Interactions entre locuteurs : de la détection de la parole superposée à la détection des interruptions

Martin Lebourdais

THÈSE DE DOCTORAT DE

De
LE MANS UNIVERSITÉ
Sous le sceau de
LA COMUE ANGERS-LE MANS

ÉCOLE DOCTORALE N° 641
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par
Martin LEBOURDAIS

Interactions entre locuteurs : de la détection de la parole superposée à la détection des interruptions.

Thèse présentée et soutenue à Le Mans, le 17/10/2023
Unité de recherche : Laboratoire d'informatique de l'université du Mans
Thèse N° : 2023LEMA1022

Rapporteurs avant soutenance :

Romain Sérizel	Maître de conférence HDR, LORIA
Ricard Marxer	Professeur, LIS

Composition du Jury :

Président :	Slim Essid	Professeur, Telecom Paris
Examinateurs :	Romain Sérizel	Maître de conférence HDR, LORIA
	Ricard Marxer	Professeur, LIS
	Slim Essid	Professeur, Telecom Paris
	Martine Adda-Decker	Directrice de recherche, CNRS
	Hervé Bredin	Chargé de recherche , CNRS
Dir. de thèse :	Sylvain Meignier	Professeur, LIUM
Encadrant(e)s :	Marie Tahon	Professeur, LIUM
	Antoine Laurent	Professeur, LIUM

Invité(s) :

Laetitia Biscarrat	Maitresse de conférence, LERASS

TABLE DES MATIÈRES

TABLE DES FIGURES

Liste des tableaux

Remerciements

C'est une des parties les plus dures à écrire parce que je suis certain que c'est celle que le plus de monde va lire.

Tout d'abord, un grand merci à ma famille, papa, maman, chez qui j'ai squatté pratiquement tous les week-ends, Timo, José, Laurine, pour les bons repas passés et les cadeaux de Noël de la panique.

Un énorme merci à tous les copain(e)s doctorants ou non. Théo, merci d'avoir été un camarade de chambre au poil en plus d'un super pote (c'est quand tu veux pour les prochaines confs). Valentin et Thibault, tic et tac, merci pour m'avoir supporté pendant plus de 3 ans (8 pour valentin ça fait beaucoup). Albane pour les discussions inutiles pendant trop longtemps (je sais que je t'ai pourri la vie pendant deux ans mais t'as survécu, félicitations). Thibault, l'autre Thibault, qui a dû supporter mes questions de math, à la limite de la philo. Simon, tu as trop souvent servi de carte linguiste pour vérifier la plausibilité de certaines hypothèses. Valentine, pour avoir tenu l'Adoum d'une main de fer pendant ces années (on s'est quand même bien marré) et tu vas la finir cette thèse. Un grand merci aussi à tous les amis que je ne vois pas aussi souvent que je souhaiterais, mais qui répondent toujours quand j'envoie mon message biannuel.

Bon pour ceux qui me connaissent un peu, je n'aurais jamais tenu aussi longtemps sans la musique, donc un autre merci spécialisé pour les musiciens avec qui j'ai eu le plaisir de jouer, Cécile, Camille, Zoé, Patrick, Lucas, Thomas, et tous les autres également (la liste aurait été un peu longue en vrai). Les soirées répétitions lézard auront été mémorables et un super point de décompression dans la semaine. C'est avec plaisir que je reviendrai dès que possible pour rejouer, ou juste faire coucou.

Pour continuer dans l'ordre arbitraire que j'ai fixé, l'ambiance du labo est essentielle pour survivre dans ce monde impitoyable. Ces trois ans auraient été beaucoup moins sympas sans tous les membres qui le composent, c'est donc tout naturel que je vous remercie tous. Un merci tout particulier à Grégor et Etienne pour arriver à réparer mes bourdes régulières, à Anne-Cécile qui a dû gérer des délais (trop) courts pour tout un tas de démarches. Un autre merci également pour Anthony qui, même sans être dans mon encadrement a à de nombreuses reprises été d'une grande aide, scientifique ou non.

Enfin, comme je l'ai dit à pas mal de reprise, une thèse qui marche, c'est au moins 50 % d'encadrement qui marche. Un énorme merci donc, à mes encadrants, Sylvain qui a toujours su poser les questions qui fâchent pour faire avancer ma réflexion, Antoine, qui

même si j'étais assez éloigné des domaines de prédilections a quand même ajouté sa patte et ses retours autant sur le fond que sur la forme (promis, je n'oublierais plus l'espace avant les deux points). Enfin, un immense merci à Marie, qui a été mon interlocutrice la plus proche pour cette thèse, que ça soit pour la définition des pistes d'expériences, les rappels à la réalité quand une piste part trop loin, le forcing pour que je finisse l'expérience jusqu'au bout, même quand ça ne marche pas, la rédaction des articles (il a dû falloir une patience monstrueuse pour ce point-là), ou même la musique à côté.

Au final, j'ai probablement oublié énormément de personnes avec qui j'ai eu des discussions passionnantes, que cela soit des personnes citées plus haut ou des personnes rencontrées en conférences ou en workshop. Ces discussions ont forgé en grande partie mes connaissances/centre d'intérêts/intuitions actuelles et méritent donc d'être mentionnées ici.

Introduction

La différence de traitement entre les femmes et les hommes est un enjeu actuel majeur. Ce problème est autant lié aux incivilités qu'à la représentation des femmes et des hommes dans la sphère publique. Partant de ce constat, le projet ANR Gender Equality Monitoring (GEM)cherche à étudier les différences de représentation et de traitement entre les femmes et les hommes dans les médias de manière transdisciplinaire.

Pour traiter un sujet aussi vaste, cinq laboratoires et deux partenaires industriels sont associés. Ce sujet est ancré dans un objet d'étude social, qui est étudié par trois laboratoires traitant des sujets du genre et des médias, le CARISM [1], le LERASS [2] et l'ENS Lyon [3]. Plusieurs possibilités existent pour traiter ce sujet et peuvent être regroupées grossièrement en deux familles : les méthodes qualitatives et quantitatives. Ces deux formes d'études apportent des points de vue complémentaires et doivent être menées en parallèle pour obtenir les résultats les plus pertinents possibles. Les méthodes qualitatives s'intéressent à un ensemble de données réduit en considérant l'historique entre les interlocuteurs, le rôle des participants, etc. . Les conclusions obtenues qualifient finement l'objet d'étude, mais peuvent être difficiles à généraliser sur de nouvelles données. Parallèlement, les méthodes quantitatives reposent sur le traitement de grandes quantités de données pour en extraire des caractéristiques numériques, incluant l'évaluation des performances des modèles proposés. Ces caractéristiques généralisent bien à de nouvelles situations mais ne représentent qu'une vision réduite du problème. L'objectif du projet GEM est de combiner les méthodes qualitatives et les méthodes quantitatives. Le choix réalisé dans ce projet pour combiner ces deux approches est d'utiliser les méthodes quantitatives pour accélérer l'analyse qualitative d'un grand nombre de données, permettant ainsi une meilleure généralisation des résultats obtenus. Pour traiter cette problématique, les laboratoires d'informatiques travaillant sur ce projet sont le LISN [4] et le LIUM [5] tous deux spécialisés en traitement automatique des langues. Pour finir, le besoin d'un grand nombre de données pour les traitements informatiques est considéré par deux partenaires industriels, Deezer [6], plateforme de streaming musical et Institut National de l'Audiovi-

1. https ://carism.u-paris2.fr/fr
2. https ://www.lerass.com/
3. http ://www.ens-lyon.fr/
4. https ://www.lisn.upsaclay.fr/
5. https ://lium.univ-lemans.fr/
6. https ://research.deezer.com/

suel (INA) [7], institut national de l'audiovisuel qui est à l'initiative de ce projet. L'INA, coordinateur du projet, est un EPIC, c'est-à-dire un organisme de droit privé avec un financement public. Il a pour mission de collecter et de stocker tous les médias diffusés en France depuis sa création en 1975. Cet organisme dispose donc d'un grand nombre de données pertinentes dans le cadre du projet.

Dans le cadre de ce projet, nous souhaitons travailler sur la détection des interruptions de parole. Elle peut avoir de multiples usages. Tout d'abord, nous pouvons imaginer une utilisation par l'ARCOM (ex CSA) pour garantir le respect des tours de parole lors des débats officiels. Une utilisation similaire pourrait être employée par des entreprises souhaitant un meilleur déroulement des réunions en invitant les participants interrompant trop à laisser la parole aux autres interlocuteurs. Ces utilisations traitent alors les interruptions comme une forme de rupture du discours sans prendre en compte les causes de ces interruptions. L'analyse des causes peut difficilement être traitée par une analyse quantitative, car elle nécessite de prendre en compte un contexte très large et très complexe. Par exemple, les membres du projet souhaitent évaluer l'influence du genre sur les interruptions. Nous pouvons par conséquent fournir des zones susceptibles d'en contenir et de laisser les spécialistes de sciences humaines faire une interprétation plus fine.

Problématique

Dans cette thèse, nous nous focaliserons sur des outils de traitement du signal qui faciliteront la caractérisation des représentations des locuteurs. Plus précisément, nous allons étudier la détection d'interruption dans le cadre de conversations issues d'émissions de débats télévisuels et proposer une méthode permettant de fournir des statistiques sur la présence d'interruption. Nous allons également développer des méthodes pour accélérer les études qualitatives sur le sujet qui seront laissées aux chercheurs en sciences des médias et du genre. Au travers de ce sujet, nous nous posons plusieurs questions.

Quelle définition pouvons-nous donner aux interruptions pour que ce phénomène soit détectable automatiquement ?

L'interruption est une notion subjective, dont la définition n'est pas consensuelle. Dans le domaine du traitement automatique, cette tâche est nouvelle, sans cadre et avec peu de ressources. Nous nous posons alors la question de la réduction de la définition d'interruption à celle d'un cas particulier de la parole superposée conformément à la littérature en

7. https ://www.ina.fr/

sociologie et en sciences du langage.

Comment segmenter la parole superposée à partir d'un signal audio ?
La définition de parole superposée permet de mettre en place des modèles de segmentation automatique. Nous posons alors les questions relatives au choix de ces modèles (architecture, apprentissage, évaluation). Les méthodes envisagées sont supervisées et nécessitent des données annotées, il nous faut donc également réfléchir à la provenance de ces données et comment évaluer leur pertinence.

Quelles utilisations peuvent être faites d'un système de détection de parole superposée ?
Par souci de cohérence avec les objectifs présentés plus haut, nous souhaitons également proposer des outils et des méthodes d'analyses pour les évaluations qualitatives à partir de nos systèmes de détection de parole superposée.

Comment détecter les interruptions ?
Dans un corpus composé de données conversationnelles de médias, nous souhaitons savoir dans quelle mesure nous sommes capables de détecter automatiquement une interruption. Comme pour la détection de parole superposée, nos méthodes étant supervisées, nous étudions également les provenances possibles de nos données annotées. Enfin, nous ne cherchons pas à interpréter la signification, ni la cause de ces interruptions. En effet, les problématiques engendrées par la caractérisation des interruptions dépassent de loin le niveau atteignable en trois ans après des études en informatique.

Organisation

Ce document se divise en deux parties principales, une première contenant un état de l'art permettant de préciser la problématique ainsi que de définir les notions utiles à la compréhension de la suite du document, et une seconde partie consacrée à nos contributions réalisées durant cette thèse.

En raison de l'importance de l'aspect sciences humaines de cette thèse, le chapitre 1 est un état de l'art commenté de l'objet d'étude. Il définit dans un premier temps la notion d'organisation de tour de parole. Dans un second temps, cet état de l'art détaille la production acoustique d'un signal de parole, afin de mieux interpréter les résultats de détection automatique aux phénomènes acoustiques. La dernière partie de ce premier chapitre aborde les notions de parole superposée, d'interruption, et enfin de l'influence du genre sur les conversations.

Le chapitre 2 présente un état de l'art du traitement automatique de la parole, élément indispensable à l'étude automatique des conversations. Cet état de l'art débute

sur les méthodes d'apprentissage automatique utilisées en traitement automatique de la parole axées particulièrement sur les méthodes de segmentation par classification. Nous poursuivons ensuite par une description des différentes tâches étudiées, qui sont la détection d'activité vocale, la détection de parole superposée, la détection de genre et enfin la détection d'interruption.

Pour finir cette partie état de l'art, le chapitre 3 fait un point sur les différents corpus qui sont disponibles pour les tâches que nous avons abordées. Nous analysons également dans une seconde partie les annotations associées à ces corpus pour en comprendre les différences et étudier leurs limites.

La seconde partie de ce manuscrit expose les différentes contributions apportées par cette thèse. Le chapitre 4, est séparé en deux parties. La première porte sur l'entraînement des systèmes de détection d'activité vocale ainsi que de détection de parole superposée, deux tâches qui ont été au cœur de nos travaux. La seconde partie propose une analyse plus poussée de nos systèmes pour en améliorer les performances.

Le chapitre 5 suivant analyse les segmentations obtenues pour étudier différents aspects de la détection de parole superposée de façon plus qualitative. Ces analyses permettent de créer des liens avec d'autres disciplines. Pour cela, nous analysons la distribution temporelle des segments de parole superposée, ainsi que les mots contenus dans ceux-ci.

Pour finir, le chapitre 6 traite de la classification d'interruption en commençant par la récolte et l'analyse d'un corpus dédié. En effet, le protocole de recueil de données y est détaillé, en commençant par le choix des données, les classes à annoter et la plate-forme utilisée pour la récolte des données. Les données récoltées sont ensuite manuellement analysées pour déterminer les possibilités d'utilisations. Ce chapitre se termine par l'entraînement d'un système de classification, et l'analyse des résultats de ce dernier.

Ces travaux s'achèvent par la présentation des différentes perspectives envisagées au travers des expériences réalisées, proposant ainsi de nouvelles pistes de travail.

Première partie

État de l'art

Chapitre 1

Interruptions dans les conversations

1.1 La parole conversationnelle

N'étant pas experts en communication ni en linguistique, nous avons dû acquérir un vocabulaire commun avec les différentes communautés qui travaillent sur ces domaines. Définissons donc les termes autour du concept de conversation qui seront nécessaires pour la suite de ce manuscrit. Ce chapitre contient une introduction aux notions de tours de parole et de locuteurs.

1.1.1 Organisation d'une conversation

Définition d'un tour de parole

Une conversation peut être définie par un "Échange de propos, sur un ton généralement familier et sur des thèmes variés, entre deux ou plusieurs personnes" (CNRTL). Cette définition permet d'introduire une distinction importante réalisée en sciences du langage entre parole spontanée, parole préparée et parole lue. Ces trois types de parole peuvent être similaires d'un point de vue acoustique, mais sont entièrement différents quand on considère la construction de cette parole. Pour notre étude, nous souhaitons nous concentrer sur la parole spontanée. Celle-ci peut-être définie comme un "énoncé conçu et perçu dans le fil de son énonciation" [Luz04]. Elle est cependant moins fréquemment rassemblée en corpus que d'autres formes de parole et est par conséquent difficile à étudier seule avec des outils informatiques. Nous allons donc étendre nos travaux à une variation de parole spontanée décrite dans [ÉM10], la parole journalistique dialogale, qui contient des éléments d'une conversation spontanée, tout en étant préalablement préparée par les interlocuteurs.

Affinons à présent les notions de conversation aux travers des travaux de Sacks, Schegloff et Jefferson, sociologues, qui ont formalisé et systématisé la notion de tour de parole

en 1974 dans [SSJ74]. Celle-ci est séparée en deux parties, le *tour* et la spécificité de celui-ci pour la parole. Le mot *tour* implique un échange social avec une organisation définie de manière implicite ou explicite. Un tour peut alors être vu comme un tour de jeu. Ou encore, dans un débat politique, le journaliste modérateur arbitre les échanges en gérant les prises de parole de chacun. Bien que non écrites, les règles du débat politique sont connues et le plus souvent respectées. Comme autre exemple, on peut également citer un croisement routier qui est régi par des règles écrites du Code de la route. Ces règles permettent alors de faciliter la bonne circulation des véhicules.

Un tour peut cependant être régulé par des règles moins définies, par exemple pour les discussions lors d'une réunion, le régulateur de parole est alors la volonté de chacun de laisser parler les autres tout en donnant son avis. L'utilisation de "tours" permet de répartir équitablement une ressource, la parole dans notre cas d'étude, formant ainsi des tours de parole qui facilitent la distribution des temps de parole de manière équitable. Le tour de parole concerne donc cette dernière qui, lors d'une discussion, est nécessaire à l'échange d'informations. La présence de ces tours permet alors de réguler le temps de parole ainsi que de favoriser l'échange d'informations. Nous définirons donc le tour de parole comme étant un ensemble de mots prononcés, d'une longueur variable, permettant à un locuteur d'exprimer un propos. Cet ensemble de tours de parole forme alors une conversation.

Organisation d'un tour de parole

En repartant des travaux présentés précédemment [SSJ74], nous pouvons relever plusieurs points pertinents à la définition systématique (modélisable par un système) d'une conversation. Premièrement, dans une conversation il existe plusieurs changements de locuteurs. On peut étendre ce raisonnement pour ainsi inférer que dans une interaction dyadique, c'est-à-dire entre deux locuteurs, les tours de parole alternent.

Un second point présenté est qu'il faut considérer la communication humaine comme un processus très développé. Les changements de tour de parole fluides, sans pause ni superposition, sont donc non seulement possible, mais fréquents. Cependant, bien que les tours de paroles soient la plupart du temps mono locuteurs, de très courts passages de parole superposée ne sont pas rares. Cette observation rend la parole superposée à la fois peu présente par rapport à la durée totale d'un enregistrement, mais fréquente si on compte le nombre de tours de parole incluant de la parole superposée. Si l'on ajoute à

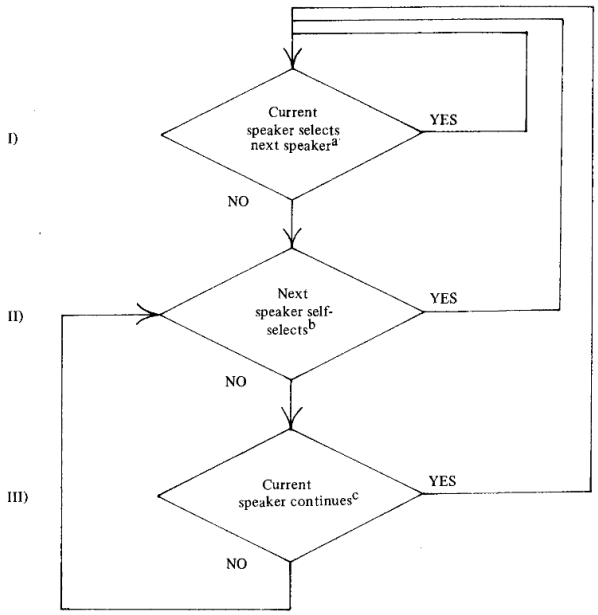


FIGURE 1.1 – Diagramme de flux de l'organisation en tour de parole de Sacks et al. [SSJ74] réalisé par Zimmerman et al. [WZ75]

ces deux types de transitions celles qui contiennent une pause, on obtient pratiquement toutes les transitions d'une conversation. Celles-ci se déroulent à des endroits convenus dans des conversations, appelés Transition Relevant Places (TRP).

L'organisation de ce modèle peut être systématisée par l'automate à états finis créé par Zimmerman et al. en 1975 [WZ75] sur la base des travaux de Sacks [SSJ74] présenté en figure 1.1. Dans le cas de ce modèle, le locuteur actif choisit le prochain locuteur à prendre la parole à partir de signaux dans la parole. Ce choix peut se manifester par des expressions spécifiques comme "qu'en pensez-vous ?", mais il s'agit d'un cas particulier. Souvent un locuteur décide de prendre la parole sans attendre l'autorisation du locuteur précédent. Dans le cas où son tour de parole n'est pas terminé, le premier locuteur peut alors soit s'arrêter, soit continuer à parler, entraînant ainsi une perte de fluidité dans l'enchaînement de la discussion, qui doit alors être réparé pour assurer la bonne continuation de la conversation.

Toutefois, cet article n'est pas le premier à paraître sur l'organisation en tour de parole. Nous pouvons également citer les travaux de Schegloff et al., qui, en 1972 [SGH72], présentent leurs travaux sur les conversations. L'apport majeur de cet article par rapport à ses prédécesseurs est qu'il ne souhaite pas restreindre son étude en considérant une

conversation comme une interaction polie entre deux locuteurs, mais comme un cas de parole spontanée. Cette définition de conversation est dans le cas de notre étude considérée comme assez large pour être utilisée.

Ce modèle n'est cependant pas exhaustif et peut parfois contenir des ruptures que nous appellerons disfluidités. Nous considérons celles-ci de deux manières distinctes. La première est celle définie dans la littérature comme disfluence. Cette notion contient les éléments rompant la fluidité d'un discours. On peut y retrouver les répétitions, les hésitations, les faux-départs, les bégaiements et autres tics de langages. Nous nous intéressons plus aux disfluidités à l'échelle d'une conversation, qui rompent le flux du modèle présenté précédemment. Parmi ces évènements, nous pouvons retrouver les notions détaillées ultérieurement de *parole superposée*, *interruptions*, mais aussi d'autres notions que nous n'aborderons pas telles que les pauses longues entre deux tours de parole.

1.1.2 Locuteur

Enfin, en suivant les discussions sur les possibles implications d'une conversation, celle-ci est invariablement conduite entre plusieurs acteurs [1], appelés locuteurs. Un locuteur ou locutrice est défini.e comme une personne ayant au minimum un tour de parole lors de la conversation. Ce locuteur peut être principal ou récurrent s'il parle beaucoup, ou bien secondaire s'il intervient en même temps qu'un autre, au second plan. Celui-ci peut être l'énonciateur d'un message, en mettant ainsi l'accent sur les questions de prosodie, détaillées en section 1.2.2 ou de contenu de message. Les informations échangées par les locuteurs peuvent également être de multiples modalités. Pour un besoin de cadre et de simplification de l'étude, nous ne nous intéressons qu'au contenu prononcé par le locuteur. Cependant, nous sommes conscients qu'une grande partie de l'information transite de manière non verbale, au travers de gestes ou postures des locuteurs.

Le locuteur peut également être considéré comme une personne présente dans la conversation que l'on peut alors chercher à identifier. Cette identification peut prendre plusieurs formes. On peut vouloir reconnaître un locuteur particulier, par exemple pour des applications liées à la biométrie. On se place alors proche du domaine de la vérification de locuteur. Mais on peut également vouloir en différencier plusieurs, ce qui simplifie le problème. La situation est donc une discrimination de locuteur et non une identification. C'est de cette dernière application que nous sommes le plus proche.

1. Le monologue est considéré hors du cadre de cette étude

Enfin, cette identification d'un locuteur au travers de sa voix nécessite de pouvoir la représenter de manière unique et déterministe au travers d'un modèle de celle-ci.

1.2 Production acoustique

Ferdinand de Saussure (1857-1913), fondateur du structuralisme, mouvement de pensée visant à représenter des notions complexes par des systèmes, définit le langage comme étant une combinaison de la langue et de la parole. Dans cette thèse, nous souhaitons utiliser le langage au sens d'un message perçu et exprimé par des interlocuteurs. Il nous faut alors travailler sur la réalisation instanciée du langage, peu importe la modalité orale ou écrite (parole) et laisser de côté la structure formelle du langage (langue). Pour restreindre le champ d'étude, nous choisissons de nous concentrer sur la modalité orale, ce qui induit de nouvelles questions sur la production vocale.

1.2.1 Production vocale

Cette notion de production vocale entraîne un passage dans les domaines de la phonétique, et plus généralement de l'acoustique. Ce domaine considère la parole comme un phénomène physique suivant un modèle : le modèle source filtre [Mil05] composé de deux parties, une source sonore convoluée par un filtre linéaire. La première partie contient la source du signal. Celle-ci est située dans le larynx et peut-être créé par les différents plis vocaux qui entrent en vibration grâce à l'air expulsé des poumons. Ces plis vibrent à une certaine fréquence que l'on appelle fréquence fondamentale, ou f0.

La seconde partie du modèle correspond à un filtre. Celui-ci est un ensemble de modifications appliquées au signal source pour en rajouter des harmoniques. Dans le cadre de la voix, ce filtrage est réalisé par le canal vocal. Ce canal est composé de plusieurs parties, la gorge, la bouche et le nez. Chacune de ces parties contient des résonateurs, aspérités et muscles permettant la modification du signal source et le contrôle de celui-ci par un humain.

Ainsi, ces filtres permettent de former les différentes voyelles et consonnes utiles à la communication dans les langues occidentales qui seront les seules types de langues étudiées dans le cadre de cette étude. Ces sons peuvent être analysés acoustiquement avec son spectre sonore, représentation intensité/fréquence, obtenue grâce à une transformée de Fourier. Les pics de l'enveloppe spectrale alors obtenus sont appelés formants. Ces pics

peuvent varier en fréquence en fonction de la voyelle prononcée. Celles-ci peuvent être placées dans une représentation appelée triangle vocalique qui représente les voyelles dans un espace défini par la fréquence des deux premiers formants.

Les voyelles sont définies par le passage d'un flux d'air contrairement aux consonnes qui sont définies par une restriction du flux d'air. Celles-ci peuvent être divisées selon plusieurs catégories. On peut tout d'abord les séparer selon leur mode d'articulation, c'est-à-dire la manière de provoquer ce son. En français, on peut distinguer les consonnes nasales, occlusives, fricatives, roulées. L'autre classification possible concerne le point d'articulation de la consonne, quelle partie du canal vocal est mobilisée pour produire le son. Cette classification contient entre autres les consonnes bilabiales, labio-dentales, dentale, vélaire (sur le palais) ou uvulaire (sur le haut du palais).

Ce modèle de la parole peut alors être utilisé pour représenter des phénomènes vocaux. Une partie de ceux-ci peuvent être étudiés par la prosodie.

1.2.2 Prosodie

La prosodie désigne un ensemble de caractéristiques vocales qui se définissent sur une durée plus longue que les sons de la parole. Parmi ces caractéristiques, on trouve l'accentuation, la fréquence, l'intonation, le rythme.

Celles-ci sont étudiées grâce aux éléments acoustiques présentés plus tôt. Par exemple, l'intonation correspond à la variation de fréquence fondamentale (f0) en fonction du temps. Elle est généralement dissociée du contenu linguistique.

Ces éléments peuvent, comme présenté précédemment, être exprimés sous la forme d'une séquence temporelle. Ils peuvent également être réduits au moyen de fonctions statistiques (moyenne, écart-type) au niveau du phonème, du mot, ou de la phrase.

L'analyse, automatique ou manuelle, de ces caractéristiques contribue à la compréhension des phénomènes vocaux. Par exemple, elles sont souvent utilisées par les systèmes de classification d'émotions [Eyb+16 ; Mac+20]

1.3 Les interruptions

1.3.1 Parole superposée

Quand plusieurs locuteurs parlent simultanément, la zone de parole ainsi créée s'appelle une zone de parole superposée. Cette définition, représentée figure 1.2, peut sembler

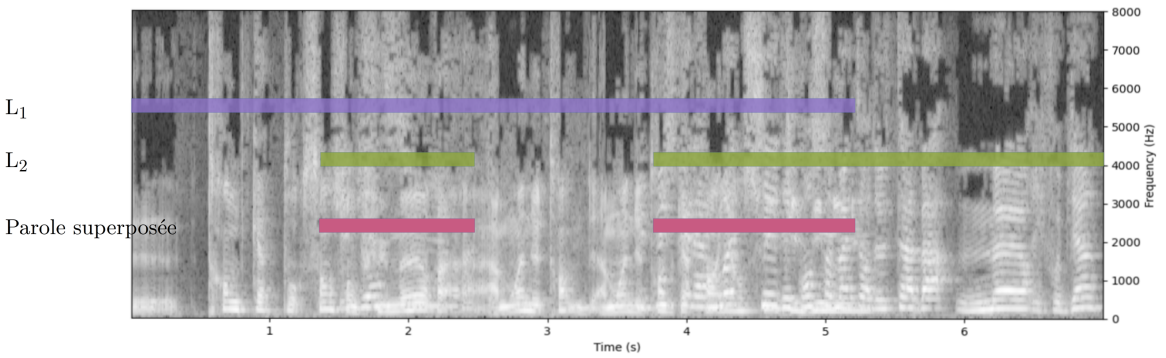


FIGURE 1.2 – Série de parole superposée entre deux locuteurs L1 et L2

intuitive, mais il existe tout de même des points à définir sur la notion d'activité d'un locuteur. La définition de parole superposée est affectée par la méthode utilisée pour l'observer, dans notre cas, des annotateurs humains qui vont écouter et relever ces zones. En effet, un tour de parole est composé de mots possiblement entrecoupés de silences. Il peut donc comporter de très courts silences entre deux syllabes ou deux mots. Les zones de paroles superposées sont donc implicitement définies par la précision de la perception humaine et peuvent être définies par un regroupement de segments dans lesquels deux locuteurs ou plus ont une activité vocale simultanée. Pour pouvoir préciser cette définition, il faut également prendre en compte la signification d'une telle zone. Celle-ci peut être composée de paroles n'appartenant pas à la même conversation. Par exemple, un locuteur au téléphone en arrière-plan d'un reportage peut-il être considéré comme de la parole superposée par rapport au présentateur ? La décision de l'annotateur est alors affectée par l'écart de volume sonore des locuteurs principaux et secondaires. Une traduction simultanée des paroles d'un locuteur crée-t-il alors de la parole superposée, alors qu'ils ne sont potentiellement ni au même endroit ni à la même période temporelle ? La compréhension de la langue de départ a alors une grande importance dans la perception de ce phénomène. Ces deux exemples montrent qu'un point important de la notion de parole superposée vient de la distinction entre parole intelligible et bruit. Ces questions ne peuvent pas avoir de réponse unique, mais doivent être considérées selon le cas d'application et les techniques utilisées.

1.3.2 Définir l'interruption

Cette parole superposée est le plus souvent observée dans les TRP présentées plus tôt. L'étude de ces zones associée à celles des zones de parole superposée permet de nous rapprocher d'une définition d'interruption. Parmi les travaux sur le sujet, nous pouvons citer par exemple ceux de Beattie et al. en 1982 [Bea82] qui ont étudié des interviews de Margaret Thatcher et Jim Callaghan au travers desquelles ils délimitent, afin de mieux interpréter les résultats de détection automatique aux phénomènes acoustiques cinq types de transitions possibles : "Départ anticipé", "Interruption simple", "Changement de locuteur fluide", "Interruption silencieuse" et "Interruption ratée". À partir de ce dénombrement, les auteurs montrent que les interruptions représentent 37.0 % de tous les changements de locuteurs. Une étude précédente des mêmes auteurs [BB79] réalisée avec une méthodologie similaire relève une proportion d'interruption de 10.6 % pour des conversations dyadiques et 6.3 % pour des conversations téléphoniques, montrant ainsi l'importance des interruptions dans les discours politiques (37.0 % contre 10.6 % pour des conversations avec le même nombre de participants). Ces cinq transitions peuvent être regroupées en deux types de transitions, celles avec parole superposée, dont "Départ anticipé", "Interruption simple", et "Interruption ratée" ainsi que celles sans parole superposée, "Changement fluide" et "Interruption silencieuse".

Ce rapport entre parole superposée, départ anticipé, interruption et interruption silencieuse est également présenté dans une étude antérieure par Ferguson et al. en 1977 [Fer77]. Cette étude porte spécifiquement sur la présence d'interruptions dans une zone de silence, donc sans parole superposée, appelées interruptions silencieuses. Les résultats de celle-ci présentent parmi 1452 zones de changement de locuteurs non fluides, 705 zones de départ anticipé (48.55 %), 342 interruptions (23.55 %), 229 interruptions ratées (15.77 %) et 176 interruptions silencieuses (12.12 %). Les interruptions peuvent alors être considérées comme des éléments d'une classification de ruptures du système de tour de parole présenté par Sacks [SSJ74] qui peuvent prendre des formes variées d'un point de vue acoustique. Dans les deux travaux présentés précédemment, la notion d'interruption n'est pas particulièrement décrite mais est considérée comme un élément pouvant être classifié de manière certaine par des annotateurs différents, un phénomène définissable objectivement.

Cette observation a depuis été remise en question par plusieurs études, [Dru89 ; Tan94]. L'étude de Drummond et al. [Dru89] propose une hypothèse sur le modèle alors utilisé d'une interruption comme simple coupure de parole. Après expérience, le modèle existant ne fonctionne pas de manière empirique, la réalité est beaucoup plus complexe et rend le

dénombrement des interruptions impossible sans faire preuve de jugement biaisé. Le seuil de décision pour savoir si une zone comporte une interruption dépend du contexte. Seule une analyse globale de la situation peut alors permettre de prendre une décision, qui sera donc forcément basée sur la sensibilité des annotateurs. Par la suite, nous considérerons donc que la notion d'interruption est subjective.

Cependant, dans le cadre d'une étude automatique, nous devons cadrer au maximum l'objet d'étude. La question la plus intéressante pour nous concerne donc la présence d'une définition locale existante d'une interruption. Une définition locale signifierait une variation déterministe d'un signal acoustique ou linguistique qui indique la présence d'une interruption de manière certaine à l'endroit de cette variation. Le second apport de l'article de Drummond et al. est donc une étude des différentes méthodes de codage des interruptions utilisées dans le texte pour du dénombrement. La première définition présentée par cet article est celle de Wiens en 1965 [Wie+65]. Dans cette définition, toutes les zones de parole superposée sont présentées comme étant une interruption, ce qui ne répond pas à la contrainte de l'hypothèse de départ de présence d'un interrompant et d'un interrompu. La seconde définition est proposée par Meltzer et al. en 1971 [MMH71]. Elle reprend les termes de la définition précédente, en définissant une interruption comme de la parole superposée mais la précise en s'intéressant à la résolution de l'interruption, c'est-à-dire qui va reprendre la parole. Cette définition a l'avantage de se baser sur des caractéristiques observables, telles que l'amplitude vocale pour étudier la résolution de l'interruption. Le problème de la présence de backchannel [H70], courte interjection pour marquer son accord/attention, dans ces zones est éludé en avançant la proportion de ces zones supposée faible. Cependant, comme présenté par [Dru89], cette proportion atteint 60 % des cas de zones de parole superposée, ce qui n'est pas négligeable. La définition de Rogers & Jones [RJ75] tente de contourner le problème en rajoutant l'exception des petits mots aux définitions précédentes. Cette exception n'est cependant pas suffisante en raison de la présence de départ simultané, deux locuteurs qui commencent le tour de parole en même temps, ce qui est assez fréquent, toujours en citant les travaux de Drummond. Enfin, de la difficulté de formalisation de ces interruptions, ce dernier propose d'officialiser la simplification du problème réalisée par les précédents auteurs, et de s'intéresser à la création et résolution de la parole superposée comme substitut d'une interruption.

Cette idée provient en partie de l'étude de Zimmerman et West [WZ75] sur les interruptions. Cette étude se limite à des dyades, supposées plus simples à analyser, et repose sur une transcription manuelle. La définition d'une interruption est encore une zone de

		Annotations (%)			
label	occ	**Backchannel**	**Ajout d'infos**	**Interruption**	**Dep. anticipé**
Backchannel	63	91.1	8.0	1.0	0.0
Ajout d'infos	50	9.2	75.8	15.0	0.0
Interruption	107	0.4	3.6	89.2	6.8
Dep. anticipé	26	0.0	0.0	24.0	76.0

TABLE 1.1 – Distribution des annotations de parole superposée, tiré du travail de Adda-Decker et al. [Add+08], avec comme abréviation, backchannel, ajout d'information complémentaire, interruption et départ anticipé

parole superposée, avec changement de locuteur. La différence présentée entre départ anticipé et interruption tient dans la proximité avec une fin de tour de parole légitime. La notion de la légitimité étant subjective, le problème n'est pas résolu. Comme présenté dans un article plus récent de Guillot [Gui22], la définition d'une interruption claire et objective d'un point de vue local n'est pas possible à formuler ce qui oblige à passer par des chemins détournés pour en détecter des caractéristiques objectives.

En dehors des communautés linguistiques et sociologiques, ce lien entre interruption et parole superposée a également des échos dans la communauté informatique. En 2007, une étude d'Adda et al. [Add+07] présente différentes catégories de parole superposée et une étude de l'impact de ces différents types de parole superposée sur les disfluences observées. Cette étude sépare la parole superposée en quatre classes. Le *backchannel*, présenté plus tôt qui se compose principalement d'onomatopées courtes comme "oui", "mhh". Le *départ anticipé* correspond à un tour de parole débutant avant la fin du tour de parole précédent, mais avec une proximité de la fin attendu de celui-ci jugée suffisante, ce qui est subjectif. Nous nous intéressons principalement à la définition de l'*interruption*. Enfin, l'*ajout d'information complémentaire* contient des interjections destinées à rajouter un élément à l'énoncé du locuteur principal. Par exemple dans le cas d'une interview, le journaliste peut rajouter le nom de famille d'une tierce personne désignée de son prénom par l'invité. Ces quatre classes reposent sur deux dimensions, l'opposition et l'intrusivité. Par exemple, le backchannel a une très faible opposition et une intrusivité faible, alors que l'ajout d'informations complémentaires a une intrusivité forte malgré une opposition faible. Des annotateurs humains ont alors été recrutés pour annoter des segments suivant ces classes, leur accord est résumé dans le tableau 1.1. La distribution des annotations montre que les classes sont relativement consensuelles à l'exception de la différence entre l'interruption et le départ anticipé qui présente une confusion de 24 %. En effet, ces deux

classes semblent proches, la différence étant la présence ou non de signal indiquant la fin de phrase. Cette définition d'interruption n'est toujours pas entièrement objective, mais semble possible à simplifier et par conséquent à automatiser.

1.3.3 Impact du genre sur les conversations

La dernière notion à aborder sur ces interruptions provient directement de l'objet d'étude du projet GEM : le genre. Cette dimension est l'objet des études menées dans le cadre du projet sur l'évolution de la représentation des femmes et des hommes dans les médias. Il est donc nécessaire d'étudier les travaux menés sur l'influence du genre dans ces conversations.

Le point de départ de cet état des lieux est l'article mentionné précédemment de Zimmerman et West de 1975 [WZ75] sur le lien entre interruptions et genre. Cet article traite de la question de la dominance dans une conversation au travers de la présence d'interruption, de silence et d'intervention de soutien du propos entre des dyades mixtes et non mixtes. Le recueil des données est effectué dans des zones publiques où les auteurs pouvaient avoir accès (cafés, parcs), à l'aide d'un enregistreur portable, les participants étant informés de l'enregistrement après celui-ci. Ces données sont donc extrêmement spontanées. Comme présentée précédemment, la définition d'interruption est réduite à la présence de parole superposée, avec un changement de locuteur avant la fin du tour, et obtiennent alors un haut accord inter annotateur entre les auteurs de 93 %. La première observation de l'article est que sans prendre en compte le genre, les interruptions et départs anticipés sont répartis équitablement entre le premier et second locuteur pour les conversations non mixtes. Il faut cependant noter le faible nombre de segments de parole superposée étudiés, soit 29 échantillons pour les conversations non mixtes et 57 zones de parole superposée pour les conversations mixtes, soit un total de 86 zones de parole superposée. Concernant les conversations mixtes, les auteurs trouvent l'intégralité des interruptions et des départs anticipés réalisés par des hommes, à l'exception de deux interruptions réalisées par une femme ayant une position supposée dominante (enseignante et élève). Cette étude a depuis été répliquée par plusieurs études plus récentes [Bro82 ; BS83]. Cependant, d'autres auteurs présentent des résultats opposés [KC83 ; Noh92] en changeant des paramètres de l'expérience. Par exemple, Nohara [Noh92] teste des dyades dans un contexte "de tous les jours" puis dans un contexte de laboratoire et trouve une inversion de la tendance de la présence d'interruption. Dans l'expérience en contexte "de tous les jours", les auteurs trouvent que les femmes interrompent plus les hommes

que l'inverse, contrairement à l'expérience réalisée en laboratoire avec les mêmes dyades. L'hypothèse présentée par Zimmerman et West de la plus fréquente interruption des femmes par des hommes est donc discutable.

De nombreuses études ont été réalisées pour tenter de prouver ou de rejeter cette hypothèse. L'étude de Anderson et al. [AL98] présente une méta-analyse de 43 de ces articles sur le lien entre interruption et genre pour en déterminer les facteurs principaux au travers de méthodes statistiques. Les conclusions de cet article sont que les hommes sont plus susceptibles que les femmes d'interrompre, mais avec une taille d'effet mesurée avec le d de Cohen négligeable (magnitude d'un effet dans une population donnée par rapport à l'hypothèse nulle). Ces résultats montrent que le genre n'est pas le facteur le plus important quant à la présence d'interruption. Cette étude montre aussi que la définition d'interruption choisie est un élément beaucoup plus déterminant sur la répartition des interruptions entre hommes et femmes que le genre des participants. En effet, les études utilisant toute la parole superposée comme interruption sont biaisées vers le nombre d'interruptions réalisées par les femmes, alors que les études n'utilisant que la parole superposée intrusive sont biaisées vers le nombre d'interruptions par des hommes. La définition d'interruption est donc encore une fois primordiale à traiter pour le bon déroulement de ce projet.

