast, to the extent that they can be effectively implemented, are better positioned: because they directly reproduce the learning outcomes associated with a given input and learning mechanism.

Broad/Narrow Scope

A theory has a broad scope if it encompasses not one single linguistic level (phonetic, morphological, syntactic, semantic, etc.) or phenomenon but several at once. Narrow Scope theories are useful in focusing on learning specific representations, assuming all other representations are fixed. However, many controversies about learning mechanisms arise because of co-dependencies between linguistic levels, making it problematic to assume all levels are fixed except one. Being able to account for how infants can learn jointly all of these levels is at the heart of solving so-called 'bootstrapping' problems that are integral to language learning.

In Table 1, we position some familiar theoretical approaches in terms of these four axes. This characterisation may seem overly simplistic or reductionist, but we hope it will help outline the specific contribution of learning simulations. Verbal frameworks typically have a broad scope and embrace the complexities of the child's real environment. They are causal to the extent that they mention specific learning principles but are not on the quantitative side. They are still the single most influential theoretical approach for infant language learning, providing insight into large quantities of experimental results. However, they resist empirical refutations or amendments because of their qualitative nature. Correlational models are on the quantitative side and integrate many variables and levels. When informed by a corpus of infant/caretaker interactions, they can reveal the relationships between input quality, quantity, and language outcome (Fernald et al., 2013; Hart & Risley, 1995). However, because they are not causal and rely on abstract variables derived from the input, they cannot directly speak to learning mechanisms. Computational/formal models (henceforth called learning simulations) are both causal and quantitative, but their ability to significantly impact controversies about learning mechanisms depends on the breadth of their scope and their degree of realism or abstraction. We discuss such models in more detail in the next section.

A brief history of learning simulations

For a long time, scientists with various backgrounds, from formal linguistics to developmental psychology and artificial intelligence, have contemplated the possibility of building mathematical models or computer simulations of language learning in infants. The hope was that building a simulated ersatz of the infant would reveal the formal conditions

Table 1. Four dimensions along which theoretical approaches of language acquisition can be sorted

Properties	Verbal Framework	Correlation Model	Learning Simulation		
Causal	✗ / ✓	✗		✓	
Quantitative	✗	✓		✓	
Realistic	✗	✗	✗	↔	✓
Broad Scope	✗	✓	✗	↔	✓

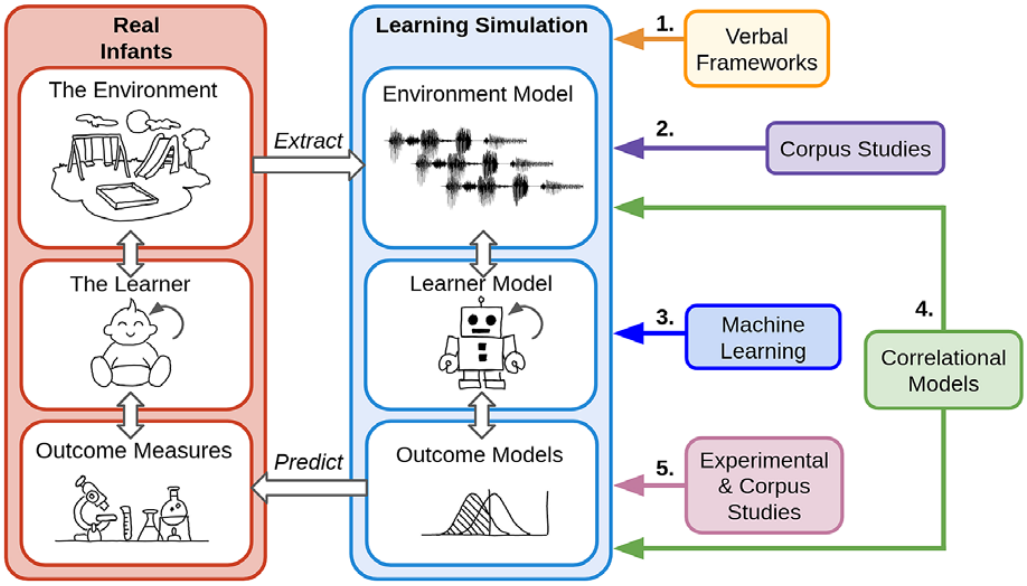


Figure 1. General outline of a realistic learning simulation (centre) in relation to real infants (left) and traditional theoretical approaches (right). 1. Verbal frameworks inspire and help set up the entire language learning simulation by describing the environment, learner, and outcome models; 2. Corpus studies of children's input help us build realistic models of the environment. In the best case, the model of the environment is a subset of a real environment, obtained through child-centred long-form recordings, for instance; 3. Machine learning provides effective artificial language learners. The learner model is relatively unconstrained as learning mechanisms used by the real learner (i.e., infants) remain largely unobservable; 4. Correlational models describe how the input should relate to the outcome measures; 5. Experimental and corpus studies of children's outcomes show how we can evaluate learning outcomes of the artificial learner. The real versus predicted outcome measures allow us to compare humans to machines and provide new predictions for correlational models that relate input to outcomes in infants.

for learning (Pinker, 1979), would allow us to better formulate hypotheses about how infants actually learn (Frank, 2011; Meltzoff, Kuhl, Movellan & Sejnowski, 2009) or would yield machines that learn in a graceful and robust fashion (Turing, 1950). Here again, the diversity of the proposed models is too large to be reviewed (see Dupoux, 2018, for an attempt). Instead, we classify the approaches based on the dimensions which we claim are central to answering key questions about learning mechanism: realism and scope.

As illustrated in Figure 1, all learning simulations consist of three components: a model of the environment, a model of the learner, and a model of the outcome measure. The model of the environment specifies the type of inputs/interactions available to the learner. The learner updates itself using a learning algorithm based on its interaction with the environment. The outcome measures of the learner are measured after exposure to speech. Where learning simulations differ is how they implement these three components.

Focusing on AI-inspired models, the most visible trend historically has been on how to implement the learner. Early models (e.g., Anderson, 1975; Kelley, 1967) were rule-based. The second phase was probabilistic models (e.g., Brent, 1996; de Marcken, 1996), followed by connectionist and deep learning models (Brown et al., 2020; Elman, 1990), each phase replacing hand-wired components with more and more powerful learning systems. As far as we are concerned, the way in which the learner is implemented is irrelevant. What counts is whether the learning mechanism actually reproduces the learning outcome or

not, given infants' input[1]. More relevant to our argument, another trend can be seen regarding the model of the environment, moving from synthetic data (e.g., Elman, 1990; Vallabha, McClelland, Pons, Werker & Amano, 2007) to transcribed corpora (e.g., Bernard et al., 2020) and, more recently, to raw audio and images or video recordings (Räsänen & Khorrami, 2019; Schatz, Feldman, Goldwater, Cao & Dupoux, 2021). Finally, the first models were focused on learning a single linguistic level (e.g., phonetic categories: Vallabha et al., 2007; word forms: Brent, 1999; word meanings: Roy & Pentland, 2002; syntax: Pearl & Sprouse, 2013), and more recent approaches would learn several levels jointly (phonemes and words: Elsner, Goldwater & Eisenstein, 2012; syntax and semantics: Abend, Kwiatkowski, Smith, Goldwater, Steedman, 2017; phonetics, words and syntax: Nguyen et al., 2020).

In other words, thanks to recent progress in machine learning and AI (Bommasani et al., 2021), learning models that are simultaneously of broad scope and able to ingest realistic data are around the corner. Obviously, a complete model that would feature maximal scope (integrating all relevant input and output modalities for language and communication) and maximal realism (using sensory data indistinguishable from what infants experience) is still out of reach. In the next section, we examine STELA, a recently proposed model (Lavechin, de Seyssel et al., 2022c) and argue that even though it is limited both on scope and realism, this work can help us make nontrivial progress on some of the long-standing controversies regarding language learning mechanisms.

Before moving on, let us clarify that we are not claiming that broad-scope realistic simulations are the only valuable approach. Narrow-scope abstract models still have valuable contributions to make (e.g., Frank, Goodman & Tenenbaum, 2009; Kachergis et al., 2021). First, contrary to many realistic and broad-scope models, abstract and narrow models are interpretable and therefore allow building bridges with verbal frameworks. They are also more tractable and can be easily modified and experimented on in a way which is more difficult with larger models. Finally, one can view abstract learning simulations as "control" experiments: by comparing an abstract and a realistic learning simulation implementing a similar learning mechanism, we can gain knowledge on the role of specific abstractions made by the learner.

What has been achieved so far?

Among the competing hypotheses regarding the learning mechanisms that underlie early language learning, the one that seems the most natural to approach with learning simulations is the statistical learning hypothesis (Pelucchi, Hay & Saffran, 2009; Saffran, Aslin & Newport, 1996). It posits that infants learn at least some linguistic levels (phonetic, lexical and morphosyntactic) through a statistical or distributional analysis of their language inputs. The idea has a long history (Rumelhart, McClelland & MacWhinney, 1987; Skinner, 1957) and has generated many controversies (Chomsky, 2013; Fodor & Pylyshyn, 1988) and mathematical investigation (Gold, 1967; Jain et al., 1999). But it is also the simplest hypothesis to implement in a learning simulation. If one equates language input to the auditory modality, the corresponding learning simulation would simplify the environment to audio recordings, and the learner to a probabilistic model

[1]Many developmental scientists worry about the so-called 'psychological plausibility' of these various kinds of models. Following Frank (2014), we believe that issues of plausibility have either to be formulated as outcome measures that the model should reproduce, or should be disregarded.

that accumulates statistics paying no attention to other modalities or context, nor interacting with its environment.

Here, we present recent work on simulating a statistical learner for language acquisition (Lavechin et al., 2022b; Lavechin, de Seyssel et al., 2022c). We present the simplifying assumptions made in these simulations and reflect on how simulated learners compare to infants. Then, we go over different use cases of such a simulation by showing how some of the skills the simulated learner has acquired through exposure can help shed light on some long-standing controversies in our understanding of language acquisition in infants.

We focus on a high-level description of this simulation as we believe it makes it easier to appreciate its lessons. However, readers interested in the technical details can refer to the original paper (Lavechin, de Seyssel et al., 2022c). We will also list specific research use cases that the framework helped deepen. By doing so, we illustrate concretely how such realistic learning simulations can help future research, both in terms of proof of feasibility and inspiration for research.

Introducing STELA

Lavechin, de Seyssel et al. (2022c) introduced STELA (STatistical learning of Early Language Acquisition), a language learning simulation that tackles the problem of discovering structure in the continuous, untranscribed, and unsegmented raw audio signal. As said above, the scope of this simulation is restricted to the statistical learning hypothesis, where infants learn passively and uniquely by extracting statistical cues from what they hear (see Table 2). In this section, we present the model of the environment, the model of the learner, and the model of the outcome measures used in STELA.

The environment

STELA specifies the environment as raw audio speech recordings. For this to remain relevant, we need to restrict the quantity of speech within a plausible range of data. Current estimates of cumulative speech experiences by one year of age vary from around 60 hours (Cristia, Dupoux, Gurven & Stieglitz, 2019) to approximately 1,000 hours (Cristia, 2022). In STELA, the data comes either from open-source audiobooks with quantities varying from 50 to 3,200 hours covering the observed range. Admittedly, the infant's language environment is different from audiobooks. On the one hand,

Table 2. Non-exhaustive list of language learning assumptions for infants and whether they are included within the STELA simulation

Assumption	STELA
Infants are statistical learners (Bulf, Johnson & Valenza, 2011; Romberg & Saffran, 2010; Saffran et al., 1996)	✓
Quantity of speech input predicts language outcome (Newman, Rowe & Ratner, 2016)	✓
Modalities other than speech can be useful in language learning (Abu-Zhaya, Seidl, Tincoff & Cristia, 2017; Seidl, Tincoff, Baker & Cristia, 2015).	×
Infants learn by **interacting** with peers – reinforcement learning (Kuhl, Tsao & Liu, 2003; Nelson, 2007; Snow, 1989; Yu, Ballard & Aslin, 2005)	×

audiobooks contain clearly articulated speech (read speech) and relatively good audio conditions, potentially facilitating learning for the model compared to the spontaneous and noisy speech available to infants (see Lavechin et al., 2022b). On the other hand, audiobooks may use larger vocabularies and more complex sentences than infants' input, potentially putting the model in a more challenging situation than infants (Gleitman, Newport & Gleitman, 1984). Nevertheless, this type of input is in the range of what infants could plausibly hear or overhear and is relatively easier to access in large quantities across languages than long-form recordings. Therefore, they are a good starting point, offering controlled conditions and replicability for the deployment and analysis of such simulations. Long-form recordings represent the extreme in realism that can be achieved in such simulations, but they are less accessible than audiobooks due to privacy concerns (Lavechin, de Seyssel, Gautheron, Dupoux & Cristia, 2022a).

The learner

Elman (1990) was perhaps the first to introduce a practical implementation of a system that learns non-trivial linguistic representations by extracting regularities from language inputs: a simple recurrent neural network trained to predict future words or characters based on past ones. Since then, this idea has been expanded with more complex and larger neural networks trained on increasingly larger datasets. The resulting so-called "language models" can be viewed as models of the probability distribution of sentences and have been shown to generalise beyond the sentences in the training set (Baroni, 2020), reaching near human performances on many language tasks (Liu, He, Chen & Gao, 2019). One major limitation of these models – as models of the infant learner – is that they only take as input words or characters, which are not entities accessible to a learning infant. However, recent breakthroughs in representation learning have made it possible to expand these models to work with raw audio inputs (Borsos et al., 2022; Dunbar et al., 2021; Lakhotia et al., 2021). In a nutshell, these so-called 'Generative Spoken Language Models' replace text with their own discrete representations learnt from the audio and learn a probabilistic model of speech directly from raw inputs.

In Figure 2a, we present the model used in STELA, which has been selected from the class of Generative Spoken Language Models (Dunbar et al., 2021) for its simplicity. From a high-level perspective, the learner can be described as the combination of two components, which are named according to the current practices in machine learning 1) an '*acoustic model*' and 2) a '*language model*'[2]. The acoustic model is fed with raw, continuous waveforms and trained using a form of predictive coding. It learns a vector representation for each slice of 10ms of signal by attempting to predict each of the twelve upcoming slices based on past ones, yielding a prediction over a 120ms time window. An exciting outcome of such a learning procedure is that the model learns representations that successfully abstract away from acoustic details and encode phonetic information. In STELA, we discretise these representations using clustering, yielding a discrete code each 10ms, which is passed onto the language model. This model is similar to Elman's

[2]Although the term 'language model' can sound counterintuitive in the context of phonological and lexical acquisition, as no language-related or language-specific heuristics are integrated into the model, which learns on its own to discover structures in the speech input, we view it from the machine learning point of view, where a language model is simply an algorithm which learns to predict, from a sequential input, the next representation (let it be text, speech or other) based on the previous representations.

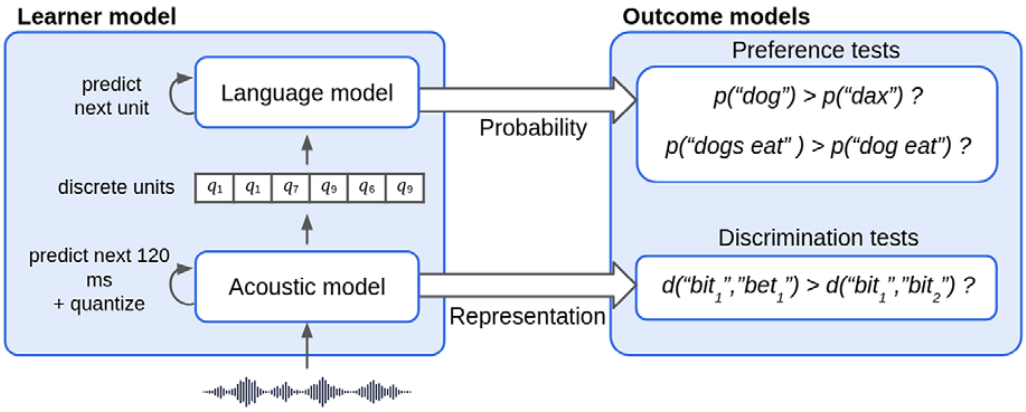


Figure 2. Overview of the STELA learner and outcome measures. a. (left): model of the learner; b. (right): add-on models for two types of outcome measures.

recurrent language model, only using an improved architecture (LSTMs) and more parameters. This model is trained to predict the next code based on past ones. Because the model's output is not a single code, but a probability distribution over all the discrete codes, one can compute the probability of an utterance as the product of the conditional probabilities of each successive code (see Appendix A).

The outcome measures

Several types of outcome measures are used in infant development. Some are provided by caregivers (like the Child Development Inventory, or CDI: Fenson, 2007), who assess whether a word is known or produced by the child, some are linked to the production of the child as attested through transcription of naturalistic corpora (mean length of utterance such as used in Miller & Chapman, 1981 for instance), and some are obtained via in-lab experiments. Here we concentrate on the last type of measure. In principle, a maximally broad language learning simulation would include all linguistic and non-linguistic components (attention, memory, eye movement, etc.) and the artificial learner could just be virtually seated in a virtual lab and be subjected to the same experiments as real babies (Leibo et al., 2018). Here, STELA only simulates a subpart of infants' linguistic competence and therefore has to specify a special add-on module to generate the equivalent of experimental outcome measures. Fortunately, experimental paradigms in infants are relatively simple and can be sorted into two main types: discrimination experiments and preference experiments[3], yielding two types of add-on modules.

Discrimination experiments can vary in how they are conducted in the lab (ABX, AXB, AX, etc.). Still, they all rely on the ability of the learner to compute a perceptual distance between two stimuli (such as 'bit' versus 'bet'). An add-on for ABX discrimination will just need to (a) extract a representation of a stimulus from the learner (typically the activation pattern of some layer) and (b) compute a distance over two representations (typically, the normalised dot product, or the angle between the vectors). In STELA (Lavechin, de

[3]This is a non-exhaustive list. Some experiments use a more complex design where infants are familiarised to some materials (for instance, an artificial language) and then tested using preference or discrimination metrics. This would require the learner to memorise or learn from the familiarisation phase, which has not been implemented in STELA so far.

Seyssel et al., 2022c), this is used to measure phonetic knowledge through a machine ABX sound discrimination task (Schatz et al., 2013) in which the learner has to choose two occurrences of, *e.g., 'bop'* as being closer than one occurrence of 'bop' and one occurrence of *'bip'*. The test is done over thousands of trials and over all possible contrasts of phonemes[4].

Preference experiments rely on the ability to compute a 'preference' or 'probability' associated with an input stimulus. Most learning algorithms learn by minimising an objective function, such as the error made in predicting the future based on the past. We can use the same objective function and apply it to test stimuli: if the stimulus is well represented or considered probable by the model, then the error should be low. Totally novel or unexpected stimuli should give a high error.

In STELA, this is used through the spot-the-word task developed in Nguyen et al. (2020). Here, the model receives a spoken word (*e.g.,* 'apple') and a spoken non-word (*e.g.,* 'attle') matched for syllabic and phonotactic structure. We then look at the model's probability of generating both words. The model is considered correct for the trial if the probability of generating the correct word is higher than the non-word. The same logic can be applied at the syntactic level using pairs of grammatical and ungrammatical sentences (i.e., 'the brother learns' versus 'the brothers learns'), in which the model has to assign a higher probability to the grammatical sentence.

In the next section, we present case studies illustrating how meeting the four above-mentioned properties in a single simulation can help us make theoretical advances.

Results

Learning simulations can either be used as "proof of concept" for particular hypotheses about learning mechanisms or to offer novel predictions, never tested experimentally. Here, we focus on the first use case by addressing three long-standing controversies on language learning mechanisms as applied to the phonetic and lexical levels. In each instance, we use a design which enables us to conduct experiments that are both developmental (obtained by training the same learner on increasing quantity of speech, from 50 hours up to 3200 hours) and cross-linguistic (obtained by training and testing the models on two languages, French and English, deriving scores for the native and non-native language).

Could infants rely exclusively on statistical learning over speech inputs to bootstrap into language?

One of the major conceptual difficulties in accounting for early language acquisition is understanding how the young learner can learn several interdependent linguistic levels simultaneously and gradually. Statistical learning (Saffran et al., 1996) seems a good hypothesis to address this, since it posits that infants gather information about the distribution of sounds. This would naturally yield gradual learning. As for simultaneous learning across levels, it could rest on the idea that probabilities can be gathered at several levels of descriptions simultaneously. Now, the evidence in favour of statistical learning is

[4]It is worth pointing out at this point that the sound contrasts presented in this task are extracted from read speech across many different contexts, while stimuli used in laboratory experiments are more controlled. Potential coarticulation effects make the machine sound discrimination task harder than typical in-lab phone discrimination tasks.

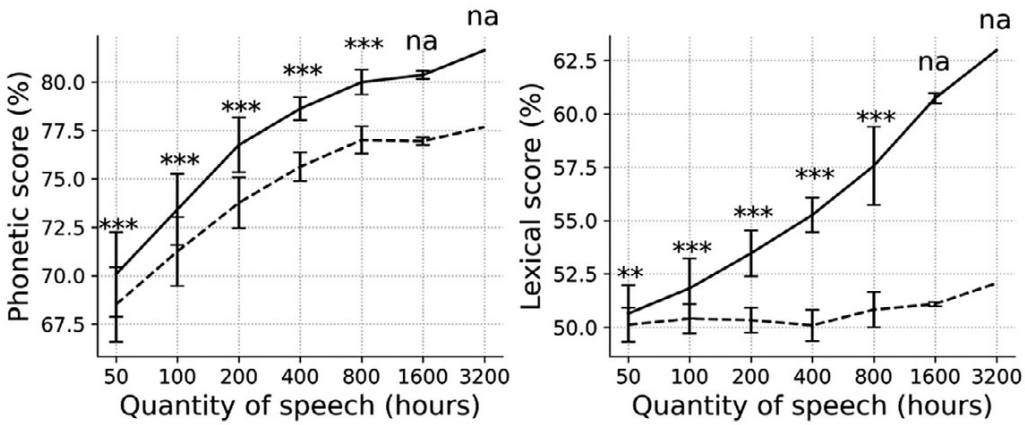


Figure 3. Phonetic (left) and Lexical (right) scores for native and non-native input at different quantities of training data. Phonetic score is expressed in terms of ABX accuracy, obtained by the discrete representations for native and non-native inputs. Lexical score is expressed in terms of accuracy on the spot-the-word task, on the high frequency words for native and non-native inputs. Error bars represent standard errors computed across mutually exclusive training sets. Two-way ANOVAs with factors of nativeness and training language were carried out for each quantity of speech. Significance scores indicate whether the native models are better than the non-native ones. Significance was only computed when enough data points were available to run sensical comparisons. Significance levels: na: not applicable, ns: not significant, * p<.05, **, p<.001, *** p<.0001. Figure taken from Lavechin, de Seyssel et al. (2022c).

itself debated. Experimental evidence in infants only rests on simplified artificial languages (synthetic stimuli, small number of sounds), and it is not clear that this would translate to audio data in which speech sounds are highly variable according to phonetic context, speaker, speaking style and rate, in addition to being potentially contaminated by non-speech background sounds.

In Figure 3, we highlight a few key results obtained by STELA when presented with raw audio from audiobooks (Lavechin, de Seyssel et al., 2022c) and tested at the phonetic level (ABX discrimination) and lexical level (spot-the-word) using the tasks presented in a previous section. The results clearly show above-chance performance on native test stimuli and gradual and parallel learning at both phonetic and lexical levels, with the system being able to discriminate sounds better, and prefer words over nonwords more, as more data is presented to the model. This improvement is weaker when tested on a non-native language (actually, not present at all for the lexical task). Further tests (not shown in Figure 3) using a syntactic task (which is also carried out on the language model component presented in Figure 2) in which the system has to show a preference for legal versus illegal sentences revealed much weaker learning. Only the model trained on the largest quantity of speech available (that is, 3200 hours) was able to show preference on an adjective-noun order task ('the nice rabbit' versus 'the rabbit nice'), with a slightly-above-chance 55% accuracy.

In brief, the STELA simulation suggests that raw speech input only, combined with statistical learning, and more precisely predictive learning, is: 1) sufficient to bootstrap the phonetic, the lexical and only very weakly the syntactic levels; 2) sufficient to reproduce the gradual and overlapping developmental trajectory observed in infants at the phonetic and lexical levels[5]. It is the first time a simulation reproduces the gradual and multilevel learning observed in infants from audio signals, at least when audiobooks are used as input.

[5]Larger models, trained with more audio data are able to pass more complex syntactic tests, and show the beginning of semantic abilities as well (Dunbar et al., 2021), suggesting that the structure of the model can itself learn at several levels beyond phonetic and lexical levels.

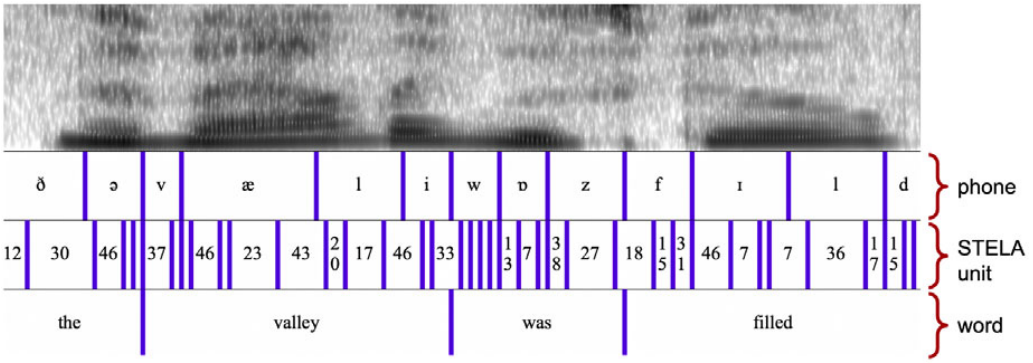


Figure 4. An example spectrogram of an English utterance, along with the corresponding phonemes (top tier) and the units discovered by a STELA model trained on 3200 hours of English. Transcription: "The valley was filled"

Do infants learn and perceive language in terms of linguistic categories?

A second debate concerns whether linguistic categories (phones, words) are necessary building blocks in early language acquisition. On the one hand, linguistic theories describe adult competence in terms of such categories. On the other hand, these categories are language-dependent and therefore need to be learned by infants, who have only access to continuous sensory information at the beginning. Schatz et al. (2021) recently proposed a learning simulation of phonetic learning from raw audio signals based on a probabilistic model using Mixtures of Gaussians. While reproducing observed native advantage effects in phonetic discrimination between Japanese and English phonemes, the learner used in this simulation did not learn phonemes or units that could be described linguistically. These results suggest that phonetic learning can occur without the existence of phonetic categories.

The STELA simulation reproduces this conclusion using a totally different learning algorithm, supporting once again the idea that phonetic categories are not necessary for phonetic learning (see also Feldman, Goldwater, Dupoux & Schatz, 2022). To dive further into this, it is interesting to reflect on how the acoustic model behaves during training concerning the duration of the learnt representations. Pre-exposure (i.e., before the model has received any input) speech is represented within the model as a string of random units. As the model receives speech, it learns to structure this discrete representation: discrete units start repeating themselves, and the sound discrimination accuracy increases. An analysis of the duration of the discrete learnt units revealed that the latter are too short to correspond to phones (43 ms for the learnt units, versus 90 ms for a typical English phone), similarly to what has been found in Schatz et al. (2021). An example of how the discovered units compare to the original phones is presented in Figure 4, where units are clearly shorter than the phones. More surprisingly, the more speech the model receives, the lower the duration of the discrete units. It is essential to note that no constraint is applied to the duration of these units. The model could, in principle, converge to phone-length discrete units, but does no such thing. In other words, the model does not converge to phone-like representations, yet it can still pass phonetic, lexical and, to a certain extent, syntactic tests for which phoneme representations are still often considered a prerequisite[6].

[6]Probing experiments using linear separation revealed however that the representations learned by the acoustic model become more and more structured according to phonetic dimensions like phonetic category

In STELA, it is also possible to ask the question of linguistic categories at higher linguistic levels. Surprisingly, even though the model can distinguish words from non-words, we could not find an indication that the model represents words as such, or would represent the boundaries between words. Yet, the continuous activations found in the hidden layers of the recurrent model contained some approximate linguistic information, as a trained linear classifier was able to classify test words into function versus content words or verb versus adjective/adverb versus noun better than chance, and the separation increased with more input data. These results show that, although the model does not learn discrete and interpretable linguistic categories internally, linguistic information increasingly structures the learnt representations (for more in-depth analyses of the types of units yielded by such models, see de Seyssel, Lavechin, Adi, Dupoux & Wisniewski, 2022; Nguyen, Sagot & Dupoux, 2022; Sicherman & Adi, 2023). Thus, our simulation promotes the view that linguistic categories could be the end product of learning, not their prerequisite.

Can statistical learning alone account for early phonetic acquisition from ecological audio?

One of the largest controversies in language learning orbits around the poverty of the stimulus argument (Chomsky, 1980). This argument states that the input available to infants is too scarce and too noisy to warrant language learning through a general-purpose learning algorithm. Therefore, only a learning algorithm with strong inductive biases would be able to reproduce human language learning. For a long time, this controversy has remained unsolved for lack of learning algorithms that can work even on rather simple inputs. With STELA, at last, we are able to address this controversy, at the level of phonetic and lexical learning. The preceding sections show that a relatively general-purpose system based on predictive coding is able to learn at both levels when fed with audiobooks, but this kind of input may not be realistic enough to correspond to the learning problem faced by infants. Indeed, the audio environment of infants, first of all, contains a majority of non-speech noises, and the little amount of speech that is heard may be under-articulated, reverberated and absorbed by the surrounding obstacles in the environment, and overlaid with various background noises. Could the relatively generic learner of STELA handle such noisy inputs?

One way in which one can revisit this simplifying assumption is by using child-centred long-form recordings, i.e., daylong recordings collected via child-worn microphones as people go about their everyday activities. Lavechin et al. (2022b) exposed the STELA contrastive predictive coding algorithm to such ecological recordings of children's language experiences and found that the discrimination gap between the native and the non-native models vanishes. It is only when supplemented with inductive biases in the form of filtering and augmentation mechanisms (restricting learning to speech parts, taking into account speaker invariance, and making the system resistant to reverberant noise) that the model could exhibit some form of perceptual attunement again (see Figure 5). In addition to this result, Lavechin et al. (2022b) showed that, even in the

(vowels, fricatives, approximants, plosives, etc.), place of articulation for consonants (bilabial, labiodental, dental, etc.), and voicing (voiced or voiceless) as a function of amount of input data. This suggests that the model is learning some phonetic structure from the data even though it is not learning interpretable categories like phonemes.

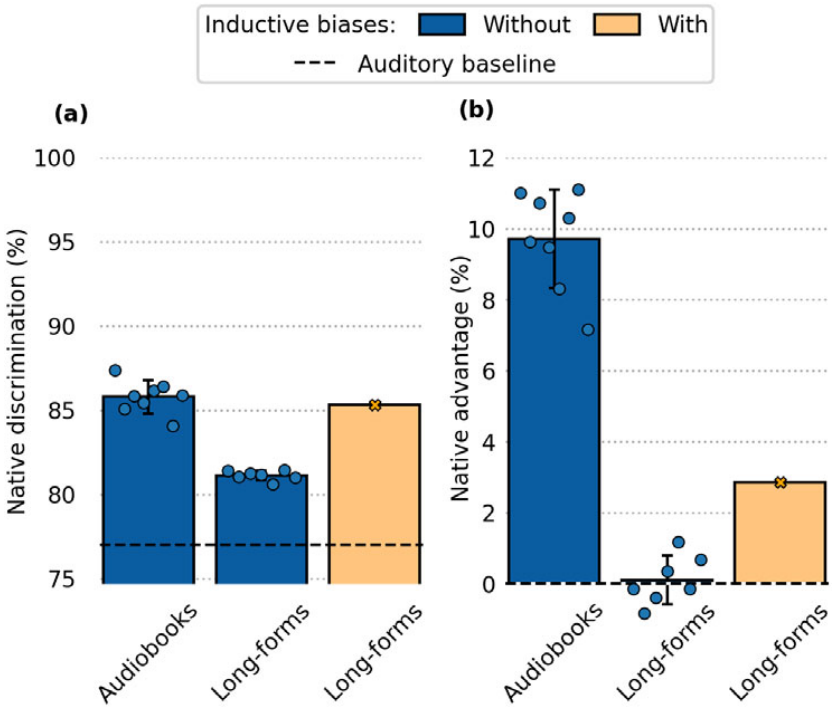


Figure 5. Panel (a) shows native discrimination accuracy, as measured in an ABX discrimination task, obtained by American English and Metropolitan French CPC models (both models are evaluated on phonemes of their native languages). Panel (b) shows native advantage, computed as the average relative difference of the native model and the non-native model, obtained by the same pairs of models (a positive native advantage indicates that the native model is better at discriminating native sounds than the non-native model). Figure adapted from Lavechin et al. (2022b).

presence of inductive biases, the learning speed of the learner was still negatively impacted by the presence of additive noise and reverberation in the training set and that this loss could not be recovered by adding more data.

Given the sparse, variable and noisy nature of the speech overheard by children, this simulation suggests that a statistical learning algorithm alone might not be sufficient to account for early phonetic acquisition. Given that linguistic input represents a small fraction of the audio environment of the child, and that even speech is itself overlapped with non-speech signals, any statistical learning algorithm will devote its resources to discovering the structure of the entire audio, thereby failing to capture the structure of speech sounds.

The three types of inductive biases that were introduced in this study are plausible and independently motivated by experimental evidence in infants: infants show an early preference for attending to speech versus non-speech sounds, and it is plausible that they would learn preferentially on such sounds. In addition, there is evidence that infants distinguish speakers and associate speakers to their voices at an early age; it is therefore plausible that their learning algorithm would be speaker-specific. Finally, the human learner has the benefit of an auditory system that has been fine-tuned by millions of years of evolution to accurately perceive sound sources in complex auditory scenes, and it is plausible that learning operates not on raw sensory data, but rather on sensory streams organised according to source and therefore resist additive noise and reverberation. It is important to note, however, that the inductive biases we implemented are not sufficient;

as subsequent testing at the lexical level showed that, even with them, no lexical learning is evidenced in STELA when fed with long-form recordings. This indicates that, as far as phonetic and lexical learning is concerned, some form of poverty of the stimulus argument is valid, and that generic learning algorithms (at least the ones we tested) need to be supplemented with strong inductive biases.

In brief

We showed that realistic learning simulations could help address some of the key controversies within language acquisition. For instance, STELA shows that statistical learning can be sufficient to reproduce some key findings in infants (phonetic attunement, preference for words over nonwords) from raw audio inputs in the total absence of multimodal grounding or social feedback. It also shows that such learning patterns can arise in the total absence of interpretable linguistic categories. However, it also shows that it has to be supplemented with inductive biases in order to deal with the noise present in naturalistic recordings that are representative of what infants really hear. Of course, these findings are only theoretical results: and, as such, can demonstrate that mechanism A is sufficient (or not needed) to observe outcome B. Whether infants really use similar mechanisms remains to be further established.

What lies ahead?

So far, we have presented evidence that learning simulations, when scaled to incorporate realistic inputs and to model more than one linguistic level, can address some long-standing controversies regarding learning mechanisms in infants. However, our demonstration was limited to testing one hypothetical learning mechanism: statistical learning, and a particularly narrow version of it that is restricted to audio inputs. While STELA could perhaps be counted as the first successful learning simulation of early language acquisition in infants when trained on audiobook data, it struggles to learn with ecological data, even with inductive biases. This suggests two directions of future work: (1) improving STELA with more inductive biases; (2) build a model that incorporates other learning mechanisms (e.g., cross-situational learning, social feedback, etc.). Either way, there is work to be done for both the psycholinguistic and AI communities, which we review below.

