L'Art de la Voix : Caractériser l'information vocale dans un choix artistique

Adrien Gresse

ACADÉMIE D'AIX-MARSEILLE
AVIGNON UNIVERSITÉ

THÈSE

Présentée à Avignon Université pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : INFORMATIQUE

École Doctorale 536 « Agrosciences & Sciences »
Laboratoire Informatique d'Avignon (EA 4128)

L'Art de la Voix

Caractériser l'information vocale dans un choix artistique

par
Adrien Gresse

Soutenue publiquement le 6 février 2020 devant un jury composé de :

M.	Emmanuel VINCENT	Directeur de recherche, INRIA-Nancy, LORIA	Rapporteur
M.	Jean-Julien AUCOUTURIER	Chargé de recherche, CNRS, IRCAM	Rapporteur
M.	Yannick ESTÈVE	Professeur à l'Université d'Avignon, LIA	Examinateur
Mme	Lori LAMEL	Directrice de recherche, CNRS, LIMSI	Présidente du jury
Mme	Julie MAUCLAIR	Maître de conférences à l'Université de Toulouse, IRIT	Examinatrice
M.	Richard DUFOUR	Maître de conférences à l'Université d'Avignon, LIA	Co-Encadrant
M.	Vincent LABATUT	Maître de conférences à l'Université d'Avignon, LIA	Co-Encadrant
M.	Jean-François BONASTRE	Professeur à l'Université d'Avignon, LIA	Directeur de thèse

LABORATOIRE INFORMATIQUE D'AVIGNON

Remerciements

Je souhaiterais remercier tout d'abord les membres du jury d'être venus à Avignon pour ma soutenance de thèse. Merci donc à madame Lori Lamel, présidente du jury et aux examinateurs : madame Julie Mauclair et monsieur Yannick Estève. Merci également aux deux rapporteurs : monsieur Emmanuel Vincent et monsieur Jean-Julien Aucouturier. Ce fut pour moi un grand honneur d'obtenir votre reconnaissance.

De manière plus formelle, je souhaiterais remercier la Fondation de l'Université d'Avignon pour m'avoir attribué la bourse Pierre Bergé et financé ce projet de recherche.

Cette thèse est aussi le fruit du travail de plusieurs personnes. Merci donc à mon directeur de thèse Jean-François Bonastre et mon premier encadrant Richard Dufour qui m'ont guidé dans la réalisation de ce travail et qui m'ont aidé à sortir de l'impasse. Je remercie également mon deuxième encadrant Vincent Labatut pour son travail de relecture, sa rigueur et ses conseils.

Toutes ces années au LIA n'auraient pas été aussi extraordinaires sans les personnes qui font le laboratoire, je pense notamment à l'équipe administrative qui fait des merveilles. Merci à Mickael pour m'avoir aidé à mettre en place ma première expérience. Merci à tous les collègues du LIA qui font sa convivialité. Je pense notamment à Driss pour les nombreuses discussions sur les réseaux de neurones, mais aussi à tous ceux qui m'ont soutenu pendant ma thèse. Je souhaiterais remercier plus particulièrement mes colocataires de bureau : Paul, Jonas, Nico et Sondes ainsi que ceux ayant fait une plus courte escale en C007. Je voudrais également remercier tous les doctorants : Moez, Imedh, Cédric, Killian, Manu, Xavier, Waad, Zack, Mohamed, Mayeul, Cyril, Elvys, Matthieu, Carlos, Titouan, Mathias, Tessnim, Anaïs, Paul-Gauthier, Antoine, Salima, Luis et tous ceux que j'oublie au moment où j'écris ces lignes, mais qui ne connaissent que trop bien les interminables pauses café du CERI. Aussi, merci Téva pour tes conseils typographiques, tu peux maintenant mettre à jour Pegasus et installer Rust sur toutes les

machines.

J'aimerais également remercier ma famille : mes parents, mes frères, mais aussi tous les cousins, cousines, oncles et tantes pour leur soutient immuable. Merci aussi à Renée pour m'avoir aidé à prendre du recul sur la difficulté de ce travail. Je souhaiterais aussi remercier mes amis de toujours : Enzo, Alex, Vincent, Laeti, Pierre et Marine. Votre amitié compte beaucoup à mes yeux. Aussi, j'aimerais remercier Carl pour avoir partagé avec moi cette passion pour le vélo qui m'a permis de souffler un peu. Merci également à tous les Fils de TugLand. Nous n'avons pas eu l'occasion de beaucoup nous voir, mais les rares fois furent pour moi un grand bol d'air frais. Enfin, j'aimerais remercier une personne en particulier, celle qui m'a accompagné dans cette épreuve. Anaïs, merci d'avoir été si patiente et de m'avoir donné la force de continuer, sans toi, je n'aurais pas été jusqu'ici.

Résumé

Pour atteindre une audience internationale, les productions audiovisuelles (films, séries, jeux vidéo) doivent être traduites dans d'autres langues. Très souvent les voix de la langue d'origine de l'œuvre sont doublées par de nouvelles voix dans la langue cible. Le processus de casting vocal visant à choisir une voix (un acteur) en accord avec la voix originale et le personnage joué est réalisé manuellement par un directeur artistique (DA). Aujourd'hui, les DAs ont une inclination pour les nouveaux « talents » (moins coûteux et plus disponibles que les doubleurs expérimentés), mais ils ne peuvent pas réaliser une audition à grande échelle. Doter les industriels de l'audiovisuel d'outils automatiques capables de mesurer l'adéquation entre une voix dans une langue source avec un contexte donné et une voix dans une langue/culture cible est donc d'un fort intérêt. De plus, au-delà du casting vocal, cette problématique du choix d'une voix fait écho aux grands enjeux scientifiques de la compréhension des mécanismes de perception de la voix.

Dans ce travail de thèse, nous utilisons des voix d'acteurs professionnels sélectionnées par un DA dans différentes langues pour des œuvres déjà doublées. Dans un premier temps, nous construisons un protocole fondé sur une méthode état-de-l'art en reconnaissance du locuteur pour mettre en évidence l'existence d'une information caractéristique du personnage dans nos données. Nous identifions également l'influence du biais linguistique sur les performances du système. Nous mettons en place, dans un second temps, un cadre méthodologique pour évaluer la capacité d'un système automatique à discriminer les paires de voix doublant un même personnage. Le système que nous avons créé repose sur des réseaux de neurones siamois. Dans ce cadre d'évaluation nous exerçons un contrôle fort des biais (contenu linguistique, genre, etc.) et nous apprenons une mesure de similarité permettant de prédire les choix du DA avec un écart significatif par rapport au hasard. Enfin, nous entraînons un espace de représentation mettant en avant l'information caractéristique du personnage, appelé p-vecteur.

Nous montrons, grâce à notre cadre méthodologique que cette représentation permet de mieux discriminer les voix de nouveaux personnages, par comparaison à une représentation orientée sur l'information locuteur. De plus, nous montrons qu'il est possible de bénéficier de la connaissance généralisée d'un modèle appris sur un jeu de données proche en utilisant les techniques de distillation de la connaissance dans les réseaux de neurones.

Cette thèse apporte un début de réponse pour la construction d'un outil d'aide au casting vocal capable de réaliser une présélection des voix pertinentes parmi un grand ensemble de voix disponibles dans une langue. Si nous avons montré dans cette thèse qu'il est possible d'extraire, à partir d'un grand volume de données, une information caractéristique d'un choix artistique souvent difficile à formaliser, il nous reste encore à mettre en évidence les facteurs explicatifs de cette décision. Nous souhaitons pouvoir fournir en complément de la sélection de voix réalisée une description des raisons de ce choix. Par ailleurs, la compréhension du processus de décision du système nous aiderait à définir la « palette vocale ». À la suite de ces travaux, nous aimerions explorer l'influence de la langue et de la culture ciblée en étendant nos travaux à plus de langues. À plus long terme, ce travail pourrait aider à comprendre comment la perception des voix a évolué depuis les débuts du doublage.

Abstract

To reach an international audience, audiovisual productions (films, TV shows, video games) must be translated into other languages. Generally, the original voice is replaced by a new voice in the target language. This process is referred as dubbing. The voice casting process aimed at choosing a voice (an actor) in accordance with the original voice and the character, is performed manually by an artistic director (AD). Today, ADs are looking for new "talents" (less expensive and more available than experienced dubbers), but they cannot perform large-scale auditions. Automatic tools capable of measuring the adequacy between a voice in a source language with a voice in a target language/culture and a given context is of great interest for audiovisual companies. In addition, beyond voice casting, this voice selection problematic echoes the major scientific questions of voice similarity and perception mechanism.

In this work, we use the voices of professional actors selected by ADs in different languages from already dubbed works. First, we set up a protocol with state-of-the-art methods in automatic speaker recognition to highlight the existence of character/role specific information in our data. We also identify the influence of linguistic bias on the performance of the system. Then, we build methodological framework to evaluate the ability of an automatic system to discriminate pairs of voices playing the same character. The system we created is based on Siamese Neural Networks. In this evaluation protocol, we apply strong constraints to avoid possible biases (linguistic content, gender, etc.) and we learn a similarity measure that reflects the AD's choices with a significant difference that is not attributed to chance. Finally, we train a new representational space highlighting the character specific information, called *p*-vector. Thanks to our methodological framework, we show that this representation allows to better discriminate the voices of new characters, in comparison to a representation oriented on the speaker information. In addition, we show that it is possible to benefit from the generalized knowledge of a model learned on a similar dataset

using knowledge distillation in neural networks.

This thesis gives a initial answer for assisted voice casting and provides automatic tools capable of preselecting the relevant voices from a large set of voices in a target language. Despite the fact that the information characteristic of an artistic choice can be extracted from a large volume of data, even if this choice is difficult to formalize, we still have to highlight the explanatory factors of the decision of the system. We would like to explain, in addition to the selection of voices, the reasons of this choice. Furthermore, understanding the decision process of the system would help us define the "voice palette". In future work, we would like to explore the influence of the target language and culture by extending our work to more languages. In the longer term, this work could help to understand how voice perception has evolved since the beginning of dubbing.

Introduction générale

La voix est un objet sonore complexe qui est défini comme l'ensemble des sons issus de l'appareil phonatoire de l'être humain. L'air expulsé par les poumons est modulé par les plis vocaux et « articulé » par le conduit vocal, c'est là que la voix se transforme en parole. La parole, et par extension, la voix, est la forme sonore du langage humain qui permet l'expression de la pensée. La voix est aussi le support d'autres moyens d'expression, comme le chant et le cri, qui sont tous deux bien distincts de la parole. De manière générale, l'amalgame entre voix et parole doit être évité. À ce titre, Billières fait une distinction plus nuancée entre la voix et la parole, d'après lui : « Voix et parole utilisent le même canal. Mais la parole s'adresse à la raison, la voix à la sensibilité. »

La voix véhicule des informations sur beaucoup d'aspects de l'existence humaine. Au-delà du message verbal en lui-même, elle communique l'attitude du locuteur envers les autres (curiosité, détermination, sincérité, etc.). D'autres informations sont transmises par la voix sur l'identité du locuteur telles que le genre, l'âge, la forme physique, l'état de santé ; mais aussi des informations d'ordre social comme les critères d'appartenance à un groupe, le statut social ; ainsi que des indices sur les aspects internes du locuteur (personnalité, état émotionnel). Cette grande richesse explique que nombre de chercheurs éprouvent un fort intérêt sur la question de la voix. Immanquablement, la définition et la délimitation de la voix, tant les fonctions et les significations qui lui sont attribuées sont nombreuses et variées, se trouve être difficile. Les études sur la voix sont donc souvent réalisées en utilisant une vue réductrice de celle-ci (Denes et al. 1993 ; Kreiman 2018), baptisée *speech chain* selon laquelle la production, l'acoustique et la perception de la voix sont étudiées indépendamment les unes des autres. Cette conception simplifiée de la voix est analysée selon différentes perspectives : acoustique, phonétique, linguistique, prosodique, psychologique, sociale, médicale, cognitive, etc. En résulte le développement d'une multitude de procédés techniques, de mesures différentes ainsi qu'une vaste terminolo-

gie qui, en définitive, accroissent encore plus la confusion à propos de ce qu'est la voix.

Dans le cas concret de la recherche appliquée à l'automatisation du traitement de la parole (au sens large), les méthodes utilisées sont fondées sur l'analyse d'une information de bas niveau, généralement l'information acoustique, portée par le signal de parole. Pourtant, il semble indéniable que des informations de plus haut niveau telles que les aspects socioculturels et émotionnels, jouent un rôle essentiel. Le traitement de la parole étant opéré sur une grande variétés de signaux, savoir comment se caractérisent les aspect émotionnels ou culturels dans la voix pour pouvoir être en mesure de démêler ces différentes informations est donc d'un très grand intérêt. Toutefois, il est difficile de définir et de mesurer l'information caractéristique des aspects socioculturels dans la voix, dont les limites ne sont pas évidentes. De plus, la caractérisation d'une information de haut niveau est d'autant plus difficile qu'elle revêt une part de subjectivité liée, par exemple à la perception de l'observateur (ce qui pose problème notamment pour l'annotation des données). Par conséquent, les chercheurs se concentrent d'avantage sur une information de bas niveau (acoustique), bien définie et facilement récupérable de surcroît.

Pour remédier à cette situation, nous souhaitons prendre en considération cette grande richesse de la voix pour ainsi déterminer de nouvelles dimensions liées à des informations de haut niveau, au même titre que les travaux effectués sur les émotions par exemple. Nous positionnons ce projet de recherche dans un cadre particulier, celui de la création artistique, notamment dans le cas de la production d'œuvres audiovisuelles. Ainsi, nous pensons qu'il est possible de tirer profit de l'information qui est générée indirectement par l'être humain, lorsqu'il associe un son particulier à un univers (un contexte) donné et cela même s'il est incapable de formaliser les raisons qui l'ont poussé vers cette décision. C'est précisément le cas lorsqu'un expert humain décide d'utiliser une voix dans un contexte donné, par exemple dans le cas du choix d'une voix de doublage. Il est ici possible de s'appuyer sur l'ensemble des productions culturelles déjà doublées (cinéma, films d'animation, jeux vidéo) pour extraire l'information de haut niveau, caractéristique des décisions prises par l'expert humain.

Ce contexte implique de prendre en considération les attentes en termes de réception d'une voix par un public visé. La voix a un rôle social indéniable, par conséquent la mise en évidence de nouvelles dimensions dans la voix est intéressante d'un point de vue sociologique. En effet, cette ca-

ractérisation de la voix doit permettre de distinguer sa fonction : « à quoi sert-elle ? » ; sa destination : « à qui s'adresse-t-elle ? » ; ainsi que les circonstances de son utilisation.

Un des objectifs principaux de nos travaux, consiste à mettre en place un cadre méthodologique scientifique permettant de mettre en évidence l'information émanant de ce choix artistique. À partir de là, nous cherchons à caractériser cette information, en supposant être en mesure de définir des types de voix (et leurs spécificités vocales propres), en accord avec les critères socioculturels du public visé par rapport à un rôle/personnage d'une œuvre donnée.

Sommaire

Chapitre 1

Introduction

Sommaire

Il est un objet qui, observé attentivement, nous permet de mettre à jour ce qui est caché au plus profond de nous-même. Cet objet, c'est la voix, « interface entre deux plus gros concepts que sont le corps et le langage ».

1.1 Qu'est ce que la voix ?

La voix est le résultat d'un mécanisme pulmonaire et laryngé permettant à l'air émis par les poumons d'être modulé pour donner naissance à la parole ou au chant. Certaines caractéristiques acoustiques permettent de décrire la voix, qu'elle soit chantée ou parlée, comme la hauteur, l'intensité et le timbre. Les chercheurs mesurent également les aspects mélodiques et rythmiques (accent, ton, pause, intensité, volume…) de la voix « qui suit inévitablement une ligne temporelle » et implique donc une notion de durée (Cornaz et al. 2014). De plus, la musicalité de la voix est « la voie privilégiée par laquelle l'enfant entre dans le langage verbal » (Golse 2005)

et notamment par le biais de la voix maternelle dès la vie intra-utérine, qui est alors le « premier objet sonore » auquel l'humain est confronté. Les conséquences irréversibles observées sur le développement et les capacités sociales de l'enfant, lorsqu'il est privé de toute expérience vocale et donc du langage par le biais de la parole (voir les cas d'études sur les « enfants sauvages » [1]), montrent le rôle primordial de la voix. Nous pouvons donc dire que la voix est essentielle pour l'être humain du fait qu'elle permet la communication parlée et participe au développement de ses aspects sociaux.

Il existe une multitude de voix et chaque voix est unique, mais bien qu'elle soit singulière, elle peut revêtir plusieurs formes (en plus des différentes modalités : parlée ou chantée). Effectivement, la voix se transforme et s'adapte en fonction de la situation. Il y a par exemple une voix qui hurle, une voix qui chuchote, une voix qui explique et une voix qui séduit. Selon Le Breton : « La voix s'écrit toujours au pluriel, comme pourrait s'écrire ainsi le visage et comme s'écrit le corps, car si la voix est toujours singulière, elle n'est jamais univoque au fil du jour et du temps pour le même individu. » (Le Breton 2011)

La voix est d'une grande richesse harmonique [2], au moins aussi riche que n'importe quel son produit par un instrument de musique. Cette richesse définit en grande partie le timbre d'une voix. Ce dernier est lié à la forme du conduit vocal ainsi qu'aux caractéristiques de la vibration glottique et à l'adaptation résonantielle (Henrich 2001). Ainsi, la voix transporte avec elle une part de l'individu. D'un point de vue plus spécifique, la voix caractérise l'identité d'une personne, elle est un formidable outil de renseignement. Elle peut nous indiquer le genre, renseigner sur l'âge et sur l'état de santé. Aussi, elle peut révéler un trait de personnalité, et des états plus éphémères, comme une émotion par exemple. La voix est également impactée par des aspects socioculturels. Elle peut, par exemple, révéler une appartenance régionale dans le cas où elle est teintée par un accent. En observant les habitudes langagières du locuteur, il est également possible de reconnaître les indices d'une origine rurale ou plutôt urbaine et de se faire une idée de son niveau d'éducation. Ainsi, la voix donne des informations sur l'appartenance sociale. Ce n'est pas pour rien que la voix est souvent désignée comme le reflet de l'âme.

Les usages de la voix suivent des codes et des conventions bien définis et

1. Confère l'expérience de Frédéric II.
2. En acoustique, un harmonique est une composante du son dont la fréquence est un multiple de la fréquence fondamentale.

partagés par un groupe d'individus. Nous sommes sans cesse en train de déchiffrer la voix qui naturellement attire l'attention, parce qu'elle est susceptible de nous concerner. Par conséquent lorsque la voix déborde des normes établies, un malaise s'installe entre les interlocuteurs. Chaque culture a ses propres codes, sa propre mélodie. La mélodie de la voix s'interprète comme une transformation, par rapport à une phrase qui serait énoncée par une voix neutre –si tant est qu'elle puisse être neutre– guidée par une grammaire et qui détermine une intention, une émotion. Cette grammaire semble plutôt indépendante de la langue (Sauter et al. 2010). En effet, si nous entendons une langue que nous ne parlons pas, nous sommes néanmoins capables de détecter des éléments de langage. Comme le souligne Jean Abitbol dans son excellent ouvrage (Abitbol 2005), il semblerait qu'il existe un méta-langage (celui des humains) utilisé dans le monde entier.

La manière dont le cerveau est organisé lui confère une efficacité redoutable pour reconnaître les unités sonores atomiques d'une langue (les phonèmes) qui, mises côte à côte, construisent les mots. Durant les premières années de la vie, le cerveau s'affûte pour la reconnaissance et pour la production des phonèmes de la langue maternelle, parce que les zones de la perception du langage (l'aire de Wernicke) et l'aire motrice (l'aire de Broca), permettant sa production, sont étroitement connectées. Au-delà d'un certain âge il devient alors difficile de percevoir de nouveaux phonèmes (provenant d'une langue étrangère) et il est souvent difficile de les produire (Oyama 1976).

À l'instar de notre acclimatation aux différents éléments du langage, le cerveau s'habitue à des voix. Cela permet notamment de reconnaître une voix familière dès les premiers mots perçus, par exemple : celles de nos parents, de nos amis, d'un professeur... De cette façon, nous sommes capables de détecter un changement dans la voix et donc de percevoir des troubles indiquant par exemple que la personne est inquiète ou fatiguée. Parmi les éléments qui aident à reconnaître le locuteur il y a le vocabulaire utilisé, la prosodie, les aspects phonétiques ou encore le timbre. Ce dernier se définit en fonction des harmoniques de la fréquence fondamentale (la fréquence la plus basse). Certains harmoniques sont particulièrement intenses et correspondent à des formants. Par exemple, un chanteur a une voix qui se distingue des personnes non-entraînées au chant par la présence de formants particuliers qui donnent une brillance et une chaleur à sa voix. Néanmoins, la particularité d'une voix se définit aussi beaucoup par ses imperfections, assimilées généralement à du bruit. Par exemple, la raucité de la voix corres-

pond à une instabilité de la fréquence fondamentale à court terme (Teston 2004).

La voix a également une dimension esthétique. La beauté est perceptible dans la voix, au même titre que pour un visage. En se plaçant dans un contexte d'écoute réduite [3], la voix peut être considérée comme un objet sonore. La voix est alors perçue en tant qu'« unité sonore dans sa matière, sa texture propre, ses qualités et ses dimensions perceptives propres » (Couprie 2001). Dans cette configuration, ni les mots, ni le locuteur n'ont d'importance, seule la voix demeure comme phénomène sonore perçu en tant que tel. Ainsi, certaines voix marquent plus que d'autres, au-delà de notre entourage même, des voix dont le timbre, le style particulier peut être perçu comme séduisant, attirant. Ces voix sont bien souvent des voix de célébrités, des chanteurs mais aussi des voix d'acteurs, généralement découvertes au travers des médias, au sens large du terme, par exemple avec le cinéma.

1.2 La place de la voix dans l'audiovisuel

La voix attire inévitablement l'attention de l'être humain, car le message qu'elle véhicule pourrait lui être destiné. Quoi de plus logique donc d'avoir intégré la voix dans le cinéma pour le captiver ?

Il existe dans le cinéma différents types de voix. Des voix dites *in*, dont le locuteur est visible à l'écran. Des voix *hors-champ* qui ne sont pas visibles, mais qui dans l'imaginaire du public, restent situées dans le même temps que l'action visible à l'écran. Enfin, les voix dites *off*, dont la source est située dans un temps autre que celui de l'action.

Depuis les débuts du cinéma parlant, la voix est devenue un objet particulier de préoccupation des cinéastes. Dans le livre intitulé *La Voix au Cinéma*, Michel Chion nous explique comment l'utilisation de la voix a évolué avec le temps, notamment avec les progrès techniques (Chion 1982). Elle évolue aussi selon les cultures, les réalisateurs et selon les publics visés. Au début du cinéma parlant, les aspects techniques avaient beaucoup plus d'influence qu'aujourd'hui. Il fallait de façon récurrente avoir recours à de la post-synchronisation [4], les micros étant trop peu sensibles pour rendre

3. Selon Pierre Schaeffer, c'est en brisant la relation de causalité qui existe entre le son et la source que l'objet sonore est perçu de manière globale.

4. Technique permettant de ré-enregistrer des dialogues ou des voix *off* en studio, après la phase de production (tournage). Ce procédé se différencie du processus de pré-synchronisation, utilisé notamment pour les films d'animation ou les jeux vidéo.

les dialogues intelligibles à l'écran. Lorsque la technique s'est améliorée, la post-synchronisation est devenue moins indispensable, bien qu'encore très employée par certains réalisateurs. Elle est notamment utilisée lorsque les acteurs tournent dans un environnement bruyant ou lorsque la Direction Artistique (DA) du film décide d'un changement nécessitant de réenregistrer un dialogue. Ce processus de post-synchronisation est appelé Automated Dialogue Recording (ADR) (souvent confondu avec un autre processus de post-synchronisation : le doublage).

De toute évidence, la question de la langue est une problématique aujourd'hui inévitable dans le monde du cinéma et finalement assez peu explorée. Selon Barnier (Barnier et al. 2013), « la langue est matériau et non simple canal d'information ». De plus, elle participe à la construction des représentations stéréotypiques. Comme le souligne Michel Chion (dans les propos rapportés par Barnier), l'emploi d'un accent dans le doublage est fréquent, notamment en Europe. En effet, l'accent colore la voix et transporte avec lui tout un imaginaire collectif lui conférant un rôle social important.

Le sous-titrage et le doublage sont les manières les plus répandues pour traduire une œuvre audiovisuelle aujourd'hui, quand elle n'est pas réadaptée et directement transposée à la culture du pays visé. Les sous-titres sont une superposition visuelle du texte à l'écran. Le sous-titrage, notamment la traduction, s'avère être une tâche particulièrement difficile et il est souvent nécessaire de supprimer une partie importante du texte original pour rester synchronisé avec les images. Le doublage est une technique de traduction qui utilise le canal audio, contrairement aux sous-titres qui sont visuels. Suivant le type de production (grand public, indépendante), les critères linguistiques pour la traduction ne sont pas les mêmes. Par ailleurs, selon Barnier : « le traducteur est au service du film et ne saurait traduire sans empathie pour son objet. »

Il est assez rare que les œuvres cinématographiques, notamment dans les films de la catégorie « art et d'essai » et provenant le plus souvent d'Europe, soient doublées. Ces films se regardent plus généralement sous-titrés pour éviter de perdre l'effet recherché. En revanche, certains films ayant bénéficié d'un certain succès ne sont jamais doublés ni sous-titrés, mais sont directement transposés à la culture visée au moyen d'une adaptation. Les américains sont sujets à cette pratique (Bonhomme 2014). C'est par exemple le cas du film *The Upside* (Burger 2017), adaptation américaine du film *Intouchable* (Toledano et al. 2011).

1.2.1 Le doublage de voix et le casting vocal

Le doublage fait référence à plusieurs processus. De manière générale, il consiste à remplacer les voix originales d'une œuvre par d'autres voix, qui plus est dans une autre langue, tout en respectant du mieux possible le phrasé et le mouvement des lèvres. Dans l'industrie du cinéma, ce processus est appelé *revoicing*. Il peut toutefois prendre la forme d'un *voiceover* (contrairement au doublage, la voix est superposée à celle d'origine) pour la narration ou pour commenter les images. Cette technique est très utilisée dans les films documentaires. Le processus de doublage fait aussi référence au processus de post-synchronisation dans lequel les acteurs enregistrent les dialogues directement en studio, soit pour des raisons techniques, soit à la demande de la réalisation. Pour désigner ce processus, les anglo-saxons utilisent le terme *dubbing*. Les comédiens francophones utilisent généralement les termes anglo-saxons *looping* ou ADR. Enfin, le doublage désigne également les étapes de « création de voix » dédiées aux films d'animation et aux jeux vidéo.

L'avantage du doublage par rapport au sous-titrage, bien qu'il ait son lot de contraintes également, est qu'il impose une faible modification du texte original. Et bien sûr, il ne divise pas l'attention du public en deux, l'une se focalisant sur les sous-titres l'autre tentant de suivre les images. Le doublage, en contre-partie, est beaucoup plus onéreux. Le choix entre sous-titrage et doublage est aussi une question culturelle. Par exemple, dans les pays d'Europe du Nord, le sous-titrage est systématique. À l'inverse, les allemands se tournent généralement vers le doublage. En France, le sous-titrage est aussi très secondaire.

Dans le cas des films d'animation, il est aujourd'hui fréquent de voir des acteurs vedettes incarner les personnages principaux. Auparavant, les personnages étaient doublés par des acteurs de doublage généralement inconnus du grand public. Le terme « doublage » n'est pas tout à fait exact ici. En effet, c'est bien la voix du personnage qui est enregistrée pour la Version Originale (VO), il s'agit donc plus de création de voix que de doublage. Il est important de distinguer les deux cas de figure : en pré-synchronisation et en post-synchronisation. Cette distinction permet de définir la valeur qui est accordée à l'acteur. Ainsi, lorsque le doublage est réalisé en pré-synchronisation, l'acteur et le personnage sont intimement liés. Dans certains cas l'acteur, sa *persona*, définit le personnage. L'acteur est dans ces cas-là souvent auditionné (« casté ») très en amont du processus de création de

l'œuvre. C'est par exemple le cas chez Disney, dans Aladdin, où les mises en scène du génie ont été largement inspirées par les improvisations de Robin Williams.

L'utilisation de voix vedettes est devenue l'essence même des gros films d'animation, surtout depuis les années 2000 où les premières images de synthèse voient le jour. À ce jeu-là, les voix vedettes, avec leur timbre particulier et leur dynamisme, constituent un avantage certain que les studios ne se privent pas d'utiliser et de mettre en avant pour la promotion de leurs films. La performance d'Eddie Murphy dans le film *Shrek* (Adamson et al. 2001) a été incontestablement appréciée par les spectateurs. Ce phénomène de « voix vedettes » dans les films d'animation a été étudié beaucoup plus en détails par Bérénice Bonhomme (Bonhomme 2014) qui présente une analyse intéressante des choix de voix réalisés dans la Version Française (VF) de différents films ayant reçu un certain succès, ou au contraire, ayant été très critiqués. L'utilisation de la voix de Gérard Lanvin pour doubler le mammouth, le « gros dur au cœur tendre » dans *L'Âge de Glace* (Wedge et al. 2002), constitue un exemple où l'accord entre le personnage et la voix est respecté. À l'inverse, certains choix sont parfois mal perçus par le public qui ressent alors le décalage entre la voix vedette (sa *persona*) et le personnage qu'elle incarne. Comme le choix de Franck Dubosc, humoriste plutôt connu pour ses rôles de séducteur macho, pour le rôle de Marin dans *Le Monde de Némo* (Stanton et al. 2003).

Un des points les plus intéressants de l'analyse de Bonhomme concerne le degré de *voice-characterization*, c'est à dire à quel niveau est situé la limite entre la voix et le personnage. Le personnage est construit avec l'acteur et les animations peuvent être réalisées selon l'attitude de l'acteur. Dans les cas les plus poussés, la ressemblance se retrouve jusqu'au niveau des traits physiques. Le casting vocal est devenu un élément clé pour le film d'animation (du moins pour les films à gros budget). Le monde du doublage, jusqu'ici presque inconnu du public, a été d'un coup propulsé au-devant de la scène, et l'utilisation des voix vedettes, au détriment des acteurs professionnels de doublage, est devenue presque systématique, du moins en ce qui concerne la VO.

Tout ce que nous venons de dire au sujet de films d'animation s'applique de la même façon pour l'industrie du jeu vidéo. Cependant, les voix vedettes ont un coût, les rendant tout simplement inaccessibles pour des productions à budget plus réduit.

Pour l'industrie du cinéma ou du jeu vidéo, la voix devient une marchandise et il est alors nécessaire de les catégoriser. Les voix sont mises en scène et elles sont évaluées en fonction de ce qu'elles disent ou selon des critères esthétiques voire des idéologies langagières (stéréotypes). Planchenault (Planchenault 2014) nous explique sur quoi repose ce marché de la voix et comment certaines voix (types de voix) se sont démarquées.

Aujourd'hui, des sociétés spécialisées ont accumulé un capital de savoirs et de connaissances fondé sur la voix. Ces entreprises proposent un large choix de voix sur catalogue et leurs prix varient en fonction du talent de l'acteur, du style de la voix, ou encore d'un accent (Planchenault 2014). Ces traits vocaux ont un rôle social important (Barnier et al. 2013). De plus, l'utilisation de ces types de voix dans les médias (au sens large), notamment dans le cinéma et aujourd'hui dans les jeux vidéo, a permis la construction d'un imaginaire collectif et sont attachés à des représentations stéréotypées. Ces voix, ou plutôt ces catégories de voix, sont un pur produit de la culture.

Pour de nouveaux acteurs désirant se lancer dans le monde du doublage ces catégories agissent par conséquent comme des attracteurs. En effet, comment peuvent-ils se positionner sur ce marché s'ils sont incapables de se comparer aux autres voix ? Ces nouveaux talents sont par conséquent incités à se plier aux exigences du marché et donc à celles de l'industrie. Ainsi, dans ce système qui s'auto-entretient, peu de place est laissée à l'originalité, à la nouveauté. Cette nouveauté même qui a autrefois séduit les différents publics et qui se trouve être à la base de ces stéréotypes.

Plusieurs facteurs influencent généralement le choix d'une voix de doublage. Il peut s'agir d'une contrainte d'ordre budgétaire ou économique, la voix désirée pouvant ne pas être disponible à ce moment. Ou bien, d'une simple volonté artistique. La DA qui opère le casting vocal a généralement une idée informelle de la voix qu'elle souhaite attribuer au personnage.

Parfois, la voix doit respecter les traits du personnage, sa *persona*. Si il s'agit d'un doublage pour une VF, il est alors possible de se référer à la voix originale et de chercher une voix similaire. Dans quelques rares cas, la VF est mieux reçue par les spectateurs que la VO. Prenons par exemple la voix de Patrick Poivey, la doublure officielle de l'acteur américain Bruce Willis. Sa voix est très appréciée en France, ce qui fait qu'elle se retrouve dans nombre de spots publicitaires. Enfin, dans le cas où la *voice-characterization* est plus poussée, la DA est plus libre et se repose donc finalement sur le talent de l'acteur.

La DA opère la sélection des voix manuellement sur un nombre restreint de voix. En effet, il n'est pas envisageable pour un opérateur humain d'écouter toutes les voix qui sont à disposition dans les bases de données. Cependant, l'utilisation d'un porte-feuille réduit de voix implique un effet de bord non souhaitable. La mise en avant systématique des mêmes voix fait qu'elles gagnent de plus en plus en popularité et qu'elles deviennent, par conséquent, de moins en moins disponibles et implicitement de plus en plus chères. Nous comprenons alors aisément pourquoi les organisations dont c'est le métier ont un fort appétit pour les nouveaux talents, peu cher et plutôt disponibles.

Le choix d'une voix de doublage est peu formalisé. L'opérateur de casting doit donc associer son expertise du doublage avec les informations que l'équipe de direction originale lui aura fournies. La DA peut avoir une idée approximative de la voix recherchée, mais en général, il est nécessaire d'essayer plusieurs voix pour en trouver une adéquate. Tout l'enjeu réside donc dans l'ensemble de voix susceptibles d'être écoutées, tout en sachant que le catalogue de voix peut être grand. De plus, il y a toujours quelque part l'espoir de trouver la « perle rare » qui propulsera l'œuvre au-devant de la scène et entraînera des retombées économiques importantes.

1.2.2 La recommandation de voix pour le casting vocal

Cette problématique du choix d'une voix de doublage peut être vue comme un problème de recommandation. Contrairement à l'humain, l'outil informatique est capable d'analyser de très grandes bases de données de voix. Il est alors possible de concevoir un système automatique d'aide au casting vocal réalisant, au préalable, la sélection d'un sous-ensemble de voix en accord avec les critères de la DA. Cette sélection peut prendre la forme d'une liste ordonnée de voix qui serait accompagnée d'une description des facteurs explicatifs de cette décision. Une telle solution permettrait aux professionnels de casting de gagner du temps ainsi que de découvrir des voix inédites.

Dans la mesure où cette pré-sélection de voix doit correspondre aux attentes du directeur artistique, il semble naturel que l'espace de représentation des voix soit construit à partir du savoir expert de l'opérateur humain. Ce dernier peut, dans l'idéal, être utilisé afin d'améliorer les recommandations, notamment à travers une boucle de rétro-action qui relie l'opérateur et le système. Comme illustré dans la figure 1.1, la DA, à la recherche d'une voix de doublage, formule sa requête à partir de la voix originale. Le sys-

tème procède à l'analyse de la requête et propose un ensemble de choix de voix possibles issus du catalogue de voix et adaptés au doublage de la voix originale. Au fur et à mesure que l'utilisateur (l'opérateur de casting) retient ou rejette les voix qui lui sont suggérées, le système affine ses recommandations.

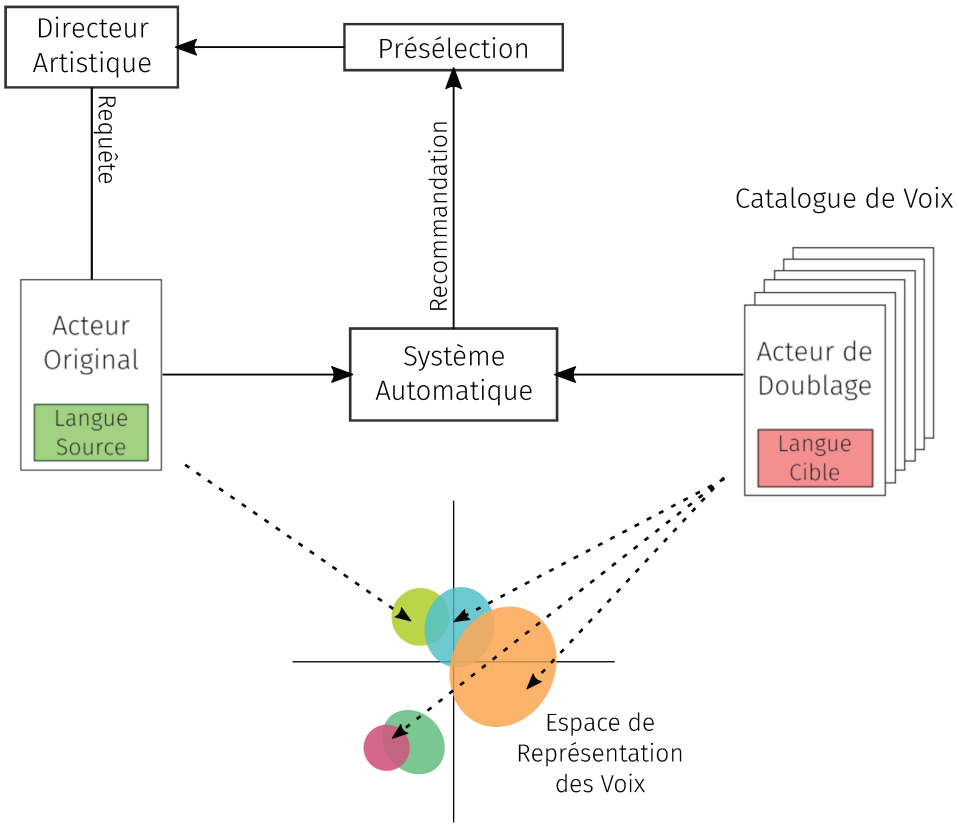

Figure 1.1 – Système de recommandation automatique de voix.