1.4 Conclusion

Cet état de l'art a permis de relever plusieurs points d'importance pour la compréhension du sujet. Tout d'abord, la conversation peut être représentée comme possédant une structure simple et modélisable, ce qui en rend possible l'analyse automatique. Grâce au modèle de tour de parole systématique de Sacks [SSJ74], les notions qui en dépendent peuvent, en théorie, être définies ou au moins restreintes par des caractéristiques objectives. Comme présenté dans la section 1.3.2, l'interruption ne dispose pas d'une définition totalement objective et exhaustive, il faudra donc prendre en compte la nature subjective de cette notion, ce qui n'est pas trivial en informatique. La prise en compte de cette subjectivité est cependant possible, comme le montre les travaux menés sur la détection automatique d'émotions qui porte également sur des concepts fondamentalement subjectifs. Nous devons donc simplifier et cadrer le sujet pour aller vers la détection automatique d'interruption en fonction du genre.

Cependant, comme montré dans la section 1.3.3, l'influence du genre semble être dis-

cutée et débattue par des chercheurs et chercheuses bien plus qualifiés que moi sur la question, il n'est donc à mon sens pas pertinent de l'inclure dans l'objectif final. Cette donnée sera donc observée sans être corrélée avec la présence d'interruption. Le soin est laissé aux partenaires souhaitant travailler sur le sujet de corréler l'information de présence d'interruptions et toutes les dimensions qu'ils jugeront appropriées.

L'interruption étant trop compliquée à détecter de zéro, nous décidons de passer, comme proposé par l'article de Zimmerman et al. [WZ75], par une détection de parole superposée, qui pourra par la suite être affinée pour en séparer les interruptions des autres catégories de parole superposée. Cependant, en utilisant cette méthode, nous ignorons consciemment les interruptions qui interviennent hors des zones de parole superposée, ces zones n'intervenant que dans 10 à 15 % des cas. Pour finir, le choix principal de cette étude est de privilégier les données aux résultats. En effet, les notions étudiées sont beaucoup trop complexes d'un point de vue sociologique pour pouvoir simplifier ce problème à un problème de chiffres. Je souhaite donc, au travers de ma participation à ce projet, ne pas fournir de statistiques sur le nombre d'interruptions, mais permettre la facilitation des études de sciences humaines, qui souffrent de volume de données souvent trop restreint pour pouvoir raisonnablement généraliser, en utilisant des méthodes automatiques pour donner des informations supplémentaires sur la présence de zone d'interruption. Cependant, ma thèse reste ancrée en informatique, et je souhaite donc contribuer également à l'avancement de la recherche dans les domaines auxquels je m'intéresserais.

En informatique, tous les éléments traitants du langage rentrent dans la catégorie du traitement automatique de la parole. Ce domaine de recherche contient plusieurs disciplines d'intérêt pour notre tâche. Tout d'abord, nous souhaitons étudier l'organisation des tours de paroles, ce qui revient à étudier la question du "qui parle ? quand ?", définition de la segmentation et regroupement en locuteur. La partie segmentation nous intéresse tout particulièrement pour deux de ses sous-tâches, la détection de parole, qui cherche à savoir si quelqu'un parle, et la détection de parole superposée qui cherche à savoir si plusieurs locuteurs parlent en même temps. C'est donc sur la tâche de parole superposée, brique principale de cette thèse que je baserais la majorité de mon travail. Enfin, bien qu'étant absent de mon objectif final, nous allons étudier la question de la détection de genre avec pour objectif pour fournir un état de l'art des méthodes actuelles aux partenaires sciences humaines du projet, sans pour autant chercher à le faire progresser.

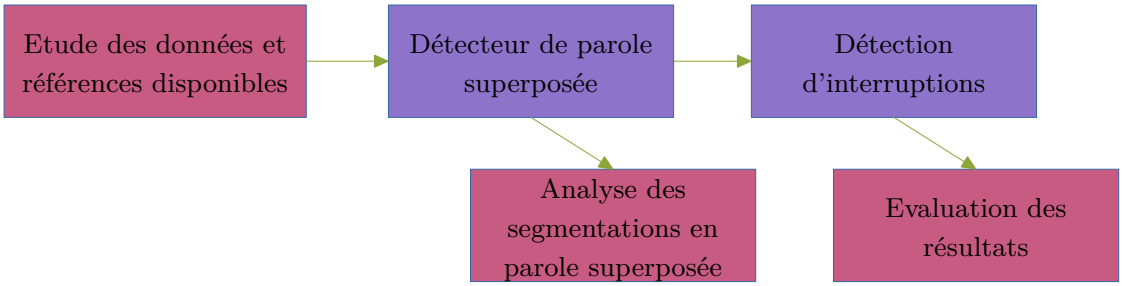


FIGURE 1.3 – Plan de recherche composé des Analyses effectuées et des Modèles développés

Mon projet de recherche peut donc être résumé par le cheminement décrit par la figure 1.3.

Développer un détecteur automatique de parole superposée, analyser ses résultats pour l'améliorer, utiliser la segmentation obtenue comme point de départ au développement d'outils d'analyse du discours ainsi que développer un modèle pouvant classifier ces zones entre interruption ou non-interruption.

Chapitre 2

Segmentation et caractérisation automatique des conversations

2.1 Architectures utilisées pour la classification

Dans la section précédente, nous avons défini la tâche traitée dans cette thèse. Celle-ci est trop compliquée pour être résolue analytiquement par des méthodes de traitement de signal et inscrit donc ces travaux dans le domaine de l'apprentissage automatique. Cependant, ce domaine reste trop vaste et nécessite une précision. Nous travaillons sur le signal audio de parole et allons donc entraîner des modèles automatiques sur des bases de données de signaux de parole. Dans un premier temps, nous aborderons donc les architectures et caractéristiques utilisées pour les problèmes de segmentation par classification que nous développerons dans une seconde partie consacrée plus précisément à nos tâches.

2.1.1 Réseaux de neurones

Le réseau de neurones est un modèle assez ancien [Ros58] tombé en désuétude à cause du manque de puissance de calcul des machines de l'époque. Récemment, l'apprentissage automatique profond a été réutilisé grâce aux avancées en matière d'architecture matérielle et est aujourd'hui omniprésente dans le domaine du traitement automatique de la parole, et plus largement dans l'intelligence artificielle. Le fonctionnement global reste cependant identique à l'idée de base.

Un réseau de neurones est un modèle cherchant à reproduire partiellement les mécanismes d'apprentissage du cerveau humain. En effet, il regroupe des neurones, reliés ensemble par des poids. Ces poids mimiquent les liaisons entre neurones dans le cerveau qui deviennent plus fortes lorsque les neurones sont activés en même temps. L'entrée du réseau de neurones peut être n'importe quelles données multi-dimensionnelles représentant la modalité à modéliser. L'objectif est donc de privilégier l'activation de certains

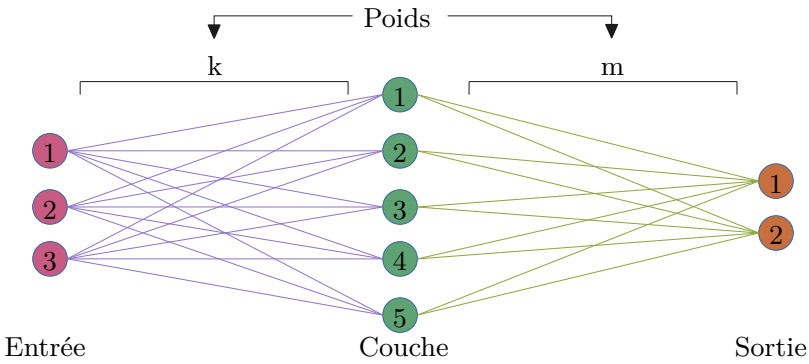


FIGURE 2.1 – Structure d'un réseau de neurone avec une seule couche intermédiaire

neurones pour transformer une entrée donnée en un résultat voulu.

La construction d'un tel réseau est un enchaînement de *briques* appelées couches qui ont plusieurs formes et propriétés pour obtenir un résultat. Concrètement, un réseau de neurones est composé d'une ou plusieurs matrices de poids que l'on cherche à optimiser de telle manière que la combinaison, souvent non linéaire de cette matrice avec l'entrée, donne la sortie souhaitée. Pour cela, on dispose de plusieurs types d'apprentissages. Celui que nous allons utiliser tout au long de cette thèse s'appelle l'apprentissage supervisé. Pour réaliser cet apprentissage, on dispose de références, c'est-à-dire de données auxquelles est déjà associé un résultat considéré vrai. Chaque itération de l'algorithme d'apprentissage se déroule en deux étapes, une passe avant, où l'on prédit une valeur à partir d'entrée et une passe arrière, où l'on corrige le modèle pour améliorer la prochaine itération.

La passe avant consiste en un enchaînement de multiplications de matrices entre différentes couches de neurones pour en extraire des liens. En considérant le réseau présenté en figure 2.1 contenant un vecteur d'entrée noté X, une couche linéaire L liée à l'entrée par une matrice de poids k, et à la sortie par une matrice de poids m. La valeur de la sortie suit l'équation suivante.

$$\begin{aligned} S[o] &= \sum_{j=1}^{5} m[j,o] \times L[j] \\ \text{avec } L[j] &= \sum_{i=1}^{3} k[i,j] \times X[i] \end{aligned} \tag{2.1}$$

Pour permettre de modéliser des phénomènes complexes et non linéaires, des fonctions avec cette dernière propriété sont introduites dans les neurones. Ces fonctions se nomment fonctions d'activations et servent principalement à la non-linéarisation du modèle.

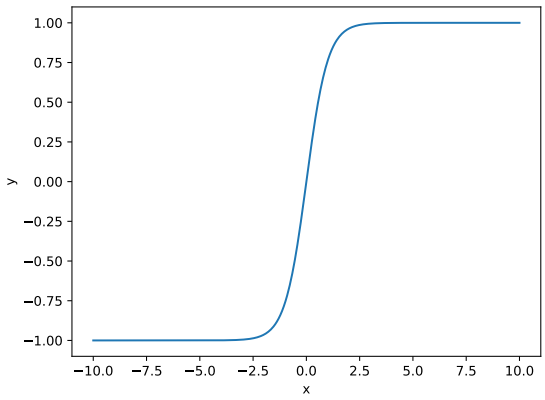


FIGURE 2.2 – Fonction Tangente hyperbolique pour x de -10 à 10

Parmi ces fonctions, nous pouvons trouver la tangente hyperbolique, notée tanh qui est représentée en figure 2.2. L'utilisation d'une telle fonction transforme l'équation 2.1 en l'équation suivante.

$$\begin{aligned} S_o &= \sum_{j=1}^{5} m_{jo} \times tanh(L_j) \\ \text{avec } L_j &= \sum_{i=1}^{3} k_{ij} \times X_i \end{aligned} \tag{2.2}$$

Une fois la valeur prédite, on peut la comparer à la valeur réelle, connue dans le cas de l'apprentissage supervisé à l'aide d'une fonction de coût. Cette fonction fournit un objectif au système : la minimiser. Parmi ces fonctions, celle que nous allons utiliser le plus souvent pour les tâches de classification (détaillées ultérieurement) est la cross-entropy. Cette fonction se base sur la notion mathématique d'entropie définie par Shannon du degré d'incertitude d'une variable aléatoire par rapport à ses possibilités de résultats. L'entropie, notée H peut être formalisée de la manière suivante.

$$H(x) = -\sum p(x) \times log(p(x)) \tag{2.3}$$

En utilisant des pseudo-probabilités prédites par un modèle de classification, nous pouvons les comparer aux valeurs de référence. La fonction de coût notée CE est alors définie de la manière suivante pour n classes avec y la cible et S la prédiction.

$$CE = -\sum_{i=1}^{n} y_i \times log(S_i) \tag{2.4}$$

La minimisation de cette fonction emploie une technique d'optimisation, qui est la descente de gradient. Tout le réseau étant dérivable localement, il est possible de calculer l'influence de chaque poids sur le résultat au moyen d'une dérivée partielle de cette fonction de coût. Avec la responsabilité de chaque poids et le sens de l'erreur, fournis par le gradient, il est possible de mettre à jour les poids pour réduire l'erreur. Cette passe arrière est formalisée en considérant les différentes couches X, L et S respectivement parcourues par des indices i, j, o et reliées par les poids k et m. On prend comme cas la mise à jour des poids de la matrice k.

$$\begin{aligned} k_{ij} &\leftarrow\!\shortmid k_{ij} - \lambda e_j^L X_i \\ \text{avec } e_j^L &= tanh'(L_j) \sum_{o=1}^{2} m_{jo} \times e_o^S \\ \text{avec } e_o^S &= S_o \times CE(softmax(S_o), y_o) \end{aligned} \tag{2.5}$$

Dans la formule précédente nous pouvons voir une fonction softmax appliquée à la couche de sortie S. Cette fonction est obligatoirement présente pour l'utilisation de la cross-entropy et assure la présence de pseudo-probabilités sur la couche de sortie. Cette fonction peut être formalisée de la manière suivante.

$$softmax(S_o) = \frac{e^{S_o}}{\sum_{j=1}^{2} e^{S_j}} \tag{2.6}$$

Le dernier paramètre à présenter dans cette formule est le λ présent dans la mise à jour des poids. Il est appelé *learning rate* et est un coefficient, généralement compris entre 0 et 1 qui modère l'erreur afin de limiter son influence sur les poids du modèle. La formule présentée considère une mise à jour des poids pour chaque exemple. Le plus souvent, pour limiter l'effet des valeurs aberrantes sur le modèle, l'erreur est moyennée avant de la propager sur une suite d'exemples appelée batch. L'apprentissage d'un réseau de neurones consiste donc en une suite de passes avant et de passes arrière jusqu'à l'arrêt du système suite à des performances satisfaisantes. Il se déroule donc un ensemble d'époques qui correspondent au passage de toutes les données d'entraînements dans le modèle.

Cette présentation regroupe les éléments classiques d'un apprentissage par réseaux de neurones, il existe cependant des extensions à celui-ci telles que ADAM [KB14] ou RMSprop [Dau+15] qui apportent différentes normalisations dans les différentes étapes, sans pour autant modifier le fonctionnement basique. Pour construire un réseau, plusieurs types de couches existent, voyons à présent les couches utilisables pour la segmentation

par classification.

2.1.2 Réseaux récurrents

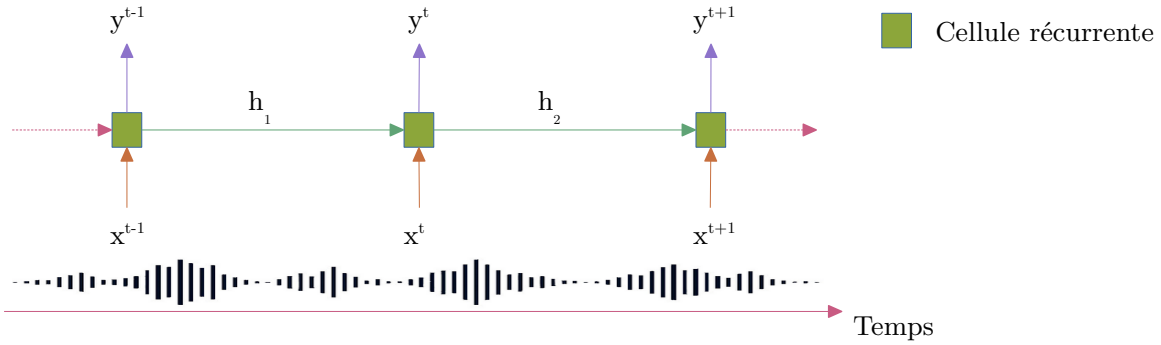


FIGURE 2.3 – Fonctionnement d'un réseau récurrent générique avec des entrées x, des sortie y et un historique h

Un réseau récurrent est un type de modèle qui prend en compte une séquence de donnée dans son ensemble comme présenté dans la figure 2.3. Pour cela, ce type de réseau effectue plusieurs itérations autour d'une couche avec une mémoire des éléments vus précédemment pour conserver un historique. L'utilisation des informations préalables est effectuée dans des blocs appelés cellules récurrentes.

Bien qu'il existe différentes cellules récurrentes, nous nous concentrerons sur la plus utilisée, le Long Short-Term Memory (LSTM) qui, par extension, donne son nom à la couche. L'avantage majeur de ce réseau est de prendre continuellement en compte la dimension temporelle d'une séquence de donnée, en prenant également en compte l'information de la position d'une donnée par rapport à la séquence complète. Le fonctionnement de ce réseau permet de gérer des séquences de données de longueur variable, tant que le nombre de caractéristiques reste stable.

Cependant, comme montré par Bengio et al. [BSF94], ces systèmes ont pour limite l'impossibilité de retrouver des liens entre vecteurs sur de longues séquences.

Ces réseaux sont utilisés principalement pour toutes les tâches nécessitant des séquences de données. Le LSTM est également très utilisé dans sa forme évoluée, le Bidirectional LSTM (BiLSTM). Ce réseau consiste en deux LSTM à la suite en sens inverse, permettant ainsi de conserver le contexte avant la valeur ainsi que le contexte après la valeur.

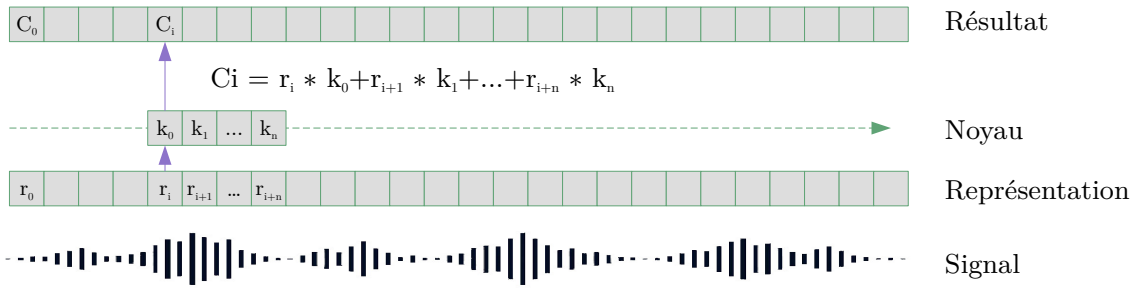


FIGURE 2.4 – Fonctionnement d'une opération de convolution en une dimension avec un noyau de taille n, un vecteur de représentation r, une sortie C et un kernel k

2.1.3 Réseaux convolutifs

Un réseau convolutif est un type de réseau particulièrement efficace pour l'extraction de motifs à partir d'un signal. Il s'agit du même principe fondamental qu'un filtre convolutif utilisé en traitement de signal, aussi appelé filtre à réponse impulsionnelle finie. Le système se base sur l'opération mathématique de convolution, réalisée en informatique de manière discrète, ce qui se rapproche d'une corrélation. Cette opération, présentée en figure 2.4 a deux parties, le signal, qui sert de base et un noyau, appelé kernel qui est appliqué. La convolution d'un signal par un kernel s'effectue par un glissement de ce kernel pour obtenir une valeur dépendant des points connexes, pondérée par un coefficient pour chacun de ces points. En traitement de signal, plusieurs filtres sont connus, par exemple une moyenne glissante pour lisser un signal qui est composée d'un kernel avec toutes les valeurs égales. Le principe d'un réseau neuronal convolutif, appelé Convolutional Neural Network (CNN) est d'apprendre les coefficients du kernel avec une rétropropagation de gradient. Une passe avant par un tel système peut être formalisée de la manière suivante, en considérant C comme étant la sortie attendue, r le vecteur d'entrée de taille m et k le kernel de taille n.

$$C_i = \sum_{j=0}^{n} r_{i+j} \times k_j, \forall i \in [0, m-n[\tag{2.7}$$

Dans l'équation 2.7 le noyau est aligné à gauche sur la représentation. Ce choix entraîne une perte d'au plus n valeurs à la fin de l'entrée. Pour pallier ce problème, une pratique classique est de rajouter du bourrage, aussi appelé padding, pour remplir les valeurs manquantes et pouvoir traiter les dernières données intéressantes.

Ce type de couche est extrêmement efficace pour l'extraction de motifs, chaque motif produisant une valeur assez facilement interprétable par des couches linéaires. Le second avantage est la rapidité de calcul. Ce système comporte peu de poids, les seuls poids devant être appris sont les poids du kernel. Le système n'est également composé que d'une suite de multiplication de matrices, ce qui est particulièrement optimisé grâce aux cartes GPU utilisées en intelligence artificielle.

Ce système a cependant quelques inconvénients. Premièrement, il manque de contexte global, là où le réseau récurrent permet de conserver beaucoup d'informations, le CNN ne reconnaît que des motifs locaux et a du mal à reconnaître des motifs globaux, même quand placé à la suite d'autres couches convolutives. Un autre point à considérer est que le CNN peut apprendre à "tricher" pour classifier des exemples [Bak+18], c'est-à-dire à isoler des signaux inhérents aux méthodes d'acquisition des données pour classifier les exemples, ce qui le rend difficile pour la généralisation. Par exemple, un système de reconnaissance d'image qui confond un loup et un husky par la neige autour de lui.

Le CNN est particulièrement utilisé en traitement d'image [Lec+98]. C'est de cette communauté d'où vient la majorité des nouvelles architectures et qui sont par la suite transférées vers le traitement d'audio, soit en considérant des spectrogrammes comme des images, soit en passant le signal comme dans les méthodes de traitement de signal plus anciennes. Pour résoudre les problèmes de ce réseau, plusieurs architectures sont développées à partir de couches convolutives, dont le TCN.

2.1.4 TCN, Temporal convoluted network

Un Temporal Convoluted Network (TCN), est un réseau convolutif proposé par Bai et al. en 2018 [BKK18]. Les auteurs souhaitent démontrer grâce à un réseau convolutif basique qu'ils appellent TCN que les réseaux récurrents ne sont pas forcément la première architecture à tester dans un cas de modélisation de séquence. Cette famille de réseaux a deux caractéristiques principales. Le réseau doit être causal, c'est-à-dire qu'un élément ne doit dépendre que des éléments passés et il doit pouvoir prendre en entrée des séquences de longueur arbitraire comme le ferait un réseau récurrent. L'efficacité de ce système vient du principe de dilatation des convolutions qui permet d'obtenir des kernels plus grand quand on considère qu'un élément a une influence sur les éléments contigus.

L'avantage principal du TCN tient dans sa clarté d'architecture qui devient beaucoup plus facile à adapter pour différentes tâches. Les auteurs présentent également des points positifs de cette architecture. Tout d'abord, les données ne dépendant pas d'un quelconque

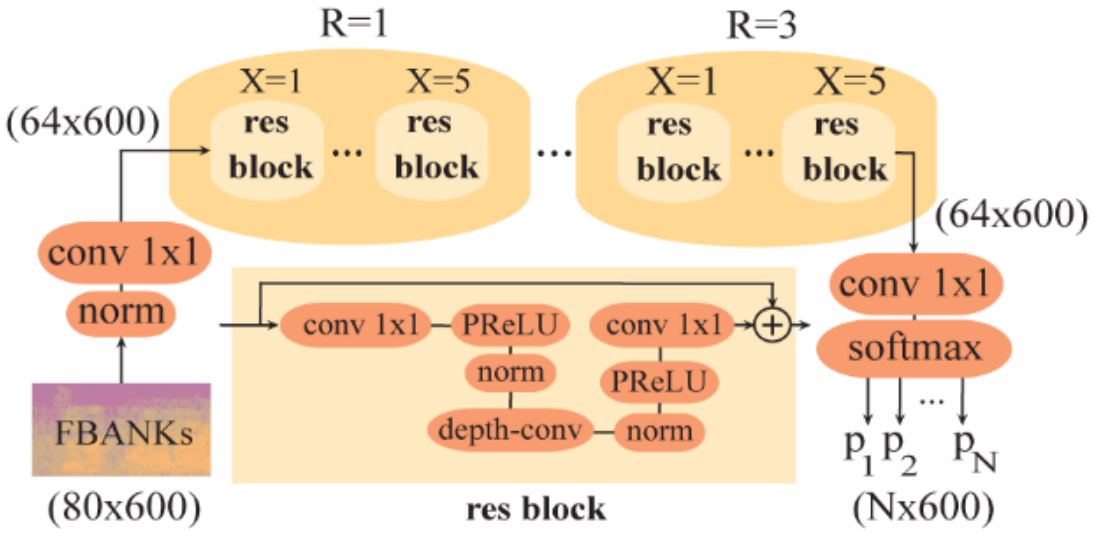


FIGURE 2.5 – Architecture de TCN utilisé par Cornell et al. [Cor+20]

historique, la séquence complète peut être calculée en parallèle contrairement à un LSTM qui doit attendre la fin de l'échantillon précédent pour avoir l'historique. La gestion de la taille des dilatations et de la fenêtre utilisée pour la prédiction est facilement adaptable à une tâche en empilant plusieurs systèmes. Autre avantage de ce système provenant de l'absence de récurrence, n'étant pas récurrent il n'a donc pas les problèmes de disparition ou d'explosion de gradient. Enfin, Le TCN n'a pas beaucoup de paramètres, ce qui le rend efficace à entraîner et utilisable avec peu de données.

Cependant, les auteurs pointent deux inconvénients au système. Premièrement, la manière de gérer le contexte dans un TCN nécessite plus de mémoire qu'un historique de réseau récurrent. Ce qui nécessite des cartes graphiques adaptées ou de gérer la taille d'entrée des données. De plus, la sélection de la taille de l'historique est entièrement laissée à l'utilisateur, ce qui peut poser des problèmes d'adaptation si des tâches, ou corpus nécessitent un historique plus long ou plus court.

L'objectif de ce système est de pouvoir remplacer tous les cas d'usage d'un LSTM avec une architecture plus rapide et plus efficace.

2.2 Extraction et compression d'informations à partir du signal audio

Ces différents systèmes extraient des informations à partir d'un signal de parole. Cependant, le format de fichier (wav) utilisé principalement contient trop d'informations parasites pour permettre l'apprentissage correct d'un modèle [1]. La première étape de toute architecture est donc d'extraire des descripteurs pertinents du signal pour en extraire des informations utilisables par les modèles.

2.2.1 Caractéristiques acoustiques

Les premières caractéristiques utilisées pour représenter un signal audio viennent de l'acoustique et du traitement de signal, adaptées pour les besoins de l'informatique. Elles sont moins utilisées actuellement, mais toujours utiles, car interprétables par un humain et corrélées à une réalité physique.

Descripteurs experts

Pour représenter un signal de parole, une technique efficace est d'extraire et agréger plusieurs caractéristiques acoustiques pour chercher à décrire au mieux le signal. Parmi ces ensembles de caractéristiques, on peut citer eGeMAPS [Eyb+16]. Cet ensemble de caractéristiques développé à Genève contient des éléments dépendant de la fréquence, comme le "jitter" variation de la fréquence fondamentale (F0) entre deux échantillons, le "pitch", logarithme de la fréquence fondamentale placé sur une échelle par demi-ton, ou encore les fréquences des premiers formants, pics de l'enveloppe spectrale du signal, caractéristiques du son prononcé. D'autres éléments sont également présents concernant l'énergie d'un signal, comme le "shimmer", variation d'énergie entre deux pics pour deux F0 consécutives, l'intensité pondérée par la variation de la perception humaine ou encore le rapport entre l'énergie des harmoniques et celle du bruit. Le développement de ces ensembles de caractéristiques nécessite une bonne connaissance à priori de la tâche pour savoir quels éléments pourraient être discriminants. C'est pourquoi, pour profiter de la capacité des modèles à trouver des relations entre descripteurs qu'un humain n'aurait pas considérées, d'autres sortes de descripteurs plus proches du signal sont utilisés.

1. Certains articles présentent des modèles prenant une forme d'onde en entrée [Sai+15] mais nous considérons qu'une extraction de caractéristique à l'aide de convolution reste une extraction de caractéristique.

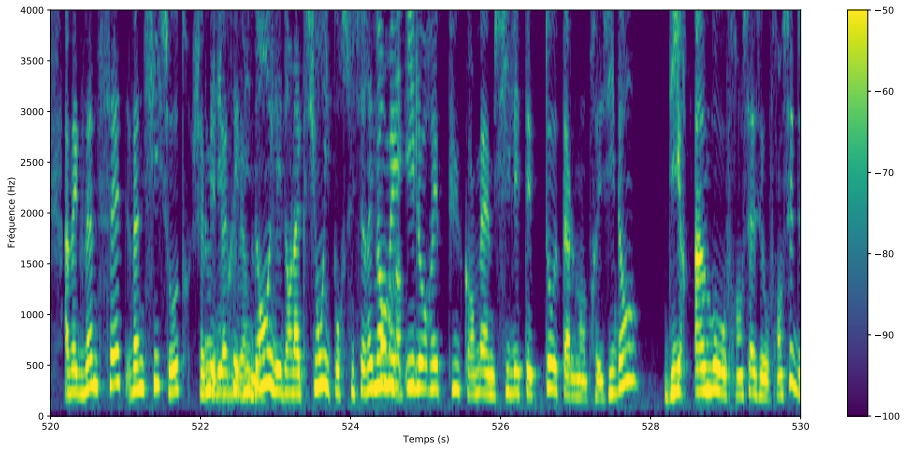


FIGURE 2.6 – Exemple d'un spectrogramme de parole

Spectrogramme

Un spectrogramme est une représentation visuelle de la puissance sonore d'un signal en 2 dimensions. Il représente la puissance sonore d'une bande de fréquence en fonction du temps. La représentation ainsi obtenue comme présentée en figure 2.6 est assez compréhensible pour pouvoir lue directement à partir de la représentation graphique par des personnes entraînées. En informatique, on peut considérer ce type de descripteurs comme un ensemble de vecteurs ou bien comme une image à traiter avec des méthodes de traitement d'image. Ces spectrogrammes sont obtenus grâce à une transformée de Fourier réalisée sur une fenêtre glissante de faible largeur.

Mel-frequency Cepstral coefficients

Les Mel-Frequency cepstral coefficients, ou MFCCs, sont des descripteurs acoustiques basés sur une échelle se rapprochant de la perception humaine, l'échelle Mel. Bien qu'assez peu utilisés par la communauté acoustique, ces coefficients sont extrêmement utilisés en traitement de la parole. Ils sont calculés en prenant la transformée de Fourier discrète du signal sur une fenêtre environ égale à 30 ms avec un pas de l'ordre de 10 ms pour obtenir un spectrogramme, puis en transformant l'échelle Hertz utilisée en échelle Mel. Deux formules de conversion principales existent : "Slaney" qui sépare la bande de fréquence en deux en appliquant une transformation linéaire aux fréquences en dessous de 200 Hz et une échelle

logarithmique au-dessus, et la formule "HTK" [2] qui n'utilise qu'une échelle logarithmique de formule $2595 * log10(1 + \frac{frequences}{700})$. Ces deux formules ont pour objectif de reproduire la meilleure discrimination des fréquences en basse fréquence par l'oreille humaine par rapport aux hautes fréquences. Par la suite, nous utiliserons la version implémentée dans torchaudio qui est par défaut la formule "HTK".

2.2.2 Caractéristiques neuronales

S'éloignant des caractéristiques acoustiques, plusieurs extracteurs d'informations ont ainsi été développés pour construire des représentations compactes d'une dimension, que ce soit une représentation de la langue, ou d'un locuteur à partir d'un signal audio. Ces vecteurs sont souvent éloignés de ceux que l'on peut obtenir par une étude acoustique d'un signal mais contiennent et compressent plus d'informations de haut niveau.

SincNet

Pour garder un lien avec les caractéristiques acoustiques précédemment présentées, SincNet [RB18] est un réseau hybride, dans le sens où il utilise des réalités physiques comme base et en apprend des paramètres à partir d'un réseau de neurones convolutif (détaillé en section 2.1.3). Ce réseau n'a été utilisé que peu de temps, en raison de l'arrivée des systèmes d'extraction de caractéristiques entièrement neuronaux qui présentent des performances supérieures. L'interprétabilité de SincNet est cependant légèrement meilleure que celle des systèmes entièrement neuronaux en permettant d'obtenir des filtres de réponse en fréquence.

Wav2Vec

Wav2Vec [Bae+20] est un exemple d'apprentissage auto supervisé. Ce type d'apprentissage est utilisé pour construire des représentations efficaces à partir de beaucoup de données non annotées. Il vient à l'origine de la communauté de traitement de texte avec la grande popularité des différents modèles BERT [Dev+18] et consiste en un apprentissage d'une représentation du langage comme sous-produit d'une tâche simple ne nécessitant pas d'annotation. La tâche utilisée par Wav2Vec est une tâche de récupération de signal masqué. À partir d'un signal, on cherche à retrouver la valeur d'une partie masquée de celui-ci, apprenant ainsi l'organisation d'un signal de parole.

2. https ://htk.eng.cam.ac.uk/

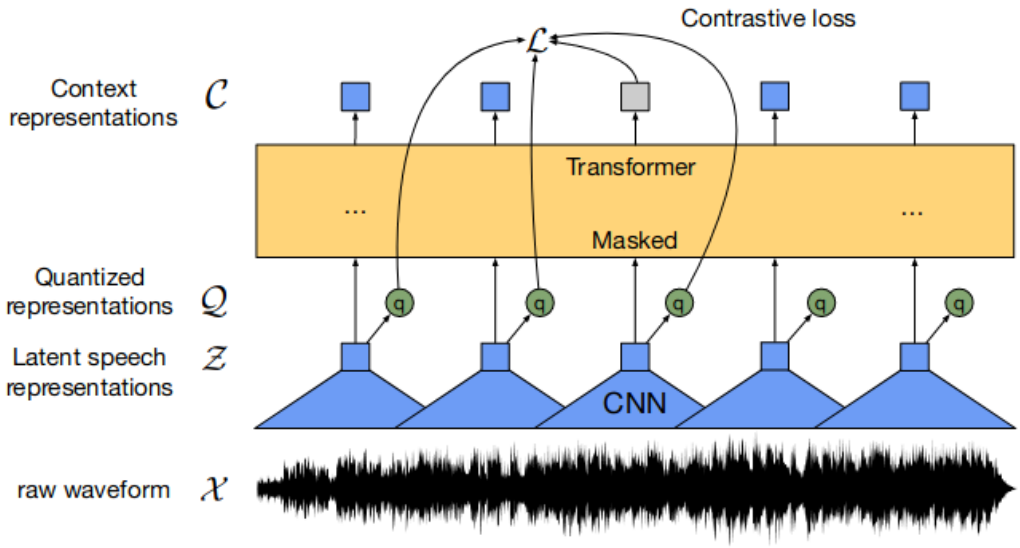


FIGURE 2.7 – Fonctionnement de l'entraînement de Wav2vec, schéma issu de l'article original [Bae+20]

Par rapport au texte, l'audio a un souci principal en l'absence de vocabulaire fini, ce qui nécessite une loss particulière. Celle-ci est une dite contrastive, c'est-à-dire qu'elle doit maximiser la similarité entre la valeur masquée et sa référence et minimiser cette similarité entre la valeur masquée et d'autres, prises dans le signal.

HuBERT

HuBERT [Hsu+21] est un modèle de représentation de la parole qui fonctionne sur le même principe global que Wav2Vec. La principale différence, et apport de HuBERT par rapport à ce dernier, se situe dans la manière de traiter le problème de vocabulaire infini. Comme présenté dans la figure 2.8, HuBERT construit au préalable un "vocabulaire" avec un système de découverte d'unités acoustiques. L'exemple de l'article présente un clustering K-means sur des MFCCs, mais ce n'est pas la seule méthode utilisable. La création de ce vocabulaire permet de considérer le problème comme une classification d'un mot parmi d'autres.

WavLM

Enfin, dans le cas de WavLM [Che+21], cette tâche est également une récupération de signal masqué. Ce système, illustré en figure 2.9 utilise la même méthode de résolution du

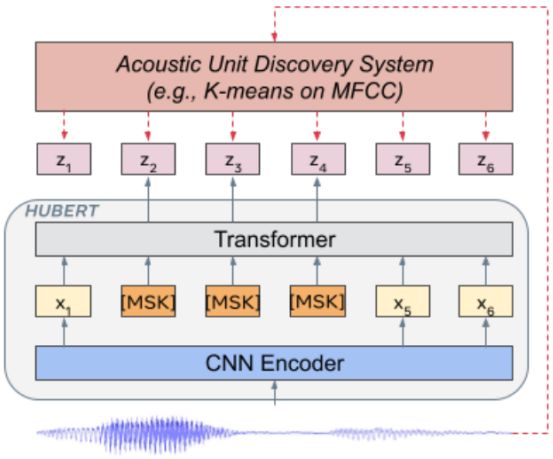


FIGURE 2.8 – Fonctionnement de l'entraînement de HuBERT, schéma issu de l'article original [Hsu+21]

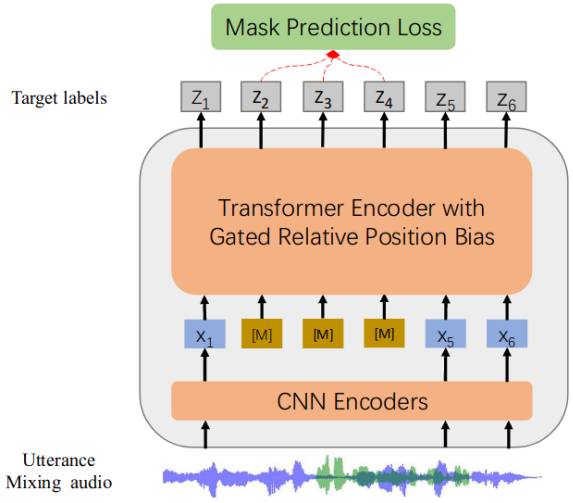


FIGURE 2.9 – Fonctionnement de l'entraînement de WavLM, schéma issu de l'article original [Che+21]

problème de vocabulaire que HuBERT. Les principaux apports de WavLM par rapport à ses prédécesseurs sont la présence de données provenant de différentes sources, ainsi que la présence de parole superposée artificielle et d'un encodage de la position (positional encoding) qui sert à pondérer les parties utiles du signal. WavLM existe en plusieurs versions, différentes par les tailles de couche et le nombre de données d'entraînement. Ce système est très efficace, et est actuellement (au moment de l'écriture de ce paragraphe) le meilleur système de représentation de signaux de parole présenté sur le benchmark SUPERB [Yan+21a] dans toutes les catégories à l'exception de la transcription dans laquelle il arrive second derrière un autre système appelé data2vec Large.

Plusieurs méthodes existent pour traiter les caractéristiques de WavLM. Il est possible de ne considérer que la dernière sortie du modèle pour obtenir la meilleure représentation phonétique ou de moyenner les sorties des différentes couches pour bénéficier de représentations de la parole de différentes granularités. Nous utilisons une moyenne non pondérée, celle-ci n'entrainant qu'une très légère perte par rapport au rajout de poids entraînables. Une vérification de cette affirmation est disponible en annexe A.1.

2.3 Tâches de segmentation en locuteur

2.3.1 Historique

"Ceux qui ne peuvent se souvenir du passé sont condamnés à le répéter", Georges Santayana.
Commençons cet état de l'art des tâches que nous allons aborder par un bref historique du traitement de la parole.

Le traitement automatique de la parole est généralement structuré en tâches applicatives, telles que la transcription, qui s'intéresse au contenu linguistique de la parole mais également la vérification de locuteur, et la segmentation et regroupement en locuteur qui respectivement visent à reconnaître un locuteur, et à savoir qui parle à quel instant. Cette liste non exhaustive peut également contenir d'autres tâches s'intéressant à la compréhension de la parole, à la reconnaissance d'émotions des locuteurs, ou encore à la génération de signal de parole. Nous souhaitons nous consacrer à la caractérisation de la parole prononcée par des locuteurs, nous aborderons donc différentes thématiques. Ces domaines sont en constante évolution avec, comme particularité pour l'intelligence artificielle, la rapidité de cette évolution, trop importante dans le cas des systèmes par rapport à la

temporalité des articles de journaux. La majorité des avancées de ce domaine sont liées des campagnes d'évaluations, qui sont des concours avec un corpus et une méthode d'évaluation commune pour faire un point sur l'état de l'art actuel d'une ou plusieurs tâches précises. La première tâche qui nous intéresse est la segmentation du signal de parole, qui se rattache à la Segmentation et Regroupement en Locuteur (SRL). Nous allons donc voir l'évolution de l'intérêt pour ces tâches au travers de différentes campagnes d'évaluations. Celles-ci sont organisées principalement par le NIST qui est l'institut américain des normes, mandaté par le gouvernement américain pour fournir toutes sortes d'évaluations.