Guidelines for psycholinguistics and AI communities

Modelling the environment

Concerning the learning environment, we believe that one challenge that lies ahead consists of collecting and characterising more ecological data. As demonstrated above, results are quite different when models are presented with audiobooks or long-form recordings. We foresee that moving towards more naturalistic training sets will increase the impact and relevance of language learning models.

As data is the crux of any language learning simulation, we believe constant efforts must be put in place to collect and share ecological learning environments. On this front, we would like to highlight important initiatives such as the privacy-preserving sharing platforms for long-form audio recordings (VanDam et al., 2016) or video data (Simon,

Gordon, Steiger & Gilmore, 2015), and the DARCLE (DAylong Recordings of Children's Language Environments, DARCLE.org, n.d.) community. We believe these initiatives must become standard practices as they can transform our understanding of language development by enabling incremental and reproducible science and fueling language learning simulations with realistic data.

In addition, most of what we know concerning language development comes from Western, Educated, Industrialised, Rich, and Democratic (WEIRD) populations (Henrich, Heine & Norenzayan, 2010; Scaff, 2019), and this bias toward WEIRD populations reflects in the type of data computational modellers have access to. Current large-scale audio datasets – whether they contain child-centred recordings or audiobooks – are primarily collected in American English (Kearns, 2014; VanDam et al., 2016). We believe this represents a significant limitation for language realistic learning simulations that can – and should – be run considering diverse socioeconomic and cultural backgrounds. Doing so would help us extract and understand universal constants taking place in the course of language development.

Finally, another challenge is to enrich the nature of the data provided to the learner by incorporating ecologically collected multimodal data, in order to address the importance of cross-situational learning in real life. Also, quantifying the nature and prevalence of social feedback (some of which is nonverbal) is very important as a first step towards building interactive models of the learning environment (Tsuji et al., 2021)

Modelling the learner

One key challenge on the learner side relates to the quantity of data needed to reach a certain level of linguistic performance. Today's most performant text-based language models are trained on roughly one thousand times the amount of linguistic input afforded to a typical child (Warstadt & Bowman, 2022). Therefore, current language models are confronted with a data efficiency problem that is doomed to be even more critical when learning from the raw audio, where other sources of variations have to be considered (speaker's identity, speech rate, acoustic conditions, etc.). Future research should focus on implementing algorithms that can reach human-like performances with the same input data available to an infant – that is, that can map the input and the output measures to those of the modelled human.

Related to this question is the challenge of improving perceptual constancy (on the difficulty of obtaining speaker-invariant representations, see van Niekerk, Nortje, Baas & Kamper, 2021) for state-of-the-art learners of audio representations. As stated above, speech sounds, words and sentences can be realised in numerous ways depending on the speaker's identity, the speech rate, or the acoustic environment. This problem is bypassed when considering the text as input, although text brings other simplifying assumptions irrelevant in the context of language acquisition. We believe normalising audio representations along all dimensions irrelevant to language represents one crucial step to bridging the performance gap between audio-based and text-based language models.

Finally, it is important to develop learners that go beyond the statistical learning hypothesis (Erickson & Thiessen, 2015; Romberg & Saffran, 2010; Saffran et al., 1996). Comparing this hypothesis with alternative ones (cross-modal grounding, social constructivism, etc.) will require developing learners with other learning mechanisms to play a more critical role. Reinforcement learning may, for instance, integrate social and interactive rewards, whereas supervised learning may integrate corrective feedback from

caregivers. Admittedly, integrating multiple learning mechanisms and modalities in a single learning simulation requires collaborative work across fields, as has been analysed in Tsuji et al., 2021.

Modelling the outcome measures

The ultimate test of any language learning simulation is the comparison to humans. Dupoux (2018) proposed to aim at cognitive indistinguishability in that setup: "a human and machine are cognitively indistinguishable with respect to a given set of cognitive tests when they yield numerically overlapping results when run on these tests". This critically assumes that cognitive tests that can be applied to the infant and the learner alike are available.

This is not an easy task, and much more can be done in this regard. As discussed above, outcome measures come in several flavours. Laboratory experiments require infants to cooperate with the setting, which is not a given. As a result, the outcome measures are loaded with non-linguistic factors. Infants' performance depends on various factors that most simulations do not currently consider (e.g., memory or fatigue). This problem is even worse when considering babies for which measures are noisier (but see Blandón, Cristia & Räsänen, 2021, who propose evaluations against meta-analyses). This measurement noise needs to be integrated into the outcome model before direct comparisons between infants and simulations can be done. We refer to this problem as the calibration problem. Some outcome measures are more ecological, and extracted directly from the speech of infants. This requires a learner that can also speak, which has not yet been developed. Other measures, like the CDI, depend on the judgement of a caretaker, which here again needs to be modelled specifically. Ultimately, the calibration of measures extracted from the machine to those extracted from the human (or vice versa) will have to be dealt with one measure at a time.

Similarly to HomeBank (VanDam et al., 2016) or Databrary (Simon et al., 2015), we believe both the AI and the psycholinguistics communities would greatly benefit from a privacy-preserving platform to share stimuli – as well as responses – used in psychology experiments. Such a platform would allow researchers to 1) re-use stimuli as new hypotheses arise; 2) revisit stimuli – or responses – to control for confounding factors, or in the context of meta-analytic studies; and 3) create benchmarks that aim at comparing humans and machines. Concerning the last point, we believe there are still too few works that directly compare human and machine performance on a common benchmark (but see Millet & Dunbar, 2020 for a sound discrimination capability study). A stimuli-sharing platform would accelerate collaborative works across the AI and the psycholinguistics community and could also extend to other domains of psychology (including decision-making or social experiments, for instance).

Conclusion

The article's main aim was to provide an extensive description of an emerging theoretical approach in the field of language acquisition: learning simulations, and especially realistic and broad-scope learning simulations. We proposed four criteria we believe are essential for such a simulation to address the current theory crisis and act as a cumulative and unifying theory of language acquisition. We then presented STELA, one such simulation, and showed how it could help shed light on long-standing controversies. Realistic

learning simulations can – and should – integrate the large body of knowledge acquired by the different approaches that comprise the field of language acquisition. Such realistic learning simulations are by no means replacements for other approaches, as all are needed to reach a unified theoretical landscape. Indeed, verbal frameworks can inspire the design of artificial learners, computational models can provide hands-on algorithms, statistical models can exhibit relationships between input and learning outcomes, and corpus studies help describe the characteristics of language environments. Of course, there remain challenges ahead of us to build more complete realistic learning simulations, and we dedicated the last section to address some of them.

Acknowledgements. We are particularly thankful to Dr. Daniel Swingley, Dr. Michael Frank and Dr. Abdellah Fourtassi for helpful insights on previous versions of the manuscript. We are grateful to CoML members for helpful discussions. All errors remain our own. E.D., in his academic role (EHESS), acknowledges funding from Agence Nationale de la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute), a grant from CIFAR (Learning in Machines and Brains), and the HPC resources from GENCI-IDRIS (Grant 2020-AD011011829). M.S. acknowledges PhD funding from Agence de l'Innovation de Défense. M.S and M.L. contributed equally to this work. Authorship order was decided by a coin flip.

Competing interest declaration. The authors declare none.

References

Abend, O., **Kwiatkowski, T.**, **Smith, N. J.**, **Goldwater, S.**, & **Steedman, M.** (2017). Bootstrapping language acquisition. *Cognition*, **164**, 116–143.

Abu-Zhaya, R., **Seidl, A.**, **Tincoff, R.**, & **Cristia, A.** (2017). Building a multimodal lexicon: Lessons from infants' learning of body part words. *GLU 2017 International Workshop on Grounding Language Understanding*, 18–21. https://doi.org/10.21437/GLU.2017-4

Anderson, J. R. (1975). Computer simulation of a language acquisition system: A first report.

Baroni, M. (2020). Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, **375**(1791), 20190307.

Bernard, M., **Thiolliere, R.**, **Saksida, A.**, **Loukatou, G. R.**, **Larsen, E.**, **Johnson, M.**, **Fibla, L.**, **Dupoux, E.**, **Daland, R.**, & **Cao, X. N.** (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Methods*, **52**(1), 264–278.

Blandón, M. A. C., **Cristia, A.**, & **Räsänen, O.** (2021). *Evaluation of computational models of infant language development against robust empirical data from meta-analyses: What, why, and how?* PsyArXiv. https://doi.org/10.31234/osf.io/yjz5a

Bommasani, R., **Hudson, D. A.**, **Adeli, E.**, **Altman, R.**, **Arora, S.**, **von Arx, S.**, **Bernstein, M. S.**, **Bohg, J.**, **Bosselut, A.**, **Brunskill, E.**, **Brynjolfsson, E.**, **Buch, S.**, **Card, D.**, **Castellon, R.**, **Chatterji, N.**, **Chen, A.**, **Creel, K.**, **Davis, J. Q.**, **Demszky, D.**, … **Liang, P.** (2021). *On the opportunities and risks of foundation models.* ArXiv Preprint ArXiv:2108.07258

Borsos, Z., **Marinier, R.**, **Vincent, D.**, **Kharitonov, E.**, **Pietquin, O.**, **Sharifi, M.**, **Teboul, O.**, **Grangier, D.**, **Tagliasacchi, M.**, & **Zeghidour, N.** (2022). *AudioLM: A language modeling approach to audio generation.* ArXiv Preprint ArXiv:2209.03143.

Brent, M. R. (1996). Advances in the computational study of language acquisition. *Cognition*, **61**(1-2), 1–38.

Brent, M. R. (Ed.). (1997). *Computational approaches to language acquisition*. Cambridge, MA: MIT Press.

Brent, M. R. (1999). Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive Sciences*, **3**(8), 294–301.

Brown, T. B., **Mann, B.**, **Ryder, N.**, **Subbiah, M.**, **Kaplan, J.**, **Dhariwal, P.**, **Neelakantan, A.**, **Shyam, P.**, **Sastry, G.**, **Askell, A.**, **Agarwal, S.**, **Herbert-Voss, A.**, **Krueger, G.**, **Henighan, T.**, **Child, R.**, **Ramesh, A.**, **Ziegler, D. M.**, **Wu, J.**, **Winter, C.**, **Hesse, C.**, **Chen, M.**, **Sigler, E.**, **Litwin, M.**, **Gray, S.**, **Chess, B.**, **Clark, J.**, **Berner, C.**, **McCandlish, S.**, **Radford, A.**, **Sutskever, I.**, & **Amodei, D.** (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33**, 1877–1901.

Bulf, H., **Johnson, S. P.**, & **Valenza, E.** (2011). Visual statistical learning in the newborn infant. *Cognition*, **121**(1), 127–132. https://doi.org/10.1016/j.cognition.2011.06.010

Chomsky, N. (1980). *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Chomsky, N. (2013). 4. A review of BF Skinner's verbal behavior. *Volume I Readings in Philosophy of Psychology, Volume* I, 48–64.

Cristia, A. (2022). A systematic review suggests marked differences in the prevalence of infant-directed vocalization across groups of populations. *Developmental Science*, e13265.

Cristia, A., **Dupoux, E.**, **Gurven, M.**, & **Stieglitz, J.** (2019). Child-directed speech is infrequent in a forager-farmer population: A time allocation study. *Child Development*, **90**(3), 759–773.

DARCLE.org. (n.d.). Retrieved 9 September 2022, from https://darcle.org/.

de Marcken, C. (1996). *Unsupervised language acquisition*. Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

de Seyssel, M., **Lavechin, M.**, **Adi, Y.**, **Dupoux, E.**, & **Wisniewski, G.** (2022). *Probing phoneme, language and speaker information in unsupervised speech representations*. In *Proc. Interspeech 2022*, doi:10.21437/Interspeech.2022-373.

Dunbar, E., **Bernard, M.**, **Hamilakis, N.**, **Nguyen, T. A.**, **De Seyssel, M.**, **Rozé, P.**, **Rivière, M.**, **Kharitonov, E.**, & **Dupoux, E.** (2021). *The zero resource speech challenge 2021: Spoken language modelling*. In *Proc. Interspeech 2021*, 1574-1578, doi: 10.21437/Interspeech.2021-1755.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, **173**, 43–59. https://doi.org/10.1016/j.cognition.2017.11.008

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, **14**(2), 179–211. https://doi.org/10.1016/0364-0213(90)90002-E

Elman, J. L., **Bates, E. A.**, & **Johnson, M. H.** (1996). *Rethinking innateness: A connectionist perspective on development* (Vol. **10**). Cambridge, MA: MIT press.

Elsner, M., **Goldwater, S.**, & **Eisenstein, J.** (2012, July). Bootstrapping a unified model of lexical and phonetic acquisition. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 184–193).

Erickson, L. C., & **Thiessen, E. D.** (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, **37**, 66–108. https://doi.org/10.1016/j.dr.2015.05.002

Feldman, N. H., **Goldwater, S.**, **Dupoux, E.**, & **Schatz, T.** (2022). Do infants really learn phonetic categories? *Open Mind*, **5**, 113–131.

Fenson, L. (2007). *MacArthur-Bates communicative development inventories*. Baltimore, MD: Brookes.

Fernald, A., **Marchman, V. A.**, & **Weisleder, A.** (2013). SES differences in language processing skill and vocabulary are evident at 18 months. *Developmental science*, **16**(2), 234–248.

Fodor, J. A., & **Pylyshyn, Z. W.** (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, **28**(1–2), 3–71.

Frank, M. C. (2011). Computational models of early language acquisition. *Current Opinion in Neurobiology*, **21**(3), 381–386.

Frank, M. C. (2014). "Psychological plausibility" considered harmful. *Babies learning language*. http://babieslearninglanguage.blogspot.com/2014/02/psychological-plausibility-considered.html

Frank, M. C., **Goodman, N. D.**, & **Tenenbaum, J. B.** (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological science*, **20**(5), 578–585.

Frank, M. C., **Bergelson, E.**, **Bergmann, C.**, **Cristia, A.**, **Floccia, C.**, **Gervain, J.**, **Hamlin, J. K.**, **Hannon, E. E.**, **Kline, M.**, & **Levelt, C.** (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, **22**(4), 421–435.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, **31**(4), 271–288.

Gleitman, L. R., **Newport, E. L.**, & **Gleitman, H.** (1984). The current status of the motherese hypothesis. *Journal of Child Language*, **11**(1), 43–79. https://doi.org/10.1017/S0305000900005584

Gold, E. M. (1967). Language identification in the limit. *Information and Control*, **10**(5), 447–474.

Hart, B., & **Risley, T. R.** (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.

Henrich, J., **Heine, S. J.**, & **Norenzayan, A.** (2010). Most people are not WEIRD. *Nature*, **466**(7302), 29–29.

Jain, S., **Osherson, D.**, **Royer, J. S.**, & **Sharma, A.** (1999). *Systems that learn: An introduction to learning theory*. Cambridge, MA: MIT press.

Jusczyk, P. W. (1993). From general to language-specific capacities: The WRAPSA model of how speech perception develops. *Journal of Phonetics*, **21**(1-2), 3–28.

Kachergis, G., **Marchman, V. A.**, & **Frank, M. C.** (2021). Toward a "standard model" of early language learning. *Current Directions in Psychological Science*, **31**, 20–27.

Karmiloff-Smith, B. A. (1994). Beyond modularity: A developmental perspective on cognitive science. *European Journal of Disorders of Communication*, **29**(1), 95–105.

Kearns, J. (2014). Librivox: Free public domain audiobooks. *Reference Reviews*.

Kelley, H. H. (1967). Attribution theory in social psychology. In: D. Levine (Ed.) *Nebraska symposium on motivation* (Vol. **15**, pp. 192–240). Lincoln: University of Nebraska.

Kuhl, P. K., **Tsao, F.-M.**, & **Liu, H.-M.** (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences*, **100** (15), 9096–9101. https://doi.org/10.1073/pnas.1532872100

Kuhl, P. K., **Conboy, B. T.**, **Coffey-Corina, S.**, **Padden, D.**, **Rivera-Gaxiola, M.**, & **Nelson, T.** (2008). Phonetic learning as a pathway to language: New data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**(1493), 979–1000.

Lakhotia, K., **Kharitonov, E.**, **Hsu, W.-N.**, **Adi, Y.**, **Polyak, A.**, **Bolte, B.**, **Nguyen, T.-A.**, **Copet, J.**, **Baevski, A.**, **Mohamed, A.**, & **Dupoux, E.** (2021). Generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, **9**, 1336–1354.

Lavechin, M., **de Seyssel, M.**, **Gautheron, L.**, **Dupoux, E.**, & **Cristia, A.** (2022a). Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics*. https://doi.org/10.1146/annurev-linguistics-031120-122120

Lavechin, M., **de Seyssel, M.**, **Métais, M.**, **Metze, F.**, **Mohamed, A.**, **Bredin, H.**, **Dupoux, E.**, & **Cristia, A.** (2022b). *Statistical learning models of early phonetic acquisition struggle with child-centered audio data*. PsyArXiv. https://doi.org/10.31234/osf.io/5tmgy

Lavechin, M., **de Seyssel, M.**, **Titeux, H.**, **Bredin, H.**, **Wisniewski, G.**, **Peperkamp, S.**, **Cristia, A.**, & **Dupoux, E.** (2022c). *Can statistical learning bootstrap early language acquisition? A modeling investigation*. PsyArxiv. https://doi.org/10.31234/osf.io/rx94d

Leibo, J. Z., **d'Autume, C. de M.**, **Zoran, D.**, **Amos, D.**, **Beattie, C.**, **Anderson, K.**, **Castañeda, A. G.**, **Sanchez, M.**, **Green, S.**, & **Gruslys, A.** (2018). *Psychlab: A psychology laboratory for deep reinforcement learning agents*. ArXiv Preprint ArXiv:1801.08116.

Liu, X., **He, P.**, **Chen, W.**, & **Gao, J.** (2019). *Improving multi-task deep neural networks via knowledge distillation for natural language understanding*. ArXiv Preprint ArXiv:1904.09482.

MacWhinney, B., & **MacWhinney, B.** (1987). The competition model. *Mechanisms of language acquisition*, 249–308. London, United Kingdom: Routledge.

McMurray, B. (2021). Categorical perception: Lessons from an enduring myth. *The Journal of the Acoustical Society of America*, **149**(4), A33–A33.

McPhetres, J., **Albayrak-Aydemir, N.**, **Mendes, A. B.**, **Chow, E. C.**, **Gonzalez-Marquez, P.**, **Loukras, E.**, **Maus, A.**, **O'Mahony, A.**, **Pomareda, C.**, & **Primbs, M. A.** (2021). A decade of theory as reflected in psychological science (2009–2019). *PloS One*, **16**(3), e0247986.

Meltzoff, A. N., **Kuhl, P. K.**, **Movellan, J.**, & **Sejnowski, T. J.** (2009). Foundations for a new science of learning. *Science*, **325**(5938), 284–288.

Miller, J. F., & **Chapman, R. S.** (1981). The relation between age and mean length of utterance in morphemes. *Journal of Speech, Language, and Hearing Research*, **24**(2), 154–161.

Millet, J., & **Dunbar, E.** (2020). The perceptimatic English benchmark for speech perception models. *CogSci 2020-42nd Annual Virtual Meeting of the Cognitive Science Society*.

Muthukrishna, M., & **Henrich, J.** (2019). A problem in theory. *Nature Human Behaviour*, **3**(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nelson, D. G. K., **Jusczyk, P. W.**, **Mandel, D. R.**, **Myers, J.**, **Turk, A.**, & **Gerken, L.** (1995). The head-turn preference procedure for testing auditory perception. *Infant Behavior and Development*, **18**(1), 111–116.

Nelson, K. (2007). *Young minds in social worlds: Experience, meaning, and memory*. Cambridge, MA: Harvard University Press.

Newman, R. S., **Rowe, M. L.**, & **Ratner, N. B.** (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language*, **43**(5), 1158–1173. https://doi.org/10.1017/S0305000915000446

Nguyen, T. A., **de Seyssel, M.**, **Rozé, P.**, **Rivière, M.**, **Kharitonov, E.**, **Baevski, A.**, **Dunbar, E.**, & **Dupoux, E.** (2020). *The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling.* ArXiv Preprint ArXiv:2011.11588.

Nguyen, T. A., **Sagot, B.**, & **Dupoux, E.** (2022). Are discrete units necessary for spoken language modeling? *IEEE Journal of Selected Topics in Signal Processing*, **16**(6), 1415–1423.

Pearl, L., & **Sprouse, J.** (2013). Syntactic islands and learning biases: Combining experimental syntax and computational modeling to investigate the language acquisition problem. *Language Acquisition*, **20**(1), 23–68.

Pelucchi, B., **Hay, J. F.**, & **Saffran, J. R.** (2009). Statistical learning in a natural language by 8-month-old infants. *Child Development*, **80**(3), 674–685.

Pinker, S. (1979). Formal models of language learning. *Cognition*, **7**(3), 217–283.

Pinker, S. (1994). *The language instinct: How the mind creates language*. New York, NY: Harper Collins.

Räsänen, O., & **Khorrami, K.** (2019). *A computational model of early language acquisition from audiovisual experiences of young infants.* ArXiv Preprint ArXiv:1906.09832.

Romberg, A. R., & **Saffran, J. R.** (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, **1**(6), 906–914.

Roy, D. K., & **Pentland, A. P.** (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, **26**(1), 113–146.

Rumelhart, D., **McClelland, J.**, & **MacWhinney, B.** (1987). *Mechanisms of language acquisition*. In B. MacWhinney (Ed.), (pp. 195–248). Erlbaum Hillsdale, NJ.

Saffran, J. R., **Aslin, R. N.**, & **Newport, E. L.** (1996). Statistical learning by 8-month-old infants. *Science*, **274**, 1926–1928. https://doi.org/10/fcqz9d

Saffran, J. R., & **Kirkham, N. Z.** (2018). Infant statistical learning. *Annual Review of Psychology*, **69**, 181–203.

Scaff, C. (2019). *Beyond WEIRD: An interdisciplinary approach to language acquisition* [PhD Thesis]. PhD thesis.

Schatz, T., **Peddinti, V.**, **Bach, F.**, **Jansen, A.**, **Hermansky, H.**, & **Dupoux, E.** (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 1781–1785.

Schatz, T., **Feldman, N. H.**, **Goldwater, S.**, **Cao, X.-N.**, & **Dupoux, E.** (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, **118**(7), e2001844118.

Seidl, A., **Tincoff, R.**, **Baker, C.**, & **Cristia, A.** (2015). Why the body comes first: Effects of experimenter touch on infants' word finding. *Developmental Science*, **18**(1), 155–164. https://doi.org/10.1111/desc.12182

Sicherman, A., & **Adi, Y.** (2023). *Analysing discrete self supervised speech representation for spoken language modeling.* ArXiv Preprint ArXiv:2301.00591.

Simon, D. A., **Gordon, A. S.**, **Steiger, L.**, & **Gilmore, R. O.** (2015). Databrary: Enabling sharing and reuse of research video. *Proceedings of the 15th Acm/Ieee-Cs Joint Conference on Digital Libraries*, 279–280.

Skinner, B. F. (1957). *Verbal behavior* (pp. xi, **478**). Acton, MA: Copley Publishing Group

Smith, L., & **Yu, C.** (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, **106**(3), 1558–1568.

Snow, C. E. (1989). Understanding social interaction and language acquisition; sentences are not enough. In *Interaction in Human Development* (pp. 83–103). Lawrence Erlbaum Associates, Inc.

Suanda, S. H., **Mugwanya, N.**, & **Namy, L. L.** (2014). Cross-situational statistical word learning in young children. *Journal of Experimental Child Psychology*, **126**, 395–411.

Swingley, D., & **Humphrey, C.** (2018). Quantitative linguistic predictors of infants' learning of specific English words. *Child Development*, **89**(4), 1247–1267. https://doi.org/10.1111/cdev.12731

Tesar, B., & **Smolensky, P.** (2000). *Learnability in optimality theory*. Cambridge, MA: Mit Press.

Tomasello, M. (2003). The key is social cognition. *Language in mind: Advances in the study of language and thought*, pp47–57. Cambridge, MA: MIT press.

Tomasello, M. (2005). *Constructing a language: A usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tsuji, S., **Cristia, A.**, & **Dupoux, E.** (2021). SCALa: A blueprint for computational models of language acquisition in social context. *Cognition*, **213**, 104779. https://doi.org/10.1016/j.cognition.2021.104779

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, **59**(236), 433.

Vallabha, G. K., **McClelland, J. L.**, **Pons, F.**, **Werker, J. F.**, & **Amano, S.** (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, **104**(33), 13273–13278.

VanDam, M., **Warlaumont, A. S.**, **Bergelson, E.**, **Cristia, A.**, **Soderstrom, M.**, **De Palma, P.**, & **MacWhinney, B.** (2016). HomeBank: An online repository of daylong child-centered audio recordings. *Seminars in Speech and Language*, **37**(02), 128–142.

Van Niekerk, B., **Nortje, L.**, **Baas, M.**, & **Kamper, H.** (2021). *Analyzing speaker information in self-supervised models to improve zero-resource speech processing.* ArXiv Preprint ArXiv:2108.00917.

Warstadt, A., & **Bowman, S. R.** (2022). *What artificial neural networks can tell us about human language acquisition*. ArXiv Preprint ArXiv:2208.07998.

Werker, J., & **Curtin, S.** (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, **1**(2), 197–234. https://doi.org/10.1207/s15473341lld0102_4

Yu, C., **Ballard, D. H.**, & **Aslin, R. N.** (2005). The role of embodied intention in early lexical acquisition. *Cognitive Science*, **29**(6), 961–1005. https://doi.org/10.1207/s15516709cog0000_40

Yu, C., & **Ballard, D. H.** (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, **70**(13–15), 2149–2165.

Yu, C., & **Smith, L. B.** (2017). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*, **41**, 5–31.

Zhang, Y., **Chen, C.**, & **Yu, C.** (2019). Mechanisms of cross-situational learning: Behavioral and computational evidence. *Advances in Child Development and Behavior*, **56**, 37–63.

Appendix A: How to derive a probability from a Language Model?

Head-turn preference experiments (Nelson et al., 1995) provide a wealth of results regarding the type of stimuli infants prefer to listen to. However, mechanisms underlying this preference remain unobservable. Computational modelling provides complementary information by assessing hypotheses about HOW statistical information might be used to exhibit similar preference patterns as those exhibited by infants, or *what* underlying information processing problem is being solved. Language models, and probabilistic models in general, offer a natural way to extract a preference measure from an artificial learner: a stimulus A is preferred to a stimulus B if A is more probable than B.

But how does one compute the probability of a stimulus from a Language Model? First, the waveform goes through the Acoustic Model which returns a discrete representation of the audio: $q_1, q_2 \ldots, q_T$. Then, the Language Model, which has been trained to predict the next discrete unit of a sequence given its past context, assigns a probability to the discrete sequence using the following chain-rule:

$$P(q_1, \ldots, q_T) = \prod_{t=1}^{T} P(q_t | q_1, \ldots, q_{t-1})$$

We compute the logarithm of the resulting probability which has the effect of increasing the difference between probabilities assigned to a minimal pair of stimuli (e. g., a word and a non-word that differ in a single phoneme). The logarithm is then normalised by the length of the input stimuli to enforce the model to not show a constant preference for the longest stimuli.

Appendix B: Overview of the learner used in STELA

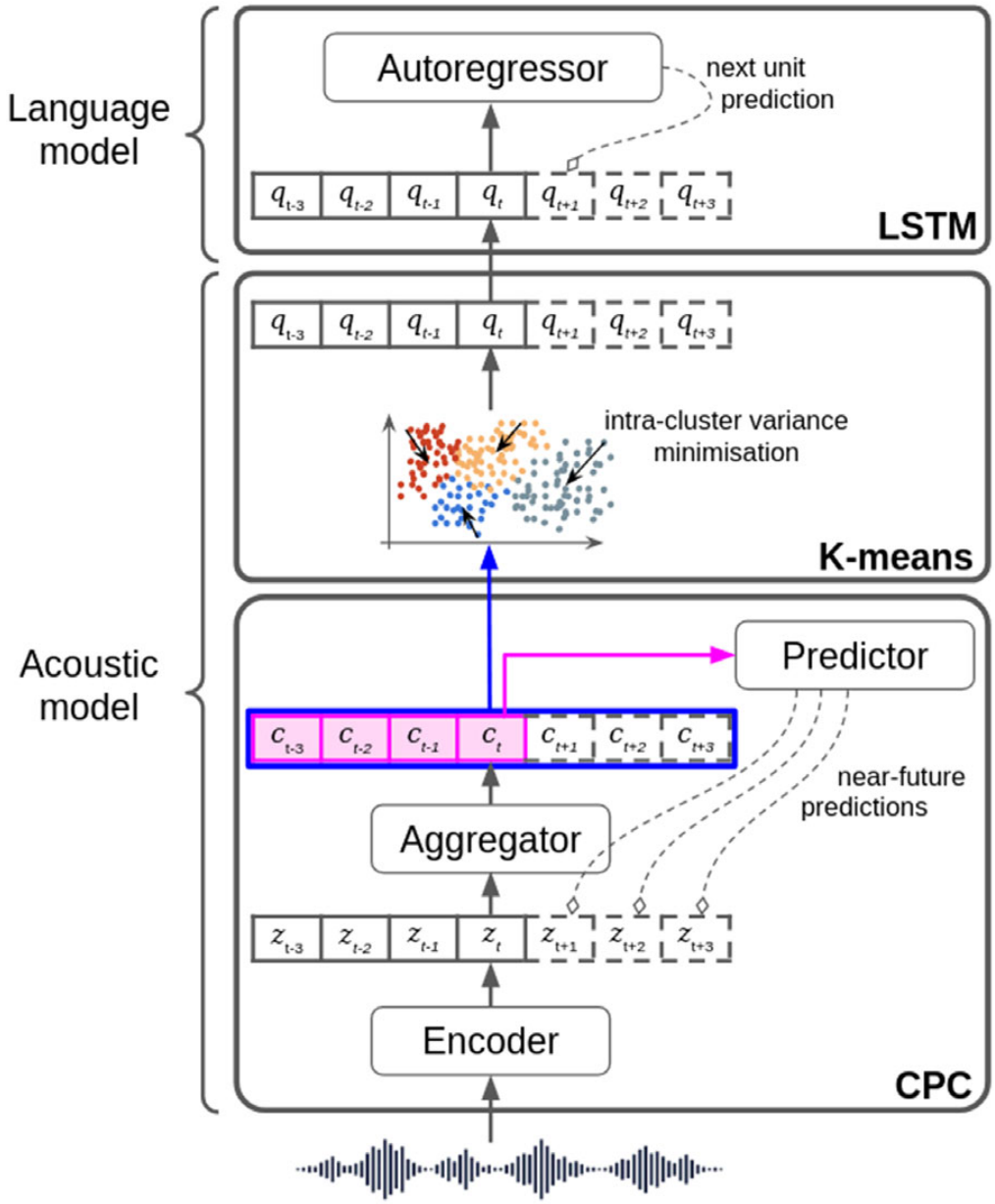


Figure A1. Model of the learner used in STELA. The Acoustic model is composed of a convolutional encoder which delivers a vector of continuous values z_t every 10ms. This is sent to a recurrent network aggregator that integrates context and delivers vectors with the same time step. Contrastive Predictive Coding is trained to predict the outputs of the encoder in the near-future (up to 120 ms). The output of the aggregator is sent to a K-means algorithm that discretise the continuous representations c_t into q_t. Then, a language model (long-short term memory (LSTM) network) is trained to predict the next q_t unit based on past ones.

Cite this article: de Seyssel M., Lavechin M., & Dupoux E. (2023). Realistic and broad-scope learning simulations: first results and challenges. *Journal of Child Language* 1–24, https://doi.org/10.1017/S0305000923000272

3.5 Discussion and conclusion of Chapter 3

This chapter introduced a learning simulation designed to simulate early language acquisition based on the STELA learner. One of the key features of this simulation is its ability to create what we refer to as "developmental" curves for different outcome metrics, which can be compared with human development. Moreover, we used English and French input and outcome measures, making it possible to analyse the nativeness effect by reusing the symmetrical testing concept introduced in the previous chapter.

Initially, our framework comprised two outcome measures that evaluated linguistic knowledge at the phonetic and lexical levels. In Section 3.3, we introduced a new metric, ProsAudit, which captures the knowledge of structural prosody. Although ProsAudit currently only exists in English, it provides valuable insights into the role of prosody in early language acquisition.

Finally, in the final section of this chapter, we discussed the potential applications of similar learning simulations.

3.5.1 Additional metrics

Section 3.4 extensively discusses phonetic and lexical outcome measures presented in Section 3.2. However, additional metrics such as the prosodic one presented in 3.3 offer similar perspectives and limitations. As part of the work done for this thesis, we also developed two new metrics that evaluate semantic and syntactic linguistic information for models of speech. These metrics were used in the 2021 edition of the Zero-Resource Speech challenge and are discussed in detail in a paper we co-authored:

Nguyen, T. A.*, **de Seyssel, M.***, Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., & Dupoux, E. (2020). The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modelling. In *Neurips Workshop on Self-Supervised Learning for Speech and Audio Processing*

Given that the linguistic levels assessed by these metrics typically emerge later in the language acquisition process, we have chosen not to include them in our evaluation of the STELA framework, which concentrates on early language acquisition. However, we provide below a brief overview of these two metrics. Please refer to the paper in Appendix D for a more comprehensive presentation of these metrics. All metrics presented in this chapter are summarised in Table 3.1.

3.5.1.1 Syntax: sBLiMP

The syntactic evaluation task is adapted from the Benchmark of Linguistic Minimal Pairs (BLiMP), targeted at evaluating syntactic knowledge in text models (Warstadt et al., 2020). In the original task, the test is composed of a high number of sentence pairs, with one grammatically correct (e.g., *the child is waiting*') and the other incorrect (e.g., *the child are waiting*') and covers multiple major grammatical phenomena. To use BLiMP for evaluating speech models, we created the

Linguistic level	Metrics	Dataset	Task	Example
structural prosody	acceptabil judge-ment	ProsAudit	$p(a) > p(b)$?	(hello <**B**> world, he<**B**>llo world) (the <**B**> girl eats her peas, the girl eats <**B**> her peas)
acoustic-phonetic	ABX	Libri-light	$d(a,x) < d(b,x)$? $a \in A, b \in B$, $x \neq a \in A$	within-speaker: $(\text{apa}_{s_1}, \text{aba}_{s_1}, \text{apa}_{s_1})$ across-speaker: $(\text{apa}_{s_1}, \text{aba}_{s_1}, \text{apa}_{s_2})$
lexicon	spot-the-word	sWuggy	$p(a) > p(b)$?	(brick, blick) (squalled, squilled)
lexical semantics	similarity judge-ment	sSIMI	$d(a,b) \propto d_h(a,b)$?	(abduct, kidnap): 8.63 (abduct, tap): 0.5
syntax	acceptabil judg-ment	sBLIMP	$p(a) > p(b)$?	(dogs eat meat, dogs eats meat) (the boy can't help himself, the boy can't help herself)

Table 3.1: Summary description of the different zero-shot metrics used in the evaluation of speech models. The metrics highlighted in light blue use a pseudo-distance d between embeddings (with d_h representing human judgements), while those in light orange use a pseudo-probability p computed over the full input sequence.
Note: Adapted from Nguyen, de Seyssel* et al. (2020).*

Synthesised Benchmark of Linguistic Minimal Pairs (sBLiMP). The main step in the adaptation involved synthesising the sentence pairs as two audio utterances to ensure natural prosodic contour and differentiability based on audio only. To create these pairs, we generated new sentences using vocabulary from the LibriSpeech dataset (Kahn et al., 2020). The metric uses a similar approach as the sWuggy and ProsAudit metrics, comparing the probability of generating both sentences from a pair, with a score of 1 assigned if the grammatically correct sentence has a higher score. The final score is the average of all pairs' scores.