La pré-sélection des voix réalisée par le système doit s'appuyer sur une modélisation de la voix spécifique à la tâche de casting vocal. Des techniques d'apprentissage automatique peuvent être utilisées pour la construction d'un espace de représentation, à partir de l'ensemble des voix de doublage et des décisions prises par le directeur artistique. Un des objectifs de ces travaux de recherche réside dans la caractérisation de l'information contenue dans la voix, sur laquelle est fondée la décision de l'opérateur. Les techniques de recherche de l'information dans le domaine du traitement automatique de la voix (ou de la parole) sont nombreuses. Il faut donc déterminer les méthodes pouvant s'adapter à cette problématique. L'existence de cette information est hypothétique, mais les capacités humaines à la percevoir laisse supposer qu'il est possible, à partir un gros volume de données, de trouver une combinaison des facteurs expliquant une part de la variabilité observée en termes de types de voix. En effet, nous supposons

que certaines caractéristiques vocales (accent, genre, timbre…) expliquent les choix de l'opérateur de casting.

L'enjeu général d'un système de recommandation automatique est important, mais difficile à percevoir au travers d'une application pour le casting vocal. Dans l'imaginaire collectif, les systèmes automatiques sont bien perçus, dans les cas où ils vont permettre de décharger l'humain de certaines tâches fastidieuses, mais ils suscitent également des craintes, comme le remplacement possible des emplois humains par des machines. Dans un système de recommandation, l'humain, de par son rôle dans la boucle, occupe une place essentielle. En cela, les systèmes de recommandation s'inscrivent en tant qu'outils d'aide à la réalisation d'une tâche manuelle.

Au-delà des bénéfices directs que peuvent en tirer les industriels de la production audiovisuelle, un tel système constitue un objet d'étude aux perspectives scientifiques intéressantes. En effet, il permet la caractérisation de nouvelles dimensions dans la voix reliées a des aspects socioculturels encore peu explorés. L'analyse de l'espace de représentation des voix appris sur un large volume de données permet notamment d'enrichir la recommandation d'une descriptions des facteurs qui ont été déterminants pour la décision. Soumettre cette description à évaluation auprès des opérateurs de casting assure ainsi l'évolution et l'amélioration continue du système.

La recommandation comme dispositif de médiation

Le choix effectué par la DA relève d'aspects socioculturels, parfois subtils et difficilement formalisables. Par conséquent, c'est une tâche qu'il paraît compliqué d'automatiser. En se référant aux nombreuses décisions prises pour le doublage des œuvres existantes, il est possible d'extraire, à partir des voix sélectionnées, l'information qui caractérise le choix subjectif de l'opérateur. De plus, cette approche permet la constitution d'un capital de connaissances.

Dans la section 1.2.1, nous avons expliqué comment, au travers des voix utilisées, sont perpétuées les représentations et les codes propres à une culture. Dans son travail, le directeur artistique opère une forme de médiation culturelle, car il s'appuie lui-même sur ces représentations et les attentes du public en termes de réception d'une voix pour réaliser son choix. À ce titre, l'outil automatique facilite cette transmission culturelle. D'autre part, il rend possible la découverte de voix inédites ce qui constitue un atout majeur pour les sociétés de doublage. Le savoir expert du directeur artis-

tique étant difficile à formaliser, le système d'aide au casting vocal assure la pérennité de ce savoir ainsi que sa succession d'un opérateur à un autre.

1.3 Ressources utilisées : le choix des jeux vidéo

En considérant les œuvres cinématographiques, les films d'animations, les séries TV et les jeux vidéo déjà doublés, le volume de données exploitables est immense et continue de croître à chaque fois qu'une nouvelle production voit le jour. De plus, un certain nombre de ces productions ont été traduites puis doublées dans plusieurs langues et pour différentes cultures, ce qui permet la réalisation d'études transculturelles. Le grand nombre de voix de doublage associées aux différentes décisions des directeurs artistiques offre donc des possibilités intéressantes en termes d'apprentissage automatique.

Nous avons choisi de nous concentrer sur l'étude des voix de doublage issues du monde du jeu vidéo. Aujourd'hui, les jeux vidéo accordent une grande importance au sentiment d'immersion du joueur. À l'instar du cinéma, ils proposent une trame narrative accompagnée d'un grand nombre de dialogues. De plus, les grosses productions font de plus en plus appel a des voix vedettes désireuses d'incarner un personnage de jeu vidéo, mais l'utilisation de professionnels du doublage reste encore très majoritaire.

Les personnages de jeux vidéo sont souvent centrés sur des représentations stéréotypées (Schröter et al. 2014). En effet, l'utilisation de stéréotypes permet de développer plus rapidement les personnages et permet à l'histoire d'avancer plus vite. Nous supposons que les voix des personnages, comme pour leur représentation graphique, sont aussi fondées sur des stéréotypes. Les personnages sont notamment distinguables selon des prototypes de voix qui renvoient à : leur rôle (soldat, commandant, héros, vieux sage, bandit, sorcier, etc.) ; leurs intentions (condescendant, sensuel, souffrant, cynique, etc.).

Encart 1.1 : Quand les stéréotypes mènent au préjudice

Certains travaux ont permis de montrer que les stéréotypes utilisés dans les jeux vidéo peuvent avoir un caractère discriminant (Dill et al. 2007 ; Jansz et al. 2007 ; Deskins 2013 ; Deskins 2013). Ainsi, les jeux vidéo, dans leur représentation graphique des personnages, semblent décrire un monde où l'homme est catégorisé comme agressif et où la femme est sexualisée. De plus, les études réalisées considèrent que les jeux vidéo sont en général destinés à un public masculin et blanc de surcroît. Il est possible qu'une modélisation par apprentissage automatique conduise à la reproduction des discriminations que connaît la société et dont sont issues les données. Il est alors légitime de se questionner sur la question du respect de l'égalité et de la diversité.

Nous nous sommes concentrés sur deux jeux vidéo et plus particulièrement sur un jeu intitulé *Mass Effect 3*. De manière plus secondaire, nous avons également utilisé dans nos travaux un jeu vidéo nommé *The Elder Scrolls V : Skyrim*. Ces deux jeux diffèrent dans leur genre : le premier suit une thématique science-fiction et le deuxième un thème médiéval-fantastique. Toutefois, ils font tous les deux partie de la catégorie « jeux de rôle » pour laquelle un effort tout particulier est mis sur la narration. De fait, ils contiennent un grand nombre d'interactions et de dialogues entre des personnages très variés.

Dans *Mass Effect 3*, les personnages sont des soldats, des officiers et autres mercenaires, mais aussi des personnages dans des rôles de médecins, d'ingénieurs, de techniciens et des scientifiques. Quelques-uns de ces personnages sont des extraterrestres ou des intelligences artificielles. Les personnages de *Skyrim* sont quelque peu différents. On rencontre des guerriers, des légionnaires et des bandits, mais aussi des sorciers, des érudits, des marchands ou des fermiers. La plus grande part des personnages sont des humains, quelques orques et des elfes sont également présents.

1.4 Organisation du manuscrit

Nous avons divisé cette thèse en deux parties : dans la première partie, un état-de-l'art est présenté, dans lequel nous posons les frontières théoriques de notre travail. La problématique du choix d'une voix de doublage étant assez peu étudiée dans la littérature, nous nous sommes intéressés aux différents domaines du traitement de la voix, proches de cette problématique.

Nous commençons par introduire dans le chapitre 2 la notion d'identité de la voix ainsi que le cadre de la reconnaissance du locuteur. Nous y présentons les avancées de ces dernières années permettant notamment de vérifier l'identité d'une voix de manière automatique à partir du signal acoustique. Nous nous sommes penchés dans le chapitre 3 sur la question de la perception des traits de personnalité et des états plus éphémères comme les émotions, au travers de la voix. D'autre part, dans le chapitre 4 nous faisons état des travaux qui questionnent le lien entre certaines caractéristiques vocales et les jugements (les impressions) de similarité. À la fin de ce chapitre, nous faisons une rapide synthèse nous permettant de clôturer la première partie de cette thèse.

Dans la deuxième partie de cette thèse, nous présentons nos travaux. Le chapitre 5 est un avant-propos où nous présentons une approche expérimentale permettant de lever le verrou scientifique. Par la suite, nous proposons dans le chapitre 6 la mise en place d'un cadre méthodologique pour évaluer la similarité entre des voix de doublage. Enfin, nous présentons dans le chapitre 7 nos travaux portés sur la modélisation d'un espace de représentation de la voix dédié à notre tâche.

Enfin, nous donnons les conclusions de nos travaux et mettons en avant quelques pistes pour de futurs travaux dans le chapitre 8.

Première partie

Les informations véhiculées par la voix

Chapitre 2

Reconnaître le locuteur au travers de la parole

Sommaire

2.1 Reconnaissance du locuteur

L'être humain possède d'excellentes capacités pour reconnaître, à partir d'un signal de parole, la personne qui parle, d'autant plus s'il s'agit d'une voix familière (un proche ou une célébrité) (Van Lancker et al. 1985). Il est évident que la reconnaissance d'une personne peu ou pas connue est beaucoup plus difficile. Selon Hansen (Hansen et al. 2015), la familiarité avec une voix dépend du temps passé à l'écouter. De plus, suivant le contexte

(environnement bruyant, téléphone) ces capacités sont parfois soumises à rude épreuve. La reconnaissance orale du locuteur peut être une tâche délicate, par exemple lors d'une expertise juridique où les échantillons de voix examinés sont souvent de courte durée et proviennent de locuteurs inconnus. Généralement, les experts ont également recours à d'autres analyses (Bonastre, Bimbot et al. 2003). Aussi, dans le cas d'une interaction homme-machine, les systèmes mettent en œuvre de la reconnaissance du locuteur notamment pour des applications de sécurité, mais aussi pour de la personnalisation de conversation. C'est par exemple le cas des assistants vocaux personnels (Google Home, Amazon Echo…). Ces technologies sont le fruit de plusieurs décennies de recherche en Reconnaissance Automatique du Locuteur (RAL) qui ont permis de faire évoluer les techniques et d'améliorer les performances des systèmes.

Dans nos travaux, nous avons eu recours aux techniques utilisées dans les systèmes de RAL pour évaluer la similarité entre des voix. Nous nous sommes intéressés plus particulièrement aux techniques d'extraction de l'information caractéristique du locuteur ainsi qu'aux méthodes de comparaison de voix. Dans cette partie, nous tâcherons donc d'expliquer en quoi consiste la reconnaissance du locuteur et détaillerons les méthodes développées à cet effet.

2.1.1 Les différentes tâches de la reconnaissance du locuteur

En premier lieu, il faut lever la confusion portée par l'intitulé du domaine. En effet, nous distinguons deux tâches dans le domaine de la RAL : l'identification et la vérification du locuteur. Cette dernière fera l'objet d'une attention particulière.

Identification du locuteur

La première tâche consiste à identifier le locuteur dans un enregistrement audio, parmi un ensemble de personnes différentes. Cette identification peut être réalisée par un système automatique ou par un humain lors d'un test de perception. Le locuteur à identifier est obligatoirement présent dans le panel de locuteurs enregistrés par le système. Nous parlons de « scénario d'identification en espace clos ». Dans le cas d'une identification réalisée par un humain, il est possible que le locuteur à identifier ne soit pas

présent dans le panel de locuteurs, nous parlons alors de « scénario d'identification en espace ouvert ».

Vérification du locuteur

La deuxième tâche consiste à vérifier si deux échantillons de voix sont prononcés par la même personne. Elle est avant tout utilisée pour l'authentification de l'utilisateur, pour des applications bancaires par exemple, mais aussi pour les expertises judiciaires (J. P. Campbell et al. 2009). L'identité du locuteur est connu seulement pour le premier enregistrement, appelé enregistrement de comparaison ou d'apprentissage. Le deuxième fait quant à lui office d'enregistrement de test. Cette tâche revient donc à vérifier si la voix du deuxième échantillon est conforme à l'identité revendiquée. La comparaison est dite *target* si les deux échantillons examinés proviennent du même locuteur. Inversement, elle est dite *nontarget* quand les locuteurs sont différents.

Encart 2.1 : Le cas de l'expertise judiciaire

Dans le cas concret d'une expertise judiciaire, la comparaison considère une observation O faite sur les deux enregistrements X et Y (par exemple : la hauteur de voix moyenne diffère de 10 kHz entre les deux échantillons). Le score qui y est associé est appelé Rapport de Vraisemblance (LR) et se calcule comme suit :

$$LR = \frac{p(O|H_0)}{p(O|H_1)} \tag{2.1}$$

où H_0 représente l'hypothèse selon laquelle X et Y sont prononcés par le même locuteur et H_1 l'hypothèse selon laquelle les échantillons sont prononcés par deux locuteurs différents. Ce score correspond donc à la probabilité d'observer O à partir des deux enregistrements compte tenu des deux *a priori*. La décision est binaire, toutefois, elle est prise à partir du score qui sert d'indice de confiance.

Un système de vérification automatique considère un échantillon de voix qui est appelé « enregistrement de test ». Celui-ci est comparé à deux modèles alternatifs : le premier modèle correspond au locuteur dont l'identité est revendiquée et le deuxième correspond au modèle général, appelé modèle du monde. Ainsi, le score donné par le système mesure la probabilité que le locuteur revendiqué soit l'auteur de cet enregistrement par rapport à

la probabilité qu'il ne le soit pas.

2.1.2 Sources de variabilité du signal de parole

Une des plus grandes difficultés auxquelles sont confrontés les chercheurs en reconnaissance automatique du locuteur réside dans la multitude de signaux acoustiques observables dans la nature. D'une part, ces variabilités sont liées directement au canal de transmission utilisé pour transporter le signal de parole. Les plus fortes variations proviennent de la technologie utilisée (téléphone, radio), la qualité du micro, la distance, la qualité des enregistrements, la quantité de parole dans le signal et les algorithmes de compression. Ces variabilités sont en plus combinées aux possibles nuisances sonores comme le bruit de fond ou la réverbération. D'autre part, la parole représente en elle-même une source importante de variabilité. Par exemple, le contenu linguistique peut différer du fait que le locuteur ne dira pas toujours la même chose (il existe même des différences dans une phrase prononcée plusieurs fois par la même personne et dans le même contexte). En effet, la voix subit des altérations provenant d'un grand nombre de facteurs : humeur, santé, intention, etc. Un des objectifs principaux des chercheurs en reconnaissance du locuteur réside donc dans le développement de méthodes de compensation permettant de réduire une partie de ces variabilités (Hansen et al. 2015).

2.1.3 Évaluation de la tâche de vérification du locuteur

La tâche de vérification du locuteur est évaluée en fonction des erreurs commises sur les différentes comparaisons effectuées. Il y a typiquement deux types d'erreurs :

Faux Rejet (FR) : lorsque les deux enregistrements sont prononcés par le même locuteur et que la décision prétend qu'ils sont différents.

Fausse Alarme (FA) : lorsque les enregistrements proviennent de locuteurs différents alors que la décision dit qu'il s'agit du même locuteur.

La décision repose sur un seuil qui valide (ou rejette), en fonction du score obtenu sur une comparaison, l'hypothèse selon laquelle les deux enregistrements sont prononcés par le même locuteur (ou non). Ce seuil a donc une influence sur le nombre de FA et de FR. Il est possible d'observer l'influence du seuil sur ces pourcentages en traçant la courbe de Detection Error Tradeoff (DET). La méthode préférée pour l'évaluation de la tâche de vérification

du locuteur correspond à l'Equal Error Rate (EER). Cette méthode consiste à déterminer le seuil de décision tel que le pourcentage de FR soit égal au pourcentage de FA.

Le cadre d'évaluation NIST-SRE

Le National Institute of Standard and Technology (NIST) propose d'évaluer, au travers d'une campagne d'évaluation Speaker Recognition Evaluation (SRE), les systèmes de vérification du locuteur. Chaque année, les performances des systèmes sont donc comparées sur un corpus spécifique contenant des enregistrements provenant généralement de conversations téléphoniques.

2.2 Extraire l'information propre au locuteur

Les travaux de Kahn (Kahn et al. 2011) montrent que les systèmes de RAL atteignent des performances comparables à celles obtenues lors de différents tests de perception. Au-delà de ça, il est intéressant de constater que l'information captée par l'humain semble complémentaire à celle utilisée par la machine. Pourtant, les études réalisées en reconnaissance du locuteur avec des vrais jumeaux (Nolan et Oh 1996 ; San Segundo et Mompean 2017), nous suggèrent que les systèmes automatiques ont moins de réussite que les humains. Ce dernier est entraîné depuis sa naissance à percevoir des voix, ce faisant, il possède d'excellentes facultés pour y déceler une information spécifique lui permettant de faire la distinction entre deux voix très similaires. L'enjeu pour les systèmes automatiques réside dans leur capacité à capter l'information propre au locuteur et à détecter les éléments idiosyncratiques. En d'autres termes, la question est donc de savoir sur quelles caractéristiques du signal de parole s'appuyer pour reconnaître une personne de manière automatique. Dans un cadre idéal, les caractéristiques extraites doivent peu fluctuer entre les différents enregistrements d'un même locuteur et au contraire faire preuve d'une plus grande variabilité pour des locuteurs différents, comme l'ont souligné Wolf et Laver (Wolf 1972 ; Laver 1980).

De plus, les systèmes automatiques ont recours à des algorithmes nécessitant que les caractéristiques tirées de l'enregistrement audio soient représentées par des vecteurs de taille fixe. Le processus que nous illustrons dans

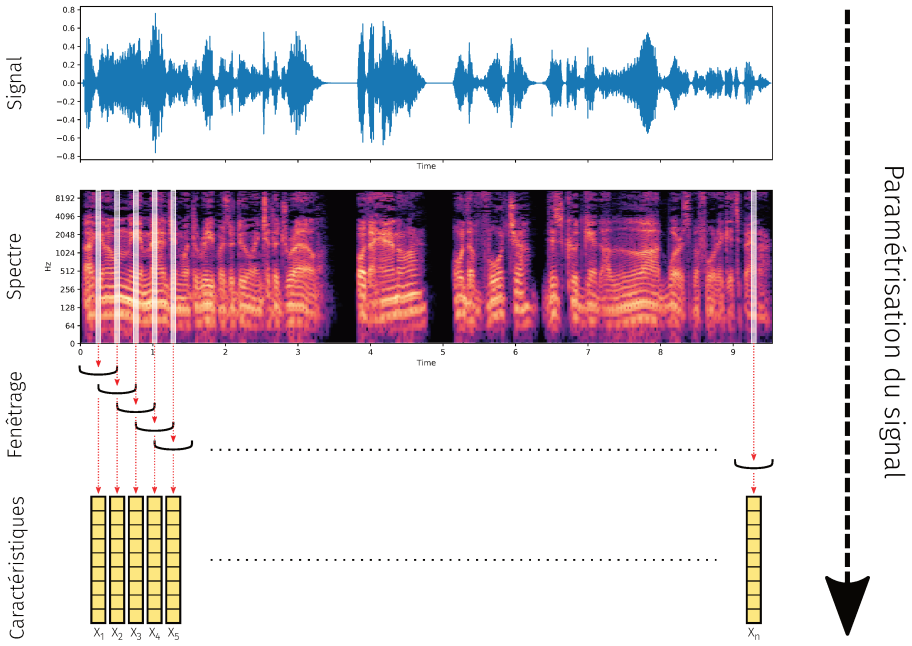

Figure 2.1 – Processus de paramétrisation du signal de parole.

la figure 2.1 est appelé « paramétrisation du signal » et permet de passer d'un signal dans un domaine temporel à une séquence de vecteurs de paramètres de dimensions fixées. Chaque vecteur de paramètres résume l'information acoustique présente dans la trame. Nous verrons dans la section 2.3 comment passer d'une séquence de taille variable à une représentation dans un espace de représentation de taille fixe.

Les systèmes automatiques s'appuient naturellement sur une analyse acoustique du signal de parole. Il existe plusieurs niveaux d'analyse du signal du parole. Généralement, la paramétrisation s'appuie sur le spectre d'amplitude calculé à partir de l'enregistrement (en ignorant la phase).

Analyse cepstrale

Le cepstre résulte d'une transformation d'un signal du domaine temporel vers un domaine analogue obtenu par transformée de Fourier appliquée au logarithme de la transformée de Fourier du signal. Pour rappeler l'analogie avec le spectre, on parle alors de cepstre et de quéfrence pour rappeler la fréquence. Il existe plusieurs paramètres issus de l'analyse cepstrale, mais les plus populaires sont les Mel Frequency Cepstral Coefficients (MFCC). En effet, ces derniers sont majoritairement utilisés en reconnaissance de la parole (Haton et al. 2006). Pour en savoir plus sur les différents paramètres acoustiques, se référer à J. P. Campbell 1997 ; Hansen et al. 2015.

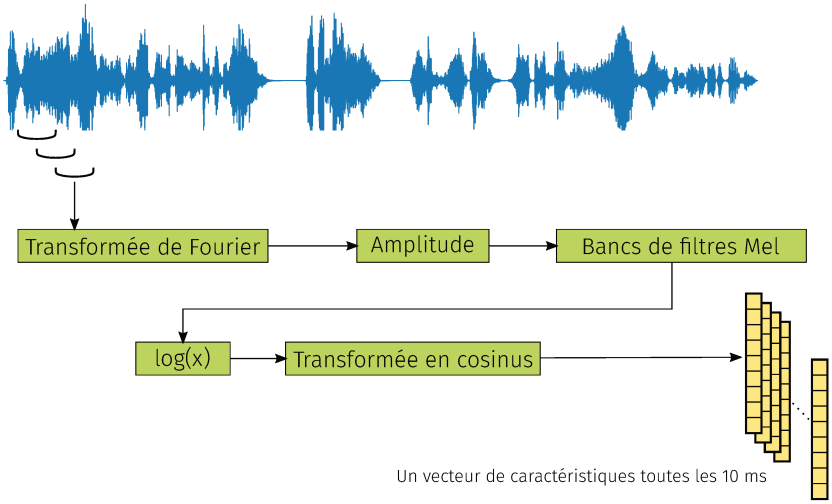

Figure 2.2 – Extraction des paramètres MFCC.

Dans la figure 2.2, nous schématisons le processus d'extraction des MFCC. Ces derniers ont la particularité de tenir compte de la perception des fréquences par l'oreille humaine qui suit une échelle non-linéaire, l'échelle Mel. Le banc de filtres Mel moyenne l'énergie spectrale pour chaque bande de fréquence. Par la suite, une transformée en cosinus discrète est appliquée. Elle permet 1) une compression de l'énergie dans un nombre réduit de coefficients ; 2) des coefficients hautement décorrélés. Cette transformation est intéressante, d'une part car elle réduit les dimensions, et d'autre part parce qu'elle supprime des composantes redondantes et donc nuisibles (Furui 1981).

En RAL, les paramètres sont extraits sur des courtes fenêtres d'une durée de 20–25 ms et en considérant une superposition de 10 ms. De plus, nombre de systèmes considèrent l'utilisation de l'information rendant compte de la variation de ces paramètres dans le temps. Cette information correspond aux dérivées première et seconde de chaque paramètre par rapport au temps (Δ, $\Delta\Delta$). Une technique alternative, appelée Shifted Delta Cepstra (SDC), initialement appliquée en reconnaissance de la langue, a aussi été proposée pour la vérification du locuteur (Kinnunen et al. 2006 ; Gonzalez et al. 2009).

Il est à noter que toutes les trames ne sont pas utilisées. En effet, la technique de Vocal Activity Detection (VAD) classifie chaque trame comme contenant de la parole ou du silence. Ainsi, les systèmes s'appuient principalement sur les trames contenant de la parole et ignorent simplement les autres. Enfin, il est possible d'effectuer une normalisation des paramètres extraits sur un enregistrement. Dans cette opération, le vecteur de para-

mètres moyen est soustrait à chaque vecteur de paramètres de sorte à centrer la moyenne sur 0. Cette technique s'appelle Cepstral Mean Normalization (CMN). En outre, il est possible de normaliser la variance à 1, il est alors question de Cepstral Mean Variance Normalization (CMVN).

Paramètres prosodiques

Alternativement aux paramètres cepstraux qui sont porteurs d'une information phonétique à court terme, nous distinguons d'autres paramètres qui caractérisent une information dite para-linguistique, telle que la prosodie. Plusieurs systèmes faisant état de l'utilisation de caractéristiques prosodiques ont été utilisés (N. Dehak, Dumouchel et al. 2007 ; Shriberg et al. 2008 ; Ferrer et al. 2010 ; Scheffer et al. 2011). Ces travaux ont montré que ces paramètres, lorsqu'ils sont combinés aux paramètres cepstraux, amènent une information complémentaire permettant d'améliorer la performance des systèmes.

Encart 2.2 : Le leurre de l'empreinte vocale

Historiquement, les premiers travaux réalisés sur des représentations spectrographiques tirés d'enregistrements vocaux ont, dans le cadre de la reconnaissance du locuteur, émergé après la seconde guerre mondiale dans les *Bell Telephone Laboratories* et notamment grâce aux travaux de Kersta 1962. Par analogie avec les empreintes digitales, Kersta clamait pouvoir comparer des enregistrements audio et parlait « d'empreintes vocales ». Nous savons maintenant que les enregistrements sont sujets à d'importantes variabilités et donc qu'une telle représentation n'est pas envisageable (Boë 2000). Aujourd'hui, ces comparaisons sont, d'une part, effectuées manuellement dans le cadre judiciaire par des phonéticiens experts, les caractéristiques qui sont utilisés variant en fonction des cas. D'autre part, ces comparaisons peuvent être faites automatiquement.

2.3 Modélisation du locuteur et comparaison

Nous avons dit plus haut que les paramètres utilisés doivent idéalement nous permettre d'observer une variabilité moindre entre différents enregistrements d'un même locuteur et inversement, une plus grande variabi-

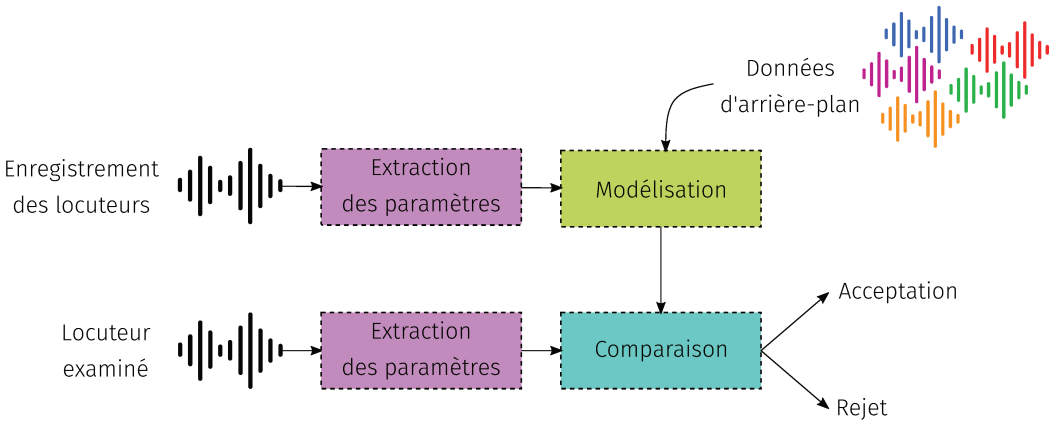

Figure 2.3 – Schématisation d'un système de vérification automatique du locuteur. La modélisation permet, entre autre, d'obtenir une représentation de taille fixe à partir de la séquence de paramètres de longueur variable.

lité quand il s'agit d'enregistrements provenant de locuteurs différents. Un grand nombre de facteurs sont susceptibles de faire varier le signal de parole, nous en avons énoncés quelques uns dans la sous-section 2.1.2. Ces dernières années, les performances des systèmes se sont nettement améliorées, notamment grâce à des techniques de modélisation du locuteur permettant de pallier une partie de ces variabilités. Deux approches globales différentes sont utilisées pour la modélisation : l'approche générative et l'approche discriminante.

Les approches discriminantes, cherchent à prédire la classe notée y à laquelle appartient une entrée x. Cela revient à calculer la probabilité $p(y|x)$ et à apprendre les frontières de décisions entre les différentes classes. Dans une approche générative il est avant tout question d'apprendre la distribution de probabilité associée à chaque classe. Classifier x revient donc à calculer la vraisemblance de chaque probabilité a priori $p(x|y)$ et à lui attribuer la classe la plus probable.

Dans la figure 2.3, nous présentons une vue synthétique d'un système de vérification du locuteur. Celui-ci contient, en plus de l'étape de paramétrisation, des phases de modélisation et de comparaison. Pour chaque locuteur enregistré dans le système, un modèle est crée à partir des séquences de paramètres extraites des enregistrements du locuteur. De plus un modèle acoustique générique est appris sur un grand nombre de locuteurs, appelé modèle du monde. La dernière étape consiste alors à calculer la probabilité que l'enregistrement de test appartienne au locuteur dont l'identité est revendiqué. Pour cela, l'échantillon est comparé à la fois au modèle du lo-

cuteur et au modèle du monde. Le score issu de cette comparaison permet d'accepter ou de rejeter l'identité du locuteur, en fonction d'un seuil prédéfini.

Nous décrivons plus bas l'étape de modélisation qui permet de passer d'une séquence de taille variable à une représentation de dimensions fixées. Cette partie n'a pas pour objet la réalisation d'un descriptif très détaillé de chacune des techniques utilisées pour la modélisation du locuteur, pour cela se référer à J. P. Campbell et al. 2009 ; Hansen et al. 2015. La section suivante présente un point introductif sur les principales techniques employées dans le domaine et insiste un peu plus sur celles que nous avons utilisées en particulier.

2.3.1 Méthodes fondées sur des GMM

Un Modèle de Mélange Gaussien (GMM) se définit comme une somme pondérée de fonctions de densité de probabilité gaussienne. Les GMM permettent de modéliser des données multivariées. La parole possède différents « états » correspondant aux différents sons qui sont produits. Un GMM peut donc être utilisé pour modéliser la distribution des séquences de vecteurs de paramètres issues d'un signal de parole. De plus, par leur nature probabiliste, les GMM rendent possible la modélisation d'une plus grande variabilité dans les données et amènent une plus grande robustesse (Reynolds et R. C. Rose 1995).

Il est possible de modéliser un vecteur aléatoire x_n par une somme pondérée de fonctions de densité de probabilité suivant :

$$p(x_n|\lambda) = \sum_{g=1}^{M} w_g \mathcal{N}(x_n|\mu_g,\Sigma_g), \tag{2.2}$$

avec M composantes $g = 1...M$, où $\mathcal{N}$ correspond à une distribution gaussienne de moyennes μ_g, de covariances Σ_g avec w_g la pondération associée à la composante g. Ainsi, la vraisemblance d'un vecteur de paramètres étant donné un GMM peut être calculée. Ici, $\lambda = \{w_g,\mu_g,\Sigma_g \mid g = 1...M\}$ fait référence au GMM. Pour une séquence de vecteurs de paramètres $\mathcal{X} = \{x_n \mid n \in 1...T\}$ la probabilité d'observer ces paramètres sachant λ est calculée selon

la formule :

$$p(\mathcal{X}|\lambda) = \prod_{n=1}^{T} p(x_n|\lambda), \tag{2.3}$$

en supposant que les vecteurs de paramètres de la séquence sont indépendants. L'aspect temporel de la séquence est neutralisé.

Généralement, un GMM est entraîné en utilisant l'algorithme Espérance-Maximisation (EM) qui augmente itérativement la vraisemblance des observations étant donnée le modèle.

Méthode UBM-GMM

La tâche de vérification du locuteur consiste à valider (ou rejeter) l'hypothèse selon laquelle un enregistrement est prononcé par un locuteur particulier. Cela revient à calculer la probabilité $p(\mathcal{X}|\lambda_{hyp})$ où $\mathcal{X}$ correspond aux paramètres de l'enregistrement, et λ_{hyp} le modèle du locuteur. L'hypothèse alternative peut être évaluée en calculant $p(\mathcal{X}|\lambda_{\overline{hyp}})$, avec $\lambda_{\overline{hyp}}$ au modèle correspondant à la distribution générale des paramètres extraits sur un grand ensemble de locuteurs. Ce modèle alternatif est appelé Modèle du Monde (Universal Background Model) (UBM) (Reynolds 1997 ; Reynolds, Quatieri et al. 2000 ; Bimbot et al. 2004).

Dans la littérature, il a été proposé de dériver le modèle propre au locuteur en mettant à jour simplement les paramètres de l'UBM, grâce à une technique d'adaptation Bayésienne (Gauvain et al. 1994) appelée Maximum A Posteriori (MAP). De cette manière, le modèle de locuteur est plus robuste que s'il était appris uniquement sur les données du locuteur. Nous illustrons dans la figure 2.4) l'adaptation d'un UBM à 4 composantes.

Compte tenu de la méthode GMM que nous avons décrite, le test de vraisemblance peut s'écrire en considérant l'observation notée O et un locuteur hypothétique noté ℓ. La tâche de vérification doit évaluer les hypothèses suivantes :

— H_0 : O provient du locuteur ℓ
— H_1 : O ne provient pas du locuteur ℓ

Dans l'approche GMM-UBM, les deux hypothèses sont représentées par le GMM du locuteur λ_{hyp} et l'UBM $\lambda_{\overline{hyp}}$. Ainsi, avec les paramètres acoustiques $\mathcal{X}$ calculés à partir de O, le test de vraisemblance sous sa forme logarithmique s'écrit :

$$LLR = log\, p(\mathcal{X}|\lambda_{hyp}) - log\, p(\mathcal{X}|\lambda_{\overline{hyp}})$$

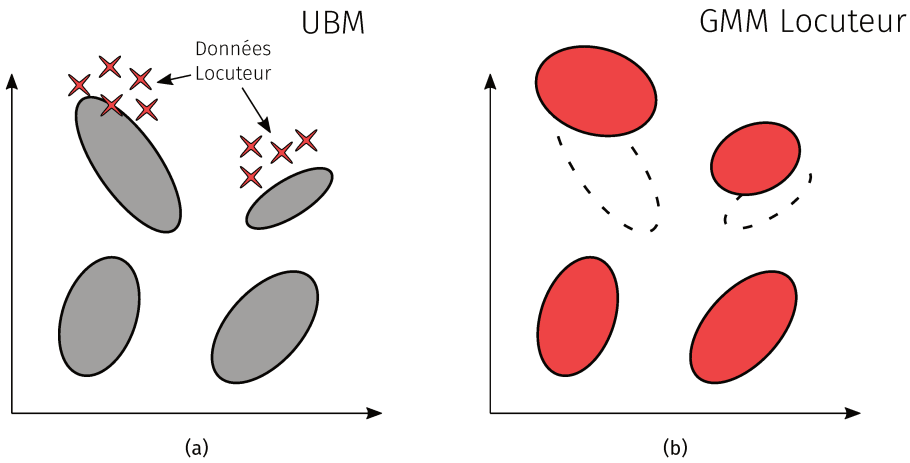

Figure 2.4 – Schématisation d'un GMM-UBM de 4 composantes et adaptation du modèle locuteur avec la procédure MAP.

Super-vecteur

La plupart des algorithmes de classification proposés dans la littérature nécessitent que les données en entrée soient de taille fixe. L'un des principaux problèmes en RAL étant la grande variabilité des enregistrements en termes de durée, il est essentiel de pouvoir représenter la séquence variable de vecteurs de paramètres par une représentation dans un espace de dimension fixe. Il est possible d'utiliser les paramètres d'un GMM (w_i, μ_i, Σ_i) pour obtenir une telle représentation. Plus précisément, il s'agit d'un vecteur concaténant les paramètres (typiquement les moyennes) mis-à-jour d'un GMM-UBM par une adaptation MAP pour un locuteur spécifique (Reynolds, Quatieri et al. 2000 ; Ben et al. 2003 ; P. Kenny, Mihoubi et al. 2003). Cette représentation est appelée « super-vecteur ». Ce dernier offre donc une solution aux chercheurs pour la réalisation de comparaisons au niveau de la session. Ce vecteur de grande dimensionnalité est particulièrement efficace quand il est utilisé conjointement avec un classifieur Machine à Vecteurs de Support (SVM), comme l'a montré W. M. Campbell et al. 2006, l'objectif étant de minimiser les taux de fausse alarme et de faux rejet durant l'apprentissage du SVM. De plus, le super-vecteur s'avère être à la base du développement de méthodes de normalisation comme Within-Class Covariance Normalization (WCCN), permettant la compensation de la variabilité intra-classe (Hatch et al. 2006).

Analyse Factorielle

Une des difficultés majeures en RAL réside dans la grande variabilité des enregistrements provenant notamment du canal de transmission, du locuteur, et de l'environnement (voir la sous-section 2.1.2). Kenny (P. Kenny et Dumouchel 2004) suppose que le signal de parole est constitué de deux composantes additives dans l'espace super-vecteur : une composante propre au locuteur et une composante de dimensionnalité réduite propre au canal. La solution proposée utilise l'Analyse Factorielle (FA). Par la suite, Kenny proposera un modèle appelé Joint Factor Analysis (JFA) :

$$s = m + Vy + Ux + Dz \tag{2.4}$$

où s représente le super-vecteur d'un locuteur donné, m le super-vecteur indépendant du locuteur (provenant de l'UBM), y les facteurs propres à la composante V dépendant du locuteur, x les facteurs de la composante U dépendant du canal, et enfin D la matrice résiduelle contrôlée par les facteurs résiduels notés z.

2.3.2 Espace de variabilité totale : approche *i*-vecteur

Selon Dehak (N. Dehak, R. Dehak et al. 2009), il est difficile d'isoler la variabilité propre au locuteur dans un sous espace du super-vecteur sans qu'une partie de l'information le concernant ne se retrouve dans la composante correspondant au canal. De plus, l'utilisation de cette approche requiert un corpus d'entraînement contenant plusieurs sessions par locuteur, ce qui est une contrainte forte relativement au gain obtenu (Matrouf et al. 2008). Dehak (N. Dehak, R. Dehak et al. 2009 ; N. Dehak, P. J. Kenny et al. 2010) propose de modéliser la variabilité avec une seule composante de faible dimensionnalité, appelée « espace de variabilité totale » et définie comme suit :

$$s = m + Tw \tag{2.5}$$

avec w correspondant au vecteur de dimension réduite [1] des facteurs de variabilité totale, chacun des facteurs contrôlant une dimension propre de la matrice de variabilité totale notée T. Cette dernière est apprise en utilisant la même méthode que pour apprendre la matrice V dite *Eigenvoice*, à

1. La matrice T a un rang fixé entre 200 et 800, ce qui est considérablement inférieur à la dimension de l'espace super-vecteur (30720 avec des vecteurs de 60 paramètres et un GMM de 512 composantes)

l'exception que chaque enregistrement est supposé provenir d'un locuteur différent.