La première occurrence indirecte de la tâche de SRL date de 1996 lors de la campagne d'évaluation Hub4[98]. Cette campagne demandait une transcription continue d'une émission sans proposer de segmentation préalable, c'est à dire de découpage de l'audio en fonction du locuteur présent, en laissant le soin de celle-ci aux participants. Le NIST incite ainsi les équipes à développer leurs propres systèmes de segmentation, bien que ce ne soit pas la tâche principale. Cette campagne d'évaluation a été réitérée avec les mêmes modalités de 1996 à 1998. La première apparition de segmentation en locuteur, proposée seule, intervient en 2000, pour la campagne Nist-SRE (Speaker Recognition Evaluation), les éditions précédentes ayant un découpage déjà réalisé pour identifier la personne qui parle. De 2000 à 2002, une tâche de SRL a donc été incluse dans la campagne consacrée à la reconnaissance de locuteur. À partir de 2002, la campagne change de tâche pour évaluer la transcription en la transformant pour rajouter des informations paralinguistiques. Ce passage montre un changement de l'état de tâche annexe utilisant des techniques similaires de reconnaissance de locuteur, à celui de prétraitement nécessaire à la transcription. La campagne NIST-RTE-02 [Gar+02], Rich transcription Evaluation comporte donc une discipline de *diarization* qui est définie comme deux sous tâches, une première segmentation en tour de parole monolocuteurs, puis un regroupement de ces segments pour les assigner à un locuteur commun. Depuis cette campagne, de nombreuses autres évaluations annexes sont apparues. Tout d'abord, le NIST continue ses challenges en modifiant le nom "Segmentation et Regroupement en Locuteur" en *diarization* qui de 2003 à 2005 inclut la musique et les bruits, pour, de 2005 à 2009 ne contenir que la voix afin de mieux cadrer la tâche à effectuer. C'est également en 2003 que la parole superposée est prise en compte pour la première fois dans l'évaluation, posant ainsi les bases d'un nouveau problème. Il faut toutefois signaler que la parole superposée n'intervient pas dans le calcul de la métrique principale, mais dans une seconde, toujours expérimentale, et cela jusqu'en 2006. Cet organisme n'est pas le seul à traiter de ces tâches. En 2003, 2005 et 2008, les cam-

pagnes ESTER et ESTER 2 [Gal+06] reproduisent le modèle d'évaluation proposé par le NIST pour des données radiophoniques et télévisuelles en français. La campagne ESTER2 de 2008 inclut également une tâche considérée "exploratoire" de détection de parole superposée. Ces campagnes d'évaluation françaises vont continuer à se développer avec, en 2012, EPAC [Est+10] qui permet une nouvelle fois l'augmentation du volume de données en français. Cette augmentation est également un des objectifs pour ETAPE [Gra+12] en 2012 ou encore REPERE [Gir+12] la même année. Depuis 2018, une nouvelle série de campagnes appelée DIHARD [Rya+21], initiées par un consortium de chercheurs de différentes structures, permet de faire avancer le domaine de la diarization sur des données en anglais et mandarin difficiles à traiter (bruitées ou contenant de la parole superposée). On peut s'attendre à ce que les évaluations futures soient également portées sur des cas limites de la diarization, avec de plus en plus de parole superposée, de bruit, et de parole spontanée.

La plupart des techniques actuelles utilisent des réseaux de neurones, mais celles-ci ne sont pas les premières à avoir été utilisées. Nous sommes conscients de l'existence de méthodes antérieures présentées sous le nom de "ségrégation de locuteur" par Gish et al. en 1991 [GSR91] et 1993 [YG93] en utilisant des modèles à mixtures de gaussiennes représentant des locuteurs. Cependant, nous ne développerons pas plus ces méthodes par souci de concision.

2.3.2 Métriques

En raison du nombre de communautés différentes travaillant sur ce sujet, plusieurs métriques sont utilisées en segmentation par classification, extraction d'une segmentation par classification de toutes les trames temporelles. Voici donc une présentation succincte des principales rencontrées.

Accuracy

L'accuracy, utilisée en anglais pour la différencier de la précision est une métrique de classification calculée à partir de l'ensemble ordonné des éléments de la prédiction $\hat{y}$ et l'ensemble ordonné des éléments de la référence y de longueur n grâce à la formule 2.8.

$$Accuracy = \frac{\#\{\hat{y}_k \mid \hat{y}_k = y_k, k \in [0; n]\}}{\#y} \tag{2.8}$$

L'accuracy est utilisé comme métrique de classification dans le cas où les données sont équilibrées. Celle-ci a pour avantage principal sa facilité de calcul et d'interprétation, elle peut en effet être interprétée par un pourcentage de bonne réponse. Le principal inconvénient de cette métrique est son biais envers la classe majoritaire. Un système favorisant la classe majoritaire aura un meilleur score qu'un système aléatoire. Pour pallier ce problème, nous pouvons utiliser d'autres métriques, par exemple le F-score.

F-score

Le F-score est une métrique de classification binaire avec $\hat{y}$ l'ensemble des prédictions de longueur n et y l'ensemble des références obtenu grâce à une combinaison des rappels et précisions. Le rappel et la précision sont deux métriques binaires complémentaires indiquant le rapport entre les faux positifs et faux négatifs. Elles dépendent du nombre de vrais positifs $\#\{\hat{y}_k \mid \hat{y}_k = y_k = 1, k \in [\![0; n]\!]\}$ ainsi que du nombre de prédictions positives pour la précision et le nombre de références positives pour le rappel.

$$Précision = \frac{\#\{\hat{y}_k \mid \hat{y}_k = y_k = 1, k \in [\![0; n]\!]\}}{\#\{\hat{y}_k \mid \hat{y}_k = 1, k \in [\![0; n]\!]\}} \tag{2.9}$$

$$\text{Rappel} = \frac{\#\{\hat{y}_k \mid \hat{y}_k = y_k = 1, k \in [\![0; n]\!]\}}{\#\{\hat{y}_k \mid y_k = 1, k \in [\![0; n]\!]\}} \tag{2.10}$$

$$F1\text{-}score = \frac{2 * \text{Rappel} * \text{Précision}}{\text{Rappel} + \text{Précision}} \tag{2.11}$$

La principale utilisation de cette métrique est dans le cadre d'une classification binaire, lorsque les deux classes ne sont pas équilibrées et que la classe positive est plus intéressante ou que l'on souhaite faire varier le coût des différents types d'erreurs. En effet, étant pondérée par les faux positifs et faux négatifs, cette métrique peut dans la plupart des cas diminuer le biais vers la classe majoritaire sans pour autant l'éliminer totalement [Fer23], ce score étant biaisé par le ratio entre le nombre de valeurs positives et le nombre de valeurs négatives. Il est également possible de régler le poids du rappel et de la précision pour se rapprocher du coût réel des faux positifs et faux négatifs. Dans le cas de cette étude, le poids est toujours égal à 1, c'est-à-dire avec le même poids pour le rappel et la précision, nous désignerons donc cette métrique par le nom F1-score.

Equal Error rate

L'Equal Error Rate (EER) est une métrique de classification binaire prenant en compte les faux positifs et faux négatifs sous forme de durée. On note $\hat{y}$ la séquence prédite et y la référence de longueur n.

$$False\ alarm = \#\{\hat{y}_k \mid \hat{y}_k = 1, y_k = 0, k \in [\![0; n]\!]\} \tag{2.12}$$

$$Miss = \#\{\hat{y}_k \mid \hat{y}_k = 0, y_k = 1, k \in [\![0; n]\!]\} \tag{2.13}$$

Pour calculer cette métrique, on fait varier un seuil d'acceptation du modèle pour obtenir $False\_alarm = Miss$. L'equal error rate est la valeur ainsi obtenue.

DetER

Le DER, dans notre cas *Detection Error Rate* ressemble au F-score, dans le sens où il dépend des faux positifs et faux négatifs, mais contrairement à celui-ci, il est calculé sur un intervalle de temps continu et non pas sur des trames de 10 ms.

$$DER = \frac{\mid Faux\ positif \mid + \mid Faux\ négatif \mid}{\mid fichier \mid} \tag{2.14}$$

Cette métrique permet de mesurer les performances d'une segmentation sans prendre en compte la précision de la prédiction. Elle permet donc d'évaluer des systèmes différents de manière équitable sans a priori sur la méthode de prédiction de la segmentation.

2.3.3 Détection de parole/non parole

Description de la tâche

La détection de parole, aussi appelée Voice Activity Detection (VAD), est une tâche qui consiste en une segmentation du signal en zones avec au moins un locuteur actif et zones sans locuteur actif. Cette tâche est très utile à toutes celles de traitement de la parole. En effet, elle agit comme première étape de pratiquement tous ces systèmes en séparant les zones de parole des zones de silence/bruit. La mauvaise réalisation de cette tâche est la cause principale des erreurs rencontrées par les systèmes de diarization, comme montrée par un article de Huijbregts et al. de 2007 [HW07].

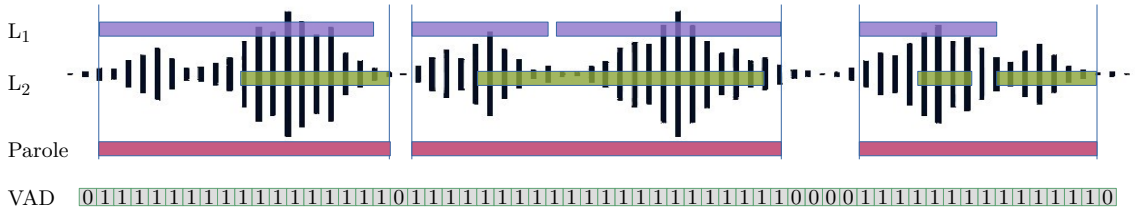


FIGURE 2.10 – Fonctionnement d'un système de détection de parole (VAD)

Dans notre étude, nous définissons la détection de parole comme une tâche dite "séquence vers séquence". Cela signifie que l'on souhaite obtenir une séquence de sortie à partir d'une séquence d'entrée. Dans notre cas, notre séquence d'entrée correspond à des caractéristiques acoustiques ou non, échantillonnée à une fréquence fixe à partir d'un signal audio. La séquence de sortie correspond à une prédiction en fonction du temps, échantillonnée pour nos travaux à 100 Hz. Cette séquence, présentée en figure 2.10, est classiquement un vecteur binaire, avec 1 indiquant la présence de parole dans la trame prédite et 0 en indiquant l'absence. Le modèle intermédiaire transforme donc une séquence de valeurs extraites sur un intervalle de temps connu et prédit une séquence de sortie de même durée temporelle en conservant l'ordre des prédictions.

2.3.4 Systèmes utilisés

De nombreux systèmes ont été utilisés pour cette tâche, obtenant des résultats présentés dans le tableau 2.1. Cette tâche étant annexe dans nos travaux, nous réaliserons un état des lieux non exhaustif reprenant différentes étapes de la recherche en détection de parole. Les premiers systèmes sortis utilisent des méthodes de traitement de signal à base d'énergie. Parmi ces travaux, nous pouvons citer ceux entrepris par Chengalvarayan et al. en 1999 [Che99] bien que ceux-ci n'aient pas été évalués sur cette tâche. Suivant ces travaux, plusieurs systèmes de détection de parole ont été développés sur une base de caractéristiques acoustiques en utilisant des modèles statistiques [NGM01 ; Woo+00 ; SKS99 ; PES01 ; SH98]. Ces systèmes sont alors utilisés en électronique pour aider à la reconnaissance de parole en environnement bruyant. Nemer et al. [NGM01] sont les seuls parmi ces systèmes à s'évaluer directement dans le cas d'une situation sans bruit et ob-

Article	**Modèle**	**Corpus**	**Métrique**	**Score**
[Ng+12]	GMM+MLP	RATS_dev	EER	1.82
[RLY13]	MLP	Havic	EER	19.82
[Tho+14]	CNN	RATS	EER	6.6
[GG17]	LSTM	AMI	DER	6.20

TABLE 2.1 – Résultats des différents modèles de VAD présentés dans l'état de l'art

tiennent une précision de 99.9 %, cible de l'expérience et un rappel de 22.2 %.

Enfin, suivant la popularité actuelle des réseaux de neurones, les derniers systèmes se basent sur ces réseaux pour prédire la présence ou l'absence de parole [Ng+12 ; RLY13 ; Tho+14 ; GG17 ; Lav+19]. Ces systèmes ont été évalués sur des données différentes, les résultats ne sont donc pas comparables, mais sont donnés à titre indicatif. Tout d'abord, Ng et al. [Ng+12] proposent un perceptron multicouche (MLP) pour un résultat mesuré en EER. Ce système a été évalué sur les données présentes dans la campagne d'évaluation RATS, mais les résultats présentés sont sur le corpus de développement. Avec ces paramètres, ce système atteint comme meilleur score 1.82 d'EER. En 2013, Ryant et al. [RLY13] présentent un nouveau système composé d'un perceptron multicouche avec des activations en ReLU sur le corpus Havic, contenant des vidéos YouTube. Ce modèle a également été évalué en EER avec un score de 19.82 obtenu sur le corpus de test. Le modèle suivant a été développé par Thomas et al. [Tho+14] en 2014. Il s'agit d'un CNN entraîné sur le corpus RATS. Les résultats obtenus sont également fournis en EER et atteignent 6.6 sur des données non vues. Plus récemment, Gelly et al. [GG17] proposent plusieurs architectures récurrentes pour cette tâche et entraînées/évaluées sur le corpus AMI. La métrique utilisée est différente, il s'agit du DER présenté précédemment. Le DER sur AMI atteint 6.20 points. Cette tâche est actuellement considérée comme ayant des résultats extrêmement fiables dans les cas généraux, les exemples de travaux plus récents proposés précédemment se concentrent sur les cas limites, avec beaucoup de bruit ou d'autres éléments pouvant perturber les systèmes.

2.3.5 Détection de parole superposée

Description de la tâche

La détection de parole superposée (OSD) est la tâche, présentée en figure 2.11, qui isole les zones avec plus de deux locuteurs actifs simultanément. La parole superposée est

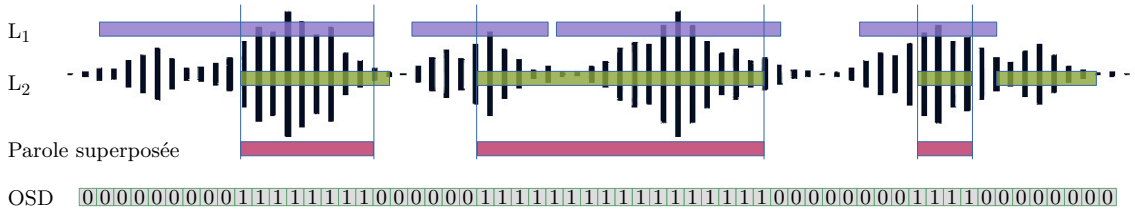


FIGURE 2.11 – Fonctionnement d'un détecteur de parole superposée (OSD)

un élément qui comporte beaucoup d'intérêts pour les sciences humaines, mais également pour l'informatique. En effet, comme montré par de nombreuses études [SSB01 ; HW07 ; Gar+20 ; ÇS06], ces zones sont sources d'erreurs importantes. Quand on traite de la transcription, l'étude de Shriberg et al. en 2001 [SSB01] pose un constat qui présente la présence de parole superposée comme source d'erreurs supplémentaires en favorisant les erreurs d'insertion, c'est-à-dire ajouter un mot qui n'existe pas. Un second article présenté par Cetin et al. en 2006 [ÇS06] étudie plus en détail les erreurs provoquées par la présence de parole superposée dans une transcription. L'hypothèse de cet article est que la présence de parole superposée dégrade les résultats de cette zone, mais également sur les zones précédant et suivant la parole superposée. L'article présente ainsi les évolutions de WER, word error rate qui doit diminuer pour signifier une amélioration, en fonction de la distance en seconde d'une zone de parole superposée. Outre le pic obtenu sur la zone de parole superposée, les scores autour sont dégradés par la présence de la superposition.

La présence de parole superposée est également étudiée en segmentation et regroupement en locuteur. En 2007, Huijbregts et al. [HW07] montrent que l'utilisation d'un système de détection de parole superposé est responsable d'une amélioration relative de 22.11 %, ce qui est le deuxième champ de gains après la détection d'activité vocale. Cette thématique est toujours actuelle. En effet, plus récemment l'équipe participant au workshop JSALT 2019 a également présenté un article [Gar+20] montrant l'importance des tâches de prétraitement sur les performances finales d'un système de reconnaissance de locuteur. Enfin, Bullock et al. [BBG20] ont proposé en 2019 une méthode d'utilisation de l'information de présence de parole superposée pour améliorer la diarization.

Systèmes utilisés

De nombreuses méthodes ont été utilisées pour détecter ces zones de parole superposée. Un résumé de ces méthodes est disponible dans le tableau 2.2. Tout d'abord, avant l'usage massif des réseaux de neurones, des systèmes ont été présentés utilisant des méthodes statistiques. En effet, certains travaux utilisent par exemple des HMM (modèles de Markov) ou des GMM (modèle à mixture de gaussienne) [Gei+13a ; ZH12 ; Boa+08] pour utiliser l'information extraite de caractéristiques acoustiques manuellement sélectionnées. Parmi ces travaux, on peut citer les travaux de Geiger et al. en 2013 qui utilisent un ensemble de caractéristiques, additionnées d'une information linguistique obtenue en comparant le mot en cours avec un modèle de langage entraîné sur de la parole superposée. Les résultats sont donnés en F1-score, présenté précédemment, et atteignent 41.71 % sur un sous-ensemble du corpus AMI. Utilisant des techniques similaires, on peut également citer les travaux de Zelenak et al. en 2012 [ZH12]. Ces travaux utilisent également des caractéristiques acoustiques sélectionnées manuellement en entrée d'un HMM pour obtenir un F1-score de 49.52 % sur un sous-ensemble du corpus AMI différent du précédent. Enfin, parmi ces travaux réalisés sans réseaux de neurones, on peut citer ceux de Boakye et al. qui utilisent également le même type d'architecture, soit un HMM/GMM avec des caractéristiques acoustiques sélectionnées. Ce système est également entraîné et évalué sur le corpus AMI et obtient pour son meilleur ensemble de caractéristiques un F1-score de 47 % sur le même sous-ensemble que le système précédent.

Lors de l'arrivée des réseaux neuronaux sur le domaine du traitement de la parole, du fait de la tâche qui est par nature "séquence vers séquence", le premier type de réseau à avoir été utilisé est un réseau récurrent, avec l'utilisation d'un LSTM par Geiger et al. en 2013 [Gei+13b]. Ce système fonctionne à partir des mêmes caractéristiques que pour son HMM/GMM [Gei+13a], est entraîné/évalué sur les mêmes données et obtient un F1-score de 38.82 %. Cette étude a été suivie de nombreuses autres utilisant également des architectures récurrentes [Hag+17 ; Saj+18]. On peut notamment citer les travaux de Hagerer et al. en 2017 [Hag+17] qui utilisent des MFCCs en entrée d'un LSTM. L'apport principal de l'article est l'introduction fondamentale de parole superposée artificielle dans les données d'entraînement. Ce système atteint de bonnes performances de 82.45 % de F1-score sur le corpus AMI, bien qu'il semblerait que l'ajout de parole superposée artificielle ait été utilisé sur le corpus de validation également, ce qui fausse les résultats. Une des méthodes les plus populaires actuellement pour la détection de parole superposée est celle implémentée dans l'ensemble d'outils pour la diarization pyannote [Bre+20]. Cette

détection se base sur l'architecture proposée par Bullock et al. [BBG20] composée de deux couches BiLSTM suivies de trois couches linéaires avec comme entrée des MFCCs ou un SincNet [RB18]. Ces travaux obtiennent des résultats, pour l'époque, à l'état de l'art sur le corpus AMI avec un F1-score de 63.41 % avec des MFCCs, 74.36 % avec une entrée SincNet, et sur le corpus DIHARD un F1-score de 27.01 % avec des MFCCs et de 37.77 % avec SincNet.

Plus récemment, suivant les travaux réalisés à partir de réseaux récurrents, une autre branche inspirée par les travaux en reconnaissance d'image utilise des réseaux convolutifs. Comme présenté précédemment, ce type de réseau est utile pour trouver des motifs dans un signal, ce qui le rend efficace pour les tâches liées au traitement de signal. Parmi ces études [YH20 ; KB18 ; ACB17 ; Kun+19], on peut citer les travaux de Kunešová et al. en 2019 [Kun+19] qui présentent un modèle basé sur trois couches convolutives suivies de trois couches linéaires pour détecter les zones de parole superposée. Ce système, évalué sur AMI obtient un F1-score de 57.28 %. En 2018, Kazimirova et al. [KB18] utilisent également un CNN avec un spectrogramme en entrée pour obtenir un F1-score de 71 % sur le corpus SSP Conflict Corpus. Enfin, parmi les travaux sur des réseaux convolutifs, nous pouvons citer ceux d'Andrei et al. en 2017 [ACB17]. Leur système utilise des caractéristiques acoustiques sélectionnées manuellement comme entrée d'un CNN. Les résultats sont obtenus sur un corpus artificiel et atteignent un F1-score de 80 %.

Des travaux ont également cherché à combiner les réseaux récurrents et convolutifs en utilisant des réseaux convolutifs ayant une notion temporelle. On peut citer notamment l'utilisation de Convolutional Recurrent Neural network(CRNN) [Jun+21 ; Pha+19] qui atteint pour les travaux de Jung et al. [Jun+21] un F1-score de 43.40 % sur le corpus DIHARD2, ou de Time-delayed Neural network (TDNN) [Man+19]. Une dernière méthode de combinaisons des intérêts des méthodes récurrentes et convolutives que nous utiliserons plus particulièrement par la suite est celle décrite par Cornell et al. [Cor+20]. Cette étude présente l'usage d'un TCN, Temporal Convoluted Network introduit par Bai et al. [BKK18] pour la tâche de comptage de locuteurs supposée très proche de la détection de parole superposée. Ce réseau est un réseau convolutif qui utilise les propriétés de celui-ci pour apprendre des motifs de plus en plus globaux et ainsi garder l'échelle temporelle du signal.

Enfin, d'autres méthodes ont également eu de bons résultats. Par exemple, l'apprentissage bout-à-bout présenté par Bredin et al. [BL21] apprend à détecter de la parole superposée en même temps qu'une détection de parole en prenant pour objectif la créa-

Article	**Caractéristiques**	**Modèle**	**Corpus**	**F1-score (%)**
[Gei+13a]	Sélectionnées	HMM/GMM	AMI_1	41.71
[ZH12]	Sélectionnées	HMM/GMM	AMI_2	49.52
[Boa+08]	Sélectionnées	HMM/GMM	AMI_2	47
[Gei+13b]	Séléctionnées	LSTM	AMI_1	38.82
[Hag+17]	MFCC	BLSTM	AMI_artificiel	82.45*
[BBG20]	MFCC	BLSTM	AMI	63.41
[BBG20]	SincNet	BLSTM	AMI	74.36
[BBG20]	MFCC	BLSTM	DIHARD2	27.01
[BBG20]	SincNet	BLSTM	DIHARD2	37.77
[Kun+19]	Spectrogramme	CNN	AMI	57.28
[KB18]	Spectrogramme	CNN	SSPConflict Corpus	71
[ACB17]	Sélectionnées	CNN	Artificiel	80
[Jun+21]	Mel-spectrogramme	CRNN	Dihard2	43.40
[BL21]	SincNet	Multi-Task	DIHARD3	59.9
[BL21]	SincNet	Multi-Task	AMI	75.3

TABLE 2.2 – Résultats des différents modèles d'OSD présentés dans l'état de l'art

tion d'une segmentation complète en locuteur. Cette méthode atteint des résultats état de l'art en détection de parole superposée sur les corpus AMI et DIHARD3 en obtenant respectivement un F1-score de 75.3 % et de 59.9 %. Une autre méthode particulière de détection de parole superposée est présentée par Shokouhi et al. [SH17]. Celle-ci utilise un nouveau type particulier de caractéristique, les pyknograms comme entrée des modèles et utilise comme classifieur, avec des résultats satisfaisants, un SVM (support vector machine) ou un apprentissage non supervisé basé sur la distance euclidienne entre deux trames consécutives.

2.4 Tâches de caractérisation

Certaines tâches que nous étudions ne sont pas des tâches de segmentation par classification, mais des tâches de caractérisation. La principale différence tient dans l'échelle de la prédiction qui est à la trame pour la segmentation et au segment pour la caractérisation. Les métriques utilisées sont cependant identiques à celles présentées précédemment.

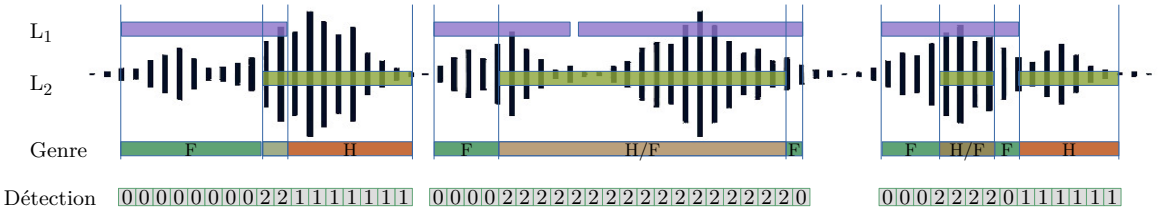


FIGURE 2.12 – Fonctionnement d'un système de détection de genre (GD)

2.4.1 Détection de genre

Description de la tâche

La détection de genre est une tâche qui traite de l'extraction d'information sur les locuteurs. Cette tâche est assez particulière, car elle ignore volontairement la définition sociale du genre pour considérer le genre comme caractéristique physiologique observable et reconnaissable dans la voix, mais l'évalue sur le genre dans sa définition sociale. Il existe donc certaines erreurs qui sont inévitables, car indiscernables pour des annotateurs humains.

La détection de genre dispose de plusieurs utilités. Tout d'abord, étant une tâche automatique, elle permet de traiter de grandes quantités de données dans un but d'état des lieux, ou de diagnostic. Par exemple, en 2021 est sorti un article du Comité Supérieur de l'Audiovisuel [21], traitant de la présence des femmes dans le paysage audiovisuel français, relevant ainsi des chiffres disparates, et une inégalité de représentation. Cette étude a été rendue possible grâce à l'utilisation de détection automatique de genre sur les données de l'INA, qui aurait été impensable avec une annotation manuelle. D'autres utilisations peu éthiques sont également possibles pour l'adaptation de publicité ou de service client en fonction du genre.

La détection de genre est également utilisée à des fins d'amélioration de système automatique de traitement de parole. Par exemple, Garnerin et al. [GRB19] présentent une amélioration des résultats de transcription en faisant usage de l'information de genre à priori. Après comparaison de l'erreur de transcription entre des locuteurs homme et femme, cet article montre une augmentation de 24 % d'erreur pour les locutrices. Un corpus équilibré aide à réduire cet écart.

Systèmes utilisés

La détection de genre fait partie des problèmes abordés avant l'âge d'or des méthodes neuronales. On trouve donc des méthodes statistiques et liées au traitement de signal. Cependant, cette tâche n'a jamais été considérée comme une tâche majeure nécessitant une collaboration approfondie entre laboratoires. Il n'existe donc pas de cadre d'évaluation répété sur plusieurs expériences. Tous les résultats obtenus sont sur des corpus différents avec des définitions de tâches différentes. Tout d'abord, on peut citer deux études menées sur un ensemble très réduit de locuteurs (52 locuteurs) atteignant une reconnaissance parfaite du corpus en utilisant des descripteurs acoustiques sélectionnés manuellement et une distance euclidienne [WC91]ainsi qu'une étude sur des caractéristiques extraites des voyelles [CW91]. Cependant, ces études sont menées au niveau du locuteur, et non d'un segment de parole, c'est à dire reconnaître le genre d'un locuteur à travers plusieurs segments. Au niveau du segment, plusieurs études [PC96 ; Boc+08] ont été menées sur la détection de genre. On peut citer notamment celle menée par Parris et al. [PC96] qui utilise deux HMM pour représenter les classes homme et femme et ainsi maximiser la vraisemblance qu'un segment étudié appartienne à l'un des deux modèles. Ce système atteint une accuracy de 98 % sur le corpus Oregon graduate institute (OGI) multi lingual corpus qui contient deux fois plus de locuteurs homme que de locuteurs femmes. Une étude plus récente de détection de genre a été menée par Bocklet et al. [Boc+08]. Cette étude cherche à détecter le genre et l'âge d'un locuteur grâce à un modèle de mixture de gaussienne(GMM) par locuteur qui sont alors classifiés grâce à un SVM. Plus récemment, des méthodes neuronales ont également été utilisées pour détecter le genre à la trame, comme une modélisation de séquence. Ce fonctionnement est décrit en figure 2.12. On peut citer notamment celle de Doukhan et al. [Dou+18] comparant l'utilisation de GMM, I-vector, qui est une méthode de compression d'information de locuteur, et d'un CNN pour la détection de genre. Cet article propose également un cadre défini pour la détection de genre dans des segments de parole monolocuteurs.

2.4.2 Détection d'interruption

Description de la tâche

La détection d'interruption est la tâche que nous souhaitons aborder comme objectif final de cette thèse. Cette tâche n'est, à ma connaissance, pas une tâche standard du traitement automatique de la parole. Il existe des travaux isolés, mais pas de définition

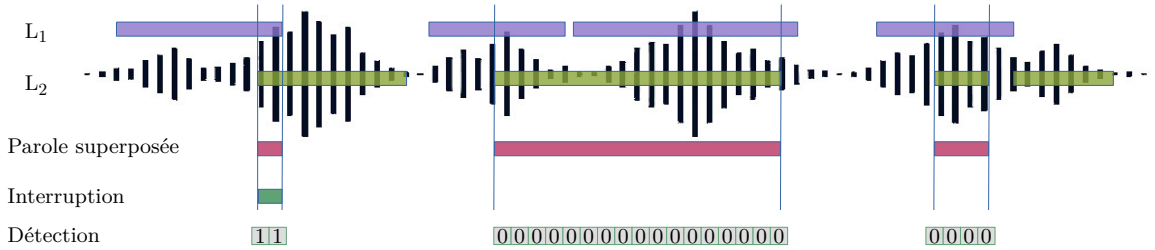


FIGURE 2.13 – Fonctionnement d'un système de détection d'interruption)

standard de cette tâche.

Suivant la définition d'interruption proposée en section 1.3.2 nous définissons la tâche de détection d'interruption comme la segmentation d'un signal de parole pour relever les zones dans lesquelles une interruption aurait eu lieu selon des annotateurs humains. La notion d'incertitude ou de réalité de présence de l'interruption est compliquée à déterminer en raison de la nature subjective de la tâche. L'évaluation de ce système est donc en partie perceptive. L'objectif attendu par notre expérience est résumé dans la figure 2.13.

La première utilité prévue à ce système provient du projet GEM. Il s'agit de la recherche de zones contenant potentiellement des interruptions pour accélérer les recherches menées sur la signification de ces interruptions, ou toute autre caractéristique sociale liée à leur présence par des laboratoires spécialisés dans ce domaine. Une seconde possible utilisation est proposée par Caraty et al. [CM15] qui proposent la détection "d'interruption" comme moyen de détecter les situations de conflit. Enfin, une dernière utilisation possible de la détection d'interruption est proposée par Fu et al. [Fu+22] comme une mesure de la fluidité des échanges dans un but de faciliter l'intervention de tous les participants d'une conversation souhaitant y prendre part.

2.4.3 Systèmes utilisés

L'état de l'art sur cette tâche est extrêmement restreint. On peut tout d'abord citer Caraty et al. [CM15] qui considère la parole superposée comme étant une interruption, ce qui est une simplification trop importante pour notre objectif. Ils ont néanmoins utilisé un SVM basé sur des descripteurs acoustiques pour déterminer la présence de parole superposée, ainsi qu'un second SVM qui utilise des descripteurs acoustiques et des carac-

téristiques issues du système de détection de parole superposée pour estimer un niveau de conflit. Le second article traitant de la détection de parole superposée est plus récent. Il s'agit d'un article de Fu et al. d'octobre 2022 [Fu+22] qui présente un système destiné aux conversations par visioconférence pour détecter automatiquement les tentatives de prises de paroles ratées. Bien que différente de notre tâche, elle est suffisamment similaire pour que des techniques utilisées par cet article puissent fonctionner dans notre cas. Le système utilisé est en effet semblable au nôtre (présenté en section 6.3), utilisant les informations fournies par un système auto supervisé pour classifier des segments selon le type de parole superposée.

2.5 Synthèse

Dans cette section, nous avons pu voir les tendances de l'état de l'art concernant les tâches que nous allons aborder durant cette thèse. Au travers de celui-ci, plusieurs éléments ressortent. Tout d'abord, pour toutes ces tâches, les derniers systèmes utilisés sont à base de réseaux de neurones, nous allons donc nous concentrer sur ces techniques. Ce choix implique notamment un travail approfondi sur les données, les réseaux de neurones étant de très grands consommateurs de ces dernières. Il est cependant nécessaire de préciser que l'utilisation de tels réseaux rend l'interprétabilité et l'explicabilité des résultats pratiquement impossible. Nous ne pourrons donc pas fournir d'aide au diagnostic des résultats obtenus.

Concernant la détection de parole superposée, la tâche étant une tâche séquence vers séquence, nous allons utiliser des systèmes à base de réseaux récurrents, les plus efficaces semblants être les BiLSTM. Une autre piste que nous souhaitons explorer est l'utilisation de TCN qui sont conçus pour permettre de remplacer les réseaux récurrents de manière efficace. Nous souhaitons utiliser deux types principaux d'entrées pour nos systèmes. Les MFCCs, qui sont assez bas niveau pour bénéficier des capacités de reconnaissance de motifs des réseaux de neurones et qui ont été utilisés avec succès dans l'état de l'art ainsi que des modèles préentrainés d'extraction de caractéristiques. Concernant ces derniers notre choix se porte sur WavLM, en effet ce système est actuellement le meilleur système de ce type pour la modélisation de langage [Yan+21a]. Le second avantage repose sur une intuition personnelle non confirmée expérimentalement. L'usage de parole superposée artificielle dans les données d'apprentissage de ce modèle améliore le comportement des systèmes utilisant ces caractéristiques face à ce type de données. Celles-ci, comme montré

par l'évolution de l'état de l'art, sont obligatoires pour obtenir des systèmes de détection de parole superposée efficaces. Cette tâche, au même titre que la détection de parole/non-parole est toujours intéressante pour les domaines qui utilisent de la parole en entrée. Cependant, de plus en plus de systèmes recherchent des alternatives à leur utilisation. En effet, dans plusieurs systèmes récents, ces tâches sont ignorées en comptant sur les architectures utilisées pour les traiter en parallèle de la tâche souhaitée. Je pense toutefois que la réalisation de ces tâches peut avoir d'autres utilités qu'être un pur prétraitement et dois par conséquent suivre les avancées techniques des autres domaines. Certains exemples seront présentés dans la suite de ce manuscrit.

Concernant maintenant la détection de genre, nous avons fait le choix de ne pas nous attarder sur cette tâche. En effet, nous avons pu voir dans l'état de l'art que les systèmes existants sont déjà très performants et donc difficiles à améliorer. La difficulté d'amélioration est accentuée d'un manque d'intérêt global pour cette tâche, en effet les systèmes actuels ne traitent pas de la détection de genre. Il n'est cependant pas exclu que nous testions quelques systèmes comme tâche annexe.

Enfin, la détection d'interruption est une tâche qui n'existe pas encore. En effet, sur les deux articles trouvés traitant de la question, un n'utilise pas du tout la même définition d'interruption que nous, et le second est intervenu après la réalisation des premières expériences sur cette tâche. L'absence de corpus pour celle-ci nous incite donc à créer un corpus répondant à notre définition d'interruption, puis d'explorer différents systèmes pour en prédire les occurrences.

Chapitre 3

Analyses des corpus pour la segmentation de signal de parole

Comme définie dans les sections précédentes, nous allons travailler sur la segmentation et caractérisation de signal de parole à l'aide de techniques d'apprentissage automatique. Celles-ci nécessitent une grande quantité de donnée annotées, c'est à dire dans notre cas, segmentées par des annotateurs humains.

3.1 Présentation des corpus utilisés

Nos travaux menés sur la détection d'interruptions dans des corpus de médias nécessitent deux types de ressources différentes. Des corpus de médias existent en français mais ne sont que très peu utilisés dans la communauté internationale pour la détection de parole superposée. Pour pouvoir s'inscrire dans cette communauté et permettre aux travaux effectués d'avoir un impact sur l'état de l'art, nous allons utiliser deux corpus en anglais ne traitant pas de médias. Ces corpus ont également l'avantage de contenir une quantité de parole superposée importante (>10 %). La seconde partie de ce tour des corpus utilisés contient une description des principaux corpus de médias qui pourraient être utilisés pour faire une chaîne de traitement complète et représentative de la tâche.

3.1.1 Corpus pour la détection de parole superposée

Les deux corpus les plus utilisés actuellement pour la détection d'activité vocale et de parole superposée sont les corpus AMI [Mcc+05] et DIHARD [Rya+21]. Nous utiliserons donc ces corpus dans notre étude. Ce choix nous permettra de pouvoir comparer nos résultats à ceux de l'état de l'art, ainsi que de pouvoir utiliser ces données sans doute sur la qualité des annotations fournies.

Corpus	**Durée (h)**	**Proportion de parole superposée (%)**	**Domaine**
AMI	100	24.7	Réunion
DIHARD	34	11.6	Multiple
ESTER1&2	260	0.67	Radiophonique
ETAPE	30	1.49	Télévisuel
EPAC	100	5.29	Radiophonique
Repere	60	3.36	Télévisuel
ALLIES	328	3.32	Médias
ALLIES_LCP_debate	48	9.85	Débats
Téléréalité	116	Non annoté	Téléréalité

TABLE 3.1 – Corpus utilisés pour la segmentation de signal de parole

AMI

Le corpus AMI (Augmented Multi-party Interaction) [Mcc+05] a été collecté grâce au projet européen du même nom en 2005. L'objectif de ce corpus est de permettre le développement de techniques visant à améliorer les interactions au sein d'un groupe de personnes. Pour cela, une campagne de recueil de données a été menée sur quatre sites distincts, Idiap, Edinburgh, TNO and Brno. Cette campagne inclut plusieurs modalités différentes comme l'audio, la vidéo ou encore des crayons connectés. L'audio, modalité sur laquelle nous nous sommes spécialisés, a été enregistré avec un micro-cravate pour chaque participant ainsi qu'une antenne de micro pour la salle, ce corpus est donc utilisable pour du traitement monocanal en champ proche et multicanal en champ distant. Ce corpus est semi-acté, dans le sens où il simule des réunions par des locuteurs qui s'expriment spontanément, mais à un rôle dont ils n'ont pas l'habitude. La partition d'apprentissage que nous utilisons provient de celle proposée par l'équipe de recherche en traitement automatique BUT de Brno. Ce corpus est en anglais et comporte approximativement 100 h de données.

DIHARD

Le corpus DIHARD [Rya+21] a été utilisé dans le cadre de la campagne d'évaluation DIHARD III qui s'est déroulée fin 2019. L'objectif de ce corpus est d'évaluer dans un cadre commun les systèmes de diarization développés dans les laboratoires dans des conditions considérées comme difficiles. Ce corpus est une extension du corpus utilisé lors de DIHARD

II contenant un regroupement de données extraites de différents corpus appartenant à différents domaines.

- Médical : Conversations enregistrées dans des contextes médicaux. Les enregistrements sont tirés du corpus ADOS.
- Face à face : Interviews dans des contextes variés, avec des qualités variés. Ce domaine utilise des données issues des corpus DASS [Kre+12], SLX [Str+03] et Youthpoint. Les enregistrements regroupent des discussions réalisées dans le cadre d'études linguistiques sur les accents régionaux américains ainsi que des interviews de personnalités par des étudiants.
- Téléphone : Utilise une partie du corpus Fisher [CMW03]. Ce corpus contient des discussions téléphoniques entre deux interlocuteurs ne se connaissant pas sur des questions définies à l'avance.
- Map task : Sous-ensemble du corpus DCIEM [T+96] où un locuteur donne des directions à un second sur une carte. Ce corpus a été récolté grâce à des micro-cravates comme partie d'une étude menée sur la privation de sommeil.
- Conversation de groupe : Conversations entre plusieurs locuteurs, dans des conditions plus ou moins bruitées. Il est composé d'un sous-ensemble des corpora ROAR, CIR et RT04 [Fis+07].
- Tribunal : Enregistrements des audiences de la Cour Suprême américaine issues du corpus Scotus.
- Livre audio : Ensemble d'enregistrements venant de LibriVox.
- Vidéos : Partie du corpus VAST qui est un recueil de vidéo en anglais, mandarin et arabe, segmentées en tour de parole.

Ce corpus comporte donc beaucoup de bruits, de qualités audio différentes ainsi que de parole spontanée, qui contient de la parole superposée. La partition d'apprentissage est celle définie par notre équipe pour la campagne d'évaluation DIHARD III séparant le corpus fourni entre train et validation, conçue pour respecter l'équilibre de domaine des données dans les distributions originales. La partition de test est la partition officielle distribuée par les organisateurs. Ce corpus contient majoritairement des données en anglais et comporte approximativement 34 h d'audio annoté.

3.1.2 Corpus de médias

Notre tâche concerne la détection automatique d'interruptions dans des médias. Nous souhaitons par conséquent travailler sur des signaux audio issus de médias audiovisuels

français. Il faut donc plusieurs corpus, répondant aux mêmes critères que les corpus précédents, sur des médias.

Ester1&2 Les corpus Ester1&2 [Gal+06] sont issus des campagnes de transcription éponymes. Ils ont pour principal objectif de fournir à la communauté de transcription automatique, des données en français. Pour cela, ils contiennent des fichiers audio provenant de 6 sources radiophoniques françaises, France Inter, France Info, RFI, RTM, France Culture et Radio Classique, collectés entre 1998 et 2004. Ils sont donc tout à fait adaptés à l'étude des interactions dans les médias. L'ensemble comporte approximativement 100 h d'audio annoté pour Ester1 et 160 h pour Ester2.

Etape Le corpus Etape [Gra+12] est issu de la campagne d'évaluation du même nom. Il prend la suite directe des corpus Ester1&2 en contenant également des zones avec de la parole spontanée et de la parole superposée. Pour cela, il contient des données provenant d'émissions télévisuelles en privilégiant les plus fortement interactives. Cette interactivité est particulièrement intéressante dans le cadre de cette thèse, car susceptible de contenir des zones d'interruption et/ou de parole superposée. Ce corpus contient 30 h de segments audio annotés.