3.5.1.2 sSimi

The lexical Semantic task is also adapted from an evaluation for text models, initially introduced in Chung and Glass (2018). The evaluation involves computing the semantic distance between word embeddings that are semantically close or distant and then comparing it to the "human-perceived" semantic distance. As in Chung and Glass (2018), we used pre-existing semantic similarity and relatedness datasets based on human ratings of similarity and relatedness scores. To ensure comparability across datasets, we normalised all scores. After removing words not present in the LibriSpeech dataset, we created two versions of the task. In the "synthesised" version, words were synthesised similarly to the sBLiMP task. In the "natural" version, we directly extracted words from natural speech in the LibriSpeech dataset. Finally, we report Spearman's rank correlation coefficient as the *semantic similarity score*, which reflects the correlation between the model's semantic distance scores and the human scores.

3.5.2 Expanding the applications of STELA: bilingual language acquisition

We have extensively discussed the perspectives, future work, and limitations of STELA in Section 3.4. This previous section serves as the comprehensive discussion for this chapter, so we will not delve deeper into these aspects here.

In the next chapter, our focus will shift to the practical applications of this work. Specifically, we will explore how the learning simulation developed in this chapter can advance research on bilingual language acquisition.

Chapter 4

More than one language: Modelling bilingual language acquisition

4.1 Introduction

In Chapter 3, we showed that our learning simulation developmental framework provides a valuable tool for exploring the complexities of early language acquisition, mainly by providing proof of concepts on specific research questions. While the framework we presented has helped to shed light on some of the fundamental processes involved in general language acquisition, there remains a significant knowledge gap concerning how infants raised in bilingual environments acquire language.

In this chapter, and in line with the guidelines on learning simulation proposed in the previous chapter, we adapt our developmental framework[1] to account for *bilingual input*, with the aim to shed new light on bilingual language acquisition and indexical speech perception. Can a model of language acquisition such as STELA reproduce the language discrimination results modelled with global, i-vector models in Chapter 2? Is this discrimination dependent on the size of the input? Are the same mechanisms used to simulate monolingual language acquisition also able to reproduce bilingual language acquisition developmental patterns? Does bilingual input affect different levels of linguistic speech perception differently? These are the questions that we will attempt to answer in this final chapter. Before we proceed, we will present a brief overview of the recent Self-Supervised Learning (SSL) speech models literature with multilingual input.

4.1.1 Multilingual SSL speech models

Multilingual models were first proposed in the context of low-resource language applications, where the scarcity of data, even unlabelled ones, makes it challenging to construct a robust system in the desired language. A potential solution to

[1]From now on, when we mention our "developmental framework", we refer to the learning simulation presented in Chapter 3 (§3.2.1), including the developmental and cross-linguistic methodology, the STELA learner (that is the combinations of the Acoustic Model - CPC and k-means - and of the Language Model - LSTM) - and the outcome measures.

this problem lies in the technique of cross-lingual model pre-training, an approach first proposed in text language models by Lample and Conneau (2019). This method relies on the capacity of multilingual features to transfer effectively across languages. In the context of speech recognition, for example, cross-linguistic pre-training relates to the initial pre-training of a speech recognition model on multiple languages with greater resource availability. This pretrained model is then fine-tuned on the target language that has a relatively smaller amount of data (Dalmia et al., 2018; Li et al., 2019).

This strategy has recently been applied to self-supervised learning (SSL) setups. SSL models are pretrained on multiple languages, which may also include the target language, potentially leading to linguistic features shared across languages. An example of such a model is XLSR-53 (Conneau et al., 2020), a wav2vec 2.0 model trained on 53 languages. When subsequently fine-tuned on speech recognition models for each specific language, the speech recognition performance was higher on the low-resource target languages than if the models had been trained only on their target language (necessarily with much less data).

Interestingly, the ability to transfer features across languages is not restricted to multilingual input. Rivière et al. (2020) showed that a CPC speech model trained exclusively on English data would reach good phone discrimination scores when tested on other languages on an ABX phoneme discriminability task. More remarkably, these scores remained relatively high even when the model was not fine-tuned on the target languages. This suggests that the learned representations carry some potentially universal information, at least at the phonetic level.

But here comes a "but". While multilingual training does appear to benefit low-resource language models by overcoming the lack of data, it does not help with high-resource languages. When models could have been otherwise trained on a large amount of monolingual input, multilingual training instead seems to have a negative impact, resulting in lower performance (Conneau et al., 2020). This *negative interference*, which was already observed in text-based multilingual models (Wang et al., 2020), appears to result from the presence of language-specific parameters within the multilingual models. Interestingly, it seems that the greater the language distance between the training language(s) and the target language, the higher the interference (de Vries et al., 2022). Similar findings have also been reported in speech models (Jacobs and Kamper, 2021; Nowakowski et al., 2023). Efforts have been made to mitigate this negative interference in multilingual speech models: one approach involves pruning some language-specific parameters (Lu et al., 2022). Additionally, increasing the number of parameters in the model has been shown to help prevent such interference (Babu et al., 2021).

However, these studies have all been carried out on models trained on very large amounts of data and rarely evaluated using zero-shot metrics (the interference is usually observed on downstream tasks). Moreover, for better analyses of such interferences, one would need to compare monolingual and multilingual models trained on exactly the same amount of data while controlling for the input data size, which is challenging to do with very large models requiring intensive computational resources.

4.1.2 Computational modelling in bilingual research

We would also like to briefly elaborate on the role of computational modelling in advancing bilingual research. In fact, numerous models have been proposed to explain and account for language learning in bilingual individuals (see Li and Xu, 2022 and Grosjean and Li, 2013 for reviews). However, most of these models primarily focus on sequential bilingualism rather than simultaneous bilingualism, which is our interest here. Additionally, these models rarely encompass more than one level of linguistic learning, with the majority focusing solely on lexical and semantic aspects (e.g. the Multilink model, Dijkstra et al., 2019). Furthermore, these models seldom leverage the recent advancements in machine learning, particularly deep learning networks. Although there is growing awareness of the importance of incorporating more ecologically valid input and using larger-scale, deep learning-based models in studying bilingualism in psycholinguistics, as called for by Li and Grant (2019). Of course, it is worth noting that some models do employ more advanced technologies, such as the Bilingual Dual-path model (Tsoukala et al., 2017, 2020, 2021), which used a Recurrent Neural Network (RNN) to simulate code-switching production. However, not only is this model supervised, but it also takes text as input instead of raw speech.

In fact, there remains a significant gap to be addressed. Currently, models have yet to be proposed to help in the understanding of early bilingual language acquisition while adhering to the guidelines outlined in the previous chapters of this thesis. For an effective simulation of bilingual language learning, the model must rely solely on raw speech as input, without any additional supervision, and be capable of simultaneously modelling multiple linguistic levels. This is what we will attempt to address in this chapter, using the learning simulation and developmental framework presented in Chapter 3.

4.1.3 Outline of Chapter 4

In the present chapter, we build on the findings from Chapters 2 and 3 by using our developmental framework introduced in the previous chapter to investigate how multilingual speech input influences speech perception in bilingual infants. This amounts to establishing a link between the linguistic and indexical information that was examined separately in the previous two chapters.

We first examine how a learner like STELA, which focuses on segmental and supra-segmental granularities, accounts for indexical information such as speaker and language identity. This type of information is typically captured using models like i-vectors, which operate at a more global level. We also explore how this indexical information is represented as the learner receives more training data. This work is presented in Section 4.3.

Then, in Section 4.4, we use our developmental framework to explore how bilingual input affects the outcome measures of language acquisition. By adding bilingual models within our framework and comparing the developmental curves of language acquisition to those obtained from the monolingual models from Chapter 3, we

aim to gain insights into the challenges associated with learning multiple languages early in life.

4.2 General Methods

This work builds on the methodology and models of our developmental framework, introduced in Chapter 3. We use the same English and French monolingual models as in the original study and introduce novel bilingual (English-French) models. Therefore, the learners and evaluation components of the simulation are identical to those introduced in Chapter 3 unless stated otherwise.

4.2.1 Training the bilingual models

We trained models on bilingual data derived from the monolingual French and English data presented in Section 3.2.1. The training sets were generated by concatenating English and French speech data from the monolingual models' training sets. This is why our smallest bilingual models start at 100 hours, as they result from the concatenation (and subsequent reshuffle) of data from the English and French 50 hours models. That is, we concatenate a French and an English "family" of 50 hours to create a bilingual "family" of 100 hours, and this continuously with all models until we have two bilingual models of 3,200 hours. Because of computational capacity limitations, we do not train a large 6,400 hours bilingual model.

For the different analyses reported in this chapter, we sometimes make the distinction between two ways of visualising the bilingual models, the "Total Matched" and the "Language Matched" models, depending on how they are matched to their monolingual counterparts: in the "Total Matched" variant, we consider the total size of the data; a 100-hours model is comprised of 50 hours of English and 50 hours of French. This is the standard approach, which we refer to if not specified otherwise. We also consider another approach, which we name here "Language Matched" (or "LangMatched") variant, where we match the amount of data per language; a Language Matched 100-hours model corresponds to a model trained on 100 hours of French and 100 hours of English. This second variant allows us to isolate and examine the influence of the quantity of exposed data as the sole predictor of potential costs in bilingual input. Ultimately, the two visualisations are simply shifted versions of each other (the 50-hour Language Matched models correspond to the 100-hour Total Matched models).

Training sets aside, the bilingual models are trained following exactly the same methods and parameters as the monolingual models (refer to the Methods in Section 3.2.1 for more information).

Balanced Language Exposure and OPOL environment The method we employed to generate the training sets results in bilingual models that receive exactly the same amount of speech in each language (50% - 50% exposure pattern). Of course, this scenario is quite different from what is typically observed in

bilingual families, where language input is seldom perfectly balanced between the two languages (Orena et al., 2020). In Section 4.4.4, we will expand on this limitation by examining the impact of varying language exposure on language learning. However, we will focus on the strictly balanced language exposure scenario for the remainder of the chapter. Another consequence of our experimental setup is that we operate within a One Parent One Language (OPOL) environment, where each speaker exclusively speaks one language[2]. While it would have been advantageous to include a condition in which the training set contained speech samples from the same speakers in both languages, as was done in Section 2.2, this proved unfeasible given the available data we possessed. It is a limitation to remember as we proceed with our analyses.

4.2.2 Evaluation

We base all of our analyses on the English and French Common Voice test sets presented in Section 3.2.1[3]. We also evaluate the models with the same outcome measures presented in Chapter 3, namely the Common Voice phone ABX (phonetic), sWuggy (lexical) and ProsAudit (prosodic).

4.3 Languages representation in bilingual language acquisition

In this section, we delve into the representations of speech in monolingual and bilingual models. We pay particular attention to how indexical (language and gender) and linguistic information are represented within these different models.

At this point, it is essential to pause and reflect on the concepts of separation and discrimination in relation to the approach we took in Chapter 2. In Section 2.2, we distinguished between "language discrimination" and "language separation", demonstrating that one does not necessarily entail the other. While the i-vector models could discriminate one language from another when presented with both (language discrimination), they would not produce two distinct clusters for each language when the number of languages was unknown (language separation). Instead, the separation would primarily occur based on speaker identity. Although we were able to analyse this lack of language separation in Section 2.2 relying on a fully bilingual setup, which allowed us to decorrelate speaker and language information completely, we could only model an One Parent One Language set up within the framework used in this chapter. Consequently, we set aside the notion of "language separation" in this chapter in the way it was presented in Section 2.2, focusing instead on the notion of discrimination. To bring the reader additional

[2]As in Chapter 2 (§2.2), this is not to say that we have a single speaker per language, but rather than we do not have any overlap between the speakers of each language.

[3]Additional details on the generation of the test set are presented in Appendix B.

information than raw discrimination scores, as we had done in Section 2.2, we also offer more qualitative analyses in the form of visualisations[4].

First, in §4.3.1, we look at what type of information is represented in the largest monolingual and bilingual models, focusing only on the Acoustic Model component (CPC and k-means) and their phoneme representations. Specifically, we ask whether these representations can suggest some language discrimination ability at the phoneme level. Our investigation here focuses on the largest models. This work is presented in the form of a paper:

de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., & Wisniewski, G. (2022) Probing phoneme, language and speaker information in unsupervised speech representations. In *Proc. Interspeech 2022*, 1402-1406, doi:10.21437/Interspeech.2022-373

In §4.3.2, we extend our analysis to include the Language Model component and investigate the model's ability to distinguish words based on their language. We also provide other quantitative analyses of language discrimination, similar to the ones presented in Chapter 2. More importantly, we explore the effect of development on the ability of the STELA learner to discriminate language (at both the Acoustic and Language model levels), providing insights into the dynamics of language representation in self-supervised models.

4.3.1 Probing phoneme, language and speaker information in unsupervised speech representations

Note: In this published paper, we occasionally use the term "separation". However, our intended meaning is more accurately conveyed by the words "discrimination" or "segregation". We apologise for any confusion caused by this imprecise usage.

Paper summary

In this study, we analyse different types of information encoded in self-supervised speech representations, focusing on the Acoustic Model from the STELA learner. We compare a bilingual model trained on 3,200 hours of speech (1,600 hours of English and 1,600 hours of French) and monolingual French and English models trained on 3,200 hours of speech, initially introduced in Chapter 3 (the models are therefore matched in term of total size, or "Total Matched").

We extracted the Acoustic Models' CPC continuous representations for English and French speech from the Common Voice-based evaluation sets before cutting them along their annotated phonemes boundaries. From these phoneme representations, we explored the presence of language information (French or English), phonetic information (phone class) and speaker information (gender). Our analysis incorporates qualitative methods, using dimension reduction and visual representation, and quantitative approaches, which involve training logistic regression

[4]We sometimes use in this context the term "segregation" to relate to data which seems visually to be discriminable.

models on each category (phone class, gender, and language) to probe the model's knowledge of these different types of information. We also replicate the analysis on the discrete representations as extracted from the additional k-means component of the acoustic model.

Our results show that all models (monolingual and bilingual) encode some phonetic and speaker information. However, only the bilingual model encodes enough language information to potentially allow language discrimination. Moreover, this additional encoding of language information for the bilingual model does not lead to a loss of information at the phonetic or speaker levels, with the bilingual model achieving scores comparable to monolingual models on all logistic regression probes.

However, we also find that there is a cost to training on multiple languages on downstream tasks, as evidenced by lower phoneme discrimination scores on the discrete units in an ABX task, which cannot be compensated by increasing the number of target clusters.

Overall, our results demonstrate that a model trained on two languages can learn enough information to discriminate them (i.e. distinguish between their phonemes in terms of language) without explicit language labels.

It may appear surprising that the phonemes in the visualisation presented in the paper are entirely segregated in terms of language, even though some phonemes are shared between the two languages. However, it is essential to consider that the phoneme representations are extracted from longer utterances containing coarticulation patterns and short-distance information. When taken together, these features can provide valuable insights into which language is spoken. These results hold true regarding cognitive plausibility, as speech is rarely if ever, perceived solely at the phoneme level but rather within spoken utterances.

Probing phoneme, language and speaker information in unsupervised speech representations

Maureen de Seyssel [1,2], *Marvin Lavechin* [1,3], *Yossi Adi* [3], *Emmanuel Dupoux* [1,3], *Guillaume Wisniewski* [2]

[1]Cognitive Machine Learning (ENS–CNRS–EHESS–INRIA–PSL Research University), France
[2]Université de Paris Cité, CNRS, Laboratoire de linguistique formelle, F-75013 Paris, France
[3]Meta AI Research, France

maureen.deseyssel@gmail.com

Abstract

Unsupervised models of representations based on Contrastive Predictive Coding (CPC) [1] are primarily used in spoken language modelling in that they encode phonetic information. In this study, we ask what other types of information are present in CPC speech representations. We focus on three categories: phone class, gender and language, and compare monolingual and bilingual models. Using qualitative and quantitative tools, we find that both gender and phone class information are present in both types of models. Language information, however, is very salient in the bilingual model only, suggesting CPC models learn to discriminate languages when trained on multiple languages. Some language information can also be retrieved from monolingual models, but it is more diffused across all features. These patterns hold when analyses are carried on the discrete units from a downstream clustering model. However, although there is no effect of the number of target clusters on phone class and language information, more gender information is encoded with more clusters. Finally, we find that there is some cost to being exposed to two languages on a downstream phoneme discrimination task.

Index Terms: unsupervised speech representation, self-supervised learning, language representation, probing

1. Introduction

Recent self-supervised models of speech representations capture linguistic features of speech, which can then be used to build language models from raw speech in the context of spoken language modelling [1, 2, 3]. Specifically, it was found that the output representations of such models encode a significant amount of phonetic information, as suggested by the high scores they yield in phoneme discrimination tasks [4, 5, 6, 7]. However, what is less studied is what other types of information are captured by such acoustic models and how they interact.

In this study, we focus on self-supervised models of speech based on Contrastive Predictive Coding (CPC) [1]. Using different probing techniques on their output representations, we want to understand better what information is present and how it is encoded. More specifically, we focus on phonetic, gender and language information. Furthermore, following the growing interest in multilingual representations, we are also interested in how models trained on multiple languages specialise in terms of language information, and for this, we compare models trained on one (monolingual) and two (bilingual) languages.

Being more aware of the types of information present in such speech representations can be of great interest for downstream applications. Depending on the task, we might want to discard some of this information. In the case of language modelling, we can hope for speaker- and gender-invariant representations, as it is irrelevant to some target tasks. For instance, [8, 9, 10] showed that speaker-specific information is present in CPC-based speech representations, while [8] additionally show that removing it can be helpful in lexical, semantic and syntactic downstream spoken language modelling tasks. Yet, other information can prove to be useful. In speech-to-speech translation, for example, multilingual self-supervised models of speech can be used as pretraining to obtain discrete speech units [11]. In this context, information about the language can benefit the downstream translation task. Still, while monolingual CPC-based models have shown to transfer well to other languages [12, 13] in the context of pretraining for Automatic Speech Recognition, no further analyses were done on whether and how language information is present in monolingual and multilingual models.

Approach. More precisely, we focus on three categories of information: phonetic class[1], gender (male and female), and language (English and French), which are all representations that can be learnt directly from raw data. We also compare two types of models: two monolingual ones (trained exclusively on either English or French) and a bilingual one (trained on both English and French). We first probe the CPC speech representations from these different models for all three categories, using qualitative and quantitative measures (t-SNE visualisation and logistic regression classification). Based on past literature, we expect to find phone class and gender-specific information in all models. Whether and how language information is present in the monolingual and bilingual models is more uncertain. Because in most downstream applications, CPC-based models are followed by a clustering step aiming at transforming the continuous output in discrete units, we will also analyse outputs from clustering k-means algorithms. Finally, we will directly look at the differences between monolingual and bilingual models and will test, on a downstream phoneme discrimination task, the impact of having been exposed to two languages.

2. Experimental setting

2.1. Models

We compare models trained on three conditions: a monolingual English set (EN), a monolingual French set (FR) and a bilingual English and French set (EN+FR). Each train set is made of 3,200h of read speech, retrieved from audiobooks (from the

[1]We consider the following phonetic classes: fricative; affricate; plosive; approximant; nasal; nasal vowel; semi-vowel; vowel

10.21437/Interspeech.2022-373

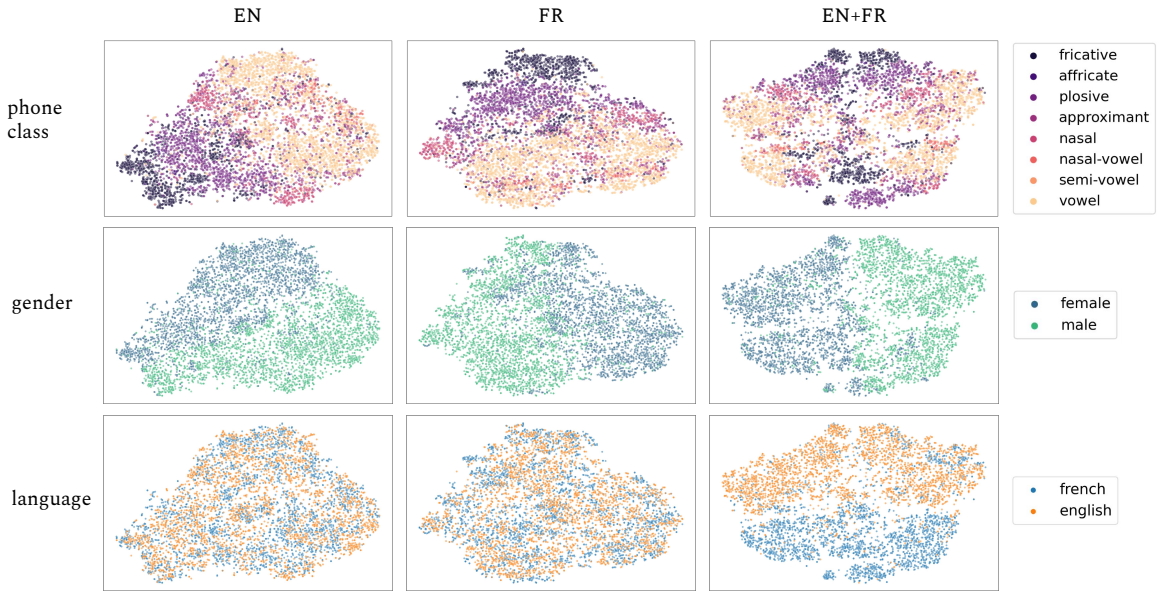


Figure 1: *T-SNE visualisation of English and French phone embeddings at the CPC level, for monolingual (EN and FR) and bilingual (EN+FR) models. Embeddings are colored based on their phone class label, gender label and language label.*

LibriVox and LitteratureAudio projects[2] for English and French respectively) and segmented into short utterances using a Voice Activity Detection Model [14]. All three train sets are matched in terms of number of speakers, genre and utterance duration.

Each train set was used to train an unsupervised acoustic model based on Contrastive Predictive Coding (CPC)[1], using the PyTorch [15] implementation from [13]. During training, this model aims at predicting the near future by selecting the correct frame representation amongst a sample of other negative examples. The architecture and hyperparameters are the same as the CPC-small baseline in [5]. Once trained, the model can be used to predict representations of an audio sequence, made of 256 dimensions feature vectors for every 10ms of audio.

2.2. Evaluation sets

We downloaded American English and Metropolitan French speech from CommonVoice 7.0 [16] to create a 20 hours test set balanced in gender, speakers and languages. We then retrieved for each phone its corresponding audio sequence from the signal, using a phone aligner that we had previously trained using the Kaldi toolkit [17]. This allowed us to have, for each extracted phone, its audio alignment and its corresponding gender label, language label, and phone class label.

3. Results

3.1. Visualising information in the CPC representations

For all three models, we extracted the output CPC speech representations of every aligned phoneme from the test set. Because we have a speech representation for every 10ms of speech, we applied a mean pooling function over the speech representations of every frame of a same phoneme to obtain a single 256-dimension feature vector per phoneme. On a subset of the data (N=6,000), we then applied a t-Distributed Stochastic Neighbor Embedding (t-SNE) algorithm to reduce the dimensions from 256 to 2 dimensions [18], using the Scikit-Learn implementation [19]. We plotted the resulting dimensions and colour-coded the data points based on phone class, gender and language for all three models, as presented in Figure 1.

First, we note that the phone classes are clearly separated for both the monolingual and bilingual models, suggesting that phonetic information is encoded as expected. Besides, the phone classes are also colour-coded based on how sonorous they are (the darkest ones being the least sonorous), and this sonority seems to be encoded as well, as we can see colour gradients in each model. Gender category also seems to be well separated in all three models, as we can see that the two colours are distinct with little overlap. Finally, this qualitative analysis does not show any clear language separation in the monolingual models, whilst there seems to be one in the bilingual model. This can indicate that being exposed to multiple languages leads to encoding some information that can be used to separate languages. More interestingly, visualisation for the bilingual model suggests that language and gender information are clearly distinct and potentially orthogonal to each other.

3.2. Probing the CPC representations

While the t-SNE visualisation allows us to get a qualitative idea of how different categories might be separated, it is still necessary to support it with quantitative measures [20]. Hence, we also use logistic regressions as probes to analyse further what properties are encoded in the speech representations [21]. As in the previous section, we used the phone-level CPC repre-

[2]https://librivox.org and http://litteratureaudio.com/

Table 1: *Logistic Regression Classification error scores (in %), on phone class, gender and language, for the EN, FR and EN+FR models. Number of active features is in italics. Inverse ℓ_1 regularisation strength factor C: $\ell_1 a$: 0.001; $\ell_1 b$: 0.0001.*

	Phone Class			Gender			Language		
	EN	FR	EN+FR	EN	FR	EN+FR	EN	FR	EN+FR
LogReg	**17.6** *(256)*	**18.0** *(256)*	**15.7** *(256)*	**5.8** *(256)*	**6.3** *(256)*	**4.6** *(256)*	**21.7** *(256)*	**20.8** *(256)*	**8.2** *(256)*
LogReg+$\ell_1 a$	24.7 *(75)*	23.7 *(62)*	22.1 *(63)*	8.2 *(17)*	8.8 *(21)*	6.4 *(15)*	27.7 *(53)*	26.6 *(45)*	12.0 *(22)*
LogReg+$\ell_1 b$	40.8 *(8)*	36.4 *(8)*	37.4 *(11)*	10.0 *(2)*	10.4 *(2)*	8.7 *(3)*	49.4 *(0)*	39.7 *(1)*	13.2 *(1)*

sentations, mean-pooled over all frames for each phone. We then split the original test set into an 85/15 train and test set and trained, for each CPC model, three logistic regression models (implementation from Scikit-Learn [19]) using either phone classes, gender or language as labels. Error scores are given in the first row of Table 1. Note that the scores between phone class and the two other categories are not directly comparable due to the differing number of labels for each category and the unbalanced distribution within the phone class category[3].

Low error scores are reached for both phone class and gender in all three models, supporting the previous results that both phonetic and gender information are encoded in both the monolingual and bilingual models. As expected, the language error score for the bilingual model is also very low (8.2%). Surprisingly, the language error scores for the two monolingual models are also relatively low (21.7% and 20.8% for EN and FR, respectively), suggesting that some of the features encoded by the monolingual models can be used to discriminate languages. This result, which was not visible from the t-SNE visualisation, could be due to the fact that the information which is used to reach these low error scores in the monolingual models is diffused within multiple dimensions rather than being strongly encoded in a specific one.

To test this hypothesis, we ran other logistic regression analyses using ℓ_1 regularisation. By doing this, we force the linear model to focus on the most important features, giving less weight to other dimensions. We would expect results with strong ℓ_1 regularisation to be closer to the pattern seen in the visualisation, with lower language scores in the language category for the monolingual models than for the bilingual one. We tested different strengths of ℓ_1 regularisation. Error scores along with the number of active features (features that do not have a coefficient of 0 in the logistic regression) are presented in the two last rows of Table 1. As expected, the language error scores rise dramatically for the monolingual models compared to the bilingual model when we add ℓ_1 regularisation. Figure 2 shows the number of active coefficients in function of accuracy score for the language category. We can see that the monolingual models need to use more features to reach relatively good accuracy scores, compared to the bilingual model, which can already reach 86.8% accuracy with a single feature. As hypothesised, it indicates that the information used to reach good language accuracy scores in the monolingual models is scattered in a number of dimensions rather than specific to a small number of them. For all other categories (and for language in the bilingual model), although the error scores slightly increase, they stay relatively low even with only a very small number of active features, suggesting that there are some specific gender, phone class, and language (for the bilingual model) dimensions. This is also supported by the fact that none of the remaining active features in the model with the strongest ℓ_1 regularisation overlap between the three categories.

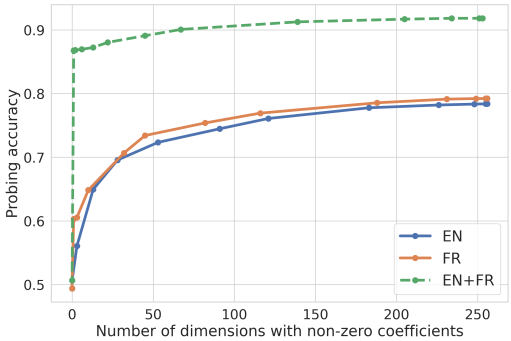


Figure 2: *Probing Accuracy on Language Logistic Regression models wrt. number of active coefficients*

Finally, we can note that the phone class and gender scores for the bilingual model are comparable to those for the monolingual models, suggesting that the additional capture of language information does not prevent the other information to be encoded.

3.3. Analyses of k-means outputs

CPC speech representations are used in the context of spoken language modelling. For this purpose, a clustering step is required to transform the continuous features into discrete units before training a language model [5]. This is why it is also interesting to analyse these discrete units in terms of phone class, gender and language. Therefore, we applied a k-means algorithm with varying number of target clusters (k=50,100,200) on the CPC representations from the EN, FR and EN+FR models. We then retrieved the k-means clusters at each frame of the test set and converted them into one-hot vectors. Finally, for each phone token from the test set, we applied mean-pooling over all its corresponding one-hot vectors to obtain a unique vector of dimension k. Following the method presented in Section 3.2, we trained logistic regression models (without ℓ_1 regularisation) on each of the phone class, gender and language labels. Classification error scores are presented in Table 2.

Error scores go up in all three categories and for all models compared to the error scores on the CPC representations, which is expected as the discrete units cannot encode all the informa-

[3]error score achieved by random labelling on the test set: phone class: 76%; gender: 50%; language: 50%

Table 2: *Logistic Regression Classification error scores (in %), on phone class, gender and language, for the EN, FR and EN+FR models. K50 indicates a k-means model of 50 clusters.*

	Phone Class			Gender			Language		
	EN	FR	EN+FR	EN	FR	EN+FR	EN	FR	EN+FR
CPC	17.6	18.0	15.7	5.8	6.3	4.6	21.7	20.8	8.2
K50	29.3	28.1	30.7	22.6	24.8	22.1	40.7	41.6	25.8
K100	28.1	27.3	28.1	16.8	20.6	15.1	39.1	40.0	25.3
K200	27.1	25.8	25.8	14.4	18.3	12.9	37.2	38.1	24.7

tion from the continuous representations. Still, phone class and gender classifications reach good scores on the k-means units for both the monolingual and bilingual models. In line with our previous analyses, the bilingual model also reaches much lower language error classification scores on the k-means outputs than the monolingual models, with the latter getting closer to the chance level. This suggests that, for the bilingual model only, the units discovered by the clustering can distinguish the two languages.

Effect of number of clusters. There is little variation on the accuracy scores when changing the number of target clusters for the phone class category with monolingual models, in line with previous studies [5]. This is also true for the bilingual model, where the effect of number of clusters is very small (16% error score decrease when considering 200 clusters rather than 50). There does not seem to be any effect of number of clusters on the language classification either. However, there is an effect of the number of clusters on gender classification, and this for all three models, with an average error score decrease of 35% when going from 50 to 200 clusters. This, along with the fact that the logistic regression scores on CPC are lower on gender than on language, suggests that gender is the most present of the three categories within the CPC speech representations. Furthermore, as with the results on CPC representations, the bilingual models show comparable discrimination scores to the monolingual ones in both phone class and gender categories, despite having also encoded language information.

3.4. Comparing monolingual and bilingual models on a phone discrimination downstream task

Our analyses show that models trained on a single language do not encode directly language-specific information to the same extent that models trained on more languages do. We also found that gender and phone information are still present in bilingual models to the same extent than for the monolingual ones. However, we are also interested in whether there is a cost on downstream tasks to multilingual training. For this, we used the phone ABX task [22] to compute discrimination scores for minimal-pairs triplets from the test set. Contrary to the probes used previously, this task is not supervised and allows us to analyse how the representations compare at the phone level without explicitly specifying the relevant features. We tested each model on the language(s) they were trained on (e.g. FR models were tested on French triplets, but EN+FR models were tested on English and French triplets). The task was run on both CPC and Kmeans outputs. Within speaker error scores are presented in Table 3, along with the MFCC baseline scores. While all three models seem to be able to perform phoneme discrimination from their trained language(s) at a similar level on the CPC representations, the lower ABX scores on k-means clusters for the bilingual model compared to monolingual ones suggest that there is indeed a cost to being exposed to more languages. Furthermore, adding more clusters does not seem to help reduce this difference, discarding the hypothesis that the lower results are due to the larger number of phonemes present in the bilingual model. Understanding where this difference comes from would be of great interest in further studies, especially with the rise of multilingual models in spoken language processing.

Table 3: *Within-speaker on phoneme ABX error scores (in %). Models are tested on the same language(s) they were trained on. K50 indicates a k-means model of 50 clusters.*

	MFCC	CPC	K50	K100	K200
EN	17.2	9.86	17.2	18.4	19.0
FR	17.3	11.0	19.5	19.5	19.1
EN+FR	17.3	11.7	25.3	25.1	25.6

4. Discussion & Conclusions

The analyses done on CPC speech representations confirm the fact that both phonetic and speaker (using gender as a proxy) information are present in the output features, replicating what was found in the past [5, 8, 9, 10]. Furthermore, these different types of information are still present after converting the CPC continuous representations into discrete units, using a clustering algorithm, even if at a lesser level. This is an important takeaway as different downstream models might need to work on some representations agnostic to one or the other category, depending on their application.

Comparing monolingual and bilingual models, we also found that models trained on multiple languages encode some language information, which seems to be disentangled from gender information. It is not the case for monolingual models, where the language information is less present and more entangled with other features. Moreover, we found no impact on the quality of phone class and gender features when using bilingual models, with the latter reaching scores comparable to the monolingual models on logistic regression probes. Yet, some of our results suggested there is a cost of being trained on multiple languages on a downstream phoneme discrimination task on the discrete units, which cannot be compensated by augmenting the number of target clusters. With multilingual self-supervised models of speech being proposed as pretraining for a series of downstream applications, more analyses of such multilingual representations should be carried out. Another benefit of further studies on the topic would be to understand which cues from the signal (acoustic, phonotactics, prosodic) carry such language information.

Finally, whilst our analyses were carried out on CPC speech representations, the use of the contrastive loss for unsupervised representation learning goes well beyond specifics of our implementation and has been used in numerous works [23, 24, 3]. Therefore, we think our findings may be applicable to other self-supervised models. In any case, the methodology proposed in this work remains relevant for probing information in any audio representations.

Acknowledgments. MS's work was partly funded by l'Agence de l'Innovation de Défense and performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011012315). ED in his EHESS role was supported in part by the Agence Nationale pour la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute) and a grant from CIFAR (Learning in Machines and Brains).

5. References

[1] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[4] E. Dunbar, X. N. Cao, J. Benjumea, J. Karadayi, M. Bernard, L. Besacier, X. Anguera, and E. Dupoux, "The zero resource speech challenge 2017," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2017, pp. 323–330.

[5] T. A. Nguyen, M. de Seyssel, P. Rozé, M. Rivière, E. Kharitonov, A. Baevski, E. Dunbar, and E. Dupoux, "The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling," *arXiv preprint arXiv:2011.11588*, 2020.