Ces facteurs sont par la suite utilisés en tant que vecteur de représentation (par exemple en entrée d'un classifieur) appelé *i*-vecteur. Ils forment ainsi une représentation compacte résumant l'information contenue dans le super-vecteur.

À partir de cet espace de faible dimensionnalité, il est possible d'appliquer des techniques d'apprentissage automatique pour d'une part compenser la variabilité intra-classe (liée au canal), d'autre part maximiser la variabilité inter-classe (liée au locuteur). Les techniques Analyse Discriminante Linéaire (LDA) et WCCN (Hatch et al. 2006) font partie des plus utilisées pour la compensation inter-session.

Méthodes de comparaison

Le développement de l'espace *i*-vecteur a rendu beaucoup plus efficace les méthodes de projection par LDA et WCCN du fait qu'elles n'opèrent plus sur des vecteurs de très grande dimension. La taille réduite et la densité d'information de cet espace rendent également possible l'utilisation de méthodes de comparaison simples qui s'appuient sur une similarité cosinus. D'autres techniques qui nécessitent une phase d'entraînement (pour l'apprentissage de la matrice de covariance inter-session), comme la distance de Mahalanobis, sont également utilisées.

De plus, Kenny (P. Kenny 2010) a proposé l'utilisation d'une Analyse Discriminante Linéaire Probabiliste (PLDA), initialement introduite pour la compensation de la variabilité inter-session en reconnaissance faciale (Prince et al. 2007), qui suit une modélisation similaire à la JFA. En effet, la PLDA opère directement sur les *i*-vecteurs qui contiennent à la fois la variabilité dépendant du locuteur et dépendant de la session. Une fois que les paramètres de la PLDA sont estimés, seuls deux *i*-vecteurs sont nécessaires pour la phase de comparaison. Le score obtenu correspond au rapport de vraisemblance associé aux deux *i*-vecteurs compte tenu des hypothèses H_0 et H_1.

Techniques de normalisation

Différentes méthodes de normalisation ont été proposées dans le but d'améliorer les performances des systèmes. Parmi elles, l'Eigen Factor Radial (EFR) (Bousquet, Larcher et al. 2012) ou la *Length Normalization* (Bousquet, Matrouf et al. 2011) qui permettent notamment d'obtenir une distribution gaussienne des données et d'utiliser la version gaussienne de la PLDA moins coûteuse que son homologue dite *heavy-tailed* (Garcia-Romero et al. 2011).

2.3.3 Méthodes fondées sur l'apprentissage profond

Les progrès réalisés ces dernières années en RAL ont amené à des modèles de plus en plus robuste, notamment au niveau des segments de courte durée (Bhattacharya et al. 2017). Ces améliorations sont reliées à l'utilisation des approches par apprentissage profond. Ces dernières ont (comme dans nombre d'autres domaines) été largement adoptées pour la reconnaissance automatique du locuteur (P. Kenny, Stafylakis et al. 2014; Mitchell McLaren et al. 2015).

Le recours à des approches discriminantes fondées sur des Réseaux de Neurones Profonds (Deep Neural Networks) (DNNs) permettent d'apprendre une représentation abstraite des données d'entrées dans les couches profondes (Bengio, Courville et al. 2013). Ces techniques ont notamment permis l'extraction de caractéristiques de haut niveau, plus connues en tant que Bottleneck Features (BNF) (voir figure 2.5), dans un espace de dimensionnalité réduite (Yaman et al. 2012; Yamada et al. 2013). Les BNF peuvent alors être utilisées pour entraîner un système i-vecteur/PLDA.

L'information contenu dans les BNF est caractéristique de la tâche réalisée pour les apprendre. Dans un contexte d'« apprentissage par transfert », il est possible d'optimiser le réseau sur une tâche de reconnaissance de la parole (par exemple en ayant en sortie les différentes unités phonétiques). De plus, l'information des BNF peut être combinée à l'information spectrale. Il est ici question d'approches dites « tandem » qui associent des caractéristiques contenant des informations de différents niveaux.

De plus, l'utilisation de l'apprentissage par transfert a permis des progrès en matière de robustesse au bruit grâce à l'influence d'un DNN appris sur une tâche de reconnaissance automatique de la parole (Lei et al. 2014).

Dernièrement, ce sont les approches dites *end-to-end* qui ont suscité le

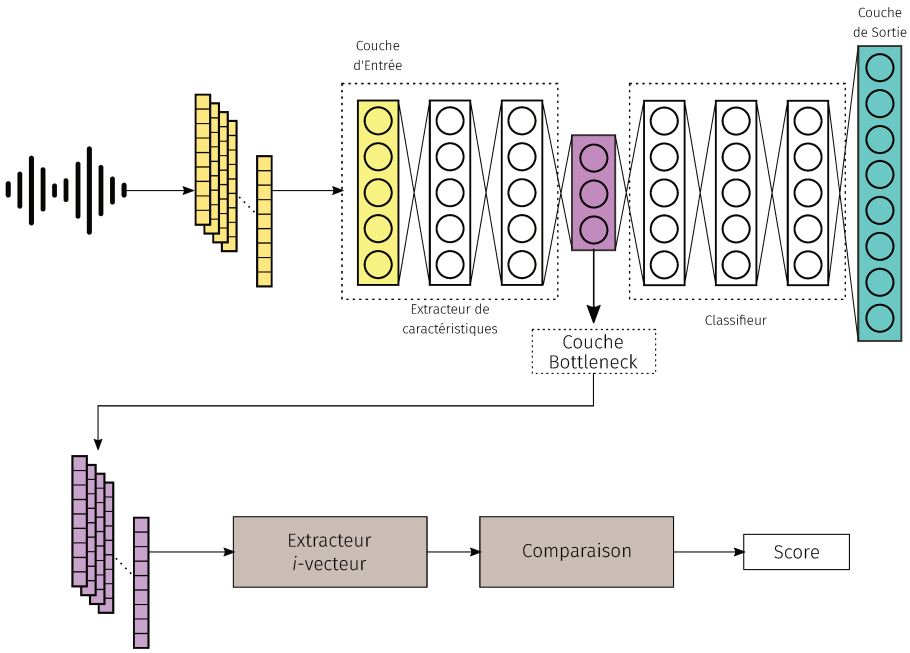

Figure 2.5 – L'apprentissage des *Bottleneck Features* est guidé par une tâche de discrimination.

plus d'attention (Variani et al. 2014; Snyder, Ghahremani et al. 2016; C. Li et al. 2017) et qui tirent profit d'un large volume de données pour apprendre une représentation dédiée à la tâche effectuée. En plus, de l'avantage évident lié au remplacement de toute la chaîne d'apprentissage par un seul processus d'apprentissage, les résultats obtenus avec ces méthodes montrent une plus grande robustesse aux tests réalisés sur des segments de plus courte durée.

Enfin, les systèmes fondées sur des réseaux de neurones profonds permettent d'apprendre une représentation dédiée au locuteur aussi appelée *speaker embedding*, notamment grâce à l'approche *x*-vecteur (Snyder, Garcia-Romero, G. Sell et al. 2018), état-de-l'art actuel.

Approche *x*-vecteur

L'approche *x*-vecteur consiste à apprendre un espace de représentation à l'aide d'un DNN dont le but est de discriminer un grand nombre de locuteurs. Cette représentation de taille fixe est apprise à partir d'enregistrements de taille variable. Chaque enregistrement correspond à une séquence composée d'une suite de vecteurs de paramètres calculés au niveau de la trame. Le passage d'une séquence de taille variable à une représentation de taille fixe est permise grâce à une technique appelée *statistical pooling*. De plus, l'espace de représentation qui est appris prend en compte une infor-

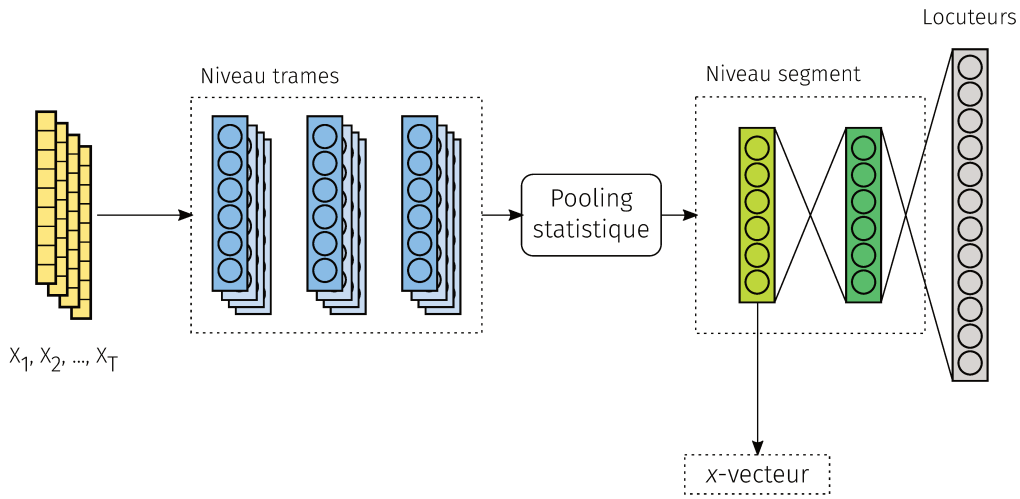

Figure 2.6 – Architecture *x*-vecteur.

mation à plus long terme, du fait du l'utilisation d'un Time Delay Neural Network (TDNN) dont chaque neurone reçoit une information contextuelle des neurones de la couche précédente.

Dans la figure 2.6 nous décrivons l'architecture *x*-vecteur proposée par Snyder. Deux parties du réseau permettent un traitement à différents niveaux de l'information. La première partie traite l'information au niveau des trames et la deuxième opère à plus long terme. Un grand nombre de trames d'un même locuteur sont agrégées par la fonction de *pooling*. Chaque *mini-batch* utilisé pour l'entraînement du réseau prend en compte un grand nombre de trames de différents locuteurs. Le réseau est optimisé pour discriminer au mieux chaque locuteur.

Enfin, l'entraînement d'un système *x*-vecteur est réalisé sur un gros volume de données qu'il est possible d'augmenter par ajout de bruit et de réverbération. L'avantage de cette approche, par rapport à l'approche *i*-vecteur, vient notamment de sa robustesse sur les segments de courte durée.

2.4 Conclusion

Ces dernières décennies ont permis une amélioration des modèles utilisés en RAL. Cela est dû notamment au cadre d'évaluation de cette tâche qui permet à chaque équipe de recherche d'avancer indépendamment tout en leur permettant de comparer leurs méthodes sur des corpus différents lors des campagnes d'évaluation annuelles.

Dans un grand nombre de cas, les systèmes automatiques sont plus per-

formants qu'un expert humain. Toutefois, dans certaines situations, l'oreille humaine conserve un net avantage sur la machine. En effet, l'humain semble capable de compenser un signal de très mauvaise qualité et de se fier à des caractéristiques para-linguistiques de haut niveau (N. Dehak, Dumouchel et al. 2007).

En ce qui concerne la compensation des nuisances (cf. sous-section 2.3.2), les méthodes proposées sont de plus en plus efficaces. Toutefois, la recherche d'une représentation mettant en avant l'identité du locuteur et éliminant les possibles nuisances reste une question ouverte. En effet, la capacité des algorithmes utilisés n'est pas adaptée à l'immense variabilité des signaux de parole. Les techniques récentes fondées sur des réseaux de neurones profonds et reposant sur de larges volumes de données permettent de caractériser une information propre au locuteur à partir d'une information extraite à court terme dans le signal. En tout cas, c'est ce que nous suggèrent les dernières avancées mettant en œuvre des approches *end-to-end*. Cependant, les systèmes fondés sur des architectures profondes ont un inconvenient : la complexité des modèles appris les rend difficiles à analyser.

En définitive, un locuteur peut être reconnu de manière automatique en se concentrant généralement sur les éléments idiosyncratiques de la parole. Ces informations correspondent en majeure partie aux variations à court terme de la prononciation des différents phonèmes de chaque locuteur. Nous pouvons donc supposer que ces mêmes variations se retrouvent chez les locuteurs ayant des voix similaires, mais nous en reparlerons plus en détails dans le chapitre 5.

Chapitre 3

Percevoir l'état émotionnel et les traits de personnalité

Sommaire

La voix est riche en information, elle véhicule notamment de l'information qui renseigne sur l'état psychologique du locuteur. Les différents ressentis émotionnels du locuteur impactent la façon dont est produite la parole et peuvent se transmettre par la voix. Nous nous intéressons ici à cette information et comment elle est perçue. L'état émotionnel du locuteur peut être perçu grâce à des éléments contenus dans le signal de parole. Une des problématiques principales de la reconnaissance des émotions concerne

la caractérisation des éléments contenus dans le signal de parole porteurs d'information sur l'état émotionnel du locuteur. Il est possible que ces éléments d'information soient localisés à un endroit précis du signal de parole. Quand la voix déraille ou vient buter sur un mot, dans une situation de stress par exemple. L'information peut également être globale. Le débit de parole par exemple se mesure sur la totalité du signal et peut renseigner sur l'état émotionnel du locuteur. Les informations locales ou « à court terme » sont calculées sur des courtes durées à l'inverse de l'information « à long terme » ou globale. Nous faisons donc un état-des-lieux succinct du domaine et de ses avancées. Nous verrons comment poser un cadre d'évaluation basé sur la perception de la voix, en termes d'états affectifs du locuteur, mais aussi sur la perception des traits de personnalité.

3.1 Les émotions : perspective psychologique

Selon Scherer (K. R. Scherer 2003a), l'étude des émotions dans la littérature remonte à l'antiquité grecque, mais a pendant longtemps manqué de techniques valides appropriées à la mesure des phénomènes de son expression. En cause, la difficulté de réaliser une étude empirique des processus émotionnels, contrairement à d'autres processus cognitifs, plus simples à appréhender de manière rationnelle. La capture des émotions en laboratoire s'avère peu commode et leur déclenchement volontaire soulève aussi des problèmes éthiques, rendant ces études d'autant plus délicates. Hors laboratoire, il est difficile d'observer les réactions émotionnelles, ce qui rend toute étude de terrain impossible à réaliser. Cela est dû au fait que dans un grand nombre de sociétés, l'exercice d'un contrôle affectif permanent prévaut au libre cours de l'expression des émotions.

La définition de l'émotion en tant que concept psychologique reste une question ouverte. Selon Fontaine et al. 2007, seul un consensus partiel autour d'une définition temporaire fait autorité et malgré cela, les chercheurs se demandent encore s'il s'agit d'un processus psychologique, de la description d'un état ou bien d'une excitation physiologique. Selon le consensus en vigueur, l'émotion se place à un niveau supérieur aux concepts de motivation et de cognition. Le concept d'émotion se compose de plusieurs parties : une activation neurophysiologique, une expression motrice, une composante d'ébauche d'action et de préparation du comportement, enfin une composante subjective correspondant à l'état émotionnel.

Les principales fonctions attribuées aux émotions sont d'ordre adaptatif et motivationnel. Les aspects les plus importants du mécanisme émotionnel sont : l'évaluation de la situation par rapport aux besoins, projets et préférences de l'organisme, la préparation physiologique et psychologique aux actions en réponse aux stimuli, enfin la communication dans l'environnement social. Les organismes vivant en société sont caractérisés par une émotivité intense et complexe. Selon Darwin, la coexistence dans une société est rendue possible grâce aux émotions. En effet, les émotions de part leur qualité expressive communiquent la réaction et l'intention d'action aux autres individus (Darwin et al. 1872). Pour justifier l'apparition d'un tel système, K. R. Scherer 1984 fait le lien entre l'augmentation des capacités de traitement d'information des organismes et la variabilité croissante des comportements. Par conséquent, il met en lumière la nécessité d'un mécanisme capable d'adapter correctement ces organismes aux différentes stimulations. Le système émotionnel répond à ce besoin et permet plus encore puisqu'il constitue un important système de signaux intra-organiques permettant l'apprentissage.

Le problème de délimitation du concept d'émotion n'est pas de définir ce qu'est une émotion, mais plutôt d'en déterminer les différentes sortes. Dans la littérature, qu'importe le domaine (*e.g.* science cognitives, neuroscience, informatique) les termes émotions, sentiments, affects sont fréquemment utilisés et cela, indépendamment de leur sens propre. L'affect est un processus non-conscient qui précède les émotions et les sentiments. Ces derniers sont des phénomènes conscients. Quant aux émotions, elles sont l'expression de l'affect (Shouse 2005). Dans Picard 2000, l'auteur propose d'utiliser de manière interchangeable les mots « émotionnel » et « affectif » comme adjectifs décrivant à la fois les composantes physiologique et cognitive des émotions. Cependant elle attribue un sens plus large au mot « affectif ». Aussi, elle propose d'utiliser les termes « sentiment » et « sensation » de manière interchangeable avec les adjectifs émotionnel et affectif. En ce qui nous concerne, nous utilisons ces termes de manière équivalente pour faire référence aux émotions perçues à travers des stimuli audio. À ce propos, une des grandes questions qui animent la communauté est de savoir si les émotions sont perçues ou induites.

Encart 3.1 : Émotions perçues et émotions induites

Les deux sont probablement vraies et dépendent de la situation. Prenons par exemple le contexte d'une audition pour un concours, le jury à l'écoute pourra alors percevoir l'état émotionnel (très probablement du stress) dans lequel se trouve l'individu. À l'inverse, en se plaçant du point de vue du spectateur à l'écoute du *Concerto pour piano n° 2* de Rachmaninov par exemple, la musique est alors susceptible d'induire chez celui qui écoute quelques émotions. Que l'émotion soit perçue ou induite, dans les deux cas, il s'agit d'une rencontre. Soit entre la personne qui perçoit et celle qui ressent l'émotion, soit entre le spectateur et l'objet esthétique.

3.1.1 Les modèles théoriques des émotions

Dans la littérature, il existe trois grandes conceptualisations différentes des émotions. En premier, sont apparues les versions discrètes ou catégorielles qui proposent de distinguer les émotions au moyen d'adjectifs variant selon les auteurs. Deuxièmement, la théorie dite dimensionnelles qui suppose que les émotions s'inscrivent dans un espace multidimensionnel. Enfin, une théorie plus en marge des deux autres supposée refléter le mécanisme d'évaluation.

Théorie discrète des émotions

Les nombreux termes associés aux émotions ont été listés et regroupés en différentes catégories. Par exemple dans le modèle catégoriel des émotions proposé par Hevner 1936 les adjectifs sont regroupés en huit groupes de manière à différencier des macro-classes émotionnelles. Les termes présents au sein d'une même classe ne font pas consensus. Au contraire, chaque terme est adapté à un contexte particulier. De plus, ce modèle est destiné aux émotions perçues dans la musique et n'a pas vocation à décrire les émotions de manière générale. Pars la suite, le modèle catégoriel a été étendu, d'autres termes ont été ajoutés, d'autres ont été enlevés. Le problème vient de la grande richesse des émotions perceptibles au travers du son, qu'il s'agisse de musique ou voix. De plus, au fur et à mesure que le nombre de termes augmente, l'attribution d'une étiquette soulève de plus en plus d'ambiguïtés. Pour éviter ce biais, les chercheurs réalisent leurs expérimentations sur un panel d'émotions très réduit. K. R. Scherer, Banse et Wall-

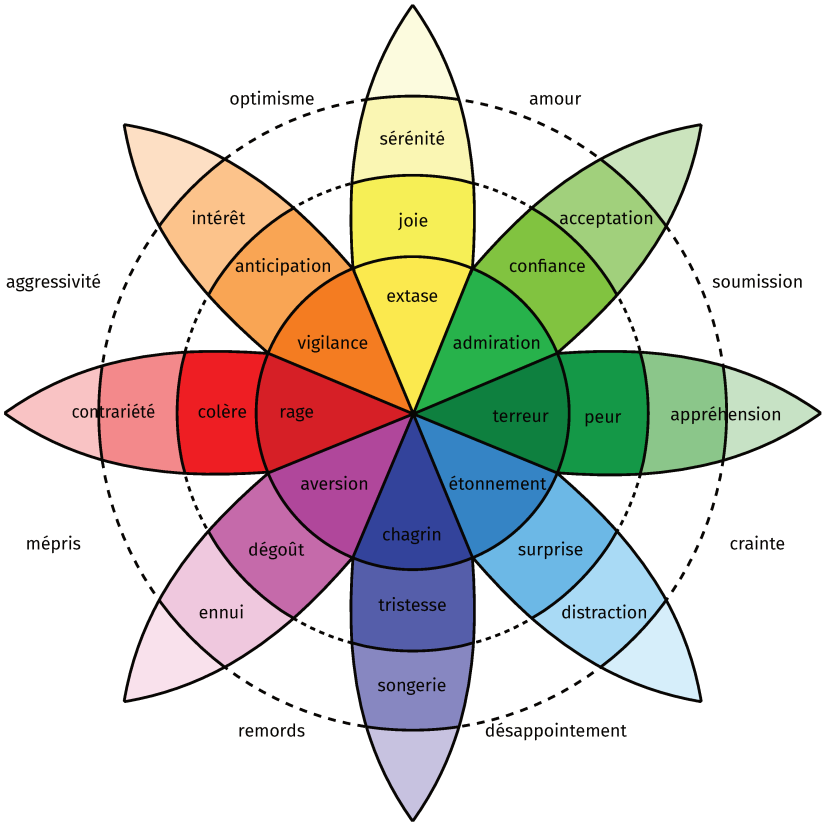

Figure 3.1 – Modèle tridimensionnel des émotions (Plutchik 1984).

bott 2001 propose d'inférer à partir d'enregistrements vocaux les émotions de joie, surprise, tristesse, peur, colère et dégoût. Ces émotions sont, comme beaucoup de théories semblent s'y accorder, considérées comme les émotions basiques et universelles (Ekman 1999). Selon Plutchik (Plutchik 1984), les émotions secondaires, plus complexes, résultent de la combinaison de ces émotions primaires (voir la figure 3.1). Cela donne lieu à une organisation par opposition des émotions primaires (qui sont au nombre de 8 chez Plutchik) et de leurs variations, représentées par différentes teintes de couleurs.

Théorie dimensionnelle des émotions

Les émotions peuvent aussi être décrites dans un espace continu à trois dimensions (Schimmack et al. 2000). La première composante affective définit de manière continue les émotions selon un axe opposant l'envie et l'aversion (les émotions positives contre les émotions négatives), c'est à dire la *valence*. La deuxième dimension oppose l'excitation et l'apaisement et dé-

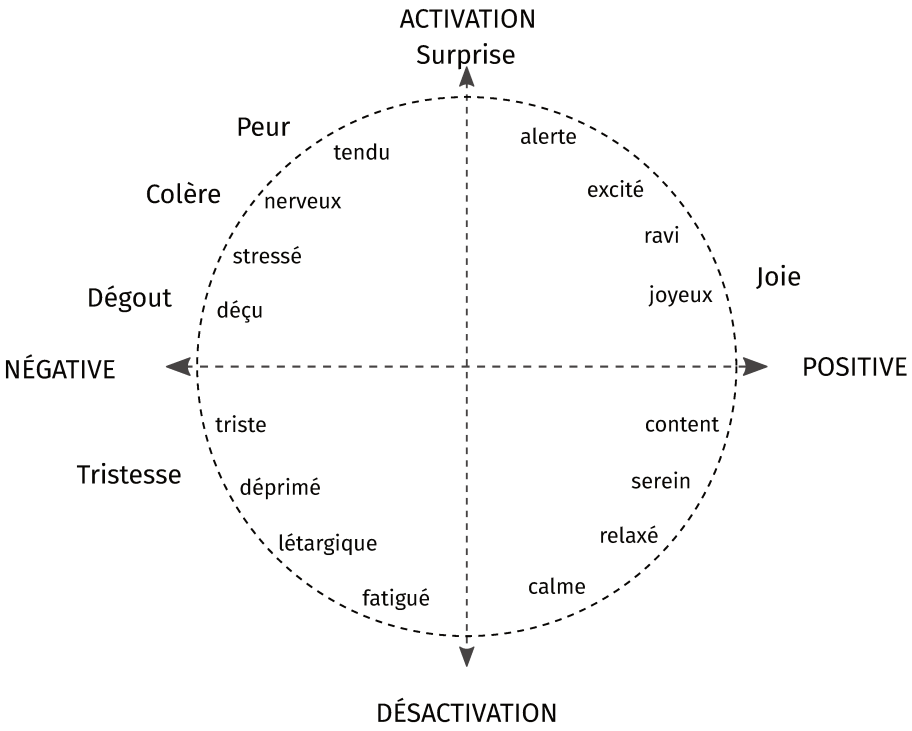

Figure 3.2 – Modèle circomplexe des émotions (Russell 1980).

fini l'*activation* générée par l'émotion. Le troisième axe décrit quant à lui la *tension* que suscite l'émotion. Cette dernière dimension étant sujette à controverse, il est plus fréquent d'utiliser ce modèle dans sa version bi-dimensionnelle, dit modèle Valence-Activation (VA), initialement proposé par Russell 1980 et que nous illustrons dans la figure 3.2. Dans le modèle circomplexe des émotions, les adjectifs correspondants aux émotions sont répartis selon ces axes le long d'un cercle.

Théorie de l'évaluation

Selon K. R. Scherer 2009, les émotions résultent d'un processus d'évaluation cognitif d'un évènement par rapport aux attentes et aux objectifs propres à l'individu, qui prépare à l'action et est accompagné d'un ressentiment subjectif. Nous ne décrivons pas plus en détail ce modèle, car bien que très complet, du fait qu'il décrive bien les aspects dynamiques de la génération des émotions, il rend difficile l'annotation de la perception des émotions.

3.1.2 Expression des émotions au travers de la voix

L'expression émotionnelle dans le cadre d'une communication par le biais de la parole ainsi que son impact sur l'auditeur sont étudiés depuis la Grèce Antique. Cependant, les études empiriques sur leurs effets ont réellement débuté au début du 20ième siècle pour finalement aboutir à un modèle spécifique de la communication des émotions par la parole (K. R. Scherer 2003a). Ce modèle est une adaptation du modèle de lentille proposé par Brunswik (Brunswik 1956). En effet, l'étude des états internes du locuteur au travers de la voix peut se faire à partir d'un modèle de codage et décodage des émotions. L'état émotionnel du locuteur altère certains attributs physiologiques et induit des modifications dans la voix (*e.g.* respiration, phonation, articulation). Ces modifications sont encodées par des indices acoustiques. L'observateur décode l'état émotionnel du locuteur par le biais des indicateurs qu'il perçoit. La transmission peut être sujette à des éléments perturbateurs pouvant nuire à l'inférence des émotions perçues par l'observateur par rapport aux émotions ressenties par le locuteur. Ces perturbations peuvent provenir de facteurs extérieurs comme le bruit environnant, la distance ou de spécificités (physiques, sociales, culturelles) propres aux interlocuteurs.

3.1.3 Discussion

Les psychologues ont mis au point différentes théories pour modéliser les émotions. La théorie de l'évaluation proposée par Scherer est trop complexe pour être utilisée dans le cadre de l'automatisation de la reconnaissance d'émotions. Les modèles catégoriels sont beaucoup plus simples, mais décrivent les émotions de manière très limitée. Les chercheurs en reconnaissance des émotions se sont concentrés sur l'utilisation du modèle bidimensionnel. Qu'il s'agisse d'émotions véhiculées par la musique ou par la voix, ce modèle semble les décrire avec suffisamment de précision tout en restant suffisamment simple pour mener à bien les étapes d'annotations, cruciales dans ce domaine.

Encart 3.2 : Évaluation du jugement

L'évaluation d'une émotion est réalisée par un observateur. Il est communément admis que l'émotion perçue par l'observateur représente une bonne approximation de l'émotion ressentie par le locuteur (Busso et Narayanan 2008). Par conséquent les bases de données utilisées contiennent des segments de voix (il peut s'agit de parole spontanée ou contrôlée) annotées par différents observateurs, selon les axes VA. Toutefois, des études ont révélé un écart entre les évaluations faites par des observateurs naïfs en comparaison de celles réalisées par le locuteur lui-même (Busso et Narayanan 2008 ; Truong et al. 2012). En effet, ce dernier a tendance à être plus sélectif dans ses évaluations tandis que les observateurs sont plus modérés. Pour expliquer cette différence, il n'est pas aberrant de supposer que la transmission provoque une altération de l'émotion perçue. Par conséquent, le locuteur ne serait pas le mieux placé pour juger des émotions qu'il ressent lui-même. Pour donner plus de valeur à ces travaux, il serait intéressant de faire une étude comparative de la perception des indices acoustiques *proximal* et *distal* (selon le modèle de communication des émotions proposé par K. R. Scherer 2003a). Selon les applications, il serait alors intéressant de faire varier le point de vue de l'évaluation des émotions.

3.2 Reconnaissance automatique des émotions

Un système de reconnaissance automatique des émotions se décompose en deux étapes. La première consiste à extraire, depuis le signal acoustique, les caractéristiques appropriées. La deuxième étape consiste a apprendre de manière supervisée la combinaison de ces caractéristiques pour reconnaître l'émotion véhiculée par le segment audio. Ce modèle peut alors être utilisé pour prédire l'état affectif du locuteur à partir d'autres segments de voix. Pour être entraîné, le système a besoin de données annotées en termes d'états affectifs. Selon que l'annotation soit numérique ou bien qualitative (qu'elle suive respectivement le modèle dimensionnel VA ou le modèle catégoriel), une régression ou une classification sera réalisée respectivement.

3.2.1 Les descripteurs acoustiques

La littérature fait état d'un grand nombre de caractéristiques utilisées, toutes varient dans leur pouvoir de discrimination des émotions. Nous n'avons

pas pour objectif de survoler toutes les caractéristiques utilisées durant les décennies de recherches dédiées à ce sujet, pour cela se référer à Petta et al. 2011. Nous évoquerons simplement les principaux descripteurs utilisés aujourd'hui et les différents types de caractéristique sur lesquelles reposent les systèmes de reconnaissance automatique des émotions.

Les MFCC sont parmi (voir la section 2.2 pour plus de détails) les plus utilisés. Ces derniers sont généralement extraits sur des fenêtres de 20–30 ms. Ce sont des caractéristiques à court terme et qui par conséquent sont qualifiées de locales, à la différence d'autres caractéristiques dites globales calculées à plus long terme, voire sur le signal complet. En général, les caractéristiques globales amènent de meilleures performances en classification des émotions que les caractéristiques locales (El Ayadi, Mohamed S. Kamel et al. 2011).

Les caractéristiques prosodiques qui font référence à la musicalité de la parole sont aussi largement étudiées dans le domaine et se révèlent bien corrélées à la communication des émotions par la voix. La fréquence fondamentale $F0$ se trouve être corrélée à la dimension *activation* du modèle bidimensionnel (Gunes et al. 2011). Le timbre ou la qualité de voix peut être mesuré à l'aide de caractéristiques spectrales (Schuller, Batliner et al. 2011). Les premiers formants, jitter, shimmer ainsi que le Rapport Signal sur Bruit (Harmonic to Noise Ratio) (HNR) décrivent la qualité vocale. Cependant leur rôle dans la perception des émotions est mal défini et en ce qui concerne leur interprétation en termes de descripteurs de qualité de voix (soufflée, sèche, tendue, etc.), il n'y a pas de réel accord établi dans la communauté (El Ayadi, Mohamed S. Kamel et al. 2011).

3.2.2 Modélisation

Approches classiques

Il n'existe pas de classificateur dédié à la tâche de reconnaissance des émotions. Les chercheurs font usage de différentes méthodes de classification en prenant en compte les expériences déjà réalisées (Koolagudi et al. 2012; Anagnostopoulos et al. 2015). Les chercheurs en reconnaissance automatique des émotions s'appuient sur des approches statistiques. Par exemple, dans Neiberg et al. 2006, les auteurs proposent d'utiliser un GMM pour modéliser la variabilité des paramètres calculés au niveau de la trame permettant de distinguer les différentes émotions. D'autres approches pro-

posent de prendre en compte les aspects dynamiques de la parole pour mieux décrire les émotions. Par exemple, Nwe et al. 2003 mettent au point un système s'appuyant sur un Modèle de Markov Caché (Hidden Markov Model) (HMM).

Selon El Ayadi, Mohamed S. Kamel et al. 2011 et Schuller, Batliner et al. 2011, les HMM ont été les plus utilisés, du fait de leur grande popularité dans le domaine du traitement de la parole. La littérature fait état de l'utilisation d'approches fondées sur un apprentissage discriminant comme les SVM (Seehapoch et al. 2013 ; Fu et al. 2008). Ces derniers offrent des performances surpassant les approches GMM et HMM. Les SVM permettent de réaliser une classification à partir d'une transformation non-linéaire des caractéristiques en entrées dans un espace de plus grande dimension. Cependant, la séparabilité des caractéristiques n'est pas assurée. D'autant plus que la sélection de ces caractéristiques semble être une problématique en elle-même (Luengo et al. 2010). En définitive, chaque méthode a ses avantages, il semble donc logique que des approches combinant ces différentes techniques aient été proposées (Lugger et al. 2009).

Comme dans beaucoup d'autres domaines, les réseaux de neurones ont été largement adoptés pour la reconnaissance des émotions, en complément des méthodes classiques. Aujourd'hui, la littérature propose beaucoup d'architectures de type DNN pour cette tâche. Dans Anand et al. 2015, les auteurs proposent d'extraire des caractéristiques de haut niveau à l'aide d'un Réseau de Neurones Convolutifs (CNN). En effet, ces derniers sont généralement utilisés pour l'extraction de caractéristiques depuis le signal de parole brut qui semble montrer de meilleur résultats (Huang et al. 2019). L'étape de reconnaissance des émotions repose en règle générale sur un Réseau de Neurones Récurrents (RNN). Par exemple, les travaux présentés dans Mirsamadi et al. 2017 reposent sur une architecture récurrente, plus précisément un Long Short-Term Memory (LSTM) combiné à un mécanisme d'attention. Ce dernier permet de se concentrer sur des parties spécifiques du signal.

De manière générale, les DNN sont plus efficaces pour apprendre des transformation non-linéaires. Cependant, leur performance dépend fortement du choix des hyper-paramètres (activation, nombre de couches, nombre d'unités, régularisation, etc.).

Approches end-to-end

Les avancées en matière de réseaux de neurones ces dernières années nous ont conduit vers les approches *end-to-end*. En reconnaissance des émotions, les caractéristiques tirées du signal doivent idéalement être robustes pour capter le contenu émotionnel depuis une grande variété de signaux. L'avantage de l'approche *end-to-end* réside dans l'apprentissage automatique de la meilleure représentation possible pour la tâche, depuis le signal de parole en lui-même (Trigeorgis et al. 2016 ; Tzirakis et al. 2018 ; Sarma et al. 2018). Toutefois, cette approche à un inconvénient majeur. Elle nécessite pour l'apprentissage un volume de données bien plus important pour un gain en performance pas forcément significatif.

3.3 Perception des traits de personnalité

Nous avons vu comment sont perçues les émotions et comment en extraire les informations qui y sont reliées à partir du signal de parole. Ces informations dépeignent un état temporaire du locuteur. Nous souhaitons nous intéresser dans cette section à un état du locuteur qui, contrairement aux états affectifs, n'est pas en constante évolution. Il s'agit de son caractère, sa personnalité.

La question de la caractérisation de le personnalité trouve une réponse dans la psychologie. Selon Matthews et al. 2009, la personnalité est une construction qui englobe des caractéristiques stables d'un individu (des traits), qui plus est des attributs quantifiables, de manière à pouvoir prédire son comportement. La compréhension de la personnalité et notamment la perception des traits de son expression est pertinente pour tout ce qui touche à la prédiction du comportement humain mais aussi pour la génération de parole. Les premières approches visant à modéliser cette dimension avaient pour objectif de l'intégrer dans les Interactions Homme-Machine (IHMs) (Nass et al. 2001). Selon Kaplan et al. 2010, ce domaine a par la suite gagné en intérêt en partie grâce à la masse d'information à caractère personnel disponible au travers des média sociaux et à l'utilisation croissante des appareils de type smartphone (Raento et al. 2009). Trois principales problématiques sont présentent dans le domaine. La reconnaissance automatique de la personnalité propre à un individu, la prédiction de la personnalité qu'un tiers attribue à un individu (on parle alors de perception) et enfin la génération de personnalités artificielles. Pour ce qui nous

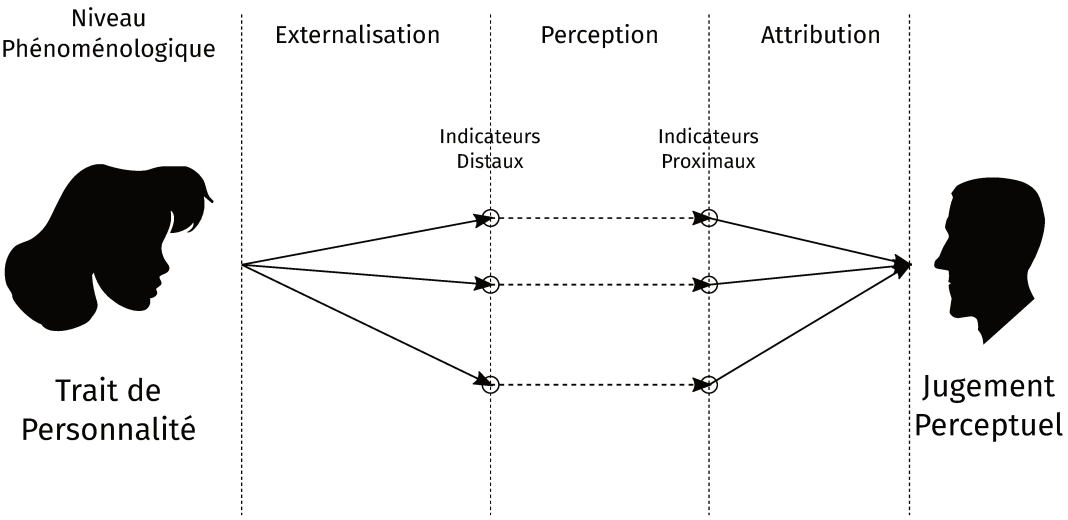

Figure 3.3 – Modèle en lentille de Brunswik (Brunswik 1956).

concerne nous resterons concentrés sur les aspects de perception des traits de personnalité.