Epac Le corpus Epac [Est+10] a été créé à l'initiative de quatre laboratoires français, IRIT (Toulouse), LI (Tours), LIA (Avignon) et LIUM (Le Mans) en 2010. L'objectif de cette initiative est de concentrer le plus d'informations possible, de manière structurée, sur des données audio pour disposer de données en français utilisables pour de nombreuses tâches, comme la segmentation, la transcription, la détection de sujet ou encore la détection d'opinion. Les données de ce corpus proviennent de zones non annotées du corpus Ester1 en privilégiant les zones avec de la parole spontanée. Ce corpus est séparé en deux parties, une première annotée, que nous utiliserons d'une longueur approximative de 100 h de données audio, ainsi que 1700 h de données transcrites automatiquement. Nous n'utiliserons que la partie annotée manuellement pour nous assurer une segmentation précise des données utilisées.

Repere Le corpus Repere [Gir+12] provient d'une campagne d'évaluation de la DGA (Direction Générale de l'Armement) dont l'objectif est d'obtenir un cadre commun de recherche sur la reconnaissance de locuteurs de manière multimodale, en utilisant l'audio et la vidéo. Les données proviennent de deux chaînes de télévision française, LCP

et BFMTV, plus précisément de leurs émissions d'information et de débats. Ce corpus contient environ 60 h de données vidéo et audio annotées.

ALLIES Le corpus ALLIES [Lar+21] est un métacorpus développé par le LIUM. Il répond à un besoin constant de quantité de données, plus précisément de données annotées de manière homogène en français pour faire de la segmentation et regroupement en locuteur. Ce corpus contient donc les données extraites des corpus Ester, Etape et Repere, accompagnées d'un nouvel ensemble de données plus récentes, avec des annotations manuellement corrigées. Le corpus ALLIES permet de centraliser un grand nombre de données et ainsi de pouvoir utiliser des systèmes assez conséquents ou d'extraire des sous-ensembles spécifiques sans manquer de données.

Ce corpus comporte approximativement 328 h d'audio annoté dans sa première version et devrait prochainement être mis à disposition de la communauté. Cette première version est celle désignée par ALLIES dans la suite de ce manuscrit. Nous tirons de ce corpus un sous-ensemble appelé ALLIES_LCP_debate qui contient des débats issus de trois émissions de la chaine LCP, *Pile et Face*, *Entre les lignes* et *Ça vous regarde*. L'objectif est de constituer un sous-ensemble fortement interactif pour en étudier les particularités lors d'études de corpus.

Ce corpus est également l'objet d'une campagne de correction d'annotation. Les noms des locuteurs sont alors harmonisés et corrigés sur l'ensemble des fichiers du corpus par une phase de comparaison de similarité obtenue par un système automatique de vérification de locuteur. Les ensembles de locuteurs susceptibles d'être mal annotés sont alors vérifiés par des annotateurs humains.

Télé-réalité Le corpus Télé-réalité est un ensemble de données actuellement non publiques, extraites par l'INA dans le cadre du projet ANR GEM. Ce corpus est utilisé comme données communes entre une étude menée par un laboratoire d'étude des médias et le LIUM pour amorcer une collaboration. Il contient l'ensemble de la saison des Marseillais à Dubaï ainsi que la première saison de Loft Story. Plusieurs sources d'audio sont disponibles : la version brute extraite de l'émission, ainsi qu'une version obtenue après passage d'un algorithme de séparation de source (Spleeter [Hen+20]) pour éliminer la musique omniprésente qui est source d'erreur dans les systèmes automatique de traitement de la parole. Ce corpus compte 84 h d'audio pour la partie des Marseillais et 32 h pour le Loft.

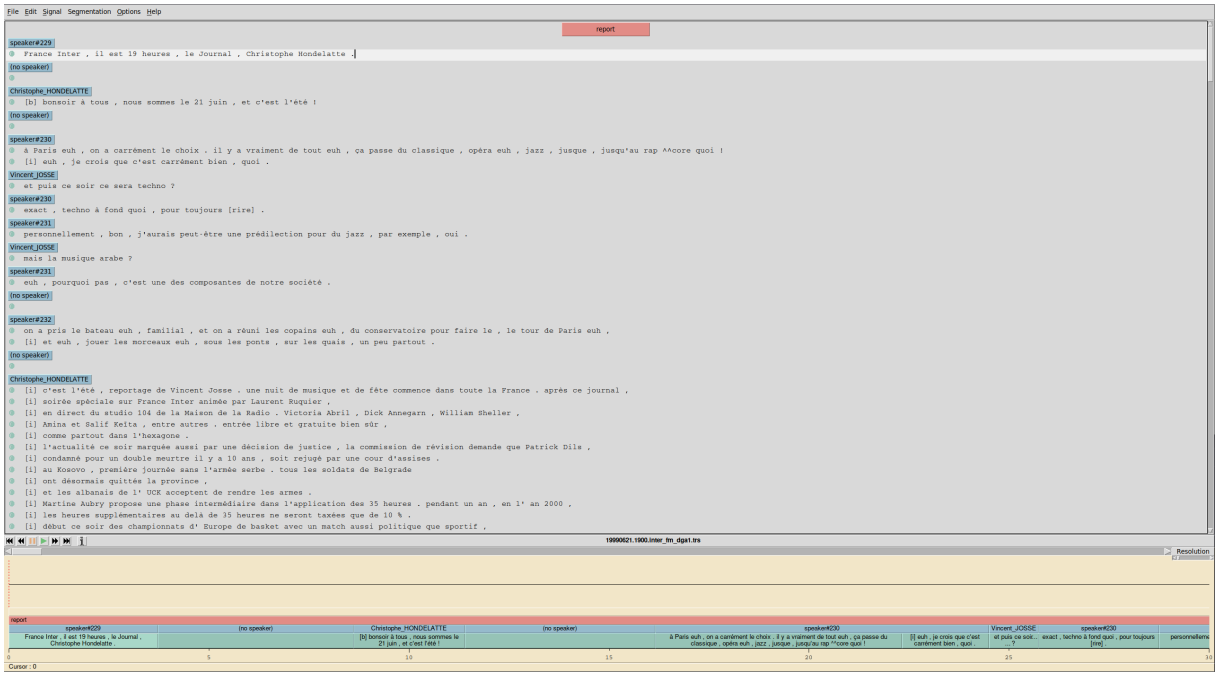


FIGURE 3.1 – Interface d'annotation pour transcriber [Bar+98]

3.1.3 Format d'annotation

Pour utiliser ces corpus, nous utilisons une segmentation en locuteur quand elle est disponible. Cette segmentation est obtenue par des annotateurs humains qui découpent un enregistrement en tours de parole. Ces annotations peuvent être réalisées grâce à des outils tels que transcriber [Bar+98]. Comme pour les corpus de médias présentés précédemment, ces annotations peuvent contenir une transcription alignée, ce qui permet d'obtenir les temps de début et de fin de tour de parole. À partir de ces outils, les références sont enregistrées sous plusieurs formats standardisés.

Parmi ces formats, nous utilisons le format *mdtm* qui suit le schéma suivant :

```
show canal début durée type NA genre identifiant
```

Pour un fichier appartenant à un des corpus utilisés, ce format prend la forme suivante :

```
20140429_Ca_vous_regarde_2220 1 333.012 24.920 speaker NA adult_male Arnaud_ARDOIN
20140429_Ca_vous_regarde_2220 1 363.916 33.176 speaker NA adult_female Brigitte_BOUCHER
```

Ce format contient un nombre limité d'informations, mais est aisément lisible et pratique à traiter. Cependant, il ne peut pas contenir de transcription, ce n'est donc pas le format utilisé pour les corpus provenant de la communauté de transcription. Ce format est également dérivé en un second qui contient les mêmes informations dans un ordre différent

appelé *rttm*.

La communauté de transcription de laquelle provient la majorité des corpus segmentés en français utilise un format provenant de l'outil transcriber. Ce format, appelé TRS suit le format suivant :

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE Trans SYSTEM "trans-14.dtd">
<Trans audio_filename="20140429.2220.LCP_CaVousRegarde.wav">
<Speakers>
    <Speaker id="spk2" name="Arnaud_ARDOIN" check="no" type="male" dialect="native" accent="" scope="local"/>
    <Speaker id="spk5" name="Brigitte_BOUCHER" check="no" type="female" dialect="native" accent="" scope="local"/>
</Speakers>
<Episode>
    <Section type="report" startTime="333.012" endTime="3264.148">
        <Turn speaker="spk2" startTime="333.012" endTime="357.932">
            <Sync time="333.012"/>
        </Turn>
        <Turn startTime="357.932" endTime="363.916">
            <Sync time="357.932"/>
        </Turn>
        <Turn speaker="spk5" startTime="363.916" endTime="397.092">
            <Sync time="363.916"/>
        </Turn>
    </Section>
</Episode>
</Trans>
```

Dans ce type de fichier, les tours de parole sont stockés dans une arborescence xml qui contient toutes les informations nécessaires à la segmentation.

3.2 Analyses de référence

Pour avoir une vue d'ensemble des corpus que nous utiliserons, nous avons analysé les corpus que nous utiliserons. Cette étude est menée dans le but de trouver des tendances à exploiter ou des hypothèses à formuler à partir des données.

3.2.1 Annotation des corpus

Les données venant de sources différentes, pour pouvoir les comparer ou les agréger, il faut dans un premier temps savoir exactement comment elles ont été collectées, et annotées.

De manière générale, les protocoles d'annotation de corpus sont décrits précisément dans des documents appelés guides d'annotation. Ces documents sont supposés couvrir tous les cas rencontrés par les annotateurs et décrire exactement les annotations effectuées.

Dans les faits, des différences subsistent entre les spécifications inscrites et l'annotation effectuée.

L'annotation des corpus de segmentation et regroupement en locuteur consiste en une segmentation en tour de parole. Elle est composée de segments contenant au minimum le temps de début du tour, la durée ou le temps de fin du tour ainsi que l'identité du locuteur sous forme d'un nom ou d'un identifiant. À ces informations peuvent s'ajouter d'autres caractéristiques comme le genre du locuteur, ou tout autre élément pouvant être ajouté au segment de parole. La parole superposée est extraite à partir de ces segmentations en récupérant l'intersection de deux tours de parole.

Les corpus que nous utilisons ont été annotés par des équipes différentes, à des périodes différentes, pour des objectifs différents. Ces besoins contraignent la création de guide d'annotation et pas conséquent apportent des différences dans l'annotation de la parole superposée. Tout d'abord les corpus ESTER 1&2, EPAC, ETAPE et REPERE sont des corpus de transcription, c'est-à-dire qu'ils sont supposés contenir la transcription linguistique des locuteurs à chaque instant. C'est sur les zones de parole superposée que la différence principale se fait. En effet, pour les corpus ESTER 1&2 ainsi que pour le corpus EPAC, en zone de parole superposée, les mots prononcés par les deux locuteurs sont souvent transcrits quand ils sont intelligibles, bien que non spécifié dans le guide d'annotation. Il s'agit ici d'une des limites de ces guides d'annotations. La qualité du travail des annotateurs est évaluée sur un minimum d'information à annoter. Toute information supplémentaire est alors considérée comme "bonus". Mais l'histoire derrière la création de ces corpus devient alors nécessaire pour savoir ce qu'il est possible d'utiliser. Il est donc nécessaire de connaître une des personnes impliquées dans la création du corpus pour en connaître tous les soucis rencontrés non inclus dans les guides d'annotations, réduisant ainsi considérablement l'intérêt de ceux-ci.

Autre exemple, le corpus ETAPE aurait pu être très intéressant pour notre tâche. En effet, il ne contient pas uniquement les zones de parole superposées, mais également des informations supplémentaires sur ces zones, notamment le type de parole superposée tel que défini par Adda et al. [Add+07]. Cependant, le choix réalisé dans l'annotation de ces zones a mis de côté une grande partie des temps de début et de fin de ces zones de parole superposée, considérés comme inutile pour un traitement manuel, mais indispensable au traitement automatique. C'est pour cette raison que la proportion de parole superposée de 1.49 % (voir tableau 3.1) est aussi faible. La nature assez interactive du corpus aurait dû entraîner une proportion supérieure à 5 % par comparaison avec des corpus similaires. Pour

finir, comme présenté dans la section 3.1.2, le corpus ALLIES est un corpus dont la seconde version est supposée sortir prochainement, qui combine les corpus cités précédemment. L'harmonisation et la correction de ces données sont les éléments principaux permettant de pouvoir entraîner et tester des systèmes sereinement, mais il s'agit d'une tâche très complexe à réaliser manuellement. L'automatisation de ces vérifications est donc une tâche intéressante à étudier et à considérer comme application des différents systèmes automatiques développés actuellement.

3.2.2 État des lieux de la parole superposée

Proportion de parole superposée

Comme présenté précédemment en section 1.1.1, la parole superposée est un évènement fréquent en occurence mais peu présent en durée par rapport à la durée de parole totale [Add+07]. Or l'apprentissage de systèmes de détection de parole superposée nécessite une proportion de parole superposée assez importante. Mesurer cette proportion de parole superposée dans nos différents corpus est donc un moyen efficace pour trouver quels corpus utiliser en priorité pour l'apprentissage d'un tel système.

Protocole Pour ce faire, nous mesurons la durée cumulée des segments de parole superposée normalisée par la durée annotée du fichier total (incluant les silences). Il s'agit donc de la proportion de parole superposée du fichier total et non pas de la proportion par rapport à la quantité de signal de parole annoté. Toutes les valeurs sont récupérées à partir des références au format RTTM.

Résultats et discussion Ces résultats sont regroupés dans le tableau 3.1. Tout d'abord, on peut voir que la proportion de parole superposée est variable, allant de 0.7 % pour les corpus Ester 1 et 2 (discours journalistique) 24.7 % pour le corpus AMI (discours spontané), ce qui confirme l'hypothèse que le type de discours présent influence grandement la proportion de parole superposée. Enfin, parmi les corpus, on peut remarquer trois ensembles de corpus. Le premier contient AMI et DIHARD (11.6 %) et pourrait contenir aussi ALLIES_LCP_debate (9.85 %) qui contient majoritairement du discours journalistique dialogal (entre spontané et préparé). Ces trois corpus contiennent donc de la parole spontanée, hautement interactive.

Le second groupe contient EPAC (5.29 %), REPERE (3.36 %) et ALLIES (3.32 %)

et correspond à des corpus de médias classiques avec des débats, interviews et journaux télévisés. On pourrait s'attendre à voir également le corpus ETAPE (1.49 %) dans cette catégorie, mais il semblerait que des différences d'annotations modifient la proportion de parole superposée comme montré en section 3.2.1.

La dernière catégorie regroupe les corpus ESTER 1&2 (0.7 %). Il s'agit de corpus contenant principalement des journaux radiophoniques avec un seul présentateur ou des magnétos (émissions préenregistrées et montées), dans lesquels il ne devrait pas y avoir de parole superposée.

Distribution statistique des durées des segments de parole superposée

Connaissant les proportions globales des segments de parole superposée, il est intéressant de regarder plus en profondeur la répartition statistique de ces segments de parole superposée, que ce soit la durée moyenne et l'écart type, ou bien la distribution complète de ces données pour étudier la forme des distributions de parole superposée et ainsi connaître le type de résultats que nous sommes supposés obtenir.

Protocole Nous réalisons cette analyse sur l'ensemble des corpus pour lesquels il existe des références. Nous traçons l'histogramme des durées des segments de parole superposée. Par souci de lisibilité, nous avons estimé la densité de probabilité à la place des histogrammes. Cette densité est obtenue à l'aide d'une interpolation Spline issue de la librairie scipy. Le nombre de bins est alors manuellement sélectionné pour avoir des variations globales de la courbe sans oscillations trop rapides.

Résultats et discussion Sur la figure 3.2 nous pouvons confirmer la présence de plusieurs types de corpus, symbolisés par des couleurs différentes. Premièrement les corpus très peu interactifs, auxquels appartiennent ESTER 1 et 2, qui, comme présentés précédemment ne contiennent que très peu de parole superposée, la courbe associée est presque plate, en ne privilégiant pas de durée particulière.

Les corpus contenant des débats réalisés par des professionnels de la communication, auxquels appartiennent EPAC, REPERE, ETAPE et le sous corpus d'ALLIES appelé ALLIES_LCP_debate. Les distributions présentent une surreprésentation des segments entre 0.5 et 1 s.

La dernière catégorie contient les corpus AMI et DIHARD. Il s'agit de corpus contenant de la parole spontanée, avec des réunions pour AMI, et des situations diverses pour

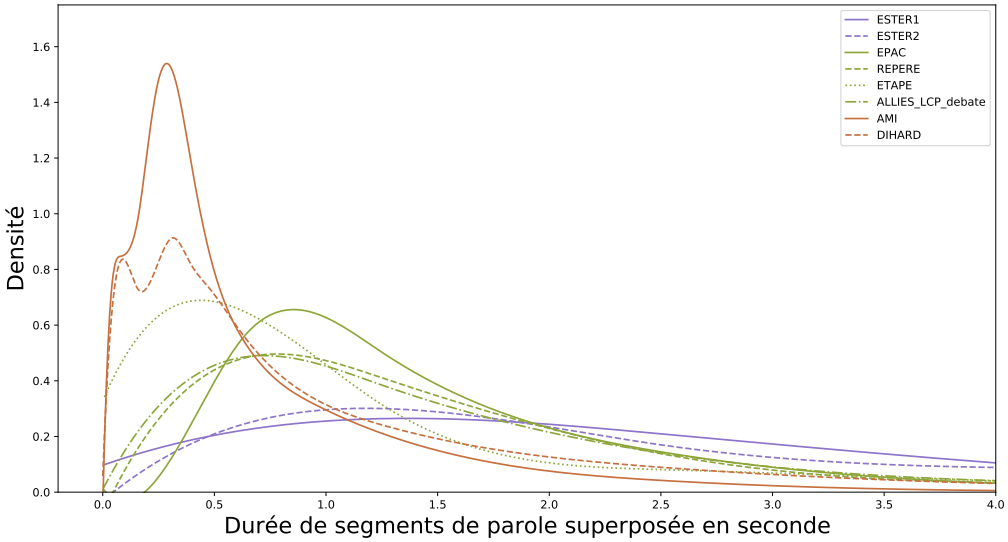


FIGURE 3.2 – Distribution des durées des segments de parole superposée par corpus

DIHARD. Ici nous supposons, l'anglais étant proche culturellement du français, que la langue n'influe pas sur la distribution de ces zones.

L'observation des distributions permet de confirmer que notre sélection d'émissions pour créer le sous-corpus ALLIES_LCP_debate est pertinente pour le rapprocher de la parole spontanée. Grâce à cette analyse, nous pouvons également confirmer l'hypothèse avancée par Adda et al. [Add+07] que la parole superposée est un phénomène court. Dans le cas des corpus spontanés, la majorité des segments de parole superposée sont d'une longueur inférieure à une seconde.

Deuxième partie

Contributions

Chapitre 4

Détecteur automatique d'activité vocale et de parole superposée

La première brique de notre détecteur d'interruptions consiste en un détecteur de parole superposée. Nous allons donc dans ce chapitre détailler les différentes expériences menées dans le but d'obtenir un détecteur de parole superposée le plus efficace possible.

4.1 Entraînement des systèmes de segmentation

4.1.1 Constitution de l'architecture de segmentation

Les tâches de détection d'activité vocale et de parole superposée étudiées dans cette partie sont similaires. En effet, elles peuvent toutes deux être définies comme des cas particuliers de comptage de locuteur, avec au moins un locuteur, ou au moins deux locuteurs. Elles sont par conséquent toutes deux des tâches de segmentation par classification binaire, cherchant des informations sur la présence de locuteurs. Les systèmes permettant de les traiter peuvent donc être harmonisés pour former une architecture générique de segmentation audio. L'ensemble de ce système de segmentation se base sur une architecture en deux parties, l'extraction de caractéristiques et le modèle de classification.

Représentations du signal audio

Pour toutes ces expériences, deux types de caractéristiques sont utilisés.

La première caractéristique utilisée se base sur les MFCCs. Cette extraction, décrite en section 2.2.1 est résumée par la figure 4.1. Les paramètres de ces MFCCs suivent ceux définis dans pyannote [Bre+20], soit 20 coefficients avec les delta et delta seconde auxquels

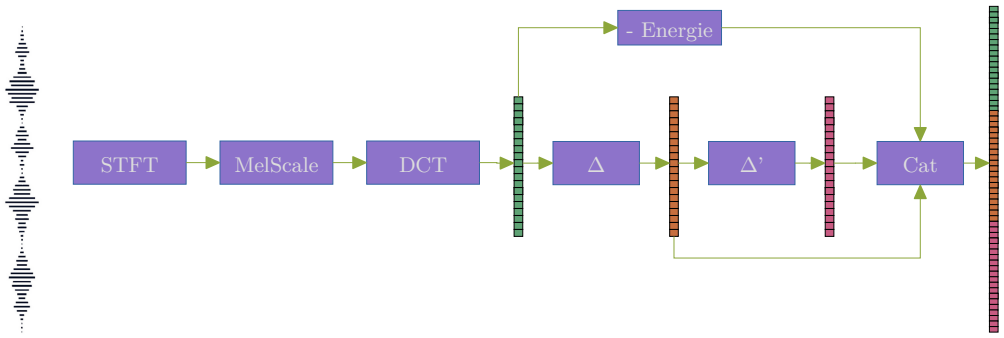


FIGURE 4.1 – Extraction de caractéristiques basée sur des MFCCs

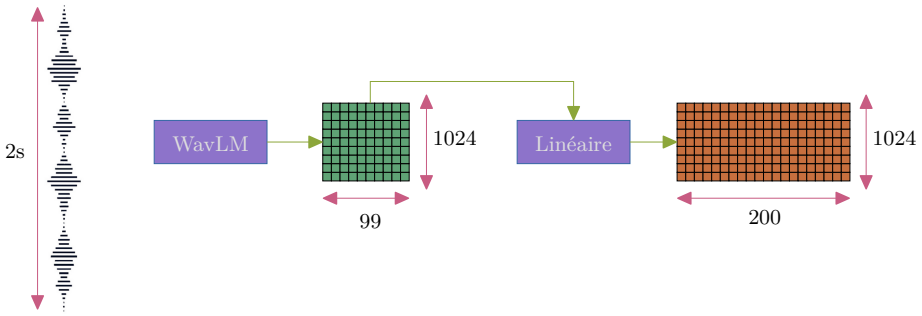


FIGURE 4.2 – Extraction de caractéristiques basée sur un modèle WavLM

on retire l'énergie pour un total de 59 dimensions. Toujours comme dans pyannote, le mel-spectogramme d'où est tiré ces coefficients est extrait tous les 10 ms sur une fenêtre de 30 ms avec 80 filtres pour obtenir des segments de 2 s, soit une représentation de dimension $MFCC \times trames$ de 59×200. Une différence notable est dans la méthode de passage à l'échelle mel qui est effectuée avec Slaney dans Pyannote et HTK dans nos expériences.

La seconde représentation utilisée est un ensemble de caractéristiques extraites par un système WavLM Large sur des fenêtres de 2 s, soit 1024 caractéristiques par trame de 25 ms. Nous obtenons alors 99 trames pour 2 s. La centième trame n'est pas extraite.

La figure 4.2 représente la méthode d'extraction de caractéristiques à partir d'un modèle WavLM. Nous introduisons également une couche de projection temporelle linéaire, désignée par PTL. La référence étant échantillonnée à 100 Hz, nous utilisons cette couche PTL afin de réaliser une interpolation quasi-linéaire des 99 trames par seconde vers 200 trames par seconde. Il existe cependant d'autres solutions pour résoudre ce problème de dimensions. Le choix réalisé par les créateurs de pyannote pour résoudre ce problème est

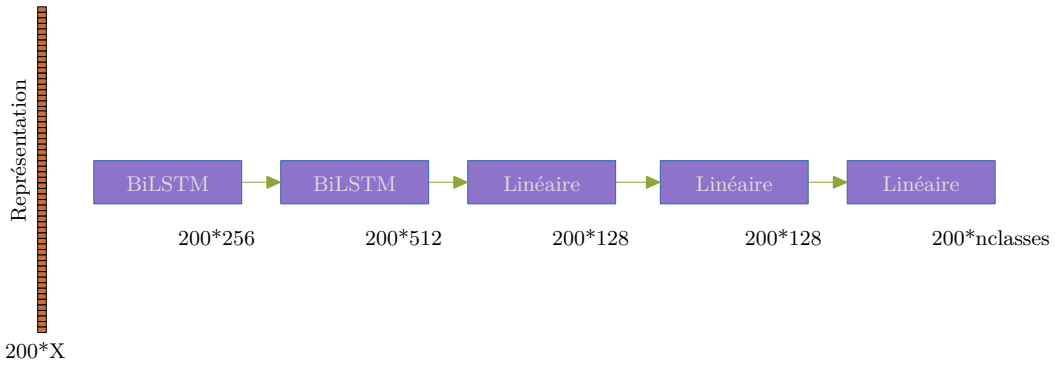


FIGURE 4.3 – Réseau récurrent inspiré du système de pyannote appelé ROSD

d'aligner la référence à la prédiction en diminuant la fréquence d'échantillonnage. Nous avons choisi la première solution pour permettre la comparaison directe avec les MFCCs et autres caractéristiques utilisées dans la communauté de détection de parole superposée qui sont par défaut échantillonnées à 100 Hz.

Par la suite, nous désignerons l'extraction de caractéristiques utilisant les MFCC par le nom MFCC et celle utilisant un modèle pré-entraîné WavLM par WavLM.

Système

Deux architectures de classifications sont comparées dans cette expérience.

Le premier appelé ROSD (*Recurrent Overlapped Speech Detector*) est basé sur le système PyanNet présent dans pyannote qui proposait au début de ces travaux des performances à l'état de l'art. Bien que le nom désigne explicitement un réseau de détection de parole superposée, ce système n'est ici pas utilisé uniquement pour cette tâche, le nom est cependant conservé car l'architecture est strictement identique pour différentes tâches. Comme présenté dans la figure 4.3, il contient deux couches BiLSTM de taille 128 suivies de deux couches linéaires de taille 128 également et d'une couche de sortie de taille égale au nombre de classes pour avoir une pseudo-probabilité de sortie pour chacune des classes prédites.

Le second système est une adaptation du TCN proposé par Cornell et al. [Cor+20]. Celui-ci est originellement un système de comptage de locuteurs, tâche qui semblait proche de la tâche de détection de parole superposée. Un système de détection de parole superposée peut en effet être dérivé d'un système de comptage de locuteur avec deux classes, $<$ deux locuteurs et $>$ deux locuteurs. Nous faisons l'hypothèse que la détection de parole

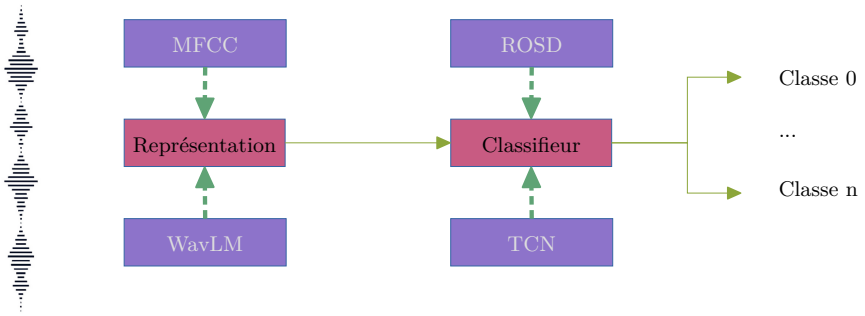


FIGURE 4.4 – Combinaisons des entrées et modèles présentés

superposée est par conséquent plus simple que le comptage de locuteur. L'architecture est donc adaptée pour avoir une couche de sortie binaire correspondant à la présence de parole superposée. Ce système semble prometteur car combinant les avantages des réseaux convolutifs de partage de poids et de reconnaissance de motifs avec la conservation de l'aspect temporel très important aux tâches séquence vers séquence. Bien que développé à l'origine pour la détection de parole superposée, ces systèmes fonctionnent également pour la détection d'activité vocale. Les noms ROSD et TCN sont conservés pour désigner les architectures, bien qu'elles seront utilisées pour des tâches différentes de la détection de parole superposée.

La combinaison de ces architectures de classifications avec les représentations se fait comme présenté dans la figure 4.4. On compare alors quatre types d'architectures pour les tâches de détection de parole superposée et d'activité vocale.

4.1.2 Détection d'activité vocale

Pour les traitements que nous allons effectuer, la première étape consiste en une détection des zones dans lesquelles au moins un locuteur est actif. Nous avons donc commencé à travailler sur la détection parole/non-parole pour obtenir une segmentation. Les travaux réalisés dans ce cadre ont été faits en collaboration avec Théo Mariotte, également doctorant au LIUM.

Objectif

Nos travaux en détection de parole superposée, ainsi que ceux menés en détection d'interruptions, se basent sur un signal de parole. Comme présenté dans la section 2.3.3,

la détection de parole/non-parole ou détection d'activité vocale est une tâche utile en prétraitement d'autres. Nous souhaitons par conséquent disposer d'un système interne pouvant effectuer cette tâche avec des performances proches de celles de l'état de l'art.

Protocole expérimental

Ces expériences sont menées sur le corpus DIHARD présenté précédemment avec comme distribution de données celle utilisée lors de la campagne DIHARD III par l'équipe formée par le consortium Intel/IRIT/LIUM. Nous avons choisi de nous concentrer sur les données de ce corpus pour voir les comportements d'un système de détection de parole dans les cas limites. En effet, comme mentionné dans la section 3.1.1, ce corpus contient des données issues de nombreuses situations d'enregistrement, avec beaucoup de bruit et de parole superposée pour une durée de 34 h. Les fichiers audio de ce corpus étant présentés au format flac, ils ont été convertis en wav à une fréquence de 16 kHz. Nous utilisons comme cible un vecteur binaire échantillonné à 100 Hz extrait des annotations en locuteur avec 1 indiquant la présence de locuteur et 0 l'absence. Pour cette tâche, l'optimisation est réalisée avec un SGD sans scheduler et appris avec une categorical cross entropy comme fonction de coût, ces paramètres étant des paramètres efficaces pour la plupart des tâches de classification[Cho+19].

Résultats et Discussion

Comme présenté lors de l'état de l'art, les protocoles utilisés en détection de parole/non-parole varient fortement d'un article à l'autre. Nous devons donc choisir un article particulier pour évaluer l'efficacité de notre modèle par rapport à l'état de l'art. Nous choisissons donc "End-to-end speaker segmentation for overlap-aware resegmentation" écrit par Hervé Bredin et Antoine Laurent en 2021 [BL21], qui à l'époque présentait les meilleurs résultats à notre connaissance en détection de parole superposée. Cette tâche étant notre objectif principal, il est pertinent de comparer nos résultats aux leurs en détection d'activité vocale, également évaluée bien que ce ne soit pas leur tâche principale. Il est cependant tout à fait possible que certains résultats en détection de parole soient meilleurs dans d'autres travaux.

Tout nos systèmes présentés dans le tableau 4.1 obtiennent des performances satisfaisantes en étant de manière significative au-dessus de notre référence. En analysant nos résultats, nous notons deux phénomènes qui étaient attendus au vu des observations réalisées dans la section 2.5. Tout d'abord, Le TCN améliore les résultats par rapport à

Input	Architecture	Paramètres	Précision	Rappel	F1-score	DetER
MFCC	ROSD	0.638 M	95.0	97.4	96.2	6.48
MFCC	TCN	0.268 M	95.8	97.2	96.5	5.89
WavLM	ROSD	1.647 M	96.9	97.3	97.1	4.84
WavLM	TCN	0.352 M	97.0	97.3	97.2	4.73
Bredin et Al. [BL21]			/	/	/	7.3

TABLE 4.1 – Résultats de détection de parole/non-parole sur le corpus de test de DIHARD avec le F1-score, la Précision, le Rappel et le FA+Miss exprimés en %

un réseau récurrent et confirme par conséquent la possibilité de remplacement de cette seconde architecture par la première. Nous remarquons un gain absolu de 0.59 points de pourcentage de DetER (Detection Error Rate) pour un TCN par rapport au ROSD quand des MFCCs sont utilisés, et un gain de 0.11 points de pourcentage pour ces mêmes systèmes quand WavLM est utilisé. L'architecture de classification utilisée a donc une importance limitée dans les résultats. Cependant, le plus gros gain se trouve dans la représentation du signal de parole utilisée. En effet, les représentations pré-entraînées donnent un gain absolu de 1.64 points de pourcentage par rapport à des MFCCs pour le réseau récurrent ainsi qu'un gain de 1.16 points de pourcentage pour un TCN comme architecture de classification.

Nous pouvons donc en conclure que pour la tâche de détection d'activité vocale, les représentations du signal de parole sont plus importantes que les architectures de classification utilisées.

4.1.3 Détection de parole superposée

Nous avons présenté dans la section précédente un détecteur d'activité vocale performant. Nous pouvons donc partir de cette architecture pour construire un système de détection de parole superposée qui, pour chaque trame, donne une prédiction binaire. Pour avoir une idée des éléments qui améliorent les résultats d'un système, plusieurs combinaisons de systèmes et de caractéristiques d'entrées décrites précédemment sont testées.

Objectif

La première brique de notre objectif final nécessite d'obtenir un système de détection de parole superposé. Notre système global est en cascade, c'est à dire que les erreurs des

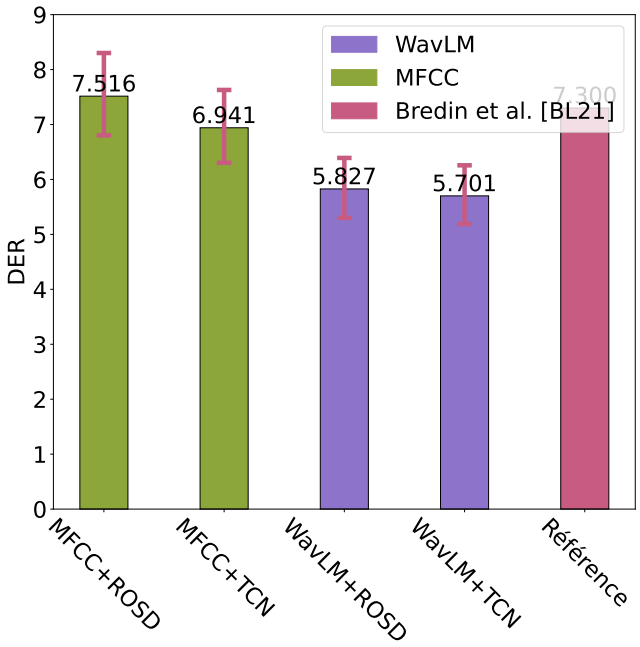


FIGURE 4.5 – DetER(FA+Miss) du détecteur d'activité vocale sur le corpus de test de DIHARD. Les barres d'erreur sont calculées sur les DetER des différents fichiers

premières briques impactent toujours les systèmes suivants. Nous avons donc besoin du meilleur système possible.

Protocole expérimental

Cette expérience est similaire à celle menée pour la détection d'activité vocale. Nous considérons donc un corpus identique à celui utilisé pour cette tâche. Les paramètres d'apprentissage, optimiseur, taille de batch et fonction de coût sont également identiques à l'expérience précédente. La seule différence tient dans la référence utilisée. Nous utilisons un vecteur binaire avec *1* indiquant la présence de 2 locuteurs et plus au lieu de 1 locuteur et plus, et *0* la présence de 0 ou 1 locuteur.

Résultats et Discussion

Pour comparer nos résultats avec l'état de l'art actuel, nous utilisons les résultats obtenus sur le même corpus, avec la même distribution par Bredin et al. [BL21]. Dans le tableau 4.2, rendue plus lisible par la figure 4.6, plusieurs éléments sont mis en valeurs. Tout d'abord les résultats obtenus avec WavLM (notés en pointillés), soit un F1-score de 62.3 et de 63.4 pour l'utilisation d'un réseau récurrent et d'un TCN respectivement, sont

Input	Architecture	Paramètres	Précision	Rappel	F1-score
MFCC	ROSD	0.638 M	34.2	60.8	43.8
MFCC	TCN	0.268 M	46.6	59.8	52.4
WavLM	ROSD	1.647 M	61.0	63.6	62.3
WavLM	TCN	0.352 M	60.1	67.1	**63.4**
	Bredin *et al.* [BL21]		57.2	62.8	59.9

TABLE 4.2 – Résultats de détection de parole superposée sur le corpus de test de DIHARD avec le F1-score, la Précision et le Rappel exprimés en %

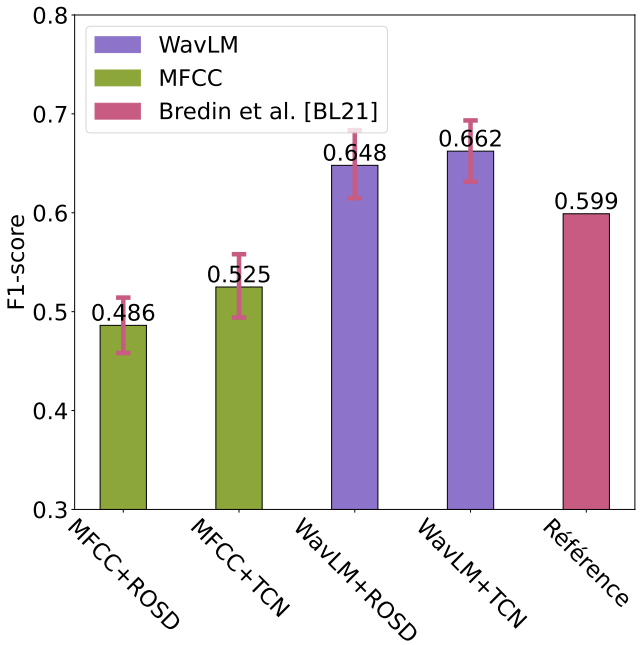


FIGURE 4.6 – F1-score du détecteur de parole superposée sur le corpus de test de DIHARD. Les barres d'erreur sont calculées sur les F1-score des différents fichiers

bien meilleurs que ceux obtenus pour des MFCCs (notés en hachuré), respectivement 43.8 et 52.4. Le TCN (codé en vert) a également des performances légèrement meilleures que le réseau récurrent (codé en violet) avec un gain faible pour l'utilisation de WavLM en entrée, de 62.3 à 63.4 mais plus conséquent pour l'utilisation des MFCCs, avec un F1-score qui passe de 43.8 à 52.4. Si on considère l'utilisation de WavLM comme la meilleure solution, l'intérêt principal du TCN n'est pas dans le gain de F1-score minime, mais plutôt dans le gain de paramètres. En effet, l'utilisation de couches convolutives permet de passer de 1.647 millions de paramètres à 352 000 soit un gain en nombre de paramètres d'environ 450 %. Cette propriété permet d'utiliser de plus petits corpus, avec des données mieux choisies, c'est-à-dire avec une plus grande variété de situation et des annotations plus propres. Cependant, avec l'usage de WavLM, on perd la possibilité d'interpréter les sorties du modèle, et de découvrir de possibles pistes d'amélioration en regardant les caractéristiques d'entrées des zones en erreur. En effet, les dimensions ne semblent pas présenter de motif interprétable pour un humain.

Pour nos deux systèmes, nous observons un gain de performances en utilisant WavLM comme couche d'entrée à la place de MFCC et un TCN à la place d'un ROSD. Ce gain est cependant plus important pour le système de détection de parole superposée, par rapport au système d'activité vocale. Nous faisons l'hypothèse que la simplicité de la tâche de détection d'activité vocale rend la marge de gain minime par rapport à la détection de parole superposée, qui est moins avancée.

4.1.4 Choix de l'optimiseur

Dans le but d'améliorer le système, une des possibilités est de changer l'optimiseur utilisé. Bien que considéré plus stable par des travaux [Cho+19], Le SGD peut être remplacé par d'autres optimiseurs, dont un très utilisé, Adam [KB14]. L'expérience suivante compare donc les résultats obtenus précédemment avec un SGD et ceux obtenus avec Adam pour une tâche de détection de parole superposée.

Protocole Experimental

Cette expérience utilise donc les données du corpus DIHARD, avec les modèles présentés dans la figure 4.4. Les résultats présentés pour le SGD sont ceux présentés précédemment et les résultats présentés pour ADAM sont obtenus après un apprentissage de 30 époques pour avoir la meilleure époque en termes de F1-score sur la partition de

Input	Architecture	Optimiseur	Précision	Rappel	F1-score
MFCC	ROSD	SGD	34.2	60.8	43.8
MFCC	TCN	SGD	46.6	59.8	52.4
WavLM	ROSD	SGD	61.0	63.6	62.3
WavLM	TCN	SGD	60.1	67.1	63.4
MFCC	ROSD	ADAM	41.9	57.9	48.6
MFCC	TCN	ADAM	47.1	59.2	52.5
WavLM	ROSD	ADAM	62.3	67.5	64.8
WavLM	TCN	ADAM	65.1	67.5	66.3

TABLE 4.3 – Résultats de la détection de parole superposée pour deux optimiseurs différences sur le corpus de test de DIHARD avec le F1-score, la Précision et le Rappel exprimés en %

développement. Les hyper-paramètres utilisés sont les paramètres par défaut de PyTorch avec un learning rate à 0.001 et une taille de batch de 32. Nous avons choisi de garder ces paramètres sans les optimiser plus, l'objectif n'étant pas de créer un modèle prêt pour la production, mais plutôt de tester l'influence de paramètres importants. Ce choix permet d'économiser du temps de calcul qui n'aurait pour seul impact d'apporter un très léger gain de performance.

Résultats et discussion

Les résultats obtenus lors de cette expérience sont présentés dans le tableau 4.3 et résumés dans la figure 4.7. Plusieurs éléments ressortent de ces résultats. Avec SGD (en violet), l'usage d'un TCN à la place d'un ROSD apporte une amélioration relative de 19.6 % pour des MFCC et de 1.8 % pour WavLM. Avec ADAM (en vert) l'usage d'un TCN à la place d'un ROSD apporte une amélioration relative de 16.6 % pour des MFCC et de 2.3 % pour WavLM. L'influence des caractéristiques d'entrée sur l'efficacité d'un optimiseur semble plus important que l'architecture utilisée. L'utilisation de WavLM favorise visiblement la convergence du modèle avec les deux optimiseurs testés. Pour tous nos cas d'utilisation, ADAM a de meilleurs résultats que SGD, nous utiliserons donc cet optimiseur à l'avenir.

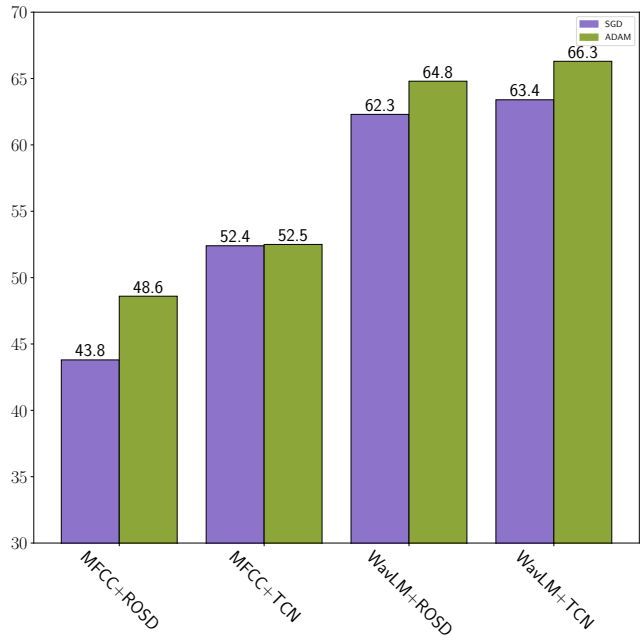


FIGURE 4.7 – F1-score d'un système de détection de parole superposée en fonction de l'optimiseur utilisé. SGD est représenté en violet et ADAM en vert.

4.1.5 Détection jointe parole, parole superposée

Pour améliorer l'efficacité et la rapidité de segmentation en parole/non-parole et en parole superposée, nous entraînons un modèle pouvant traiter ces deux tâches. Pour traiter des tâches jointes, plusieurs stratégies sont possibles. Nous avons considéré le multilabel qui peut sortir plusieurs labels pour une trame et multitâche qui ne sort qu'un label par trame. Nous avons choisi d'utiliser un système multi-tâche pour simplifier le passage des systèmes à deux classes vers le système à trois classes.

Système

La création d'un modèle à trois classes, comme pour le modèle à deux classes compare deux modèles et deux ensembles de caractéristiques. Le premier modèle utilisé est à base de TCN. Ce système, présenté dans la figure 4.4, ressemble fortement au TCN à deux classes utilisé précédemment, la couche de sortie finale est remplacée pour sortir trois classes au lieu de deux. Le second système utilise des BiLSTM et couches linéaires pour classifier les caractéristiques extraites. Comme pour l'expérience décrite en 4.1.1 les caractéristiques comparées sont un ensemble de MFCC et un système WavLM contenant la couche PTL. Le système utilisé tente de prédire une des trois classes suivantes : aucun locuteur, un

	MFCC			WavLM		
Architecture	Précision	Rappel	F1-score	Précision	Rappel	F1-score
ROSD_3c	43.97	58.32	50.14	59.09	71.31	64.63
TCN_3c	44.66	64.16	52.66	59.45	71.96	65.11
ROSD_2c	41.9	57.9	48.6	62.3	67.5	64.8
TCN_2c	47.1	59.2	52.5	65.1	67.5	66.3
Bredin *et al.* [BL21]	57.2	62.8	59.9			

TABLE 4.4 – Résultats de la détection de parole superposée pour les différents systèmes à trois classes sur le corpus de test de DIHARD avec le F1-score, la Précision et le Rappel exprimés en %. Les résultats présentés dans le tableau 4.3 sont reportés pour comparaison.

	MFCC				WavLM			
Archi	Précision	Rappel	F1-score	FA+Miss	Précision	Rappel	F1-score	FA+Miss
ROSD_3c	95.42	97.17	96.29	6.3	96.95	97.09	97.02	4.95
TCN_3c	96.04	97.21	96.62	5.68	96.45	97.49	96.97	5.06
ROSD_2c	95.0	97.4	96.2	6.48	96.9	97.3	97.1	4.84
TCN_2c	95.8	97.2	96.5	5.89	97.0	97.3	97.2	4.73

TABLE 4.5 – Résultats de la détection d'activité vocale pour les différents systèmes à trois classes sur le corpus de test de DIHARD avec le F1-score, la Précision et le Rappel exprimés en %. FA+Miss correspond au DetER présenté. Les résultats d'un système à deux classes présentés dans le tableau 4.1 sont reportés pour comparaison.

locuteur, ou plus d'un locuteur. Pour extraire des segmentations à partir de ces résultats, on considère que la classe 2 est une zone de parole superposée et les classes 1 et 2 sont des zones de parole.

Protocole

Comme montré précédemment, Adam est plus performant que SGD pour notre tâche, nous utiliserons donc cet optimiseur pour toute cette expérience. Pour rester comparable avec le système à deux classes, le réseau est entraîné et testé sur le corpus DIHARD, avec un système WavLM en entrée, et une cross-entropy en loss. Le système est évalué séparément pour parole superposée et parole/non-parole.

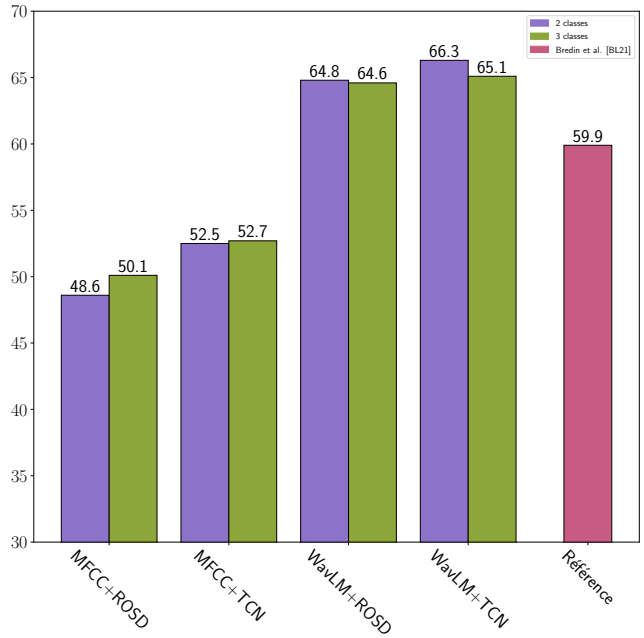


FIGURE 4.8 – F1-score de détection de parole superposée obtenu avec un modèle à trois classes comparé avec un système à deux classes.

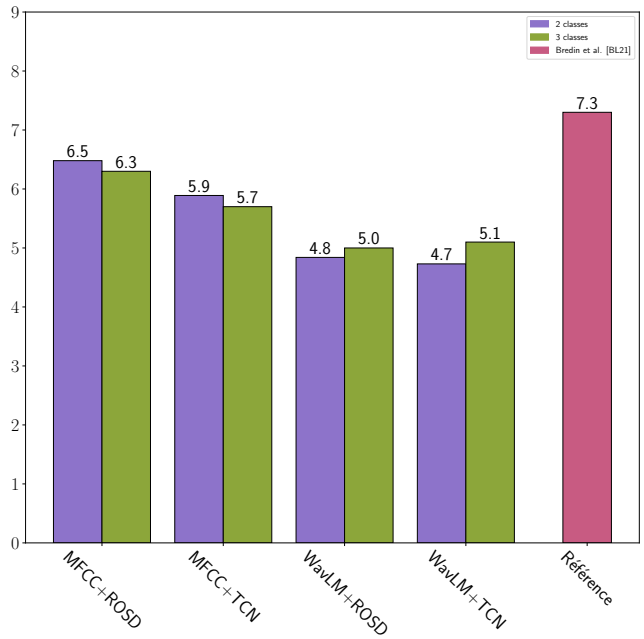


FIGURE 4.9 – DetER d'un détecteur d'activité vocale obtenu avec un modèle à trois classes comparé avec un système à deux classes.

Résultats et discussion

Les résultats en VAD, présentés dans le tableau 4.5 et dans la figure 4.9 ne varient que très peu entre les systèmes à trois classes et deux classes. Nous nous concentrerons donc sur la détection de parole superposée. Les résultats obtenus lors de cette expérience sont présentés dans le tableau 4.4 et résumés dans la figure 4.8. Nous pouvons les comparer avec les résultats obtenus avec les systèmes à deux classes obtenus également avec un optimiseur ADAM pour permettre la comparaison. Ces résultats ont été reportés dans le même tableau pour plus de lisibilité.

Tout d'abord, les résultats obtenus par le système joint sont légèrement moins bons que ceux obtenus par des systèmes spécialisés. Le F1-score pour un système utilisant une classification basée sur un TCN et un WavLM comme représentation passe de 66.3 à 65.11 soit une perte de 1.2 points de F1-score. Cependant, pour le cas d'une utilisation de MFCC avec une architecture récurrent, le F1-score obtenu avec le système joint est meilleur qu'avec des systèmes séparés.

Enfin, bien que présentant une perte de performance absolue, cette diminution peut être négligeable sur des applications peu sensibles. Le gain de temps alors obtenu est intéressant et justifie la préférence de ce système à trois classes par rapport à 2 systèmes à deux classes.

4.2 Analyses Complémentaires

Dans les sections précédentes, nous avons présenté des systèmes de détection de parole/non-parole et de parole superposée obtenant des performances satisfaisantes. Pour aller plus loin, cette section regroupe un ensemble d'expériences menées sur ces systèmes pour en comprendre mieux les résultats.

4.2.1 Influence du domaine

Objectifs

Dans ces deux premières expériences, nous souhaitons analyser l'importance du domaine des données, c'est à dire la situation d'acquisition. Pour cela, nous étudions dans un premier temps nos résultats sur les différents domaines présent dans le corpus DIHARD. Dans un second temps, nous étudions les résultats de systèmes appris sur un corpus et

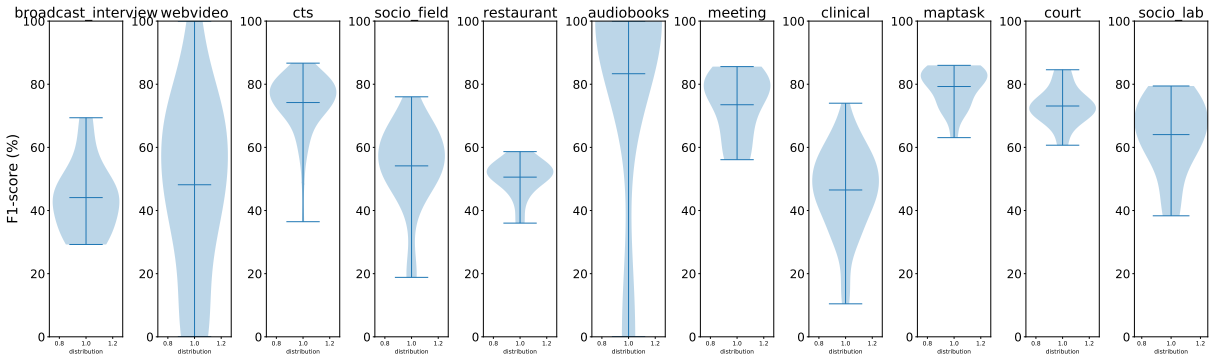


FIGURE 4.10 – F1-score obtenu sur les différents domaines du corpus DIHARD

testés sur un second. Cette expérience varie de la première par l'absence de données du domaine d'entraînement dans le domaine de test.

Études des sous corpus de DIHARD

Protocole La première expérience est une comparaison des résultats obtenus sur le corpus DIHARD à partir du TCN_3c pour les différents corpus contenus dans DIHARD. Pour cela nous utilisons les metadatas fournies par les créateurs du corpus qui contiennent le domaine de chaque show pour obtenir 11 classes. Le domaine "interview"(3.8 locuteurs par émission en moyenne) contient des entretiens entre plusieurs locuteurs dans un contexte télévisé. Le domaine "webvideo"(4 locuteurs par émission en moyenne) contient un recueil de vidéos obtenues sur YouTube avec des sujets et des formats divers. Le domaine "cts" (2 locuteurs par émission) contient des conversations téléphoniques. Le domaine "socio lab" (2 locuteurs par émission) contient des conversations avec des qualités d'enregistrements mauvaises menés par un interviewer. Le domaine "restaurant" (7.2 locuteurs par émission en moyenne) contient des conversations dans des restaurants. Le domaine "audiobook" (1 locuteur par émission en moyenne) contient de la lecture. Le domaine "meeting" (5.4 locuteurs par émission en moyenne) contient des enregistrements de réunion. Le domaine "clinical" (2 locuteurs par émission) regroupe des entretiens réalisés dans un contexte médical. Le domaine "maptask" (2 locuteurs par émission) contient un jeu ou un participant tente de guider un second sur une carte. Le domaine "court"(6.9 locuteurs par émission en moyenne) qui contient des enregistrements de tribunal. Et enfin le domaine "sociofield" (3.5 locuteurs par émission) contient des entretiens menés avec des groupes de personnes dans un environnement contrôlé.

Résultats/Discussion La figure 4.10 présente le F1-score obtenu sur différents domaines du corpus DIHARD avec des conditions différentes. Par exemple, les zones de paroles superposées contenues dans des discussions avec un objectif défini, comme le domaine "cts" dans lequel les participants devaient discuter d'un sujet précis, le domaine "map task" qui est clairement orienté autour d'une tâche requérant une attention des deux participants, et le domaine "court" qui traite de données réelles issues de situations de travail, de même que pour le domaine "meeting". Les données "clinical", "broadcast_interview", "restaurant" ainsi que "scociolab" et "sociofield" sont des conversations plus spontanées dans laquelle l'attention des deux participants n'est pas requise. Le domaine "phone" aurait pu se trouver dans cette catégorie, l'hypothèse peut être faite que le fait de ne pas voir l'interlocuteur intervient également dans la structuration du tour de parole. Cette hypothèse est également valable pour le domaine "maptask" dans lequel les participants ne se voient pas. Le domaine "audiobook" est particulier car ne contenant normalement pas de parole superposée. Le F1-score est donc supposément 100 % si le système n'a pas prédit de parole superposée ou 0% si n'importe quel segment est prédit. Les segments présents dans les domaines "sociofield", "clinical" et "restaurant" ont également beaucoup de bruit et de variété dans les types d'enregistrements. On peut donc en conclure que la détection de parole superposée dépend en partie du type de données utilisé, ou encore de la modalité de la conversation, par exemple dyadique, en groupe, en présentiel, à distance. On peut également émettre l'hypothèse que la qualité du signal enregistré influe sur les performances, le bruit posant encore des problèmes au détecteur de parole superposée. Cette conclusion dessine naturellement des pistes d'améliorations et nouveaux objectifs pour cette tâche. En effet, comme montré par la campagne d'évaluation DIHARD, les systèmes de traitement de parole manquent de robustesse aux dégradations de signal et au bruit ambiant.

Étude en corpus croisé

Protocole La seconde expérience porte sur un apprentissage et une évaluation en corpus croisé. C'est à dire évaluer une même architecture apprise sur différents corpus sur les autres corpus disponibles. Pour cette expérience, nous utilisons le TCN à 3 classes présenté dans la section 4.1.5 ainsi les corpus ALLIES, DIHARD et AMI.

Résultats et discussion Cette expérience permet de montrer plusieurs phénomènes. Premièrement comme présenté dans le tableau 4.6 ainsi que dans la figure 4.11 résumant

Corpus		Score		
Entraînement	**Test**	**Précision**	**Rappel**	**F1-score**
DIHARD	DIHARD	65.88	67.89	**66.87**
DIHARD	AMI	85.35	61.95	71.79
DIHARD	ALLIES	64.84	74.23	69.22
AMI	DIHARD	62.74	50.79	56.14
AMI	AMI	81.55	79.23	**80.38**
AMI	ALLIES	61.58	86.95	72.10
ALLIES	DIHARD	74.07	18.61	29.75
ALLIES	AMI	81.52	54.87	65.59
ALLIES	ALLIES	75.93	74.82	**75.37**

TABLE 4.6 – Résultats de la détection de parole superposée pour différents modèles et différents corpus avec le F1-score, la Précision et le Rappel exprimés en %

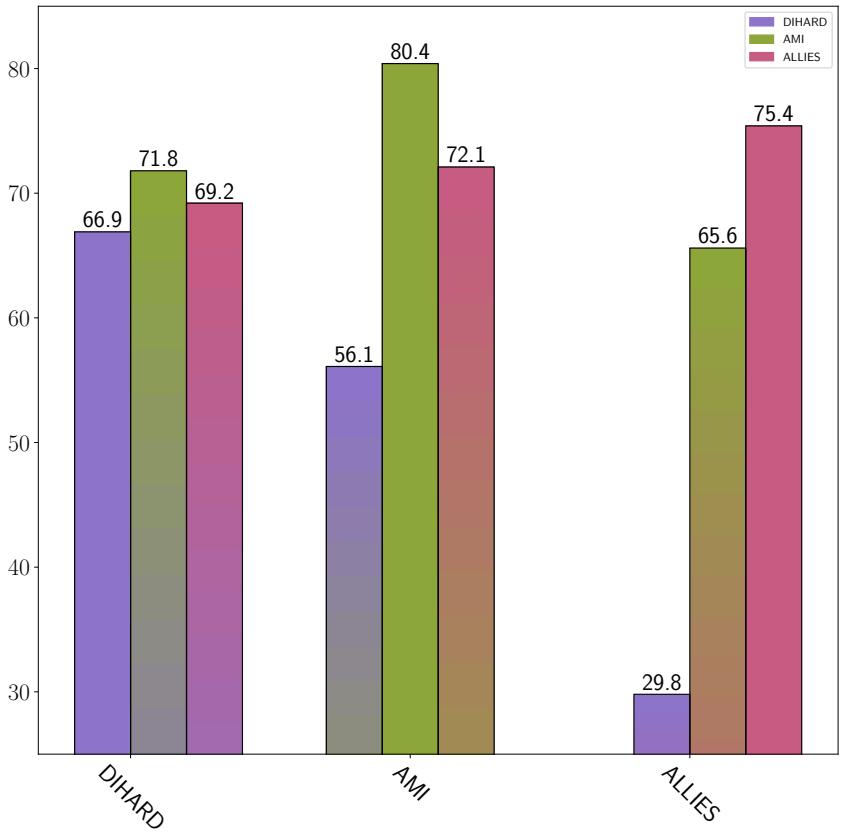


FIGURE 4.11 – F1-score de détection de parole superposée pour des modèles à 3 classes entraînés sur un corpus et testé sur un autre

les f1-score de ce tableau, ceux-ci sont obtenus pour des modèles entraînés sur les mêmes corpus que ceux sur lesquels ils sont testés. Ce résultat montre encore une fois l'importance du domaine des données sur les performances d'un système. Le second phénomène intéressant à observer concerne les précisions et rappels. La quantité de parole superposée des corpus originaux influe sur la balance précision rappel de la prédiction. Ce phénomène est particulièrement visible pour le corpus ALLIES qui comporte très peu de parole superposée. Le modèle appris sur ALLIES (en rose) a donc une bonne précision sur les deux autres corpus, dans le cas de DIHARD (en violet), la précision est même meilleure que celle obtenue pour le modèle issu de ce corpus. En contrepartie, le rappel obtenu est très bas. Cela montre donc l'importance d'avoir plusieurs types de situations d'acquisitions de données dans un corpus d'apprentissage, que des fichiers avec peu de parole superposée ne sont pas nécessairement moins intéressants que des fichiers avec une haute proportion de parole superposée.

4.2.2 Adaptation au domaine

Objectifs

L'influence du domaine maintenant constatée, il peut être intéressant de trouver des moyens pour améliorer un système sans en réentraîner un nouveau de zéro. La technique la plus évidente est l'adaptation au domaine, c'est à dire entraîner un système pré-appris sur un domaine avec des données issues d'un nouveau domaine. Un système plus général permettrait de limiter la masse nécessaire pour traiter différents types de données. Notre cascade ayant déjà deux systèmes distincts, limiter le nombre de modèles du premier système permet de faciliter l'intégration dans le second.

Source/cible	DIHARD	AMI	ALLIES
DIHARD	**66.87**	61.22	48.64
AMI	60.17	56.14	48.15
ALLIES	60.12	57.31	29.75

TABLE 4.7 – Résultats en F1-score exprimé en % sur le test de DIHARD pour des corpus adaptés depuis un corpus source vers un corpus cible. La diagonale correspond au modèle sans adaptation.

Protocole

Pour cette expérience, nous utilisons les corpus DIHARD, AMI et ALLIES, tentons différentes combinaisons d'adaptation et les évaluons en F1-score sur les partitions de test des 3 corpus sources. Les corpus DIHARD et AMI contiennent beaucoup de parole superposée par rapport au corpus ALLIES, mais sont beaucoup plus courts. Pour l'adaptation, les corpus entiers n'ont pas été utilisés, un sous-ensemble caractéristique a à chaque fois été extrait manuellement. Pour le corpus DIHARD, un fichier par domaine a été choisi pour un total de 1 h 15 d'audio. Pour le corpus ALLIES, le sous-ensemble caractéristique extrait contient 15 fichiers contenant de la parole superposée, provenant d'émissions différentes, pour un total de 7 h d'audio. Pour AMI, le corpus ne contenant pas de différents domaines, l'accent a été porté sur la diversité acoustique en choisissant deux fichiers par lieu d'enregistrement pour un total de 6h20 d'audio.

Pour cette expérience, nous choisissons d'utiliser le modèle à 3 classes présenté en section 4.1.5 appris sur l'un des 3 corpus. Nous relançons alors l'apprentissage à partir de l'extrait du corpus cible sans figer de couches du modèle original. Cela nous donne donc 9 combinaisons de modèles entraînés sur des données d'un domaine et adapté aux données d'un autre modèle, que nous évaluons sur chacun de ces trois corpus.

Résultats et Discussion

Le premier résultat sorti de cette expérience montre que dans le cas d'ALLIES, présenté dans le tableau 4.8 ou le cas d'AMI, dans le tableau 4.9, les résultats obtenus après une adaptation sont similaires à ceux obtenus avec un modèle pur, voir, dans le cas d'un modèle pré-entraîné sur AMI adapté sur ALLIES, surpasse les résultats originaux. Ces résultats sont obtenus pour une quantité de données très faible par rapport au corpus total, moins d'un quarantième des données pour ALLIES, un quinzième pour AMI et un

Source/cible	DIHARD	AMI	ALLIES
DIHARD	69.27	69.36	75.35
AMI	70.4	72.25	**75.76**
ALLIES	70.12	71.51	75.37

TABLE 4.8 – Résultats en F1-score exprimé en % sur le test de ALLIES pour des corpus adaptés depuis un corpus source vers un corpus cible. La diagonale correspond au modèle sans adaptation.

Source/cible	DIHARD	AMI	ALLIES
DIHARD	71.79	79.4	69.11
AMI	72.82	**80.38**	70.03
ALLIES	72.01	79.38	65.59

TABLE 4.9 – Résultats en F1-score exprimé en % sur le test de AMI pour des corpus adaptés depuis un corpus source vers un corpus cible. La diagonale correspond au modèle sans adaptation.

vingt-cinquième pour DIHARD. Ce résultat montre l'importance de la qualité/diversité des données d'apprentissage par rapport à la quantité. Le second résultat à noter concerne les performances sur le corpus source après adaptation. Si on considère que DIHARD et AMI sont des corpus proches, on peut remarquer que l'adaptation entre corpus proche dégrade moins les performances que l'adaptation à un corpus plus simple. On peut penser, dans le cas du 48.64 et 48.15 obtenu sur le corpus DIHARD (tableau 4.7) par des systèmes adaptés sur ALLIES, que des zones considérées comme parole superposée dans DIHARD, ne le sont pas dans ALLIES, ce qui dégrade hautement les performances. Enfin, pour l'adaptation en vue d'améliorer les performances sur un corpus cible, aucune tendance quant à la proximité ou difficulté du corpus source n'est clairement visible dans cette expérience.

4.2.3 Analyse de la couche PTL (Projection temporelle linéaire)

Lors de la présentation des différents systèmes de segmentation en parole superposée, nous avons présenté brièvement la couche PTL à la sortie de WavLM.

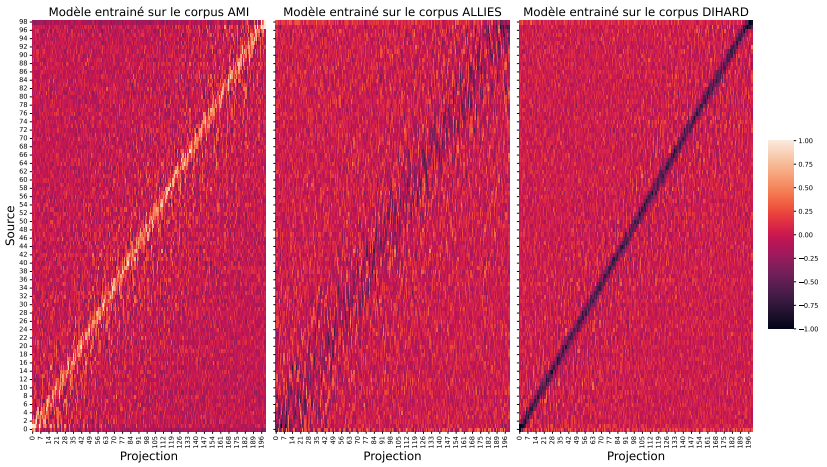


FIGURE 4.12 – Comparaison des poids de la couche PTL en sortie de WavLM pour les corpus ALLIES, AMI et DIHARD

Pourquoi utiliser cette couche ?

La présence de cette couche est due à un verrou technique. Un des soucis de WavLM pour notre utilisation est sa fréquence d'échantillonnage. En effet, celle utilisée est de 50 Hz, qui est différente de notre référence échantillonnée à 100 Hz. Nous ne pouvions également pas simplement dupliquer les échantillons en raison de la disparition de l'un d'eux, supposé être le dernier. Nous obtenons donc un vecteur de taille 99 pour un segment de 2 s d'audio, contre une référence à 200 points. La méthode choisie pour synchroniser ces deux vecteurs est de passer par une couche linéaire censée "interpoler" la sortie de WavLM pour la synchroniser avec la référence.

Un an et de multiples expériences plus tard, ce système fonctionnant, deux questions se posent.

— La couche a-t-elle le fonctionnement attendu ?

— Cette couche est-elle remplaçable.

Pour répondre à ces questions, nous avons observé les points de la couche PTL sur plusieurs modèles entraînés lors de l'expérience d'adaptation au domaine présentée en section 4.2.2.

Comme nous pouvons le voir dans la figure 4.12, toutes ces couches semblent converger vers des diagonales, qui ont le comportement attendu, soit de réaliser une pseudo-interpolation des caractéristiques WavLM vers un vecteur synchronisé avec la référence. Cependant, certaines couches sont inversées comme celles entraînées sur DIHARD et sur AMI, ce qui oblige le modèle de classification à apprendre une inversion à la suite de cette couche. Nous pouvons donc conclure que cette couche remplie partiellement la tâche pour

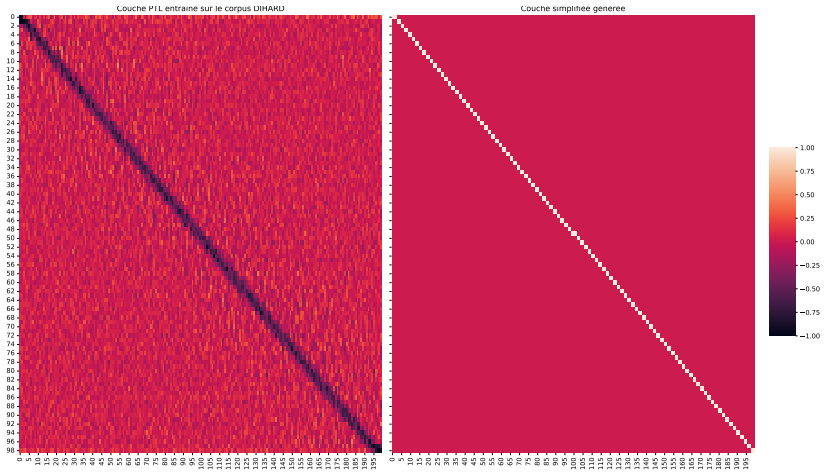


FIGURE 4.13 – Comparaison entre la couche PTL simplifiée et une couche linéaire entraînée sur le corpus DIHARD.

laquelle elle a été conçue.
La seconde partie de cette expérience consiste donc à la remplacer par une couche fixe crée à priori. Pour cela, nous utilisons l'algorithme suivant pour créer une matrice pseudo-diagonale à multiplier à notre entrée :