[6] J. Millet and E. Dunbar, "Do self-supervised speech models develop human-like perception biases?" 2022.

[7] J. Chorowski, R. J. Weiss, S. Bengio, and A. Van Den Oord, "Unsupervised speech representation learning using wavenet autoencoders," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 12, pp. 2041–2053, 2019.

[8] B. van Niekerk, L. Nortje, M. Baas, and H. Kamper, "Analyzing speaker information in self-supervised models to improve zero-resource speech processing," *arXiv preprint arXiv:2108.00917*, 2021.

[9] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W.-N. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," *arXiv preprint arXiv:2104.00355*, 2021.

[10] E. Kharitonov, J. Copet, K. Lakhotia, T. A. Nguyen, P. Tomasello, A. Lee, A. Elkahky, W.-N. Hsu, A. Mohamed, E. Dupoux *et al.*, "textless-lib: a library for textless spoken language processing," *arXiv preprint arXiv:2202.07359*, 2022.

[11] A. Lee, H. Gong, P.-A. Duquenne, H. Schwenk, P.-J. Chen, C. Wang, S. Popuri, J. Pino, J. Gu, and W.-N. Hsu, "Textless speech-to-speech translation on real data," *arXiv preprint arXiv:2112.08352*, 2021.

[12] K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. v. d. Oord, "Learning robust and multilingual speech representations," *arXiv preprint arXiv:2001.11128*, 2020.

[13] M. Rivière, A. Joulin, P.-E. Mazaré, and E. Dupoux, "Unsupervised pretraining transfers well across languages," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7414–7418.

[14] M. Lavechin, R. Bousbib, H. Bredin, E. Dupoux, and A. Cristia, "An open-source voice type classifier for child-centered daylong recordings," *arXiv preprint arXiv:2005.12656*, 2020.

[15] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[16] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[17] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. CONF. IEEE Signal Processing Society, 2011.

[18] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[19] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," in *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, 2013, pp. 108–122.

[20] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-sne effectively," *Distill*, 2016. [Online]. Available: http://distill.pub/2016/misread-tsne

[21] G. Alain and Y. Bengio, "Understanding intermediate layers using linear classifier probes," *arXiv preprint arXiv:1610.01644*, 2016.

[22] T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux, "Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline," 2013.

[23] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[24] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *arXiv preprint arXiv:1910.05453*, 2019.

4.3.2 Is there a developmental effect of language discrimination in SSL bilingual models?

In the paper presented previously, we demonstrated that bilingual models can learn to represent phonemes in a way that segregates languages. In this section, we investigate whether this potential language discrimination ability occurs even with very small models or whether there is an effect of the amount of training data. We will examine the three types of representations yielded by our STELA learner (CPC and k-means for the Acoustic Model and LSTM for the Language Model).

We will also carry out two types of analyses. In the first type of analysis, we analyse qualitatively whether the representations yielded by the different models show some discrimination ability in terms of language. In other words, we look for evidence of language-based segregation, similar to the visualisations from Section 4.3.1. We focus on representations at the level of phonemes (Acoustic Model) and words (Language Model).

In the second analysis, we go back to the language discrimination experiments first introduced in Chapter 2(§2.2) and compute language discrimination scores using a language-based ABX task.

4.3.2.1 Acoustic Model

Visualising language discrimination at the phone level We first focus on the representations of phones as output by the CPC models, which yield continuous speech representations. We follow the same methods presented in Section 4.3.1, extracting phonemes representations from the Common Voice test set at the CPC level and applying t-SNE to visualise the representations. We provide visualisations for one bilingual model of each "train size".

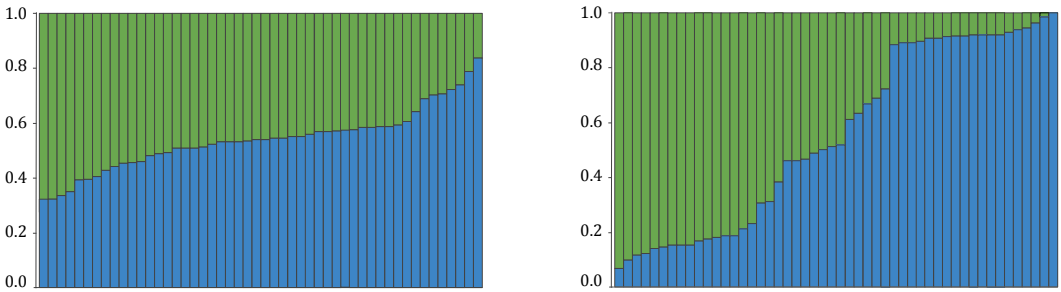
Figure 4.1 displays a 2D visualisation of French and English phoneme representations, with each phoneme colour-coded according to the spoken language. This visualisation represents bilingual models with increasing training sizes, ranging from 100 to 3,200 hours of data. While we already established in the previous section that larger models could discriminate languages based on their phonemes, this does not appear to be true for models trained with less data. Indeed, the ability for language-based discrimination seems to increase with the amount of training data. From the visualisation, it appears that the models begin to produce features allowing for phoneme segregation when trained on a dataset containing between 800 and 1,600 hours of data.

To confirm this finding, we examine the discretised phoneme representations produced by the k-means algorithm. For this purpose, we assess, for each of the 50 clusters generated by k-means, the probability of the cluster to correspond to an English or a French phoneme. That is, we extract for all phonemes from the test set their corresponding cluster as computed by the k-means step. We then calculate for each extracted cluster the probability of being matched to a French or English phoneme. This analysis is shown for a small (100 hours) and a large (3,200 hours) bilingual model in Figure 4.2. For the 100h bilingual model, most clusters are equally likely to be assigned to French or English phonemes. However,

Figure 4.1: T-SNE visualization illustrating the English and French phone CPC representations for the bilingual models.

a contrasting pattern emerges in the bilingual 3,200h model, where some clusters are almost exclusively assigned to either French or English phonemes.

Findings on the CPC (continuous) representations and the k-means (discrete) representations suggest that contrary to monolingual models, bilingual models yield features which allow segregating phonemes in terms of language, but only when trained on enough data.



(a) 100h **(b)** 3,200h

Figure 4.2: Probability for each cluster to correspond to an English (blue) versus a French (green) phoneme, arranged in ascending order of likelihood for English. The left panel showcases the results from the bilingual 100h model, while the right panel presents the findings from the bilingual 3,200h model.

Language discrimination at the utterance level We then focus on ABX-based language discrimination scores, similar to the ones computed on i-vector

representations in Chapter 2 (§2.2). As for the i-vectors, we move on from the phonemes and words representations to a slightly more holistic representation of the utterance by averaging the representations over longer speech snippets[5]. For this purpose, we created a specific subset, sampling utterances from the Common Voice evaluation set so that the distribution of utterances is perfectly balanced between French and English in terms of duration and the number of speakers.We then adapted the ABX task for language discrimination without bilingual speakers: when in Chapter 2, all of our three utterances A, B, and X belonged to the same speaker; here, due to the lack of bilingual speakers in our evaluation set, we force A, B and X to belong to three different speakers, so that the task does not become a speaker-discrimination task. The setup is depicted in Figure 4.3. Because the evaluation dataset is as much as possible matched in terms of utterance duration, speakers and gender between the two languages, we assume that the ABX task should have little bias towards one language over the other.

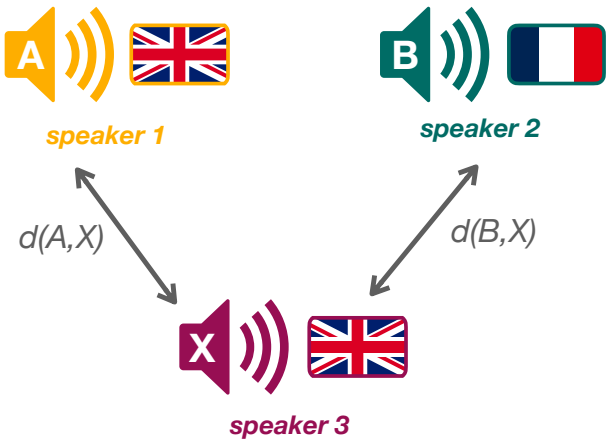


Figure 4.3: ABX setup used for the language discrimination task used in Chapter 4.

We compute the language discrimination error rates on all monolingual and bilingual models ranging from 100 to 3,200 hours at both the CPC and k-means levels. The results are shown in Figure 4.4. Notably, monolingual models consistently perform at chance level, indicating their inability to discriminate between languages (with only minor improvement as data increases). In contrast, bilingual models initially perform at chance level but progressively improve in language discrimination as the volume of training data increases. Specifically, these models begin to achieve satisfactory discrimination levels after 800h, which aligns with the emergence of language discrimination ability observed in our initial analyses. Furthermore, bilingual models demonstrate significantly higher discrimination scores at the k-means level compared to the CPC level, suggesting further that clustering from k-means somewhat follows language information, which is expected since we know that the k-means units already show language-specific information (see Section 4.3.1).

[5]For the computation of the ABX task, we only use up to the first 4 seconds of each utterance, to avoid excessive computation time.

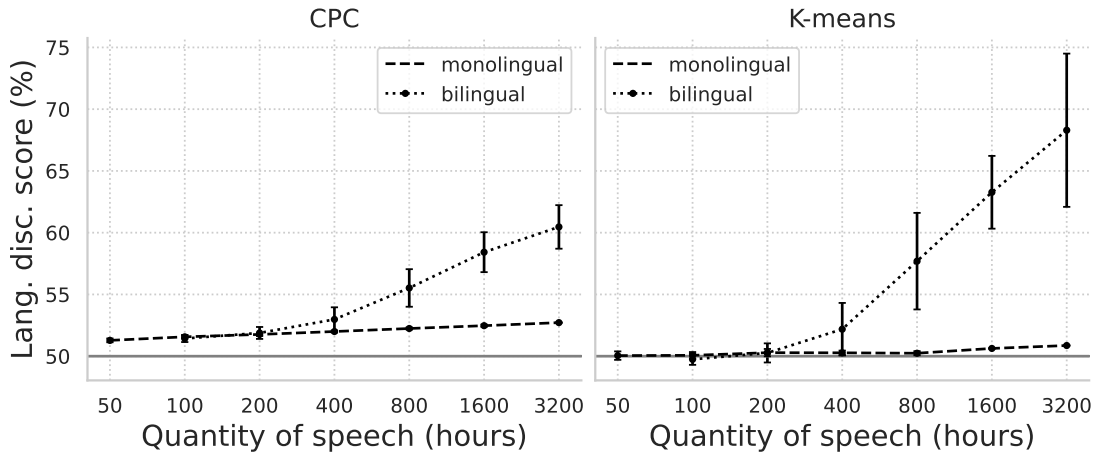


Figure 4.4: Language discrimination scores on the CPC and k-means representations, observed in both monolingual and bilingual models, with regard to the size of the training data.

4.3.2.2 Language Model

Lastly, we turn our focus to how *words* are represented within the STELA learner by extracting word representation from the LSTM. To achieve this, we pooled representations at the word level using a random subset of words from the Common Voice evaluation subset. Besides that, the t-SNE visualisation is done the same way as for the phone representations at the CPC level. The resulting visualisation carried out on a small (100h) and a large (3,200h) bilingual model is presented in Figure 4.5. The results are in line with previous findings on the Acoustic Model. The large bilingual model seems able to separate the word representations based on the two languages (panel b), whilst the small model does not (panel a). Consistently with our previous findings, and although not presented here, the monolingual models always fail to show similar language discrimination based on the word representations[6].

To conclude, we found that bilingual models become more and more able to discriminate speech representations (both at the phone and word level) in terms of language, with only the models trained on large amounts of data suggesting clear language-based discrimination. Moreover, this pattern is reflected within the three components of our model (CPC and k-means for the Acoustic Model and LSTM for the Language Model). From our analysis, it seems that this discrimination starts to occur somewhere between 400 hours and 800 hours of training data, with strong discrimination patterns emerging between 800 hours and 1,600 hours of training data.

[6]We do not conduct the same ABX language discrimination analysis at the Language Model level than the one provided for the Acoustic Models' representations, as the models already demonstrated discrimination at the first two levels; thus, it is reasonable to assume that similar results would be observed at the third level.

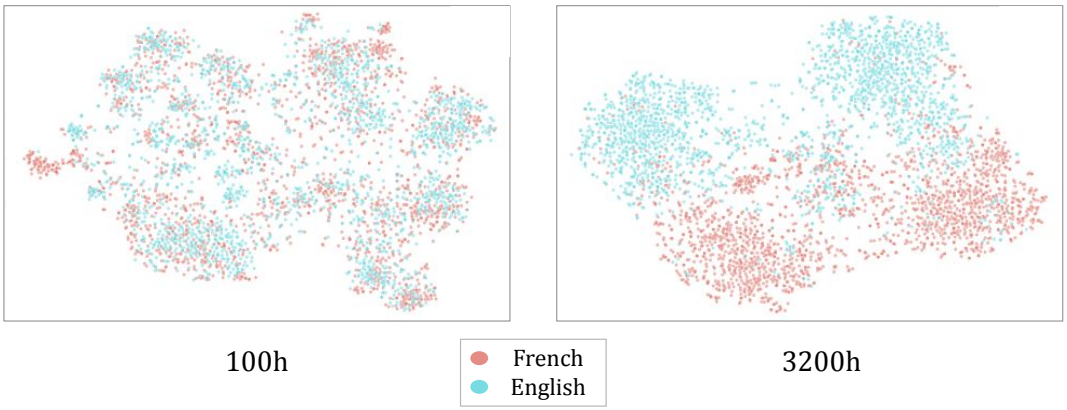


Figure 4.5: T-SNE visualisation presenting the English and French word representations at the LSTM level for both a small (100 hours) and a large (3,200 hours) bilingual model

4.3.2.3 Discussion

In this study, we found that bilingual models learn enough information to exhibit language discrimination, but this is only when enough data is provided (seemingly between 800h and 1,600h). The discrimination is evident and transfers to all levels of our learner: the acoustic model (CPC and k-means) and the language model. The monolingual models, on the other hand, never seem to be able to discriminate between languages, which is not surprising given that they were only presented with one language.

Although large bilingual models exhibit language discrimination, the fact that smaller ones do not, and most importantly, that monolingual ones never do, contradicts our results with the i-vector models and the findings in the infant psycholinguistics literature. Indeed, we know that infants, both raised in monolingual or bilingual environments, can differentiate languages from birth (Mehler et al., 1988; Byers-Heinlein et al., 2010), and the i-vector models presented in Chapter 2 were capable of modelling these patterns even with minimal amounts of training data. A possible explanation for such results may lie in the differences between the training methods of the i-vector models and our STELA learner. First of all, while the i-vector model learns to represent information at the indexical level (capturing the distinctive features of an entire utterance), our present learner focuses on the linguistic level, examining a more fine-grained window of information, which may result in a loss of indexical information. Second, and more significantly, the i-vector model, by constructing a subspace of acoustic variability, learns to capture the variability between the acoustic features (often indexical). As a result, if a novel utterance with very different acoustic features (e.g., a new language) is introduced into the subspace, it will be positioned far from the representations of other utterances the model was trained on. On the other hand, with a model like CPC, the notion of an acoustic subspace is not present, and therefore the features from an utterance within a novel language will be represented as closely as possible

to other features the model was trained on. We will discuss the psycholinguistics implications in Chapter 5.

Now, we know that the fact that the monolingual models presented in the last chapter do not reproduce infants' discrimination abilities does not impede the ability of the model to reproduce language acquisition patterns, as shown in Chapter 3. But do these results hold with bilingual models, where the language indexical information can be fundamental in linguistic learning? Can bilingual models exhibit the same patterns without discriminating languages? If smaller models that do not exhibit such discrimination patterns still undergo linguistic learning, does this imply that language discrimination is not necessary for language learning in bilingual settings? These are some of the questions we will focus on in the next section of this chapter.

4.4 Language acquisition in bilingual learning simulation

In Chapter 3, we presented a developmental framework for modelling language acquisition that enables us to assess linguistic knowledge at different levels using a series of outcome measures. In Section 3.4 specifically, we demonstrated how this framework could enhance our understanding of language acquisition by leveraging the reverse engineering approach proposed by Dupoux (2018), as further discussed in Lavechin et al. (2022). By altering components in the model's environment, learner, and outcomes aspects, we can gain valuable insights into the underlying cognitive processes involved in language acquisition in humans.

In this section, we explore the effects of bilingual input on linguistic representation by applying similar reasoning as in the previous chapters. Specifically, we investigate whether having speech input in two languages would impact the acquisition of language at different levels. We can conduct a modelling study by comparing models trained on bilingual data with those trained on monolingual data using the same metrics presented in Chapter 3. We examine phonetic learning (ABX; Section 4.4.1), lexical learning (sWuggy; Section 4.4.2), and prosodic learning (ProsAudit; Section 4.4.3) levels. We also investigate the role of language exposure proportion in Section 4.4.4. Finally, we discuss our results in Section 4.4.5.

Except if stated otherwise, we compute the different outcome measures in exactly the same way as we did in Section 3.2.1. We also present the results primarily in terms of "Total Matched" curves; that is, if not stated otherwise, the scores presented for the bilingual and monolingual models' comparison correspond to the total amount of data the model was trained on, regardless of the language. Finally, while, when possible, we only present here the results averaged over the two evaluation languages, keeping close to the symmetrical testing approach first presented in Chapter 2, we also present and discuss the phonetic and lexical results separately for each language in Appendix C. Because for prosodic learning, we only have an evaluation set available in English, we leave these results in the present chapter.

4.4.1 Phonetic learning in bilingual models

First, we focus on phonetic learning using the machine phone ABX task. While in Section 3.2.1, we only presented the scores calculated on the k-means representations, that is, the discretised output of the Acoustic Model, in this section, we will also present the scores for the upstream continuous CPC representations. Indeed, although the scores for the patterns in the monolingual models remained identical between the two representations (except for an expected drop in performance when calculated on the discretised units; see Nguyen*, de Seyssel* et al., 2020), differences in patterns in the bilingual models made it necessary for us to report and discuss both representations in this section.

4.4.1.1 Phonetic learning at the CPC level

We use the same ABX task as we used on the monolingual k-means representations, presented in Chapter 3(§3.2.1). However, instead of computing the distance between the one-hot discretised representations, we do it in this section over the CPC continuous representations. Implementation details are the same as for the Zero-Resource Speech baseline and therefore are supplied within the related paper in Appendix D (Nguyen*, de Seyssel* et al., 2020)

	Evaluation	Monolingual		Bilingual
		Native	Non-native	
ABX CPC	Average	+ **0.83%**	+ 0.70%	+ 0.74%
ABX k-means	Average	+ 2.59%	+ 2.12%	+ **2.71%**
ABX k-means (bil k=100)	Average	+ 2.59%	+ 2.12%	+ **2.67%**
sWuggy	Average	+ **3.70%**	+ 0.63%	+ 1.65%
sWuggy (bil k=100)	Average	+ **3.70%**	+ 0.63%	+ 1.97%
ProsAudit (protosyntax)	English	+ 2.24%	+ 1.25%	+ **2.73%**
ProsAudit (lexical)	English	+ 2.98%	+ 0.33%	+ **3.46%**

Table 4.1: Average relative percentage increase observed when the size of the training set is doubled for all phonetic, lexical, and prosodic development curves (equivalent to the slope of the development curves). Both the k-means and sWuggy models were trained using 50 clusters (k=50), except for the instances labelled as 'bil k=100', which indicates that the bilingual models alone were trained using 100 clusters.

Figure 4.6 shows the scores on the ABX metrics for the bilingual and monolingual (native and non-native) models, calculated on the CPC representations. The slopes of the developmental curves, which reflect the average percentage increase when doubling the size of the training set, are reported in Table 4.1. One notable observation is that all models show a progressive improvement in discriminating phonemes with increasing amounts of training data. Additionally, it appears that the discrimination scores of the bilingual models are initially as poor as those of the

non-native models with the smallest amount of data. However, the bilingual curve has a higher slope than the non-native one, indicating a faster rate of improvement, and therefore become increasingly better than the non-native model (the native one, nonetheless, has the steepest developmental curve's slope). However, the bilingual curve never reaches the level of the average monolingual score, which is the average between the native and non-native scores. This suggests a negative interference resulting from bilingual training at this point.

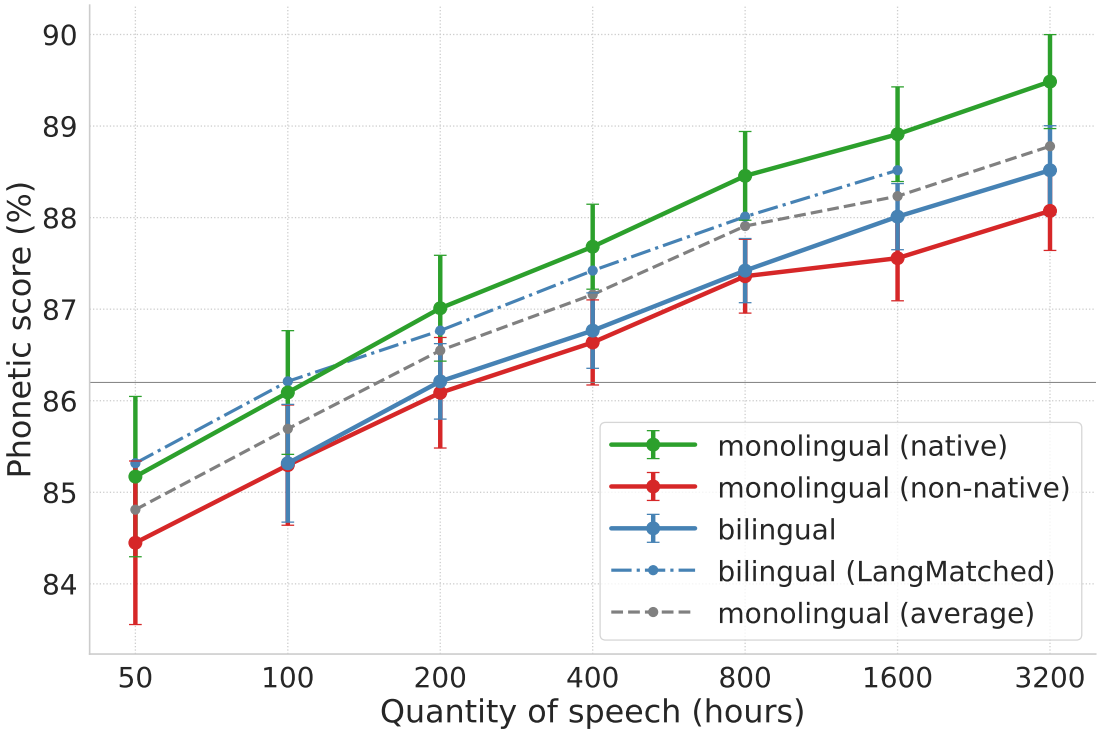


Figure 4.6: Comparison of ABX phonetic scores on monolingual and bilingual CPC representations with respect to the size of the training data. The scores are averaged on the English and French test sets.

When comparing the scores by the total amount of speech heard in each language (as reflected in the bilingual Language Matched curve in Figure 4.6), we observe that the bilingual scores still do not reach the performance of the monolingual ones, despite being closer. That is, a monolingual model trained on x hours of language A is always better than a bilingual model trained with x hours of language A and x hours of language B. This finding suggests that, even when considering models trained on the same amount of language, having exposure to another language still impedes the bilingual models' performance.

4.4.1.2 Phonetic learning at the k-means level

We then proceed to analyse the phonetic scores calculated using the k-means discretised units (as was done in §3.2). From previous work and analyses, we know that these scores are generally lower than those calculated using continuous representations due to the loss of information during the discretisation process

(Nguyen*, de Seyssel* et al., 2020; de Seyssel et al., 2022a). However, the overall nativeness pattern remains similar. The left panel of Figure 4.7 presents the resulting developmental curves when calculating the ABX scores using 50 k-means units (as was the case in §3.2). Surprisingly, there is a significant drop in performance for the bilingual models when compared to the monolingual ones, which fail to reach even the scores of the non-native model, despite having this time a steeper slope than any monolingual model (native or non-native). That is, the bilingual models exhibit a consistently better percentage increase, but their performance starts from a very low point, as reported in Table 4.1 and Figure 4.7.

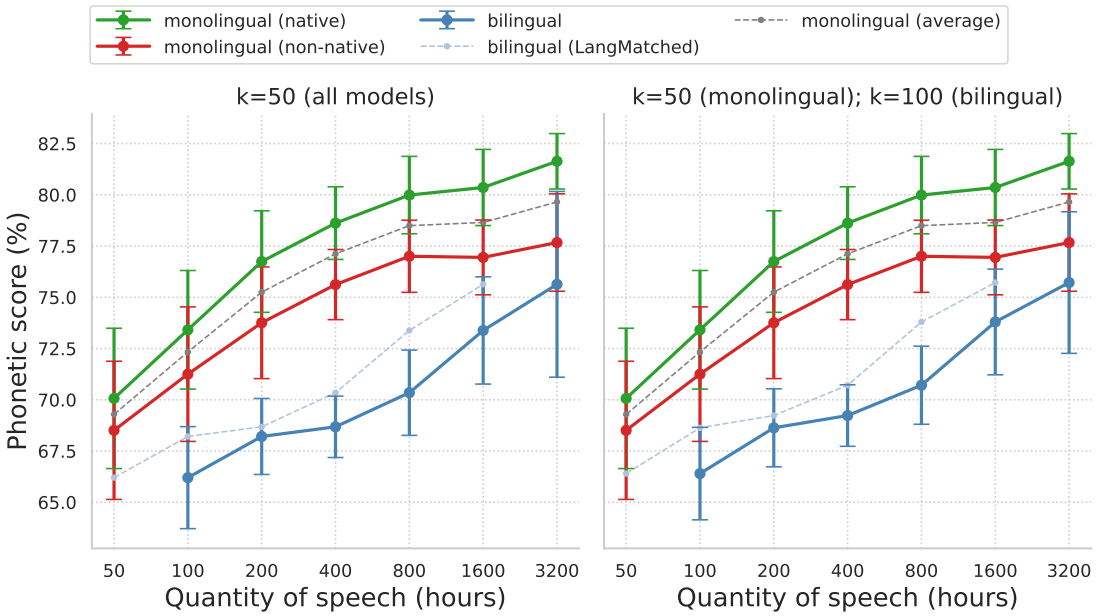


Figure 4.7: Comparison of ABX phonetic scores on monolingual and bilingual (k-means) representations with respect to the size of the training data. The scores are averaged on the English and French test sets.

This result differs significantly from the one obtained using the CPC representations, leading us to suspect that the k-means method might lose too much information relevant to the task during the discretisation process for the bilingual model. A plausible hypothesis is that since more phonemes need to be encoded in the bilingual models, doubling the number of clusters for these models would make sense. However, as shown in the right panel of 4.7, there is almost no difference when shifting to 100 clusters for the bilingual model, which still yields worse results than the non-native models. This result is consistent with our findings in Section 4.3.1, where increasing the number of units led to a better encoding of information, such as speaker details, but did not improve significantly capturing of phoneme information.

4.4.2 Lexical learning - sWuggy

We next move on to analysing the performance of bilingual models on lexical learning, using the sWuggy lexical task with the French and Evaluation test sets

presented in Section 3.2.1. This evaluation is carried out on the language model (LSTM) representations. If not specified otherwise, the LSTM was trained on 50 k-means units. As for the previous section, the monolingual curves presented here are the same as those in Section 3.2.1, with the reported results being calculated on the high-frequency band of the sWuggy test sets.

4.4.2.1 General results

The resulting developmental curves, which evaluate the different models, are presented in the left panel of Figure 4.8, and their corresponding curve slopes in Table 4.1. The resulting pattern of developmental curves is similar to the pattern observed in the ABX task on the CPC representations. Specifically, the bilingual models begin with performance levels as low as those of non-native models but gradually improve and approach the average performance of monolingual models. Yet, the slope of improvement for bilingual models is not as steep as for native models, resulting in the bilingual models never reaching the native models' scores. With the same reasoning from the previous section, we wanted to see whether doubling the number of units for the bilingual models would help fill the gap with the native models. Results are presented in the right panel of Figure 4.8 and in Table 4.1. We see that augmenting the number of clusters slightly improves the results for the bilingual models, with a slightly steeper curve, but they still do not reach the average monolingual (native and non-native) scores.

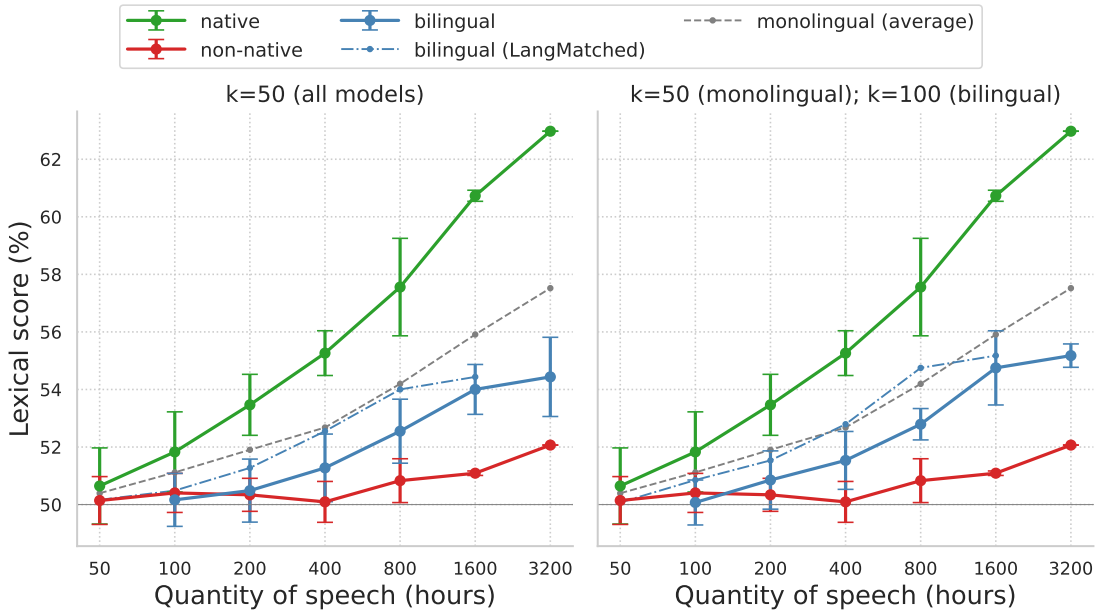


Figure 4.8: sWuggy lexical scores on monolingual and bilingual LSTM representations with regards to the size of the training data train size, averaged on the English and French test sets. The scores are averaged on the English and French test sets and calculated exclusively on the high-frequency band (t=64). The left panels depict scores computed on models trained with 50 clusters, while the right panel showcases bilingual models trained on 100 clusters.

We can also examine the "LangMatched" bilingual curve (which corresponds to the bilingual curve shifted to the left) and compare the bilingual points to their monolingual counterparts based on the amount of speech matched per language. Despite this adjustment, the bilingual models only achieve average monolingual scores and remain far below native speaker performance.

4.4.2.2 Effect of model size - scalability

The present results do not reflect what has been observed in infants, where no delay at the lexical level is observed for infants exposed to two languages (as long as they are compared in terms of total vocabulary size, see Chapter 1, §1.4). A pragmatic and mechanistic explanation could be attributed to the limited number of parameters available in the model, which may not be sufficient for encoding the larger volume of information required. It is worth noting that these models are usually trained on monolingual experiments, and we have not done additional grid searches for hyperparameters. Therefore, increasing the number of parameters could potentially improve their performance. In fact, as discussed in Section 4.1, similar effects have been observed in other studies where a higher number of parameters led to better multilingual outcomes, though this connection requires further investigation (Babu et al., 2021).

To test this hypothesis, we trained new language models with 3,200 hours of data, both monolingual (English and French) and bilingual models. We roughly doubled the number of trainable parameters by adjusting the decoder hidden size, the number of layers, and the number of target clusters the LM was trained on. Specifically, we created new language models trained on 100 clusters with a decoder hidden size of 1,200 (originally 1,024) and 4 decoder layers (originally 3), which increased the number of parameters from 22 million to 42 million[7]. The comparison between the lexical scores for the sWuggy high-frequency band, computed on the original and these new, large models are presented in Table 4.2. Next, we examined the percentage difference between the various conditions when comparing the original models with the ones featuring doubled parameters.

Condition	sWuggy (original)	sWuggy (large)
Native	62.97 %	63.21 %
Non-native	52.07 %	51.00 %
Bilingual	54.44 %	55.88 %

Table 4.2: sWuggy scores for the original and large Language Models, calculated on the most frequent band.

As shown in Table 4.2, doubling the number of parameters improves lexical scores and reduces the difference between the monolingual average and the bilingual model. More specifically, in the original models, there was a 15.68% increase in lexical scores when transitioning from bilingual to native models and a 5.66% increase

[7]We chose these hyperparameters for capacity limitations reasons, but different choice in hyperparameters could be made.

from bilingual models to the monolingual average (the average scores between the non-native and native models). However, with the larger, doubled-parameter models, the difference between the bilingual and native models' scores decreased to 13.13%, and the difference between the bilingual models and the average of monolingual models dropped to 2.2% (nearly halved compared to the smaller models). This demonstrates that increasing the number of parameters effectively reduces the negative impact of bilingual exposure, suggesting that an improved architecture could yield better outcomes. This finding aligns with expectations from a machine learning perspective but is more challenging to explain from a cognitive science standpoint. Therefore, our initial STELA learner reproduces monolingual results effectively but may not provide sufficient computational space for bilingual modelling - at least when it comes to monolingual and bilingual comparison.

4.4.2.3 Frequency effect

In Chapter 3 (§3.2.1), we showed that we could reproduce the frequency effect (i.e., the fact that more frequent words are better recognised than less frequent ones) within our developmental framework for monolingual models. We conducted the same analysis on the bilingual models; results are shown in Figure 4.9 (bottom row). When tested on both the English and French sWuggy evaluation sets, we find a clear frequency effect, with words from the highest frequency bands being, on average, better recognised than words from the lowest bands. However, this effect is smaller than the one observed in the monolingual (native) models. In fact, the frequency effect is highly prominent for the highest frequency bands, where we can see a clear difference between one frequency band and the next. Yet, the lower bands are not as distinct, with almost no score difference between the three rarest frequency bands.