Le modèle en lentille de Brunswik (Brunswik 1956) est utilisé pour décrire la perception des traits de personnalité –de la même manière qu'il a été repris par Scherer dans sa théorie de l'évaluation par composante pour décrire le processus de perception des émotions. Il décrit sur différents niveaux les processus d'expression, de transmission et d'attribution d'un trait de personnalité pour un individu donné. Le premier niveau correspond à l'externalisation des indices correspondant au caractère expressif du trait de personnalité du locuteur. Dans le cas du traitement de la parole, il s'agit d'indicateurs acoustiques. Ces indices sont qualifiés de *distaux* dans le sens où ils sont éloignés de l'observateur. Le deuxième niveau est celui de la perception des indices par l'observateur. Les indices perçus sont dits *proximaux*. Enfin le troisième niveau décrit l'attribution d'un trait de personnalité. Cela se fait par le biais de l'interprétation et du jugements de l'observateur, en fonction des indices proximaux. Nous illustrons le processus de perception et d'attribution dans la figure 3.3.

3.3.1 Mesure de la personnalité

Le paradigme dominant dans le domaine de la personnalité est basé sur le modèle des Big Five (BF). Il fait office de modèle de référence pour les traits de personnalité (voir la table 3.1) et est construit à partir de jugements humains à propos de la similarité sémantique et relationnelle des

adjectifs généralement utilisés pour décrire sa personnalité ou celle des autres (McCrae 2009).

3.4 Prédiction automatique des traits de personnalité

Cette tâche a pour objectif de prédire la personnalité qu'un observateur attribue à un individu donné selon les indices proximaux qu'il perçoit dans la parole, à la différence de la tâche de reconnaissance automatique qui cherche à inférer la personnalité propre d'un individu, reconnue par lui-même (Vinciarelli et al. 2014). Si pour la tâche de reconnaissance, il est nécessaire d'avoir les annotations personnelles des individus considérés, il faut pour la tâche de perception automatique, disposer des évaluations faites par les observateurs à propos de l'individu examiné. La perception automatique consiste à prédire la personnalité attribuée en moyenne par différents observateurs. Cette tâche est donc intéressante, car elle vise à mettre en lumière ce que les observateurs ont en commun. À en juger par la littérature, les approches proposées en perception automatique utilisent (en complément d'autres caractéristiques) des indicateurs non-verbaux (Tim et al. 2012 ; Vinciarelli et al. 2014). La hauteur, l'énergie et le tempo de la voix sont parmi les aspects les plus importants de la prosodie. On extrait donc la hauteur de la voix, également les deux premiers formants ainsi que la durée des segments parlés et non-parlés qui reflètent indirectement le débit de parole. Les caractéristiques sont ici calculées sur des fenêtres de signal acoustique d'une durée de 40 ms.

Pour réaliser des prédictions automatiques à partir des caractéristiques acoustiques du signal, des régressions logistiques ou des SVM sont principalement utilisés (Mohammadi et Vinciarelli 2012 ; Mohammadi et Vinciarelli p. d.). Les modèles de prédictions dépassent en général les 70 % de

Traits	Adjectifs
Ouverture	Artistique, Curieux, Imaginatif, Perspicace
Conscienciosité	Organisé, Responsable, Fiable, Minutieux
Extraversion	Actif, Énergique, Loquace, Démonstratif
Agréabilité	Appréciable, Généreux, Sympathique, Affectueux,
Névrosisme	Anxieux, Tendu, Instable, Inquiet

Table 3.1 – Les traits de personnalité du modèle BF.

bonnes prédictions (en considérant uniquement l'information acoustique du signal de parole), ce qui est certes encourageant, mais montre cependant qu'il s'agit encore d'une question ouverte. Plus récemment, l'utilisation des CNN a permis d'améliorer la prédiction des traits de personnalité (Su et al. 2017). Surtout, ils permettent de s'affranchir du processus d'extraction des caractéristiques spectrales. En effet, il est possible d'entraîner ces modèles à capturer l'information leur permettant de résoudre au mieux la tâche demandée, et cela, directement depuis le signal brut.

Ce domaine suscite beaucoup d'intérêt dans la communauté, notamment avec la recrudescence de campagnes d'évaluation qui offrent à la communauté un cadre d'évaluation pour expérimenter des méthodes diverses et variées, sur la tâche de prédiction des traits de personnalité du locuteur au travers de la parole. Par exemple, le *Paralinguistic Challenge* (Schuller, Steidl, Batliner, Nöth et al. 2012).

Ainsi, il est possible de prédire automatiquement les traits de personnalité du BF dans une certaine mesure, en supposant qu'ils soient bien indépendants. Cependant, il est fort probable que les annotateurs soient sujets à des biais d'ordre culturel et cognitif (subjectivité). Parmi les autres problèmes du domaine, il est légitime de se questionner par rapport aux données collectées. En effet, il n'est pas certain que les données soient récoltées selon une méthodologie rigoureuse d'un point de vue psychologique. La tâche de perception automatique des traits de personnalité nécessite des données annotées par des observateurs multiples ce qui requiert donc un engagement financier important et s'avère être chronophage.

3.5 Conclusion

La reconnaissance automatique des émotions et des traits de personnalité, au travers de la voix sont toutes deux des tâches difficiles. Les dernières décennies de recherches ont permis de sélectionner différentes caractéristiques acoustiques contenant de l'information pouvant y ramener. Nous assistons aujourd'hui a un transfert du travail d'ingénierie pour l'extraction et les prétraitements de ces caractéristiques vers un travail d'hyperparamétrisation des architectures neuronales. Il a été montré que ces approches sont tout aussi efficaces que les approches classiques et qu'elles permettent d'apprendre une représentation robuste dédiée à la tâche. Toutefois, un des principaux défis de la reconnaissance automatique des émotions

au même titre que des traits de personnalité, porte sur la captation des informations contextuelles à long terme. En effet, les états affectifs s'étendent le long de l'axe temporel et leurs frontières sont souvent floues.

Enfin, les modèles développés restent dépendants des données utilisées qui sont loin d'être parfaites. Les annotations sont sujettes à la subjectivité des observateurs qui diffèrent selon les corpus. Aussi, le manque de transcriptions phonétiques, parfois, rend d'autant plus difficile la mise au point de modèles génériques.

Ces domaines ont suscité de plus en plus d'intérêt dans la communauté ces dix dernières années, et plus particulièrement pour les émotions. La validité méthodologique est certes questionable, dans le sens où il est légitime de se demander : « Est-ce qu'on mesure vraiment ce que l'on dit évaluer ? » Toutefois, cela a permis de mettre en relation des chercheurs (psychologues, informaticiens) qui n'étaient pas habitués à travailler ensemble.

Chapitre 4

Percevoir et évaluer la similarité entre des voix

Sommaire

Nous avons introduit dans le chapitre précédent des aspects de la perception de la voix qui touchent aux émotions et à la personnalité du locuteur. Cependant, au-delà de ces deux considérations, nous avons omis un grand nombre d'informations perceptibles à travers la voix. Ce chapitre sera donc l'occasion pour nous d'en dire un peu plus sur ce sujet et sur le phénomène de perception en lui-même.

Nous verrons donc ce que la perception de la voix peut apporter comme renseignements. Plus particulièrement, nous verrons ce qui fait qu'une voix

peut être perçue comme plus ou moins attractive ce qui, comme nous allons voir, impacte directement les impressions (les jugements) qui sont émises à propos du locuteur. Aussi, la question de perception de la similarité sera centrale dans ce chapitre, notamment dans le contexte du doublage de voix.

4.1 Perception de la voix

Différentes informations peuvent être perçues au travers de la voix. Kreiman (Kreiman 2018) classe ces informations dans trois catégories distinctes : linguistiques, identitaires (l'identité, mais aussi le genre, l'âge, l'accent, etc.) et affectives.

En premier lieu, il faut noter que la perception de la voix consiste en une interaction entre le signal sonore émis par l'organe phonatoire et l'observateur. La perception est un processus cognitif qui peut se définir comme un processus de transformation des stimuli en information et cela en fonction des connaissances de l'individu (l'observateur). En effet, ce processus est fortement relié à la mémoire. Comme tout processus cognitif, il nécessite un temps pour être exécuté et suscite une réaction qui peut être d'ordre émotionnel, il s'agit donc d'un processus subjectif.

La question de la perception de la voix en elle-même est finalement assez peu étudiée dans la littérature par rapport à la perception de la parole. La manière dont le cerveau reconnaît et différencie une voix peut nous aider à mieux comprendre comment la voix est perçue. Les travaux réalisés dans ce domaine adoptent généralement une approche fondée sur l'analyse des variations des caractéristiques acoustiques de la voix ayant un impact sur les jugement et les impressions des observateurs.

Dans le domaine de la parole pathologique, pour évaluer les différents aspects de la dysphonie, un expert phonéticien utilise l'échelle GRBAS pour décrire la voix selon plusieurs facettes : *Grade*, correspond au degré de sévérité de la dysphonie ; *Rough*, exprime la raucité de la voix ; *Breathy*, fait référence à l'impression de souffle ; *Asthenic*, correspond à une voix hypotonique, qui manque de puissance ; *Strained*, correspondant à une voix hypertonique, serrée, souvent aiguë. Cette échelle a été élaborée dans le but de décrire les voix dysphoniques objectivement (Sáenz-Lechón et al. 2006). Elle n'est donc pas adaptée à la description des voix dites normale.

Par ailleurs, il est possible d'utiliser un espace multidimensionnel pour

observer la similarité des jugements émis par différents observateurs sur la voix. Pour cela, la technique de Positionnement Multidimensionnel (Multidimensional Scaling) (MDS) est fréquemment appliquée. Cette méthode est employée généralement pour évaluer le degré de similarité entre les différents exemples de données. Concrètement, les données sont projetées dans un espace de faible dimensionnalité en préservant au mieux les distances calculées, dans l'espace des données, pour chaque paire d'exemples. Ainsi, les voix recevant des jugements similaires se retrouvent proches dans cet espace et inversement. Ce type d'approche permet –en considérant l'utilisation d'outils de transformation de la voix et donc en ayant le contrôle de certains paramètres acoustiques– de mesurer l'impact de certaines caractéristiques acoustiques sur la perception de la voix (Baumann et al. 2010).

Enfin, les professionnels de la voix, comme les artistes de doublage, les journalistes de radio ou encore les coaches vocaux ont recours à un vocabulaire très riche pour décrire les différentes voix. Ces personnes sont elles-mêmes généralement catégorisées en fonction de la « couleur » de leur voix et ils utilisent habituellement des termes métaphoriques et des analogies pour la décrire. Toutefois, il n'y a pas de réel consensus entre ces professionnels de la voix, ce qui rend difficile la construction d'une terminologie clairement définie. Il est souvent d'usage, d'emprunter des expressions du langage courant pour décrire une voix et notamment pour les personnes chargées de sélectionner des voix de doublage dans le but de stéréotyper un rôle. Ces expressions renvoient généralement à un trait du locuteur ou une émotion.

La voix possède donc de multiples facettes, ce qui la rend difficile à décrire objectivement. La compréhension des mécanismes de perception ainsi que la compréhension des biais cognitifs doivent pouvoir nous aider dans ce sens.

4.1.1 Point de vue des neurosciences

La perception de la voix étant un processus cognitif, il semble légitime de se questionner sur le positionnement d'une voix par rapport aux représentations mentales. Nous avons tous été au moins une fois confrontés, voire étonnés, à l'écoute de notre propre voix (par le biais d'un dispositif d'enregistrement par exemple). Il s'agit là d'un exemple parfait du décalage possible entre nos attentes en termes de perception et la réalité entendue.

Dans un article remarquable, Belin présente la notion de « visage auditif » (Belin, Fecteau et al. 2004). À l'instar du visage, la voix se caractérise par une combinaison unique de caractéristiques acoustiques directement liées aux attributs physiques de l'appareil vocal. De plus, il suggère que des aires spécifiques du cerveau sont impliquées dans le traitement des différentes informations perçues dans la voix (Belin, Zatorre et al. 2000). Il en déduit un modèle d'organisation fonctionnelle du cerveau, similaire à celui proposée pour la perception des visages. Une question très importante reste ouverte, celle de la combinaison des informations linguistiques et non-linguistiques de la voix.

D'un point de vue cognitif, il est intéressant de comprendre comment la représentation mentale d'une voix se crée. Le paradigme actuellement en vigueur veut que chaque nouvelle voix soit encodée en fonction de sa déviation par rapport à un prototype de voix de référence (Papcun et al. 1989). Une voix familière serait encodée uniquement sur la base de ses différences par rapport au prototype. Ce dernier est construit au cours du temps, à partir des voix rencontrées et donc en majorité des voix issues de l'entourage. Ce qui semble mettre en évidence un avantage pour la reconnaissance de voix familières (proches, célébrités) par rapport à des voix inconnues (Schweinberger, Herholz et al. 1997).

4.1.2 Discussion

La perception d'une voix provoque une réponse émotionnelle qui selon l'intensité induit potentiellement un ressenti. En supposant que le ressenti puisse être fonction de l'accord (ou désaccord) avec les attentes, en termes de perception et de réception de l'auditeur, le jugement émis par l'observateur sera alors en faveur (ou défaveur) de la voix et par extension, du locuteur.

De plus, la question du positionnement de la voix perçue par rapport aux représentations mentales de l'observateur, nous amène à supposer l'existence de représentations mentales construites à partir de certaines voix (des prototypes de voix). L'impact d'une voix en termes de ressenti et de jugement émis sont des considérations importantes dans le cadre du choix d'une voix de doublage (films, jeux vidéo) en accord avec les attentes du public. Ce point résume l'enjeu majeur de cette thèse : définir ce qui caractérise ces types de voix et de comprendre les raisons pour lesquelles certaines associations (voix, personnages) sont plus appréciées que d'autres.

4.1.3 Attractivité de la voix

Un grand nombre d'études ont montré que les personnes perçues comme ayant une voix attirante obtiennent des jugements plus favorables que les individus avec une voix moins attirante. Ce phénomène est mis en mots par le célèbre stéréotype : « What sounds beautiful is good » (Zuckerman et Robert E. Driver 1989). Au même titre que l'apparence, la voix a aussi un impact sur les impressions que les autres peuvent avoir, comme l'ont montré Berry 1992 ; Zuckerman et Kunitate 1993. Ces derniers ont également montré que plus le niveau d'attractivité est élevé, plus l'observateur est susceptible de s'identifier au locuteur et plus le désir d'affiliation est grand.

Un des enjeux de ce domaine de recherche est d'identifier les caractéristiques vocales impliquées dans le phénomène que nous décrivons. Les chercheurs considèrent les variables vocales extraites à partir de l'analyse du spectre sonore et les annotations obtenues par des évaluateurs humains. Les annotateurs ont recours à des adjectifs qui renvoient aux traits de personnalité (K. R. Scherer 1974 ; K. R. Scherer 1978). Par ailleurs, Latinus et al. 2011 montrent des corrélations entre l'attractivité de la voix et certains descripteurs de qualité vocale. Hill et al. 2016 nous expliquent que suivant que ces descripteurs varient en fonction du sexe.

Attractivité chez les hommes : Pour l'homme, l'attractivité pour son homologue féminin est corrélée avec la fréquence fondamentale, qui fait référence à la hauteur de voix. Ainsi, une voix grave sera plus attirante. La voix renseigne sur la santé, mais donne aussi des informations sur la forme corporelle de l'homme (A. Sell et al. 2010). Par exemple, les voix graves et masculines sont en relation avec une plus grande taille corporelle (Pisanski et al. 2014). Ces associations semblent avoir permis à un ancêtre masculin d'avoir plus de succès auprès de ses partenaires (Hodges-Simeon et al. 2011).

Attractivité chez les femmes : Selon Wheatley et al. 2014, une voix de femme atteindrait un pic d'attractivité entre 25 et 30 ans, là où la femme est le plus fertile. La fréquence fondamentale de la voix chez les femmes décroît avec l'âge. Ce basculement est entre autres lié à des changements dans la production des hormones reproductives. Dans Puts et al. 2011, les auteurs montrent que les femmes perçues comme plus attirantes par les hommes possèdent une voix qui présente une plus grande dispersion des

formants ainsi qu'une fréquence fondamentale plus haute.

D'autres facteurs ont une influence sur l'attractivité de la voix. Comme l'ont montré Kempe et al. 2013, l'articulation tient aussi son lot de responsabilités et les voix d'hommes articulant moins sont perçues comme moins attractives et élicitent une moins bonne impression chez l'observateur. Les travaux effectués dans Babel et al. 2014 soulignent également l'importance de la prise en compte des indices sociophonétiques dans l'évaluation de l'attractivité vocale qui peuvent indiquer une volonté du locuteur à se conformer aux normes de la société. Par exemple, une voix indiquant un état émotionnel stable aura tendance a être favorisée par rapport à une voix qui laisserait percevoir le contraire. Enfin, les travaux de Bruckert et al. 2010 indiquent que l'attractivité de la voix augmente en fonction de sa proximité, en termes de distance dans un espace multidimensionnel (MDS), avec une voix moyenne. Pour ce faire, les auteurs ont transformé artificiellement des voix et les ont présentées à différents observateurs. Il en résulte que les évaluateurs ont tendance à préférer la voix moyennée à celle d'origine. Ce phénomène peut s'expliquer par une réduction des irrégularités temporelles du spectre de la voix et donc une réduction des apériodicités. Ces dernières étant corrélées avec la perception de l'âge et de la santé du locuteur, il est donc plausible que cela ait un impact sur l'attractivité.

En dehors du fait que certaines voix sont plus enclines à recevoir des jugements favorables et inversement, il y a aussi la question que nous avons évoquée plus haut : le positionnement des voix perçues par rapport aux représentations mentales et de leur positionnement les unes par rapport aux autres. Cela nous amène donc vers la problématique de la perception de la similarité des voix.

4.2 Similarité de la voix

La perception de similarité se définit comme la capacité de percevoir, parmi un groupe d'individus, des voix présentant certaines similitudes, par rapport aux autres. Idéalement, il s'agira de locuteurs de même genre et de même tranche d'âge, pour éviter tout biais à ce niveau-là. La perception de similarité diffère de la notion de similarité perceptuelle. Cette dernière est quant à elle, dans la littérature, reliée à la similarité des jugements émis par différents observateurs. Ce qui définit une voix comme étant similaire à une autre demeure encore aujourd'hui, vague. Comme nous l'avons dit plus

haut, la « qualité » d'une voix est associée, le plus souvent, à des expressions courantes du langage. Généralement, la similarité entre des voix est évaluée par la mise en évidence des caractéristiques acoustiques qui expliquent la variabilité de la voix en termes de qualité de voix.

4.2.1 Mesurer la similarité

Plusieurs tentatives pour comprendre les mécanismes de perception de la similarité sont proposées dans la littérature (Amino et al. 2006 ; Remez et al. 2007 ; Lindh et al. 2010 ; Nolan, McDougall et Hudson 2011 ; Feiser et al. 2012 ; McDougall 2013). Les approches mises en place dans ce but ont recours à des enregistrements de voix et des évaluateurs qui notent la différence entre des paires de voix. La méthode MDS est généralement utilisée pour visualiser, dans un nombre de dimensions réduit, le niveau de similarité entre des paires d'individus. Cette méthode a été utilisée avec des annotations de voix issues de différents corpus, tirés de la base de données DyViS (Nolan, McDougall, De Jong et al. 2009), où un contrôle est exercé sur le genre et l'âge des locuteurs, mais aussi sur leur accent (Nolan, French et al. 2011 ; McDougall 2014). Dans chaque cas, il a été montré que le jugement de similarité, en lien avec la F_0 moyenne, est fortement corrélé avec la première dimension $d1$ et que les moyennes des trois premiers formants $F1, F2, F3$ sont corrélées avec les dimensions $d4, d3, d1$ respectivement. Toutefois, ces corrélations n'expliquent pas les caractéristiques de qualité vocale qui sont capturées par ces indicateurs acoustiques. Plus récemment, les travaux de Segundo (San Segundo et Mompean 2017 ; San Segundo, Foulkes et al. 2018) proposent d'appliquer des méthodes de clustering pour identifier des cohortes de locuteurs perceptuellement similaires sur la base d'annotations de qualités vocales. Les méthodes de classification sont en général cantonnées aux caractéristiques acoustiques (Kelly et al. 2016) ou cepstrales (Adachi et al. 2008a ; Obin, Roebel et Grégoire Bachman 2014 ; Obin et Roebel 2016).

Mesurer la similarité d'un point de vue phonétique

La *Qualité Vocale* est un cadre d'évaluation phonétique de la voix qui selon Laver (Laver 1980) se veut décrire les « caractéristiques auditives qui colorent la voix d'un individu ». Il s'agit d'un système de description de la voix à plusieurs composantes (nasale, soufflée, craquée, tendue, etc.) regrou-

pées selon différents axes. Elles peuvent se cumuler et elles reflètent donc les tendances vocales d'un locuteur, par rapport à une valeur neutre hypothétique. Bien que certaines de ces composantes décrivent un mécanisme de production de la voix, elles sont définitivement liées à la perception de l'évaluateur (Nolan, French et al. 2011).

4.2.2 Le doublage de voix : une question de similarité perceptuelle

Les stéréotypes vocaux jouent, à leur façon, un rôle dans les productions culturelles (films, films d'animation, jeux-vidéos). En effet, les stéréotypes sont partagés comme des conventions entre le public à qui est destiné l'œuvre, les acteurs et la direction artistique. De plus, ils reflètent les traits physiques et de personnalité des personnages et sont donc utilisés à ces fins. Les premiers travaux portés sur l'étude des caractéristiques acoustiques impliquées dans la perception des traits de personnages de fiction font référence à Teshigawara 2003. Ces travaux dressent leur portrait vocal :

Les héros montrent une absence de constriction pharyngale et la présence de voix soufflées.

Les méchants exhibent une constriction pharyngale et une voix rauque. De plus, la majorité des personnages féminins et quelques personnages masculins, présentent une expansion pharyngale et un larynx abaissé.

Nous avons parlé du fait que des observateurs différents infèrent les mêmes attributs et les mêmes traits de personnalité à partir d'une voix. Les résultats de ces travaux montrent que les évaluateurs attribuent plus fréquemment des traits (physiques, personnalité et émotionnels) moins favorables aux voix de personnages qui exhibent une constriction pharyngale. De plus, ils montrent que le genre joue un rôle dans la perception du rôle du personnage. Dans la continuité de ces travaux, Teshigawara propose une analyse statistique de la relation entre les caractéristiques acoustiques et les évaluations des traits du locuteur. De cette manière, il montre la pertinence d'au moins deux facteurs, interprétables en termes de qualité vocale sur la formation des impressions (Teshigawara 2004). D'autres travaux ont étudié la similarité perceptuelle en se basant sur des voix d'acteurs de cinéma japonais. Adachi et al. 2008a proposent pour cela de combiner, de manière linéaire, huit différentes caractéristiques acoustiques dans le but d'identifier parmi un panel de voix d'acteurs, celle qui est la plus similaire à la

voix de leurs utilisateurs. Ces derniers sont dans leur cas, les utilisateurs du *Future Cast System*, un système de divertissement japonais avec lequel les participants se retrouvent immergés dans un film en y incarnant un acteur.

Il est toutefois important de noter que les différences culturelles peuvent induire un changement au niveau des caractéristiques acoustiques qui sont associées aux traits perçus. À ce sujet, Rilliard a étudié les variations interculturelles au niveau des expressions conventionnelles sociales (Rilliard, Shochi et al. 2009 ; Rilliard, Moraes et al. 2013). Ses travaux se concentrent notamment sur l'étude des variations au niveau de la prosodie et leur impact sur la perception des valeurs affectives véhiculées par les expressions de politesse. En revanche, la question de l'interculturalité dans cette problématique de perception des traits et des stéréotypes vocaux, à partir des caractéristiques acoustiques, est à ce jour sans réponse.

Le doublage de voix dans une autre langue implique une sélection de la voix la plus à même de remplacer la voix d'origine et de véhiculer les traits du personnage. Nicolas Obin (Obin, Roebel et Grégoire Bachman 2014 ; Obin et Roebel 2016) propose d'évaluer la similarité entre la voix d'origine et la voix dans la langue cible à partir de caractéristiques para-linguistiques. Pour cela il utilise des voix de personnages de jeux vidéo annotées selon différentes modalités para-linguistiques (genre, age, émotion, qualité de voix, prosodie…). Il combine cette information avec un modèle acoustique optimisé sur une tâche de reconnaissance du locuteur grâce à l'utilisation d'un SVM ayant pour objectif de discriminer la représentation acoustique générée pour chaque segment de voix de personnages, sur la base de leurs annotations para-linguistiques propres. Finalement, l'agrégation des probabilités associées à chaque classe en sortie du SVM, forme un vecteur de représentation à partir duquel il est possible d'effectuer une recherche par similarité. Dans une expérience d'évaluation subjective, il montre l'importance des caractéristiques para-linguistiques dans la perception de similarité des personnages.

Encart 4.1 : Remarque

Dans les travaux de Nicolas Obin, une comparaison subjective est réalisée dans le but d'évaluer l'apport du contenu para-linguistique perçu, par rapport à la représentations acoustique issue d'un système de reconnaissance du locuteur, pour la mesure de la similarité de voix d'acteurs. Cette évaluation subjective suggère que leur système multi-labels surpasse le système de reconnaissance du locuteur. Bien qu'elle ait le mérite d'exister, le nombre d'évaluateurs est au demeurant trop faible pour écarter l'idée d'utiliser un système de reconnaissance du locuteur pour cette tâche.

4.3 Synthèse

Les travaux que nous avons réalisés et que nous décrivons dans la partie suivante sont à la croisée des chemins des différents domaines que nous venons de présenter. Il était donc nécessaire de faire un état de l'art des approches utilisées et du cadre d'évaluation mis en place dans ces différents domaines afin de pouvoir nous positionner dans ce paysage varié. Ce tour d'horizon permet de se rendre compte des différentes informations perçues au travers de la voix. Bien sûr, il n'est pas exhaustif de l'ensemble des éléments inférables automatiquement ou par le biais de la perception. Nous sommes restés concentrés sur les points qui permettent de déceler de l'information sur le locuteur.

Nous avons vu dans le chapitre 2 les éléments d'information qui renvoient à l'identité même du locuteur et les informations qui permettent de le reconnaître. Nous avons fait un état de l'art des techniques en donnant un aperçu général des avancées de ces dix dernières années. Nous avons vu que l 'être humain utilise des informations de différentes natures pour reconnaître une personne. La bonne compréhension des mécanismes humains sur lesquels repose cette faculté de reconnaissance pourrait nous amener à une amélioration des systèmes automatiques. Nous avons parlé dans le chapitre 3 de la dimension affective perçue dans la voix. Nous nous sommes placés principalement du point de vue de la perception de la voix et à aucun moment du point de vue de la production de la parole. Nous avons passé en revue les principaux axes de la perception des traits affectifs et de la personnalité, en partant des fondements psychologiques. De grands efforts ont été faits dans ce domaine de recherche, notamment sur la compréhension

des caractéristiques acoustiques ayant un impact sur la perception de cette dimension. En matière de similarité, il semble que le paradigme actuel soit tourné vers une analyse des impressions et des traits perçus par les observateurs. Il semble aussi nécessaire de considérer les jugements de plusieurs et non pas d'un seul observateur, du fait que la majorité des approches se réfèrent en matière d'évaluation à l'accord (ou désaccord) entre ces derniers.

Nous assistons aujourd'hui à un décalage de cet effort vers l'utilisation massive des réseaux de neurones artificiels et donc sur un travail de recherche à ce niveau-là. Mais il s'agit là de généralité, car cela est vrai pour de nombreux domaines utilisant des techniques d'apprentissage automatique. Il est aussi important, en ce qui nous concerne, de bien savoir dans quel cadre d'évaluation se positionner. Du côté de la reconnaissance du locuteur, le protocole d'évaluation est assez bien établi et éprouvé avec les nombreuses campagnes d'évaluations régulières. À la différence de la reconnaissance automatique du locuteur, où les données utilisées ne nécessitent pas d'annotations expertes (dans le sens où seule l'identité du locuteur est requise), il y a en revanche, un réel besoin de savoir expert dans les domaines qui touchent à la perception. Nous avons vu un engouement de la communauté, en particulier autour des émotions, que l'on ne peut que saluer, puisqu'il est le moteur de nouvelles collaborations scientifiques. Toutefois, ces domaines souffrent d'un manque au niveau du cadre d'évaluation, du fait des coûts financiers et du temps nécessaire pour collecter, annoter et valider les données. Il est donc plutôt légitime de se questionner sur ce qui est vraiment évalué, d'autant plus qu'à côté de ça les avancées réalisées en termes de performance, et notamment avec l'essor du *deep learning* masquent certains de ces problèmes.

Nous constatons que la signification de l'information perçue est toujours attribuée du point de vue de celui qui écoute. En pratique, la voix n'a de sens (du point du vue du locuteur) que lorsqu'elle permet de communiquer un message à un individu tiers. La manière dont la voix est produite traduit donc ce qui est requis (les codes) pour obtenir l'effet escompté. Comme le dit Kreiman (Kreiman 2018) : « voice production is impossible to understand out of the context of voice perception. » Kreiman remet en doute la segmentation actuelle des recherches effectuées sur la voix, basée sur la chaîne de la parole. Il est donc pour elle impossible de comprendre la vraie nature de la voix en étudiant séparément les aspects de production, de transmission et de perception. En effet, la bonne compréhension de la relation entre les mécanismes de production et de perception de la voix la replace dans

un cadre d'évaluation de sa nature même pouvant donc nous aider à mieux comprendre son rôle social.

Deuxième partie

Caractériser la dimension « personnage » dans les voix de doublage

Introduction

Les méthodes d'apprentissage automatique nécessitent souvent un grand volume de données pour attendre un niveau de performances raisonnable. Dans le cas d'un apprentissage supervisé, ces données doivent en plus être annotées. Par exemple, pour des tâches telles que la reconnaissance automatique du locuteur ou des émotions, les données (des signaux de parole) sont étiquetées en fonction de l'identité du locuteur ou de l'émotion perçue. À partir de cette information, les méthodes d'apprentissage automatique sont capables d'extraire une connaissance générale permettant de réaliser des prédictions sur de nouveaux exemples qui répondent à ces annotations.

Dans le cadre du casting vocal, nous avons ainsi besoin de données vocales d'acteurs associées à des rôles pour apprendre des modèles capables d'automatiser cette tâche. Mais à la différence d'autres domaines, la part de subjectif dans l'association d'un rôle à un acteur de doublage est ici très forte. En effet, la sélection qui est opérée par le directeur artistique est peu formalisée, car elle repose sur des facteurs culturels divers ainsi que sur une considération artistique. En dehors de ces considérations, d'autres facteurs (économique, logistique) sont susceptibles d'influencer ce choix. Notre problématique de recommandation automatique de voix nécessite de passer par une phase d'extraction d'information. Ce qui nous amène à la question suivante : « Est-il possible, à partir d'un signal audio (une voix), d'extraire l'information permettant de faire le lien avec un personnage ou un rôle donné ? »

Pour tenter de répondre à cette question et pour éviter d'introduire un biais, nous avons choisi de nous appuyer sur les choix artistiques déjà réalisés par l'opérateur de casting, dans le cadre de productions audiovisuelles déjà doublées. Plus précisément, nous utilisons des voix issues d'un jeu vidéo. Dans un premier temps, nous cherchons à vérifier la présence, dans ces données, de l'information caractéristique de la décision prise par l'opérateur humain. Plus généralement, nous cherchons à montrer qu'il est possible de mettre en évidence une information de haut niveau sémantique à partir d'un gros volume de données, quand bien même la tâche sous-jacente n'est pas clairement définie et l'opérateur humain n'est pas capable d'expliquer les raisons de son choix.

Les données que nous avons à disposition représentent les voix de doublage utilisées pour le jeu vidéo *Mass Effect 3*. Par conséquent, ces voix appartiennent toutes à un même « univers ». Derrière ces voix se cachent des

acteurs américains (anglophones) et français qui ont été sélectionnés par la DA pour jouer un même rôle et ce, bien que leur voix d'acteur (actée) utilisée pendant les enregistrements diffère de leur voix usuelle (neutre). Nous posons l'hypothèse qu'il est possible de mettre en relation les voix de la VO du jeu avec celles de la VF pour mettre en évidence une information caractéristique du choix de voix pour le personnage doublé. Dans l'idéal, nous souhaiterions retrouver les éléments de comparaison qui ont permis à l'opérateur de casting d'effectuer ce choix. Nous supposons que ces éléments correspondent à des spécificités vocales qui pourraient nous aider à définir des prototypes vocaux. La caractérisation d'une telle information représenterait, pour nous, un moyen de mesurer l'accord entre la voix sélectionnée et les attentes du public, en termes de perception pour un rôle ou un personnage donné.

La première étape de nos travaux consiste donc à mettre en évidence l'existence de cette information dans nos données. Cette première tâche repose sur une mesure de similarité de voix. Nous avons choisi de nous appuyer sur notre savoir-faire en reconnaissance du locuteur et de détourner une approche de RAL pour nos besoins. Nous faisons usage des méthodes existantes pour extraire, à partir du signal de parole, l'information propre au locuteur. Ainsi, en mesurant la similarité entre des voix, nous supposons être en mesure de prédire automatiquement les choix de voix réalisés par l'opérateur de casting. Toutefois, cette procédure doit faire l'objet d'un contrôle méthodologique strict. De cette façon, nous serons en mesure de valider la découverte de cette information.

De manière pratique, cette étape correspond à une tâche d'appariement automatique de voix. Nous cherchons à apprendre un modèle capable de retrouver les voix sélectionnées pour doubler un personnage, à partir d'un grand nombre de paires de voix constituées de manière aléatoire. Nous proposons d'évaluer cette tâche en tant que classification binaire : « Est-ce que le modèle réussit à retrouver les choix de l'opérateur oui ou non ? » Ainsi, nous distinguons les paires représentant les voix doublant un même personnage et les paires construites aléatoirement. Nous nous appuyons sur l'observation d'une différence significative entre les scores attribués à ces deux groupes de paires. Cette expérience est complétée par une étude des biais potentiels et leur prise en compte, lorsque c'est possible, par une mise sous contrainte du processus de création des paires de voix.

Après avoir fourni la preuve de l'existence de cette information dans nos données, nous cherchons dans un deuxième temps à caractériser cette infor-

mation. Pour ce faire, nous utilisons des méthodes d'apprentissage discriminant pour, à partir de l'ensemble de voix et des labels dont nous disposons, extraire une connaissance particulière des voix associées aux différents personnages. Au vu des limitations de notre corpus de données en termes de voix et de personnages. Pour pallier ce problème, nous proposons de tirer partie des voix et des labels provenant d'un autre univers (un autre jeu vidéo), grâce à des méthodes de distillation et de transfert de connaissance.

Chapitre 5

Mise en évidence de l'information caractéristique d'un choix artistique

Sommaire

5.1 Introduction

Idéalement, nous aimerions modéliser un espace de représentation de la voix dans lequel les dimensions « locuteurs » et « personnages » seraient indépendantes. Nous pourrions y observer la couverture d'un locuteur donné sur la dimension « personnage », ce qui nous informerait sur son aptitude à jouer plusieurs rôles. Malheureusement, les données dont nous disposons représentent des œuvres doublées en plusieurs langues, deux dans notre cas. Ces données associent un même personnage à deux locuteurs, jouant dans des langues différentes. Cette organisation des données par paires de locuteurs amène un risque de confusion entre la dimension « personnage » que nous souhaitons caractériser et la dimension locuteur en tant que tel : un système d'apprentissage automatique peut trouver plus aisé de mémoriser les appariements de locuteurs plutôt que d'extraire la dimension commune, la dimension « personnage ». Pour extraire l'essence même du personnage dans le signal de parole, il faut être en mesure de maîtriser l'impact de ces deux sources de variabilités (le locuteur et la langue) sur le modèle.

Nous proposons de réaliser une expérience préliminaire dans laquelle nous appairons automatiquement voix originales et voix de doublage en nous appuyant sur les différents segments de voix des personnages et sur l'association réalisée par le directeur artistique. Nous étendons pour cela l'usage d'un système de reconnaissance automatique du locuteur pour la caractérisation des informations de haut niveau, permettant d'estimer la similarité entre des voix de personnages. Nous utilisons une approche fondée sur les méthodes i-vecteur/PLDA et nous proposons une transformation de l'espace i-vecteur afin de réduire la variabilité reliée à la langue, en supposant qu'elle représente une source de nuisances pour la captation de l'information propre au personnage.

Les systèmes de reconnaissance automatique du locuteur s'appuient sur l'appréciation d'un rapport de vraisemblance mettant en jeu deux hypothèses, ou plus exactement une hypothèse et son contraire. Selon l'hypothèse de référence, la voix observée est particulièrement proche du locuteur cible. L'hypothèse inverse nous dit que la voix observée n'appartient vraisemblablement pas à ce locuteur. L'adaptation d'un système de RAL pour notre tâche consiste simplement à reformuler ces hypothèses en considérant le personnage joué à la place du locuteur. Étant donné un segment de voix noté x et un personnage donné p_i, nous calculons le rapport de vraisemblance entre les deux hypothèses suivantes :

H_0 : x provient du personnage p_i.

H_1 : x n'appartient pas au personnage p_i.

Ce chapitre est organisé de la manière suivante. Premièrement, nous présentons dans la section 5.2 l'approche que nous proposons de suivre pour apparier automatiquement les voix. Les détails concernant la mise en place des expériences que nous avons menées sont donnés dans la section 5.3. Nous présentons dans les sections 5.4 et 5.5 les résultats que nous avons obtenus. Enfin, nous faisons une brève conclusion sur ce travail dans la section 5.6.