```
def create_mat(dim_in,dim_out):
    mat = torch.zeros((dim_in, dim_out))
    x = torch.arange(dim_out)
    x_mapped = x / dim_out * dim_in
    x_int = x_mapped.long()
    mat[x_int, x] = 1
    return mat
```

Le principe de cet algorithme est de projeter un vecteur de la taille de sortie vers l'intervalle de la couche d'entrée, ce qui a pour effet de le compresser en flottant. Une troncature de ces valeurs est alors utilisée pour obtenir la matrice pseudo-diagonale souhaitée. Cette matrice peut être également comparée à une seconde entraînée. Nous pouvons alors comparer les résultats obtenus avec cette couche simple et une couche linéaire apprise représentée en figure 4.13.

Couche	**Corpus**	**Précision**	**Rappel**	**F1-score**
Linéaire	DIHARD	65.88	67.89	66.87
Linéaire	AMI	81.55	79.23	80.38
Linéaire	ALLIES	75.93	74.82	75.37
Simple	DIHARD	64.05	68.48	66.19
Simple	AMI	77.95	80.34	79.13
Simple	ALLIES	71.15	77.25	74.08

TABLE 4.10 – Résultats en détection de parole superposée obtenus sur les trois corpus DIHARD AMI et ALLIES avec un système entraîné sur le corpus correspondant au test évalué. La couche linéaire est apprise durant l'entraînement et la couche simple correspond à la diagonale présentée. Les précisions, rappels et F1-scores sont présenté en %.

Résultats et Discussion

Au travers du tableau 4.10, nous pouvons tout d'abord observer que les résultats obtenus avec la couche simple sont similaires à ceux obtenus avec la couche linéaire, bien que très légèrement inférieurs. Cette analyse de couche permet donc un gain de 20 000 paramètres dans le modèle pour des résultats similaires. Nous supposons qu'une analyse similaire de différentes couches dans d'autres modèles permettrait une meilleure compréhension du fonctionnement de ceux-ci et un allégement de certains des plus gros modèles. Ce cas est relativement simple, étant une interpolation linéaire, mais nous pouvons imaginer utiliser des formes de filtres provenant du traitement de signal pour effectuer des opérations plus compliquées, ou encore pour comprendre le fonctionnement de certaines couches plus complexes. Le seul élément nécessaire à cette analyse est la présence de plusieurs modèles entraînés sur des corpus différents avec la même architecture pour pouvoir en relever les motifs globaux.

Chapitre 5

Cas applicatif de la détection de parole superposée.

Une évaluation purement quantitative n'est pas suffisante pour juger si le modèle est utilisable hors de son contexte d'évaluation. Nous souhaitons donc dans ce chapitre faire un tour de différentes expériences menées à partir des systèmes présentés précédemment pour faire des ponts avec d'autres disciplines ou entre modèles. Ces expériences permettent alors de nouveaux angles d'approche pour étudier les conversations.

5.1 Analyse de la distribution temporelle de la parole superposée

Nous commençons cette exploration par une application qualitative d'un système de détection de parole superposée qui pourrait être utile aux études ayant des conversations pour objet. À partir d'une prédiction de parole superposée, nous pouvons également choisir de segmenter un extrait audio à un plus haut niveau pour obtenir une nouvelle information. Cette expérience propose d'étudier le niveau d'interactivité d'un extrait, défini par sa proportion de parole superposée.

5.1.1 Objectif

L'information de présence de parole superposée est intéressante en elle-même mais peut manquer de sens pour une étude qualitative. En effet, ce phénomène reste courant, sa présence n'apporte donc que peu de nouvelles informations. Nous souhaitons par conséquent apporter des informations plus haut niveau grâce à des visualisations. Notre intuition est que les zones de parole superposée interviennent dans des segments de parole fortement interactifs et spontanés. Une mesure du taux de parole superposée sur un temps donné pourrait donc servir d'intermédiaire à la mesure de l'interactivité d'une conversation.

5.1.2 Protocole

Pour calculer le taux de parole superposée, nous récupérons la prédiction de parole superposée et isolons les temps de départ de chaque segment ainsi que leurs durées. À partir de ces informations, nous calculons la durée cumulée des segments de parole superposée, notée d_{cumul} en fonction de leur temps de début, noté t_{deb} pour chaque émission. Nous obtenons N couples (d_{cumul}, t_{deb}) qui suivent l'ordre chronologique de l'émission. Nous calculons ensuite la dérivée Δ pour chaque segment n, grâce à l'équation 5.1 où N est le nombre total de segments de parole superposée.

$$\Delta[n] = \frac{d_{cumul}[n+h] - d_{cumul}[n-h]}{t_{deb}[n+h] - t_{deb}[n-h]} \quad \forall n \in [0; N] \tag{5.1}$$

La valeur de $h = 9$ a été manuellement sélectionnée de manière empirique afin de réaliser un lissage de Δ sur une large fenêtre temporelle. Une valeur plus petite augmenterait le nombre de variations et une valeur plus grande la diminuerait

Pour permettre une utilisation simple de ce système, nous avons développé une interface web qui réalise tous les traitements nécessaires à partir d'un fichier audio au format wav.

5.1.3 Exemples de résultats

Grâce à ce traitement, nous pouvons obtenir des courbes comme celle présentée en figure 5.1. Dans cette courbe, nous pouvons retrouver des pics correspondant à des moments de forte intensité de conversation.

Pour vérifier que notre système automatique prédit des résultats exploitables pour la prédiction de telles courbes, nous avons utilisé la segmentation de référence afin d'obtenir une courbe similaire qui pourra être comparée avec le contenu de l'émission pour en valider la pertinence. Cette courbe, présentée en figure 5.2 montre des similitudes claires dans la position des pics entre la courbe prédite (en figure 5.1) et la courbe de référence. La possibilité de générer la courbe automatiquement est donc validée. Il faut toutefois signaler que la validité de la courbe a été évaluée de manière qualitative et non quantitative. Une métrique de similitude de courbe telle qu'un RMSE, ou un coefficient de corrélation serait probablement basse en raison des différences de variations locales, ou d'amplitudes, mais assez peu pertinente au vu de l'usage envisagé de ces courbes. En effet, l'aspect intéressant de celle-ci se situe dans sa capacité au premier coup d'œil à repérer des zones fortement

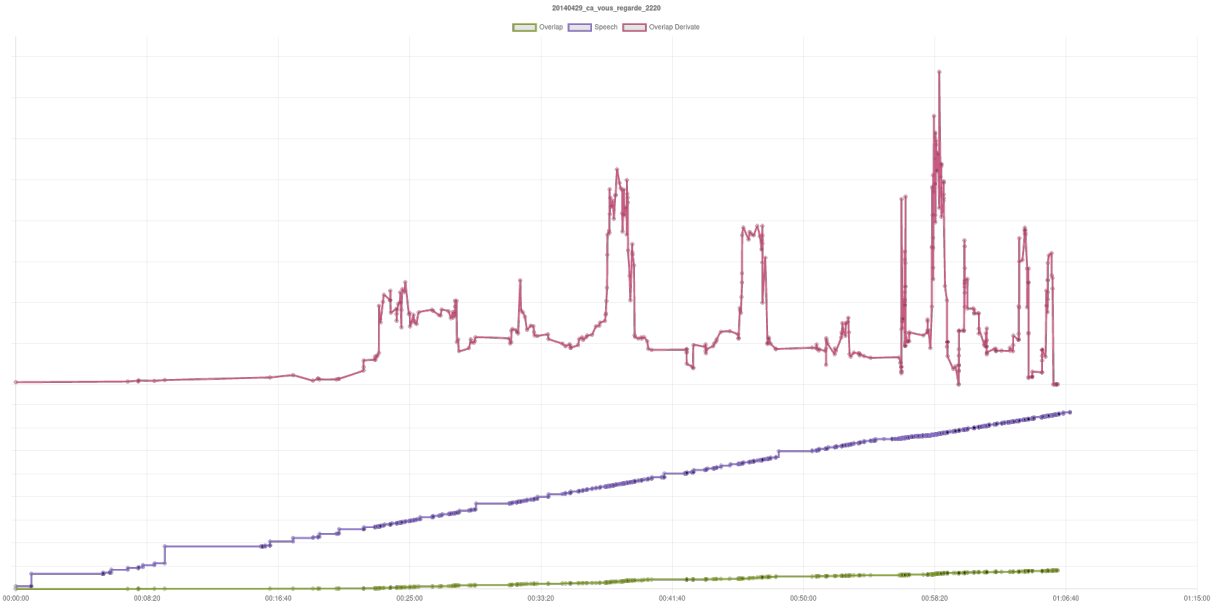


FIGURE 5.1 – Courbe d'interactivité pour l'émission *Ça vous regarde* du 29/04/2014, réalisée à partir d'une prédiction de parole superposée

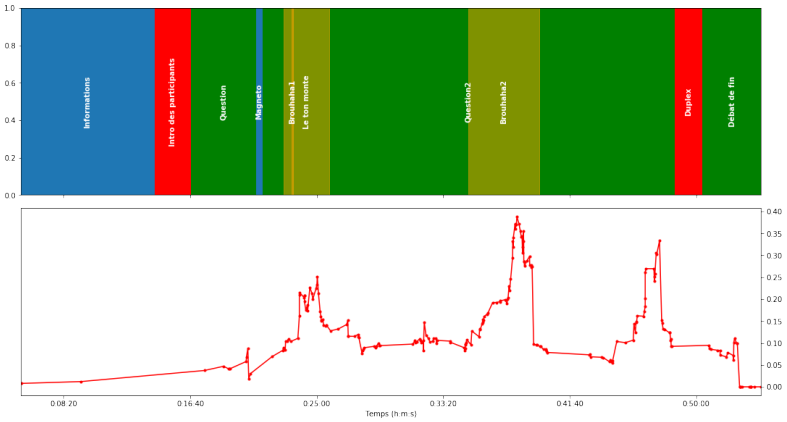


FIGURE 5.2 – Courbe d'interactivité pour l'émission *Ça vous regarde* du 29/04/2014 (identique à la figure 5.1), réalisée à partir d'une segmentation de référence avec une annotation manuelle de l'émission en parallèle.

interactives. Un décalage de quelques secondes, ou un pic moins élevé ne dégradent que très peu l'interprétation à faire de cette courbe, mais peuvent faire tomber des métriques quantitatives.

Par exemple, à partir de la figure 5.2, on peut voir grâce à la frise chronologique associée à la référence que les pics sont en principe associés à des zones de brouhaha, à l'exception du dernier pic qui présente une zone avec beaucoup d'interactivité, mais sans gêner la compréhension du sujet par les auditeurs.

En conclusion, l'usage d'une prédiction de parole superposée peut être pertinent pour accélérer l'étude d'émissions télévisuelles ou radiophoniques.

5.2 Analyse des durées des zones de parole superposée

Pour terminer cette étude de la parole superposée, nous souhaitons observer la distribution des durées de segments de parole superposée sur nos prédictions réalisées, afin de pouvoir valider que nos prédictions reproduisent une distribution existante comme calculée dans la section 3.2.2 sans créer d'effet de bord trop important. Une seconde partie de cette expérience est consacrée à un effet observé lors d'une collaboration avec Laetitia Biscarrat, chercheuse sur les thématiques de média et de genre au laboratoire LIRCES à Nice, sur un corpus de téléréalité pour montrer une utilisation possible d'une analyse similaire pour mettre en évidence des phénomènes liés au montage télévisuel.

5.2.1 Protocole

Pour pouvoir comparer les distributions avec des corpus dont nous sommes certains des annotations, nous utiliserons les corpus DIHARD ainsi qu'AMI. Nous utilisons ensuite une segmentation automatique réalisée à l'aide du système présenté en section 4.1.3, puis appliquons la même séquence de traitement que pour une référence de parole superposée afin de pouvoir les comparer de manière équitable.

5.2.2 Résultats/Discussion

Les distributions ainsi trouvées sont tracées dans la figure 5.3. On peut y voir que la distribution globale est respectée avec néanmoins des disparités dans les superpositions de

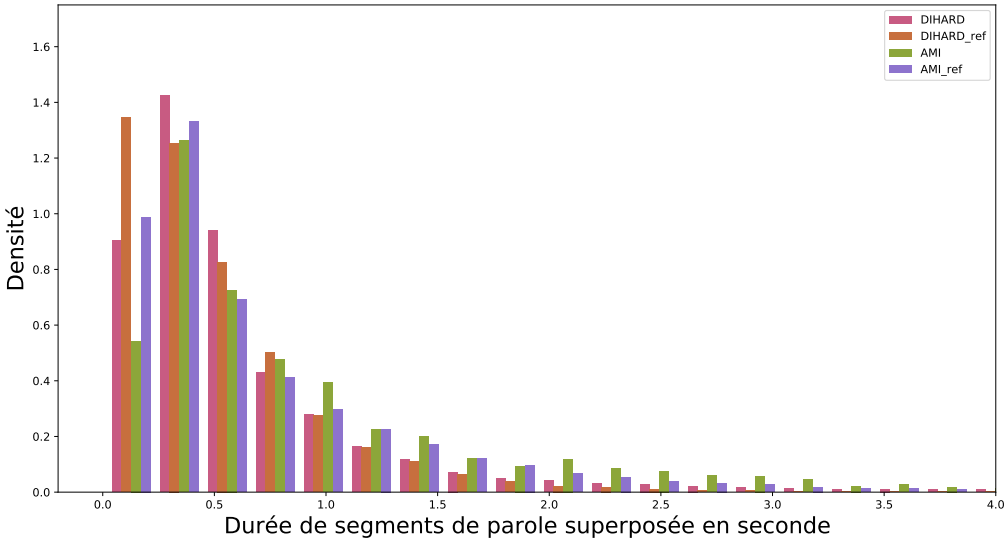


FIGURE 5.3 – Distribution des durées de parole superposée pour les corpus DIHARD et AMI avec la prédiction et la référence (notée ref).

très courte durée (inférieures à 0.3 s). Nous pouvons en conclure que notre système n'est pas adapté à la récupération de segments de cette durée. La présence de tels segments est cependant surprenante car très difficile à annoter pour des annotateurs humains.

Extension de l'expérience à des données sans référence

La seconde partie de cette expérience est une extension du même protocole sur un corpus sans références afin d'en étudier les caractéristiques. Cette étude n'est possible que grâce à la validation faite précédemment de la reproduction correcte de la distribution des durées de parole superposée par nos systèmes de prédiction. Pour cela, nous utilisons le corpus téléréalité décrit dans la section 3.1.2 dans l'objectif d'obtenir de premières analyses de données.

Comme présenté dans la figure 5.4, la distribution de parole superposée présente une grosse différence par rapport à une distribution classique. Les pics saillants observés sur la figure sont situés toutes les 0.5 s exactement. Ces nombres ont du sens pour un humain mais ne devraient pas statistiquement être privilégiés. C'est ici qu'intervient le travail d'interprétation des résultats obtenus. Ces pics observés sont très probablement des effets de bords de consignes du montage réalisé par les maisons d'éditions. Les segments sont

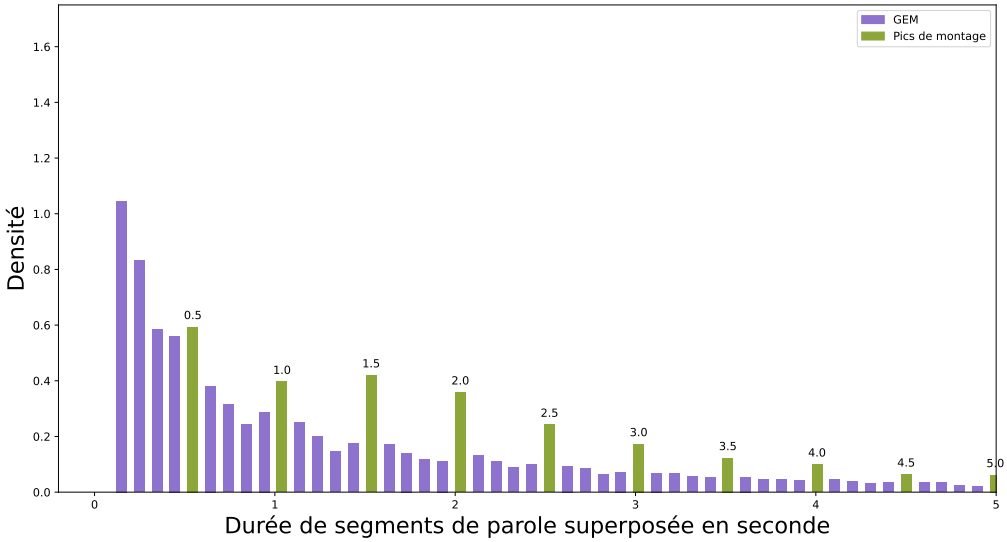


FIGURE 5.4 – Distribution des longueurs de parole superposée pour un corpus de télé-réalité

découpés pour obtenir un montage cohérent, le monteur choisissant alors une durée "ronde" de 0.5 s en 0.5 s. L'observation est triviale, mais il s'agit d'un autre exemple de comment l'observation des résultats obtenus au-delà d'un simple score de performance peut mener à des informations utilisables dans d'autres domaines.

5.3 Analyse des mots présents dans la parole superposée

5.3.1 Objectif

La parole superposée est un évènement qui se produit lorsque plusieurs personnes parlent simultanément. Nous souhaitons étudier le contenu linguistique de ces zones. Pour cela, notre intuition de départ est que la présence de certains mots montrant un désaccord fort ou un acquiescement devrait être plus importante dans ces zones de parole que dans le reste du corpus. Les mots que nous nous attendons donc à trouver le plus souvent sont les marques d'accord ou de et désaccord telles que oui/ouais/non qui indiqueraient la présence de backchannel. Nous allons donc créer un lien dans cette expérience avec le

domaine de la reconnaissance de parole. Nous aurions pu également traiter les disfluences qui sont présentes dans ces zones, mais la méthode employée ne permet pas de les relever.

5.3.2 Protocole

Nous effectuerons cette analyse sur un corpus d'extraits de dialogue courts (<20 s) contenant de la parole superposée récolté dans le cadre de nos travaux sur les interruptions. Nous détaillerons ce corpus plus précisément dans le chapitre 6.1. Il s'agit cependant d'un ensemble de données provenant d'émission télévisuelles et radiophoniques entre plusieurs locuteurs contenant à priori au moins une superposition de parole. Nous disposons alors de 5072 extraits à analyser.

Pour cette expérience, nous avons besoin d'une annotation textuelle. Celle-ci se base sur une transcription automatique obtenue grâce à un système développé au LIUM [Her+22]. Cette segmentation est alors alignée au niveau du mot et stockée dans des fichiers CTM au format :