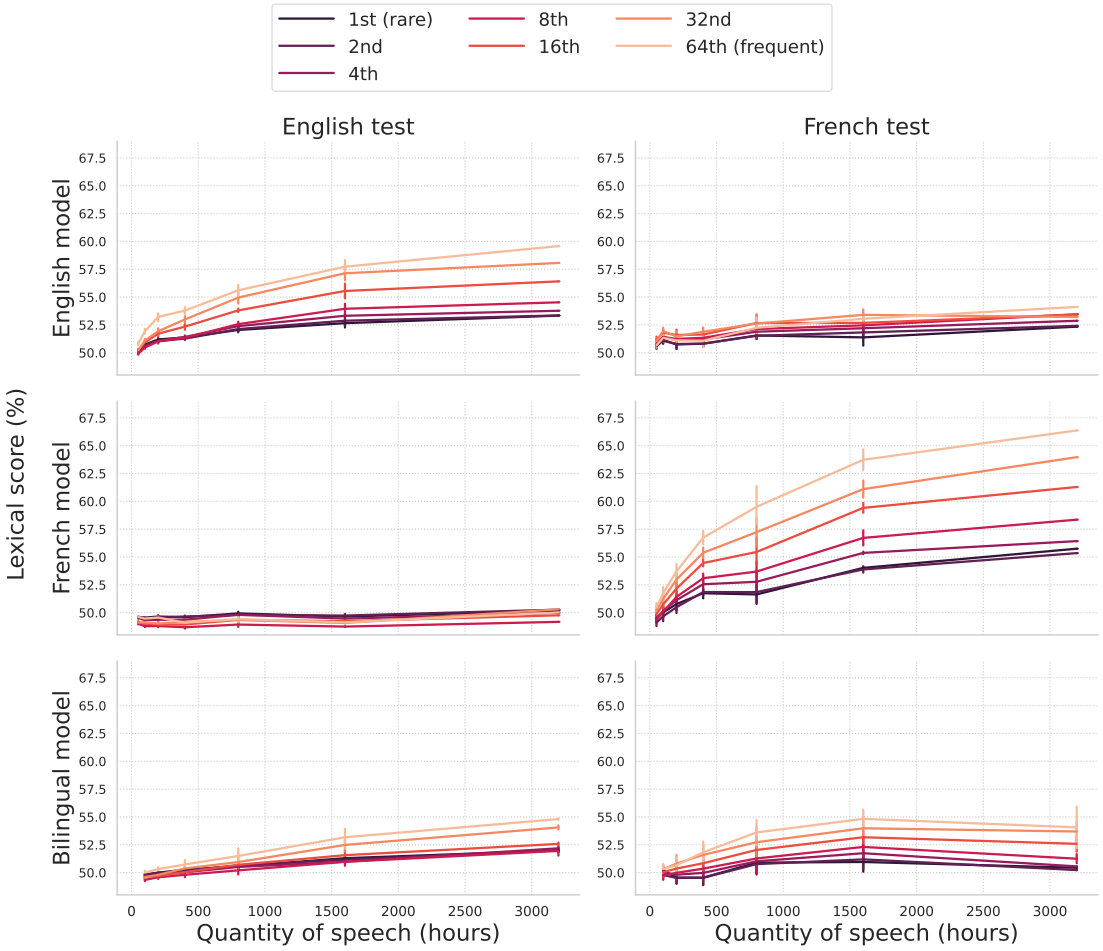


Figure 4.9: sWuggy lexical scores for LSTM models (k=50) across different frequency bands

4.4.3 Prosodic learning - ProsAudit

We turn to the third linguistic level of language acquisition studied in this thesis: that of prosodic learning and structural prosody in particular. We use the ProsAudit metric introduced in Section 3.3and evaluate our bilingual models on the protosyntax subtask (breaks at high or low prominence prosodic boundaries) and the lexical subtask (breaks between or within words). The model must accurately assign a higher probability to phrases with correct pause boundaries to succeed at the tasks. It is important to reiterate here that, contrary to the phonetic and lexical evaluations presented in the previous sections, the prosodic evaluation currently only exists in English. Consequently, the setup is not entirely symmetrical, with all the resulting limitations discussed in Chapter 2.

The developmental curves for the monolingual and bilingual models are presented in Figure 4.10 (and their corresponding slopes in Table 4.1). Focusing on the protosyntax subtask (left panel), we can note that the slope of the bilingual models' developmental curve is much steeper than those of the monolingual models. The smaller bilingual models start with scores similar to the non-native models,

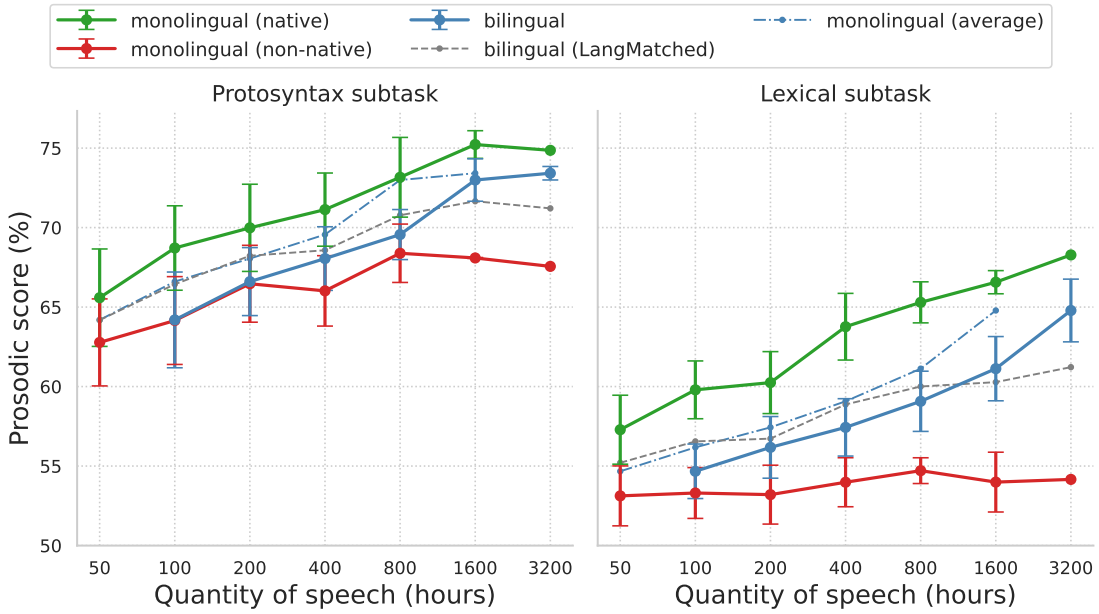


Figure 4.10: ProsAudit prosodic scores for the protosyntax (left) and lexical (right) subtasks, evaluated on both monolingual and bilingual models. Please note that the setup is not entirely symmetrical, as the native models consistently refer to the English ones, while the non-native models pertain to the French ones. This distinction arises from the ProsAudit task being exclusively available in English.

but the larger models score nearly as high as the native ones. This suggests that although the bilingual models need slightly more data than the native models to achieve good scores, they can overcome the fact that they are exposed to two languages on this specific measure. This is not surprising, as our first results, discussed in de Seyssel et al. (2023b), suggest that there are cross-linguistic cues in the protosyntax patterns between English and French, leading prosody in one language to be helpful for the other language at this level.

This cross-linguistic pattern was not found in the lexical subtask, with the non-native models showing nearly no learning. However, as shown in the second panel of Figure 4.10, the bilingual models still have a much steeper curve than even the native ones, and the larger models approach the performance of the native models. These results are on par with the lexical sWuggy results on the English evaluation test (see Appendix C), which is unsurprising as the two rely on lexical knowledge. Overall, there does not seem to be much negative interference from being exposed to two languages at the structural prosodic level, at least on English evaluation. It would also be interesting to evaluate other aspects of prosody.

4.4.4 Effect of language exposure

Until now, we have analysed the consequences of exposure to two languages on multiple aspects of language acquisition using our developmental framework. However, we only considered the "ideal" case of a bilingual model in which the model is exposed to an equal proportion of both languages. This approach has the

advantage of not introducing bias towards either language, allowing us to examine the overall effects of bilingualism. Nonetheless, this does not provide much insight into the effects of varying language exposure. Is this effect, if any, linear? We already know this is not the case between non-native, bilingual, and native models, as bilingual models do not necessarily fall perfectly between non-native and native scores. Here, we explore the question of language exposure further. Specifically, we trained two additional models on 3,200 hours of data, with one trained on 25% English speech and 75% French speech and the other with the opposite proportions.

The results are presented in Figure 4.11. Firstly, considering the phonetic results based on continuous CPC representations, language exposure appears to have a moderate impact. Models exposed to only 25% of the tested language perform nearly as poorly as the non-native models, and there is only a slight increase in scores depending on the increased amount of native language exposure after that. However, the scores are already relatively high, resulting in minimal differences between all models, making it difficult to draw substantial conclusions. When examining phonetic scores calculated using k-means units, the negative effect of exposure to two languages is evident, with only the model exposed to 75% of the tested language matching the non-native results scores, consistent with what we found in Section 4.4.2.

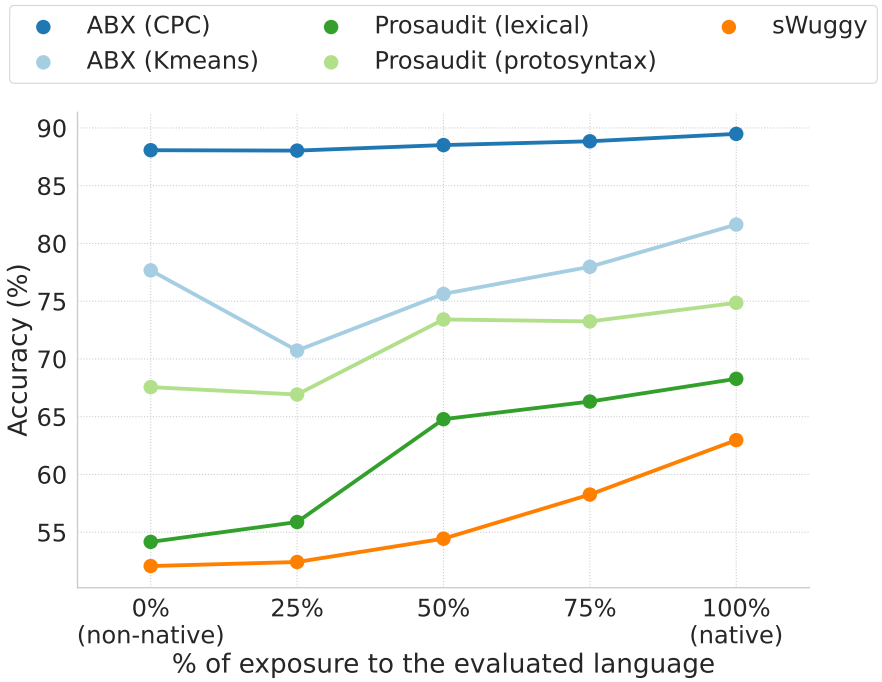


Figure 4.11: Effect of language exposure on the phonetic, lexical and prosodic metrics on the 3,200h models. For the lexical metric, only the high-frequency words are considered. The prosodic metrics are evaluated solely for English, whereas the phonetic and lexical analyses are averaged over both languages. To facilitate comparison, the evaluation is conducted on only one of the two bilingual models.

In terms of lexical learning, exposure to 25% of the evaluated language is insufficient to achieve scores higher than those of non-native models. However, there

appears to be a linear relationship between language exposure and performance *after* the 25% exposure threshold, as 75% exposure leads to scores approximately midway between those achieved by the bilingual (50%) model and the native model. This indicates that although a certain level of exposure is necessary for improvement, once this threshold (in this case, 50%) is reached, the increase in performance is directly proportional to the exposure amount.

Regarding prosodic evaluations, a similar pattern is observed in both the protosyntax and lexical subtasks. For the ProsAudit metrics, 25% exposure to a language is inadequate to demonstrate learning in that language. However, in both subtasks, a significant improvement is seen when comparing 25% exposure to 50% exposure, suggesting the threshold has been surpassed. Following that, there is only a slight effect of increasing exposure for the protosyntax subtask, possibly due to a ceiling effect. In contrast, for the lexical subtask, greater language exposure always leads to better scores.

In summary, for all metrics, 25% exposure to the evaluated language is insufficient to achieve a notable improvement over non-native scores. Beyond this minimal threshold, the importance of language exposure varies between linguistic levels.

4.4.5 Discussion

In this section, we analysed how models trained on bilingual data performed compared to models trained on monolingual data when evaluated on various aspects of linguistic knowledge, using our developmental framework coupled with the STELA learner introduced in Chapter 3. The main takeaway for our study is the systematic cost of being trained on bilingual data, with bilingual models never reaching the performance of the native models, regardless of the measure, size of training data and representations used. Moreover, this cost is not simply the result of the models receiving twice less the amount of data in each language, as even when comparing the models in terms of matched language size, the bilingual models would still score below the native monolingual ones in terms of performance (the only exception being for the protosyntax task where, when matched in terms of language size, the bilingual models sometimes reached the native models' performance).

On some metrics (ABX on k-means and ProsAudit subtasks), the bilingual developmental curves displayed a steeper slope than the monolingual native curves, suggesting they could reach the native models' scores with more training data. Of course, it is not to say that this is not the case for the other measures, as the curves themselves are not linear, and the steepness can vary. However, except maybe for the prosodic level, the models would probably need to be trained on a much larger quantity of input for the bilingual cost to disappear. All in all, at least with the amount of data we are working on in this study, there seems to be an actual negative interference of bilingual exposure, which is not found in psycholinguistics studies, suggesting that our learner is not a good model of the bilingual language acquisition process. Finally, we found some strong asymmetries between the languages at the phonetic and lexical levels, asymmetries which were

presented and further discussed in Appendix C, and which could also suggest some asymmetrical negative interference.

4.4.5.1 How can we explain the performance gap between bilingual and monolingual native language models?

Multiple factors could explain the discrepancy we find in results between the models trained on bilingual data and monolingual data, and we will briefly discuss here some of those directly linked to the modelling aspect of our work. Cognitive implications will be further discussed in Chapter 5.

Phonetic and lexical vocabulary size One of the initial factors to consider when examining linguistic learning with bilingual input is the fact that roughly double the amount of knowledge might be needed across various linguistic categories. For instance, in phonetic and lexical learning, twice the number of phonemes and words should be required to achieve results comparable to those in both languages' evaluation tests when using native models. This hypothesis led us to double the number of k-means units for bilingual models, yet as we have seen above, this did not result in any significant improvement. There are two possible explanations for this. First, as discussed in Section 4.3.1, increasing the number of units does not necessarily enhance the phonetic information they contain. Instead, it helps at the indexical level, particularly in capturing speaker-related information. Another aspect worth considering is that this doubling makes sense with definite linguistic categories. However, following the results in Chapter 3 (§3.2.1), it appears that the models exhibit language learning without creating linguistic categories (or at least not categories as traditionally defined in linguistics). In light of such findings, it is unsurprising not to find a clear effect of doubling units.

Limited model capacity (scalability) Another possibility we have already discussed is related to the capacity of the models. Indeed, we did not fine-tune the hyper-parameters of the models when switching from monolingual to bilingual models, primarily to ensure better comparability. However, these hyper-parameters were initially determined based on training on a single language, specifically English. Since two languages need to be learned in bilingual training, it is logical to hypothesise that the model would require more capacity to account for these two languages and that the original hyper-parameters do not allow for this additional learning. Indeed, the model's capacity might be insufficient to fully accommodate the complexity of both languages. In Section 4.4.2, we showed that doubling the number of hyper-parameters of the language model helped reduce the difference in results between the native and the bilingual models, suggesting that it was indeed a contributing factor, a factor which had already been found in multilingual SSL speech models (Babu et al., 2021). Although we did not conduct this experiment here, it would be interesting to reproduce it at the acoustic model level. However, while this factor can play a role, it is unlikely sufficient to explain the effect entirely.

Negative interference Another potential factor is more a consequence of training on two languages, *negative interference*, that is the negative transfer that exists between two languages, where knowledge of one interferes with the acquisition of the other, already discussed in Section 4.1 (Wang et al., 2019). This interference can potentially occur at various levels and is heavily dependent on the nature and distance between the two languages. At the phonetic level, such interference could occur, for example, if the coverage of phonetic variability in one language differs from that of another and if two phonemes in one language correspond roughly to one phoneme in the other. At the lexical level, such interference could result in a smaller frequency effect for bilingual models, as we have observed, as the model might struggle to differentiate between high-frequency words in one language and their low-frequency counterparts in the other language. Furthermore, negative interference within a language pair does not have to be symmetrical, which could explain the asymmetry observed in our experiments when evaluating bilingual models' behaviours in French and English (Appendix C). It is worth noting that such language transfers are not always negative; one language can potentially aid in learning another language in certain aspects (for instance, if the phonology is closely related).

As discussed in Section 4.1, this negative interference also seems to be linked to language distance, with languages more distant from English resulting in greater negative interference when evaluated in English. These implications highlight a significant limitation of our setup, which is even more prominent when studying bilingualism: the restriction to only two languages. As demonstrated in Chapter 2, language distance has a significant impact on the perception of indexical information, and both cognitive and modelling results suggest that this effect will also play a substantial role in modelling bilingual language acquisition. Another insight that emerges from the potential presence of negative (and positive) language transfer is that the two languages' acquisition processes do not behave entirely independently. If these processes were completely independent, one would not affect the other. Yet, we observe that language interference is present (less than the average monolingual score) even when we consider the languages to be discriminable, examining models trained on larger amounts of data. This suggests that even when we have such segregation, there is still some leaking between processes, or simply some shared learning processes, which suggests that there is not a proper impermeable separation, supporting further our findings from Chapter 2 (§2.2).

4.4.5.2 Role of language exposure

The results discussed in Section 4.4.4 indicate that the quantity of exposure to each language in bilingualism impacts language performance, and this effect is not necessarily linear. These findings align with the field of psycholinguistics, which has also observed that language dominance in bilingual individuals is influenced by the amount of exposure to each language. Specifically, the less dominant language, or the one with less exposure during infancy, tends to lead to lower performance in processing linguistic information compared to the more dominant language (Molnar et al., 2014b; Liu and Kager, 2015). Of course, the advantage of a computational

modelling approach like ours resides in the possibility of testing more strictly and with more controlled variables such hypotheses.

These findings emphasise the importance of considering the proportion of language exposure in bilingual studies, as it can significantly contribute to the variability observed among different bilingual families. Researchers should consider the quantity of exposure to each language when investigating bilingualism, as it can substantially impact language development and processing. Similarly, in the field of speech processing, it is crucial to consider language proportion whenever possible. When the quantity of exposure to a language is too small, having an additional language may not only lack the expected benefits but could also lead to negative interference. Therefore, understanding the relative exposure levels and considering language proportion is essential for leveraging the benefits of multilingual (pre-)training.

4.5 Discussion and conclusion of Chapter 4

In this last chapter of the thesis, we focused on the effect of bilingual input in unsupervised models of speech in an attempt to relate the results to psycholinguistics studies in bilingual early speech perception and learning. Playing with one of the prime indexical features, language, into a developmental simulation of language acquisition, we tied together some of the core works from Chapters 2 and 3.

In summary, we found that although speaker information is represented similarly in monolingual and bilingual speech models, this is not the case for language identity. In fact, only the bilingual models can discriminate between the two languages they were trained on, and this is only when exposed to large enough quantities. Furthermore, the performance of bilingual models was not on par with monolingual models across the linguistic levels studied. There was a performance gap between bilingual and monolingual models, although the size of this gap varied depending on the linguistic level and the amount of input data available. Finally, we found some effect of the proportion of language exposure on linguistic performance.

Although the main findings of our work have already been discussed in the discussions of the corresponding sections (development of language representations in Section 4.3 and bilingual language learning in Section 4.4), we would like to dive further here into the potential relationship found between language separation and performance, which connects analyses from the two main sections of this chapter. In Chapter 5, we will discuss the implications, both in psycholinguistics and speech processing, of the different results presented in this chapter.

4.5.1 Language separation and performance

We want to discuss the main findings on the development of language discrimination and general linguistic learning. Indeed, we found that the STELA learner gradually acquires language discrimination capacities as the size of the training data increases when trained on bilingual speech. More specifically, it appears that language discrimination capacities emerge somewhere between 400 and 1600 hours of training

data. In parallel, we also observed, examining the language discrimination curves presented in Section 4.4, that the slopes of these curves often gradually accelerated with increasing training data. A notable steepness for the bilingual models appeared after 400 hours in most language learning evaluations: phonetic when tested on k-means discrete representations, lexical learning (sWuggy), and prosodic learning (both the ProsAudit protosyntax and lexical subtasks). This sudden steepness results in the curves getting closer to the native models' curves. In summary, as the models become better at distinguishing between languages, their ability to learn individual languages improves, ultimately yielding performance levels closer (and potentially with more training data akin) to those of native models.

One possible reason for the observed improvement in language learning capabilities following the development of language discrimination capacities is the reduction of negative interference. When the models are unable to differentiate between languages effectively, there is a higher likelihood of confusion and interference between the linguistic elements of the two languages. However, once the bilingual models develop the ability to segregate their language representations, they can focus on learning each language in a more isolated manner, minimising the potential for negative interference.

These findings underscore the importance of providing sufficient training data to both languages for unsupervised models trained on bilingual data, as it is through this exposure that they develop the necessary discrimination capacities. Moreover, the results highlight the significance of language discrimination in the overall language learning process, as it plays a crucial role in reducing negative interference and promoting the efficient acquisition of linguistic elements.

Of course, these consequences are difficult to transfer as they are to real bilingual learning in infants, as we do know that discrimination can be made from birth, and therefore more work will need to be carried out on the subject (see Chapter 5, §5.2 for a discussion on the topic).

Chapter 5

General Discussion

In this thesis, we aimed to address the question of how an infant's language environment, particularly exposure to multiple languages, can affect their speech perception of indexical and linguistic information. We were also interested in whether how speech is represented at one level of speech perception also affects how it is represented at the other level.

We employed a computational modelling approach to address these questions, harnessing some of the latest advancements in unsupervised speech processing. As a result, our findings span both the fields of psycholinguistics and speech processing. We will, therefore, discuss the implications, limitations, and future research separately, first in psycholinguistics (§5.2) and then in speech processing (§5.3). Given that the bulk of the discussion has already been presented in the different chapters, we conclude here only with a general discussion spanning the thesis' three core chapters.

5.1 Summary of our contributions

Before we get into this general discussion, we first provide in Table 5.1 a summary of the key contributions and findings from the different chapters that make up our thesis. We do not delve further into the different contributions as they were already discussed in the relevant sections of the thesis.

5.2 Implications, limitations, and future research in psycholinguistics

The research presented in this thesis has significant implications for the field of psycholinguistics, particularly in the context of speech perception and language acquisition in infants. However, more work will have to be carried out with behavioural analyses to confirm or infirm some of the hypotheses resulting from our simulations. We will now discuss some key findings from our work and their implications in psycholinguistics.

Chap.	Novelties	Conclusions
2	* Modelling of indexical speech perception through speaker and language discrimination tasks * Comparison of different input patterns (language, speaker) * Development of a tool to automatically compute acoustic language distance	* Speaker and language information are intrinsically related in global models of speech perception * Effect of language distance in modelling of indexical information * Importance of notions of stability and graduality in computational modelling of speech perception
3	* Design of a developmental framework for modelling early language acquisition, which allows the generation of models' "developmental curves" based on a combination of SSL models called "STELA" * Creation of multiple zero-shot metrics and benchmarks for measuring linguistic knowledge in models (lexical and prosodic but also semantic and syntactic levels) * Proposal of criteria and guidelines for learning simulations in modelling language acquisition	* We can simulate parallel and gradual learning at the phonetic, lexical and prosodic levels using our developmental framework * Learning can arise without sharp linguistic categories * Statistical learning mechanisms are enough to explain such developmental patterns * Phonetic, lexical but also prosodic and somewhat syntactic information can be represented in our STELA learner
4	* Analysis of indexical and linguistic information in monolingual and bilingual STELA learners * Analysis of language discrimination patterns in monolingual and bilingual STELA learners * Comparison of linguistic developmental curves between monolingual and bilingual STELA learners at the phonetic, lexical and prosodic levels	* Bilingual models can discriminate languages but monolingual models cannot * Language discrimination in bilingual models is dependent on the size of input data * There is a cost to bilingual input on linguistic learning * Effect of proportion in bilingual language exposure

Table 5.1: A comprehensive summary of the principal contributions and novel insights presented within this thesis

Do bilingual infants categorise their native languages? In Chapter 2, we were able, using i-vector models, to model the language discrimination findings observed in newborns, both with monolingual and bilingual data. Our results also suggested that language discrimination does not necessarily imply a language separation process as we understand it, paving the way for the idea that a common learning mechanism can be shared between the two languages, an idea which is already discussed in psycholinguistics (see Byers-Heinlein, 2014). In fact, in Chapter 4, we found that models trained on bilingual speech input could indeed learn linguistic information at multiple levels, even when the models were not able to discriminate languages. To our knowledge, this is the first proof of concept as a learning simulation to support the hypothesis that language learning in a bilingual setting can occur without any separate language categories. More sophisticated models could be developed to refine our understanding of language discrimination and separation processes in bilingual infants. Moreover, it would be interesting to see how these findings translate to trilingual or multilingual environments, thereby expanding the scope of this work.

Towards a unified model of indexical and linguistic speech perception? As mentioned above, we could not replicate the language discrimination findings on the STELA learner when provided with small amounts of bilingual speech input, nor with any monolingual data. This can be seen as problematic, as it does not reproduce newborn findings, nor the findings from Chapter 2 when using i-vector models. However, we can provide some additional hypotheses about why this is the case. First, whilst newborns are indeed exposed to speech input (albeit filtered) before birth in their mother's womb, our models do not have such pre-exposure. Pre-training the models on some speech may help accelerate this language discrimination process. Secondly, and possibly most importantly, the STELA learner is a frame-based model, which is why we primarily used it as a model of linguistic speech perception, in contrast to i-vector models, which operate at the utterance level and can, therefore, better capture indexical information. An interesting future line of research would be to combine global and frame-based models into a single learner to model both indexical and linguistic information representations simultaneously.

Is statistical learning enough for bilingual language learning? Another intriguing outcome of our research is the systematic cost associated with bilingual input in linguistic learning, as opposed to monolingual input - a cost which has yet to be observed to such an extent in behavioural studies. We will not expand too much on the subject here as we have already discussed it in depth in Chapter 4 (§4.5), especially regarding the modelling aspects which can lead to such results. However, we have yet to consider a potential hypothesis in terms of cognition: the insufficiency of statistical learning for bilingual input. In Chapter 3, we introduced STELA as a proof of concept for statistical learning in early language acquisition. It showcased the ability to simulate monolingual infants' linguistic developmental patterns. However, as we do not reproduce the patterns observed in psycholinguistics for bilingual infants with the same statistical learner, it is possible

that these statistical learning mechanisms, which were sufficient with monolingual input, are insufficient with bilingual input. In fact, numerous studies have suggested that bilingual infants also rely on additional mechanisms in their speech perception process, such as acute attention and executive control - mechanisms less relied upon in monolingual infants (Bialystok et al., 2009; Bialystok, 2015; D'Souza et al., 2020; D'Souza and D'Souza, 2021). We also know that even monolingual infants make use of more mechanisms than mere statistical learning in their language learning process, such as additional modalities (Abu-Zhaya et al., 2017; Seidl et al., 2015) and reinforcement learning through peer interaction (Nelson, 2007; Yu et al., 2005). Hence, our findings could support the hypothesis that while statistical learning is enough to explain language learning in monolingual infants, such additional mechanisms may also be needed in bilingual infants. Therefore, future work should focus on integrating such mechanisms, which have been proposed in the speech processing field as effective learners of linguistic features. These mechanisms can be found in visually grounded models (e.g., Zhang et al., 2020; Merkx et al., 2023) and reinforcement learning approaches (e.g., Gao et al., 2020). This integration is necessary to confirm our hypothesis and advance our current models.

Interactions between indexical and linguistic information. We have seen throughout the thesis that there is a high interaction within and across different types of indexical and linguistic information. This interaction has been observed multiple times in psycholinguistic research (see Levi, 2021). Computational modelling is an ideal setup to provide a systematic review of such dependencies, and while we have touched upon some of them, more work would be needed to explore these interactions further. For instance, in Chapter 2, we highlighted the effect of language distances on speech perception. However, we have yet to apply these insights to modelling bilingual language acquisition as proposed in Chapter 4. Moreover, given that we already know asymmetries exist within languages at the linguistic learning level, different languages will likely impact linguistic learning differently. Such an analysis is uniquely feasible within computational modelling, where we maintain extensive, if not complete, control over all external variables and can solely modify the language identity. Potential extensions of this work could also examine the role of diverse input patterns, particularly regarding speaker strategies, on linguistic information, merging further our findings from Chapters 2 and 4. Nevertheless, these proposals would necessitate novel corpora to be collected, as we have been unable to extend beyond our current setup in large part due to the scarcity of suitable resources.

Mapping models and humans developmental curves. Finally, to reiterate a point addressed earlier in this thesis, an important area of future work involving learning simulations is the necessity for improved correlation between the models and the humans' learning patterns. This essentially means that there is a need to establish a precise mapping of developmental trajectories between humans and machines, considering both the quantity of input and the results derived. In other words, ideally, we would get similar success rates on models and infants exposed to n amount of input data and, therefore, better map the state of our simulations

with the corresponding age of the simulated infant. The ability to establish such a mapping would enhance the validity of our simulation outcomes, reinforcing them as robust hypotheses for predicting human behaviour. Of course, this mapping also entails a better alignment of the environment and the outcome measures. While we have already made significant advances in using input as close to what infants hear, using raw (or minimally processed for i-vector models) speech, the next step should include moving to more ecological data with speech closer to what the infant hears, compared to clean audiobooks as we are using here. Similarly, one could always find better-aligned measures of outcomes with those used in psycholinguistics studies and work on reducing the intrinsic biases resulting from our models (and from the infants') noise (see Blandón et al., 2021 for a discussion on the topic).

5.3 Implications, limitations, and future research in speech processing

Besides the implications of our research in cognitive science, and psycholinguistics in particular, our work also found its roots in machine learning and automatic speech processing due to our computational modelling approach. As a result, our findings also have significant implications for automatic speech processing, of which we will now lay out some of the most important.

Language distance in speech processing models. We would first like to highlight the role of language distance in speech processing, which we have already touched upon in §5.2. Indeed, our research has revealed that the distance between languages can influence how speech models extract different types of information. This effect is observable when training in a multilingual setup or when the language varies during testing (language transfer). Leveraging this characteristic, we proposed a method to automatically compute acoustic language distances using i-vector models, as detailed in Chapter 2. Following our paper on automatic acoustic language distance, Guillaume and Wisniewski (2022) extended this concept to the wav2vec 2.0 (SSL) speech model, further emphasising the need for such tools. The automatic computation of acoustic language distance could be particularly beneficial in multilingual training setups. Studies have found that the distance between training languages in SSL multilingual speech models and a target language can influence downstream tasks (Jacobs and Kamper, 2021; Nowakowski et al., 2023), and even directly affect word-level representations (Abdullah et al., 2021). However, these initial studies used typological language families (groups of languages with a shared origin) as a measure of language distance. Let us suppose we can extend the work on automatic language distance to these tasks. In that case, we could select languages used in pre-training more effectively based on the target (low-resource) language.

Other aspects of multilingual training Still related to the language aspect is the issue of negative transfer, that is, the fact that multilingual input can

negatively affect performance on linguistic measures and that we have already extensively discussed in Chapter 4. This interference appears to be dependent, at least at some linguistic levels, on the volume of input data processed by the models. Therefore, further investigation is needed to understand this relationship better and determine whether a threshold of input data is required to prevent such interference. Furthermore, it would be interesting to probe how the number of languages in the training data can affect these interferences. In SSL speech research, multilingual models are trained on dozens of languages rather than a couple of them, as we did here. How would the dynamics change when we extend our scope from bilingual to trilingual or even quadrilingual models? Is it better to train on dozens of languages to avoid interference, maximising the quantity of input data? These are questions that future work should address in order to comprehend better the benefits and disadvantages of multilingual pre-training, which is becoming increasingly prevalent in SSL. Finally, more work should be done to assess the impact of language discrimination on language learning in multilingual models, as we have hinted in Chapter 4. If there is a direct correlation between language discrimination and linguistic performance, adding a language discrimination objective in multilingual training could reduce interference and help obtain language-specific features. In fact, recent work on the topic seems to support this hypothesis (Ding et al., 2022).

Representations of indexical information in speech models Our work on monolingual and bilingual SSL models has revealed that language identity is not the only indexical information represented in speech representations. We have discovered that speaker information is also prominent in CPC representations, potentially more than language identity (in bilingual models) and phonetic information. While this might be seen as a fair representation of an infant's speech perception process, given that young infants often fail to recognise words as identical when spoken by individuals of different genders (Houston and Jusczyk, 2000), such a feature is likely undesirable in most speech processing applications. Indeed, whether the task at hand is spoken language modelling or automatic speech processing, the goal typically involves achieving generalised linguistic representations separate from speaker identity. Future research should therefore prioritise removing or disentangling this speaker information from desired features. Recently, researchers have started addressing this issue (Qian et al., 2022; Polyak et al., 2021), but a significant amount of work remains to obtain such results on models trained on small amounts of data.

Generalisation to other speech processing models Finally, in this thesis, we focused on i-vector models for speech representation at the utterance level and a combination of CPC, k-means, and LSTM models for speech representations at the frame level. To generalise our conclusions, it would be interesting to explore other types of models, especially as there is already recent evidence that many of the conclusions described below hold for other unsupervised speech representations models, as discussed throughout the thesis.

5.4 Conclusion

In conclusion, our findings provide a substantial contribution to the fields of psycholinguistics and speech processing. They offer a better understanding of infant speech processing, both at the indexical and linguistic level and their potential dependencies. They also highlight the potential challenges and opportunities in multilingual training and speech representation of indexical and linguistic information. Our work serves as a crucial bridge between these two fields, leveraging the strengths of each to enhance the other. The insights gained through this research not only underscore the value of computational modelling in studying psycholinguistics but also demonstrate how psycholinguistic theories can inspire and inform the design of more effective speech processing models. Furthermore, the research outcomes and the future research directions laid out in this thesis could serve as a roadmap for both communities. By promoting cross-disciplinary dialogue and collaboration, we believe our work will inspire further research that pushes the boundaries of what we know about language acquisition and speech processing, ultimately leading to more sophisticated and effective models.

Taking our work forward, a promising direction that we have hinted to in §5.2 lies in a better simulation of bilingual speech perception. By integrating both the global (*indexical*) and (supra-)segmental (*linguistic*) models of speech perception into a more comprehensive framework, and by incorporating additional learning mechanisms demonstrated by bilingual infants, we hope to mirror the developmental trajectory of bilingual language acquisition more accurately. Such advancements could not only lead to a deeper understanding of bilingual language acquisition but also offer a potential solution to the multilingual input cost observed in our research. This could, in turn, have significant implications for multilingual speech processing.

Bibliography

Abboub, N., Bijeljac-Babic, R., Serres, J., and Nazzi, T. (2015). On the importance of being bilingual: Word stress processing in a context of segmental variability. *Journal of Experimental Child Psychology*, 132:111–120.

Abdullah, B. M., Zaitova, I., Avgustinova, T., Möbius, B., and Klakow, D. (2021). How familiar does that sound? cross-lingual representational similarity analysis of acoustic word embeddings. *arXiv preprint arXiv:2109.10179*.

Abu-Zhaya, R., Seidl, A., Tincoff, R., and Cristia, A. (2017). Building a multimodal lexicon: Lessons from infants' learning of body part words. In *Proc. GLU 2017 International Workshop on Grounding Language Understanding*, pages 18–21.

Adams, O., Cohn, T., Neubig, G., Cruz, H., Bird, S., and Michaud, A. (2018). Evaluating phonemic transcription of low-resource tonal languages for language documentation. In *LREC 2018 (Language Resources and Evaluation Conference)*, pages 3356–3365.

Alain, G. and Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Albareda-Castellot, B., Pons, F., and Sebastián-Gallés, N. (2011). The acquisition of phonetic categories in bilingual infants: New data from an anticipatory eye movement paradigm. *Developmental science*, 14(2):395–401.

Anderson, J. L., Morgan, J. L., and White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and Speech*, 46(2-3):155–182.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Aslin, R. N., Jusczyk, P. W., and Pisoni, D. B. (1998). Speech and auditory processing during infancy: constraints on and precursors to language. *Handbook of Child Psychology, Cognition, Perception, and Language*, 2:147–198.