5.2 Approche

Dans ces travaux, nous nous sommes cantonnés à l'utilisation de voix issues du domaine du jeu vidéo et plus précisément du type Jeu de Rôle (Role Playing Game) (RPG). Nous avons fait ce choix pour deux raisons. La première réside dans le côté pratique de l'extraction des segments vocaux de l'œuvre. En effet, chaque interaction vocale entre les personnages du jeu est enregistrée dans un fichier audio distinct et ne contient rien d'autre que la voix du personnage. L'extraction des segments de voix, depuis le jeu, est donc grandement facilitée, contrairement au cinéma, où le processus de mastérisation nous oblige à traiter une seule et unique bande sonore contenant les dialogues, l'ambiance sonore et la musique. La deuxième raison que nous invoquons pour justifier ce choix est moins d'ordre pratique qu'hypothétique. En effet, nous supposons que les voix des personnages, à l'instar des représentations graphiques (Schröter et al. 2014), sont exagérées voire caricaturales dans les jeux vidéo par comparaison avec le cinéma, exception faite pour les films d'animation, ce qui rend la dimension « personnage » plus aisée à caractériser.

Dans l'approche que nous proposons, nous supposons qu'il est possible d'extraire de l'information à partir de la relation qui existe entre les voix de doublage associées aux personnages. La similarité entre des voix de langues différentes peut être estimée par le biais d'un apprentissage automatique dans lequel le choix artistique qui relie ces voix joue le rôle de supervision.

Pour étudier la dimension personnage de la voix, nous proposons la création d'un système ayant pour objectif de mesurer la similarité entre deux voix. La figure 5.1 illustre une vue simplifiée du système. Ce dernier prend en entrée deux fichiers audio correspondant à deux segments de voix : le

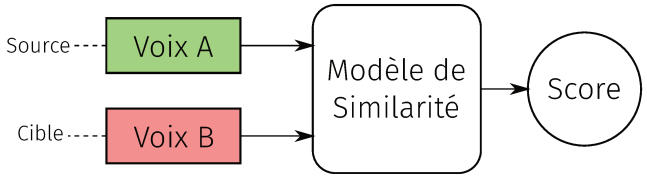

Figure 5.1 – Vue simplifiée du système de similarité.

premier dans la langue d'origine et le second dans la langue cible. En guise de sortie, le système produit un score estimant la proximité de ces deux voix dans la dimension personnage. Ainsi, en considérant deux voix doublant un personnage identique, le système doit produire un score maximal. À l'inverse, il doit attribuer un score minimal voire nul lorsque deux voix n'ayant aucun rapport en termes de personnage lui sont présentées.

5.2.1 Une approche *i*-vecteur / PLDA

Pour répondre à cette question de similarité de voix. Nous avons choisi de créer un système reposant sur l'approche éprouvée *i*-vecteur/PLDA, que nous utilisons en tant qu'outils « clés en main » pour l'estimation de la similarité.

Représentation des entrées

De manière générale, le choix de la représentation des données en entrées (des segments de voix), a un impact sur les performances globales du système. Étant donné que nous traitons des segments de voix de durées variables, nous avons opté pour une représentation des données d'entrée s'appuyant sur l'espace des *i*-vecteurs. Ces derniers ont été initialement introduits en RAL et par la suite repris dans d'autres tâches comme la reconnaissance de la langue (N. Dehak, Torres-Carrasquillo et al. 2011 ; Martinez et al. 2012) et la reconnaissance des émotions (Xia et al. 2012).

Un *i*-vecteur est une représentation compacte extraite à partir d'une séquence de paramètres acoustiques de taille variable issue d'un signal de parole. Un grand volume de données provenant de différents locuteurs, enregistrés sur plusieurs sessions, est utilisé pour entraîner la matrice de variabilité totale notée T. Les segments audio sont alors projetés dans cet espace et sont caractérisés par les *i*-vecteurs. Plus de détails sont donnés dans la sous-section 2.3.2.

Normalisation des entrées

Les i-vecteurs que nous extrayons sont normalisés selon la méthode dite Length Whitening Normalization (LW-Norm) proposée par Bousquet, Matrouf et al. 2011. Cette dernière permet notamment d'obtenir une distribution normale des données en entrée de la PLDA (voir la sous-section 2.3.2).

Comparaison des voix

Comme dit ci-dessus, nous représentons chaque segment de voix par un i-vecteur et nous estimons la similarité pour une paire de segments en utilisant l'approche PLDA. Cette dernière nous donne un score pour la paire de voix en entrée correspondant au rapport de vraisemblance noté LR et reflétant la similarité des segments qui la composent. Considérée comme étant la méthode de comparaison état-de-l'art en RAL, cette technique consiste à projeter les données d'entrée (i-vecteurs) dans un espace de dimensionnalité réduite tout en minimisant la perte de leur pouvoir de discrimination et en maximisant le rapport de la variance inter- et intra-locuteur (voir la sous-section 2.3.2).

À titre de comparaison, nous estimons aussi la similarité au travers du calcul d'une distance Euclidienne et d'une distance de Mahalanobis ainsi qu'en utilisant la méthode WCCN.

5.2.2 Neutralisation de la langue par appariement des voix

Le corpus dont nous disposons est composé de segments de voix dans différentes langues. Ainsi, chaque segment de la langue source et de la langue cible est utilisé afin d'en extraire une représentation dans l'espace des i-vecteurs. Le système que nous mettons en place permet la comparaison de deux segments appairés de la façon suivante : (x_m, x_n) tel que x_m correspond au vecteur de représentation d'un segment dans la langue source et x_n correspond à celui d'un segment dans la langue cible. En neutralisant la variabilité au niveau de la langue, nous serons en mesure de dire si elle a un impact sur le système, et par conséquent, sur la capacité à distinguer des voix d'acteurs similaires.

Passer de la langue source à la langue cible

La comparaison de deux éléments ne peut se faire qu'à condition qu'ils proviennent d'un même espace. Le cœur de cette approche consiste à projeter les i-vecteurs de l'espace de langue cible vers l'espace de langue source (et inversement) pour rendre possible la comparaison. Pour cela, nous nous sommes inspirés d'une approche utilisée en traduction automatique du langage (Mikolov et al. 2013). Dans ces travaux, les auteurs exploitent les similitudes structurelles propres aux plongements lexicaux (plus connus sous l'appellation *word embeddings*) pour apprendre une transformation linéaire permettant de passer d'un espace de mots représentatif d'une langue à une autre. Pour l'adapter à notre problématique, nous apprenons une transformation à partir d'un ensemble de paires de segments de voix noté $\{x_m, x_{m'}\}$ où x_m fait référence au i-vecteur d'un segment de voix dans la langue source et $x_{m'}$ correspond au i-vecteur du segment équivalent dans la langue cible. Ici, « équivalent » signifie que les segments x_m et $x_{m'}$ sont prononcés par le même personnage et véhiculent le même message. Nous voulons donc trouver la fonction de transformation G telle que :

$$x_m \approx G_W(x_{m'}). \tag{5.1}$$

Nous utilisons la méthode des moindres carrés (Adrien Marie Legendre 1805) pour trouver la matrice W qui paramétrise cette fonction. Nous sommes donc face à un problème relativement simple visant à minimiser :

$$\| x_m - G_W(x_{m'}) \|^2. \tag{5.2}$$

Ici G_W est assimilé à une transformation linéaire permettant de passer de l'espace de langue cible vers l'espace de langue source.

Variantes du système

L'approche mise en œuvre comprend plusieurs phases : représentation des séquences, normalisation, neutralisation de la langue et comparaison. Nous proposons donc trois variantes de ce système notées A, B et C que nous illustrons dans la figure 5.2. La variante A n'opère aucune compensation de la langue et nous sert donc de système de référence pour évaluer notre approche, mise en œuvre dans les variantes B et C. Ces deux dernières diffèrent uniquement dans l'ordre de réalisation des étapes de normalisa-

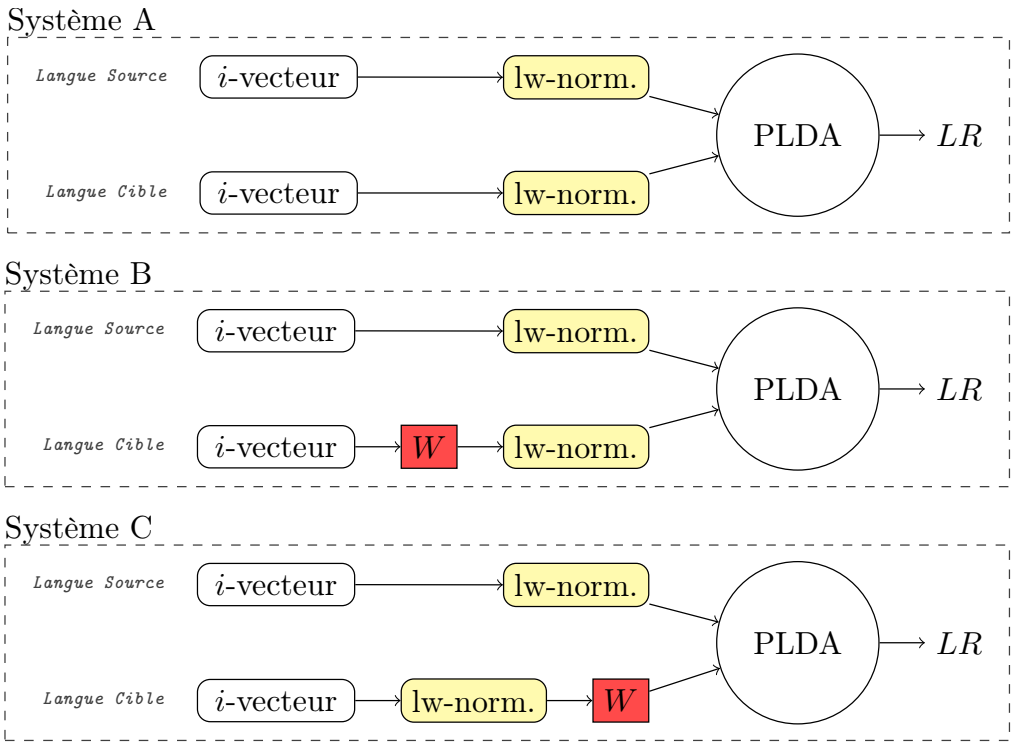

Figure 5.2 – Illustration du système de référence (A) et des deux variantes proposées dans notre approche (B et C). La *Probabilistic Linear Discriminant Analysis* (PLDA) permet d'estimer la similarité entre deux i-vecteurs au moyen d'un *Likelihood Ratio* (LR).

tion et de neutralisation. En effet, nous savons l'importance de la normalisation avant l'étape de comparaison sur les systèmes des vérification du locuteur (Bousquet, Matrouf et al. 2011). L'idée est ici de mesurer expérimentalement l'impact de la place du processus de normalisation au sein de cette chaîne.

Configurations des espaces de langue

Notre système considère deux entrées correspondant à des segments de voix : la première dans la langue source de l'œuvre et la seconde dans la langue cible. Chaque segment en entrée fait référence à un i-vecteur provenant de l'espace de variabilité totale lui-même représenté par la matrice T. Il a été observé que la spécialisation de l'UBM pour une langue en particulier donne de meilleurs résultats (Kleynhans et al. 2005). Par conséquent, nous avons modélisé deux espaces de langue indépendamment dans le but de mieux prendre en compte les variations propres au locuteur lui-même. L'UBM et la matrice T sont donc entraînés sur des signaux de parole provenant d'une langue spécifique.

Ainsi, nous considérons plusieurs configurations des entrées du système, en termes d'espaces de langue d'où sont extraits les i-vecteurs. En les faisant varier, nous pourrons observer l'impact de la langue sur le système.

EN → EN : Les segments sources et cibles sont représentés par des i-vecteurs provenant de l'espace de variabilité totale appris sur de l'anglais.

FR → FR : Les segments sources et cibles sont représentés par des i-vecteurs provenant de l'espace de variabilité totale appris sur du français.

EN → FR : Les segments sources sont représentés par des i-vecteurs issus de l'espace de variabilité totale appris sur l'anglais et les segments cibles sont représentés par des i-vecteurs issus de l'espace de variabilité totale appris sur du français.

FR → EN : Les segments sources sont représentés par des i-vecteurs issus de l'espace de variabilité totale appris sur le français et les segments cibles sont représentés par des i-vecteurs issus de l'espace de variabilité totale appris sur de l'anglais.

5.3 Protocole expérimental

Dans cette section, nous détaillons le protocole expérimental que nous avons suivi durant nos expériences. Nous commençons par décrire le corpus de données dans la sous-section 5.3.1 ainsi que la méthode d'extraction des paramètres acoustiques dans la sous-section 5.3.2. Enfin, le protocole d'évaluation est décrit dans la sous-section 5.3.4.

5.3.1 Description du corpus

Les données que nous utilisons dans ce cadre expérimental sont issues d'un jeu vidéo appelé *Mass Effect 3*. Nous utilisons les interactions vocales entre les différents personnages du jeu (les dialogues du jeu). Ce jeu a été initialement proposé en anglais, mais a été traduit et doublé dans d'autres langues, notamment le français. Nous avons donc les fichiers audio qui constituent la VO (anglais) et la VF des dialogues.

Nous disposons de chaque interaction d'un personnage sous la forme d'un fichier audio respectant la nomenclature décrite dans la figure 5.3. La

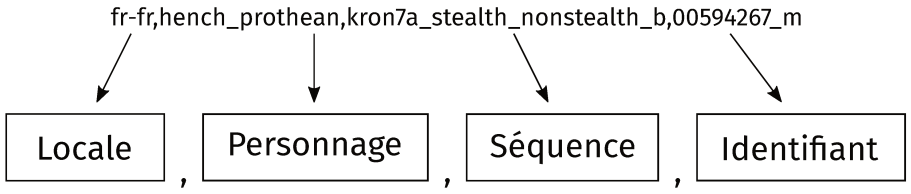

Figure 5.3 – Nomenclature des fichiers de segments de voix.

première partie du nom définit la langue utilisée, la seconde correspond au personnage qui parle, la troisième correspond à la séquence dans le jeu où se trouve le dialogue et la dernière partie donne un identifiant au segment. Par correspondance, il est possible de déterminer toutes les paires de segments VO-VF d'un personnage particulier. Grâce à cette information, nous avons pu extraire 10 000 segments vocaux dans les deux langues. Ce sont tous des enregistrements de haute qualité réalisés en studio (échantillonnés à 44,1 kHz et codés sur 16 bit).

Le corpus que nous avons recueilli se compose de 31 personnages joués par 62 acteurs. Deux acteurs distincts (l'un dans la VO, l'autre dans la VF) incarnent un personnage. Nous considérons chaque segment de voix comme appartenant à un personnage p_i prononcé dans une langue notée l_1 ou l_2. Nous considérons également le message m contenu dans le segment, ainsi si un segment délivre le message m dans la langue l_1, nous notons m' le segment délivrant le message équivalent à m dans la langue l_2.

Au total, nous disposons d'un peu plus de sept heures de parole dans chaque langue. Les segments ont une durée comprise entre 0,1 s et 54 s et la durée moyenne des segments est de 3 s. La figure 5.4 décrit la répartition des durées des segments. Toutefois, le temps de parole n'est pas uniformément réparti entre les différents personnages, du fait de leur importance relative dans le scénario.

De manière générale, un segment correspond à une phrase. Même si peu fréquent, il peut également s'agir d'une interjection (cri, rire). La part d'information importante dans ces données réside dans le lien qui unit un acteur de la VO avec un acteur de la VF dans l'incarnation d'un même personnage. Nous travaillons donc sur des paires de segments. Nous considérons l'ensemble des paires résultant de la combinaison de tous les segments en anglais avec tous ceux en français. Nous distinguons de fait deux groupes de paires différentes, les paires :

Target : constituées de deux segments provenant d'un même personnage.

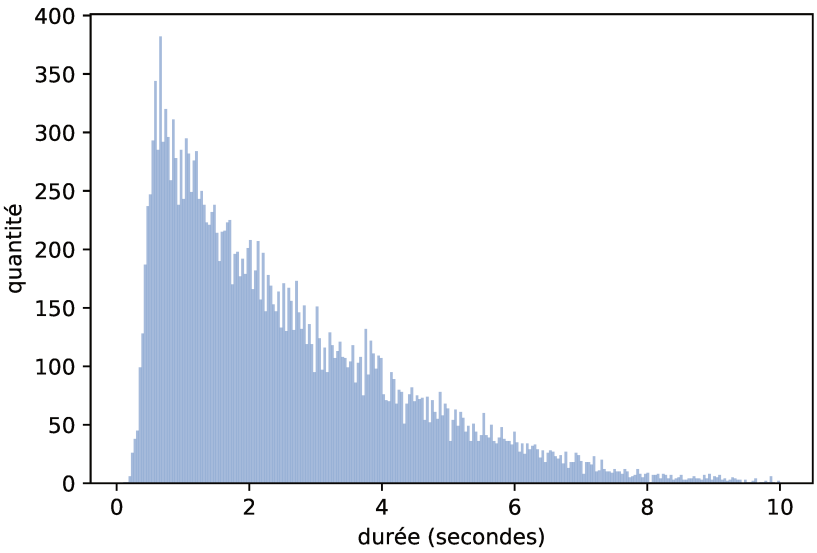

Figure 5.4 – Histogramme des durées des segments de voix du corpus *Mass Effect 3*. Ici, les segments d'une durée supérieure à 10 s (42 segments) ne sont pas représentés pour des raisons pratiques.

Nontarget : composées de segments de personnages différents.

En termes de quantité, il y a naturellement un grand déséquilibre entre ces deux groupes. Pour chaque personnage nous avons un ensemble de paires *target* (les combinaisons de ses segments en anglais avec ses segments en français) et un ensemble de paires *nontarget* (combinant ses segments en anglais avec tous les segments français des autres personnages).

5.3.2 Extraction des paramètres acoustiques

Nous réalisons cette expérience en nous appuyant sur une méthodologie consensuelle en RAL et sur la boîte à outils Kaldi (Povey, Ghoshal, Boulianne, Burget et Glembek 2011). Le système mis en place est similaire au système proposé par l'équipe RAL du LIA pour la campagne d'évaluation NIST SRE 2016 (Rouvier et al. 2016). Les paramètres sont calculés sur des fenêtres de 20 ms avec un chevauchement de 10 ms. Nous calculons pour chaque fenêtre 19 paramètres MFCC déterminés sur 40 bancs de filtres Mel et accompagnés par le logarithme de l'énergie. À cela, nous ajoutons les dérivées de premier (Δ) et second ordre ($\Delta\Delta$) : nous obtenons ainsi un vecteur de paramètres acoustiques de dimension 60 pour chaque trame du signal.

Nous effectuons une normalisation des paramètres par CMN. Nous ignorons volontairement les trames de faible énergie correspondant principalement à du silence grâce à un algorithme de VAD qui décide des trames

à conserver en fonction d'un seuil fixé par consensus entre les 10 trames contextuelles et la trame courante. Enfin, nous entraînons le modèle du monde (UBM) de 2048 composantes indépendant du genre ainsi que la matrice de variabilité totale de rang 400 grâce à quoi nous pouvons extraire les i-vecteurs.

5.3.3 Apprentissage du système

Nous avons entraîné deux systèmes d'extraction i-vecteur, un spécifique à chaque langue (anglais et français). Pour le système anglais, l'UBM, la matrice T ainsi que la PLDA sont appris sur les corpus NIST SRE 2004, 2005 et 2006. Pour apprendre le système sur du français, nous avons utilisé les corpus de parole ESTER-1, ESTER-2, EPAC, ETAPE et REPERE.

Concernant l'entraînement de la matrice W qui opère la projection d'un espace de langue à l'autre, nous utilisons une méthode de validation croisée. Nous considérons un découpage en 3 plis différents au niveau des personnages et nous prenons en compte la contrainte du nombre de segments par personnage. Ainsi la matrice W est entraînée indépendamment sur les 3 plis, chacun considérant un ensemble de paires de segments VO-VF restreint à un ensemble de personnages (le reste des personnages sont destinés à l'évaluation).

5.3.4 Méthode d'évaluation

Dans cette étude, nous cherchons à prouver l'existence d'une information présente dans le signal de parole qui soit caractéristique du personnage et non du locuteur. Notre méthode d'évaluation consiste donc à vérifier que le système est capable de discriminer les paires *target* des paires *nontarget*.

Nous comparons les segments de voix de la langue l_1 avec ceux de la langue l_2 et demandons au système si oui ou non ils proviennent du même personnage. L'évaluation consiste donc à vérifier pour chaque comparaison, la prédiction faite par le système. La réussite, au sens strict du terme, suppose que pour un personnage donnée les paires de segments *target* (même personnage) possèdent toutes un meilleur score que les paires de segments *nontarget* (personnages différents).

Nous considérons l'ensemble des paires possibles qui se définit comme

suit :

$$P_{total} = \{(X_{i,m}^{l_1}\ ,\ X_{j,n}^{l_2})\} \tag{5.3}$$

quelque soit les segments m, n et quelque soit les personnages évalués p_i et p_j.

Soit un segment $x_{i,m}$ associé au personnage p_i appartenant à l'ensemble des segments X^{l_1}. Pour évaluer une prédiction, nous prenons compte de toutes les combinaisons impliquant $x_{i,m}$. Ainsi nous considérons le sous-ensemble de paires suivant :

$$P_{i,m} = \{(x_{i,m}\ ,\ X_{j,n}^{l_2})\}. \tag{5.4}$$

Étant donné les scores associés à chaque paire contenue dans $P_{i,m}$, il suffit de retenir la paire ayant obtenu le score maximal. Celle-ci nous donne le segment $x_{j,n}$ dans la langue l_2 le plus similaire à $x_{i,m}$. La prédiction du système est donc correcte si cette paire vérifie que $p_i = p_j$, autrement il s'agit d'une erreur. Cependant, nous pensons que cette contrainte est beaucoup trop forte compte tenu de la difficulté de la tâche. De plus, un système de recommandation tel que nous le concevons doit être capable de faire plusieurs suggestions à l'utilisateur. En suivant cette idée, nous introduisons un degré de liberté dans l'évaluation du système. De fait, nous ne considérons pas seulement le meilleur score, mais les k meilleurs scores obtenus sur l'ensemble des paires $P_{i,m}$.

En définitive, nous considérons que le système effectue une prédiction correcte pour le segment $x_{i,m}$ si parmi les paires $P_{i,m}^{\star}$ ayant obtenu les k meilleurs scores, il existe une paire telle que $p_i = p_j$.

Nous évaluons la justesse des prédictions du système pour chacune des comparaisons. Nous effectuons autant de comparaisons qu'il y a de segments dans X^{l_1}. Le taux de réussite du système correspond à la proportion du nombre de prédictions correctes par rapport au nombre total de comparaisons effectuées. Il se définit également en termes de positifs et de négatifs :

$$\text{Taux de réussite} = \frac{VP + VN}{VP + VN + FP + FN} \tag{5.5}$$

De plus, nous évaluons notre approche sur la base d'une validation croisée en 3 plis. Ainsi, l'évaluation considère à chaque fois les paires de segments VO-VF d'un ensemble de personnages indépendant de l'apprentissage, ce qui rend cette tâche très difficile.

k=3	Système *A*	Système *B*	Système *C*
EN → EN	0,59	0,60	0,61
EN → FR	~~0,54~~	0,60	0,53
FR → EN	~~0,52~~	0,60	0,49
FR → FR	0,60	0,61	0,63
k=1	Système *A*	Système *B*	Système *C*
EN → EN	0,37	0,35	0,36
EN → FR	~~0,34~~	0,35	0,34
FR → EN	~~0,28~~	0,35	0,28
FR → FR	0,39	0,35	0,38

TABLE 5.1 – Taux de réussite des prédictions de la similarité des différents systèmes ($k = 3$). Tient compte des résultats cumulés sur les différents plis.

5.4 Mettre en évidence l'information caractéristique du personnage

Dans cette section, nous présentons les résultats que nous avons obtenus avec les trois systèmes que nous avons présentés plus haut (voir figure 5.2). Nous proposons également une première analyse des résultats et commençons une discussion à partir de ces derniers.

5.4.1 Système de référence

Le résumé des résultats obtenus avec les différents systèmes est présenté dans la table 5.1. Notre système de référence (le système *A*) est évalué sur la base des deux configurations impliquant la même langue (EN → EN et FR → FR). En effet, la comparaison de deux *i*-vecteurs n'est possible que si ils sont extraits à partir d'un même espace de variabilité totale. Une distance calculée entre des vecteurs provenant de deux espaces différents n'a pas d'intérêt. Aucune projection n'étant effectuée dans la variante *A* du système, les résultats indiqués pour les configurations impliquant deux espaces différents (FR → EN et EN → FR) ne sont pas pertinents, d'où le fait qu'ils soient rayés.

Concernant les deux seules configurations pertinentes pour le système *A*, nous obtenons d'assez bons résultats compte tenu de la difficulté de la tâche avec un taux de réussite de 0,59 et 0,60.

Nous avons illustré dans la figure A.1 les différents résultats, en termes

de réussite, suivant différentes valeurs de k. Globalement, nous voyons sur les différentes courbes un point d'inflexion situé approximativement aux alentours de $k = 3$. Nous avons fixé k à 3 en conséquence. De plus, cela semble un bon compromis dans l'optique finale de recommander à l'utilisateur non pas une, mais plusieurs voix.

5.4.2 Compensation de la langue

Nous comparons maintenant les deux systèmes utilisant la projection apprise sur le corpus *Mass Effect 3* (les systèmes B et C) avec le système de référence. D'une manière générale, nous obtenons dans les deux cas de meilleurs résultats lorsque le même espace de langue est utilisé.

Étonnamment, la projection ne semble pas compenser les différences au niveau de la langue pour les cas EN → FR et FR → EN. Au contraire, les résultats nous laissent supposer que l'utilisation d'un même espace de variabilité totale pour les langues de départ et d'arrivée permet de mieux apprendre les caractéristiques vocales qui caractérisent le personnage. Cette supposition est induite par le taux de réussite de 0,63 obtenue avec le système C, pour lequel l'apprentissage de la matrice de projection intervient en bout de chaîne et bénéficie donc de la normalisation. Nous notons également que lorsque la normalisation intervient après l'étape de projection (système B) nous constatons une variation moins importante des résultats.

Nous illustrons dans les figures A.1, A.2 et A.3 les résultats suivant des valeurs de k comprises entre 1 et 10 pour les différentes méthodes de comparaison utilisées. À ce propos, nous observons une nette distinction entre les méthodes s'appuyant sur le calcul d'une distance euclidienne, ou distance de Mahalanobis, avec les méthodes de comparaison WCCN et PLDA. Nous supposons que WCCN et PLDA sont plus robustes aux confusions entre locuteurs, notamment grâce à leur effet de compensation de la variabilité inter-session. Cela montre également combien il est difficile de s'affranchir complètement du locuteur.

En définitive, la configuration FR → FR est celle qui nous donne les meilleurs résultats quel que soit le système utilisé avec la PLDA comme méthode d'évaluation. Dans le cas d'une réussite stricte ($k = 1$), WCCN donne des résultats similaires.

Les résultats bruts du système de RAL sont quant à eux présentés dans la

table A.1. Ces résultats font état du pourcentage d'EER [1] en considérant les scores de chaque paire de segments de voix (VO / VF) évaluée. Cela correspond à 37 787 933 comparaisons pour chaque système appris, tous personnages confondus, chaque paire étant annotée *target* (même personnage) ou *nontarget* (personnages différents). De plus, nous faisons apparaître les résultats de chacun des plis de notre validation croisée ainsi que les résultats moyennés.

Encart 5.1 : Pour aller plus loin

Nous avons supposé pouvoir entraîner la PLDA en nous appuyant sur nos personnages, en utilisant pour chaque personnage les sessions des deux locuteurs anglais et français, mais sans succès. Si nous disposions d'un corpus comprenant plusieurs locuteurs par personnage dans chaque langue, nous pensons qu'il serait possible de compenser la variabilité inter-locuteur au niveau de chaque classe de personnage. À défaut d'avoir plus de données, nous observons toutefois des résultats corrects dans notre cas, ce qui nous renvoie à la question suivante : « Est-ce que c'est le locuteur (l'acteur) qui fait le personnage ou l'inverse ? »

5.5 Contrôler le contenu linguistique

Dans l'approche que nous proposons demeure une ambiguïté relative à la capacité du système à discriminer les voix en fonction du personnage ou en fonction du contenu linguistique des segments de voix. D'une manière générale, un personnage effectue plusieurs dialogues différents et l'information linguistique est la même dans VO et dans la VF. Nous cherchons ici à savoir si le message véhiculé en lui-même n'a pas un impact sur le système. Il est possible que le modèle associe simplement certains messages à un personnage voire qu'il soit capable d'identifier les segments qui véhiculent le même message. Cela constitue donc un biais potentiel. Nous proposons une méthode pour le vérifier. Pour cela nous avons recours à une approche contrastive.

1. L'EER est le point où le False Positive Rate (FPR) et le True Positive Rate (TPR) sont égaux.

5.5.1 Mise en évidence du biais

Expérience contrastive : protocole

En considérant les trois tests ci-dessous, nous souhaitons observer les confusions possibles du système. En s'appuyant sur la moyenne des scores obtenus dans les différents cas de test, nous sommes capables de dire si le pouvoir de discrimination du système se trouve bien au niveau du personnage et non pas au niveau du segment.

1. ***Strict equivalence test*** **:** X_1 représente un segment de voix noté m du personnage i dans la langue source l_1 et X_2 correspond au segment équivalent m' dans la langue l_2.
2. ***Soft equivalence test*** **:** X_1 représente un segment m du personnage i dans la langue l_1 et X_2 correspond à un segment n différent de m provenant du personnage i dans la langue cible l_2.
3. ***Contrastive test*** **:** X_1 représente un segment de voix noté m du personnage i dans la langue source et X_2 correspond à un segment n appartenant à un personnage j avec $i \neq j$ dans la langue l_2.

Expérience contrastive : résultats

Nous présentons dans la figure 5.5 les résultats des tests comparatifs que nous avons mis en place. Les résultats montrent bien une distinction sur les scores moyens (graphiques du haut) et leurs écarts-type respectifs (graphiques du bas) entre les différents tests, avec en jaune les scores moyens obtenus en test de comparaison avec équivalence stricte (même personnage, et même message), en bleu les scores moyens qui résultent des comparaisons d'équivalences dites *soft* (même personnage, quel que soit le message) et en vert ceux issus des tests comparatifs (personnages différents).

Les résultats de cette expérience tendent à montrer que le contenu linguistique impacte notre système, étant donné que les scores moyens des tests de comparaison avec équivalence stricte des segments sont meilleurs que ceux avec équivalence simple du personnage. Il est également possible qu'il s'agisse d'un biais sur la durée des segments, lui-même potentiellement lié au biais du contenu linguistique.

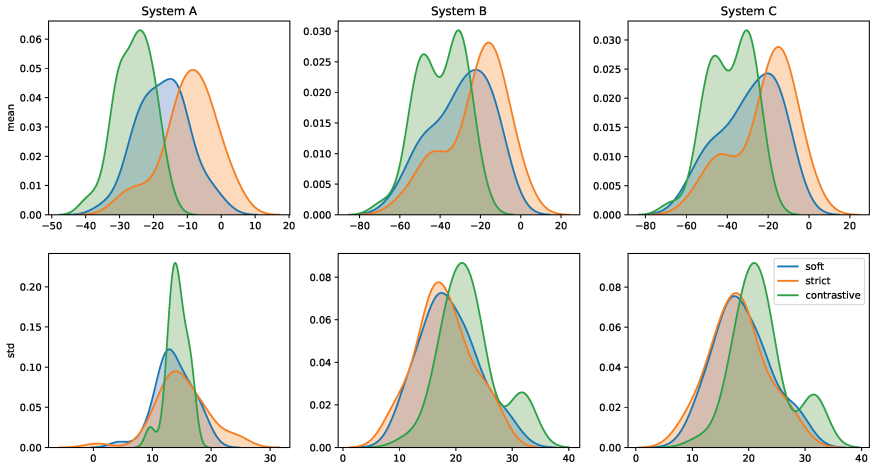

Figure 5.5 – Distributions des scores moyens obtenus sur les différents systèmes en configuration FR → FR. Les graphiques du haut illustrent les scores moyens des tests effectués sur les différents systèmes. Ceux du bas montrent leurs écarts-type respectifs.

5.5.2 Briser l'équivalence VO-VF de l'information linguistique

L'apprentissage de la matrice de projection W est réalisé à partir des paires de même personnage et composées de segments équivalents en termes de contenu linguistique. Pour plus de détails, il faut retourner à la sous-section 5.2.2. Les résultats du test comparatif montrent l'impact du contenu linguistique sur le système.

Protocole

Nous proposons ici de voir dans quelle mesure l'information linguistique impacte le système présenté dans la figure 5.6. Pour cela, nous mettons en place un protocole spécifique à l'apprentissage de la matrice de projection dans lequel nous neutralisons la relation d'équivalence entre les segments de voix de la version anglaise et française par le biais d'un appariement aléatoire des segments d'un même personnage.

Nous avons $X_i^{l_1}$ et $X_i^{l_2}$ les deux ensembles de segments du personnage p_i dans les langues l_1 et l_2. Il existe une correspondance bijective entre ces deux ensembles. Nous apprenons la matrice de projection notée W de telle sorte que $x_m \approx G_W(x_n)\ \forall m, n$ tel que $m \neq n$ avec $x_n \in X_{p_i}^{l_1}$ et $x_m \in X_{p_i}^{l_2}$.

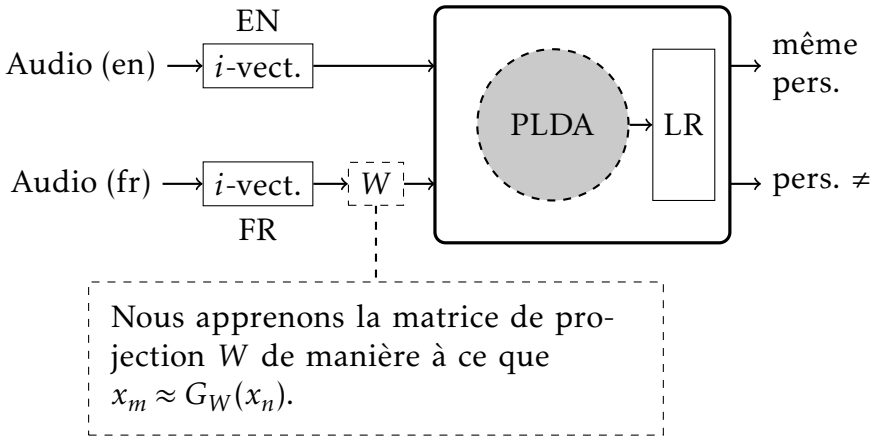

Figure 5.6 – Méthode d'apprentissage de la matrice de projection avec neutralisation du biais linguistique. La *Probabilistic Linear Discriminant Analysis* (PLDA) permet d'estimer la similarité entre deux *i*-vecteurs au moyen d'un *Likelihood Ratio* (LR).

Résultats

Les résultats de l'expérience réalisée avec neutralisation de la composante linguistique sont présentés dans la table 5.2. La colonne *A* reste naturellement inchangée par rapport aux résultats de l'expérience précédente (se référer à la table 5.1) puisqu'elle n'a pas recours à cette transformation. Nous observons une nette amélioration des résultats obtenus sur le système *C* pour les configuration EN → FR et FR → EN. À première vue, le fait d'apprendre la projection sur des segments appareillés au hasard pour chaque personnage permet une meilleure projection lorsque les espaces de variabilité sont différents.

k=3	Système *A*	Système *B*	Système *C*
EN → EN	0,59	0,59	0,59
EN → FR	~~0,54~~	0,60	0,57
FR → EN	~~0,52~~	0,58	0,58
FR → FR	0,60	0,53	0,62
k=1	Système *A*	Système *B*	Système *C*
EN → EN	0,37	0,33	0,35
EN → FR	~~0,34~~	0,36	0,32
FR → EN	~~0,28~~	0,30	0,36
FR → FR	0,39	0,29	0,38

Table 5.2 – Taux de réussite des prédictions de la similarité des différents systèmes ($k = 3$) pour le test de la composante linguistique. Tient compte des résultats cumulés sur les différents plis.

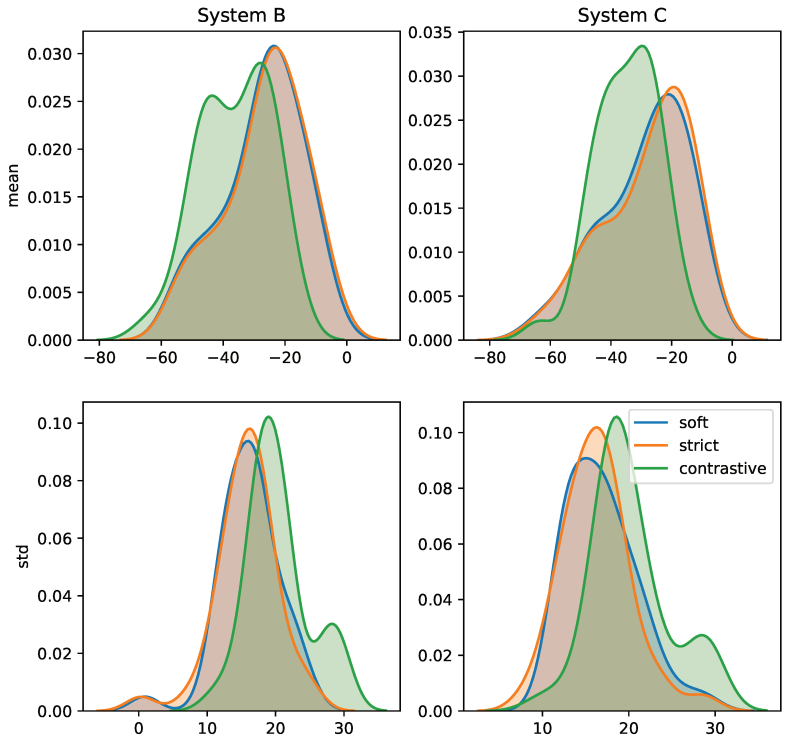

Figure 5.7 – Scores moyens obtenus sur les systèmes *B* et *C* pour le test de la composante linguistique en configuration FR → FR.