```
show canal début durée mot confiance
```

La confiance correspond à la certitude du modèle en ses résultats. Grâce à cette segmentation en mot ainsi qu'une segmentation en parole superposée obtenue grâce au système 3classes TCN avec WavLM présenté précédemment en section 4, nous pouvons chercher les liens entre ces deux segmentations pour observer les mots présents dans les zones de parole superposée.

Pour cela, nous encodons les mots présents sur un vecteur contenant une valeur représentant l'indice du mot toutes les 10 ms ainsi qu'un vecteur binaire encodant la présence de parole superposée. Cette représentation permet d'obtenir l'intersection entre cette dernière et les mots prononcés. Nous considérons qu'un mot partiellement dans une zone de parole superposée compte dans cette catégorie. L'étude alors menée se base sur les occurrences de ces mots. Il faut toutefois noter que la présence de parole superposée dans les segments analysés gène la transcription et par conséquent biaise notre distribution

5.3.3 Résultats et conclusion

Nous comparons le rang, par nombre d'occurrences, entre les mots présents dans les zones de parole superposée ainsi que ceux présents dans l'ensemble du corpus. La figure 5.5 représente cette fréquence d'apparition et permet de comparer visuellement les différences

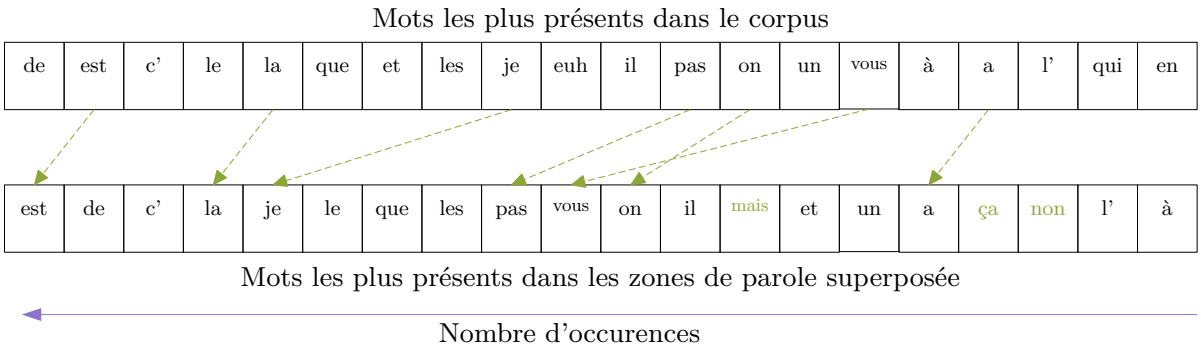


FIGURE 5.5 – Fréquence d'apparition des mots dans et hors des zones de parole superposée.

Mot	Position globale	Position overlap	Gain
est	2	1	+1
la	5	4	+1
je	9	5	+4
pas	12	9	+3
on	13	10	+3
vous	15	11	+4
a	17	16	+1
mais	22	13	+9
ça	23	17	+6
non	31	18	+13

TABLE 5.1 – Rang d'occurrence des mots dans le corpus global et dans les zones de parole superposées. Les mots présentant une opposition et les pronoms sont mis en valeurs

de rang en fonction de celle-ci. Nous faisons l'hypothèse que les différences rencontrées sont dues à la présence de zone de parole superposée. Plusieurs mots sont proportionnellement plus fréquents dans la parole superposée que dans le corpus global. Le tableau 5.1 représente les gains de rang dans le classement d'occurrence des mots dans les zones de paroles superposées par rapport à leur fréquence d'apparition globale. Parmi ces mots, nous pouvons extraire plusieurs phénomènes. Tout d'abords les pronoms, tels que *je* (+4 rangs) ou *vous* (+4 rangs) sont surreprésentés dans les zones de parole superposée. Cette fréquence est logique, une parole superposée étant un début de proposition superposée à une autre, nous pouvons nous attendre à voir des mots marquant le début de celle-ci présent dans les zones de parole superposée. La surreprésentation des pronoms *je* et *vous* peuvent également être marqueurs d'une interpellation (avec le *vous*) ou d'une mise en

avant, avec le *je*.

D'autres mots surreprésentés, sont les mots exprimant une opposition, par exemple *pas* (+3 rangs), *mais* (+9 rangs) ou encore *non* (+13 rangs). Cette différence entre les mots présents dans la parole superposée et ceux présents dans le corpus global peut s'expliquer par le fait que les interruptions sont par définition compétitives 1.3.2 et diffèrent par ce fait d'une conversation normale.

Cette expérience préliminaire apporte de nouvelles intuitions sur le fonctionnement de la parole superposée. Elle peut par conséquent être poursuivie dans plusieurs directions. Nous sommes convaincus cependant que d'autres tendances peuvent être observées avec des connaissances plus étendues de la linguistique. Cet aspect ouvre une possibilité très intéressante de collaboration avec cette dernière discipline. Enfin, nous n'avons pris en compte que les mots plus fréquents dans les zones de parole superposée que dans le reste du corpus. L'absence de certains mots de ces zones peut également être porteuse de sens et pourrait être étudiée.

Chapitre 6

Classification d'interruptions

Nous voici à la dernière étape de cette recherche. Cette section contiendra une description des travaux réalisés en détection automatique d'interruption. Dans un premier temps, nous verrons les problématiques de données rencontrées au travers de la collection d'un corpus puis de l'analyse de celui-ci. Puis dans un second temps, nous traiterons de la détection automatique de ces interruptions.

6.1 Annotation des données

La détection d'interruptions est un champ de recherche très peu exploité. Cela signifie que nous avons une grande liberté quant aux métriques et données utilisées, n'ayant pas de comparaison à l'existant à prendre en compte. Mais le revers de cette situation est qu'il n'existe a priori pas de corpus, ni de protocole expérimental sur lequel se baser pour cette tâche.

6.1.1 Description de la tâche

Comme présenté dans l'état de l'art, nous choisissons de restreindre cette tâche en ne considérant possibles que les interruptions présentes avec zone de parole superposée. Cette approximation retire environ 10 % des zones réelles d'interruption [Fer77].

Grâce à cette hypothèse, nous pouvons passer d'un problème de segmentation à un problème de classification de parole superposée. En effet, nous pouvons considérer que les zones d'interruptions sont un sous-ensemble des zones de parole superposée et donc supposés séparables d'autres classes de parole superposée par classification. Notre tâche sera donc d'apposer une étiquette à des segments de taille fixe contenant au moins une zone de parole superposée.

Classes de parole superposées choisies

Comme présenté en section 1.3.2, la parole superposée et son lien avec les interruptions ont été étudiés à plusieurs reprises et des ontologies ont été réalisées. Parmi celles-ci, nous avons choisi d'utiliser celle proposée par Adda-Decker et al. [Add+07]. En effet, nous utilisons des méthodes automatiques, et avons besoin d'une classification ayant déjà été testée sur un grand nombre de données avec des méthodes automatiques également. Cette ontologie a été définie sur le corpus Etape, elle semble donc utilisable pour notre cas d'utilisation. Cette classification comporte quatre classes. Nous considérons le premier locuteur qui parle comme le locuteur A et le second locuteur à parler le locuteur B

Départ anticipé Cette classe contient les segments de parole superposée crée par l'anticipation par B de la fin du tour du locuteur A. Elle ressemble à une interruption mais n'est pas perçue comme telle par les deux interlocuteurs car elle intervient alors que le sens de la proposition de A est déjà totalement connu, et l'intention de terminer sa phrase clairement marquée.

Backchannel Le backchannel n'est pas traduisible. Il contient les très courts segments d'activité vocale montrant un accord/désaccord, dans le but de donner un retour au locuteur principal. Cette classe est très présente dans les conversations téléphoniques pour compenser l'absence de communication non verbale.

Ajout d'informations complémentaires Ce type de parole superposée intervient souvent dans les interviews. Par exemple, un invité va designer une connaissance par un prénom, et le présentateur va compléter la proposition pour l'expliciter. Ce type de parole superposée peut également être présent dans des interactions quotidiennes pour montrer sa compréhension de manière plus complète qu'un simple backchannel.

Interruption Nous ne souhaitons pas borner la définition d'interruption, et préférons laisser aux annotateurs la liberté de considérer l'interruption selon leurs propres critères. La définition fournie aux annotateurs est "Un locuteur coupe la parole du précédent, n'importe quand dans son discours, avec une volonté claire d'interrompre. Il ne conserve pas forcément la parole ensuite.". Cette définition inclut la notion d'interruption définie plus tôt mais ne la borne pas. Ceci dans le but de conserver la dimension subjective de cette notion.

Justification des ajouts par rapport à la théorie

Cependant, en commençant l'annotation, certaines classes qui n'étaient pas présentes dans la théorie originale semblent pertinentes à ajouter.

Départ simultané Cette classe regroupe les instants où deux locuteurs commencent leur tour de parole simultanément. Cette situation n'est généralement pas présente dans une conversation à deux locuteurs, mais assez fréquente à partir de trois.

Brouhaha Cette classe décrit une zone de parole dans laquelle les propos sont trop confus pour être compris. Cette classe est présente pour pouvoir classer ces zones de paroles superposée particulières. Bien que potentiellement intéressante à traiter, la parole extrêmement bruitée de ces zones rend les traitements automatiques difficiles.

Pas de parole superposée Cette dernière classe contient les segments sans zone de superposition. Elle est purement pratique. Puisque les annotations en segmentation de locuteurs ne sont pas infaillibles, certaines zones notées comme telles pourraient ne pas contenir de parole superposée, et donc avec notre définition, pas d'interruption.

Choix des classes d'émotions

Ce corpus contient également un ensemble d'annotations en émotions, celles-ci n'ont également pas été clairement définies pour conserver la subjectivité des annotateurs. La consigne fournie aux annotateurs est la suivante.