Association, I. P. (1999). *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge University Press.

Babu, A., Wang, C., Tjandra, A., Lakhotia, K., Xu, Q., Goyal, N., Singh, K., von Platen, P., Saraf, Y., Pino, J., et al. (2021). Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Bahrick, L. E. and Pickens, J. N. (1988). Classification of bimodal english and spanish language passages by infants. *Infant Behavior and Development*, 11(3):277–296.

Bansal, P., Kant, A., Kumar, S., Sharda, A., and Gupta, S. (2008). Improved hybrid model of hmm/gmm for speech recognition. *Technologies and Applications*, page 69.

Best, C. T., McRoberts, G. W., and Goodell, E. (2001). Discrimination of non-native consonant contrasts varying in perceptual assimilation to the listener's native phonological system. *The Journal of the Acoustical Society of America*, 109(2):775–794.

Bialystok, E. (2015). Bilingualism and the development of executive function: The role of attention. *Child development perspectives*, 9(2):117–121.

Bialystok, E., Craik, F. I., Green, D. W., and Gollan, T. H. (2009). Bilingual minds. *Psychological science in the public interest*, 10(3):89–129.

Bijeljac-Babic, R., Höhle, B., and Nazzi, T. (2016). Early prosodic acquisition in bilingual infants: The case of the perceptual trochaic bias. *Frontiers in Psychology*, 7:210.

Bijeljac-Babic, R., Serres, J., Höhle, B., and Nazzi, T. (2012). Effect of bilingualism on lexical stress pattern discrimination in french-learning infants. *PLoS One*, 7(2):e30843.

Blandón, M. A. C., Cristia, A., and Räsänen, O. (2021). Evaluation of computational models of infant language development against robust empirical data from meta-analyses: what, why, and how?

Bloomfield, L. (1933). Language.. new york: Henry holt and company. *P21*.

Bortfeld, H., Morgan, J. L., Golinkoff, R. M., and Rathbun, K. (2005). Mommy and me: Familiar names help launch babies into speech-stream segmentation. *Psychological science*, 16(4):298–304.

Bosch, L., Figueras, M., Teixidó, M., and Ramon-Casas, M. (2013). Rapid gains in segmenting fluent speech when words match the rhythmic unit: evidence from infants acquiring syllable-timed languages. *Frontiers in psychology*, 4:106.

Bosch, L. and Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, 65(1):33–69.

Bosch, L. and Sebastián-Gallés, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy*, 2(1):29–49.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Bruderer, A. G., Danielson, D. K., Kandhadai, P., and Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences*, 112(44):13531–13536.

Burns, T. C., Yoshida, K. A., Hill, K., and Werker, J. F. (2007). The development of phonetic representation in bilingual and monolingual infants. *Applied Psycholinguistics*, 28(3):455–474.

Byers-Heinlein, K. (2014). Languages as categories: Reframing the "one language or two" question in early bilingual development. *Language Learning*, 64(s2):184–201.

Byers-Heinlein, K., Burns, T. C., and Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychological science*, 21(3):343–348.

Byers-Heinlein, K., Morin-Lessard, E., and Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *Proceedings of the National Academy of Sciences*, 114(34):9032–9037.

Carbajal, M. J. (2018). *Separation and acquisition of two languages in early childhood: A multidisciplinary approach*. PhD Thesis, Université de recherche Paris Sciences et Lettres.

Carbajal, M. J., Dawud, A., Thiollière, R., and Dupoux, E. (2016). The "language filter" hypothesis: A feasibility study of language separation in infancy using unsupervised clustering of i-vectors. In *2016 joint ieee international conference on development and learning and epigenetic robotics (icdl-epirob)*, pages 195–201. IEEE.

Carbajal, M. J., Peperkamp, S., and Tsuji, S. (2021). A meta-analysis of infants' word-form recognition. *Infancy*, 26(3):369–387.

Carnegie Mellon University (1993). CMU pronouncing dictionary. `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`. Accessed on 2023-05-31.

Carnegie Mellon University (2019). French CMU pronouncing dictionary. `https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/French/fr.dict`. Accessed on 2021-06-01.

Caucheteux, C., Gramfort, A., and King, J.-R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature Human Behaviour*, pages 1–12.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al. (2022). Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Christophe, A., Millotte, S., Bernal, S., and Lidz, J. (2008). Bootstrapping lexical and syntactic acquisition. *Language and speech*, 51(1-2):61–75.

Chung, Y.-A. and Glass, J. (2018). Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*.

Chung, Y.-A., Wu, C.-C., Shen, C.-H., Lee, H.-Y., and Lee, L.-S. (2016). Unsupervised learning of audio segment representations using sequence-to-sequence recurrent neural networks. In *Proc. Interspeech*, pages 765–769.

Cole, J. (2015). Prosody in context: A review. *Language, Cognition and Neuroscience*, 30(1-2):1–31.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Conneau, A., Baevski, A., Collobert, R., Mohamed, A., and Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Dalmia, S., Sanabria, R., Metze, F., and Black, A. W. (2018). Sequence-based multilingual low resource speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4909–4913. IEEE.

De Houwer, A. (2007). Parental language input patterns and children's bilingual use. *Applied psycholinguistics*, 28(3):411–424.

de Seyssel, M. and Dupoux, E. (2020). Does bilingual input hurt? A simulation of language discrimination and clustering using i-vectors. In *CogSci 2020-42nd Annual Virtual Meeting of the Cognitive Science Society*.

de Seyssel, M., Lavechin, M., Adi, Y., Dupoux, E., and Wisniewski, G. (2022a). Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022*.

de Seyssel, M., Lavechin, M., and Dupoux, E. (2023a). Realistic and broad-scope learning simulations: first results and challenges. *Journal of Child Language*.

de Seyssel, M., Lavechin, M., Titeux, H., Thomas, A., Virlet, G., Revilla, A. S., Wisniewski, G., Ludusan, B., and Dupoux, E. (2023b). Prosaudit, a prosodic benchmark for self-supervised speech models. *arXiv preprint arXiv:2302.12057*.

de Seyssel, M., Wisniewski, G., and Dupoux, E. (2022b). Is the Language Familiarity Effect gradual? A computational modelling approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

de Seyssel, M., Wisniewski, G., Dupoux, E., and Ludusan, B. (2022c). Investigating the usefulness of i-vectors for automatic language characterization. In *Speech Prosody 2022-11th International Conference on Speech Prosody*.

de Vries, W., Wieling, M., and Nissim, M. (2022). Make the best of cross-lingual transfer: Evidence from pos tagging with over 100 languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7676–7685.

DeCasper, A. J. and Fifer, W. P. (1980). Of human bonding: Newborns prefer their mothers' voices. *Science*, 208(4448):1174–1176.

Decasper, A. J. and Prescott, P. A. (1984). Human newborns' perception of male voices: Preference, discrimination, and reinforcing value. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 17(5):481–491.

Dehaene-Lambertz, G. and Pena, M. (2001). Electrophysiological evidence for automatic phonetic processing in neonates. *Neuroreport*, 12(14):3155–3158.

Dehak, N., Torres-Carrasquillo, P. A., Reynolds, D., and Dehak, R. (2011). Language recognition via i-vectors and dimensionality reduction. In *Twelfth annual conference of the international speech communication association*.

Dijkstra, T., Wahl, A., Buytenhuijs, F., Van Halem, N., Al-Jibouri, Z., De Korte, M., and Rekké, S. (2019). Multilink: a computational model for bilingual word recognition and word translation. *Bilingualism: Language and Cognition*, 22(4):657–679.

Ding, F., Wan, G., Li, P., Pan, J., and Liu, C. (2022). Improved self-supervised multilingual speech representation learning combined with auxiliary language information. *arXiv preprint arXiv:2212.03476*.

Dryer, M. S. and Haspelmath, M. (2013). The world atlas of language structures online (max planck institute for evolutionary anthropology, leipzig). *Available at wals. info. Accessed October*, 9:2014.

D'Souza, D., Brady, D., Haensel, J. X., and D'Souza, H. (2020). Is mere exposure enough? the effects of bilingual environments on infant cognitive development. *Royal Society open science*, 7(2):180191.

Dunbar, E., Algayres, R., Karadayi, J., Bernard, M., Benjumea, J., Cao, X.-N., Miskic, L., Dugrain, C., Ondel, L., Black, A. W., et al. (2019). The zero resource speech challenge 2019: Tts without t. *arXiv preprint arXiv:1904.11469*.

Dunbar, E., Bernard, M., Hamilakis, N., Nguyen, T. A., de Seyssel, M., Rozé, P., Rivière, M., Kharitonov, E., and Dupoux, E. (2021). The Zero Resource Speech Challenge 2021: Spoken Language Modelling. In *Proc. Interspeech 2021*, pages 1574–1578.

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., and Dupoux, E. (2017). The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. IEEE.

Dunbar, E., Karadayi, J., Bernard, M., Cao, X.-N., Algayres, R., Ondel, L., Besacier, L., Sakti, S., and Dupoux, E. (2020). The zero resource speech challenge 2020: Discovering discrete subword and word units. *arXiv preprint arXiv:2010.05967*.

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59.

D'Souza, D. and D'Souza, H. (2021). Bilingual adaptations in early development. *Trends in Cognitive Sciences*, 25(9):727–729.

Eimas, P. D., Siqueland, E. R., Jusczyk, P., and Vigorito, J. (1971). Speech perception in infants. *Science*, 171(3968):303–306.

Endress, A. D. and Hauser, M. D. (2010). Word segmentation with universal prosodic cues. *Cognitive psychology*, 61(2):177–199.

Erickson, L. C. and Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, 37:66–108.

Fecher, N. and Johnson, E. K. (2018). The native-language benefit for talker identification is robust in 7.5-month-old infants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(12):1911.

Fecher, N. and Johnson, E. K. (2019). Bilingual infants excel at foreign-language talker recognition. *Developmental science*, 22(4):e12778.

Fecher, N. and Johnson, E. K. (2022). Revisiting the talker recognition advantage in bilingual infants. *Journal of Experimental Child Psychology*, 214:105276.

Fecher, N., Paquette-Smith, M., and Johnson, E. K. (2019). Resolving the (apparent) talker recognition paradox in developmental speech perception. *Infancy*, 24(4):570–588.

Floccia, C., Nazzi, T., and Bertoncini, J. (2000). Unfamiliar voice discrimination for short stimuli in newborns. *Developmental Science*, 3(3):333–343.

Floccia, C., Sambrook, T., Delle Luche, C., Kwok, R., Goslin, J., White, L., Cattani, A., Sullivan, E., Abbot-Smith, K., Krott, A., et al. (2018). Vocabulary of 2-year-olds learning english and an additional language.

Fourtassi, A. (2023). Understanding children's multimodal conversational development: Challenges and opportunities.

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, 31(4):271–288.

Friederici, A. D. and Wessels, J. M. (1993). Phonotactic knowledge of word boundaries and its use in infant speech perception. *Perception & psychophysics*, 54(3):287–295.

Gao, S., Hou, W., Tanaka, T., and Shinozaki, T. (2020). Spoken language acquisition based on reinforcement learning and word unit segmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6149–6153. IEEE.

Gasparini, L., Langus, A., Tsuji, S., and Boll-Avetisyan, N. (2021). Quantifying the role of rhythm in infants' language discrimination abilities: A meta-analysis. *Cognition*, 213:104757.

Gervain, J. and Werker, J. F. (2013). Prosody cues word order in 7-month-old bilingual infants. *Nature communications*, 4(1):1490.

Gogate, L. J., Walker-Andrews, A. S., and Bahrick, L. E. (2001). The intersensory origins of word-comprehension: an ecological–dynamic systems view. *Developmental science*, 4(1):1–18.

Goggin, J. P., Thompson, C. P., Strube, G., and Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & cognition*, 19(5):448–458.

Gomez, R. L. and Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, 70(2):109–135.

Gorman, K. and Bedrick, S. (2019). We need to talk about standard splits. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 2786–2791.

Goto, H. (1971). Auditory perception by normal japanese adults of the sounds" l" and" r.". *Neuropsychologia*.

Grosjean, F. and Byers-Heinlein, K. (2018). *The listening bilingual: Speech perception, comprehension, and bilingualism*. John Wiley & Sons.

Grosjean, F. and Li, P. (2013). *The psycholinguistics of bilingualism*. John Wiley & Sons.

Guillaume, S. and Wisniewski, G. (2022). Déterminer la similarité entre deux langues à l'aide des modèles pré-entraînés de la parole. une étude pilote. In *Journées Jointes des Groupements de Recherche Linguistique Informatique, Formelle et de Terrain (LIFT) et Traitement Automatique des Langues (TAL)*, pages 67–73. CNRS.

Guion, S. G., Flege, J. E., Akahane-Yamada, R., and Pruitt, J. C. (2000). An investigation of current models of second language speech perception: The case of japanese adults' perception of english consonants. *The Journal of the Acoustical Society of America*, 107(5):2711–2724.

Hallé, P. A. and Best, C. T. (2007). Dental-to-velar perceptual assimilation: A cross-linguistic study of the perception of dental stop+/l/clusters. *The Journal of the Acoustical Society of America*, 121(5):2899–2914.

Hallé, P. A., Durand, C., and de Boysson-Bardies, B. (2008). Do 11-month-old french infants process articles? *Language and Speech*, 51(1-2):23–44.

Hepper, P. G., Scott, D., and Shahidullah, S. (1993). Newborn and fetal response to maternal voice. *Journal of Reproductive and Infant Psychology*, 11(3):147–153.

Hirschberg, J. and Manning, C. D. (2015). Advances in natural language processing. *Science*, 349(6245):261–266.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hoff, E. (2018). Bilingual development in children of immigrant families. *Child development perspectives*, 12(2):80–86.

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., and Parra, M. (2012). Dual language exposure and early bilingual development. *Journal of Child Language*, 39(1):1–27.

Höhle, B. (2009). Bootstrapping mechanisms in first language acquisition. *Linguistics*, 47(2):359–382.

Höhle, B., Bijeljac-Babic, R., and Nazzi, T. (2020). Variability and stability in early language acquisition: Comparing monolingual and bilingual infants' speech perception and word recognition. *Bilingualism: Language and Cognition*, 23(1):56–71.

Houston, D. M. and Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, 26(5):1570.

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Kettermann, A., Siebert, C., et al. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition*, 87(1):B47–B57.

Jacobs, C. and Kamper, H. (2021). Multilingual transfer of acoustic word embeddings improves when training on languages related to the target zero-resource language. *arXiv preprint arXiv:2106.12834*.

Johnson, E. K., Bruggeman, L., and Cutler, A. (2018). Abstraction and the (Misnamed) Language Familiarity Effect. *Cognitive Science*, 42(2):633–645.

Johnson, E. K., Westrek, E., Nazzi, T., and Cutler, A. (2011). Infant ability to tell voices apart rests on language experience. *Developmental Science*, 14(5):1002–1011.

Jusczyk, P. W. and Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*, 29(1):1–23.

Jusczyk, P. W., Bertoncini, J., Bijeljac-Babic, R., Kennedy, L. J., and Mehler, J. (1990). The role of attention in speech perception by young infants. *Cognitive Development*, 5(3):265–286.

Jusczyk, P. W. and Hohne, E. A. (1997). Infants' memory for spoken words. *Science*, 277(5334):1984–1986.

Jusczyk, P. W. and Luce, P. A. (2002). Speech perception and spoken word recognition: Past and present. *Ear and hearing*, 23(1):2–40.

Jusczyk, P. W., Luce, P. A., and Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of memory and Language*, 33(5):630–645.

Kahn, J., Rivière, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., et al. (2020). Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

Kearns, J. (2014). Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1):7–8.

Kedar, Y., Casasola, M., and Lust, B. (2006). Getting there faster: 18-and 24-month-old infants' use of function words to determine reference. *Child Development*, 77(2):325–338.

Kisilevsky, B. S., Hains, S. M., Brown, C. A., Lee, C. T., Cowperthwaite, B., Stutzman, S. S., Swansburg, M. L., Lee, K., Xie, X., Huang, H., et al. (2009). Fetal sensitivity to properties of maternal speech and language. *Infant Behavior and Development*, 32(1):59–71.

Kovács, Á. M. and Mehler, J. (2009). Flexible learning of multiple speech structures in bilingual infants. *science*, 325(5940):611–612.

Kremin, L. V., Alves, J., Orena, A. J., Polka, L., and Byers-Heinlein, K. (2022). Code-switching in parents' everyday speech to bilingual infants. *Journal of Child Language*, 49(4):714–740.

Kuhl, P. K. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, 66(6):1668–1679.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., and Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science*, 9(2):F13–F21.

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., and Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, 255(5044):606–608.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

Lany, J. and Saffran, J. R. (2011). Interactions between statistical and semantic information in infant language development. *Developmental science*, 14(5):1207–1219.

Lavechin, M., Bousbib, R., Bredin, H., Dupoux, E., and Cristia, A. (2020). An open-source voice type classifier for child-centered daylong recordings. *arXiv preprint arXiv:2005.12656*.

Lavechin, M., de Seyssel, M., Gautheron, L., Dupoux, E., and Cristia, A. (2022). Reverse engineering language acquisition with child-centered long-form recordings. *Annual Review of Linguistics*, 8:389–407.

Lavechin, M., de Seyssel, M., Métais, M., Metze, F., Mohamed, A., Bredin, H., Dupoux, E., and Cristia, A. (2023). Statistical learning models of early phonetic acquisition struggle with child-centered audio data.

Lavechin*, M., de Seyssel*, M., Titeux, H., Bredin, H., Wisniewski, G., Cristia, A., and Dupoux, E. (2022). Can statistical learning bootstrap early language acquisition? a modeling investigation. *PsyArXiv preprint PsyArXiv:rx94d*.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., et al. (2019). Robust modeling in cognitive science. *Computational Brain & Behavior*, 2:141–153.

Levi, S. V. (2019). Methodological considerations for interpreting the Language Familiarity Effect in talker processing. *WIREs Cognitive Science*, 10(2):e1483.

Levi, S. V. (2021). Perception of indexical properties of speech by children. *The Handbook of Speech Perception*, pages 465–483.

Li, P. and Grant, A. (2019). Scaling up: How computational models can propel bilingualism research forward. *Bilingualism: Language and Cognition*, 22(4):682–684.

Li, P. and Xu, Q. (2022). Computational modeling of bilingual language learning: Current models and future directions. *Language Learning*.

Li, X., Dalmia, S., Black, A. W., and Metze, F. (2019). Multilingual speech recognition with corpus relatedness sampling. *arXiv preprint arXiv:1908.01060*.

Liu, L. and Kager, R. (2015). Bilingual exposure influences infant vot perception. *Infant Behavior and Development*, 38:27–36.

Lu, Y., Huang, M., Qu, X., Wei, P., and Ma, Z. (2022). Language adaptive cross-lingual speech representation learning with sparse sharing sub-networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6882–6886. IEEE.

Ludusan, B., Morii, M., Minagawa, Y., and Dupoux, E. (2021). The effect of different information sources on prosodic boundary perception. *JASA Express Letters*, 1(11):115203.

Macnamara, J. (1967). Problems of bilingualism. *Journal of social issues*, 23(2):n2.

Mandel, D. R., Jusczyk, P. W., and Pisoni, D. B. (1995). Infants' recognition of the sound patterns of their own names. *Psychological Science*, 6(5):314–317.

Marcus, G. F., Vijayan, S., Bandi Rao, S., and Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, 283(5398):77–80.

Marr, D. (1983). Vision: A computational investigation into the human representation and processing of visual information.

Martinez, D., Burget, L., Ferrer, L., and Scheffer, N. (2012). ivector-based prosodic system for language identification. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4861–4864. IEEE.

Martinez, D., Plchot, O., Burget, L., Glembek, O., and Matějka, P. (2011). Language recognition in ivectors space. In *Twelfth annual conference of the international speech communication association*.

Mattys, S. L. and Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78(2):91–121.

Maurer, D. and Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental psychobiology*, 56(2):154–178.

Maxwell, S. E., Lau, M. Y., and Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, 70(6):487.

Maye, J., Werker, J. F., and Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, 82(3):B101–B111.

McMurray, B. (2022). The acquisition of speech categories: Beyond perceptual narrowing, beyond unsupervised learning and beyond infancy. *Language, Cognition and Neuroscience*, pages 1–27.

McPhetres, J., Albayrak-Aydemir, N., Mendes, A. B., Chow, E. C., Gonzalez-Marquez, P., Loukras, E., Maus, A., O'Mahony, A., Pomareda, C., Primbs, M. A., et al. (2021). A decade of theory as reflected in psychological science (2009–2019). *PloS one*, 16(3):e0247986.

Mehler, J., Bertoncini, J., Barriere, M., and Jassik-Gerschenfeld, D. (1978). Infant recognition of mother's voice. *Perception*, 7(5):491–497.

Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., and Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29(2):143–178.

Merkx, D., Scholten, S., Frank, S. L., Ernestus, M., and Scharenborg, O. (2023). Modelling human word learning and recognition using visually grounded speech. *Cognitive Computation*, 15(1):272–288.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. *Pattern recognition and artificial intelligence*, 116:374–388.

Miller, C. L. (1983). Developmental changes in male/female voice classification by infants. *Infant behavior and development*, 6(2-3):313–330.

Miller, C. L., Younger, B. A., and Morse, P. A. (1982). The categorization of male and female voices in infancy. *Infant Behavior and Development*, 5(2-4):143–159.

Mills, M. and Melhuish, E. (1974). Recognition of mother's voice in early infancy. *Nature*, 252(5479):123–124.

Minagawa-Kawai, Y., Mori, K., Naoi, N., and Kojima, S. (2007). Neural attunement processes in infants during the acquisition of a language-specific phonemic contrast. *Journal of Neuroscience*, 27(2):315–321.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., and Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of japanese and english. *Perception & Psychophysics*, 18(5):331–340.

Mohamed, A., Lee, H.-y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., et al. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*.

Molnar, M., Gervain, J., and Carreiras, M. (2014a). Within-rhythm class native language discrimination abilities of basque-spanish monolingual and bilingual infants at 3.5 months of age. *Infancy*, 19(3):326–337.

Molnar, M., Lallier, M., and Carreiras, M. (2014b). The amount of language exposure determines nonlinguistic tone grouping biases in infants from a bilingual environment. *Language Learning*, 64(s2):45–64.

Moon, C., Cooper, R. P., and Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant behavior and development*, 16(4):495–500.

Munson, W. and Gardner, M. B. (1950). Standardizing auditory tests. *The Journal of the Acoustical Society of America*, 22(5):675–675.

Muthukrishna, M. and Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, 3(3):221–229.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998). Language discrimination by newborns: toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human perception and performance*, 24(3):756.

Nazzi, T., Jusczyk, P. W., and Johnson, E. K. (2000). Language discrimination by english-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, 43(1):1–19.

Nelson, K. (2007). *Young minds in social worlds: Experience, meaning, and memory*. Harvard University Press.

Nespor, M. and Vogel, I. (2012). *Prosodic phonology*. De Gruyter Mouton.

New, B., Pallier, C., Brysbaert, M., and Ferrand, L. (2004). Lexique 2: A new french lexical database. *Behavior Research Methods, Instruments, & Computers*, 36(3):516–524.

Nguyen, T. A., de Seyssel, M., Algayres, R., Roze, P., Dunbar, E., and Dupoux, E. (2022). Are word boundaries useful for unsupervised language learning? *arXiv preprint arXiv:2210.02956*.

Nguyen*, T. A., de Seyssel*, M., Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., and Dupoux, E. (2020). The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

Nowakowski, K., Ptaszynski, M., Murasaki, K., and Nieuważny, J. (2023). Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Orena, A. J., Byers-Heinlein, K., and Polka, L. (2020). What do bilingual infants actually hear? evaluating measures of language input to bilingual-learning 10-month-olds. *Developmental science*, 23(2):e12901.

Orena, A. J. and Polka, L. (2019). Monolingual and bilingual infants' word segmentation abilities in an inter-mixed dual-language task. *Infancy*, 24(5):718–737.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Pelucchi, B., Hay, J. F., and Saffran, J. R. (2009). Statistical learning in a natural language by 8-month-old infants. *Child development*, 80(3):674–685.

Place, S. and Hoff, E. (2011). Properties of dual language exposure that influence 2-year-olds' bilingual proficiency. *Child development*, 82(6):1834–1849.

Place, S. and Hoff, E. (2016). Effects and noneffects of input in bilingual environments on dual language skills in 2 1/2-year-olds. *Bilingualism: Language and Cognition*, 19(5):1023–1041.

Polka, L. and Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human perception and performance*, 20(2):421.

Polyak, A., Adi, Y., Copet, J., Kharitonov, E., Lakhotia, K., Hsu, W.-N., Mohamed, A., and Dupoux, E. (2021). Speech resynthesis from discrete disentangled self-supervised representations. *arXiv preprint arXiv:2104.00355*.

Poulin-Dubois, D., Bialystok, E., Blaye, A., Polonia, A., and Yott, J. (2013). Lexical access and vocabulary development in very young bilinguals. *International Journal of Bilingualism*, 17(1):57–70.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C.-I., Cox, D., Hasegawa-Johnson, M., and Chang, S. (2022). Contentvec: An improved self-supervised speech representation by disentangling speakers. In *International Conference on Machine Learning*, pages 18003–18017. PMLR.

Quine, W. and Van, O. (1960). Word and object: An inquiry into the linguistic mechanisms of objective reference.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Ramus, F. (2002a). Acoustic correlates of linguistic rhythm: Perspectives. In *Speech Prosody 2002, International Conference*.

Ramus, F. (2002b). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition*, 2(1):85–115.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., and Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science*, 288(5464):349–351.

Rivera-Gaxiola, M., Silva-Pereyra, J., and Kuhl, P. K. (2005). Brain potentials to native and non-native speech contrasts in 7-and 11-month-old american infants. *Developmental science*, 8(2):162–172.

Rivière, M., Joulin, A., Mazaré, P.-E., and Dupoux, E. (2020). Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418. IEEE.

Romberg, A. R. and Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(6):906–914.

Saffran, J. R., Aslin, R. N., and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, 274(5294):1926–1928.

Saffran, J. R. and Kirkham, N. Z. (2018). Infant statistical learning. *Annual review of psychology*, 69:181–203.

Samuelson, L. K. and McMurray, B. (2017). What does it take to learn a word? *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(1-2):e1421.

Schatz, T. (2016). *ABX-discriminability measures and applications*. PhD thesis, Université Paris 6 (UPMC).

Schatz, T., Feldman, N. H., Goldwater, S., Cao, X.-N., and Dupoux, E. (2021). Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7):e2001844118.

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., and Dupoux, E. (2013). Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association*, pages 1–5.

Schierwagen, A. (2012). On reverse engineering in the cognitive and brain sciences. *Natural Computing*, 11:141–150.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

Sebastián-Gallés, N. and Bosch, L. (2002). Building phonotactic knowledge in bilinguals: role of early exposure. *Journal of Experimental Psychology: Human Perception and Performance*, 28(4):974.

Sebastián-Gallés, N. and Bosch, L. (2009). Developmental shift in the discrimination of vowel contrasts in bilingual infants: Is the distributional account all there is to it? *Developmental science*, 12(6):874–887.

Sebastian-Galles, N. and Santolin, C. (2020). Bilingual acquisition: The early steps. *Annual Review of Developmental Psychology*, 2:47–68.

Seidl, A., Tincoff, R., Baker, C., and Cristia, A. (2015). Why the body comes first: Effects of experimenter touch on infants' word finding. *Developmental science*, 18(1):155–164.

Shi, R. and Melançon, A. (2010). Syntactic categorization in french-learning infants. *Infancy*, 15(5):517–533.

Shi, R. and Werker, J. F. (2001). Six-month-old infants' preference for lexical words. *Psychological Science*, 12(1):70–75.

Shi, R. and Werker, J. F. (2003). The basis of preference for lexical words in 6-month-old infants. *Developmental Science*, 6(5):484–488.

Shi, R., Werker, J. F., and Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition*, 72(2):B11–B21.

Silverman, D., Blankenship, B., Kirk, P., and Ladefoged, P. (1995). Phonetic structures in jalapa mazatec. *Anthropological linguistics*, pages 70–88.

Singh, L., Morgan, J. L., and White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language*, 51(2):173–189.

Singh, L., White, K. S., and Morgan, J. L. (2008). Building a word-form lexicon in the face of variable input: Influences of pitch and amplitude on early spoken word recognition. *Language Learning and Development*, 4(2):157–178.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE.

Steedman, M. (1991). Structure and intonation. *Language*, pages 260–296.

Streeter, L. A. (1976). Language perception of 2-month-old infants shows effects of both innate mechanisms and experience. *Nature*, 259(5538):39–41.

Sundara, M., Ngon, C., Skoruppa, K., Feldman, N. H., Onario, G. M., Morgan, J. L., and Peperkamp, S. (2018). Young infants' discrimination of subtle phonetic contrasts. *Cognition*, 178:57–66.

Sundara, M., Polka, L., and Molnar, M. (2008). Development of coronal stop perception: Bilingual infants keep pace with their monolingual peers. *Cognition*, 108(1):232–242.

Sundara, M. and Scutellaro, A. (2011). Rhythmic distance between languages affects the development of speech perception in bilingual infants. *Journal of Phonetics*, 39(4):505–513.

Thorburn, C., Feldman, N., and Schatz, T. (2019). A quantitative model of the language familiarity effect in infancy. In *2019 Conference on Cognitive Computational Neuroscience*, Berlin, Germany. Cognitive Computational Neuroscience.

Thordardottir, E. (2011). The relationship between bilingual exposure and vocabulary development. *International Journal of Bilingualism*, 15(4):426–445.

Tincoff, R. and Jusczyk, P. W. (1999). Some beginnings of word comprehension in 6-month-olds. *Psychological science*, 10(2):172–175.

Tsoukala, C., Broersma, M., Van den Bosch, A., and Frank, S. L. (2021). Simulating code-switching using a neural network model of bilingual sentence production. *Computational Brain & Behavior*, 4:87–100.

Tsoukala, C., Frank, S. L., and Broersma, M. (2017). "he's pregnant": Simulating the confusing case of gender pronoun errors in l2 english. In *the 39th annual meeting of the cognitive science society (cogsci 2017)*, pages 3392–3397. Cognitive Science Society.

Tsoukala, C., Frank, S. L., van den Bosch, A., Valdes Kroff, J., and Broersma, M. (2020). Simulating spanish-english code-switching. el modelo está generating code-switches. pages 20–29.

Van Den Oord, A., Vinyals, O., et al. (2017). Neural discrete representation learning. *Advances in neural information processing systems*, 30.

Versteegh, M., Thiolliere, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., and Dupoux, E. (2015). The zero resource speech challenge 2015. In *Sixteenth annual conference of the international speech communication association*.

Vihman, M. M., Thierry, G., Lum, J., Keren-Portnoy, T., and Martin, P. (2007). Onset of word form recognition in english, welsh, and english–welsh bilingual infants. *Applied Psycholinguistics*, 28(3):475–493.

Wagner, M. and Watson, D. G. (2010). Experimental and theoretical advances in prosody: A review. *Language and cognitive processes*, 25(7-9):905–945.

Wang, D., Wang, X., and Lv, S. (2019). An overview of end-to-end automatic speech recognition. *Symmetry*, 11(8):1018.

Wang, Z., Lipton, Z. C., and Tsvetkov, Y. (2020). On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450, Online. Association for Computational Linguistics.

Ward, C. D. and Cooper, R. P. (1999). A lack of evidence in 4-month-old human infants for paternal voice preference. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, 35(1):49–59.

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., and Bowman, S. R. (2020). Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Wennerstrom, A. (2001). *The music of everyday speech: Prosody and discourse analysis*. Oxford University Press.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., and Amano, S. (2007). Infant-directed speech supports phonetic category learning in english and japanese. *Cognition*, 103(1):147–162.

Werker, J. F. and Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant behavior and development*, 7(1):49–63.

Werker, J. F. and Yeung, H. H. (2005). Infant speech perception bootstraps word learning. *Trends in cognitive sciences*, 9(11):519–527.

Wester, M. (2010). The emime bilingual database. Technical report, The University of Edinburgh.

Wojcik, E. H., Zettersten, M., and Benitez, V. L. (2022). The map trap: Why and how word learning research should move beyond mapping. *Wiley Interdisciplinary Reviews: Cognitive Science*, 13(4):e1596.

Ylinen, S., Bosseler, A., Junttila, K., and Huotilainen, M. (2017). Predictive coding accelerates word recognition and learning in the early stages of language development. *Developmental science*, 20(6):e12472.

Yoshida, K. A., Pons, F., Maye, J., and Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy*, 15(4):420–433.

Yu, C., Ballard, D. H., and Aslin, R. N. (2005). The role of embodied intention in early lexical acquisition. *Cognitive science*, 29(6):961–1005.

Zhang, M., Tanaka, T., Hou, W., Gao, S., and Shinozaki, T. (2020). Sound-image grounding based focusing mechanism for efficient automatic spoken language acquisition. In *Interspeech*, pages 4183–4187.

Zhang, Y., Kuhl, P. K., Imada, T., Kotani, M., and Tohkura, Y. (2005). Effects of language experience: neural commitment to language-specific auditory patterns. *NeuroImage*, 26(3):703–720.

Appendix A

Machine ABX Phoneme Discriminability Task

The machine ABX phoneme discriminability task (Schatz et al., 2013) builds upon the ABX task, often used in psycholinguistics and speech perception (Munson and Gardner, 1950). In the original psycholinguistic counterpart, a participant is first presented with two sound segments, A and B, differing in the tested feature (e.g. a different phoneme in the case of phoneme discriminability), and then with a third sound segment X, which either corresponds to the segment A or the segment B (for example, another realisation of the same phoneme as A). The participant must then decide which of A or B the X segment resembles. This setup is then repeated over multiple triplets to achieve statistical stability. In the case of phoneme discrimination, the norm is to use minimal pairs that contrast only in the central phoneme (e.g. [b ɛ t] and [b æ t]), in order to see whether the participant can discriminate between the phones [ɛ] and [æ].