Comme le montre la figure 5.7, la distribution des scores *strict* et *soft* est très similaire grâce à la neutralisation de la dimension linguistique. Ces résultats confirment le biais lié au message. Nous supposons que dans des langues proches comme l'anglais et le français, il est possible de retrouver des similitudes au-delà même du contenu linguistique, par exemple, les inflexions que la voix de doublage cible serait tentée d'imiter dans le but de maintenir une certaine cohérence avec le jeu d'acteur de la voix d'origine. Il serait très intéressant de réaliser la même expérience en contraste avec une langue très éloignée de celles dont nous disposons, par exemple le japonais.

Encart 5.2 : Remarque

L'apprentissage de la fonction de *mapping* s'appuie sur les choix réalisés par l'opérateur de casting. Chaque segment de voix est représenté par un i-vecteur où chaque facteur contrôle une dimension propre de la matrice de variabilité totale notée T. L'apprentissage du *mapping* de l'espace de variabilité totale de la langue cible vers l'espace de variabilité totale de la langue source par le biais des i-vecteurs de ces espaces est difficile dans le cas où les langues sont différentes (EN $\rightarrow$ FR et FR $\rightarrow$ EN) puisqu'il n'existe aucune correspondance entre les facteurs qui les composent. Une simple transformation linéaire peut difficilement permettre de passer d'un espace à un autre. Il semble de fait plus logique que les résultats soient meilleurs dans les cas où la langue est la même (EN $\rightarrow$ EN et FR $\rightarrow$ FR). En effet dans ce cas là, chaque facteur correspond bien à une dimension propre de l'espace de variabilité puisque les i-vecteurs proviennent du même espace. De cette manière la fonction de projection n'a pas à compenser la différence d'espaces et peut s'appuyer directement sur l'information permettant de lier la voix de doublage à celle d'origine.

5.6 Conclusion

Dans ce chapitre, nous avons présenté un système permettant d'estimer de manière automatique la similarité entre des voix de doublage. Nous avons réalisé ce système dans l'intention de vérifier les possibilités d'automatisation de l'appariement de voix originales et de voix de doublage dans un cadre plus général. L'approche proposée dans ces travaux préliminaires étend l'utilisation du couple i-vecteur/PLDA à la comparaison de voix actées sur une tâche nouvelle qui plus est dans des langues différentes.

Le cœur de cette approche réside dans l'apprentissage d'une fonction de *mapping* ayant pour but de faire le lien entre la voix d'un personnage dans la langue originale et la voix de ce même personnage dans la langue cible. Cette approche tient compte pour cela des choix réalisés par l'opérateur de casting vocal, et cherche donc à neutraliser la variabilité qui différencie ces deux langues.

Les travaux que nous venons de présenter ont fait l'objet d'une publication dans une conférence internationale (Gresse, Rouvier et al. 2017). Ces résultats initiaux nous ont permis de vérifier l'existence d'une information dans le signal de parole caractéristique du personnage. De plus, nous avons

mis en évidence un biais lié au message qui est véhiculé par le segment de voix. Nous avons supposé qu'il s'agit du contenu linguistique. Ce dernier faisant office de référence lors de l'apprentissage pour l'approximation de l'appariement VO-VF, de notre fonction de *mapping*, nous avons ainsi proposé une méthode permettant de le neutraliser.

De plus, si nous utilisons les mêmes espaces de variabilité pour l'extraction des *i*-vecteurs (utilisés pour l'apprentissage de l'appariement VO-VF), nous observons de meilleurs résultats. En effet, il semblerait que le passage d'un espace de variabilité totale (espace des *i*-vecteurs) à un autre, que nous rendons possible grâce à la transformation apprise, n'apporte pas de meilleures performances. Il se pourrait donc que la compensation de la langue ne soit pas d'une grande utilité ici. Ce qui pourrait s'expliquer par le fait que la langue française et la langue anglaise ne diffèrent par tant que ça du point de vue de l'information à court terme que nous extrayons.

Enfin, l'expérience contrastive nous a permis d'observer une distinction au niveau des scores obtenus sur des tests mettant en jeu deux segments provenant d'un même personnage, par rapport à ceux appartenant à deux personnages différents. Cette différence nous conforte dans l'idée qu'il existe dans la voix une information caractéristique du choix artistique réalisé par l'opérateur de casting.

Chapitre 6

Un cadre méthodologique pour évaluer l'appariement des voix de doublage

Sommaire

6.1 Introduction

Dans le chapitre précédent, nous avons utilisé un système de vérification automatique du locuteur pour estimer la similarité entre des voix de doublage. Ces premiers résultats nous ont permis d'une part de confirmer notre hypothèse sur la présence d'une information caractéristique de la dimension personnage, et d'autre part, de valider l'approche *i*-vecteur en tant que représentation des données en entrée. De plus, nous avons observé les limites de l'approche PLDA avec l'utilisation de notre corpus de données et par conséquent. La PLDA ne nous permet pas de nous affranchir complètement du locuteur. Enfin, le cadre d'évaluation utilisé dans le domaine de la vérification du locuteur n'est pas adapté à notre tâche. Nous avons proposé une méthode d'évaluation spécifique reposant sur l'utilisation des k meilleurs scores, mais ce degré de liberté n'est pas suffisamment objectif selon nous.

Nous avons donc choisi de travailler sur la mise en place d'un cadre méthodologique dédié à la mesure de la similarité de voix. Le protocole que nous proposons est spécifique à notre tâche. Il nous permet d'automatiser l'appariement des voix de doublage en approximant les choix de l'opérateur de casting , mais surtout il nous offre un cadre d'évaluation strict. Nous avons conçu ce protocole dans le but minimiser l'influence de biais liés à notre corpus.

Nous avons choisi de nous appuyer sur une architecture neuronale fondée sur les réseaux siamois. Nous pensons que cette architecture est bien adaptée à notre problème, car elle est généralement utilisée pour apprendre une mesure de similarité entre des éléments qui partagent une notion abstraite de similarité. L'utilisation d'un grand nombre de paires de voix et d'une méthode d'apprentissage automatique doit nous permettre de modéliser un espace de représentation capable de démêler l'information permettant la discrimination des paires de voix similaires et des paires non-similaires.

Nous supposons pouvoir extraire l'information caractéristique de cette notion de similarité à partir de la relation qui existe entre des paires de voix, ces paires reflétant les choix de l'opérateur de casting vocal. Nous pensons que ces choix reposent sur une combinaison d'éléments ou de facteurs explicables grâce à l'analyse de la voix. En d'autres termes, nous supposons qu'il ne s'agit pas d'un choix purement aléatoire, mais qu'il est fondé sur des mécanismes de perception/réception de la voix. En fonction du posi-

tionnement de la voix évalué par rapport aux représentations mentales de l'opérateur, des traits différents y sont attribués. De plus, nous supposons l'existence de certaines caractéristiques vocales associées aux représentations stéréotypées et qui peuvent donc se retrouver dans les différentes voix des personnages.

Ce chapitre sera donc l'occasion pour nous de présenter le protocole expérimental que nous avons élaboré et son application à l'apprentissage de cette notion de similarité des voix de doublage. Comme nous le verrons, un point critique du protocole proposé concerne l'évaluation sur un jeu de données limité en taille. Néanmoins, la méthodologie que nous avons mis en place, bien que spécifique à notre corpus, est transposable à l'utilisation d'un plus gros corpus de données.

Nous commençons par introduire notre approche dans la section 6.2 puis nous présentons en détails le protocole que nous avons élaboré dans la section 6.3. Les expériences réalisées ainsi que les résultats sont décrits dans la section 6.4. Enfin, nous livrons nos conclusions dans la section 6.5.

6.2 Mesurer la similarité entre des voix

Nous avons présenté, dans le chapitre précédent, une vue générale du système de similarité tel que nous l'avons imaginé (voir la figure 5.1). Ce système prend alors en entrée une paire de voix. La première voix fait référence à une voix dans la langue source (vert), la deuxième fait quant à elle référence à une voix dans la langue cible (rouge). De la même façon que pour les travaux présentés dans le chapitre précédent, nous nous appuyons sur l'approche *i*-vecteur pour la représentation des données en entrées (voir la sous-section 5.2.1). De plus, les expériences précédentes nous ont montré qu'il n'est pas utile d'utiliser un espace de variabilité totale spécifique à la langue.

En guise de sortie, le système nous fournit un score correspondant à la similarité, ou plutôt au degré d'appariement entre ces deux voix. En d'autres termes, ce score représente la capacité de la voix cible à doubler la voix source. Le modèle de similarité est par conséquent entraîné à partir d'un ensemble de voix provenant de deux langues différentes. Nous avions comparé le modèle de similarité à une « boîte noire ». Dans le chapitre précédent, nous avons utilisé la PLDA pour estimer la similarité, approche éprouvée en reconnaissance du locuteur. Ici, c'est une architecture neuronale qui oc-

cupe cette place, dont la proposition semble, a priori, plus adaptée à notre tâche.

6.2.1 Mesure de similarité

L'apprentissage d'une mesure de similarité revient à apprendre un espace de représentation dédié à une tâche particulière, tout en étant guidé par une métrique qui fait référence à une fonction de distance. Le terme similarité est utilisé par abus de langage. Ce problème devient difficile lorsqu'il est appliqué à des espaces de grande dimensionnalité.

Différentes méthodes existent pour apprendre une mesure de similarité. La question qui supporte ces travaux est : comment estimer la similarité en considérant une paire d'exemples (dans notre cas, des voix) ? En effet, suivant ce qui est mesuré, une similarité au niveau de l'accent ou du genre des locuteurs, ce ne sont pas les mêmes caractéristiques qui sont considérées. Dans le premier cas, par exemple, ce sont les aspects linguistiques et phonétiques de la parole qui expliquent le mieux cette similarité. Dans l'autre cas, le timbre et la hauteur de voix seront plus porteurs d'information que d'autres caractéristiques.

L'apprentissage d'une mesure de similarité peut se faire de manière supervisée avec des approches linéaires utilisant des métriques telles que la distance de Mahalanobis ou des approches non-linéaires utilisant soit le *kernel trick*, soit des réseaux de neurones profonds. Le but n'est pas ici de faire état de ces différentes méthodes (pour plus de détails se référer à Kulis et al. 2013).

L'apprentissage d'une mesure de similarité grâce à des réseaux de neurones est possible notamment avec l'utilisation d'un Réseaux de Neurones Siamois (Siamese Neural Networks) (SNN). L'architecture siamoise a initialement été proposés par Bromley et al. 1994 pour de la vérification de signature dans une tâche de classification binaire. La mesure de similarité entre deux exemples permet, en fonction d'un seuil défini, de dire si il s'agit de la même signature ou non. Les SNN sont devenus populaires depuis leur application à la vérification des visages (Chopra et al. 2005). De plus les travaux réalisés par Koch et al. 2015 ont montré, dans le cadre d'une classification multi-classes, que les SNN peuvent amener à une bonne généralisation.

Pour le traitement de la parole, les SNN se sont aussi montrés efficaces,

notamment dans les travaux de Zeghidour (Zeghidour, Synnaeve, Usunier et al. 2016 ; Zeghidour, Synnaeve, Versteegh et al. 2016). Ce dernier propose une méthode d'apprentissage joint pour démêler l'information du locuteur et l'information phonétique dans un espace de représentation dédié. Un des avantages de cette architecture réside dans la faible supervision requise pour l'apprentissage, puisqu'il suffit de savoir si deux exemples sont similaires ou différents. De manière générale, les SNN permettent d'apprendre, à partir d'exemples et d'annotations simples, une mesure de similarité dans différents domaines.

6.2.2 L'architecture siamoise

Les SNN consistent en deux réseaux de neurones identiques partageant les mêmes paramètres. Ces réseaux prennent deux entrées distinctes et ils sont joints dans une couche commune. Cette dernière calcule une métrique à partir de la représentation de haut niveau donnée en sortie de chacun des deux réseaux de neurones. Il s'agit généralement de réseaux multi-couches qui réalisent une transformation non-linéaire depuis l'espace des caractéristiques d'entrées vers un espace latent, appelé espace de représentation ou *embedding*. Un exemple d'architecture est illustré dans la figure 6.1. La sortie peut prendre deux formes différentes. Il peut s'agir soit d'une valeur réelle indiquant à quel point les deux entrées sont similaires [1], soit d'une variable binaire, 0 quand la distance entre les exemples de la paire est minimale (similaires) ou 1 quand la distance est maximale (différentes).

Cette architecture particulière permet l'apprentissage d'une mesure de similarité à partir des relations paires à paires qui existent entre des entrées partageant une notion de similarité abstraite. Les deux propriétés clés de cette architecture selon Koch (Koch et al. 2015) sont :

— Le partage des paramètres assure la consistance des prédictions du fait que les deux réseaux modélisent la même fonction.
— La mesure de similarité est symétrique. L'ordre des exemples qui composent la paire n'impacte pas la mesure.

L'apprentissage d'un SNN se fait de manière classique, c'est-à-dire en minimisant une fonction de perte. La mise à jour des paramètres partagés se fait aussi par rétro-propagation en accumulant ou en moyennant l'erreur dans les deux réseaux.

1. Il s'agit en réalité d'une dissimilarité étant donné que l'on se réfère à une distance.

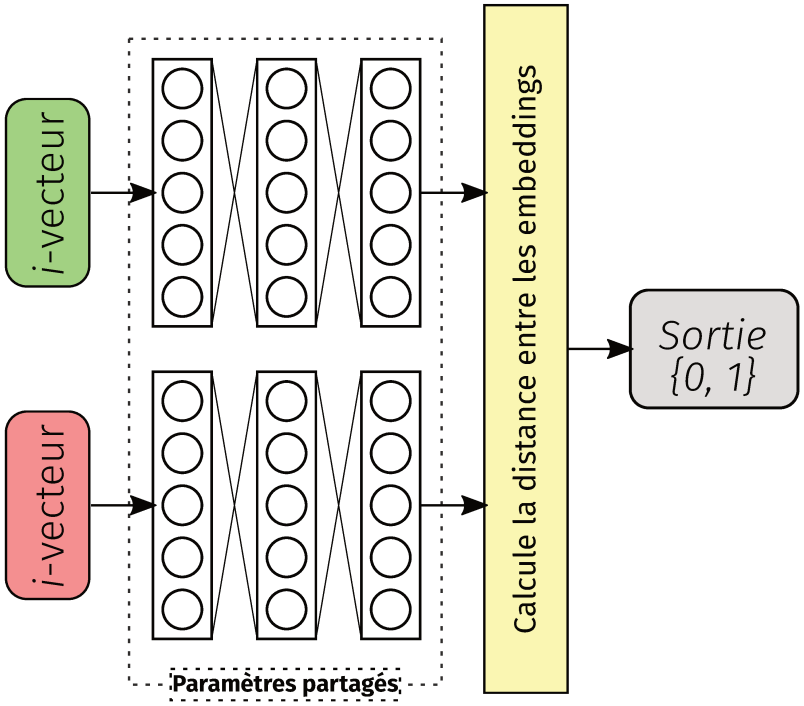

Figure 6.1 – Réseaux de neurones siamois prenant deux représentations *i*-vecteurs en entrées.

La fonction contrastive

En considérant deux entrées similaires x_1 et x_2, il faut faire en sorte que la distance $D_W(x_1, x_2)$ soit la plus petite possible de sorte à minimiser la fonction de coût notée L. N'importe quelle mesure de distance peut être utilisée pour calculer D, mais nous avons choisi d'utiliser la même distance que Koch et al. 2015, soit une distance euclidienne :

$$D_W = \| G_W(x_1) - G_W(x_2) \|_2 \tag{6.1}$$

où G correspond à la fonction calculée par les deux réseaux paramétrés par W.

Toutefois, avec ce seul terme, le modèle apprendrait à placer chaque entrée sur un point unique ce qui conduirait à prédire une sortie constante. Par conséquent, toutes les paires seraient catégorisées comme similaires. Pour éviter ce problème, un deuxième terme doit être introduit au calcul de la fonction de coût pour faire en sorte que la distance soit grande dans le cas où x_1 et x_2 sont des entrées différentes. La fonction de coût s'écrit sous la forme suivante :

$$L_{\text{totale}} = \sum L_{\text{similaires}} + \sum L_{\text{différentes}}.$$

La fonction de coût initialement proposée par Chopra et al. 2005, plus connue sous l'appellation *contrastive loss*, est définie comme suit :

$$\begin{aligned} L(x_1, x_2, Y) = (1 - Y) \times D_W(x_1, x_2)^2 \\ + Y \times max\{0, \alpha - D_W(x_1, x_2)\}^2 \end{aligned} \tag{6.2}$$

avec Y une variable binaire telle que $Y = 0$ lorsque les entrées sont similaires, et $Y = 1$ lorsqu'elles sont différentes. La constante α strictement positive est définie de sorte que $D_W(x_1,x_2) + m < D_W(x'_1, x_2)$ où x_1 et x_2 sont des entrées similaires et x'_1 une entrée différente. Cette constante α est assimilée à une marge.

D'autres fonctions de coût ont été proposées, notamment la fonction triplet, décrite en annexe B.1. De plus, jusque-là, nous avons mentionné uniquement la distance euclidienne. Toutefois il n'est pas exclu d'utiliser des métriques différentes (voir annexeB.2).

Encart 6.1 : Stratégie de création des paires

Un des problèmes majeurs des SNN concerne la sélection des paires données en entrée. Suivant le nombre d'exemples disponibles dans le corpus de données, l'ensemble des paires composées à partir de toutes les combinaisons possibles d'exemples peut très rapidement devenir immense. De plus, au-delà de deux catégories d'exemples représentées dans les données, le nombre des paires *nontarget* surpasse largement celui des paires *target*. Dans Riad et al. 2018, les auteurs ont montré que la présence d'un trop grand nombre d'entre-elles peut avoir de lourdes conséquences sur l'apprentissage et mener à des résultats très variables. Ils partent du principe qu'un nombre important de paires n'apporte que peu d'information et de variabilité pour la construction d'un espace de représentation robuste. Limiter le nombre de paires peut alors être bénéfique, mais encore faut-il disposer d'un moyen de filtrer les paires en fonction de leur capacité informationnelle.

6.3 Cadre méthodologique

Cette partie décrit la méthodologie que nous avons mise en place. Elle est essentielle à la bonne compréhension de nos travaux, nous allons donc tâcher de rendre sa description la plus limpide possible. L'apprentissage (ainsi que l'évaluation) se fait à partir de paires de segments de voix. Nous décrivons la méthodologie de construction des paires et nous détaillons également les contraintes que nous avons mises en place afin d'éviter certains biais à l'apprentissage, mais surtout à l'évaluation.

Notre tâche consiste à appairer de manière automatique les voix de la VO (anglaises) avec les voix de la VF (françaises). Nous évaluons la capacité du système à associer les voix qui ont été sélectionnées, par la DA, pour doubler le même personnage. Toutes les voix que nous utilisons proviennent du jeu vidéo appelé *Mass Effect 3*. Ce corpus a été présenté en détails dans la sous-section 5.3.1.

6.3.1 Contrôle des biais

Nous détaillons ici les méthodes utilisées pour contrôler les différents biais liés aux données et à la création des paires utilisées pour l'apprentissage et l'évaluation du modèle de similarité.

Un acteur peut en cacher un autre

Un personnage est doublé par deux acteurs, l'acteur original qui parle en anglais et l'acteur français. Il est essentiel de pouvoir tester notre système avec des voix sur lesquelles le système n'aura pas été entraîné, afin d'évaluer sa capacité à généraliser. Le corpus de données, issu du domaine du jeu vidéo, amène un problème lié aux acteurs. En effet, un acteur peut doubler plusieurs personnages (généralement un acteur double un des personnages principaux et potentiellement quelques personnages secondaires).

Pour pouvoir attribuer des personnages différents à l'ensemble d'apprentissage et d'autres à l'évaluation, nous devons nous assurer qu'aucun acteur ne double plusieurs personnages, parmi les voix que nous utilisons. Ainsi, en évitant toute contamination de l'ensemble d'évaluation, nous sommes bien disposés à évaluer la capacité du système à discriminer le personnage et non pas le locuteur.

Comme nous l'avons dit dans la sous-section 5.3.1, nous avons 31 personnages pour 62 acteurs dans notre corpus. Ces personnages là sont, en réalité, les seuls personnages du jeu pour lesquels nous avons la garantie qu'ils satisfont cette contrainte.

Fréquence d'apparition

Le corpus dont nous disposons montre un très fort déséquilibre au niveau du nombre de segments par personnage. Cela est dû à l'importance relative de chaque personnage dans le jeu. Nous uniformisons donc la distribution des segments par tirage aléatoire, ce faisant, nous évitons d'avantager (ou de désavantager) un personnage. Cependant, cette opération doit tenir compte de la correspondance des segments que nous avons pris soin d'extraire auparavant. En effet, en supposant que nous souhaitions ramener le nombre de segments par personnage à 100 pour chaque langue. Il est important de veiller à sélectionner les segments par paires (c'est à dire les segments équivalents des langues l_1 et l_2) et non pas en tirant au hasard 100 segments pour l_1 et 100 autres pour l_2. La raison sous-jacente (au-delà du fait que la concordance linguistique anglais-français entre les segments est conservée) concerne la présence de filtres [2] appliqués aux voix des personnages dans certaines situations. Ainsi, nous évitons de nous retrouver dans une situation où le personnage dans la VO possède une voix naturelle tandis que ce même personnage dans la VF a une voix filtrée.

Durée des segments

Les interactions vocales des personnages n'ont pas de durée fixe. Ainsi, il peut s'agir d'une phrase complète ou d'un morceau de phrase comme d'une simple exclamation. Il est donc possible de tomber sur un fichier de très courte durée (voir la figure 5.4). Comme nous le savons, le manque d'information peut fortement altérer la qualité de la représentation en entrée (ici, i-vecteurs). Nous avons donc, par principe de précaution, retiré au préalable tous les segments du corpus ayant une durée inférieure à une seconde.

2. Principalement pour ajouter un effet radio sur la voix

Contrôle du genre

Certains personnages sont des personnages féminins, d'autres masculins. L'équilibre n'est pas parfait (13 femmes, 18 hommes). Toutefois, l'utilisation d'un système *i*-vecteur appris indépendamment du genre nous permet de réaliser des comparaisons homme-femme.

Nous pouvons facilement supposer que notre système aura peu de mal à faire la distinction entre des voix masculines et féminines. Nous aimerions qu'il soit capable de distinguer des personnages différents malgré des voix d'un genre identique. Pour ce faire, chaque paire que nous présentons au système est mono-genrée, c'est-à-dire que nous créons des paires de segments de voix de même genre uniquement. Les paires *target* (de même personnage) satisfont cette contrainte naturellement, nous ignorons donc les paires *nontarget* (de personnages différents) composées de personnages de genres différents. Cette contrainte est nécessaire, car nous associons une seule voix par langue à un personnage. En supposant que nous puissions associer plusieurs voix de genres différents à un personnage, cette contrainte ne serait pas nécessaire.

Le contenu linguistique

Le message linguistique véhiculé dans le segment de voix représente un biais potentiel sur lequel nous pouvons agir. En effet, nous avons montré dans la section 5.5 l'impact de la composante linguistique sur les performances du système.

Pour neutraliser ce biais, nous évitons toute paire composée par un segment anglais et son équivalent en français en mélangeant aléatoirement les ensembles de segments de chaque personnage dans les deux langues, avant de les appairer, là aussi, aléatoirement.

Encart 6.2 : Le biais de l'imitation

Nous trouvons une explication à ce phénomène dans le processus même de développement et de doublage des jeux vidéo en général. Les dialogues *in-game* sont pour la VO enregistrés généralement en pré-synchronisation. C'est-à-dire que le rendu visuel est synchronisé avec la voix a posteriori. À l'inverse, le doubleur de la VF doit synchroniser sa voix en s'appuyant sur les images, et donc sur le rythme imposé par la VO. Exception faite des scènes cinématiques, dont le rendu est effectué en amont du doublage et où les doubleurs VO comme VF sont tous deux contraints de suivre la bande rythmographique. Nous pensons que cette « imitation » et le fait que le contenu linguistique soit le même, sont à l'origine de ce biais.

6.3.2 Validation croisée

Nous n'avons qu'un nombre limité de données qui satisfont les contraintes que nous venons d'énoncer. Nous avons dû fixer le nombre de segments par personnage et par langue à 90. Ce choix nous permet d'avoir en définitive 16 personnages qui respectent nos contraintes. Diviser cet ensemble de personnages en deux sous-ensembles (apprentissage / évaluation) semble poser un problème en termes de sélection de données, compte tenu de leur faible nombre, quand bien même le découpage serait effectué de manière aléatoire. En effet, rien ne garantit que l'ensemble d'évaluation soit représentatif de notre corpus.

Pour pallier ce biais, nous proposons de faire une validation croisée. L'utilisation de ce procédé nous permet d'évaluer tour à tour, sur un ensemble réduit de personnages (les autres servant à l'apprentissage du modèle), jusqu'à ce que nous ayons couvert l'ensemble des personnages. Nous illustrons notre découpage dans la figure 6.2. Ce découpage est conservé à l'identique tout au long de nos expériences pour garder une cohérence dans nos résultats et dans notre analyse.

6.3.3 Apprentissage et évaluation par les paires

Comme nous l'avons déjà présenté, les paires composées de segments correspondant au même personnage sont dénotées *target*, les autres paires sont dites *nontarget*. Cette annotation est la seule information utilisée en

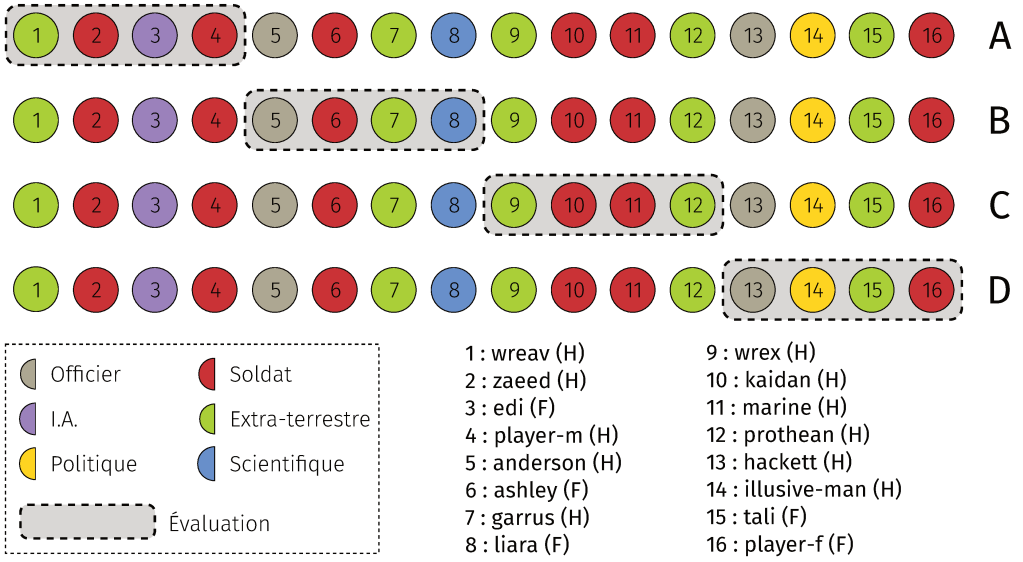

Figure 6.2 – Découpage en 4 ensembles d'évaluation notés A, B, C et D. La liste des personnages $1, 2, ..., 16$ est auparavant mélangée. Les étiquettes (soldat, officier, extra-terrestre...) sont attribuées aux personnages selon notre propre interprétation des voix et ne font en aucun cas office de supervision.

guise de supervision de l'apprentissage. Ainsi, notre tâche consiste en une simple classification binaire d'un ensemble de paires de voix.

En considérant l'ensemble de toutes les paires possibles, le nombre de paires *nontarget* dépasse largement le nombre de paires *target*. Nous ramenons leur nombre à égalité de manière à avoir une équiprobabilité a priori sur nos deux classes.

Les 4 plis de notre validation croisée, notés A, B, C et D, contiennent chacun 32 400 paires *target* et autant de paires *nontarget*, toutes destinées à l'évaluation (tirées aléatoirement en prenant en compte les contraintes que nous avons citées plus haut). Nous appliquons le même procédé afin de créer l'ensemble des paires d'apprentissage dans les 4 cas, excepté que nous les scindons en deux sous-ensembles (apprentissage : 80 % et développement : 20 %). Nous faisons cela en veillant à garder l'équilibre en termes de paires *target-nontarget* ainsi qu'en termes de fréquence d'apparition des personnages dans les paires. Au total, nous avons 194 400 paires d'apprentissage pour chaque pli.

En définitive, nous avons 4 cas d'apprentissage/évaluation différents, chacun étant constitué d'un ensemble de paires d'entraînement, un ensemble de développement (pour la validation de l'apprentissage) et un ensemble de test pour l'évaluation finale du modèle.

6.3.4 Méthode d'évaluation

Dans l'objectif d'évaluer la fiabilité de la mesure de similarité que nous apprenons, nous nous appuyons sur les scores obtenus avec les paires de l'ensemble de test. La mesure utilisée ici correspond à la distance euclidienne. Elle est calculée entre les segments qui composent la paire, projetés dans l'espace de représentation construit par le modèle. Compte tenu de la nature des deux catégories de paires (*target*/*nontarget*), nous supposons qu'une distinction significative entre leurs scores respectifs doit être observée. D'où l'utilisation d'un test statistique d'hypothèse, le t-test, dit « test de Student ». Nous comparons donc, intuitivement, les moyennes des scores obtenus dans les deux classes. Il s'agit d'un test binomial où l'hypothèse nulle suppose que les moyennes des scores des deux groupes sont égales.

En plus du test statistique, nous évaluons cette tâche de classification binaire, c'est-à-dire en nous référant aux métriques classiques, telles que le taux de réussite du système. Pour cela, nous nous appuyons sur les distances calculées sur chaque paire et déterminons le seuil d'EER. Nous utilisons ce seuil pour prédire la classe (*target* si la distance mesurée est strictement inférieure au seuil, *nontarget* sinon). Le taux de réussite est donc mesurée à partir de ces prédictions.

6.4 Expériences

Nous présentons ici les expériences que nous avons réalisées à partir du protocole proposé dans la section précédente. Nous détaillons les traitements effectués sur les données ainsi que les différents modèles qui sont entraînés. Enfin, nous présentons les résultats et évaluons l'apport de l'architecture siamoise sur cette tâche en comparaison des architectures neuronales classiques.

6.4.1 Extraction des séquences

Nous transformons le signal acoustique en vecteurs de caractéristiques en suivant la même méthodologie détaillée dans la sous-section 5.3.2, à une exception. Nous apprenons ici un extracteur i-vecteur indépendant du langage en utilisant pour l'anglais les corpus NIST SRE 2004, 2005 et 2006 et pour le français, les corpus ESTER (1 et 2), EPAC, ETAPE et REPERE. Nous

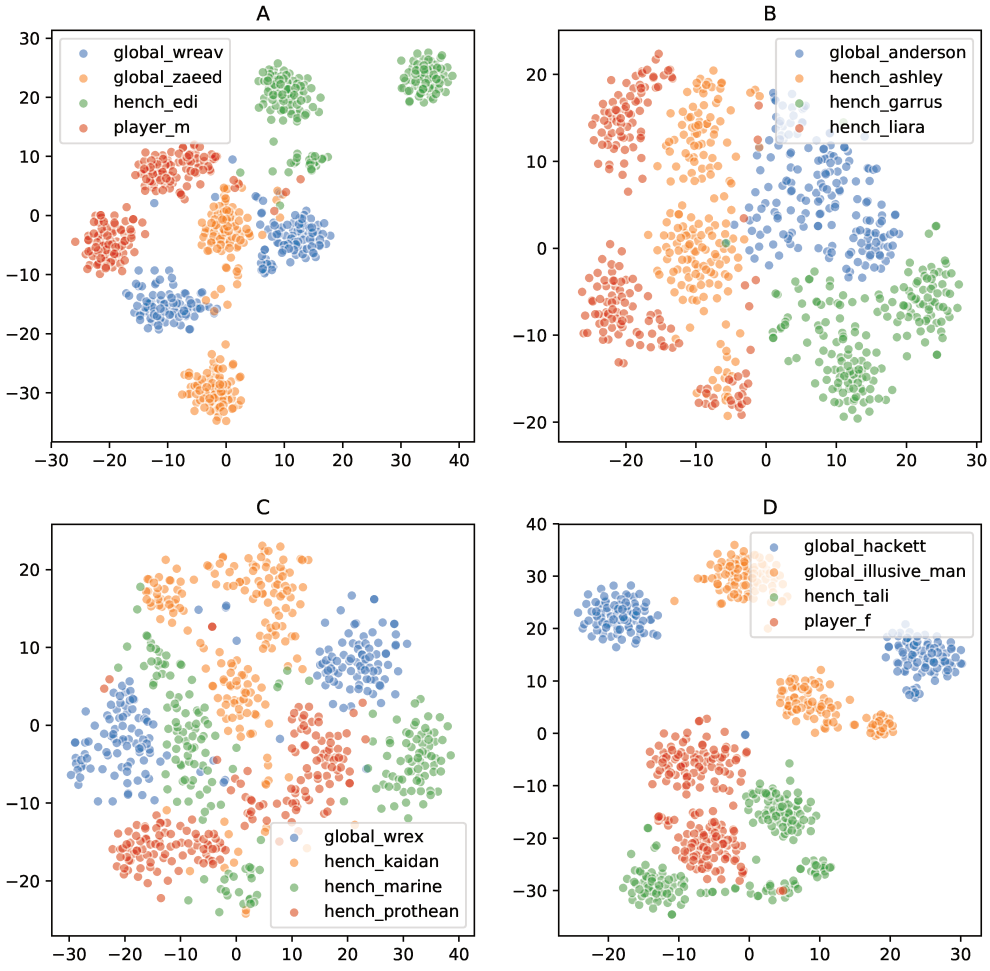

Figure 6.3 – Représentation dans l'espace i-vecteur des personnages pour les cas A, B, C et D. Illustration obtenue avec t-SNE.

extrayons pour chaque segment de notre corpus les i-vecteurs correspondants. Dans la figure 6.3, nous proposons un aperçu de la projection des différents segments de voix des personnages dans chacun des 4 cas de test. Cette illustration montre bien pour chaque personnage, la distinction entre l'acteur de la VO et celui de la VF.

6.4.2 Définition du modèle

Notre architecture siamoise s'appuie sur deux réseaux convolutifs. Nous utilisons la bibliothèque Keras (Chollet et al. 2015) pour la mise en place des réseaux et leur entraînement. Les deux réseaux sont construits suivent la structure définie dans la table 6.1. Les hyper-paramètres du réseaux ont été fixés de manière à obtenir les meilleures performances sur le développement.

Input 400×1
Conv1D(filters-32, size-10)-BatchNorm-LeakyReLU MaxPooling + Dropout(0.25)
Conv1D(filters-64, size-7)-BatchNorm-LeakyReLU MaxPooling + Dropout(0.25)
Conv1D(filters-128, size-4)-BatchNorm-LeakyReLU MaxPooling + Dropout(0.25)
Conv1D(filters-256, size-4)-BatchNorm-LeakyReLU MaxPooling + Dropout(0.25)
Flatten Dense(2048)-BatchNorm-LeakyReLU + Dropout(0.5)
Dense(256)-BatchNorm-Tanh

TABLE 6.1 – Structure des réseaux convolutifs utilisés dans le SNN suivant la nomenclature de Keras (CHOLLET et al. 2015).

Les réseaux de neurones siamois sont particulièrement difficiles à entraîner. L'hyper-paramétrisation a un impact considérable sur le temps nécessaire pour entraîner le modèle ainsi que sur sa robustesse. Nous ne sommes jamais à l'abri de trouver une configuration pour laquelle nous obtenons de très bons résultats avec un modèle même des plus basiques aux premiers abords. Surtout après plusieurs mois (voire années) d'acharnement.

Initialisation des paramètres : Nous utilisons une initialisation normale de Xavier (GLOROT et al. 2010) pour les poids du réseau dont l'écart-type est défini en fonction du nombre de poids entrants de chaque neurone. Pour l'initialisation des biais, nous utilisons une distribution normale avec $\mu = 0$ et $\sigma = 0.1$.

Optimisation : L'objectif de minimisation de l'optimiseur est fixé par la *contrastive loss* (voir la section 6.2.2) et est combiné à la rétro-propagation pour la mise à jour des paramètres. Nous utilisons pour cela l'optimiseur appelé Adadelta (ZEILER 2012) qui adapte le pas d'apprentissage au cours du temps relativement à la fréquence de mise à jour d'un paramètre. Cette mise à jour est additive du fait des deux réseaux. Il est aussi possible de la moyenner, mais nous observons une convergence plus lente dans ce cas. De plus, pour limiter le sur-apprentissage, nous avons utilisé une couche de *Dropout* (SRIVASTAVA et al. 2014) en guise de régularisation entre chaque convolution.

	A	*B*	*C*	*D*
Développement	0,72	0,71	0,70	0,71
Test	0,55	0,59	0,62	0,50

TABLE 6.2 – Taux de réussite des prédiction du modèle d'appariement VO-VF

Apprentissage : Nous avons fixé la taille du *mini-batch* à 128 paires de voix. Le processus d'entraînement est réalisé durant 50 époques où nous surveillons la justesse des prédictions effectuées sur les exemples de l'ensemble de développement de façon à éviter un sur-apprentissage. Les paramètres du modèle sont enregistrés et nous conservons l'état dans lequel le modèle obtient les meilleures performances sur le corpus de développement.

Encart 6.3 : Curriculum learning

Nous avons mentionné auparavant la nécessité d'effectuer une sélection des paires. À ce sujet, nous avons étudié la possibilité d'utiliser une méthode de *curriculum learning*[a] afin d'établir une sélection des paires par niveau de difficultés et ainsi réaliser un apprentissage progressif (BENGIO, LOURADOUR et al. 2009). Tout l'enjeu est donc de définir la difficulté propre à chaque paire.
Pour cela, nous avons proposé de mesurer cette difficulté en utilisant une méthode de classification non-supervisée (k-moyennes). Cette dernière est réalisée sur la base des segments de voix projetés dans l'espace i-vecteur. Nous mesurons la difficulté à partir du coefficient de silhouette obtenues sur les différents exemples classifiés. Cette approche n'a toutefois pas montrée d'amélioration sur notre tâche. La mesure de la difficulté d'une paire est un problème difficile que nous n'avons pas exploré plus en profondeur. Néanmoins c'est une piste qui reste envisageable.

a. Le *curriculum learning* est une technique d'apprentissage qui consiste à commencer par les exemples les plus simples pour ainsi augmenter la difficulté de la tâche graduellement.