```
Vous aurez à annoter l'état émotionnel des intervenants.
Pour cela vous disposez d'une liste de qualificatifs à attribuer
- À la partie de l'audio avant la superposition
- Au second locuteur intervenant dans la superposition
- À la partie de l'audio après la superposition
Les émotions sont assez proches les unes des autres,
sélectionnez seulement celle qui vous semble le plus proche de la réalité.
```

L'objectif de cette annotation est de voir le changement émotionnel engendré par la superposition. Pour cela une liste d'émotions a été fournie. Cette liste a été choisie à partir du cercle des émotions de Russel [Rus80] afin d'obtenir des émotions ayant plus de sens

dans un contexte de débat télévisuel que les 6 émotions habituelles [EF71] (Peur, colère, tristesse, joie, surprise, dégoût). Nous avons sélectionné 17 émotions sur ce cercle. A posteriori, celle-ci était trop longue pour obtenir un accord inter-annotateur correct.

— Frustré
— Outré
— Agacé
— Impatient
— Calme
— Attentif
— Neutre
— Confiant
— Amical
— Hésitant
— Sentiment de supériorité
— Enthousiaste
— Inquiet
— Sentiment de culpabilité
— Convaincu
— Décontenancé
— Arrogant

Ces émotions sont volontairement assez abstraites et ambiguës pour permettre de considérer les situations complexes rencontrées dans le corpus.

6.1.2 Corpus

Ces classes définies, nous savons quels types de segments rechercher, nous pouvons désormais choisir les données à faire annoter. Nous présentons ici le processus de sélection des données.

Choix des données

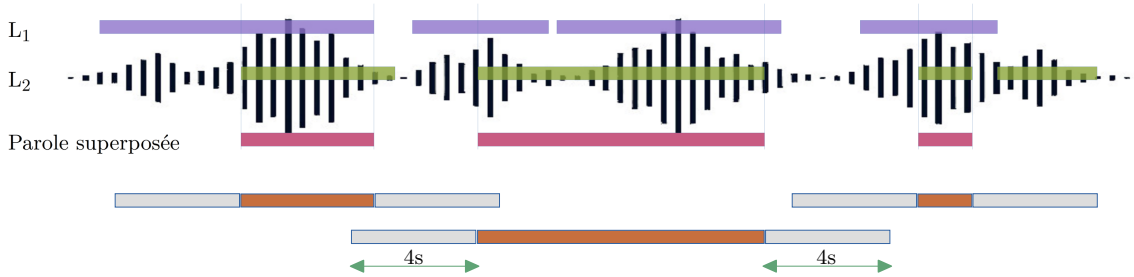
Les données à utiliser pour cette tâche doivent répondre à certains critères. Tout d'abord, le corpus doit disposer d'informations permettant une segmentation en parole superposée, c'est à dire être annoté avec des tours de parole précis et pouvant se superposer. En effet, la tâche réalisée est déjà trop complexe pour introduire des erreurs

de segmentations automatiques dès la création du corpus. Enfin, le projet GEM utilise les médias pour objet d'étude. Pour se rapprocher de cet objectif, nous souhaitons par conséquent travailler sur des données issues de médias audiovisuels.

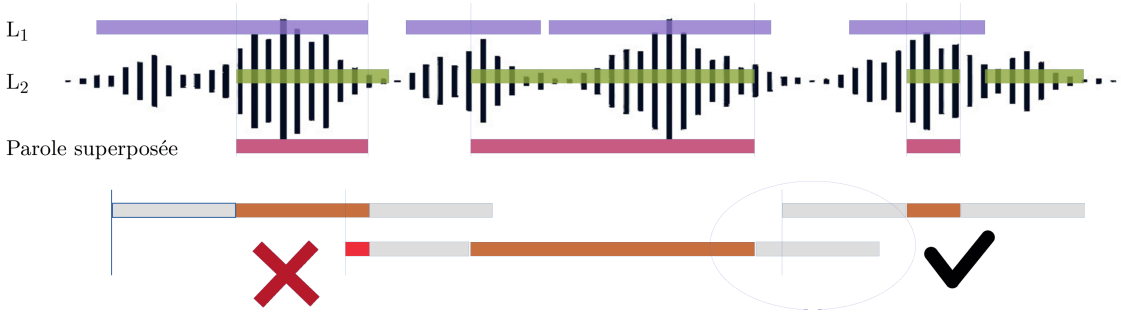
Parmi les corpus déjà utilisés, nous disposons notamment du corpus ALLIES qui répond à ces deux critères. Il est segmenté en tour de parole, ce qui permet d'obtenir une segmentation en parole superposée. Il contient des émissions issues de médias, le domaine est donc respecté. Enfin, le corpus ALLIES est actuellement en cours de correction et de vérification, ce qui assure une bonne qualité des annotations.

Pour réduire le nombre de fichiers à traiter et ainsi faire un premier tri, nous retirons tous les fichiers mono-locuteur.

Préparation des données



(a) Etape 1 : Récupération des contextes avant et arrières



(b) Etape 2 : Suppression des segments avec parole superposée dans le contexte avant

FIGURE 6.1 – Processus de collection des segments annotables. Le second segment ne peut pas être utilisé en raison de parole superposée dans son contexte avant.

Une fois les données sélectionnées, nous les traitons afin obtenir des extraits pertinents

pour l'annotation et par la suite classifiables. Nous émettons l'hypothèse que l'interruption étant complexe, elle dépend d'un contexte plus large qu'une zone de parole superposée. Nous faisons donc le choix de prendre une durée arbitraire de 4 s avant et après la zone de parole superposée en limitant cependant la taille maximum à 20 s dans le cas d'une zone de parole superposée très longue. Le protocole représenté en figure 6.1a consiste donc en la récupération de toutes les zones de paroles superposées des fichiers sélectionnés puis à un élargissement de la fenêtre pour contenir le contexte.

La zone considérée pour l'annotation étant la zone de parole superposée au centre de l'extrait, pour éviter les distractions pour les annotateurs, nous ne conservons pas les extraits avec une seconde zone de parole superposée située dans le contexte avant. Ce fonctionnement est également présenté dans la figure 6.1b. La parole superposée située dans le contexte après n'est supposée pas gênante pour l'annotation.

Les fichiers alors obtenus sont enregistrés avec l'audio, le nom du show initial, les temps de début et de fin de la superposition, ainsi que la présence de parole superposée dans le contexte arrière.

6.1.3 Plate-forme d'annotation

À partir de ces données, et des consignes présentées dans la section 6.1.1, nous pouvons élaborer un guide d'annotation. Pour pallier la subjectivité de nos données, nous recrutons plusieurs annotateurs. Ceux-ci sont donc quatre étudiants qui ont pour tâche d'annoter ces segments. Deux étudiants ont été recrutés en L2 informatique, et deux étudiantes recrutées en langue, la première en L3 et la seconde en M1.

Choix de l'outil

La tâche n'est pas une tâche connue et nécessite donc un certain degré de personnalisation de l'outil d'annotation. Il a donc fallu développer un nouvel outil pour répondre à tous nos besoins. En effet, nous souhaitons pouvoir écouter l'audio à annoter, mettre en valeur la partie du segment contenant la parole superposée sur laquelle l'annotateur doit se concentrer. Concernant les annotations à réaliser, elles doivent être réalisées au niveau du segment entier avec plusieurs catégories discrètes ainsi qu'une valeur continue. Enfin l'annotation doit pouvoir être réalisée en ligne, hors de l'université pour permettre aux annotateurs de travailler à leur rythme.

Pour disposer d'une base pour la collecte des annotations et pour l'interface, nous

avons choisi d'utiliser et de modifier l'outil FlexEval [Fay+20] développé à l'IRISA pour la réalisation de tests perceptifs. Cet outil nous fournit une base qui permet de soumettre des échantillons à des utilisateurs tout en sauvegardant les réponses dans une base de données ainsi de gérer la notion de progression et de compte, ce qui est une partie majeure du processus d'annotation.

Cet outil est développé en python pour le front et le back-end avec un système de template HTML, proposé par Django et Jinja. Nous avons donc développé une nouvelle interface d'annotation, en HTML/CSS/JS, décrite en section 6.1.3 pour permettre d'annoter toutes les informations que nous souhaitions avec les visualisations de l'audio et des consignes adaptées.

FlexEval est cependant dédiée aux tests perceptifs, certaines modifications de la logique de distributions des exemples ont par conséquent été nécessaires. Nous souhaitons tout d'abord que tous nos annotateurs voient les mêmes segments, et voient tous les segments. Nous devons donc contrôler les segments proposés pour privilégier les segments encore non annotés par l'annotateur actuel. Cette méthode conserve certains doublons qui seront par la suite traités comme indicateur de la qualité d'annotation.

Interface

Notre interface peut se découper en deux pages principales :
Une première page, présentée en figure 6.2 est un accueil pour les annotateurs. Sur cet accueil, ils ont accès à un récapitulatif des consignes ainsi qu'à des exemples pour les classes de parole superposée à utiliser. Cette page sert également de phase d'entraînement pour les annotateurs, en leurs présentant des exemples considérés comme sans ambiguïté. Les catégories d'émotions n'ont pas fait l'objet d'un entraînement.
La seconde page, accessible après connexion par les annotateurs est présentée figure 6.3. Cette page se compose de trois parties principales :

1. Le lecteur audio a été réalisé à l'aide de WavesurferJS, un plugin qui permet de générer des formes d'ondes, de jouer du son, de naviguer dans l'extrait en cliquant sur la forme d'onde et de générer des régions colorées comme affichée en rouge. Le chargement du fichier est particulier en raison de la présence des fichiers audio sur le serveur et non sur le client et de la conversion de langage du serveur en python au client WavesurferJS en javascript. La zone représentée en rouge est un indicateur visuel de la zone de parole superposée à annoter pour faciliter le travail des annotateurs.

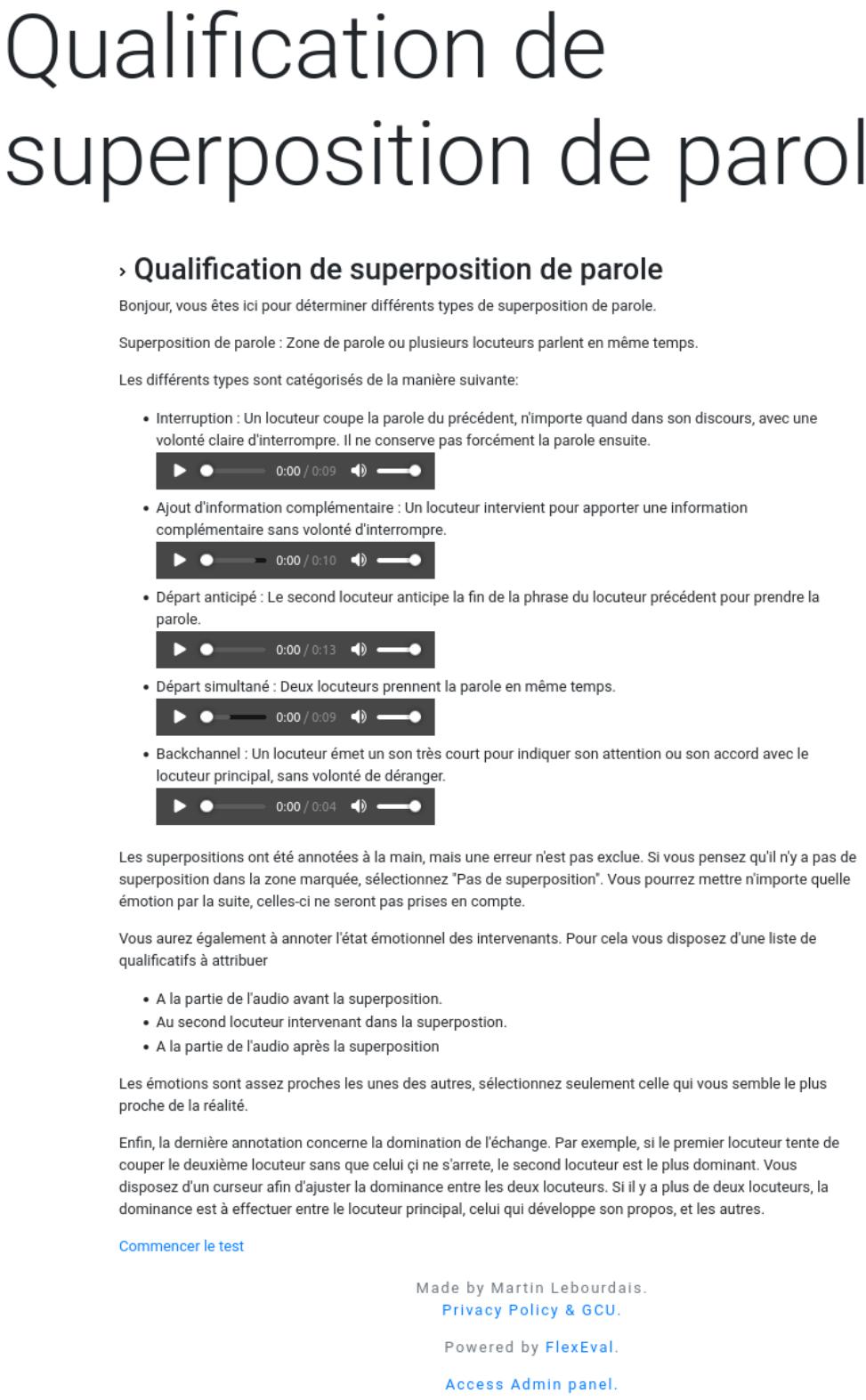


FIGURE 6.2 – Page d'accueil des annotateurs pour la plate-forme d'annotation. Cette page contient des consignes d'annotation ainsi que des extraits audio comme exemples non ambigus.

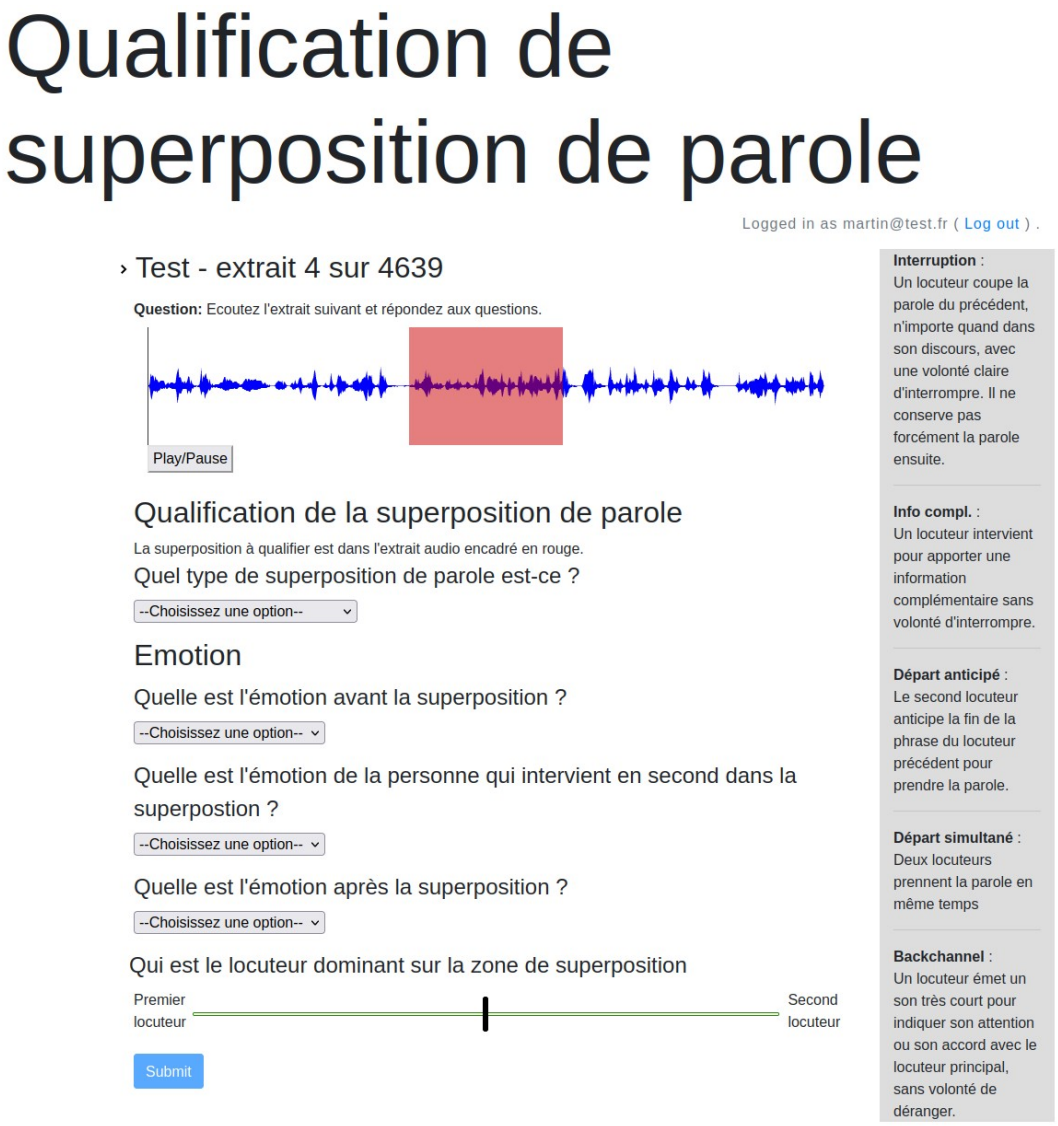


FIGURE 6.3 – Page d'annotation de la plate-forme d'annotation, contenant un lecteur audio de l'extrait à annoter, des menus déroulant pour choisir les classes et un rappel des consignes

2. La seconde partie de la page, située à droite, est un récapitulatif de certaines des catégories de parole superposée pour servir de rappel. Les catégories "Pas de superposition" et "Brouhaha" n'ont pas été incluses car leur sens est assez évident.

3. Enfin, la troisième partie de la page contient les champs à remplir. Les menus déroulants contiennent les différentes possibilités de classes de parole superposée et d'émotions. Une première version changeait d'ordre à chaque extrait pour diminuer le biais envers les premiers éléments de la liste, mais a été modifiée pour améliorer l'ergonomie. Le dernier curseur permet d'annoter de manière continue la dominance. La partie des annotations concernant les émotions n'ont cependant pas été traitées. Les classes d'émotions sont cependant conservées dans le corpus et pourront faire l'objet de traitements dans des études ultérieures.

6.2 Analyse des annotations

6.2.1 Contrôle des annotations

Les annotations faites, nous devons contrôler la qualité de ces annotations.

Cette section contient des définitions et des formalisations de métriques. Pour faciliter la lecture, nous utiliserons les notations suivantes :

Soit $N = 2393$ Nombre d'extraits annotés par tous les annotateurs, indexé par i

$M = 4$ Le nombre d'annotateurs, indexé par j

$K = 7$ Le nombre de catégories, indexé par k

Soit A Ensemble des annotations effectués

A_{ik} Annotations réalisées pour l'extrait i avec la catégorie k

A_{ijk} Annotations d'un annotateur j réalisées pour l'extrait i avec la catégorie k

B Ensemble des annotations du point de vue d'un annotateur

$B_{ij} \in \omega = \{1, ..., K\}$ Annotation d'un annotateur j pour l'extrait i

Annotateur	**Nombre de segments**	**Nombre de doublons**	**Cohérence**
Annotateur 1	4331	308	0.594
Annotateur 2	4346	293	0.608
Annotateur 3	2444	171	0.860
Annotateur 4	4334	305	0.613

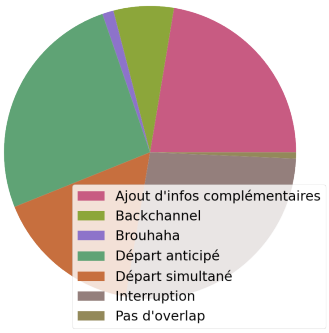
TABLE 6.1 – Cohérence des annotateurs pour la tâche d'annotation d'interruption.

Analyse intra-annotateur

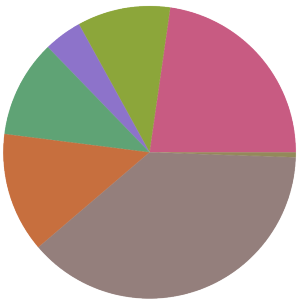
Les quatre annotateurs devaient traiter l'ensemble des segments, soit 4639 segments afin d'avoir la possibilité de gérer le côté subjectif de la tâche. Ils n'ont cependant pas réalisés le même nombre de segments. En effet, ils ont annoté 4331 segments pour l'annotateur 1, 4346 pour l'annotateur 2, 2444 pour l'annotateur 3, et 4334 pour l'annotateur 4 comme indiqué dans le tableau 6.1. Les annotateurs 1,2 et 4 ont pourtant eu l'impression de faire l'intégralité des segments. Une erreur dans l'implémentation de notre plate-forme a eu pour effet de parfois re-proposer des segments déjà annotés. Cette erreur est cependant à notre avantage car elle permet de vérifier la cohérence de ces annotations, c'est à dire la capacité d'un annotateur à reproduire une même annotation sur un même échantillon.
Ce que nous définissons comme cohérence d'un annotateur j est donc le nombre d'annotations identiques par extrait i rapporté au nombre total d'annotations de cet annotateur pour chaque extrait i.
La valeur obtenue correspond à la probabilité qu'un annotateur donne la même réponse pour un même échantillon présenté deux fois. Elle doit donc être proche de 1. Comme présenté dans le tableau 6.1, la cohérence pour les annotateurs 1,2,et 4 est proche de 0.6, ils ont donc pratiquement une chance sur deux de changer d'avis sur un même segment. L'annotateur 3 n'a pas terminé l'annotation mais a apparemment fait attention à la qualité de celles-ci. Il faut cependant prendre en compte qu'un annotateur qui répond systématiquement la même classe aura une cohérence de 1 mais des annotations inutiles. Il faut par conséquent également évaluer la répartition des classes annotées.

Répartition des annotations

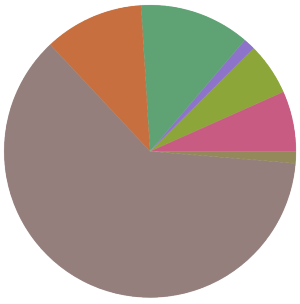
La figure 6.4 récapitule la répartition des classes de parole superposée annotées par les quatre annotateurs. À première vue, les distributions peuvent sembler similaires, mais



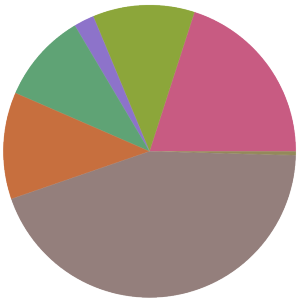
(a) Annotateur 1



(b) Annotateur 2



(c) Annotateur 3



(d) Annotateur 4

FIGURE 6.4 – Répartition des classes d'annotation en fonction des annotateurs

plusieurs phénomènes sont à noter. Premièrement, la classe "Interruption" en marron, semble occuper une part importante des distributions des quatre annotateurs, l'annotateur 1 utilise moins cette classe que les autres, en privilégiant les départs anticipés qui occupent une grande place de sa distribution. Cette situation n'est pas étonnante, en effet ces deux classes ont également montrées une grande confusion dans les travaux de Adda-Decker et al. [Add+08]. Nous ne souhaitons cependant pas fusionner ces classes. Nos travaux portent sur les interruptions qui ont une grande différence de sens avec les départs anticipés. L'annotateur 3 a une distribution d'annotations très différentes des autres, cependant il a également effectué moins d'annotations ce qui biaise la répartition selon l'ordre de présentation des segments. Cette étude nous permet de confirmer l'absence d'une classe mal définie qui aurait entravé le bon déroulement de la suite des analyses et de la prédiction. Ayant analysé les annotations de chacun des annotateurs séparément, la suite logique est de regarder l'accord entre ces utilisateurs.

Accord inter-annotateur

L'accord entre quatre annotateurs peut être mesuré grâce au kappa de Fleiss, nous utilisons donc cette métrique pour étudier l'accord des annotateurs sur les différentes classes.
Cette métrique se base sur deux proportions, la proportion d'attribution des extraits une classe, notée p_k et la moyenne de l'accord entre les annotateurs, noté P_i. La proportion d'attribution p à une catégorie k est donc tel qu'exprimé dans l'équation 6.1.

$$p_k = \frac{1}{N * M} \sum_{i=1}^{N} \sum_{j=1}^{M} Card(A_{ijk}) \tag{6.1}$$

Il faut également connaître la moyenne de l'accord entre les annotateurs noté P pour un extrait i, défini par l'équation 6.2

$$P_i = \frac{1}{M * (M-1)} \sum_{k=1}^{K} \sum_{j=1}^{M} a_{ijk}(a_{ijk} - 1) \tag{6.2}$$

Nous avons alors l'accord moyen pour chaque catégorie ainsi que pour chaque extrait que nous pouvons ainsi combiner grâce au Kappa de Cohen, dans l'équation 6.3.

$$\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$$

Avec :

$$\bar{P} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{6.3}$$

$$\bar{P}e = \sum_{k=1}^{K} p_k^2$$

Nous obtenons donc pour les fichiers remplis par les 4 annotateurs sur 7 classes un kappa de $\kappa = 0.311$. Ce kappa est faible, cependant, le grand nombre de catégories et leur déséquilibre biaise cette métrique en notre défaveur.

La faible quantité de segments annotés par l'annotateur 3 pose également un problème pour la poursuite de l'analyse et de l'utilisation des annotations. Nous souhaitons par conséquent étudier l'option de le retirer de l'annotation. Nous pouvons donc calculer ce kappa sur les trois autres annotateurs pour avoir un plus grand nombre d'extraits ($N = 4277$). Nous obtenons donc un kappa de 0.353, ce qui est toujours très faible, mais meilleur que celui à 4 annotateurs et plus homogène, au regard des distributions d'annotations de la figure 6.4.

Il n'est cependant pas possible, de manière simple de connaître l'accord détaillé par catégorie. Pour connaître cette donnée, nous pouvons calculer un accord par paire d'annotateurs pour une classe précise.
Nous définissons l'accord entre deux annotateurs 1 et 2 sur une classe k comme étant tel que défini dans l'équation 6.4.

$$Accord_{1,2,k} = \frac{Card(\{A_{i1k}, A_{i1k} = A_{i2k}\})}{Card(\{A_{i1k}\} \cup \{A_{i2k}\})} \tag{6.4}$$

Pour plus de clarté, nous pouvons rassembler ces scores dans des matrices d'accord.

La figure 6.5 contient l'accord entre les trois annotateurs pour la classe interruption. L'accord de la diagonale est un accord des annotateurs avec eux-mêmes, cette valeur ne peut différer de 1 et est donc non informative. Cet accord, qui a une valeur qui augmente en fonction de l'accord entre les paires d'annotateurs varie entre 0.51 et 0.57. Les annotateurs sont dans l'ensemble d'accord plus d'une fois sur deux pour assigner un segment comme

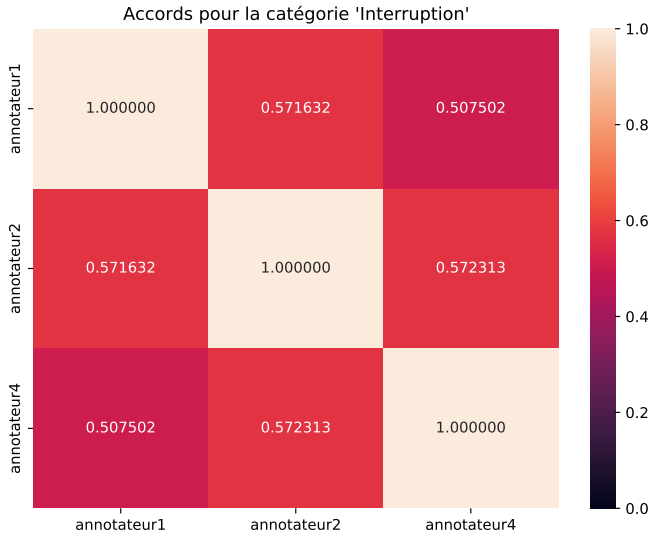


FIGURE 6.5 – Confusion inter-annotateur pour la classe Interruption avec trois annotateurs

interruption.

Voyons à présent la confusion pour quelques autres classes, mesurée avec la même méthode. La classe "Pas de superposition", présenté en figure 6.6 est beaucoup plus mitigée dans l'accord. Ce désaccord peut sembler étrange en raison de la simplicité de définition de cette classe, mais deux facteurs principaux rentrent en compte. Tout d'abord, définir la présence ou l'absence de superposition pour les cas limites, c'est à dire d'une durée très courte ou d'une intensité faible, est très difficile et relève du subjectif et de la sensibilité à la parole superposée. Le second facteur vient d'un autre type de parole superposée que nous n'avons pas anticipé. En effet, des journaux télévisés étant présents dans le corpus, il y a parfois de la traduction simultanée. Ce type de parole superposée peut être interprété de manière différente selon les annotateurs et classé dans plusieurs catégories. Les deux locuteurs ne sont pas sur le même plan, et donc l'un peut être considéré comme un bruit de fond, d'où l'annotation en "pas de superposition". On peut également considérer que la présence des deux discours simultanés trouble la compréhension et le classer en "Brouhaha" et enfin, si les deux locuteurs démarrent simultanément, il s'agit d'un départ simultané. Nous pouvons donc retrouver cette même incertitude dans les classes "Brouhaha" et "Départ simultanés" présentés respectivement en figure 6.7a et figure 6.7b. Le score plus élevé du départ simultané vient du fait qu'il existe de nombreux exemples de dé-

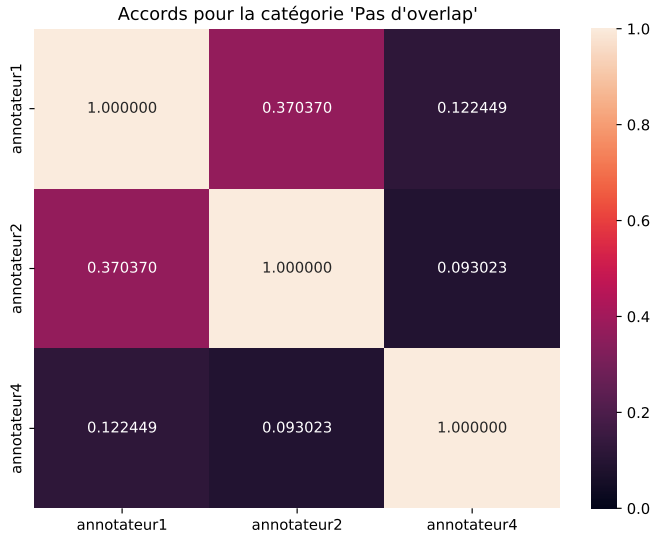
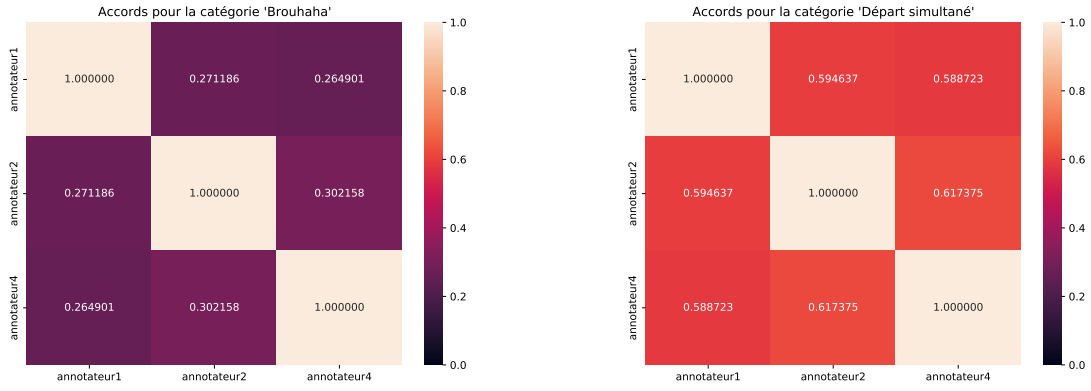


FIGURE 6.6 – Confusion inter-annotateur pour la classe 'Pas de superposition' avec trois annotateurs



(a) Confusion inter-annotateur pour la classe 'Brouhaha' avec trois annotateurs

(b) Confusion inter-annotateur pour la classe 'Départ simultané' avec trois annotateurs

parts simultanés clairs pour lesquels les annotateurs n'ont aucun problème pour trancher.

La tâche qui nous intéresse concerne la détection d'interruption. Il s'agit donc d'une classification binaire entre la classe 'Interruption' et 'Pas d'interruption' qui contiendrait toutes les classes à l'exception de la classe interruption. L'accord sur cette fusion est contenu dans la figure 6.8. Celui-ci est correct, variant de 0.7 à 0.8 en raison du grand

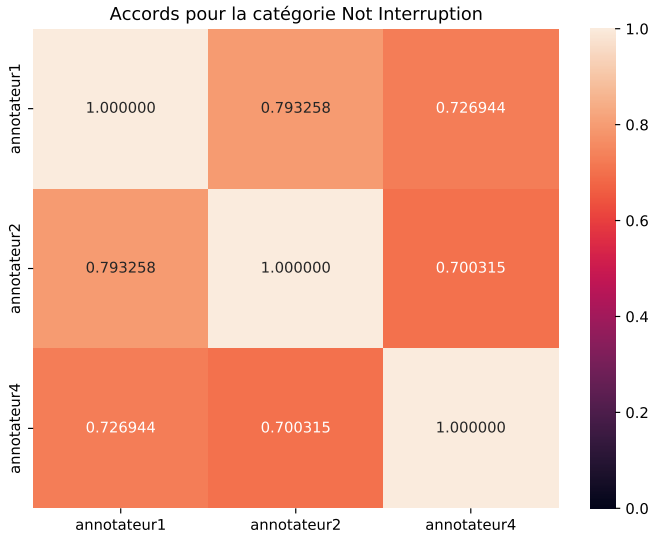


FIGURE 6.8 – Confusion inter-annotateur pour la classe 'Pas d'interruption' avec trois annotateurs

Fusion	# Segments totaux	Segments d'interruption	Part de test
Unanime	1448	571	449
Vote Majoritaire	4277	1499	1309
Pondération par l'accord	12831	4679	1309

TABLE 6.2 – Nombre de segments d'entraînement et de tests disponible pour chaque méthode de fusion des annotations en interruption.

nombre d'exemples en accord. La tâche de classification en interruption semble donc être possible à partir de ce corpus.

6.2.2 Fusion d'annotation

Nous souhaitons faire un système de classification de parole superposée en deux classes, interruption ou non interruption. Nous disposons de trois annotations réalisées pour un même segment. L'objectif est d'obtenir une unique référence par segment à partir de celles-ci. Pour cela, nous disposons de plusieurs stratégies de fusion.

Vote majoritaire

La première stratégie possible considère qu'une bonne réponse existe et est trouvable par la majorité des annotateurs. Un annotateur qui répond différemment des deux autres n'est donc par pris en compte. Pour cela, nous pouvons utiliser une stratégie de vote majoritaire. Nous considérons comme référence la classe la plus représentée pour l'échantillon. Si aucun des annotateurs n'est d'accord, le premier annotateur est désigné comme référence car le plus sérieux dans son travail.

Cependant, cette solution ne nous satisfait pas totalement car elle supprime le caractère subjectif de la perception d'une interruption. Une personne peut percevoir une interruption qu'une seconde personne ne percevra pas comme telle, sans pour autant de l'un des deux ait tort. C'est pourquoi nous privilégierons d'autres solutions de fusion d'annotation.

Sélection à l'unanimité

La seconde solution tente de résoudre le souci de la subjectivité en sélectionnant les exemples minimisant celle-ci. En sélectionnant les extraits pour lesquels les trois annotateurs sont d'accord, nous réduisons l'effet de la subjectivité en supprimant les cas limites. Cette solution réduit cependant le nombre d'annotations et nécessite des ajustements dans la suite du protocole. Comme présenté dans la figure 6.9 la fusion par unanimité réduit beaucoup le nombre de segments mais conserve une répartition des classes de parole superposée similaires aux répartitions des annotateurs séparés représentée en figure 6.4. Nous avons fait le choix de ne garder que les segments unanimes pour toutes les classes, bien que nous regroupons toutes les classes qui ne sont pas interruptions par la suite. En effet, si des annotateurs ne sont pas d'accord sur une classe précise, rien ne nous dit qu'un 5ème ou 6ème annotateur n'hésiterait pas à classer cet extrait en interruption.

Cependant, bien qu'efficace, le choix de cette méthode de fusion perd en nuance. En effet, n'ayant jamais vu de cas limites, un système entraîné sur cette fusion perdrait en efficacité dans un cadre réel et ne serait bon qu'à détecter des interruptions flagrantes.

Pondération par l'accord

Une dernière stratégie est donc proposée. Celle-ci se base sur les propriétés de l'apprentissage automatique pour apprendre à classer des éléments. L'idée est de présenter plusieurs fois au système le même extrait avec des labels différents dans le but de gêner la

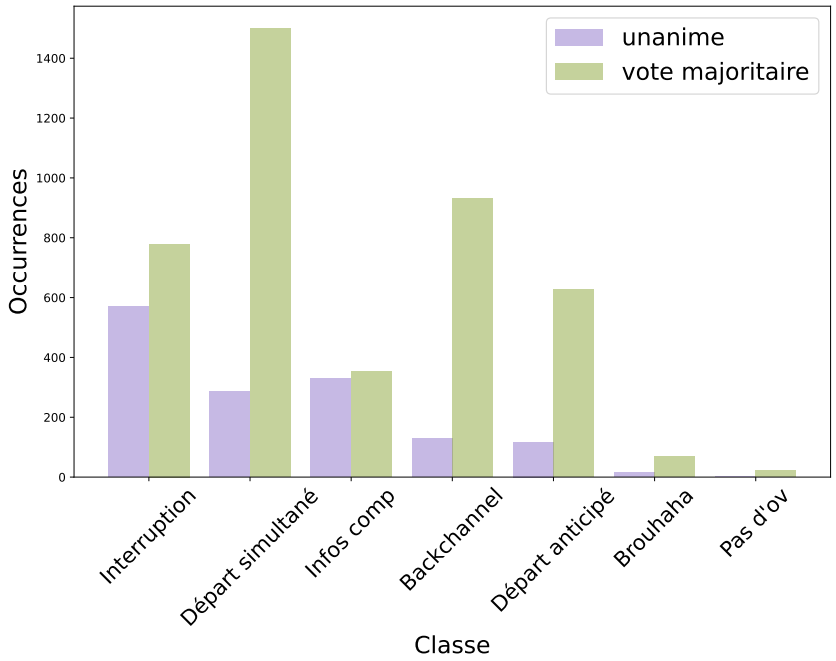


FIGURE 6.9 – Répartition des classes de segments de parole superposée pour un accord unanime et un vote majoritaire

convergence des exemples non unanimes vers une classe précise. Pour cela, nous considérons tous les labels des annotateurs comme référence pour l'entraînement et appliquons un vote majoritaire pour l'évaluation.

Cette méthode dispose d'une variante qui consiste à pondérer les extraits pour apporter plus de poids aux extraits unanimes. Pour cela, nous dupliquons le nombre d'occurrences d'un extrait suivant une puissance de 2. Par exemple, un extrait sans accord verra chacune de ses annotations placées 1^2 fois, un extrait avec une classe majoritaire aura sa classe majoritaire placée 2^2 fois dans le corpus et un extrait unanime aura sa classe placée 3^2 fois dans le corpus. Ce procédé a pour avantage d'augmenter artificiellement le nombre d'exemples du corpus mais nécessite plus de traitements et ajoute une incertitude quant à la réussite de l'apprentissage. En conclusion, nous allons dans un premier temps réaliser un système avec des annotations unanimes, pour servir de premier pas dans ce nouveau domaine, puis, tenter l'approche avec pondération par l'accord.

6.3 Système de classification

Une fois les données acquises, la dernière étape est la construction d'un système de prédiction automatique des zones d'interruption.

6.3.1 Système

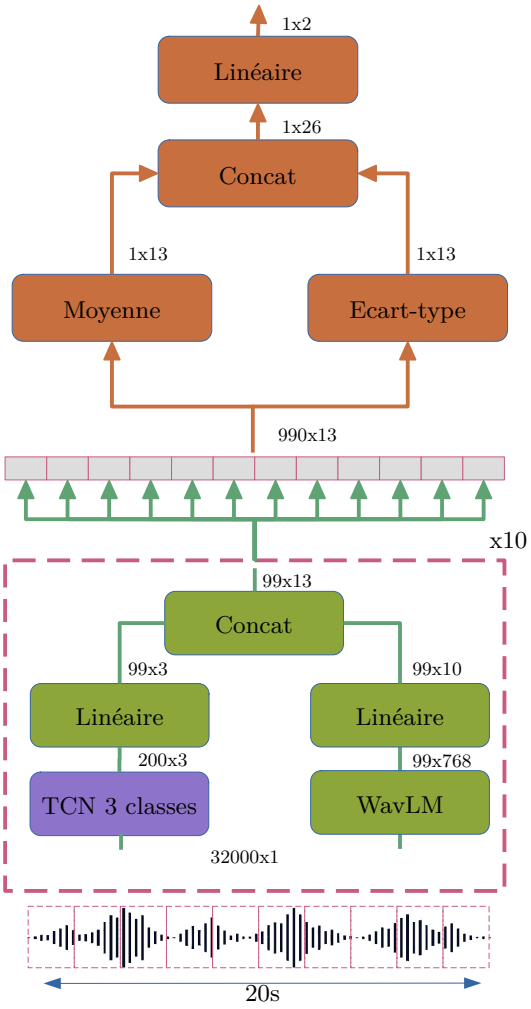


FIGURE 6.10 – Architecture du détecteur d'interruption

Les données étant peu nombreuses, le système utilisé ne peut pas avoir un grand nombre de paramètres pour éviter le sur-apprentissage, c'est à dire l'apprentissage de la

reconnaissance des données et non pas de la résolution de la tâche. Pour cela, nous proposons un modèle présenté figure 6.10. La première brique du modèle est un détecteur de parole superposée TCN à 3 classes entraîné sur le corpus ALLIES et figé. L'idée d'intégrer ce système est de rajouter la pseudo-probabilité de présence de parole superposée comme caractéristique d'entrée du modèle. Les segments d'entrée sont d'une longueur variable pouvant aller de 8 s à 20 s, ils sont complétés de 0 pour arriver aux 20 s (zero-padding). Pour pouvoir être traité par un système de parole superposé, ce segment de 20 s est découpé en 10 segments de 2 s sans recouvrement. La seconde branche d'extraction de caractéristiques consiste en une variante de WavLM proposée dans l'article original appelé *Base plus*. Cette variante est une version réduite du modèle utilisé précédemment avec le même corpus d'apprentissage et 768 dimensions de sortie. Nous utilisons ce système pour permettre de récupérer des informations différentes de la présence de parole superposée.

Afin d'éviter que les 3 caractéristiques de présence de parole superposées ne soient noyées dans les 768 caractéristiques de WavLM, une couche linéaire est donc entraînée à compresser l'information de WavLM en 10 dimensions. Pour finir sur cette extraction de caractéristiques, nous devons prendre en compte que le détecteur de parole n'étant pas échantillonné à la même fréquence que WavLM. Nous réalisons une interpolation pour sous-échantillonner la détection de parole superposée à la fréquence de WavLM. Les deux types de caractéristiques sont concaténés d'abord suivant l'axe des caractéristiques puis concaténés suivant la dimension temporelle afin de retrouver des segments de 20 s. On obtient alors un segment de 990 échantillons et 13 caractéristiques pour un extrait audio de 20 s. Par intuition, vérifiée par la suite en section 6.3.4, 13 caractéristiques semblent manquer de correlations pour discriminer un évènement aussi complexe qu'une interruption. Nous avons donc choisi de prendre la moyenne et l'écart type dans le temps pour ce vecteur, puis de concaténer les résultats pour obtenir 26 caractéristiques que nous pouvons classifier avec une couche linéaire à deux sorties. Cette technique est également souvent utilisée en reconnaissance des émotions [Mac+20].

Système entraîné sur des données unanimes

Un des principaux freins à l'utilisation de ce modèle est le manque de données (présenté dans le tableau 6.2). Pour pallier ce problème, nous utilisons une cross-validation à 5 plis sélectionnés aléatoirement sans recouvrement entre les plis. Un ensemble de test est conservé hors de cette cross-validation pour pouvoir la comparer aux autres méthodes de fusion.

Partition	F1-score	F1-score aléatoire	Gain
Unanime Fold 1	75.10	51.52	1.458
Unanime Fold 2	72.65	45.38	1.601
Unanime Fold 3	71.93	45.86	1.568
Unanime Fold 4	73.08	45.91	1.592
Unanime Fold 5	74.55	38.13	**1.955**

TABLE 6.3 – Résultats de la classification d'interruption sur 5 plis avec le F1-score en % et un F1-score calculé sur des résultats aléatoires pour comparaison. L'évaluation est réalisée sur la partition de développement du K-fold dans le but de sélectionner le meilleur modèle possible. L'efficacité est mesurée en gain relatif par rapport à l'aléatoire.

Ce modèle est entraîné avec l'optimiseur Adam et une fonction de coût cross-entropy. Nous utilisons comme métrique un F1-score calculé séparément sur chacun des plis. Enfin, ce modèle n'ayant pas de précédent état de l'art pour en juger les capacités, nous en comparons les résultats avec les résultats obtenus en replaçant la sortie du modèle par une génération de nombres aléatoires de même dimension avant application de la softmax. Pour sélectionner le meilleur modèle, nous comparons le gain relatif par rapport à l'aléatoire et prenons le plus haut. Chaque pli est entraîné sur 10 époques et la meilleure validation est conservée comme modèle pour ce pli.

Nous regroupons les résultats obtenus par le système appris sur des données unanimes pour chaque pli dans le tableau 6.3. Les partitions n'étant pas toutes aussi compliquées à prédire, nous considérons la performance du pli comme étant le gain relatif par rapport à l'aléatoire, notre meilleur pli est donc le pli 5 (gain de 1.955), que nous désignerons donc comme le modèle *unanime*.

Système entraîné sur une pondération par l'accord

Ce système ne manque pas de données en raison de l'augmentation de données réalisée pour la fusion des annotations. Nous pouvons par conséquent utiliser une méthode d'apprentissage classique avec un corpus de développement fixe. Les paramètres d'entraînement de ce modèle sont identiques au modèle précédent, utilisant également Adam avec un momentum de 0.9, une cross-entropy et une évaluation réalisée avec un F1-score. Nous sélectionnons l'époque de la même manière que le système précédent, en prenant la meilleure validation sur 20 époques.

6.3.2 Résultats pour des données unanimes

Système	F1-score	Précision	Rappel
Unanime	76.57	79.76	73.63
Pondération par l'accord	77.04	74.11	80.22
Aléatoire	44.53±0.62	40.88±0.73	48.90±0.53

TABLE 6.4 – Résultats de la classification d'interruption sur les deux modèles avec le F1-score, le rappel et la précision exprimé en % et un système donnant des résultats aléatoires pour comparaison. Le corpus d'évaluation ne contient que des segments unanimes.

Pour faire un premier état des lieux des performances de nos systèmes, nous effectuons une première classification sur des données annotées unanimement par nos annotateurs. Cette fusion permet alors d'obtenir 449 segments. Nous comparons dans le tableau 6.4 les résultats obtenus par le modèle *unanime* et le modèle *pondéré par l'accord* sur ce corpus de test.

Tous les résultats présentés sont supérieurs à l'aléatoire. Contrairement à ce que nous attendions, le modèle unanime est légèrement moins bon que le modèle pondéré par l'accord. Cependant, les résultats des deux modèles restent proches avec respectivement 76.57 % et 77.04 % de F1-score pour les modèles unanimes et pondérés par l'accord.

6.3.3 Résultats sur des données complètes

Système	F1-score	Précision	Rappel
Unanime	59.79	54.88	65.65
Pondération par l'accord	61.48	53.66	71.96
Aléatoire	39.60±0.42	32.76±0.32	50.05±0.63

TABLE 6.5 – Résultats de la classification d'interruption sur 5 plis avec le F1-score en % et un F1-score calculé sur des résultats aléatoires pour comparaison. Le corpus d'évaluation contient un vote majoritaire sur les annotations.

Nos résultats préliminaires sur les données unanimes ne reflètent pas la difficulté de la tâche de classification d'interruption. Nous allons donc dans un second temps évaluer les deux systèmes précédents sur des données fusionnées à l'aide du vote majoritaire.

Le tableau 6.5 contient les résultats d'une évaluation sur le corpus de test, traité avec cette technique. Cette méthode est supposée apporter des cas limites, par rapport à l'évaluation précédente. Les résultats obtenus sont moins bons que ceux précédemment obtenus mais la tâche réalisée comporte plus de subjectivité et plus de segments sont pris en compte (1309 contre 449). Comme prévu, le modèle entraîné sur les données pondérées par l'accord est plus performant que son équivalent entraîné sur les données unanimes avec un F1-score de 61.48 % contre 59.79 % respectivement.

6.3.4 Analyse du système

Comme précédemment pour la détection de parole superposée, les études purement quantitatives ne se suffisent pas. Maintenant que nous avons un système de classification d'interruption, deux routes principales sont possibles. Nous pouvons regarder pourquoi ce modèle marche, et que faire avec ces résultats. Voici donc trois études sur ces sujets pour compléter nos travaux sur la détection automatique d'interruption.

Vérification du besoin de 26 dimensions

Système	F1-score	Précision	Rappel
Pondération par l'accord 26 dimensions	77.04	74.11	80.22
Pondération par l'accord 13 dimensions	54.94	43.92	73.36

TABLE 6.6 – Influence du nombre de dimensions avant la couche de classification sur les performances du système de classification d'interruption.

Une première vérification à effectuer est notre intuition de départ du manque d'information pour la classification dans 13 dimensions. Pour la vérifier, nous entraînons notre meilleur système actuellement, celui entraîné avec pondération par l'accord sur les 13 dimensions avant application de moyenne écart type sur des données unanimes. Le tableau 6.6 présente les résultats obtenus pour ces deux systèmes. La différence de F1-score est importante, avec un écart absolu de 22.6 points de F1-score. L'utilisation de la moyenne variance est donc nécessaire aux bons résultats de nos modèles.

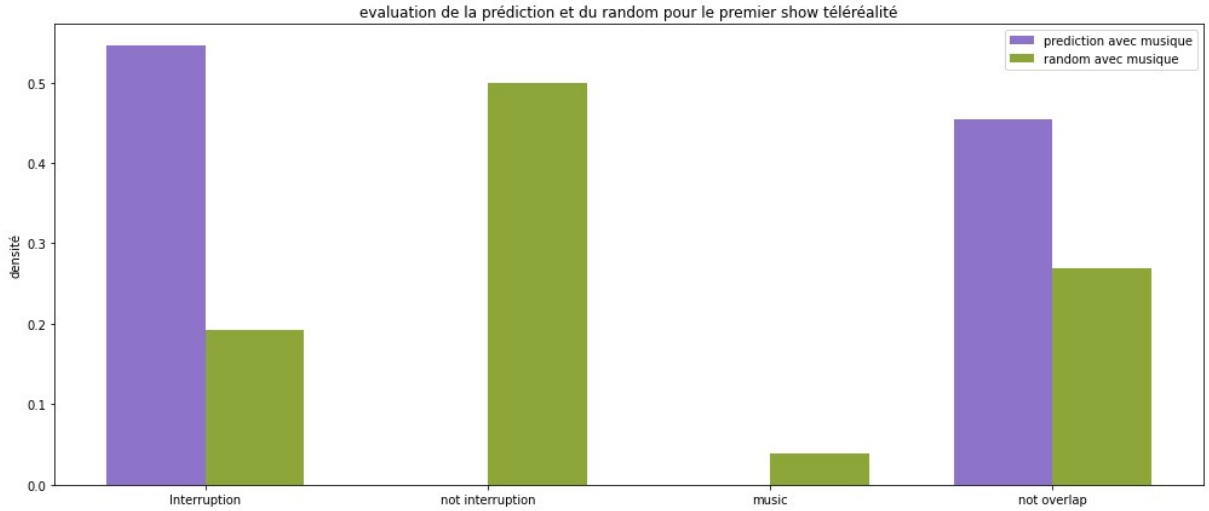


FIGURE 6.11 – Evaluation manuelle des segments prédits par le détecteur d'interruption et de segments aléatoires prédits comme ayant de la parole superposée sur des données sans prétraitement

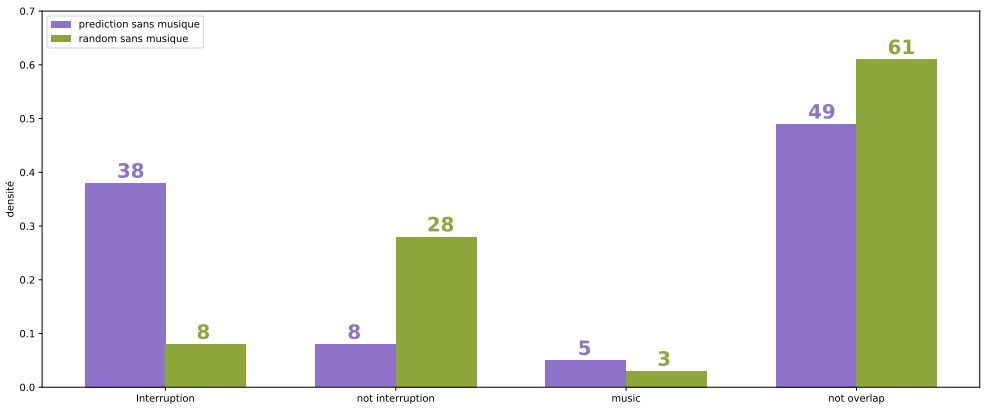


FIGURE 6.12 – Evaluation manuelle des segments prédits par le détecteur d'interruption et de segments aléatoires prédits comme ayant de la parole superposée sur des données prétraitées par Spleeter [Hen+20] provenant de la première émission du corpus téléréalité

Evaluation manuelle

Nos scores sont satisfaisants sur la partition de test du corpus créé. Il n'existe cependant pas d'autres corpus pour en évaluer la généralisation. Nous proposons donc une évaluation manuelle réalisée sur un fichier du corpus téléréalité pour lequel nous n'avons pas d'annotation. Ce test consiste en une évaluation de segments prédits comme interruption et de segment de parole superposée pris de manière aléatoire uniforme pour en comparer les distributions. Pour cela, nous avons prédit les interruptions de ce fichier et annoté les résultats obtenus en 4 classes : "Interruption", "Non-interruption", "Musique", "Pas d'overlap". Ces deux dernières classes ont été ajoutées pour séparer les erreurs dues au détecteur de parole superposée de celles du détecteur d'interruption. Ce traitement permet d'extraire 32 segments comme contenant une interruption. Nous ne disposons pas d'annotation de référence, nous choisissons donc d'évaluer les performances du système par rapport à l'aléatoire. Pour cela, nous annotons 100 extraits aléatoires parmi ceux avec de la parole superposée. Nous effectuons la même annotation sur ces segments que sur les segments réellement prédits comme interruption. La figure 6.11 contient les densités de chaque classe pour les segments prédits et aléatoires.

Grâce à ces deux distributions, nous pouvons voir que la majorité des erreurs viennent toujours du détecteur de parole superposée qui a des difficultés sur ce corpus, en particulier en raison de la présence constante de musique de fond.

Nous pouvons également calculer la significativité statistique de la distribution prédite, par rapport à la distribution aléatoire en utilisant un T-test. Nous souhaitons savoir s'il est possible que la distribution des données que nous avons annotées soit due à la répartition a priori des classes dans le corpus. Nous utilisons pour cela la p-value entre la distribution prédite et la distribution aléatoire telle que calculée par ttestind de scipy supposée mesurer la probabilité d'obtenir une distribution en prenant en compte une distribution aléatoire. Nous obtenons donc une p-value de (0.0378), qui est inférieur à 0.05 et montre donc que nos résultats sont significatifs. Il existe cependant certains biais quant à la qualité de l'annotation qui aurait dû être réalisée par plusieurs annotateurs.

Pour essayer d'améliorer ces résultats, nous utilisons Spleeter [Hen+20] pour retirer la musique. Nous supposons que la précision de la détection de parole superposée sera améliorée dans les zones contenant originellement de la parole superposée. Avec le même protocole que précédemment nous annotons 100 des segments prédits comme ayant des interruptions avec 100 segments aléatoires resélectionnés. Nous obtenons alors les résultats représentés Figure 6.12. Ces résultats nous donnent également une significativité

statistique suffisante (s=2.9158, p=0.0039).

Cette expérience permet de montrer la possibilité de généralisation du système de détection d'interruption. En effet, bien que toujours très sensible au détecteur de parole superposée, et par conséquent aux données, la détection d'interruption obtient des meilleurs résultats que l'aléatoire. Ces résultats pourront à l'avenir être améliorés.

6.4 Conclusion

Ce chapitre conclu ce projet de recherche et traite de la question du passage de la parole superposée à la détection d'interruption. Le domaine de la détection d'interruption ne dispose d'aucun cadre ou ressources, il a donc fallu le créer de zéro. Pour cela, nous avons créé un corpus avec des données de débats issues du corpus ALLIES.

Ce corpus dispose d'annotations utilisables pour la détection d'interruption, cependant plusieurs points seraient à améliorer. Le trop grand nombre de classes diminue sans doute le biais vers la classe interruptions, en répartissant les segments incertains vers les autres classes, mais il diminue aussi grandement l'accord inter-annotateur, rendant les annotations difficiles à fusionner. Concernant les classes d'émotions, nous n'avons pas pu exploiter les annotations obtenues. Cependant, au vu du grand nombre de classes, des fusions de classes seront nécessaires pour rendre les annotations exploitables. Enfin, d'un point de vue purement technique, l'utilisation de FlexEval a été très compliquée, et source d'erreur de programmations de ma part, utiles par la suite mais pas assez rigoureuses. Depuis l'expérience, nous avons expérimenté avec différents outils d'annotations et trouvé d'autres tel que LabelStudio qui sont plus simples d'utilisation et plus ergonomique pour les annotateurs.

Dans la deuxième partie de ce chapitre, nous avons présenté un classifieur d'interruptions présentant une capacité de classification supérieur à l'aléatoire. Ce résultat montre que la prédiction d'une dimension résolument subjective est possible avec des réseaux de neurones. Nous devons cependant nuancer les résultats, ceux-ci étant proches de l'aléatoire. Les principaux freins aux bons résultats de modèles de classification d'interruption se situent dans la quantité/qualité de données disponibles, ainsi que dans la définition d'interruption. Ce que nous avons défini comme interruption pour notre étude correspond à la notion d'interruption de nos annotateurs. Celle-ci n'est donc pas universelle. C'est pourquoi une étude sur une plus grande quantité de données, avec un plus grand nombre d'annotateurs devrait améliorer les résultats sur cette tâche.

Chapitre 7

Conclusion et perspectives

Au travers de cette thèse, nous avons mené un travail à la frontière entre des techniques et recherches ancrées dans le traitement automatique de la parole, et un objet d'étude qui est indéniablement social. La détection d'interruptions, objectif initial de cette thèse, s'est révélée au final un point annexe, mais apporte un point de vue important par rapport à une thèse purement technique.

Une des premières conclusions que nous souhaitons tirer de ces travaux, est l'importance capitale des définitions de tâche. Nous avons proposé une définition des interruptions adaptée au traitement automatique. Malgré nos efforts pour réduire les paramètres de cette définition, nous avons montré qu'elle reste très subjective. Nous ne considérons à aucun moment que notre définition d'interruption soit meilleure que les autres proposées. Elle est seulement plus adaptée à notre tâche. Nous pouvons citer comme exemple une deuxième thèse menée en parallèle par Rémi Uro, doctorant à l'INA, qui travaille également sur la détection d'interruption avec une méthodologie totalement différente en raison de différences de définition d'une interruption. Toutefois, si nous devions proposer la nôtre, nous définirions une interruption par l'exemple. Nous avons collecté un corpus annoté par quatre annotateurs en nous remettant à leur perception pour définir les interruptions. Notre définition est par conséquent la notion d'interruption de ces quatre annotateurs.

7.1 Contributions

7.1.1 Étude des corpus pour la segmentation de parole superposée.

Le chapitre 3 traite des différents corpus que nous avons pu utiliser pour entraîner et évaluer nos systèmes automatiques. Au vu des méthodes neuronales que nous avons utilisé, l'échelle de taille des corpus à notre disposition est suffisante. Il n'est sans doute pas

nécessaire de chercher à agrandir ces corpus tant que l'entraînement reste supervisé. Le problème majeur rencontré au cours de cette thèse est l'absence d'homogénéité des corpus du point de vue des annotations. Comme montré dans la section 3.2.1, la variabilité entre corpus est assez importante, alors que les corpus utilisés sont pour la plupart issus de la même communauté de recherche. Nous avons pu également constater le manque de précision de certaines annotations qui, pour des campagnes d'évaluation plus anciennes, ne pose pas de soucis, mais qui, avec des méthodes actuelles, limite les performances des modèles, ainsi que nos capacités à les évaluer.

7.1.2 Développement de modèles de détection d'activité vocale et de parole superposée

Le chapitre 4 présente les contributions apportées dans le domaine de la segmentation de parole, et plus précisément de la détection d'activité vocale et de détection de parole superposée. Tout d'abord, nous avons présenté plusieurs systèmes performants sur ces deux tâches. Ces systèmes atteignent des scores à l'état de l'art sur des corpus compétitifs tels que DIHARD ou AMI. Pour cela, nous avons comparé des systèmes à deux et trois classes utilisant des réseaux adaptés au traitement séquence vers séquence, un biLSTM et un TCN, avec différentes représentations d'entrée, MFCC et WavLM et différents optimiseurs. Aux travers de ces expériences, nous avons montré que les caractéristiques pré-entraînées, dans notre cas WavLM, sont extrêmement puissantes pour les tâches de segmentation audio. En effet, au cours de toutes nos expériences, le type de représentation utilisé a toujours été le premier facteur de performance. L'architecture choisie par la suite permet principalement de favoriser l'usage de ces caractéristiques en réduisant rapidement la dimension très importante de celles-ci.

Au travers de ces expériences, nous avons également montré l'importance du choix de l'optimiseur sur les performances d'un modèle, ADAM ayant donné un gain absolu de 3 points de f1-score par rapport à SGD. Enfin, l'expérience menée sur le classifieur à trois classes n'a pas apporté un grand gain en performance, mais permet de faciliter les opérations de prétraitement et d'accélérer les cas ou l'information de parole superposée et d'activité vocale sont nécessaires.

7.1.3 Étude de l'importance du domaine des données