Schatz et al. (2013) proposed to adapt this task to machine evaluation. The adapted setup is presented in Figure A.1: distances are calculated between the speech representations of A and X ($d(A, X)$) and of B and X ($d(B, X)$). If $d(A, X) < d(B, X)$, the machine successfully discriminates the pair on the presented triplet and is given a score of 1 on this triplet, otherwise 0. This is then replicated on thousands of triplets, and the average score over all triplets corresponds to the average accuracy on the phoneme discriminability task for the machine (see Schatz (2016) for a review). As for humans, the task can also be performed under different conditions, with within-speaker and across-speakers setups, as explained in Figure A.1. Of course, this task can also be adapted to discrimination tasks other than focused on phonemes, such as language and speaker discrimination, as we see in Chapter 2. It should be noted that the ABX psycholinguistics task can only be used in adults and older children since infants cannot provide verbal responses. Alternatives have been proposed to test infants, such as the Conditioned Head Turn and the High Amplitude sucking paradigms, where in that case, familiarisation to an item X is performed before switching to either a similar (A) or different (B) segments, and surprise reaction of the infant is monitored. Nevertheless, because these different paradigms are only used due to the impossibility of infants

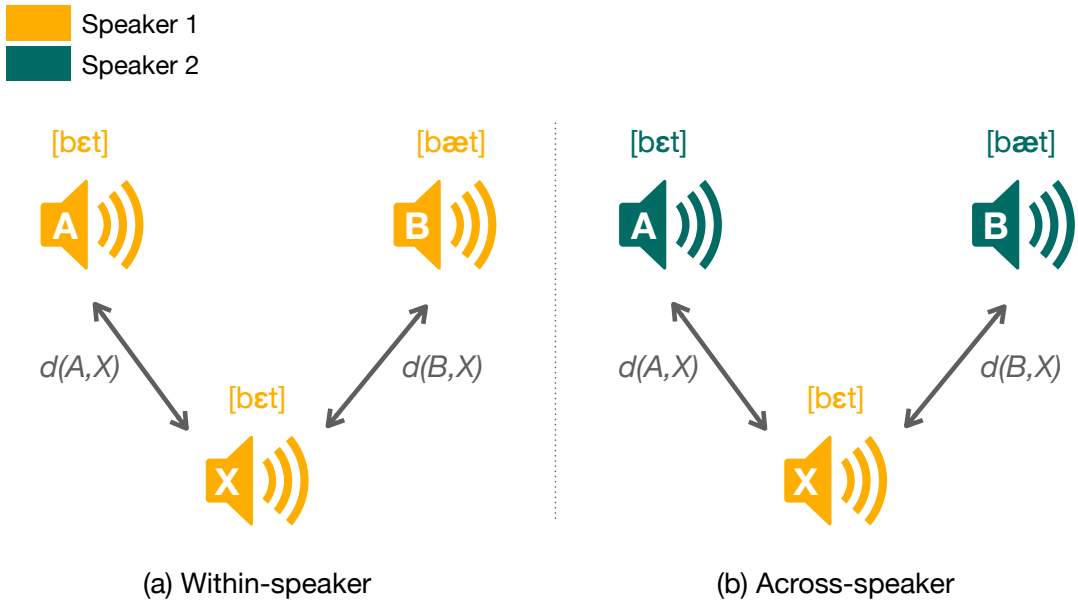


Figure A.1: Machine ABX Phoneme Discriminability task. The system is presented with three speech segments, A, B and X, with A and X sharing the same central phoneme (e.g. [ɛ]) but B differing in the central phoneme (e.g. [æ]). Distances between A and X ($d(A, X)$) and between B and X ($d(B, X)$) are then computed. The machine successfully discriminates the central phonemes if $d(A, X) < d(B, X)$. We consider two setups, the "within-speaker" setup (left), with the speech segments being uttered by the same speaker, and the "across-speaker" setup (right), with A and B being uttered by the same speaker but X being uttered by a different speaker.

responding to the original ABX task, we believe the machine ABX task still holds for the simulation of infants' discrimination capacities.

Appendix B

Phone Sets and Evaluation Sets for Chapters 3 and 4

B.1 Generating the evaluation sets

In this section, we give a better overview of how we generated the French and English CommonVoice-based evaluation sets presented in Chapter 3 and further used in Chapter 4. The aim was to have test sets as clean and balanced in terms of gender, voice, and accents as possible, made of English and French utterances, to be used for generating ABX test sets and additional probing analyses.

We used the CommonVoice 7.1 dataset (Ardila et al., 2019), which gathers read speech in different languages. We restrained our selection to utterances for which accent information was available and tagged as "US" for the English test set and "France" for the French test set. We then selected 10 hours of data split equally between 24 speakers (12M) in each language.

Retrieving phones and words alignments We then retrieved word and phone alignment for all utterances. We used Kaldi (Povey et al., 2011), as it allows for multiple pronunciations based on acoustic input, which is particularly important in French due to the multiple cases of liaisons. We used the English CMU Dictionary and the French CMU Dictionary (Carnegie Mellon University, 2019) for the English and French grapheme-to-phoneme lexicons, which were subsequently converted in line with the phone sets presented in Section B.2). To ensure the validity of our test sets, we opted not to proceed with a typical Grapheme-to-Phoneme (G2P) step (Povey et al., 2011), in which a G2P model is commonly used to automatically infer the phonetic transcription of words not present in the lexicon. By bypassing this step, we ensure that all words without transcription will be mapped to a <JUNK> token[1]. Acoustic models based on HMM-GMMs Bansal et al. (2008) are then iteratively trained on the test set, and phone alignment is iteratively inferred. We also computed word alignments.

[1]Some caveats apply to these techniques, as they may slightly reduce the quality of the acoustic model on which the alignment will be based.

Generating ABX evaluations Finally, we used the retrieved phone alignments to create a French and an English ABX task. This was done by generating all possible phone triplets from the French and English sets, following the approaches in Dunbar et al. (2017, 2019, 2020); Nguyen*, de Seyssel* et al. (2020).

B.2 Phone sets

The phone sets used for all analyses carried out in Chapters 3 and 4 are presented in Tables B.1 (French) and B.2 (English). We also provide the mapping between phone transcriptions from the CMU Pronunciation Dictionary (Carnegie Mellon University, 1993) (English), the Lexique pronunciation dictionary (New et al., 2004) and the International Phonetic Alphabet (IPA) (Association, 1999).

CMU	Lexique	IPA
gn	N	ɲ
nn	n	n
mm	m	m
jj	Z	ʒ
ss	s	s
ll	l	l
bb	b	b
kk	k	k
vv	v	v
zz	z	z
gg	g	ɡ
ww	w	w
pp	p	p
ff	f	f
ch	S	ʃ
rr	R	ʁ
yy	j	j
dd	d	d
tt	t	t
ou	u	u
ei	e	e
ii	i	i
aa	a	a
ai	E	ɛ
on	§	ɔ̃
an	@	ɑ̃
oo	o	o
oe	9	œ
eu	2	ø
ee	°	ə̃
un	1	œ̃
uu	y	y
in	5	ɛ̃
uy	8	ɥ

Table B.1: French Phonetic inventory used in this thesis, with their corresponding CMU Pronouncing Dictionary, Lexique (New et al., 2004) and IPA phone transcriptions

CMU	Lexique	IPA
K	k	k
P	p	p
T	t	t
S	s	s
NG	N	ŋ
V	v	v
F	f	f
D	d	d
TH	T	θ
B	b	b
SH	S	ʃ
JH	dZ	ʤ
G	g	g
R	r	ɹ
N	n	n
Z	z	z
M	m	m
HH	h	h
ZH	Z	ʒ
DH	D	ð
CH	tS	tʃ
Y	j	j
L	l	l
W	w	w
UH	U	ʊ
EH	E	ɛ
AH0	@	ə
IY	i:	iː
AO	o	ɔ
IH	I	ɪ
AY	aI	aɪ
AA	Q	ɑ
AE	{	æ
OW	@U	oʊ
AH	V	ʌ
EY	eI	eɪ
AW	aU	aʊ
UW	u:	uː
ER	3:	ɜː
OY	OI	ɔɪ

Table B.2: French Phonetic inventory used in this thesis, with their corresponding CMU Pronouncing Dictionary, Lexique (New et al., 2004) and IPA phone transcriptions

Appendix C

Language asymmetries in the bilingual developmental curves

In this Appendix, we present results on the ABX (CPC and k-means) and sWuggy metrics separately for each evaluation language and briefly discuss the results. We also present the average relative percentage increase (the developmental curves' slopes) for each condition and each evaluation language in Table C.1.

	Evaluation	Monolingual		Bilingual
		Native	Non-native	
ABX CPC	English	+ 0.82%	+ 0.72%	+ **0.84%**
ABX CPC	French	+ **0.83%**	+ 0.69%	+ 0.64%
ABX CPC	Average	+ **0.83%**	+ 0.70%	+ 0.74%
ABX k-means	English	+ 2.44%	+ 2.18%	+ **3.91%**
ABX k-means	French	+ **2.76%**	+ 2.07%	+ 1.51%
ABX k-means	Average	+ 2.59%	+ 2.12%	+ **2.71%**
sWuggy	English	+ **2.69%**	+ 0.14%	+ 1.86%
sWuggy	French	+ **4.68%**	+ 1.11%	+ 1.46%
sWuggy	Average	+ **3.70%**	+ 0.63%	+ 1.65%

Table C.1: Average relative percentage increase observed when the size of the training set is doubled for all phonetic, lexical, and prosodic development curves (equivalent to the slope of the development curves), for each evaluation language. Both the k-means and sWuggy models were trained using 50 clusters (k=50), except for the instances labelled as 'bil k=100', which indicates that the bilingual models alone were trained using 100 clusters.

Phonetic learning on the CPC representations (ABX)

We report the ABX scores on the CPC representations for each evaluation language separately in Figure C.1 and Table C.1. Interestingly, the nativeness effect (i.e., the difference between the native and non-native models) is much smaller when

tested on the French set. Moreover, there does not seem to be any significant difference between the scores of the bilingual and non-native monolingual models on the French set. In contrast, on the English set, the bilingual model significantly outperforms the non-native (French) model in discriminating English phonemes, even reaching the average performance of the monolingual models on the larger training set.

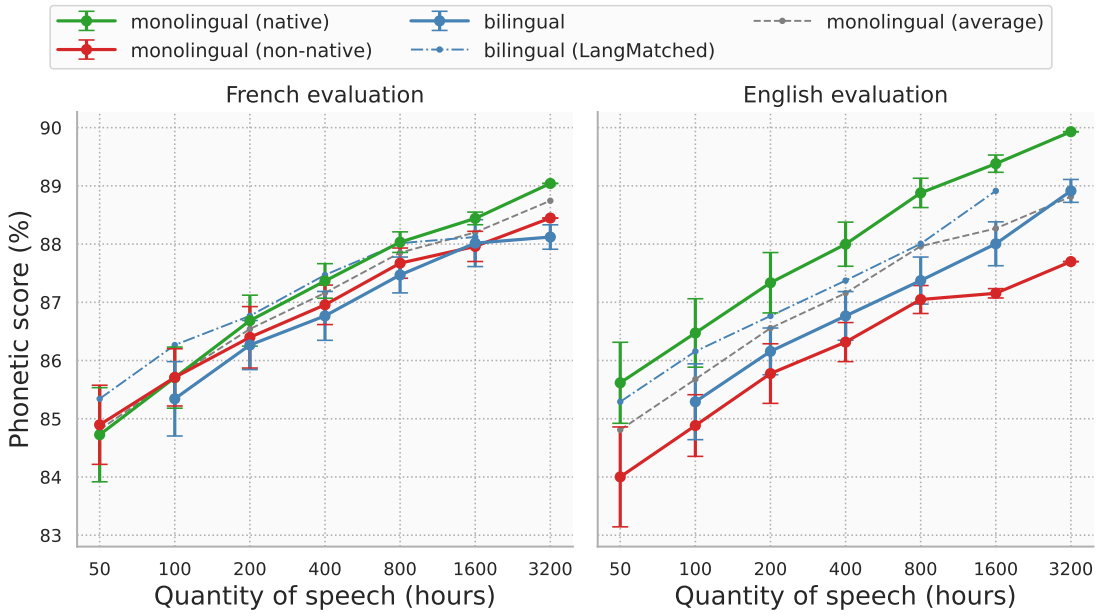


Figure C.1: ABX phonetic scores on monolingual and bilingual CPC models with regards to the size of the training data. Left: models evaluated on the French set. Right: models evaluated on the English set.

Phonetic learning on the k-means representations (ABX)

The ABX scores calculated on the k-means representations, presented in Figure C.2 and Table C.1, also reveal a strong asymmetry between the languages, with very different patterns depending on the evaluation set. When tested on French, the bilingual model consistently performs much worse than the non-native model, with even a slight drop in performance when increasing the training set size from 1,600 to 3,200 hours. However, when tested on the English set, the pattern differs: the slope of the bilingual curve is considerably steeper than that of the two monolingual models, resulting in scores for the bilingual model that are much better than the non-native model for the largest (3,200h) training size, despite starting significantly lower for the smaller model sizes. Indeed, in the 3,200h case, the bilingual models reach the average score of native and non-native models, akin to what was found for the CPC representations, eliminating any detrimental effect of bilingualism at this point. It is highly likely that if we had been able to train the bilingual models with more data, they would have eventually caught up with the scores of the native ones.

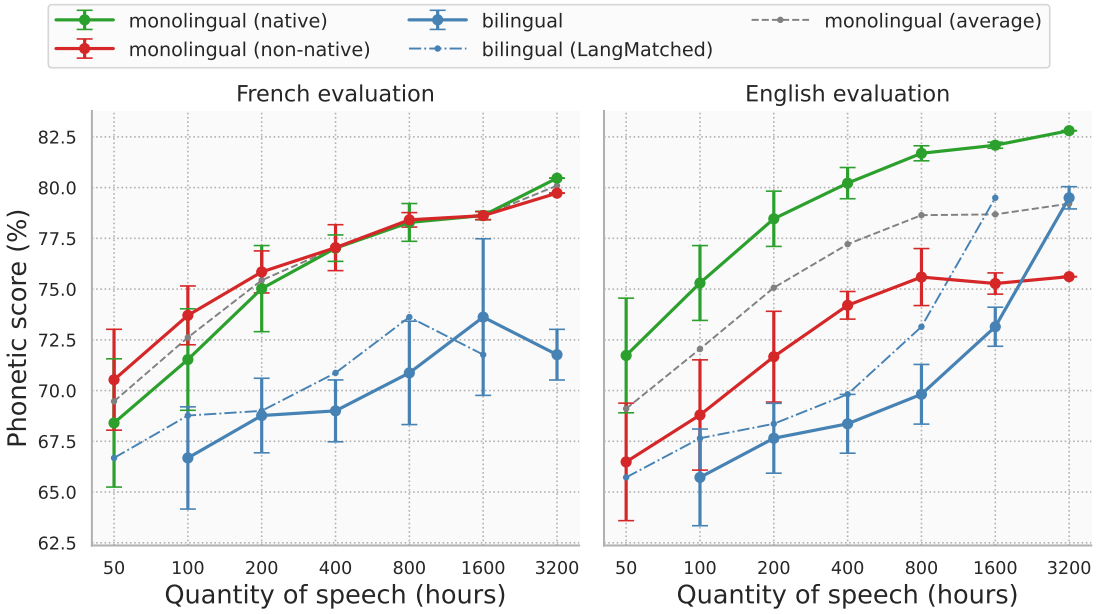


Figure C.2: ABX phonetic scores on monolingual and bilingual k-means representations w.r.t. train size. Left: models evaluated on the French set. Right: models evaluated on the English set.

What stands out from this asymmetry is the possibility that the French test set is much more challenging than the English one. This difficulty could be attributed to two different factors: first, the quality of the test set itself (e.g., the phoneme alignment or the acoustics), although we did our best to ensure it was as clean as possible (see Appendix B.1). Additionally, the French test set was created similarly to the English one. The second factor could be the greater difficulty in phoneme discrimination within the French language. It is also possible that there is a significant overlap, with many English phonemes being similar to those in French (the reverse may not necessarily be true, as the French language has more phonemes and therefore, the percentage of phonemes in English might not be as large).

Lexical learning (sWuggy)

We present the results on sWuggy separately for each evaluation language in Figure C.3 and Table C.1. The patterns observed differ between the two languages and align with our findings in the previous sections at the phonetic level. Specifically, for the English evaluation, we notice a much steeper learning curve for the bilingual models, with an increase in slope between 800h and 1,600h (when language separation begins to appear), and the bilingual models reach the monolingual average lexical scores at 3,200h. In contrast, for the French evaluation, the slope is more moderate, even displaying a slight decrease around the higher models. This is partly caused by a much more significant variance between the models trained on the same amount of speech. Nonetheless, it is reasonable to assume that these patterns result from lagging at the French phonetic level rather than an

additional detrimental effect (negative interference) of bilingualism at the lexical level. Moreover, the high variance between the models suggests that a significant amount of noise is inherent to the learning process, whether at the phonetic or lexical (by cascade) level.

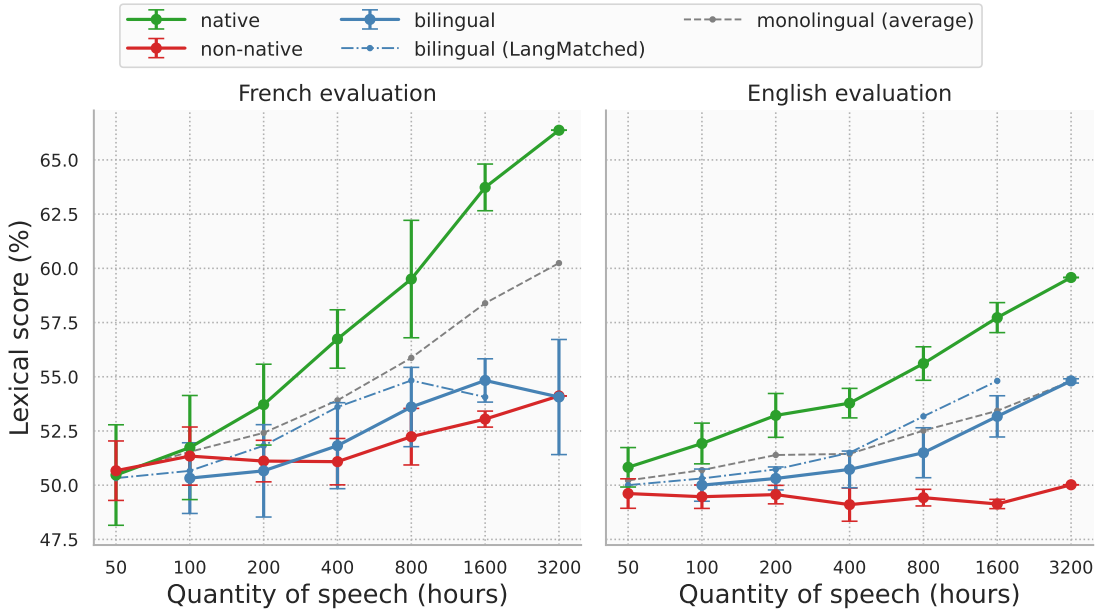


Figure C.3: sWuggy lexical scores on monolingual and bilingual LSTM representations with regards to the size of the training data, averaged on the English and French test sets. Scores are calculated on the high-frequency band (t=64) only. All scores are computed on models trained with 50 clusters. Left: models evaluated on the French set. Right: models evaluated on the English set.

Discussion

Results presented here suggest a strong asymmetry in bilingual models as a result of the language evaluated, at least at the phonetic and lexical levels. We had already observed this asymmetry at the monolingual level in Chapter 3. We first saw this at the phonetic level, where the French phonetic evaluation failed to yield a significant nativeness effect while the English test did. We then observed this at the lexical level, with the French evaluation leading to higher native lexical scores than the English evaluation.

However, these asymmetries could result from multiple factors, and it is difficult to disentangle the different types of biases contributing to them. There could be differences in the training sets, where one set could lead to better linguistic learning than the other without reflecting the language asymmetries. Similarly, the evaluation sets themselves might differ in their difficulty. Such biases could result from, among other things, differing acoustic conditions or a larger diversity in one language than in the other, leading to better learning. To address these issues, it is usually safer to look at the results in a symmetrical testing setup, as

we advocated in Chapter 2, as it helps minimise any biases that may arise due to the asymmetries in the studied languages.

Finally, the results could result from intrinsic asymmetries between the languages themselves. Indeed, the complexity of a language's structures, phone sets, or vocabulary size can also contribute to asymmetries in bilingual language learning. For example, a language with a more complex grammar or phonetic inventory may require more time and effort to acquire than a language with simpler structures. Similarly, a language with a larger vocabulary may take longer than another language with a smaller vocabulary to be fully mastered. These asymmetries, however, are difficult to calculate in real-life settings due to various factors, such as diverse language exposure and variations in input quality, which can skew the calculations.

Nonetheless, our study's primary focus is examining general bilingualism rather than individual language differences. Consequently, we can focus on the general results from the symmetrically tested designs. Despite such asymmetries, one lesson we should take from this study and most studies in this thesis, is the need for caution when drawing conclusions. In psychology, it is rare to see studies with crossed-linguistic designs. Therefore, we potentially conclude too quickly when, in fact, the effects may not necessarily exist or may be artefacts of the design. It is crucial to consider all potential sources of bias and meticulously design experiments to address them to draw accurate conclusions.

Appendix D

The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling

In this appendix, we present work carried out primarily with co-first author Tu Anh Nguyen, in the form of a paper:

Nguyen, T. A.*, **de Seyssel, M.***, Rozé, P., Rivière, M., Kharitonov, E., Baevski, A., Dunbar, E., & Dupoux, E. (2020). The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *Neurips Workshop on Self-Supervised Learning for Speech and Audio Processing.*

This paper presents the initial baseline models and metrics which were used for the 2021 edition of the Zero-Resource Speech Challenge (Spoken Language Modelling track). The track is accessible at the following url: `https://zerospeech.com/tasks/task_4/tasks_goals`.

The Zero Resource Speech Benchmark 2021: Metrics and baselines for unsupervised spoken language modeling

Tu Anh Nguyen[*]
Facebook AI Research & EHESS,
ENS-PSL Univ., CNRS, INRIA, France
nguyentuanh208@gmail.com

Maureen de Seyssel[*]
EHESS, ENS-PSL Univ., CNRS, INRIA
& U. Paris, France
maureen.deseyssel@gmail.com

Patricia Rozé
ENS-PSL Univ., CNRS
France
patricia.roze@ens.fr

Morgane Rivière
Facebook AI Research
France
mriviere@fb.com

Evgeny Kharitonov
Facebook AI Research
France
kharitonov@fb.com

Alexei Baevski
Facebook AI Research
France
abaevski@fb.com

Ewan Dunbar[†]
U. Paris Diderot, France
& U. Toronto, Canada
ewan.dunbar@utoronto.ca

Emmanuel Dupoux[†]
Facebook AI Research & EHESS,
ENS-PSL, CNRS, INRIA, France
emmanuel.dupoux@gmail.com

Abstract

We introduce a new unsupervised task, spoken language modeling: the learning of linguistic representations from raw audio signals without any labels, along with the Zero Resource Speech Benchmark 2021: a suite of 4 black-box, zero-shot metrics probing for the quality of the learned models at 4 linguistic levels: phonetics, lexicon, syntax and semantics. We present the results and analyses of a composite baseline made of the concatenation of three unsupervised systems: self-supervised contrastive representation learning (CPC), clustering (k-means) and language modeling (LSTM or BERT). The language models learn on the basis of the pseudo-text derived from clustering the learned representations. This simple pipeline shows better than chance performance on all four metrics, demonstrating the feasibility of spoken language modeling from raw speech. It also yields worse performance compared to text-based 'topline' systems trained on the same data, delineating the space to be explored by more sophisticated end-to-end models.

1 Introduction

In recent work, self-supervised techniques from vision and NLP have been applied to large datasets of raw audio, giving rise to very effective methods of pretraining for downstream ASR tasks, particularly in the low resource scenario (Schneider et al., 2019; Baevski et al., 2019; Chung and Glass, 2019; Baevski et al., 2020b; Rivière et al., 2020; Kawakami et al., 2020; Wang et al., 2020). The approaches based on transformers and masking objectives, strikingly similar to the models used to train language models, are especially intriguing. The fact that these approaches yield excellent ASR performance (less than 10% WER) with as little as 10 minutes of labels plus a language model (LM), or with 10 hours of labels but no LM (Baevski et al., 2020b), suggests that these systems may actually go beyond acoustic modeling, learning their own LM from raw audio. Such work therefore connects with

[*]Equal contribution as first authors. [†] Equal contributions as last authors.

Self-Supervised Learning for Speech and Audio Processing Workshop @ NeurIPS 2020.

Table 1: **Summary description of the four Zero Resouce Benchmark 2021 metrics**. The metrics in light blue use a pseudo-distance d between embeddings (d_h being from human judgments), the metrics in light orange use a pseudo-probability p computed over the entire input sequence.

Linguistic level	Metrics	Dataset	Task	Example
acoustic-phonetic	ABX	Libri-light	$d(a,x) < d(b,x)$? $a \in A, b \in B$, $x \neq a \in A$	within-speaker: $(\text{apa}_{s_1}, \text{aba}_{s_1}, \text{apa}_{s_1})$ across-speaker: $(\text{apa}_{s_1}, \text{aba}_{s_1}, \text{apa}_{s_2})$
lexicon	spot-the-word	sWUGGY	$p(a) > p(b)$?	(brick, blick) (squalled, squilled)
lexical semantics	similarity judgement	sSIMI	$d(a,b) \propto d_h(a,b)$?	(abduct, kidnap) : 8.63 (abduct, tap) : 0.5
syntax	acceptability judgment	sBLIMP	$p(a) > p(b)$?	(dogs eat meat, dogs eats meat) (the boy can't help himself, the boy can't help herself)

research into the *zero resource* setting, which aims at learning linguistic representations from scratch for language with little or no textual resources. However, up to now, there exists no established benchmark to analyse the representations learned by such models beyond the acoustic/phonetic level.

Typically, language models trained from text are evaluated using scores like perplexity. Unfortunately, this simple approach cannot be used here, since perplexity scores computed from learned discrete units vary according to granularity, making model comparison impossible. This is why we chose to follow a black-box NLP strategy: our metrics do require expert linguistic labels for the dev and test sets, but are *zero-shot* in that they do not require training a classifier, they use *simple tasks* enabling direct human/machine comparison, and they give *interpretable* scores at each linguistic level. As seen in Table 1, they can be divided into two types: distance-based and probability-based metrics. Distance-based metrics require models to provide a pseudo-distance computed over pairs of embeddings. The ABX score (Schatz et al., 2013), already used for the evaluation of *acoustic/phonetic* representations, falls in this category and provides a measure of how well separated phonetic categories are in a given embedding space. Here, we use the ABX score developed in Libri-light (Kahn et al., 2020). Distance-based methods can also be used to evaluate the *semantic* representation of words, by computing the correlation between these distances and human semantic similarity judgements (see Schnabel et al., 2015; Faruqui et al., 2016). Chung and Glass (2018) adapted this metric to speech, which we compiled into our sSIMI dataset. Probability-based metrics require models to compute a pseudo-probability for a given test input (non-normalized non-negative number for a given input waveform). The pseudo-probabilities are computed over pairs of inputs, one of which is acceptable in the tested language and the other not. Such methods have been used in NLP to evaluate the *syntactic* abilities of language models, by comparing the probabilities of grammatical versus ungrammatical sentences(Warstadt et al., 2019), and we built the sBLIMP dataset upon this work. Finally, in our sWUGGY dataset, we extend this logic to the *lexical* level by comparing the pseudo-probability associated to words and nonwords. The four metrics are presented in more details in Section 3.2.

Next, we apply these metrics to a simple baseline system (Section 3.3), built on contrastive pretraining (Contrastive Predictive Coding, CPC, van den Oord et al., 2018; Rivière et al., 2020), followed by k-means clustering, which we use to decode a speech dataset (LibriSpeech, Panayotov et al., 2015) into pseudo-text. This pseudo-text is used to train a language model varying in compute budget: an LSTM (smaller budget) or BERT (larger budget) model. We show (Section 4) that such simple baseline models give better than chance performance on all 4 metrics, demonstrating that it has learned representations at the four corresponding linguistic levels. However, comparison with a text-based BERT topline system trained on the phonetic transcription of the same training data shows that the speech input raises challenges for the LM component of the model that need to be addressed in further work. Datasets and baselines will be open sourced to encourage bridging the gap between speech and text-based systems.

2 Related work

Zero Resource Speech Challenge Series. Previous work (Versteegh et al., 2016; Dunbar et al., 2017, 2019, 2020) has focused on establishing benchmarks for unsupervised learning of an entire dialogue system, but has so far remained at a rather low level (acoustic, lexical). Acoustic modeling

has used two metrics: ABX, a distance-based metric to be discussed later, and opinion scores on TTS output (whereby the discovered units are used to resynthesize speech). As for the lexical level, past work has focused on using the NLP metrics developed for word segmentation (Ludusan et al., 2014). However, these metrics assume that the models should discover words explicitly. The success of character-based language models suggests that it is possible to learn high-level linguistic concepts without explicitly segmenting words (see Hahn and Baroni, 2019).

Black box NLP. Among the variety of black-box linguistic tasks, psycholinguistically-inspired ones enable direct comparison of models and humans. Grammaticality judgments for recurrent networks have been investigated since Allen and Seidenberg (1999), who use closely matched pairs of sentences to investigate grammatical correctness. This approach has recently been adopted to assess the abilities of RNNs, and LSTMs in particular, in capturing syntactic structures. For instance, Linzen et al. (2016) and Gulordava et al. (2018) use word probes in minimally different pairs of English sentences to study number agreement. To discriminate grammatical sentences from ungrammatical ones, they retrieve the probabilities of the possible morphological forms of a target word, given the probability of the previous words in the sentence. Practically, in the sentence "the boy is sleeping", they assume the network has detected number agreement if $\mathbf{P}(w = is) > \mathbf{P}(w = are)$. This methodology has also been adapted by Goldberg (2019) to models trained with a masked language-modeling objective. Similarly, Ravfogel et al. (2018) use word probes to examine whether LSTMs understand Basque agreement and Godais et al. (2017) to test the lexical level in character-based LM.

3 Methods

3.1 Training set

We used as a training set the LibriSpeech 960h dataset (Panayotov et al., 2015). We also included in this work the clean-6k version of the Libri-light dataset (Kahn et al., 2020) which is a huge collection of speech for unsupervised learning. A phonetic transcription of the LibriSpeech dataset was also employed. To obtain this, we used the original LibriSpeech lexicon, as well as the G2P-seq2seq toolkit[2] to generate the phonetic transcriptions of words lacking from the lexicon. We generated a forced-alignment version of Librispeech using the abkhazia library[3]. This enabled us to provide comparative text-based topline systems along with the speech baseline.

3.2 Metrics

We set up four metrics with their accompanying datasets, to evaluate the sLMs at four levels: phonetic (the Libri-light ABX metrics), lexical (the sWUGGY spot-the-word metrics), syntactic (the sBLIMP acceptability metrics) and semantic (the sSIMI similarity metric). The 4 datasets are composed of speech sounds extracted from LibriSpeech (sSIMI), or synthetic stimuli constructed with the Google API[4] using 4 different voices, two males and two females (sWUGGY, sBLIMP, sSIMI)[5]. When synthetic, the stimuli were subsequently force-aligned to retrieve the phonetic boundaries. The datasets containing words or sentences were filtered to only contain the LibriSpeech vocabulary, and are split into dev and test sets.

Phonetics: Libri-light ABX metrics. The ABX metric consists in computing, for a given contrast between two speech categories A and B (e.g., the contrast between triphones 'aba' and 'apa'), the probability that two sounds belonging to the same category are closer to one another than two sounds that belong to different categories. Formally, we compute an asymmetric score, with a and x, different tokens belonging to category A (of cardinality n_A) and b belonging to B (n_B), respectively:

$$\hat{e}(A,B) := \frac{1}{n_A(n_A-1)n_B} \sum_{\substack{a,x\in A \\ x\neq a}} \sum_{b\in B} \left[\mathbb{1}_{d(b,x)<d(a,x)} + \frac{1}{2}\mathbb{1}_{d(b,x)=d(a,x)}\right] \quad (1)$$

[2] https://github.com/cmusphinx/g2p-seq2seq

[3] https://github.com/bootphon/abkhazia

[4] https://cloud.google.com/text-to-speech

[5] We use WaveNet voices A, C, D and F. All dev set stimuli are synthesised in all four voices. Stimuli in the sSIMI and sBLIMP test sets are split evenly among the four different voices, and sWUGGY uses all four for each test set stimulus.

The score is symmetrized and aggregated across all minimal pairs of triphones like 'aba', 'apa', where the change only occurs in the middle phoneme. This score can be computed within speaker (in which case, all stimuli a, b and x are uttered by the same speaker) or across speaker (a and b are from the same speaker, and x from a different speaker). This score requires a pseudo-distance between acoustic tokens computed by averaging along a dynamic time warping path a framewise distance (KL or angular distance). This metric is agnostic to the dimensionality of the embeddings, can work with discrete or continuous codes, and has been used to compare ASR speech features (Schatz, 2016). Here, we run this metric on the pre-existing Libri-light dev and test sets, which has been already used to evaluate several self-supervised models (Kahn et al., 2020; Rivière et al., 2020).

Lexicon: sWUGGY spot-the-word metrics. We built on Godais et al. (2017) which used the 'spot-the-word' task. In this task, networks are presented with a pair of items, an existing word and a matching nonword, and are evaluated on their capacity to attribute a higher probability to the existing word. The spot-the-word metric corresponds to the average accuracy of classifying the words and nonwords correctly across each pair.

The nonwords are produced with WUGGY(Keuleers and Brysbaert, 2010), which generates for a given word, a list of candidate nonwords best matched in phonotactics and syllabic structure. Because we were aiming at speech stimuli, we needed additional constraints to ensure that (i) the audio synthesis of the pairs would be of good quality, and (ii) that the pairs would have matching unigram and bigram scores relative to their phonemes. On a sample of 100 word/nonword pairs, and with feedback from a native English speaker informant, we designed a synthesis-quality rule. The rule consists of testing whether the original phonetic transcription matches the output of a back-to-back phoneme-to-grapheme (p2g) and grapheme-to-phoneme encoding (g2p).[6] Only pairs where both the words and nonwords passed this test were kept. We added additional constraints using a stochastic sampler to also match unigram and bigram phoneme frequencies (see Supplementary Material A). The final sWUGGY test and development sets consists of 20,000 and 5,000 pairs respectively, with the existing words being part of the LibriSpeech train vocabulary. We also prepared additional OOV-sWUGGY test and development sets consisting of 20,000 and 5,000 pairs respectively, with existing words which do not appear in the LibriSpeech training set.

The spot-the-word accuracy is the average of the indicator function $1_{PP(word_k)>PP(nonword_k)}$ over the set of pairs $(word_k, nonword_k)$, where *PP* is a pseudo-probability (a possibly non-normalized non-negative number) assigned to each input file by the model.

Syntax: sBLIMP acceptability metrics. This part of the benchmark is adapted from BLIMP (Warstadt et al., 2019), a dataset of linguistic minimal sentence pairs of matched grammatical and ungrammatical sentences. Similarly to the preceding test, the task is to decide which of the two members of the pair is grammatical based on the probability of the sentence. We adapted the code used to generate the BLIMP dataset (Warstadt et al., 2019) in order to create sBLIMP, specifically tailored for speech purposes. In BLIMP, sentences are divided into twelve broad categories of syntactic paradigms. These categories are themselves divided into 68 specific paradigms containing 1000 sentence pairs each, automatically generated using an expert hand-crafted grammar (this includes an additional subcategory which was added to the code subsequent to Warstadt et al. (2019).

To make this dataset 'speech-ready,' we discarded five subcategories and slightly modified the grammar for nine additional subcategories in order to ensure sentences had appropriate prosodic contours. We also removed from the vocabulary all words absent from the LibriSpeech train set (Panayotov et al., 2015), as well as compound words and homophones that could cause further comprehension issues once synthesised. 5000 sentence pairs were then generated for each of the 63 remaining subcategories. We sampled sentence pairs from the generated pool to create a development and a test set, ensuring that the larger linguistic categories were sampled so as to balance the n-gram language model scores (see Supplementary Material A). The test and development sets contain 63,000 and 6,300 sentence pairs respectively, with no overlap in sentence pairs. Stimuli were then synthesized and force-aligned as described at the beginning of the section.

Similar to the spot-the-word metric, the acceptability judgment metric requires a pseudo-probability for each given input file. The sentence acceptability accuracy is reported similarly to the spot-the-word accuracy with the pairs of grammatical and ungrammatical sentences in the sBLIMP dataset.

[6] We used the G2P-seq2seq toolkit.