6.4.3 Résultats

Les résultats obtenus pour la tâche de classification des paires sont présentés dans la table 6.2. Comme nous pouvons le voir, le taux de réussite du modèle se situe au-dessus de 70 % sur le corpus de développement. Pour

rappel, ce dernier est composé de paires de segments de voix prononcés par des personnages de l'ensemble d'apprentissage, les segments sont quant à eux inédits. Ces performances montrent que le système est capable de reconnaître des personnages déjà rencontrés, indépendamment du contenu linguistique véhiculé.

En ce qui concerne les performances obtenues sur le corpus de test, nous observons, en termes de prédictions, des performances moins bonnes. Toutefois, cela n'est pas étonnant puisque ce corpus nous permet d'évaluer les capacités du système à généraliser sur des voix nouvelles. Par rapport au corpus de développement, la tâche est ici beaucoup plus difficile, puisqu'aucune des voix, donc, aucun des personnages, du corpus de test n'est utilisée dans l'apprentissage. Compte tenu de cette difficulté, nous notons que le système s'en sort assez bien dans 3 plis sur 4 –pour une classification binaire avec données équilibrées, la probabilité a priori est de 0,5. Dans les trois cas *A*, *B* et *C*, les prédictions sont ainsi meilleures que le hasard. Les prédictions réalisées sur le cas d'évaluation noté *D* sont, en termes de performance, équivalentes avec le hasard.

Encart 6.4 : Remarque sur les écarts de performance

Dans les résultats, nous pouvons observer un écart important en termes de performance sur certains cas d'évaluation (par exemple le cas *D*). Pour expliquer cet écart, il semble compliqué de mettre en cause le corpus d'apprentissage, puisqu'une part importante des données est commune avec les autres cas. Il est néanmoins possible de s'appuyer sur les personnages impliqués dans l'évaluation. Dans ce cas précis, les quatre personnages impliqués dans l'évaluation sont tous des personnages masculins ayant des voix centrées sur l'archétype du guerrier. Sachant cela, nous comprenons mieux pourquoi le système a eu du mal à les différencier.
Ici, la sélection des paires *nontarget* pose peut-être un problème. En effet, nous avons veillé à ce que la fréquence d'apparition de chaque personnage soit équilibrée. Toutefois, il est possible qu'un plus grand nombre de contre-exemples de voix « proches » soit nécessaire pour que le système apprenne à les différencier.

Nous avons évalué la mesure de similarité apprise à partir des paires de segments de voix de personnages, en nous basant sur sa capacité à distinguer, à travers de nouvelles voix, les paires *target* (même personnage) des

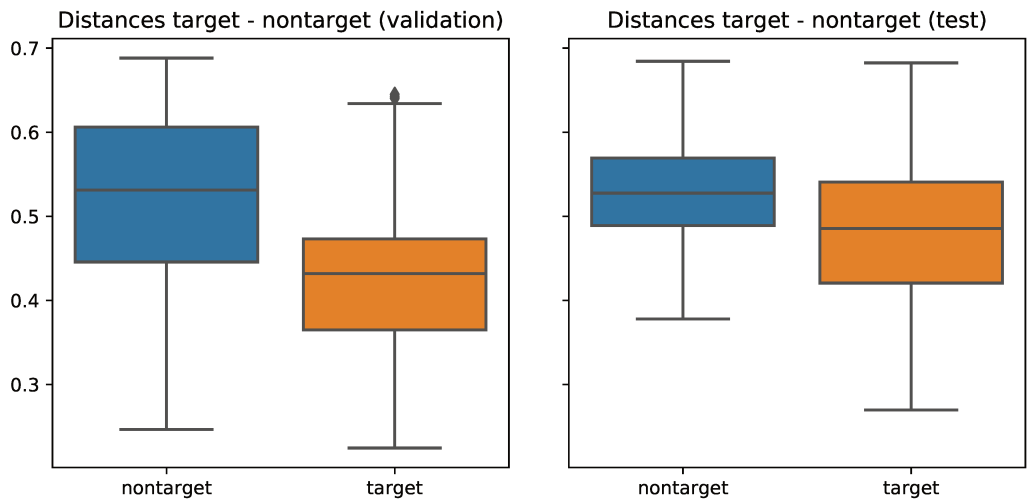

Figure 6.4 – Illustration en boîte à moustaches des distances mesurées entre les paires *target* (bleue) et *nontarget* (orange) dans le cas d'évaluation *C*. À gauche, les mesures faites sur le corpus de développement et à droite celles effectuées sur le corpus de test.

paires *nontarget* (personnages différents). Les résultats du test de *Student* (test t) sont rapportés dans la table 6.3.

La statistique du test t correspond à un ratio entre la variance *inter-* et la variance *intra*-classe et permet d'affirmer qu'il existe (ou non) une différence significative entre les moyennes des scores des deux classes. Nous observons les valeurs les plus hautes sur le corpus de test, ce qui nous laisse suggérer que la différence entre le score des paires *target* et *nontarget* est plus importante sur le corpus de test que sur le corpus de développement. Cependant, ces résultats à eux seuls ne nous autorisent pas à faire cette affirmation, bien qu'ils soient prometteurs. Les valeurs-p associées aux différents tests sont toutes inférieures au seuil (arbitrairement fixé à 0,01) de rejet de l'hypothèse nulle, à l'exception de celui effectué sur le corpus de test dans le cas *D*. Ceci nous permet donc d'affirmer qu'il existe une différence significative entre les scores des paires *target* et *nontarget* pour les cas *A*, *B* et *C* uniquement. La figure 6.4 illustre bien cette différence.

Attention néanmoins à l'interprétation : ici, nous observons une dissimilarité et non pas une similarité. Plus le score est haut, plus la distance mesurée entre les segments de la paire examinée est grande et donc au moins les voix sont jugées similaires.

Certains des personnages sont plus faciles à reconnaître que d'autres. Il est alors intéressant de voir quels sont les personnages qui amènent le plus

	A	*B*	*C*	*D*
Développement	44,90	52,77	45,18	44,46
Test	52,18	77,99	86,17	1,87

TABLE 6.3 – Valeurs de la statistique du test de *Student* pour la discrimination des paires *target* et *nontarget*

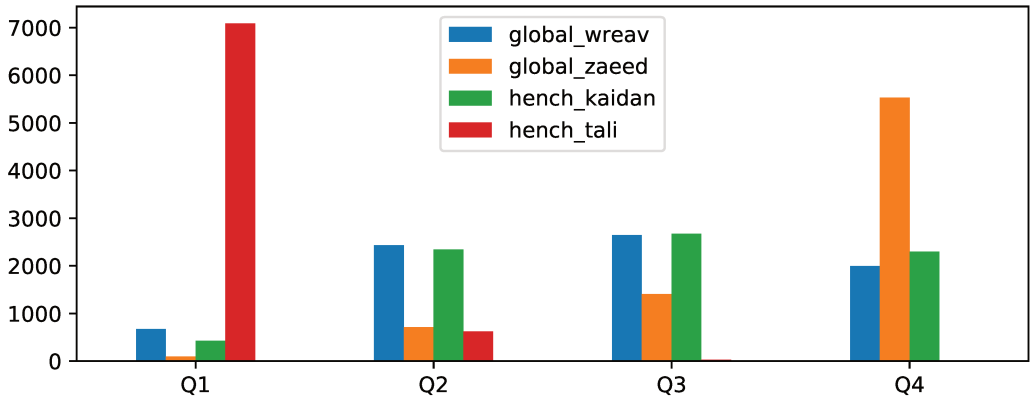

Figure 6.5 – Occurrence des personnages (impliqués dans l'évaluation *C*) dans les différents quartiles calculés sur les erreurs de prédictions.

souvent des erreurs de prédictions. L'erreur de prédiction pour une paire notée ϵ_i est calculée de la manière suivante :

$$\epsilon_i = |Y_i - score_i|. \tag{6.3}$$

Nous considérons ici l'écart entre le score prédit et la valeur théorique Y_i pour chaque paire évaluée. Nous illustrons dans la figure 6.5 le nombre d'apparitions des personnages dans les différents quartiles calculé sur les erreurs de prédictions en ce qui concerne le cas d'évaluation *C*. Ce dernier impliquant les voix d'un personnage féminin (rouge) et les voix de trois personnages masculins. Nous observons une présence très marquée du personnage féminin dans le premier quartile, ce qui correspond à une erreur très faible pour ce personnage. Inversement, nous constatons la présence très marquée d'un des personnages masculins (orange) dans le dernier quartile qui à ce jour est inexpliquée.

Encart 6.5 : Point faible

Les erreurs observées sur le seul personnage féminin de ce cas d'évaluation sont faibles. La contrainte d'unicité du genre appliquée à la création des paires nous amène précisément dans une situation où il n'existe aucune paire négative pour ce personnage. C'est-à-dire qu'on ne soumet à évaluation aucune paire impliquant le personnage féminin avec des voix de personnages masculins. Nous sommes donc face à une impasse, en ce qui concerne le protocole mis en place. En effet, la suppression de la contrainte du genre sur les paires, reviendrait à donner la possibilité au modèle de prendre en compte ce biais pour discriminer les paires. La différence (particulièrement au niveau de la fréquence fondamentale) entre les voix de femmes et les voix d'hommes étant généralement significative, cela conduirait vraisemblablement à construire un système de détection du genre. Une façon d'éliminer ce problème serait d'ajouter des données supplémentaires pour augmenter la variabilité, particulièrement en termes de locuteurs, ce qui nous permettrait de réaliser nos tests sur un plus grand panel d'acteurs.

Discussion

En considérant toutes les contraintes que nous avons appliquées à l'évaluation (et à l'apprentissage), il n'y a, selon nous, aucune chance pour que la différence mesurée soit le fruit du hasard. Cette observation nous permet donc d'affirmer que nous sommes en présence d'une information nous permettant de réaliser de manière automatique un appariement de voix (VO-VF) respectant les choix de l'opérateur de casting vocal. Nous mettons ainsi en évidence l'information, véhiculée au travers du signal acoustique, qui caractérise la dimension « personnage ».

6.4.4 Identifier l'apport des architectures siamoises

Nous sommes en droit de nous questionner sur l'intérêt d'utiliser des SNN plutôt qu'un architecture plus classique de réseaux de neurones. En effet, nous avons montré jusqu'ici l'efficacité des SNN pour l'apprentissage d'une mesure de similarité, mais nous ne l'avons pas soumise à comparaison. Nous avons donc réalisé une expérience ayant pour objectif de mesurer la différence en termes de performances entre différentes architectures neuronales.

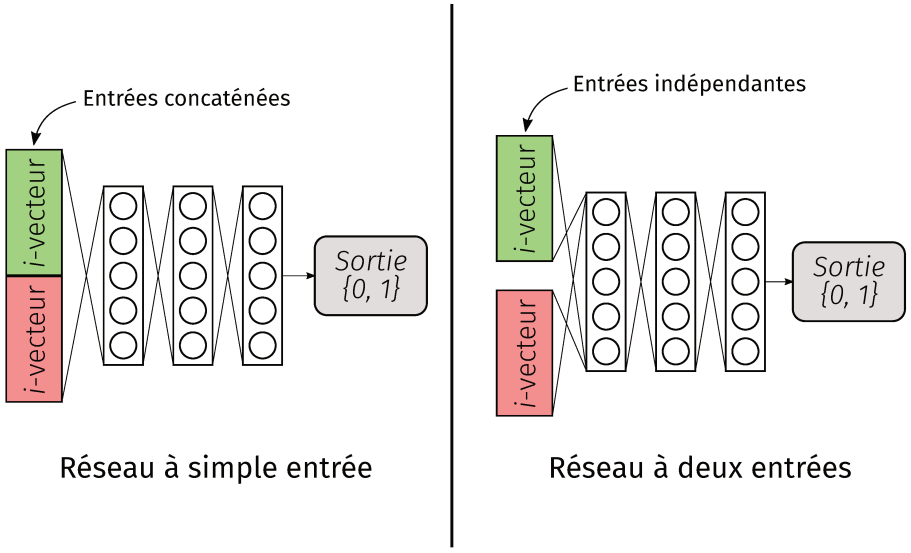

Figure 6.6 – Architectures utilisées en guise de comparaison.

		A	*B*	*C*	*D*
Entrées concaténées	Développement	0,94	0,96	0,93	0,96
	Test	0,49	0,49	0,51	0,53
Entrées indépendantes	Développement	0,93	0,94	0,93	0,96
	Test	0,52	0,50	0,53	0,52
SNN	Développement	0,72	0,71	0,70	0,71
	Test	0,55	0,59	0,62	0,50

Table 6.4 – Comparaison des performances obtenues avec les différentes architectures.

Nous avons soumis les SNN à comparaison avec deux autres architectures. La première reprend l'architecture détaillée dans la table 6.1 excepté que nous utilisons un seul réseau (au lieu de deux) et nous concaténons les vecteurs de représentation des deux segments (de taille d), qui composent la paire d'entrées, en un seul vecteur de taille $2d$ connecté à la couche d'entrée du réseau. La seconde reprend également le même réseau, mais au lieu d'utiliser la concaténation des entrées, nous connectons les vecteurs d'entrées à deux couches d'entrées du réseau indépendantes. Nous illustrons les deux architectures dans la figure 6.6. En résumé, nous utilisons le même réseau et ils diffèrent simplement dans leur manière de gérer les entrées.

Les résultats de la comparaison sont présentés dans la table 6.4. En considérant l'évaluation faite sur le corpus de test, nous observons de moins bonnes performances de manière générale, par comparaison avec les SNN qui généralisent bien mieux que les architectures classiques à l'exception du cas D.

De manière surprenante, nous avons obtenu des résultats très intéressants sur le corpus de développement avec les architectures classiques, des résultats qui surpassent de loin les SNN à chaque fois. L'explication la plus vraisemblable suppose que ces deux modèles ont plus de capacités de mémorisation au niveau des couples de locuteurs qui définissent chaque personnage. En revanche, ils semblent moins aptes à généraliser, au vu de leurs mauvaises performances lorsqu'ils sont faces à des nouvelles voix.

Ces mesures comparatives montrent l'intérêt d'une approche fondée sur les architectures siamoises. En effet, le partage des paramètres rend possible l'apprentissage d'un espace latent guidé par le calcul d'une distance entre les représentations de haut niveau et permet une meilleure généralisation.

6.5 Conclusion

Nous avons émis l'hypothèse selon laquelle il est possible de tirer parti de la comparaison de paires de voix pour l'apprentissage d'une mesure de similarité. La particularité de cette mesure est qu'elle doit rendre compte de la relation qui existe entre les voix qui constituent les paires ainsi que du savoir expert de l'opérateur de casting vocal. La mise en place du cadre d'évaluation et de la méthodologie que nous avons utilisés pour répondre à ce problème constitue une part importante de cette thèse. Nous avons ainsi prouvé l'efficacité des réseaux de neurones siamois pour modéliser cette similarité entre les voix de doublage. De plus, cette méthode permet de réaliser un apprentissage faiblement supervisé, qui nécessite une simple annotation des paires en tant que *target* ou *nontarget*.

Les résultats que nous avons présentés et qui ont fait l'objet d'une publication internationale (Gresse, Quillot et al. 2019), montrent que l'hypothèse que nous avons formulée est correcte, dans le sens où nous avons appris une mesure de similarité abstraite à partir des paires de voix (VO-VF). De plus, nous avons montré que les réseaux de neurones siamois offrent des capacités de généralisation qui surpassent celles des architectures dites « simples », au vu des résultats obtenus sur de nouvelles voix. En effet, la mesure de similarité apprise rend possible la discrimination *target* et *nontarget* des paires et ce même lorsque nous utilisons des paires de voix de personnages inédits.

Nous avons veillé à prendre toutes les précautions possibles pour l'exploitation de notre corpus de données de manière à neutraliser les biais

potentiels. Cela signifie qu'il est donc possible de construire un espace de représentation latent permettant de mettre en évidence la présence d'une information abstraite qui caractérise la dimension du personnage dans la voix. Toutefois, nous sommes conscients des limites qu'impose notre corpus et il existe sûrement d'autres biais que nous n'avons pas envisagés dans ce protocole expérimental. Pour pouvoir aller plus loin, il est primordial de recourir à d'autres données, en compléments de celles-ci. Ainsi, nous pourrons accroître la variabilité en termes de personnages, mais surtout en termes de voix incarnant des personnages similaires.

Enfin, nous avons, dans ces travaux, utilisé l'espace des i-vecteurs pour la représentation des données d'entrées. Nous avons fait ce choix par pure commodité. En effet, celui-ci ayant prouvé sa robustesse pour différentes tâches, nous l'avons donc utilisé afin d'obtenir une représentation de taille fixe de nos extraits audio de durée variable. De manière générale, le choix de la représentation des entrées joue un rôle important, il est alors tout naturel que nous travaillions sur cette question.

Dans le chapitre qui suit, nous proposons de travailler sur la caractérisation de la dimension personnage, à travers l'apprentissage d'un espace de représentation dédié à la tâche du casting vocal.

Chapitre 7

Le p-vecteur : un espace de représentation du personnage

Sommaire

7.1 Introduction

Bien que nous ayons observé une différence significative entre les deux catégories de paires et ce même sur des paires de voix nouvelles, il ne nous est pas permis d'affirmer que cet effet est représentatif de la dimension « personnage ». En effet, il peut s'agir d'autres biais que nous n'aurions pas envisagés et donc que nous n'avons pas neutralisés. Néanmoins, il est aussi tout à fait possible que ces biais potentiels soient quelque part liés au personnage. Prenons par exemple la hauteur de voix. Ainsi pour des nouvelles voix que le modèle n'aura pas appris à associer (au travers des paires qui lui sont présentées), le système peu prédire deux voix de F_0 proches comme étant similaires en termes de « personnage ».

Le modèle de similarité que nous avons présenté dans le chapitre 6 est évalué sur la base d'une classification binaire des paires de voix dites *target* et *nontarget*. Il semble tout de même difficile d'extraire à partir de la relation entre des paires de voix, une information caractéristique du personnage. En effet, nous avons observé, sur les tests effectués sur des nouvelles voix, des résultats différents suivant les cas de tests et même parfois proches d'un comportement aléatoire.

Étant donné que l'interpolation de la fonction de similarité ne peut être basée que sur des paires de contre-exemples, il semble, en effet, difficile de pouvoir s'attendre à une bonne généralisation. Nous avons donc supposé qu'il est possible d'apprendre un espace de représentation optimisé pour la discrimination en termes de personnage. De plus compte tenu des limites du corpus *Mass Effect 3* il serait intéressant de voir les données d'un autre jeu vidéo peuvent être exploitées, notamment les voix issues des personnages de *Skyrim*.

Dans ce chapitre, nous présentons donc le travail que nous avons effectué sur l'apprentissage d'un espace de représentation dédié à la dimension « personnage » que nous appelons *p*-vecteur dans la section 7.2. Mais avant cela, nous montrons comment nous tirons profit des voix du jeu *Skyrim*, en utilisant la technique dite de distillation de la connaissance que nous détaillons dans la section 7.2.3. enfin, nous décrivons le protocole expérimental et les résultats que nous avons obtenus dans les sections 7.3 et 7.4. Finalement, nous présentons nos conclusions dans la section 7.5.

7.2 Approche

7.2.1 Le p-vecteur : une représentation de l'information caractéristique du personnage

Nous savons qu'une part importante du succès d'un algorithme d'apprentissage automatique dépend de la représentation des données en entrée. La paramétrisation acoustique du signal de parole (MFCC) permet de représenter l'information cepstrale contenue dans une trame. En revanche, la paramétrisation ne répond pas au problème de variabilité de la durée du signal de parole. En RAL, une modélisation plus poussée est réalisée (i-vecteur). L'objectif est de résumer l'information contenue dans la séquence de paramètres acoustique dans une représentation de taille fixe.

Ces dernières années, ce sont les approches fondées sur des architectures neuronales profondes qui ont bénéficié d'un intérêt particulier. En effet, les DNN peuvent être utilisés pour l'apprentissage d'un espace de représentation capable de démêler les facteurs explicatifs d'une variabilité plus ou moins bien cachés dans les données (Bengio, Courville et al. 2013).

De récents travaux dans le domaine de la reconnaissance automatique du locuteur (Variani et al. 2014 ; Snyder, Ghahremani et al. 2016 ; Snyder, Garcia-Romero, Povey et al. 2017 ; Snyder, Garcia-Romero, G. Sell et al. 2018) ont montré que en s'appuyant sur une stratégie d'apprentissage discriminant, il est possible d'apprendre un espace de représentation dédié à la tâche en question (se référer à la section 2.3.3). De plus, l'apprentissage d'une telle représentation, aussi appelé *embedding*, peut être guidé par une mesure de similarité. En RAL, les méthodes fondées sur des DNN surpassent généralement les méthodes alors état-de-l'art, s'appuyant sur les i-vecteurs. Ces approches sont qualifiées de « bout-en-bout » (*end-to-end*). Elles requièrent cependant un volume important de données propres au domaine. Cette contrainte peut alors s'avérer problématique pour des tâches sous-représentées en termes de données.

Dans ces travaux, nous proposons d'utiliser une approche similaire. Ainsi, en bénéficiant d'un apprentissage discriminant, nous supposons qu'il est possible d'extraire l'information caractéristique du personnage dans un espace de représentation latent, appelé p-vecteur (p fait ici référence à « personnage »), optimisé pour la discrimination des voix incarnant différents personnages. Une telle représentation est avantageuse, car elle permet de

projeter les vecteurs de représentation des données en entrée dans un nouvel espace qui maximise la variabilité en termes de personnage. À l'inverse, nous supposons que cela permet d'atténuer les facteurs ayant un faible impact sur cette tâche. Cette approche, fondée sur le *representation learning* décrit par Bengio (Bengio, Courville et al. 2013), suppose que nos données soient révélatrices d'une variabilité latente et décomposable selon différents facteurs.

Apprentissage disjoint fondé sur une représentation orientée locuteur

Les méthodes proposées dans la littérature dites bout-en-bout ont l'inconvénient de nécessiter un volume de données conséquent. Ce manque de données est un problème dans de nombreuses applications, ce qui est le cas dans la problématique que nous traitons dans ce manuscrit, en particulier sur le jeu de données utilisé. Pour pallier cette limitation, nous utilisons une approche disjointe dans laquelle nous apprenons, à partir du signal de parole, une représentation intermédiaire (*x*-vecteur) dont l'apprentissage est supervisé par un gros volume de données annotées selon l'identité du locuteur (voir la figure 7.1). Cette représentation intermédiaire est ensuite utilisée pour construire l'espace des *p*-vecteurs grâce à un apprentissage discriminant.

De la même façon que pour les travaux présentés auparavant, nous avons opté pour une représentation des données dédiée au locuteur, à l'exception que nous considérons ici la méthode état-de-l'art (*x*-vecteur).

7.2.2 Homogénéisation de l'information par distillation

Dans notre contexte de travail, nous disposons d'un ensemble réduit de données annotées respectant les contraintes fixées par notre protocole d'évaluation (voir section 6.3). La seule information dont nous disposons correspond aux associations de voix de doublage que la DA a réalisé par le passée. Apprendre de manière supervisée une représentation dédiée à notre tâche, en nous appuyant sur une architecture neuronale profonde semble donc difficile compte tenu de cette limitation.

Idéalement, nous souhaiterions construire un espace de représentation dédié aux voix de doublage utilisées dans les œuvres audiovisuelles. Cependant, l'espace de représentation que nous entraînons est spécifique à notre corpus *Mass Effect 3*, lui-même représentatif d'un « univers sonore »

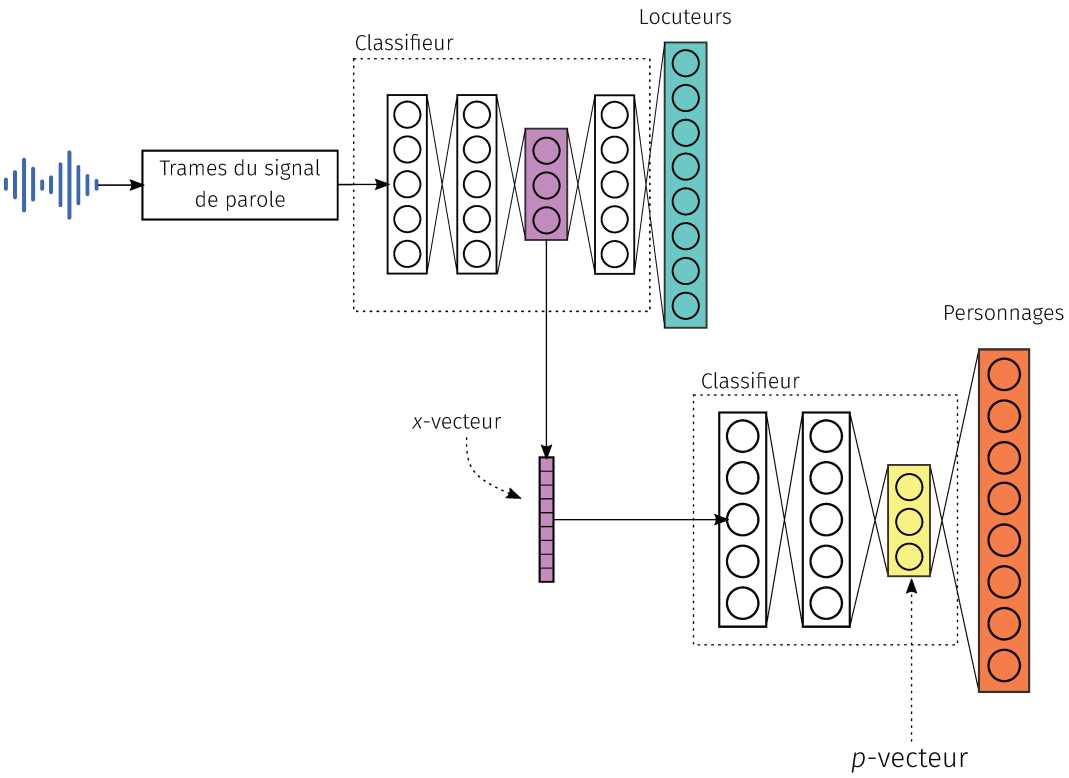

Figure 7.1 – Illustration de l'approche disjointe utilisée pour l'apprentissage du *p*-vecteur.

propre à cette œuvre. Se pose donc la question de la représentativité des données d'entraînement utilisées par rapport à l'ensemble des productions audiovisuelles existantes. Au-delà du fait que l'utilisation d'un corpus spécifique (qui plus est réduit) pose inévitablement un problème de sélection de données, nous sommes ici face à un biais lié à la subjectivité des données d'apprentissage. La grande variabilité des productions culturelles pose donc un problème pour la construction d'un corpus d'entraînement homogène et consistant. En effet, les points de vues diffèrent d'une équipe de DA à une autre et il n'y a pas de consensus, autre que les aspects socioculturels établis, sur les jugements portés à l'égard des voix.

Afin de remédier à cette situation, nous proposons d'utiliser une méthode appelée « distillation de la connaissance ». Cette technique nous permet d'extraire la connaissance généralisée d'un modèle entraîné sur un corpus additionnel et d'en bénéficier lors de l'apprentissage d'un nouveau modèle. Ce dernier est quant à lui destiné à l'apprentissage de notre espace de représentation (*p*-vecteur) à partir du corpus *Mass Effect 3*. La connaissance généralisée du premier modèle guide l'apprentissage du *p*-vecteur.

7.2.3 Distillation de la connaissance

La distillation de la connaissance dans un réseau de neurones proposée par Hinton (Hinton et al. 2014) s'inscrit dans le paradigme d'apprentissage automatique appelé « Maître-Élève » permettant à la machine d'apprendre à la machine à l'instar de l'apprentissage humain dans lequel le maître « intelligent » transmet sa connaissance en guidant, conseillant et commentant l'élève dans son apprentissage. Vapnik (Vapnik et al. 2015) qualifie ce qui émerge de l'interaction entre le maître et l'élève d'« information privilégiée » et propose une méthode d'apprentissage automatique mettant en jeu deux modèles : un « Maître » et un « Élève ». Nous donnons plus de détails à ce sujet dans l'annexe C.

Initialement, Hinton a présenté la méthode d'apprentissage Maître-Élève par distillation comme un moyen de réduire la complexité des modèles utilisés, par exemple, pour l'apprentissage d'une tâche de reconnaissance de la parole dans un contexte de déploiement dans un cadre contraint en termes de latence et de ressources matérielles.

L'apprentissage d'un modèle de classification, qui consiste généralement à maximiser la probabilité de prédiction de la bonne classe, a un effet de bord intéressant. Le modèle assigne une probabilité à la bonne réponse mais aussi à toutes les mauvaises réponses. Même si ces probabilités sont infimes, certaines sont plus importantes que d'autres. Leur différence relative est donc une source d'information importante en ce qui concerne le comportement et la généralisation du modèle. En guise d'exemple, une image de voiture a peu de chance d'être confondue avec une image de camion, mais il est encore moins probable qu'elle soit confondue avec une image de vélo. Le principe de distillation consiste donc à utiliser les probabilités assignées aux différentes classes, que Hinton appelle les *soft-targets*, comme références pour l'entraînement d'un autre modèle. Dans le cas où les *soft-targets* ont une entropie élevée, ces probabilités a posteriori fournissent plus d'informations que les simples *hard-targets*. Ces dernières correspondent aux classes encodées de façon binaire, avec un 1 pour la bonne classe, toutes les autres étant à 0, aussi appelé encodage *one-hot*. Le modèle entraîné en premier lieu, supposé détenir une connaissance générale du domaine, joue le rôle du Maître. Le modèle entraîné à partir des *soft-targets* fournies par le Maître fait alors référence à l'Élève.

Un des aspects intéressants de cette technique réside dans la possibilité de recourir à des données non annotées pour l'apprentissage du modèle

Élève. Seules les *soft-targets* obtenues à partir du Maître sont utilisées pour superviser l'apprentissage de l'Élève. Plus généralement, le Maître permet à l'Élève d'apprendre une représentation des frontières de décisions qui n'est pas contenue dans les données (Lopez-Paz et al. 2016).

Principe de distillation

Concrètement, nous utilisons un réseau de neurones optimisé sur une tâche de discrimination des personnages pour apprendre notre espace de représentation. Ce modèle neuronal assigne aux différentes classes (personnages) une probabilité grâce à sa couche de sortie qui est associée à une fonction de *softmax*. Cette fonction convertit chaque sortie de la couche, le z_i (aussi appelé « logit »), en le comparant aux autres logits, en une probabilité q_i de la manière suivante :

$$q_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)} \tag{7.1}$$

où T fait référence à la température (fixée à 1 dans la fonction *softmax* usuelle) et $i,j \in [0...C]$ avec C le nombre de classes. La distillation consiste à augmenter la valeur de T dans l'apprentissage du modèle Maître, pour ainsi produire une distribution de probabilités entre les classes plus progressive. Le corpus utilisé pour apprendre le modèle Maître est appelé corpus de transfert. Ce dernier peut être identique au corpus principal, mais en ce qui nous concerne nous utilisons un corpus différent. L'Élève est entraîné sur le corpus principal (*Mass Effect 3*) avec, en guise de supervision, les vecteurs de probabilités donnés par le Maître pour chaque voix qui compose ce corpus. Notons que la même valeur de T est utilisée pour l'entraînement des deux modèles.

Dans le cas où les données utilisées bénéficient d'une annotation spécifique, il est alors possible d'optimiser le réseau de neurones Élève selon deux fonctions objectives : la première calculée sur les *soft-targets* données par le Maître ; la deuxième calculée sur les *hard-targets* en référence à l'annotation des données. Un paramètre λ est alors introduit pour contrôler le poids attribué à chacune des deux fonctions objectives grâce à la formule suivante

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} [(1-\lambda)L(y_i, q_i) + \lambda L(s_i, q_i)] \tag{7.2}$$

avec $i \in [0...N]$, N est le nombre d'exemples et où L dénote la fonction de

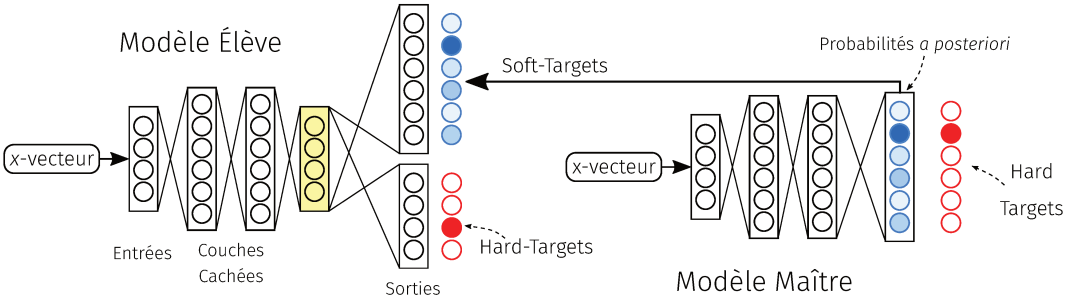

Figure 7.2 – Dans cette illustration, le modèle Maître apprend à discriminer les vecteurs donnés en entrée selon différentes classes, jusqu'à ce qu'il produise les *soft-target* requises pour l'apprentissage du modèle Élève. Les deux modèles peuvent être entraînés sur différents jeux de données.

coût appelée « entropie-croisée », s_i fait référence aux *soft-targets* et y_i aux *hard-targets*.

Selon Hinton, de meilleures performances sont obtenues en entraînant l'Élève à la fois sur les *soft-targets* fournies par le Maître et les *hard-targets*. Nous illustrons cette stratégie dans la figure 7.2. En résulte une approche qui finalement, rejoints le cadre d'apprentissage avec information privilégiées proposé par Vapnik. Dans Lopez-Paz et al. 2016, les auteurs proposent d'unifier l'apprentissage par information privilégiée et la distillation dans un cadre général de distillation se résumant en trois étapes :

1. Apprendre le modèle Maître avec les paires d'entrées-sorties $(x_i^\star, y_i)$ (se référer à l'annexe C).
2. Calculer les *soft-targets* avec le modèle entraîné.
3. Entraîner le modèle Élève en utilisant les deux paires entrées-sorties $(x_i, y_i), (x_i, s_i)$ et $\lambda \in [0,1]$.

En définitive, le cadre d'apprentissage Maître-Élève nous permet d'utiliser la connaissance du modèle Maître apprise à partir d'un corpus externe, d'un domaine proche. Cette connaissance est utilisée pour guider l'apprentissage du modèle Élève sur notre domaine spécifique. Ce dernier est assez restreint en termes de personnages, d'où l'utilisation d'un corpus additionnel contenant plus de personnages. La méthode Maître-Élève a été utilisée dans différents travaux et sur des tâches variées (Price et al. 2016 ; Markov et al. 2016 ; J. Li et al. 2017 ; Watanabe et al. 2017 ; Asami et al. 2017 ; Joy et al. 2017). Nous proposons d'étendre cette méthode dans le cadre de l'apprentissage d'une représentation dédiée à la caractérisation de l'information du personnage/rôle.

7.3 Expériences

7.3.1 Corpus

Jusqu'ici, nous avons fait un usage exclusif du corpus *Mass Effect 3*. Ce dernier est décrit en détails dans la section 5.3.1. Dans ces travaux, nous avons recours également au corpus *Skyrim* issu du jeu vidéo ayant le même nom. Il s'agit donc de données additionnelles que nous pouvons considérer comme de nature similaire aux données de *Mass Effect 3*.

Le corpus *Skyrim* est composé des dialogues anglais (VO) et français (VF) du jeu vidéo et totalise pas moins de 120 heures de parole. La nomenclature des fichiers audio respecte une structure similaire à celle décrite dans la figure 5.3. Cela nous permet d'extraire 50 000 équivalences de segments anglais-français. Nous disposons de 30 personnages différents (7 femmes et 23 hommes). Un personnage étant défini par une paire d'acteurs anglais-français. Enfin, il s'agit de segments audio de haute qualité enregistrés en studio.

Ce jeu de donnée est utilisé en tant que corpus de transfert pour le processus de distillation. Toutefois, l'évaluation de notre système ne s'appuie pas sur ce corpus, car nous n'avons pas de garanties suffisantes en termes d'équivalence VO-VF des segments. En effet, dans certains cas, les identifiants des segments de la VO correspondent aux segments de la VF, mais leur contenu est différent. Ce décalage est sans doute lié au processus de traduction. N'étant pas en mesure d'effectuer des vérifications approfondies sur l'ensemble du jeu de données, nous préférons éviter d'utiliser ce corpus en guise d'évaluation. De plus, nous n'avons pas de garanties sur les rôles attribués aux différents acteurs. Il est donc possible qu'un acteur joue plusieurs personnages. Par chance, nous ne notons aucune intersection entre les acteurs de *Skyrim* et ceux de *Mass Effect 3*, ce qui nous préserve de tout biais en termes de locuteur au niveau de l'évaluation (le corpus *Mass Effect 3* servant pour l'évaluation).

7.3.2 Préparation des données

Contrôle des biais

Nous n'utilisons pas tous les segments de voix du corpus *Skyrim* car nous voulons éviter d'introduire des biais dans l'apprentissage du modèle Maître.

Nous appliquons donc sur ce corpus les même contraintes qu'auparavant (voir la section 6.3). À l'exception de la contrainte sur le genre, puisque nous n'effectuons pas d'apprentissage ni d'évaluation par paires de voix avec le corpus *Skyrim*. Autrement, nous neutralisons le biais de fréquence, non contrôlons la durée des segments audio utilisées, ainsi que le contenu linguistique.

Découpage des données

Dans ce travail, nous réutilisons la méthode d'apprentissage par validation croisée mise en place dans les travaux que nous avons présentés précédemment (voir la section 6.3.2). Pour rappel, nous avons 4 plis notés *A*, *B*, *C* et *D*, chacun contenant 12 personnages d'apprentissage et 4 pour l'évaluation isolés de l'entraînement. Notons que cela concerne uniquement le corpus principal (*Mass Effect 3*) sur lequel nous évaluons le système.

En ce qui concerne le corpus de transfert (*Skyrim*), nous ne faisons pas de découpage en plusieurs plis de validation, étant donné que nous n'évaluons pas le système sur ce jeu de données. Toutefois, nous prenons en compte le même nombre de segments par étiquette de personnage pour éviter un quelconque déséquilibre en termes de fréquence d'apparition. De plus, nous divisons l'ensemble des segments de chaque personnage de *Skyrim* en deux. Une partie destinée à l'entraînement et une seconde désignée pour la validation de l'apprentissage (développement). Ce dernier compte 20 % des segments utilisés, les 80 % restant servent à l'entraînement.