Enfin, la dernière partie de ce chapitre présente une série d'expériences sur le système à trois classes montrant l'importance du domaine (téléphone, restaurant, etc.) dans ces prétraitements. L'adaptation d'un modèle à un autre domaine permet d'améliorer les performances dans ce domaine mais ne généralise pas aux domaines non rencontrés. Il s'agit donc d'une méthode efficace, mais nécessitant des données du domaine cible, ce qui a un intérêt certain pour une mise en application. L'expérience réalisée sur la comparaison des résultats en fonction du domaine de données des corpus DIHARD, AMI et ALLIES montre qu'un système performant est possible à partir du moment où des données d'entraînement adéquates sont fournies. En effet, nous obtenons des résultats similaires avec les modèles adaptés et des modèles entraînés depuis la base. Il faut cependant noter que l'adaptation au domaine a été réalisée à partir d'un échantillonnage manuel des données conçu pour représenter au mieux la variabilité du corpus. Les données d'adaptation pour cette expérience sont par conséquent beaucoup moins fournies que les corpus originaux utilisés pour un entraînement complet. Nous pouvons alors confirmer que la qualité et diversité des données est plus importante que la quantité. Une sélection précise de données provenant de différents corpus peut apporter une première réponse au problème de généralisation. Cependant, l'usage d'un grand nombre de corpus pose des soucis d'hétérogénéité des annotations. Le développement de techniques pour utiliser des données hétérogènes est par conséquent nécessaire.

7.1.4 Analyse des résultats de la détection de parole superposée

Le chapitre 5 propose différentes utilisations d'un système de détection de parole superposée. Dans ce chapitre, nous avons tout d'abord présenté une visualisation des zones contenant de la parole superposée pour assister le travail d'étude des conversations. Cette représentation montre une corrélation entre les pics présentés et le niveau d'intensité/interactivité d'une conversation.

La seconde expérience présente la distribution statistique des durées des segments de parole superposée détectés. En les comparant avec la distribution statistique de durée des segments de parole superposée de la référence, nous pouvons trouver d'éventuels biais dans les prédictions ou vérifier le bon fonctionnement du système. Nous avons également présenté un résultat obtenu dans une étude menée en parallèle sur un corpus contenant de la téléréalité. L'étude de ces durées met en évidence la présence de montage sur les

données grâce à une surreprésentation des segments d'une durée multiple de 0.5 s.

La troisième expérience présente une analyse des mots contenus dans les zones de parole superposée. Durant cette expérience préliminaire, nous avons pu relever certains phénomènes relatifs à la fréquence des mots dans, et hors de la parole superposée. Tout d'abord, nous notons une surreprésentation des pronoms personnels, servant à commencer un tour de parole ou à interpeller l'interlocuteur. Cette observation confirme l'hypothèse d'une variation des mots présent ou absent de ces zones. Ensuite, nous observons une surreprésentation des mots indiquant un désaccord avec l'interlocuteur. Nous faisons l'hypothèse que celle-ci montre la présence importante d'interruptions, et interactions conflictuelles dans les zones de parole superposée de notre corpus.

Les travaux présentés pour cette expérience ne sont que préliminaires et mériteraient un approfondissement au travers d'une collaboration avec les sciences du langage pour obtenir une analyse plus poussée des fréquences d'apparitions observées.

7.1.5 Création d'un corpus de détection d'interruptions

Pour finir, le chapitre 6 est consacré à l'étude menée sur la détection automatique d'interruptions. Cette tâche ne dispose que de peu de ressources disponibles. La première partie de ce chapitre est donc consacrée à la collection d'un corpus. Le corpus collecté contient des extraits audio de 10 à 20 secondes composés d'un segment de parole superposée, ainsi que quatre secondes de contexte avant et arrière. Les données utilisées pour ce corpus proviennent du corpus ALLIES et sont donc des extraits issus de médias audiovisuels en français. Les segments ont été annotés par quatre annotateurs grâce à une plate-forme d'annotation que nous avons développé à partir de l'outil FlexEval. Ces annotations seront mises à disposition en même temps que le corpus ALLIES.

Nous avons, dans une deuxième section de chapitre, analysé les annotations obtenues afin d'en extraire des intuitions pour la suite de l'expérience. Ces annotations n'ont tout d'abord qu'un faible accord inter-annotateur de 0.353 (mesuré avec le kappa de Fleiss). Cela montre encore une fois la subjectivité de la tâche ainsi que la difficulté de prédiction attendue. Nous présentons enfin dans une dernière partie les méthodes de fusion des annotations que nous avons utilisées et évaluées pour la détection automatique d'interruptions.

7.1.6 Développement d'un détecteur d'interruptions

Notre dernière contribution est le développement d'un système de détection d'interruptions à partir du corpus collecté. Pour cette tâche, nous avons dû composer avec la taille réduite du corpus d'apprentissage en proposant un modèle tirant parti des capacités des modèles pré-entrainés, dans notre cas WavLM, en les associant à la connaissance de présence de parole superposée fournie par notre modèle de détection jointe d'activité vocale et de parole superposée. Le modèle proposé ne dispose alors que de trois couches avec des paramètres pouvant être appris, deux pour permettre de fusionner les deux caractéristiques d'entrées, et une pour la classification finale.

L'évaluation des performances de notre système a été réalisée par comparaison avec l'aléatoire pour servir de référence naïve et montre une amélioration de 32 et 33 points de pourcentage pour une cible unanime et de 20 et 22 points de pourcentage pour une cible définie par vote majoritaire. Cette évaluation ne montre pas la possibilité d'utilisation de notre système de détection d'interruptions, mais démontre la possibilité de détecter automatiquement une interruption suivant notre définition.

Nous avons par la suite cherché à évaluer les performances du système de détection d'interruptions dans des conditions plus réalistes, en évaluant manuellement des prédictions sur un nouveau corpus et en cherchant à les comparer à une annotation de segments aléatoires, représentant la distribution a priori des interruptions dans le corpus. Cette étude présente des biais indéniables en raison de l'ajout d'une nouvelle annotation, mais permet de montrer la possible généralisation de la détection d'interruptions sur un nouveau corpus.

7.2 Perspectives

7.2.1 Limite des références de segmentation en locuteur

Tout apprentissage supervisé nécessite des annotations. Lors de nos travaux, nous avons été confrontés au manque de données ainsi qu'à l'imprécision des annotations et aux différences de consignes d'annotations. Plusieurs pistes sont à creuser. Tout d'abord, il faudrait corriger les corpus historiques majeurs pour en éliminer les erreurs qui ont très probablement déjà été observées par les différents acteurs utilisant ces corpus. Cela permettrait de continuer l'utilisation de méthodes supervisées tout en enrichissant les connaissances actuelles de ces corpus. Cependant, la correction des annotations ne retire

pas le principal problème de l'approche supervisée : pour entraîner un système sur une nouvelle tâche il faut récupérer de nouvelles données, ce qui ne favorise pas l'homogénéité des annotations. Une autre possibilité serait de changer de forme de données. Comme pour les derniers systèmes de diarization end-to-end nécessitant une très grande quantité de données, il est possible d'utiliser des conversations simulées et non réelles. Cette forme de données introduit de nouvelles problématiques à considérer mais pourrait permettre d'atténuer les problèmes actuels des corpus, sans pour autant retirer l'obligation d'avoir des données de test annotées de manière certaine. Enfin, une dernière solution possible serait le passage du paradigme supervisé à auto-supervisé. Ce paradigme utilise une référence générée à partir des données, sans besoin d'avoir des annotations.

7.2.2 Détection jointe multi-classe

Comme montré grâce au détecteur joint parole/parole superposée, la résolution de plusieurs tâches de segmentation simultanément n'offre pas nécessairement de gain de résultat/score. Cependant, cette détection offre deux avantages principaux. Premièrement, elle apporte un gain de temps et de simplicité, permettant de réaliser les prétraitements nécessaires à des tâches telles que la segmentation et regroupement en locuteurs en une passe de système. Le second avantage de prédire plusieurs classes simultanément est la possibilité d'utiliser ces informations dans un système tiers. De telles informations injectées dans un extracteur de plongements de locuteurs permettrait d'éliminer d'éventuelles trames bruitées. Le rajout de classes supplémentaires à un tel système est également possible et apporte de nouvelles informations pour cette dernière utilisation. Cette perspective a été étudiée lors du Workshop JSALT 2023 dans le projet de diarization interprétable. Un système multilabel prédisant une segmentation activité vocale / musique / bruit / parole superposée a été conçu, et les sorties de ce modèle sont utilisées pour améliorer l'extraction de représentation de locuteur.

7.2.3 Perspectives de la détection d'activité vocale et de la parole superposée

La détection d'activité vocale et de parole superposée sont deux tâches de prétraitement arrivant aux limites de leur définition. La détection d'activité vocale atteint actuellement des scores proches de 100 % de bonnes réponses et arrive au niveau des erreurs d'annotations. Des améliorations ne sont donc plus significatives sans redéfinition des ob-

jectifs pour mieux répondre à une tâche. La détection de parole superposée atteint également un mur de performance. Les gains de performances mesurés en f1-score ne sont alors plus nécessairement significatifs. La majorité des erreurs obtenues avec nos systèmes ont plusieurs sources clairement définies. Tout d'abord, la qualité de l'enregistrement (bruit de fond, musique) est toujours un élément limitant. Des progrès dans la représentation des données pourraient alors améliorer significativement les résultats. La seconde source d'erreur provient de la définition même de parole superposée et à sa différence avec le bruit. On peut considérer un bruit comme un signal non désirable. La question qui se pose est alors quand est-ce que de la parole superposée doit être considérée comme telle, et quand il ne s'agit que d'un bruit. Une meilleure classification de la parole superposée pourrait alors être une amélioration intéressante si l'on considère l'information de présence de parole superposée comme une tâche finale et non comme un traitement d'une chaîne. En effet, si l'on souhaite utiliser la détection de parole superposée pour la diarization, la formulation du problème par les systèmes end-to-end [Fuj+19] supprime le besoin de détecter ces zones en amont. Une connaissance de la tâche finale est donc primordiale pour pouvoir définir la tâche de détection de parole superposée et par conséquent les métriques à utiliser ou la définition de la classe positive.

7.2.4 Limite de la création du corpus de détection d'interruptions

De cette annotation et de son analyse ressortent plusieurs conclusions. Premièrement, nous avons un corpus utilisable pour la détection d'interruptions, ce qui a permis de commencer les expériences sur ce sujet. Cette expérience met en valeurs plusieurs écueils pour la création de corpus pour ce type de tâche. Pour la plateforme d'annotation, nous avions choisi de privilégier l'ergonomie pour les annotateurs à la robustesse et avons donc choisi une plateforme entièrement personnalisable. Cependant, l'outil utilisé était adapté à l'évaluation de système de synthèse de parole qui présente de grandes différences avec notre tâche, notamment au niveau de la présentation des segments aux annotateurs. Nous sommes toujours persuadés que l'ergonomie de la plateforme est primordiale, mais une expérience ultérieure d'annotation nous a permis de tester LabelStudio, également open-source, qui serait plus adapté pour la création de corpus de détection d'interruption. Un second écueil à éviter est la surcharge cognitive des annotateurs. Nos données sont annotées de manière très fine, avec plusieurs tâches à réaliser simultanément (interruption

et émotions). La finesse des annotations permet d'éviter une classe neutre non informative, mais rend la tâche plus compliquée aux annotateurs. Séparer les tâches d'annotation est par conséquent un élément important pour les futures annotations. Les annotations en émotions auraient quant à elles dû se trouver dans une seconde phase d'annotation pour permettre de les exploiter au maximum.

7.2.5 Perspectives pour la détection d'interruptions

Pour compléter notre expérience sur la détection d'interruptions, il faudrait réaliser un test perceptif à plus grande envergure avec des extraits prédits comme interruptions avec des degrés de confiance différents pour pouvoir déterminer un seuil à partir duquel les annotateurs présentent un accord sur la présence ou l'absence d'interruptions. Ces degrés de confiance ne sont possibles à définir qu'après application d'une calibration de nos modèles, c'est à dire une correspondance entre les probabilités a posteriori du modèle et les sorties de la couche de classification. L'amélioration du système de détection d'interruptions serait possible grâce à plus de données annotées dans de bonnes conditions, tout en faisant attention à conserver une tâche identique à celle-ci pour rester comparable. L'intérêt de la communauté pour cette tâche reste cependant l'élément décisif des progrès possibles sur cette dernière.

BIBLIOGRAPHIE

[21] *La représentation des femmes à la télévision et à la radio*, rapp. tech., Conseil supérieur de l'audiovisuel, 2021.

[98] *The 1998 Hub-4 Evaluation Plan for Recognition of Broadcast News, in Spanish and Mandarin*, 1998.

[ACB17] V. Andrei, H. Cucu et C. Burileanu, « Detecting Overlapped Speech on Short Timeframes Using Deep Learning », en, in : *Proc. ISCA Interspeech*, Stockholm, Sweden : ISCA, août 2017, p. 1198-1202, doi : 10.21437/Interspeech.2017-188.

[Add+07] G Adda et al., « Speech overlap and interplay with disfluencies in political interviews », in : *International Workshop on Paralinguistic Speech-between models and data, ParaLing* (2007), p. 41-46.

[Add+08] M. Adda-Decker et al., « Annotation and analysis of overlapping speech in political interviews », in : *Proc. Language Resources and Evaluation Conference (LREC)*, Marrakech, Morocco, 2008, p. 3105-3111.

[AL98] K. J. Anderson et C. Leaper, « Meta-Analyses of Gender Effects on Conversational Interruption : Who, What, When, Where, and How », en, in : *Sex Roles* 39.*3* (août 1998), p. 225-252, issn : 1573-2762, doi : 10.1023/A:1018802521676.

[AW03] J. Ajmera et C. Wooters, « A robust speaker clustering algorithm », in : *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, nov. 2003, p. 411-416, doi : 10.1109/ASRU.2003.1318476.

[Bae+20] A. Baevski et al., « wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations », en, in : *arXiv* (oct. 2020).

[Bak+18] N. Baker et al., « Deep convolutional networks do not classify based on global object shape », en, in : *PLOS Computational Biology* 14.*12* (déc. 2018), e1006613, issn : 1553-7358, doi : 10.1371/journal.pcbi.1006613.

[Bál+51] R. F. Bales et al., « Channels of Communication in Small Groups », in : *American Sociological Review* 16.*4* (1951), p. 461-468, ISSN : 0003-1224, DOI : 10.2307/2088276.

[Bar+98] C. Barras et al., « Transcriber : a Free Tool for Segmenting, Labeling and Transcribing Speech », in : *Proc. Language Resources and Evaluation Conference (LREC)*, Granada, Spain, 1998, p. 1373-1376.

[BB79] G. W. Beattie et P. J. Barnard, « The temporal structure of natural telephone conversations (directory enquiry calls) », in : *Linguistics* 17.*3-4* (1979), p. 213-230, ISSN : 0024-3949, DOI : 10.1515/ling.1979.17.3-4.213.

[BBG20] L. Bullock, H. Bredin et L. P. Garcia-Perera, « Overlap-Aware Diarization : Resegmentation Using Neural End-to-End Overlapped Speech Detection », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, p. 7114-7118, DOI : 10.1109/ICASSP40776.2020.9053096.

[Bea82] G. Beattie, « Turn-taking and interruption in political interviews : Margaret Thatcher and Jim Callaghan compared and contrasted », in : *Semiotica* 39 (1982), p. 93-114, DOI : 10.1515/semi.1982.39.1-2.93.

[BKK18] S. Bai, J. Z. Kolter et V. Koltun, « An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling », en, in : *arXiv :1803.01271 [cs]* (2018).

[BL21] H. Bredin et A. Laurent, « End-to-end speaker segmentation for overlap-aware resegmentation », in : *Proc. ISCA Interspeech*, Brno, Czech Republic, 2021.

[Boa+08] K. Boakye et al., « Overlapped speech detection for improved speaker diarization in multiparty meetings », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Las Vegas, NV, USA : IEEE, mars 2008, p. 4353-4356, DOI : 10.1109/ICASSP.2008.4518619.

[Boc+08] T. Bocklet et al., « Age and gender recognition for telephone applications based on GMM supervectors and support vector machines », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2008), p. 1605-1608, DOI : 10.1109/ICASSP.2008.4517932.

[Bre+20] H. Bredin et al., « Pyannote.Audio : Neural Building Blocks for Speaker Diarization », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, p. 7124-7128, DOI : 10.1109/ICASSP40776.2020.9052974.

[Bro82] V. R. Brooks, « Sex differences in student dominance behavior in female and male professors' classrooms », en, in : *Sex Roles* 8.*7* (juill. 1982), p. 683-690, ISSN : 1573-2762, DOI : 10.1007/BF00287565.

[BS83] E. Bohn et R. Stutman, « Sex-Role Differences in the Relational Control Dimension of Dyadic Interaction », in : *Women's Studies in Communication* 6.*2* (oct. 1983), p. 96-104, ISSN : 0749-1409, DOI : 10.1080/07491409.1983.11089656.

[BSF94] Y. Bengio, P. Simard et P. Frasconi, « Learning long-term dependencies with gradient descent is difficult », in : *IEEE Transactions on Neural Networks* 5.*2* (mars 1994), p. 157-166, ISSN : 1941-0093, DOI : 10.1109/72.279181.

[Che+21] S. Chen et al., « WavLM : Large-Scale Self-Supervised Pre-training for Full Stack Speech Processing », in : *arXiv* (2021), arXiv : 2110.13900 [cs.CL].

[Che99] R. Chengalvarayan, « Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition », in : *Sixth European conference on speech communication and technology*, 1999.

[Cho+19] D. Choi et al., « On Empirical Comparisons of Optimizers for Deep Learning », en, in : (déc. 2019).

[CK02] A. de Cheveigne et H. Kawahara, « YIN, a fundamental frequency estimator for speech and musica) », en, in : *The Journal of the Acoustical Society of America* 111.*4* (2002), p. 14.

[CM15] M.-J. Caraty et C. Montacie, « Detecting Speech Interruptions for Automatic Conflict Detection », in : *Conflict and Multimodal Communication : Social Research and Machine Intelligence* (fév. 2015), p. 377-401, ISSN : 978-3-319-14080-3, DOI : 10.1007/978-3-319-14081-0_18.

[CMW03] C. Cieri, D. Miller et K. Walker, « The Fisher Corpus : a Resource for the Next Generations of Speech-to-Text », en, in : (2003).

[Cor+20] S. Cornell et al., « Detecting and Counting Overlapping Speakers in Distant Speech Scenarios », in : *Proc. ISCA Interspeech*, Shanghai, China, 2020, p. 3107-3111, doi : 10.21437/Interspeech.2020-2671.

[ÇS06] O. Çetin et E. Shriberg, « Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site : insights for automatic speech recognition », in : *Proc. ISCA Interspeech*, Pittsburgh, USA, 2006, paper 1915-Mon2A2O.6.

[CW91] D. G. Childers et K. Wu, « Gender recognition from speech. Part II : Fine analysis », eng, in : *The Journal of the Acoustical Society of America* 90.*4 Pt 1* (oct. 1991), p. 1841-1856, issn : 0001-4966, doi : 10.1121/1.401664.

[Dau+15] Y. N. Dauphin et al., « RMSProp and equilibrated adaptive learning rates for non-convex optimization. », in : *CoRR* abs/1502.04390 (2015).

[Dev+18] J. Devlin et al., « BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding », in : *CoRR* abs/1810.04805 (2018).

[Dou+18] D. Doukhan et al., « An Open-Source Speaker Gender Detection Framework for Monitoring Gender Equality », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, avr. 2018, p. 5214-5218.

[Dru89] K. Drummond, « A backward glance at interruptions », en, in : *Western Journal of Speech Communication* 53.*2* (août 1989), p. 150-166, issn : 0193-6700, doi : 10.1080/10570318909374297.

[EF71] P. Ekman et W. V. Friesen, « Constants across cultures in the face and emotion », in : *Journal of Personality and Social Psychology* 17 (1971), p. 124-129, issn : 1939-1315, doi : 10.1037/h0030377.

[ÉM10] C. Émond et L. Ménard, « Les marques prosodiques des styles de parole dans les téléjournaux québécois », in : *Communication Vol. 27/2* (mars 2010), p. 150-165, issn : 1189-3788, 1920-7344, doi : 10.4000/communication.3107.

[Est+10] Y. Estève et al., « The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news », in : *Proc. Language Resources and Evaluation Conference (LREC)*, Valetta, Malta, 2010, p. 1686-1689.

[Eyb+16] F. Eyben et al., « The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing », in : *IEEE Transactions on Affective Computing* 7.*2* (avr. 2016), p. 190-202, ISSN : 1949-3045, DOI : 10.1109/TAFFC.2015.2457417.

[Fay+20] C. Fayet et al., « FlexEval, création de sites web légers pour des campagnes de tests perceptifs multimédias », in : *6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)*, sous la dir. de Christophe Benzitoun et al., Nancy, France : ATALA, 2020, p. 22-25.

[Fer23] L. Ferrer, *Analysis and Comparison of Classification Metrics*, en, juin 2023.

[Fer77] N. Ferguson, « Simultaneous speech, interruptions and dominance », en, in : *British Journal of Social and Clinical Psychology* 16.*4* (1977), p. 295-302, ISSN : 2044-8260, DOI : 10.1111/j.2044-8260.1977.tb00235.x.

[Fis+07] J. G. Fiscus et al., *2004 Spring NIST Rich Transcription (RT-04S) Development Data*, déc. 2007, DOI : 10.35111/FEHG-1397.

[Fu+22] S.-W. Fu et al., « Improving Meeting Inclusiveness using Speech Interruption Analysis », in : *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, New York, NY, USA : Association for Computing Machinery, oct. 2022, p. 887-895, ISBN : 978-1-4503-9203-7, DOI : 10.1145/3503161.3548379.

[Fuj+19] Y. Fujita et al., « End-to-End Neural Speaker Diarization with Self-Attention », in : *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2019, p. 296-303, DOI : 10.1109/ASRU46091.2019.9003959.

[Gal+06] S. Galliano et al., « Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News », in : *Proc. Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, mai 2006, p. 139-142.

[Gar+02] J. Garolofo et al., *NIST Rich Transcription 2002 Evaluation : A Preview*, rapp. tech., 2002.

[Gar+20] L. P. Garcia Perera et al., « Speaker Detection in the Wild : Lessons Learned from JSALT 2019 », en, in : *Proc. ISCA Speaker and Language Recognition Workshop (Odyssey)*, 2020, p. 415-422, DOI : 10.21437/Odyssey.2020-59.

[Gei+13a] J. T. Geiger et al., « Using linguistic information to detect overlapping speech », in : *Proc. ISCA Interspeech*, Lyon, France, 2013.

[Gei+13b] J.T. Geiger et al., « Detecting overlapping speech with long short-term memory recurrent neural networks », in : *Proc. ISCA Interspeech*, Lyon, France, 2013, p. 1668-1672.

[GG17] G. Gelly et J.-L. Gauvain, « Optimization of RNN-based speech activity detection », in : *IEEE/ACM Transactions on Audio, Speech and Language Processing* 26.*3* (2017), p. 646-656.

[Gha+10] H. Ghaemmaghami et al., « Noise robust voice activity detection using features extracted from the time-domain autocorrelation function », in : *11th Annual Conference of the ISCA*, 2010, p. 3118-3121.

[Gir+12] A. Giraudel et al., « The REPERE Corpus : a multimodal corpus for person recognition », in : *Proc. Language Resources and Evaluation Conference (LREC)*, Istanbul, Turkey, 2012, p. 1102-1107.

[Gra+12] G. Gravier et al., « The ETAPE corpus for the evaluation of speech-based TV content processing in the French language », in : *Proc. Language Resources and Evaluation Conference (LREC)*, Istanbul, Turkey, 2012, p. 114-118.

[Gra+15] S. Graf et al., « Features for voice activity detection : a comparative analysis », in : *Proc. European Association For Signal Processing (EURASIP)* 2015.*1* (2015), p. 1-15.

[GRB19] M. Garnerin, S. Rossato et L. Besacier, « Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance », in : *AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, New York, NY, USA, 2019, p. 3-9, ISBN : 978-1-4503-6917-6, DOI : 10.1145/3347449.3357480.

[GSR91] H. Gish, M.-H. Siu et R. Rohlicek, « Segregation of speakers for speech recognition and speaker identification », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1991), 873-876 vol.2, DOI : 10.1109/ICASSP.1991.150477.

[Gui22] M.-N. Guillot, « Revisiting the methodological debate on interruptions : From measurement to classification in the annotation of data for cross-cultural research », en, in : *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)* (juill. 2022), p. 25-47, ISSN : 1018-2101, 2406-4238, DOI : 10.1075/prag.15.1.02gui.

[H70] Y. V. H, « ON GETTING A WORD IN EDGEWISE », in : *Chicago Linguistics Society, 6th Meeting, 1970* (1970), p. 567-578.

[Hag+17] H. Hagerer et al., « Enhancing LSTM RNN-based Speech Overlap Detection by Artificially Mixed Data », en, in : (2017), p. 8.

[Hen+20] R. Hennequin et al., « Spleeter : a fast and efficient music source separation tool with pre-trained models », in : *Journal of Open Source Software* 5.*50* (2020), p. 2154, DOI : 10.21105/joss.02154.

[Her+22] N. Hervé et al., « Using ASR-Generated Text for Spoken Language Modeling », in : *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, virtual+Dublin : Association for Computational Linguistics, mai 2022, p. 17-25, DOI : 10.18653/v1/2022.bigscience-1.2.

[Hsu+21] W.-N. Hsu et al., « HuBERT : Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units », en, in : *arXiv* (juin 2021).

[Hui09] M. Huijbregts, « Speech Overlap Detection in a Two-Pass Speaker Diarization System », en, in : *Proc. ISCA Interspeech* (2009), p. 4.

[HW07] M. Huijbregts et C. Wooters, « The blame game : performance analysis of speaker diarization system components. », in : *Proc. ISCA Interspeech*, Antwerp, Belgium, 2007.

[Itu96] A Itu, « silence compression scheme for G. 729 optimized for terminals conforming to recommendation V. 70 », in : *ITU-T Recommendation G* 729 (1996).

[Jun+21] J.-W. Jung et al., « Three-Class Overlapped Speech Detection Using a Convolutional Recurrent Neural Network », en, in : *Proc. ISCA Interspeech*, ISCA, août 2021, p. 3086-3090, DOI : 10.21437/Interspeech.2021-149.

[KB14] D. P Kingma et J. Ba, « Adam : A method for stochastic optimization », in : *arXiv preprint arXiv :1412.6980* (2014).

[KB18] E. Kazimirova et A. Belyaev, « Automatic Detection of Multi-speaker Fragments with High Time Resolution », in : *Proc. ISCA Interspeech*, Hyderabad, India, 2018, DOI : 10.21437/Interspeech.2018-1878.

[KC83] C. W. Kennedy et C. T. Camden, « A new look at interruptions », in : *Western Journal of Speech Communication* 47.*1* (avr. 1983), p. 45-58, ISSN : 0193-6700, DOI : 10.1080/10570318309374104.

[KDO05] T. Kristjansson, S. Deligne et P. Olsen, « Voicing features for robust speech detection », in : *Entropy* 2.*2.5* (2005), p. 3.

[Kim+21] M. Kim et al., « North America Bixby Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021 », in : *arXiv* (2021).

[Kre+12] W. Kretzschmar Jr. et al., *Digital Archive of Southern Speech*, fév. 2012, DOI : 10.35111/5BNT-R659.

[Kun+19] M. Kunešová et al., « Detection of Overlapping Speech for the Purposes of Speaker Diarization », en, in : *Speech and Computer*, Lecture Notes in Computer Science, Cham : Springer International Publishing, 2019, p. 247-257, DOI : 10.1007/978-3-030-26061-3_26.

[Lar+21] A. Larcher et al., « Speaker Embedding For Diarization Of Broadcast Data In The ALLIES Challenge », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021, p. 5799-5803.

[Lav+19] M. Lavechin et al., « End-to-end domain-adversarial voice activity detection », in : *arXiv preprint arXiv :1910.10655* (2019).

[Lec+98] Y. Lecun et al., « Gradient-based learning applied to document recognition », in : *Proceedings of the IEEE* 86.*11* (nov. 1998), p. 2278-2324, ISSN : 1558-2256, DOI : 10.1109/5.726791.

[LLM16] A. Larcher, K. A. Lee et S. Meignier, « An Extensible Speaker Identification SIDEKIT in Python », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, p. 5095-5099, DOI : 10.1109/ICASSP.2016.7472648.

[Luz04] D. Luzzati, « Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané », in : *MIDL 2004*, Paris, France, 2004.

[Mac+20] M. Macary et al., « Multi-corpus Experiment on Continuous Speech Emotion Recognition : Convolution or Recurrence ? », in : *SPECOM* (2020), p. 304-314, DOI : 10.1007/978-3-030-60276-5_30.

[Man+19] V. Manohar et al., « Acoustic Modeling for Overlapping Speech Recognition : Jhu Chime-5 Challenge System », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, United Kingdom : IEEE, mai 2019, p. 6665-6669, ISBN : 978-1-4799-8131-1, DOI : 10.1109/ICASSP.2019.8682556.

[Mcc+05] I. Mccowan et al., « The AMI Meeting Corpus », in : *Proceedings on the Conference on Methods and Techniques in Behavioral Research*, Wageningen, Netherlands, 2005, p. 4.

[Mil05] R. L. Miller, « Nature of the Vocal Cord Wave », in : *The Journal of the Acoustical Society of America* 31.*6* (juill. 2005), p. 667-677, ISSN : 0001-4966, DOI : 10.1121/1.1907771.

[MMH71] L. Meltzer, W. N. Morris et D. P. Hayes, « Interruption outcomes and vocal amplitude : Explorations in social psychophysics », in : *Journal of Personality and Social Psychology* 18 (1971), p. 392-402, ISSN : 1939-1315, DOI : 10.1037/h0030993.

[Moo96] T.K. Moon, « The expectation-maximization algorithm », in : *IEEE Signal Processing Magazine* 13.*6* (nov. 1996), p. 47-60, ISSN : 1558-0792, DOI : 10.1109/79.543975.

[MŽ20] J. Málek et J. Žďánský, « Voice-Activity and Overlapped Speech Detection Using x-Vectors », in : *Text, Speech, and Dialogue*, Cham : Springer International Publishing, 2020, p. 366-376, ISBN : 978-3-030-58323-1.

[Ng+12] T. Ng et al., « Developing a speech activity detection system for the DARPA RATS program », in : *Thirteenth annual conference of the international speech communication association*, 2012.

[NGM01] E. Nemer, R. Goubran et S. Mahmoud, « Robust voice activity detection using higher-order statistics in the LPC residual domain », in : *IEEE Transactions on Audio, Speech, and Language Processing* 9.*3* (2001), p. 217-231.

[Noh92] M. Nohara, « Sex differences in interruption : An experimental reevaluation », en, in : *Journal of Psycholinguistic Research* 21.*2* (mars 1992), p. 127-146, issn : 1573-6555, doi : 10.1007/BF01067991.

[PC96] E.S. Parris et M.J. Carey, « Language independent gender identification », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1996, p. 685-688, isbn : 978-0-7803-3192-1, doi : 10.1109/ICASSP.1996.543213.

[PES01] T. Pfau, D. PW Ellis et A. Stolcke, « Multispeaker speech activity detection for the ICSI meeting recorder », in : *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, 2001, p. 107-110.

[Pha+19] H. Phan et al., « Unifying Isolated and Overlapping Audio Event Detection with Multi-Label Multi-Task Convolutional Recurrent Neural Networks », en, in : *arXiv :1811.01092 [cs, eess, stat]* (fév. 2019).

[RB18] M. Ravanelli et Y. Bengio, « Speaker Recognition from Raw Waveform with SincNet », in : *Speech and Language Technology SLT* (2018), p. 1021-1028, doi : 10.1109/SLT.2018.8639585.

[RJ75] W. T. Rogers et S. S. Jones, « Effects of Dominance Tendencies on Floor Holding and Interruption Behavior in Dyadic Interaction1 », en, in : *Human Communication Research* 1.*2* (1975), p. 113-122, issn : 1468-2958, doi : 10.1111/j.1468-2958.1975.tb00259.x.

[RLY13] N. Ryant, M. Liberman et J. Yuan, « Speech activity detection on youtube using deep neural networks. », in : *Proc. ISCA Interspeech*, Lyon, France, 2013, p. 728-731.

[Ros58] F. Rosenblatt, « The perceptron : A probabilistic model for information storage and organization in the brain », in : *Psychological Review* 65 (1958), p. 386-408, issn : 1939-1471, doi : 10.1037/h0042519.

[Ros77] M. Z. Rosaldo, « Barrie Thorne and Nancy Henley, eds., Sex and language Difference and dominance. Rowley, Mass. : Newbury House, 1975. Pp. xii–311. », in : *Language in Society* 6.*1* (1977), p. 110-113, DOI : 10.1017/S0047404500004930.

[Rou+06] J.E. Rougui et al., « Fast Incremental Clustering of Gaussian Mixture Speaker Models for Scaling up Retrieval In On-Line Broadcast », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, t. 5, mai 2006, p. V-V, DOI : 10.1109/ICASSP.2006.1661327.

[Rus80] J. A. Russell, « A circumplex model of affect. », en, in : *Journal of Personality and Social Psychology* 39.*6* (1980), p. 1161-1178, ISSN : 0022-3514, DOI : 10.1037/h0077714.

[Rya+21] N. Ryant et al., « The Third DIHARD Diarization Challenge », in : *Proc. ISCA Interspeech*, Brno, Czechia, 2021, p. 3570-3574.

[Sai+15] T. N. Sainath et al., « Learning the speech front-end with raw waveform CLDNNs », in : *Proc. ISCA Interspeech*, Dresden, Germany, 2015, p. 1-5, DOI : 10.21437/Interspeech.2015-1.

[Saj+18] N. Sajjan et al., « Leveraging LSTM Models for Overlap Detection in Multi-Party Meetings », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, avr. 2018, p. 5249-5253, DOI : 10.1109/ICASSP.2018.8462548.

[Sch78] G. Schwarz, « Estimating the Dimension of a Model », in : *The Annals of Statistics* 6.*2* (mars 1978), p. 461-464, ISSN : 0090-5364, 2168-8966, DOI : 10.1214/aos/1176344136.

[SCP15] D. Snyder, G. Chen et D. Povey, « Musan : A music, speech, and noise corpus », in : *arXiv preprint arXiv :1510.08484* (2015).

[SGH72] E. Schegloff, J. Gumperz et D. Hymes, « Sequencing In Conversational Openings », in : *American Anthropologist - AMER ANTHROPOL*, déc. 1972, p. 346-380, ISBN : 978-3-11-088043-4, DOI : 10.1515/9783110880434-006.

[SH17] N. Shokouhi et J. H. L. Hansen, « Teager–Kaiser Energy Operators for Overlapped Speech Detection », en, in : *IEEE/ACM Transactions on Audio, Speech and Language Processing* 25.*5* (mai 2017), p. 1035-1047, DOI : 10.1109/TASLP.2017.2678684.

[SH98] R. Sarikaya et J. HL Hansen, « Robust detection of speech activity in the presence of noise », in : *Proc. ICSLP*, t. 4, Citeseer, 1998, p. 1455-8.

[SKS99] J. Sohn, N. S. Kim et W. Sung, « A statistical model-based voice activity detection », in : *IEEE Signal Processing Letters* 6.*1* (1999), p. 1-3.

[Sny+18] D. Snyder et al., « X-Vectors : Robust DNN Embeddings for Speaker Recognition », en, in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB : IEEE, 2018, p. 5329-5333, doi : 10.1109/ICASSP.2018.8461375.

[SSB01] E. Shriberg, A. Stolcke et D. Baron, « Observations on overlap : findings and implications for automatic processing of multi-party conversation », in : *Proc. ISCA Interspeech*, Aalborg, Denmark, 2001.

[SSJ74] H. Sacks, E. A. Schegloff et G. Jefferson, « A Simplest Systematics for the Organization of Turn-Taking for Conversation », in : *Language* 50.*4* (1974), p. 696-735, issn : 00978507, 15350665.

[Str+03] S. Strassel et al., *SLX Corpus of Classic Sociolinguistic Interviews*, nov. 2003, doi : 10.35111/109X-K373.

[T+96] Martin T. et al., *DCIEM/HCRC*, 1996, doi : 10.35111/4540-J072.

[TAE10] F. Torreira, M. Adda-Decker et M. Ernestus, « The Nijmegen Corpus of Casual French », in : *Speech Communication* 52 (2010), p. 201-212.

[Tan94] D. Tannen, *Gender and Discourse*, en, Oxford University Press, juill. 1994, isbn : 978-0-19-972782-7.

[Tho+14] S. Thomas et al., « Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2014, p. 2519-2523.

[Wan+21] K. Wang et al., « The ByteDance Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021 », in : *arXiv* (2021).

[WC91] K. Wu et D. G. Childers, « Gender recognition from speech. Part I : Coarse analysis », eng, in : *The Journal of the Acoustical Society of America* 90.*4 Pt 1* (oct. 1991), p. 1828-1840, issn : 0001-4966, doi : 10.1121/1.401663.

[Wie+65] A. N. Wiens et al., « Interview interaction behavior of supervisors, head nurses, and staff nurses », eng, in : *Nursing Research* 14.*4* (1965), p. 322-329, ISSN : 0029-6562.

[Woo+00] K-H. Woo et al., « Robust voice activity detection algorithm for estimating noise spectrum », in : *Electronics Letters* 36.*2* (2000), p. 180-181.

[WZ75] C. West et D. H. Zimmerman, « Sex roles, interruptions and silences in conversation », in : *Language and Sex : Difference and Dominance*, sous la dir. de Barrie Thorned et Nancy Henley, Rowley, Mass. : Newbury House, 1975, p. 105-129.

[Yan+21a] S.-W. Yang et al., « SUPERB : Speech Processing Universal PERformance Benchmark », en, in : *Proc. ISCA Interspeech*, 2021, p. 1194-1198, DOI : 10.21437/Interspeech.2021-1775.

[Yan+21b] Y.-Y. Yang et al., « TorchAudio : Building Blocks for Audio and Speech Processing », in : *arXiv* (2021).

[YG93] G. Yu et H. Gish, « Identification of speakers engaged in dialog », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (1993), 383-386 vol.2, DOI : 10.1109/ICASSP.1993.319319.

[YH20] M. Yousefi et J. H.L. Hansen, « Frame-Based Overlapping Speech Detection Using Convolutional Neural Networks », in : *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, p. 6744-6748, DOI : 10.1109/ICASSP40776.2020.9053108.

[ZH12] M. Zelenák et J. Hernando, « Speaker overlap detection with prosodic features for speaker diarisation », en, in : *IET Signal Processing* 6.*8* (oct. 2012), p. 798-804, ISSN : 1751-9683, DOI : 10.1049/iet-spr.2011.0233.

GLOSSAIRE

ARCOM Autorité de régulation de la communication audiovisuelle et numérique. 18

BiLSTM Bidirectional LSTM. 41

CE Cross entropy. 39

CNN Convolutional Neural Network. 42, 43, 56

CNRTL Centre National de Ressources Textuelles et Lexicales. 22

CSA Conseil Supérieur de l'Audiovisuel. 18

DER Detection Error Rate. 54, 56

EER Equal Error Rate. 54, 56

GEM Gender Equality Monitoring. 17, 32

INA Institut National de l'Audiovisuel. 17, 18, 61

LSTM Long Short-Term Memory. 41, 44

OSD Overlapped Speech Detection. 9, 56, 57

TCN Temporal Convoluted Network. 9, 43, 44

TRP Transition Relevant Places. 24, 29

VAD Voice Activity Detection. 9, 12, 54–56

Annexe A

Annexes

A.1 Moyenne pondérée sur WavLM

Cette section a pour objectif de tester succinctement l'efficacité d'apprendre les poids de la moyenne pondérée par rapport à une moyenne non pondérée qui est utilisée dans le manuscrit. Pour cela nous comparons la même architecture, WavLM large suivie d'un TCN appris sur le corpus DIHARD avec et sans geler la couche de poids.

Modèle	Précision	Rappel	F1-score	Confiance 95 %
Appris	0.719	0.647	0.681	[0.652, 0.709]
Gelé	0.661	0.698	0.679	[0.653, 0.703]

Table A.1 – Comparaison entre une couche apprise ou non pour les poids de la moyenne pondérée des couches de WavLM

Les résultats obtenus sont présentés dans la table A.1. Cette table montre un très léger gain en utilisant une couche apprise avec le modèle, sans pour autant être significatif dans notre situation.

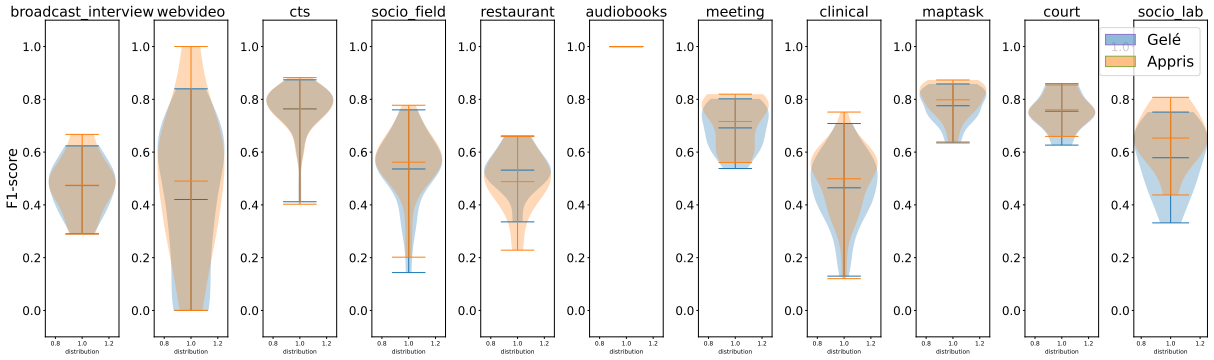


Figure A.1 – Score par domaine pour une couche de poids apprise et une couche figée

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Le Mans Université

Titre : Interactions entre locuteurs : de la détection de la parole superposée à la détection des interruptions.

Mot clés : interruption, parole superposée, traitement de la parole, intelligence artificielle

Résumé : Le projet ANR GEM, à l'initiative de l'institut national de l'audiovisuel, vise à étudier les différences de traitement et de représentation entre les femmes et les hommes dans les médias. Ce projet encourage la collaboration entre la recherche menée en sciences des médias et du langage et celle menée en informatique. Un des objectifs du projet est de favoriser la création d'outils automatiques afin de généraliser et favoriser les études SHS sur de larges corpus.

Dans cette thèse, nous nous focaliserons sur des outils de traitement du signal qui faciliteront la caractérisation des représentations des locuteurs. Plus précisément, nous proposons des méthodes pour détecter et caractériser automatiquement les interruptions au cours d'une conversation issue d'émissions de débats télévisuels.

L'interruption est une notion subjective, dont la définition n'est pas consensuelle. Dans notre domaine du traitement automatique, cette tâche est nouvelle, sans cadre et avec peu de ressources. Nous proposons, dans un premier temps, de réduire la définition des interruptions au cas particulier de la parole superposée conformément à la littérature en sociologie et en sciences du langage. Un outil de détection de la présence d'activité vocale mono et multi-locuteur a été développé dans ce contexte. Le développement d'un tel outil pose la question au-delà d'une évaluation quantitative. À partir des segments multi-locuteurs, plusieurs études ont été réalisées portant sur leur durée ainsi que sur leur contenu linguistique.

Dans un second temps, nous nous sommes intéressés spécifiquement à la détection des interruptions. L'apprentissage de modèles neuronaux dédiés a nécessité la collecte et l'annotation d'un corpus. En guidant les annotateurs, nous avons abouti à une définition de l'interruption par l'exemple. La création d'un tel corpus a permis de développer un modèle de classification binaire d'interruption pour qualifier les segments multi-locuteurs précédemment détectés.

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Le Mans Université

Title: Speaker interactions : from overlapped speech to interruption detection.

Keywords: Overlapped speech, Speech processing, Artificial intelligence, Interruption

Abstract: The ANR GEM project, initiated by the National Audiovisual Institute, aims to study the differences in treatment and representation between women and men in the media. This project encourages collaboration between research in media and language sciences and research in computer science. One of the project's objectives is to promote the creation of automated tools to generalize and facilitate social sciences and humanities studies on large corpora.

In this thesis, we will focus on signal processing tools that facilitate the characterization of speaker representations. Specifically, we propose methods to automatically detect and characterize interruptions during conversations from television debate programs.

Interruption is a subjective concept with no consensus on its definition. In our field of automatic processing, this task is new, lacks a framework, and has limited resources. We propose, initially, to narrow down the definition of interruptions to the specific case of overlapped speech, following the literature in sociology and language sciences. A tool for detecting the presence of single and multiple speakers' vocal activity has been developed in this context. Developing such a tool raises questions beyond quantitative evaluation. Several studies have been conducted on the duration and linguistic content of the multi-speaker segments.

Subsequently, we specifically focused on interruption detection. Training dedicated neural models required the collection and annotation of a corpus. By guiding the annotators, we arrived at an example-based definition of interruption. Creating such a corpus enabled the development of a binary interruption classification model to characterize the previously detected multi-speaker segments.