Lexical semantics: sSIMI similarity metrics. Here, the task is to compute the similarity of the representations of pairs of words and compare it to human similarity judgements. Based on previous work (Chung and Glass, 2018), we used a set of 13 existing semantic similarity and relatedness tests to construct our similarity benchmark. The similarity-based datasets include WordSim-353 (Yang and Powers, 2006), WordSim-353-SIM (Agirre et al., 2009), mc-30 (Miller and Charles, 1991), rg-65 Rubenstein and Goodenough (1965), Rare-Word (or rw) (Luong et al., 2013), simLex999 (Hill et al., 2015), simverb-3500 (Gerz et al., 2016), verb-143 (Baker et al., 2014) , YP-130 Yang and Powers (2006) and the relatedness-based datasets include MEN (Bruni et al., 2012), Wordsim-353-REL (Agirre et al., 2009), mturk-287 (Radinsky et al., 2011), mturk-771 (Halawi et al., 2012). All scores were normalised on a 0-10 scale, and pairs within the same dataset containing the same pair of words but in the opposite order were averaged. Pairs containing a word not in the LibriSpeech train set Panayotov et al. (2015) were discarded.

We selected as a development set the mturk-771 dataset, which was, in preliminary study using character- and word-based LMs, both highly correlated with all other datasets and was large enough to be used as a development set. It was also ensured that no pair from the development set was present in any of the test sets. All other twelve datasets were used as test sets. We then created two subsets of audio files, one synthetic, one natural. For the first, we followed the synthesis and forced alignment procedures described at the beginning of the section. For the second, we retrieved the audio extracts from LibriSpeech corresponding to each word, following the process presented in (Chung and Glass, 2018). The natural subset is therefore smaller than its synthesized counterpart as we had to discard pairs from the test and dev sets which were not present in the LibriSpeech test and dev sets respectively. However, in this natural subset, each word may appear in multiple tokens, providing phonetic diversity; duplicated scores are averaged in the analysis step. The synthesised subset is composed of 9744 and 705 word pairs for the test and dev sets respectively, and the LibriSpeech subset is composed of 3753 and 309 pairs for the test and dev sets.

The semantic similarity score is reported as the Spearman's rank correlation coefficient ρ between the semantic distance scores given by the model and the true human scores in the dataset. Note that in this work all the semantic similarity scores are multiplied by 100 for clarity.

3.3 Models

Baseline models. Our baseline models are a composite of three components: an acoustic model (CPC), a clustering module (k-means) and a language model (LSTM or BERT) varying in size.

The acoustic model is built upon Contrastive Predictive Coding (CPC, van den Oord et al. (2018)), where the representation of the audio is learned by predicting the future through an autoregressive model. In more detail, given an input signal $\mathbf{x}$, the CPC model embeds $\mathbf{x}$ to a sequence of embeddings $\mathbf{z} = (z_1, \dots, z_T)$ at a given rate through a non-linear encoder g_{enc}. At each time step t, the autoregressive model g_{ar} takes as input the available embeddings $z_1, \dots, z_t$ and produces a context latent representation $c_t = g_{\text{ar}}(z_1, \dots, z_t)$. Given the context c_t, the CPC model tries to predict the K next future embeddings $\{z_{t+k}\}_{1 \leq k \leq K}$ by minimizing the following constrastive loss:

$$\mathcal{L}_t = -\frac{1}{K} \sum_{k=1}^{K} \log \left[\frac{\exp\left(z_{t+k}^\top W_k c_t\right)}{\sum_{\tilde{z} \in \mathcal{N}_t} \exp\left(\tilde{z}^\top W_k c_t\right)} \right] \tag{2}$$

where $\mathcal{N}_t$ is a random subset of negative embedding samples, and W_k is a linear classifier used to predict the future k-step observation. We used a PyTorch implementation of CPC[7] (Rivière et al., 2020), which is a modified version of the CPC model that stabilizes the CPC training by replacing batch normalization with a channel-wise normalization and improves the CPC model by replacing the linear classifier W_k in equation (2) with a 1-layer Transformer network (Vaswani et al., 2017). The encoder g_{enc} is a 5-layer 1D-convolutional network with kernel sizes of 10,8,4,4,4 and stride sizes of 5,4,2,2,2 respectively, resulting in a downsampling factor of 160, meaning that the embeddings have a rate of 100Hz. The autoregressive model g_{ar} is a multi-layer LSTM network, with the same hidden dimension as the encoder. For this baseline, we trained two different versions of CPC: CPC-small and CPC-big. Details are given in Table 2.

After training the CPC model, we then train a k-means clustering module on the outputs of either the final layer or a hidden layer of the autoregressive model. The clustering is done on the collection of

[7] https://github.com/facebookresearch/CPC_audio

Table 2: **Characteristics of the baseline acoustic CPC models**. We took the last LSTM layer of CPC-small and the second LSTM hidden layer of CPC-big as inputs to the clustering as they give the best ABX scores (Supplementary Table S1).

Model	CPC configuration		Training data	Input to kmeans
	autoregressive	hidden units		
CPC-small	2-layer LSTM	256	LibriSpeech clean-100	LSTM level 2
CPC-big	4-layer LSTM	512	Libri-light clean-6k	LSTM level 2

Table 3: **Characteristics of the baseline LMs.** L refers to the number of hidden layers; ED, HD and FFD refer to the dimension of the embedding layer, hidden layer, and feed-forward output layer respectively; H refers to the number of attention heads in the BERT case.

Model	Architecture					nb parameters	Train data	Compute Budget
	L	ED	HD	FFD	H			
BERT	12	768	768	3072	12	90M	LS960	48h - 32 GPUs
BERT-small	8	512	512	2048	8	28M	LS960	60h - 1GPU
LSTM	3	200	1024	200	-	22M	LS960	60h- 1GPU

all the output features at every time step of all the audio files in a given training set. After training the k-means clustering, each feature is then assigned to a cluster, and each audio file can then be discretized to a sequence of discrete units corresponding to the assigned clusters. The k-means training was done on the subset of LibriSpeech containing 100 hours of clean speech.

Finally, with the discretized version of the audio files, we train language models on the discretized units. We establish two 'low budget' and two 'high budget' baselines, based on the number of parameters and the compute resources necessary to train them. The high budget used a BERT-based architecture (Devlin et al., 2019) trained either on CPC-small or CPC-big plus k-means-50 pretrained units. The low budget architectures were a two-layer LSTM and a small BERT architecture (see Table 3 for details); they both used the units from the CPC-big pretraining. Following Baevski et al. (2020a), we trained the BERT models with only the masked token prediction objective. We also followed Baevski et al. (2020a) by masking a span of tokens in the input sequence instead of a single token (otherwise the prediction would be trivial to the model as discretized units tend to replicate). We masked M consecutive tokens for each span, where $M \sim \mathcal{N}(10, 10)$, with a total masking coverage of roughly half of the input tokens (spans may overlap). All models were trained on LibriSpeech 960h. The BERT models were trained with a total batch size of 524k tokens, and the LSTM model was trained with a total batch size of 163k tokens. The learning rate was warmed up to a peak value of 1×10^{-5}. All the implementation was done via fairseq (Ott et al., 2019).

The Topline models. For topline comparison, we trained a BERT model on force-aligned phonemes using the gold transcription of the LibriSpeech dataset. We also employed the span masking similarly to the baseline model. In addition to the BERT trained on forced alignments, we also included a BERT model trained on the gold phonetic transcription of the LibriSpeech dataset, with the difference that we only mask one token instead of a span of tokens. For an absolute topline comparison, we used the pretrained RoBERTa large model (Liu et al., 2019), which was trained on 50K subword units on a huge dataset of total 160GB, 3000 times bigger than the transcription of the LibriSpeech 960h dataset.

4 Results

4.1 Libri-light ABX

Computing distances. We used the average angular distance (arccos of the normalized dot product) of the representations along the DTW-realigned path, as used by default in previous challenges (Versteegh et al., 2016; Dunbar et al., 2017, 2019). For our baseline models, we computed the ABX scores over one-hot representations of discretized units of the audio files.

Results. We first ran experiments varying the number of clusters. As seen in Supplementary Table S2, too few or too many clusters gives rise to worse ABX performance, with a sweet spot at 50 clusters, which is the number we retain for the remainder of the paper. In Table 4, we present the result of the ABX for our two models (CPC-small and CPC-big), before and after clustering. One can see that the CPC-big model yields better performance than the CPC-small model (we retain the big

Table 4: **Within and Across Speaker ABX error** (lower is better) on Libri-light dev-clean and -other for two unsupervised models, before and after clustering (1-hot representations).

	within		across	
Embedding	dev-clean	dev-other	dev-clean	dev-other
MFCC	10.95	13.55	20.94	29.4
CPC-small	6.24	8.48	8.17	13.55
+kmeans-50	10.26	14.24	14.17	21.26
CPC-big	3.41	4.85	4.18	7.64
+kmeans-50	6.38	10.22	8.26	14.86

model for the rest of the experiments), and the clustering step yields an increase in error of between 60-100%. Still, the performances are better than for an MFCC representation, with a much more compact code.

4.2 sWUGGY spot-the-word

Computing the pseudo-probability. Given an audio file x, we first discretize x into a sequence of discretized units $q_1...q_T$. Then, following Salazar et al. (2020), we propose the following pseudo-probability score for our BERT models trained with a span-masked token prediction objective:

$$\text{span-PP}_{M_d,\Delta t}(q_1..q_T) = \prod_{\substack{i=1+j\Delta t \\ \lfloor (T-1)/\Delta t \rfloor \geq j \geq 0}} P(q_i..q_{i+M_d}|q_1..q_{i-1}q_{i+M_d+1}..q_T),$$

where M_d is a chosen decoding span size, and Δt is a temporal sliding size. For the LSTM model, we computed the probability of the discretized sequence with the classic left-to-right scoring style obtained by the chain rule: $P(q_1..q_T) = \prod_{i=1}^{T} P(q_i|q_1..q_{i-1})$.

Results. We determined the optimal masking (Supplementary Table S3) to be $\Delta t = 5$ and $M_d = 15$. We kept this setting for all other experiments involving pseudo-probabilities. Table 5 presents the average of the four baseline systems and in Figure S1, the detailed performances of the baseline compared to n-gram controls and toplines. The performance of all four baselines is consistently better than chance and n-gram controls.

4.3 sBLIMP acceptability

Computing the pseudo-probability. We computed the pseudo-probability as in Section 4.2.

Results. The aggregate results are shown in Table 5 and the detailed ones on the best system in Table S4. The results of this test, while above chance are considerably lower than the text-based toplines.

4.4 sSIMI semantic similarity

Computing the distance. We computed the semantic distance between two audio files x and y as the similarity between the two corresponding discretized sequences $q_1^x...q_T^x$ and $q_1^y...q_S^y$. To obtain this, we extracted outputs from a hidden layer of the LM to the two discretized sequences, aggregating them with a pooling function to produce a fixed-length representation vector for each sequence, and computed the cosine similarity between the two representation vectors:

$$d_{SEM}(x,y) = sim\left(f_{pool}\left(h^{(i)}(q_1^x...q_T^x)\right), f_{pool}\left(h^{(i)}(q_1^y...q_S^y)\right)\right),$$

where f_{pool} is the pooling function and $h^{(i)}(\cdot)$ is the output of the i^{th} hidden layer of the LM.

As each word consists of possibly several voices, we averaged the similarity distance over pairs of the same voice for the synthetic subset, and all possible pairs for the LibriSpeech subset.

Results. For each model, we chose the pooling function and the hidden level that give the best score on the dev set, and computed the score on the corresponding test set. The aggregate results are in Table 5, and a detailed layer-by-layer analysis in Table S5. The scores for semantic similarity are overall modest, compared to BERT systems trained on larger units (BPE). However, one can observe that the best layers for semantic similarity occur towards the first third of the transformer, and that max pooling seems to be best. This contrasts with the best layers for acoustic similarity (as indexed by ABX), which occur at the extremities.

4.5 Model comparison

The overall results are in Table 5. They show that the four baseline models are above chance in the four tasks, even low budget ones, although there is substantial variation between tasks. While task at the lexical level is substantially above chance, the syntactic and semantic tasks show room for improvement compared to text-based toplines trained on similar amounts of data.

5 Discussion

Table 5: Overall performance of our baseline and topline models on dev and test sets on our four zero-shot metrics. For baseline models, the k-means training (k=50) was performed on LibriSpeech clean-100h, and the LSTM/BERT models was trained on discretized units of LibriSpeech 960h. For topline comparisons, we included a BERT model trained on the forced aligned frames of LibriSpeech 960h, a BERT model trained on the gold phonetic transcription of LibriSpeech 960h, and a RoBERTa large model pretrained on a text dataset 3000 times bigger than the transcription of LibriSpeech 960h.

		ABX within		ABX across		sWUGGY	sBLIMP	sSIMI	
System	Set	clean	other	clean	other			synth.	libri.
Low budget baseline systems									
CPC-big+km50+BERT-small	dev	6.38	10.22	8.26	14.86	65.81	52.91	3.88	5.56
	test	6.71	10.62	8.41	15.06	65.94	53.02	3.02	0.06
CPC-big+km50+LSTM	dev	6.38	10.22	8.26	14.86	66.13	53.32	4.42	7.56
	test	6.71	10.62	8.41	15.06	66.22	52.89	7.35	6.66
High budget baseline systems									
CPC-small+km50+BERT	dev	10.26	14.24	14.17	21.26	70.69	54.26	2.99	6.68
	test	10.07	14.71	13.45	22.42	70.50	54.61	8.96	-1.55
CPC-big+km50+BERT	dev	6.38	10.22	8.26	14.86	75.56	56.14	6.25	8.72
	test	6.71	10.62	8.41	15.06	75.51	56.16	5.17	1.75
Topline systems									
Forced align BERT	dev	0.00	0.00	0.00	0.00	92.19	63.72	7.92	4.54
	test	0.00	0.00	0.00	0.00	91.88	63.16	8.52	2.41
Phone BERT	dev	-	-	-	-	97.90	66.78	9.86	16.11
	test	-	-	-	-	97.67	66.91	12.23	20.16
RoBERTa large	dev	-	-	-	-	96.58	81.56	32.28	28.96
	test	-	-	-	-	96.25	82.11	33.16	27.82

We introduced the new Zero Resource Speech Benchmark 2021 for spoken language models. It is composed of 4 zero-shot tests probing 4 linguistic levels: acoustic, lexical, syntactic and semantic. We showed that a simple CPC+clustering+LM trained on LibriSpeech can perform above chance on all of these tests, outperforming n-gram models, while being worse than text-based models trained on the same data. This shows both that the spoken LM task is feasible, and that there is room for improvement.

Obvious directions for research include improving the representation learning component, the clustering methods, and the transformer, which have not been particularly tuned for this benchmark. There are also end-to-end models like wav2vec (Baevski et al., 2020b) and other masking systems (Wang et al., 2020) that could be tried in this context. The performance gap between the RoBERTa large system and our toplines trained on LibriSpeech suggest that much is to be gained by increasing the size of the training set, which can be obtained by large unlabelled audio datasets like LibriVox. Finally, even though this benchmark is intended for developing speech technology for low resource languages, significant resources are still required to construct the test sets and metrics (phonetic dictionary, aligned speech, grammar, TTS or trained speakers to make the stimuli). More work is needed to reduce this footprint and scale up this benchmark to languages other than English.

Broader Impact

The metrics developed here may help improve interpretability of unsupervised systems. Research within the Zero Resource setting may help for developing speech technology for low resourced languages, or for languages with no textual resources, which cannot be addressed in the supervised setting. Even for high resource languages, learning a language model from raw speech would help address dialect variation, including minorities, making speech technology more inclusive. Broadening the reach of speech technology might be used to increase the economic dominance of already-large actors if developed with proprietary resources. To minimize this, we engage the community through an open source benchmark.

Acknowledgments

The work for MS, PR and for EDupoux and TAN in their EHESS role was supported by the Agence Nationale de la Recherche (ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute) and grants from CIFAR (Learning in Minds and Brains) and Facebook AI Research (Research Grant). The work for EDunbar was supported by a Google Faculty Research Award and by the Agence Nationale de la Recherche (ANR-17-CE28-0009 GEOMPHON, ANR-18-IDEX-0001 U de Paris, ANR-10-LABX-0083 EFL).

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches.

Joseph Allen and Mark S Seidenberg. 1999. The emergence of grammaticality in connectionist networks. *The emergence of language*, pages 115–151.

Alexei Baevski, Michael Auli, and Abdelrahman Mohamed. 2019. Effectiveness of self-supervised pre-training for speech recognition. *arXiv preprint arXiv:1911.03912*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2020a. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Simon Baker, Roi Reichart, and Anna Korhonen. 2014. An unsupervised model for instance level subcategorization acquisition. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 278–289.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 136–145.

Yu-An Chung and James Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. *arXiv preprint arXiv:1803.08976*.

Yu-An Chung and James Glass. 2019. Generative pre-training for speech with autoregressive predictive coding. *arXiv preprint arXiv:1910.12607*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL*.

Ewan Dunbar, Robin Algayres, Julien Karadayi, Mathieu Bernard, Juan Benjumea, Xuan-Nga Cao, Lucie Miskic, Charlotte Dugrain, Lucas Ondel, Alan W. Black, Laurent Besacier, Sakriani Sakti, and Emmanuel Dupoux. 2019. The zero resource speech challenge 2019: Tts without t

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017.

Ewan Dunbar, Julien Karadayi, Mathieu Bernard, Xuan-Nga Cao, Robin Algayres, Lucas Ondel, Laurent Besacier, Sakti Sakriani, and Emmanuel Dupoux. 2020. The zero resource speech challenge 2020: Discovering discrete subword and word units. In *INTERSPEECH, perception;bootstrapping/modeling;clustering/bootphon*.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.

Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. Simverb-3500: A large-scale evaluation set of verb similarity. *arXiv preprint arXiv:1608.00869*.

Gaël Godais, Tal Linzen, and Emmanuel Dupoux. 2017. Comparing character-level neural language models using a lexical decision task pages 125–130.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint 1901.05287*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically.

Michael Hahn and Marco Baroni. 2019. Tabula nearly rasa: Probing the linguistic knowledge of character-level neural language models trained on unsegmented text. *Transactions of the Association for Computational Linguistics (Accepted)*.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

J. Kahn, M. Riviere, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazare, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, and et al. 2020. Libri-light: A benchmark for asr with limited or no supervision. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

K. Kawakami, L. Wang, C. Dyer, P. Blunsom, and A. van den Oord. 2020. Learning robust and multilingual speech representations.

Emmanuel Keuleers and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior research methods*, 42(3):627–633.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *TACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Bogdan Ludusan, Maarten Versteegh, Aren Jansen, Guillaume Gravier, Xuan-Nga Cao, Mark Johnson, and Emmanuel Dupoux. 2014. Bridging the gap between speech technology and natural language processing: an evaluation toolbox for term discovery systems. In *Proceedings of LREC*, pages 560–567.

Minh-Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. 2011. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, pages 337–346.

Shauli Ravfogel, Francis M Tyers, and Yoav Goldberg. 2018. Can LSTM learn to capture agreement? the case of basque. *arXiv preprint 1809.04022*.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

T. Schatz, V. Peddinti, F. Bach, A. Jansen, H. Hermansky, and E. Dupoux. 2013. Evaluating speech features with the minimal-pair abx task: Analysis of the classical mfc/plp pipeline. *INTERSPEECH*.

Thomas Schatz. 2016. *ABX-discriminability measures and applications*. Ph.D. thesis, Paris 6.

Tobias Schnabel, Igor Labutov, David Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 298–307.

S. Schneider, A. Baevski, R. Collobert, and M. Auli. 2019. wav2vec: Unsupervised pre-training for speech recognition. *arXiv:1904.05862*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Maarten Versteegh, Xavier Anguera, Aren Jansen, and Emmanuel Dupoux. 2016. The zero resource speech challenge 2015: Proposed approaches and results. *Procedia Computer Science*, 81:67–72.

Weiran Wang, Qingming Tang, and Karen Livescu. 2020. Unsupervised pre-training of bidirectional speech encoders via masked reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6889–6893. IEEE.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2019. Blimp: A benchmark of linguistic minimal pairs for english. *arXiv preprint arXiv:1912.00582*.

Dongqiang Yang and David Martin Powers. 2006. *Verb similarity on the taxonomy of WordNet*. Masaryk University.

Supplementary Materials

A Sampling method to balance ngram scores

We describe here our sampling method to balance ngram scores for sWUGGY and sBLIMP datasets. We first show the algorithm that we applied to sWUGGY, then we just modify slightly the algorithm for the sBLIMP dataset.

For sWUGGY, let's assume that we have N words $w_1, \dots, w_N$; and for each word w_i, we have a list of K matching nonword candidates $nw_i^1, \dots, nw_i^K$. We also assume that each word or nonword w has M scores $s_1(w), \dots, s_M(w)$ (this might be unigram/bigram char/phone scores). We aim to choose, for each word w_i, a matching nonword nw_i^* such that the proportion of the pairs where the score of the word is higher than the score of nonword is close to 50% as possible, for each of M scores.

In other words, we want to build a list of word-nonword pairs $L = \{(w_1, nw_1^*), \dots, (w_N, nw_N^*)\}$ such that the objective function

$$\text{obj}(L) = \sum_{m=1}^{M} |\text{accuracy\_of\_score\_m(L)-0.5}| \tag{S1}$$

is as close to zero as possible.

We thus deduce a simple sampling method as follows: We first initialize a list L of chosen pairs of word and nonword. At each iteration, we randomly choose an unchosen word. Then we sample a nonword candidate in the list of matching nonword candidates, update the list with the new pair, and compute the objective function of the new list as given in S1. If the objective increases, we remove this newly added element, and resample a new nonword from the list of candidates. If we encounter all the nonword candidates but cannot find a new pair, we randomly choose a nonword from the list of candidates. We then continue to the next word until all the words are chosen.

We found afterwards that if we sample all the words at the same time, we can obtain an overall score very close to 50%, but then words with high frequency or with short length tended to have higher accuracy than others. We then decided to divide the words into sub-categories by frequency and word length, and then do the sampling on each of the sub-categories, which gives a more balanced score on all the length and frequency levels.

For sBLIMP, the candidates are slightly different. We now have a list of N pairs of grammatical and non-grammatical sentences and we want to choose K pairs among them such that the accuracy of the chosen pairs is as close to 50% as possible as for sWUGGY. We can then use the same sampling method as described above, with the exception that instead of choosing a word and sampling the nonword candidates at each iteration, we sample an unchosen pair in the list of candidates, and add that pair to the chosen list if we succeed to decrease the objective function.

As we also found that there is a huge difference in the accuracy scores of linguistic paradigms, we tried to do the sampling by each sub-paradigm. However, there were still some paradigms for which we were not able to perfectly balance the score.

B Supplementary ABX methods and results

Given two sounds x and y with two sequences of representations $\mathbf{r}^x = r_1^x, \dots, r_T^x$ and $\mathbf{r}^y = r_1^y, \dots, r_S^y$ respectively, the ABX distance between x and y is computed as follows:

$$d_{ABX}(x, y) = \frac{1}{|\text{path}_{\text{DTW}}(\mathbf{r}^x, \mathbf{r}^y)|} \sum_{(i,j) \in \text{path}_{\text{DTW}}(\mathbf{r}^x, \mathbf{r}^y)} sim(r_i^x, r_j^y). \tag{S2}$$

where $sim(x, y)$ is the arc cosine of the normalized dot product between the embeddings x and y.

Table S1 shows the ABX error on Libri-light dev-clean as a function of different hidden layer of the autoregressive network. We found that as long as we have a big autoregressive network, it is generally not the last layer that brings the best phonetic information of the audio file.

Table S1: Within and Across Speaker ABX error (lower is better) on Libri-light dev-clean at different level of the autoregressive network of CPC-small and CPC-big models. Best layer for each model in bold.

	CPC-small		CPC-big			
LSTM layer	1	2	1	2	3	4
within	10.26	**6.24**	9.62	**3.41**	4.65	9.50
across	14.17	**8.17**	14.73	**4.18**	5.40	9.95

Table S2 reports the ABX scores for different number of clusters, we also included multiple-group clustering in our experiences as similar to Baevski et al. (2020a). We found that the best score is obtained with 50 clusters. Using multiple groups do not further improve the quality of the discretized units, this may be due to the fact that we only used one-hot information of the multiple groups (for example, the two codes 26-20 and 26-10 represent two different one-hot units without any correlation).

Table S2: Within and Across-Speaker ABX error rate (lower is better) on the LibriSpeech dev-clean dataset for CPC-small+kmeans (one-hot vectors embeddings) with different number of units (clusterings). Optimal number of clusters in bold.

nunits	20	50	200	500	2000	50 x 2gr	320 x 2gr
within	11.3	**10.3**	12.5	13.4	17.0	12.6	18.3
across	14.5	**14.2**	16.8	19.9	27.2	17.7	27.7

C Supplementary spot-the-word results

Table S3: **Spot-the-word accuracy** (higher is better) on sWUGGY dev as a function of the masking parameters to compute the pseudo-probabilities. The runtime is estimated based on the evaluation time with the base parameters $M_d = \Delta t = 10$. In bold the compromise we selected between accuracy and speed.

M_d	5		10			15			20		
Δt	5	1	10	5	1	15	5	1	20	5	1
scores	59.14	62.59	64.59	68.23	70.85	66.45	**70.69**	72.52	64.38	69.04	71.33
runtime (est.)	× 2	× 10	× 1	× 2	× 10	× 0.66	× **2**	× 10	× 0.5	× 2	× 10

Table S3 investigates the effect of the masking parameters M_d and Δt to the spot-the-word metrics. We found that the way of computing log-probability can greatly influence the evaluation scores. We see that as long as we overlap the masking spans more, the performance is better. In addition, given that we masked spans of $M \sim \mathcal{N}(10, 10)$ tokens during training, the best decoding masking size was found to be 15. Considering the evaluation time, it is theoretically inversely proportional to Δt, and we thus decided to choose $M_d = 15$ and $\Delta t = 5$ for an accuracy and speed trade-off.

Figure S1 shows the performance of the CPC-big system on the BERT-large architecture: they are worse than the toplines but well above chance. We reproduce the frequency effects (more frequent words giving rise to better accuracies) and the length effect (longer words giving rise to better accuracies). This may be due to the fact that the phonetic space is sparser for long than for short words. As a consequence, a short nonword like "tup" could be continued as a real word in multiple ways ("tuple", "tupperware", etc.). In contrast, a long nonword can rarely be salvaged into a word (eg, 'rhanoceros' is a nonword very early on).

D Supplementary grammaticality results

Table S4 shows the detailed results on the various subsets of sBLIMP of our best model. Almost all of the subsets show better than chance scores (11/12), and of the phoneme ngrams controls (11/12),

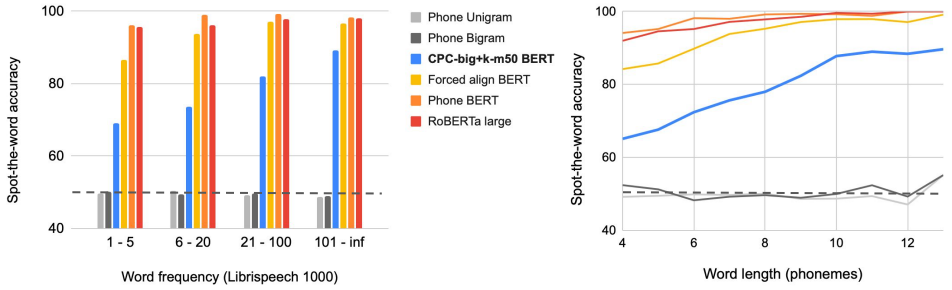


Figure S1: **Spot-the-word accuracy** (sWUGGY dev set, higher is better, chance level at 50%) for our best CPC+clustering+BERT model (blue), compared to phone ngram baselines (gray) and text-based transformer toplines (orange). Left, word frequency effect. Right, word length effect.

and most are better than the word ngrams controls (9/12 for unigram models, and 10/12 for bigram models).

Table S4: **Sentence acceptability accuracy** (sBLIMP dev set, higher is better, chance level at 50%) for our best CPC+kmeans 50+BERT model, compared to phone ngram baselines, text-based transformer toplines, and human scores (from Warstadt et al. 2019).

	Overall	Ana. Agr.	Agr. Str.	Binding	Ctrl. Rais.	D-N Agr.	Ellipsis	Fill. Gap.	Irregular	Island	NPI Li.	Quantifiers	S-V Arg.
Phone Unigram	**48.29**	50.00	50.00	52.90	50.00	50.00	50.00	50.00	45.50	50.00	38.36	39.33	50.00
Phone Bigram	**50.20**	50.50	50.11	52.40	49.80	50.12	50.00	49.88	50.00	49.93	50.00	50.00	50.00
Word Unigram	**54.40**	50.50	50.06	65.20	49.90	50.06	49.50	75.00	51.00	50.00	49.79	50.00	49.92
Word Bigram	**51.64**	50.00	50.06	66.50	50.00	50.06	49.00	50.00	50.00	50.07	50.00	57.00	49.92
CPC-big+km50 BERT	**56.14**	61.50	51.10	62.30	51.62	60.66	74.75	59.91	55.44	56.64	48.29	63.25	51.62
Forced phone BERT	**63.72**	72.62	56.40	63.80	54.90	80.47	69.00	66.34	79.94	58.71	54.29	61.00	65.12
Phone BERT	**66.78**	72.50	59.89	54.40	62.20	92.25	75.00	63.75	82.50	57.71	54.57	81.67	70.17
RoBERTa large	**81.56**	98.50	74.33	80.40	78.20	95.88	99.00	73.62	89.50	68.71	80.71	90.67	87.83
Human (on BLIMP original)	**88.60**	97.50	90.00	87.30	83.90	92.20	85.00	86.90	97.00	84.90	88.10	86.60	90.90

E Supplementary semantic similarity results

Table S5 shows the detailed sSIMI results, layer by layer of the best BERT model together with the detailed ABX results on the same layers. This shows a complementarity of these two metrics (the best layers for acoustics/phonetics are the worst for semantics and vice versa).

Table S5: Comparison of **Semantic similarity scores** (Spearman's correlation with human judgement, higher is better) on the sSIMI synthetic dev set and **ABX scores** on Libri-light dev-clean on different embedding levels of our CPC-big+kmeans50+BERT model. *CPC* refers to the outputs of the second LSTM hidden layer of the CPC-big model, *kmeans* and *outs* refers to 1-hot representations before and after the BERT model repsectively. The semantic similarity scores are also evaluated with different pooling function (mean, max, min). Higher error rates than MFCC baseline in ABX and negative SIMI scores are in red. Note that all the semantic similarity scores are multiplied by 100.

	Score	CPC	kmeans	BERT Layer 0	1	2	3	4	5	6	7	8	9	10	11	12	logits	outs
ABX	within	3.41	6.38	11.82	21.97	35.02	42.54	47.40	44.46	43.71	41.73	33.76	19.67	15.91	15.93	3.30	3.65	5.65
	across	4.18	8.26	13.77	24.59	36.95	43.90	47.94	45.52	44.76	43.12	36.29	23.13	18.92	18.84	4.11	4.59	7.32
sSIMI	mean	-	-	-0.58	-1.97	-1.54	0	1.47	-0.38	1.04	2.26	1.71	2.26	1.47	2.96	-0.57	-	-
	max	-	-	-1.79	0.25	0.51	5.02	6.25	4.03	2.61	1.86	1.69	0.83	1.78	1.78	0.09	-	-
	min	-	-	-3.3	-1.12	-0.93	0.86	6.21	1.9	0.96	0.12	3.53	5.03	0.71	3.41	-0.9	-	-

RÉSUMÉ

La parole, qui est essentielle à l'acquisition du langage, véhicule différents types d'informations. Parmi elles, les informations linguistiques (propres au sens du message communiqué) et indexicales (liées à l'identité du locuteur, dont la langue parlée). Dans cette thèse, nous nous intéressons à la manière dont les nourrissons traitent ces deux types d'informations. Nous explorons de quelle façon les spécificités de l'environnement linguistique d'un nourrisson, en particulier l'exposition à des langues multiples et diverses, façonnent leur perception de la parole. Nous nous demandons également comment la façon dont les informations indexicales sont représentées influence l'apprentissage linguistique. En adoptant une approche de modélisation computationnelle, nous modélisons la représentation des informations indexicales et linguistiques lors de la perception de la parole chez le nourrisson, en tirant parti des avancées récentes en matière d'apprentissage automatique et de traitement de la parole. Par conséquent, nos contributions ont des implications significatives à la fois pour les sciences cognitives et pour le traitement de la parole.

Tout au long de cette thèse, nous modélisons tour à tour la perception indexicale de la parole et la perception linguistique de la parole (qui implique la simulation de l'acquisition du langage) à partir de parole comme seule donnée d'entrée, sous différentes conditions, et en mettant particulièrement l'accent sur l'entrée de parole multilingue. Cette modélisation est passée par le développement de structures et de mesures appropriées pour des simulations d'apprentissage linguistique. Notre travail nous permet de souligner les avantages de la modélisation informatique dans la perception de la parole et l'apprentissage du langage chez le bébé, en fournissant des lignes directrices pour une telle approche. Ces simulations nous permettent également d'éclairer certaines hypothèses sur le traitement de la parole chez les nourrissons en servant de preuves de concept. Nous avons constaté que les mécanismes d'apprentissage statistique étaient suffisants pour simuler l'acquisition précoce du langage chez les nourrissons monolingues. Cependant, bien que nous ayons constaté un apprentissage linguistique avec les mêmes mécanismes avec des données d'entrées bilingues, nous n'avons pas pu reproduire les tendances observées chez les nourrissons bilingues. Ceci pourrait suggérer que ces mécanismes statistiques ne sont pas suffisants dans leur processus d'apprentissage du langage. Pour finir, nous examinons également les implications de notre travail dans le domaine du traitement de la parole, en discutant de l'effet de la distance linguistique et de l'interférence négative.

MOTS CLÉS

traitement automatique non supervisé de la parole; multilingue; sciences cognitives; psycholinguistique; apprentissage machine

ABSTRACT

Speech, serving as a key input in the early language acquisition process, carries different types of information. This includes linguistic information - denoting the inherent meaning of the communicated message - and indexical information - which is tied to the speaker's identity (including the identity of the language spoken). In this thesis, we are interested in how infants process these two types of information. We explore how the specificities of an infant's language environment, particularly exposure to multiple and diverse languages, shape their speech perception abilities. We also question whether the representation of indexical information, and language(s) in particular, can influence linguistic learning. Adopting a computational modelling approach, we model infant speech perception for indexical and linguistic information, leveraging recent advancements in machine learning and speech processing. Consequently, our contributions have significant implications for both cognitive science and speech processing.

Throughout this thesis, we, in turn, model indexical speech perception and linguistic speech perception (which involves simulating early language acquisition) from raw speech input, with varying input patterns and conditions, with a particular focus on multilingual speech input. This modelling became feasible due to our development of suitable frameworks and measures for appropriate learning simulations of speech perception and early language acquisition. Our work allows us to underscore the advantages of computational modelling in speech perception in infants, providing guidelines for such an approach. These simulations also enable us to shed light on some hypotheses behind infants' speech processing by serving as proofs of concept. We found that statistical learning mechanisms were enough to simulate early language acquisition in monolingual infants. However, although we found some linguistic learning with the same mechanisms when given bilingual input, we could not replicate bilingual infants' patterns, potentially suggesting that these statistical mechanisms are not sufficient in their language learning process. We also discuss how our work has implications in the speech processing field, discussing the effect of language distance and negative interference.

KEYWORDS

unsupervised speech processing; multilingual; cognitive science; psycholinguistics; machine learning