Extraction des séquences

Nous extrayons les paramètres acoustiques à partir du signal en suivant le procédé décrit dans la sous-section 5.3.2. Cependant, nous entraînons ici un extracteur *x*-vecteur indépendant de la langue en utilisant le corpus Voxceleb 1 et 2 (Chung et al. 2018). Enfin, nous calculons pour chaque segment de notre corpus les *x*-vecteurs correspondants. Toutes ces opérations sont réalisées grâce à la boîte à outils Kaldi (Povey, Ghoshal, Boulianne, Burget et Glembek 2011).

Nous illustrons dans la figure 7.3 une vue des segments de voix des différents personnages dans l'espace des *x*-vecteurs. Nous observons une bonne distinction des acteurs de la VO et de la VF. Chaque locuteur (acteur) étant bien représenté, excepté dans le cas *C*. Nous visualisons également une forte

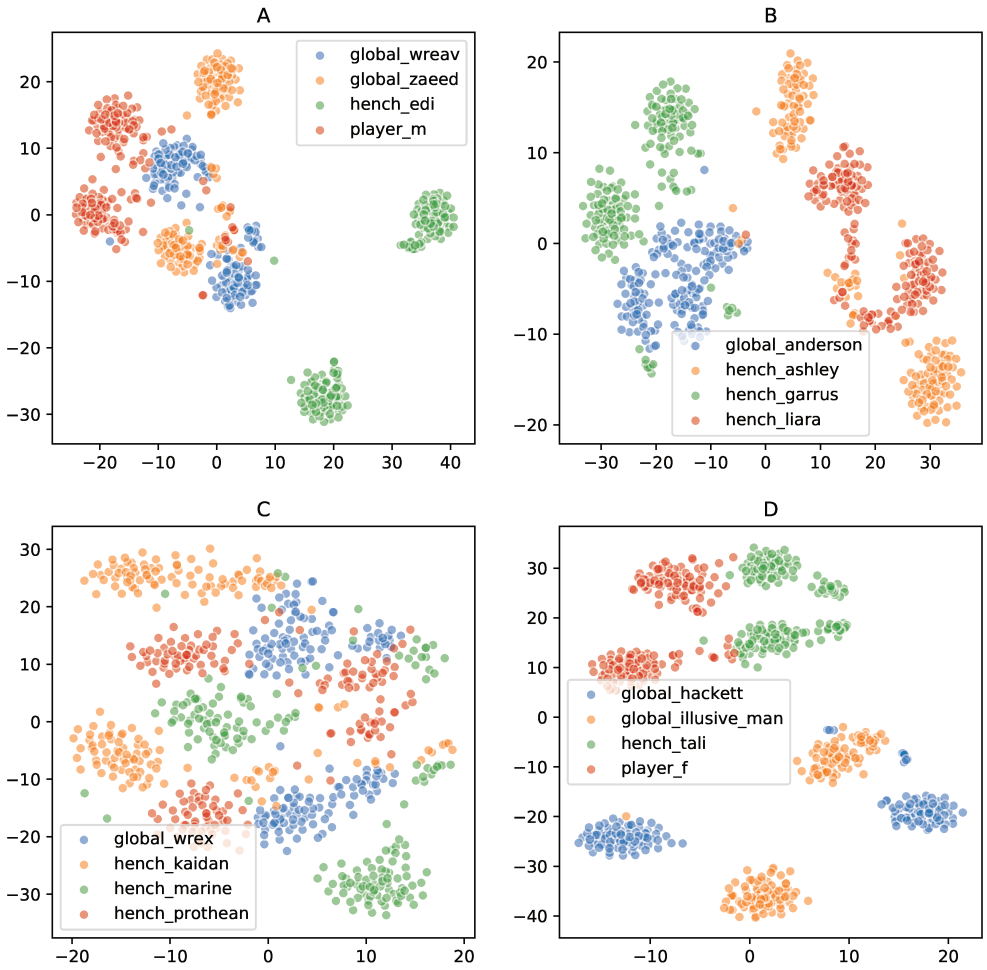

Figure 7.3 – Représentation dans l'espace des x-vecteurs des segments de voix des différents personnages.

discrimination au niveau du genre. De plus, il est important de noter qu'il semble y avoir une proximité chez différents acteurs, ce qui est révélateur d'une certaine similarité au niveau acoustique de leur voix en elles-même. Cette illustration étant obtenue grâce à t-SNE qui permet de réduire les espaces de grande dimensionnalité, à seulement deux ou trois dimensions afin des mieux les visualiser. Il est donc difficile de dire sur quelles dimensions reposent la proximité des points dessinés.

7.3.3 Évaluation

Notre objectif est de créer une représentation qui caractérise la dimension du personnage/rôle dans la voix. Pour cela, nous devons tester le système sur un ensemble de voix auxquelles il n'est pas acclimaté. C'est le rôle de notre approche par validation croisée. Étant donné le faible nombre de personnages dans notre jeu de données, ce protocole d'évaluation nous per-

met de couvrir l'ensemble des personnages.

Analyse par clustering

Comme dans les travaux présentés auparavant, nous avons 4 plis d'évaluation notés A, B, C et D contenant chacun 4 personnages différents. Une analyse par clustering s'appuyant sur la représentation apprise, le p-vecteur, permet d'évaluer sa qualité. En effet, nous souhaitons évaluer si il est possible de retrouver les 4 personnages du pli à partir des clusters appris avec un algorithme de classification non-supervisée opérant dans l'espace des p-vecteurs.

Nous utilisons pour cela l'algorithme des k-moyennes qui nous permet de fixer le nombre de groupes désirés a priori. Nous fixons $k = 4$ afin de refléter le nombre de personnages effectivement présents dans le test. Nous assignons tous les segments réunis dans un groupe au label (personnage) le plus représenté dans ce même groupe. Ainsi, nous pouvons estimer une F-mesure, calculée sur l'hypothèse d'appartenance du segment au label. Il est possible que plus d'un groupe soit assigné à un même personnage, ce qui constitue un inconvénient de cette méthode. Toutefois, cela serait la conséquence d'une mauvaise classification nous permettant de dire que la représentation apprise ne discrimine pas bien les personnages du test.

Évaluation par similarité

En complément de l'évaluation par clustering, nous proposons d'évaluer le p-vecteur dans le cadre de la tâche d'appariement des voix de doublage en nous appuyant sur les scores de similarité obtenus sur les paires de voix des personnages du test. Pour cela, nous reprenons le protocole d'évaluation détaillé dans le chapitre 6. Pour rappel, l'évaluation consiste à mesurer la capacité de discrimination des paires *target* et *nontarget*, à l'aide d'un modèle de classification binaire entraîné avec les p-vecteurs. Le fait d'apprendre un modèle mesurant l'appariement de voix dans des langues différentes et de l'évaluer sur des voix inconnues rend cette tâche particulièrement difficile.

7.3.4 Définition des modèles

Nous utilisons la bibliothèque Keras (Chollet et al. 2015) pour l'implémentation des modèles que nous détaillons ci-dessous.

Architecture

Nos modèles Maître et Élève s'appuient sur un Perceptron Multi-Couche (MLP). La couche d'entrée du réseau de dimension 512 est connectée à trois couches cachées de 256 unités neuronales. Les couches cachées sont associées à une tangente hyperbolique en guise d'activation. À cela, nous ajoutons une dernière couche cachée de 64 unités qui correspondent au p-vecteur. Cette dernière est finalement connectée à la couche de sortie du réseau qui est quant à elle associée à une fonction d'activation *softmax*. Le nombre d'unités de cette couche correspond au nombre de classes que nous voulons discriminer. Ainsi, lorsque nous entraînons le modèle Maître, nous discriminons les 30 personnages de *Skyrim* (corpus de transfert). Autrement, l'apprentissage de l'Élève considère 12 personnages de *Mass Effect 3* (corpus principal).

Optimisation et régularisation

Nous utilisons un optimiseur *Adadelta* pour gérer le pas d'apprentissage de manière dynamique. L'optimiseur doit minimiser la fonction objective, ici de l'entropie-croisée. De plus, nous intercalons entre chaque couche cachée un *Dropout*. Le taux d'unités désactivées dans les trois premières couches cachées est fixé à 0,25 et monte à 0,5 pour la dernière (p-vecteur).

Initialisation et apprentissage

Tous les paramètres des modèles sont initialisés selon une distribution normale de Xavier (Glorot et al. 2010). La taille du batch est fixée à 12 individus et nous entraînons les modèles sur 300 époques en prenant soin de surveiller les performances des modèles à partir des évaluations faites sur le corpus de développement. Nous enregistrons les paramètres des modèles dans leur configuration donnant les meilleures performances sur cette validation.

Nous entraînons le modèle Maître avec les données de *Skyrim*, considéré comme information privilégiée. Une fois ce modèle appris, nous lui soumettons les données de *Mass Effect 3* afin d'obtenir les *soft-targets*. Ces dernières, combinées aux *hard-targets* de ce même corpus nous servent pour entraîner le modèle Élève en contrôlant leur impact respectif avec le paramètre λ. Enfin, nous extrayons les p-vecteurs directement depuis l'*embedding* du

modèle Élève.

Modèle de similarité

Comme expliqué dans la sous-section 7.3.3, nous évaluons la qualité des *p*-vecteurs que nous avons extrait, au moyen du système de similarité. Contrairement à l'architecture convolutive proposée pour les réseaux siamois utilisés dans les travaux précédents, nous utilisons une architecture plus simple, un MLP (semblable à celui détaillé plus haut) plus rapide à entraîner tout en nécessitant moins d'hyper-paramètres.

L'architecture pour les réseaux siamois est constituée de seulement deux couches cachées de 256 unités, combinées à une fonction d'activation tangente hyperbolique. Un *Dropout* de 0,25 est utilisé pour la régularisation. De plus, nous utilisons la version binaire de l'entropie-croisée au lieu de la fonction contrastive, comme précédemment, qui converge plus rapidement. Nous nous retrouvons finalement avec un modèle beaucoup moins onéreux à entraîner.

7.4 Analyse des résultats

Dans cette partie, nous livrons les résultats de nos travaux sur la représentation *p*-vecteur. Nous présentons à chaque fois les résultats obtenus avec et sans l'utilisation du processus de distillation. À titre de comparaison, nous donnons également les résultats du système *baseline* (sans *p*-vecteur), c'est-à-dire en utilisant la représentation intermédiaire état-de-l'art optimisée sur une tâche de reconnaissance du locuteur, à savoir le *x*-vecteur. À titre d'information, nous ajoutons également les résultats obtenus sur une représentation *i*-vecteur.

7.4.1 Analyse par clustering

Nous commençons par présenter les résultats de l'évaluation fondée sur l'analyse par clustering. Nous pouvons les analyser à l'aide de la table 7.1. Nous présentons ainsi les *F*-mesures obtenues sur chacun des 4 plis de la validation croisée en considérant les segments des personnages de test.

Premièrement, nous observons qu'en comparaison de la représentation orientée locuteur (*x*-vecteurs), les performances avec les *p*-vecteurs appa-

	A	*B*	*C*	*D*
baseline (*i*-vector)	0,45	0,57	0,40	0,53
baseline (*x*-vector)	0,54	0,52	0,36	0,71
p-vector (sans distillation)	0,66	0,72	**0,59**	0,66
p-vector + distillation	**0,78**	**0,78**	0,40	**0,77**

TABLE 7.1 – *F*-mesure calculée sur l'analyse clustering effectuée à partir des *p*-vecteurs de test.

raissent meilleurs. Il n'est pas surprenant que ceux-ci aient de meilleurs résultats en comparaison avec une représentation qui est, elle, optimisée pour discriminer l'identité du locuteur. Étant donné qu'un personnage est défini par deux voix d'acteurs, et que l'espace *x*-vecteur est construit de telle sorte à pouvoir maximiser la variance entre les locuteurs, un simple partitionnement avec la méthode des *k*-moyennes ne permet vraisemblablement pas de distinguer les différents personnages. Contrairement aux *x*-vecteurs, les *p*-vecteurs sont eux optimisés pour mettre en avant les facteurs explicatifs de la variation en termes de personnages. Sur ce point, nous pouvons dire qu'ils ont un effet bénéfique, compte tenu des résultats observés, dans trois cas de test sur quatre (dans le cas du système *p*-vecteur sans distillation).

En termes de distillation, nous avons utilisé différentes températures $T \in [1,...,5]$ pour la distillation des *p*-vecteurs. Nous avons noté que les meilleurs résultats en moyenne s'obtiennent en utilisant $T = 4$. Il s'agit d'une température plutôt haute étant donné que le gradient de l'adoucissement de la distribution de probabilité est de l'ordre de $1/T^2$.

Toujours en termes de distillation, nous avons utilisé différentes valeurs pour le contrôle de l'imitation des *soft-targets* et des *hard-targets*, $\lambda \in [0,1]$. En moyenne, nous observons les meilleurs résultats dans le cas où $\lambda = 0,3$. Cela suggère que l'apport des *soft-targets* est substantiel.

Ces résultats attestent donc de la qualité de la représentation que nous avons apprise, au vu des résultats observés sur les voix de test qui sont, dans chaque cas, des voix inédites auxquelles le système n'a jamais été confronté. Nous donnons en guise d'illustration une vue de l'espace de représentation *p*-vecteur sur nos différents cas de test (voir la figure 7.4) grâce à la technique *t*-SNE.

Nous observons bien dans les cas *A*, *B*, et *D* la distinction entre les *p*-vecteurs des personnages féminins et masculins (*C* ne contient que des personnages masculins). Nous voyons ici à l'œil nu des groupes bien agen-

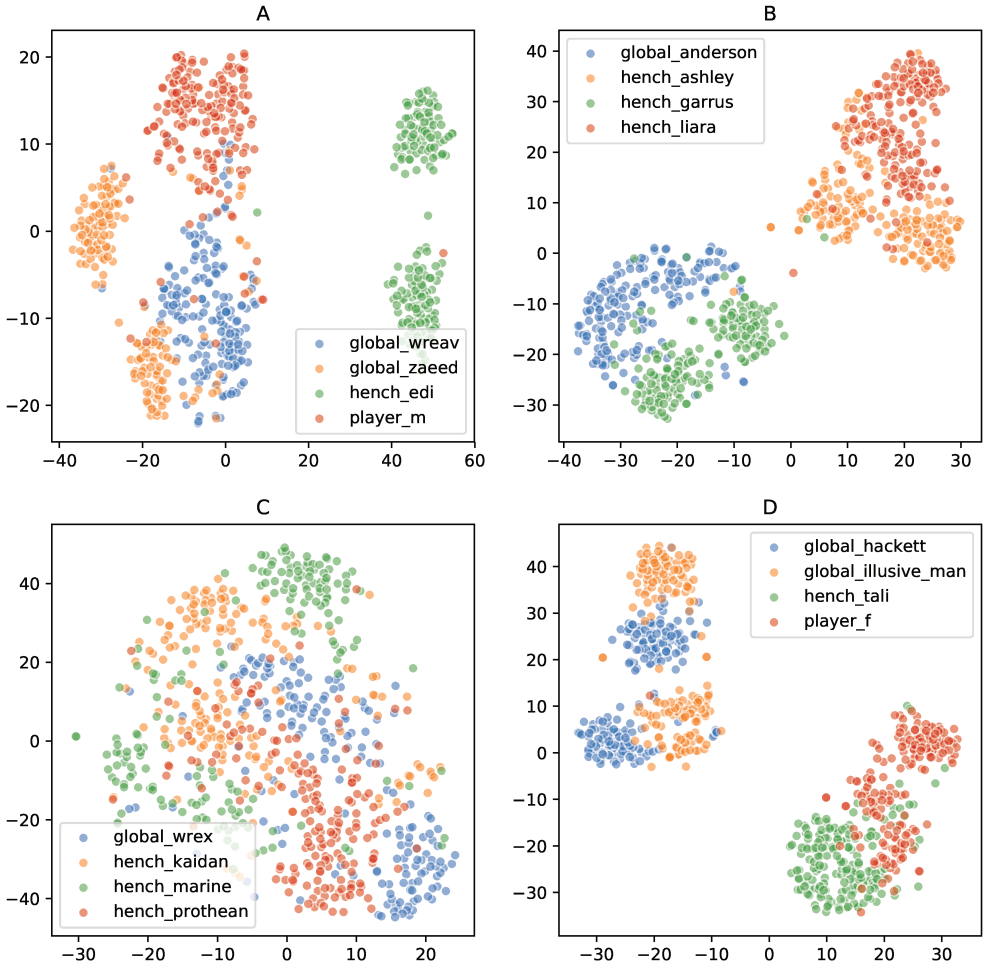

Figure 7.4 – Projection des *p*-vecteurs dans un espace à deux dimensions appris avec l'algorithme *t*-SNE. Les axes n'ont pas de signification particulière.

cés, notamment dans le cas *B* et dans le cas *D*, pour les personnages féminins (*player_f* et *hench_tali*). Les personnages masculins du cas *D* semblent confondus à première vue. Cela pourrait être dû au fait que ces deux personnages (*global_hackett* et *global_illusive_man*) ont des rôles similaires, tous les deux étant des décideurs.

7.4.2 Système de similarité

Nous continuons de dérouler le cadre méthodologique d'évaluation que nous avons présenté dans la section 6.3. Cette méthodologie de comparaison par paires de voix nous permet de vérifier la présence de l'information caractéristique du personnage dans la représentation *p*-vecteur.

Nous présentons les résultats de cette évaluation dans la table 7.2. Ces

		Taux de réussite	Test de Student
baseline *i*-vecteur	*A*	0,60	64,58
	B	0,52	20,63
	C	0,54	26,86
	D	0,49	-6,19
	mean	0,54	26,47
baseline *x*-vecteur	*A*	0,60	61,76
	B	0,54	29,99
	C	0,52	10,81
	D	0,49	-4,07
	mean	0,54	24,62
p-vecteur	*A*	0,58	53,82
	B	0,54	20,70
	C	**0,57**	**49,86**
	D	0,54	23,34
	mean	0,55	36,93
p-vecteur + distillation	*A*	**0,63**	**80,00**
	B	**0,55**	**36,46**
	C	0.55	28,33
	D	**0,55**	**34,24**
	mean	0,57	44,79

Table 7.2 – Mesure de la performance du classificateur de paires *target* et *nontarget* à partir des *p*-vecteurs de test. Les performances sur le corpus de développement (absentes du tableau) tournent généralement aux alentours de 85 % de réussite.

résultats sont présentés de deux points de vue : 1) en termes de réussite des prédictions sur les paires, et 2) avec le test statistique d'hypothèse nulle qui permet de voir s'il existe une différence significative entre les scores des paires *target* et ceux des paires *nontarget*.

En termes de résultats, nous constatons que les valeurs-*p* associées aux tests statistiques permettent de rejeter l'hypothèse nulle dans tous les cas (< 0.001). Pour rappel, l'hypothèse nulle nous dit que la moyenne des deux groupes est identique. Nous pouvons donc affirmer qu'il existe une différence significative entre les scores des paires de voix de mêmes personnages et celles composées de voix de personnages différents. C'est le point le plus important de ces résultats. En effet, l'évaluation réalisée avec notre système de similarité à partir des *p*-vecteurs montre la présence d'information permettant cette distinction. Si nous regardons maintenant l'apport de cette représentation, au niveau des résultats, nous constatons une amélioration des performances en moyenne en utilisant l'approche *p*-vecteur en compa-

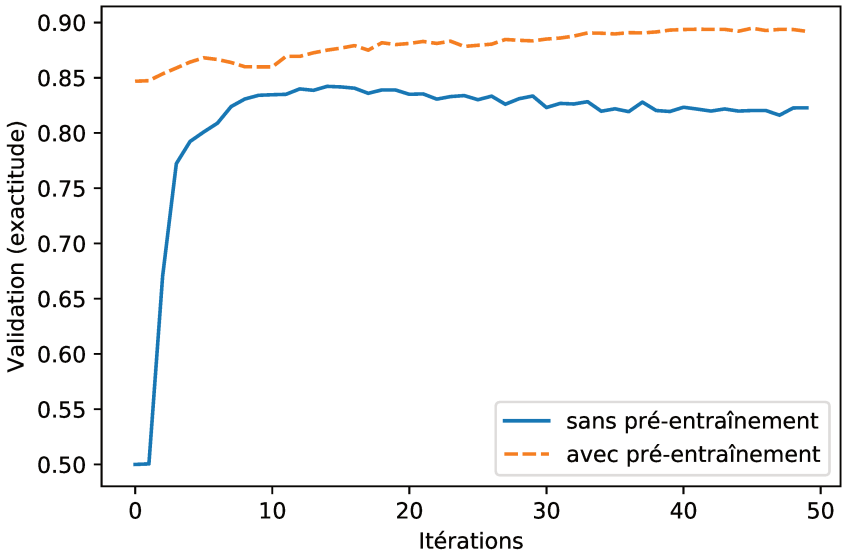

Figure 7.5 – Comparaison de la courbe d'apprentissage du modèle de similarité en fonction du pré-entraînement.

raison d'une représentation orientée locuteur.

Pré-entraînement

Il est important de noter que nous avons ajouté ici une étape de pré-entraînement pour l'apprentissage du modèle de similarité. En effet, nous pouvons utiliser le corpus de transfert pour pré-entraîner les réseaux siamois et bénéficier d'une initialisation des paramètres reflétant la distribution des *p*-vecteurs.

Nous illustrons dans la figure 7.5 les performances du modèle d'appariement des voix sur les données de validation. Nous voyons que le pré-entraînement surpasse nettement le modèle appris avec une initialisation aléatoire des paramètres. Le taux de réussite mesuré pendant la validation du modèle pré-entraîné semble meilleur. Toutefois, il est difficile de vraiment comparer cela en termes de performances, puisque nous constatons une différence significative avec ou sans pré-entraînement. Il semble délicat de dire que la différence est dans un cas plus significative que dans un autre. Nous retenons alors simplement que cette méthode permet de réduire la durée de l'apprentissage du modèle de similarité.

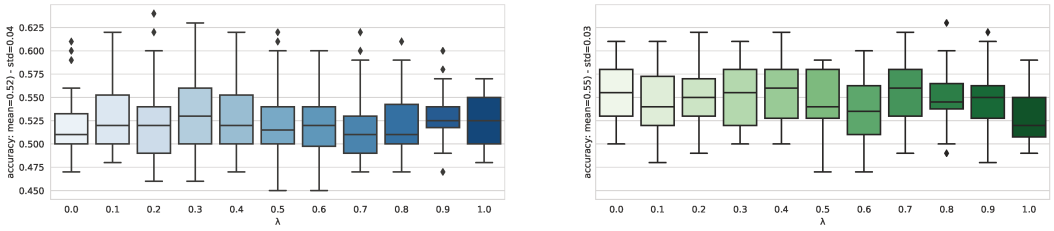

Figure 7.6 – Taux de réussite des prédictions en fonction du paramètre λ, résultats obtenus avec les p-vecteurs de test sur la tâche d'appariement des voix. À gauche l'approche originale basée sur les i-vecteurs et à droite l'approche basée sur les x-vecteurs (utilisés pour l'approche p-vecteur). Les losanges représentent les valeurs aberrantes.

Comparaison de la représentation intermédiaire

Le but initial de ce travail n'est pas de réaliser une étude comparative entre les x-vecteurs et les i-vecteurs. En effet, il ne fait aucun doute que nous pouvons valider cette approche comme nous le confirment les résultats présentés dans la table 7.1. Toutefois, la qualité de la représentation p-vecteur obtenue avec les x-vecteurs semble meilleure, compte tenue des valeurs de la F-mesure.

Nous montrons dans la figure 7.6 les performances du système de similarité en fonction du paramètre d'imitation. Pour rappel, quand $\lambda = 0$ le modèle prend en compte seulement les *hard-targets*. Inversement quand $\lambda = 1$ le modèle cherche à imiter uniquement les *soft-target* provenant du modèle Maître. Cette illustration nous permet d'observer d'une manière générale que les performances ont tendances à être plus élevées avec l'approche x-vecteurs. Surtout, nous observons une moins grande variabilité des résultats avec l'approche p-vecteurs fondée sur les x-vecteurs. Cette dernière conduit donc à un gain de robustesse. En nous référant aux valeurs extrêmes de λ, nous observons de moins bonnes performances. Il semble donc qu'un compromis entre les *soft-* et les *hard-targets* est préférable.

7.5 Conclusion

Dans ce travail, nous avons proposé un nouvel espace de représentation, appelé p-vecteur. Ce dernier est appris de manière à mettre en avant les facteurs de variabilité qui caractérisent le personnage/rôle perçu dans la voix. Cet espace de représentation permet d'évaluer la similarité d'une voix

indépendamment de la langue. Nous avons montré qu'il est possible de tirer profit de la connaissance acquise par le biais d'un jeu de données similaire, en la transférant par distillation à notre modèle.

Ces travaux permettent de valider le cadre d'évaluation que nous avons mis en place et que nous avons présenté auparavant. Les contraintes que nous avons établies dans ce cadre nous permettent d'affirmer que l'espace *p*-vecteur contient de l'information propre au personnage. En effet, nous avons montré que nos modèles ont suffisamment de capacité à généraliser, compte-tenu des résultats obtenus sur de nouvelles voix.

Nous avons en plus confirmé les résultats obtenus lors d'une expérience utilisant les *x*-vecteurs, approche aujourd'hui état-de-l'art en reconnaissance automatique du locuteur. De plus, nous avons observé une amélioration par rapport à l'utilisation de *i*-vecteurs pour représenter nos données d'entrées.

La méthode de distillation nous apporte en partie une réponse aux limitations de notre corpus d'évaluation. Certes, il nous faudrait confirmer les résultats présentés jusqu'ici en utilisant un plus gros corpus. À défaut d'avoir plus de données, nous avons en attendant, mis en place tout un protocole expérimental dédié à l'évaluation de la tâche liée au problème du casting vocal. Par ailleurs, au vu des résultats observés, la distillation semble être un bon début de réponse sur le problème de consistence des données. L'apport d'une connaissance généralisée sur un corpus de données proche de celles utilisées pour l'évaluation semble bénéfique pour la caractérisation du personnage.

Chapitre 8

Conclusion et perspectives

Sommaire

8.1 Conclusion

Le choix d'une voix de doublage en remplacement de la voix originale d'un acteur jouant un rôle particulier dans une œuvre donnée est un processus sensible, réalisée par un expert humain. De plus, le processus de casting vocal visant à sélectionner les voix candidates, en accord avec la voix originale et les traits du personnage, est une tâche difficile à formaliser.

Nous avons supposé, en amont des travaux que nous avons présentés, que les choix de voix de doublage utilisées pour les personnages dans la VO et la VF sont guidés, soit par les impressions de similarité dans la voix (notamment pour le choix de la voix en français) en référence à la voix originale, soit parce que les caractéristiques vocales reflètent des traits (biologiques, physiologique, psychique) du personnage. Au travers de ce choix nous retrouvons les grandes questions scientifiques de perception et de similarité de la voix. Ainsi, nous avons mis en place un système nous permettant, à partir des choix de l'opérateur, de mettre en évidence l'information qui caractérise cette notion abstraite de similarité.

La tâche générale que nous nous sommes fixée dans cette thèse, à savoir la mise en avant de la dimension « personnage » dans la voix actée, est une

tâche difficile en elle-même. De plus la caractérisation d'une information de haut niveau représentative de cette dimension de la voix n'a été que très peu étudiée.

Les expériences que nous avons effectuées dés le début de ces travaux nous ont permis de mettre en évidence les principaux biais de nos données. Par exemple, la question du genre et du contenu linguistique. Les contraintes que nous avons alors établies pour la neutralisation de ces biais nous ont aussi amené à un corpus d'apprentissage et d'évaluation très réduit, nous obligeant à définir un cadre d'évaluation adapté. Les résultats que nous avons présentés ne sont pas suffisants pour affirmer avoir résolu la tâche. Cependant, ils sont suffisants pour affirmer que nous avons réussi à extraire de l'information qui explique une part de la variabilité au niveau de la dimension personnage.

Le système de similarité présenté dans le chapitre 6 nous permet de prouver la présence de cette information caractéristique. Puisque nous avons appris un modèle de similarité qui permet de prédire les choix de l'opérateur sur de nouveaux personnages et, donc, sur de nouvelles voix, avec un taux de réussite allant jusqu'à 60 %. Au vu des tests statistiques employés et des valeurs-p obtenues, il apparaît nettement que ces performances ne sont pas le fruit du hasard, mais il est possible que d'autres biais que nous n'aurions pas anticipés aient un impact sur le système. De plus, nous avons supposé qu'un apprentissage par paires, bien qu'il nous permette d'évaluer cet appariement VO-VF, n'est cependant pas adapté pour l'apprentissage d'un espace de représentation caractéristique de la dimension « personnage ». En effet, il semble difficile d'apprendre un espace de représentation mettant en avant les facteurs explicatifs de la variabilité en termes de personnage sur la base des paires de contre-exemples uniquement.

Par la suite, nous avons considéré l'apprentissage d'une telle représentation. De plus, nous avons eu recours à des voix supplémentaires, en complément de notre corpus limité en termes de volume de données et très spécifique en termes de personnages. Ainsi, l'apport de nouvelles voix et de nouveaux labels, associés à l'utilisation de méthodes de transfert de connaissance, nous ont permis d'apprendre une nouvelle représentation, dédiée à notre tâche. Nous avons appelé cette représentation « p-vecteur » pour « personnage-vecteur ».

En définitive, nous avons, dans ces travaux, proposé un cadre méthodologique complet, à la fois pour la projection des voix dans un espace de

représentation propre au personnage et pour l'appariement automatique des voix de doublages. Nous avons ainsi mis en place un premier système de casting vocal automatique. Cependant, les performances obtenues par ce système restent à évaluer dans le cadre d'une utilisation réelle. Des compléments apparaissent nécessaires pour mettre en production un tel système, notamment, en termes d'explicabilité du modèle. Cela est nécessaire pour inscrire notre système en tant que composant d'un système plus large de recommandation automatique de voix, tel que nous l'avions imaginé au départ.

Nous avons eu un aperçu général de l'immense richesse de la voix. Nous avons vu que la voix est un objet très complexe qui véhicule des informations à différents niveaux sur des aspects très divers, en plus d'être aussi le véhicule de la parole. Cela fait de la voix un des facteurs les plus importants de la communication. De plus, la voix n'est utile que si quelqu'un est là pour l'écouter (le locuteur pouvant être son propre auditeur). Cela met en évidence un lien social entre le locuteur et l'auditeur ayant une culture commune, qui s'exprime au travers du langage (au sens large) et donc de la voix. Dans nos travaux, nous avons proposé un moyen technique pour extraire une part d'information caractéristique de ce lien d'appartenance socioculturelle, grâce à un grand volume de données et aux capacités de généralisation des algorithmes d'apprentissage automatique.

Les travaux que nous avons présentés dans cette thèse sont orientés vers le casting vocal. Mais au-delà de ce contexte d'application, nous pouvons tout à fait le transposer à d'autres domaines. En effet, nos travaux montrent plus généralement que nous sommes capables d'extraire des informations de haut niveau sémantique, liées à des caractéristiques et des dimensions abstraites, provenant directement des associations réalisées par le cerveau humain. Dans le cas du casting vocal, cela est rendu possible par le biais d'une décision, celle de l'opérateur de casting détenteur du savoir expert, prise dans le contexte particulier du choix d'une voix adaptée à un personnage particulier. Ainsi, au travers de son choix, l'opérateur de casting exerce un rôle social, celui de médiateur culturel, du fait qu'il permet, entre autres, de perpétuer l'existence de représentations stéréotypées de la voix.

8.2 Perspectives

Dans les travaux que nous avons présentés, nous avons mis en place un cadre de travail dédié à l'extraction d'information caractérisant la dimension « personnage ». Nous n'avons cependant exploré qu'une petite partie des possibles. Si nous prenons le problème à la base, c'est-à-dire en termes d'information brute, nos modèles reposent sur l' information cepstrale à court terme, généralement utilisée pour le traitement automatique de la parole. Il serait donc intéressant de mener des recherches à partir d'autres descripteurs acoustiques.

Nous avons identifié les limites de notre corpus à plusieurs reprises, notamment sur des cas d'évaluations précis. Typiquement les cas où les voix évaluées sont proches ou lorsqu'un seul personnage féminin est présent contre plusieurs personnages masculins. Il nous faudrait pouvoir évaluer sur un plus grand nombre de personnages et ainsi se libérer des contraintes mises en place dans le protocole. Pour l'apprentissage de la représentation *p*-vecteur, nous avons eu recours à des données supplémentaires, cependant, rien n'indique qu'un acteur y joue un personnage unique. Cela nous a donc restreint à une utilisation de ses données en tant que données de transfert du processus de distillation. Effectuer un tri de ces données complémentaires permettrait la constitution d'un corpus d'entraînement indépendant du corpus d'évaluation, ce dernier pouvant alors être utilisé dans sa totalité.

De plus, nous avons observé une variabilité au niveau des résultats obtenus avec le système de similarité. Les réseaux siamois sont, en effet, connus pour leur instabilité. Nous pensons que cette instabilité est causée par le choix des segments de voix (sélectionné aléatoirement) pour construire les paires. L'utilisation d'une technique comme le *curriculum learning* peut aider à la généralisation dans le sens où elle permet de sélectionner les paires plus intelligemment en apprenant progressivement à classifier des exemples de plus en plus difficiles. Le problème auquel nous nous sommes confrontés durant nos expérimentations réside dans l'estimation même de la difficulté des paires. Aujourd'hui, nous pourrions mesurer la difficulté des paires grâce à l'espace des *p*-vecteurs. Par exemple, il serait envisageable de mesurer la difficulté d'une paire de segments de voix en se basant sur leurs distances relatives par rapport aux frontières de décision apprise par le modèle *p*-vecteur.

Une part importante du système de recommandation que nous avions

imaginé réside dans l'explicabilité des prédictions du modèle. Un des problèmes de l'apprentissage des réseaux de neurones profonds est lié aux aspects « boîte-noire » de ces modèles. Du fait de leur structure, les dimensions de l'espace de représentation latent ne sont pas indépendantes. Idéalement, nous aimerions attribuer une dimension de cet espace à un facteur explicatif bien défini. Sur cette question, il serait intéressant d'entraîner différents classifieurs sur un ensemble de tâches visant à reconnaître par exemple les états biologiques du locuteur (genre, age), les aspects dynamiques de la parole comme la prosodie (le débit de parole, le registre et les variations de la F_0) et l'intention (exclamation, affirmation, interrogation, ordre, etc.) voire les états internes du locuteur (joie, peur, tristesse, colère, etc.). Ainsi, les sorties données par les différents classifieurs sont indépendantes et peuvent servir (en les concaténant) de représentation en complément du p-vecteur pour les questions d'explicabilité. Par ailleurs, il serait possible d'utiliser des approches par apprentissage multitâche et *end-to-end* plutôt que plusieurs classifieurs différents. Cela nécessite cependant d'utiliser plusieurs corpus de données provenant de différents domaines ce qui augmente la complexité du système et donc la possibilité de faire des erreurs.

Durant nos travaux, nous avons toujours considéré une voix en tant que personnage individuel et nous n'avons jamais pu nous baser sur une classification plus générale de ces personnages. Il serait intéressant donc de considérer des catégories de personnages. La question qui se profile là, fait donc référence aux catégories de voix et aux types de personnages. Est-il possible de mettre toutes les voix de soldats sous la même étiquette ? Une analyse sociologique devrait permettre de répondre à la question de la réception des voix. De plus la mise en place d'un protocole expérimental reposant sur la perception de différents stimuli vocaux et réalisé sur un échantillon d'individus représentatifs devrait permettre de mieux délimiter et de définir les différentes catégories de personnages.

Nous pourrions aussi voir dans quelle mesure la voix est perçue de la même façon d'une culture à une autre. Typiquement, est-ce qu'il est possible de s'appuyer sur les mêmes stéréotypes vocaux indépendamment de la culture visée et de l'effet recherché ?

À plus long terme, il serait également intéressant de soumettre les voix évaluées à modifications. En effet, les travaux réalisés en perception proposent fréquemment d'étudier l'impact, en termes de perception, de certaines caractéristiques acoustiques en les modifiant indépendamment. Par

exemple, en neutralisant la F_0 en la fixant à un niveau identique sur toutes les voix évaluées.

L'évaluation que nous avons proposée est basée sur les choix effectués par l'opérateur de casting, sensés refléter la vérité en terme de perception de similarité. Pour évaluer notre système de similarité, il serait également intéressant de soumettre les prédictions du système de similarité à un ensemble d'évaluateurs humains. L'évaluation serait ainsi basée, comme dans le cas des travaux en perception, sur l'accord inter-évaluateurs.

Enfin, nos travaux se sont en tout point basés sur l'analyse du signal de parole. Nous n'avons pas traité notre problématique du point de vue classique d'un système de recommandation. Or, nous savons que pour les systèmes de recommandation, trois approches sont communément admises. En premier lieu, les approches basées sur le contenu. L'information est extraite directement depuis le contenu lui-même, dans notre cas il s'agit donc de la voix. Deuxièmement, il y a les approches dites collaboratives qui reposent sur les méta-données (souvent associées aux différents contenus) et les informations issues de l'analyse des profils d'utilisateurs (les avis, les comportements, etc.). Enfin, les approches hybrides qui allient les deux types d'approches. Généralement, les approches basées uniquement sur le contenu sont moins performantes que les approches collaboratives. Toutefois, ces dernières présentent certains problèmes, par exemple le biais de popularité ou le problème du *cold start*, qui conduisent à un manque d'originalité des recommandations du système. En ce qui nous concerne, nous pourrions considérer l'exploitation des bases de données telles qu'IMDb (*Internet Movie Database*). Cette dernière fournit des informations sur un très grand nombre d'acteurs, comme par exemple les personnages qu'ils ont déjà doublés. Il serait donc possible de construire un graphe d'acteurs-personnages et de s'en servir pour faire des rapprochements. De plus, les avis des internautes sur leurs performances d'acteurs, leurs rôles, etc. sont exploitables, notamment grâce à l'utilisation de méthodes de recherche d'information textuelle.

Annexes

Annexe A

Résultats complémentaires

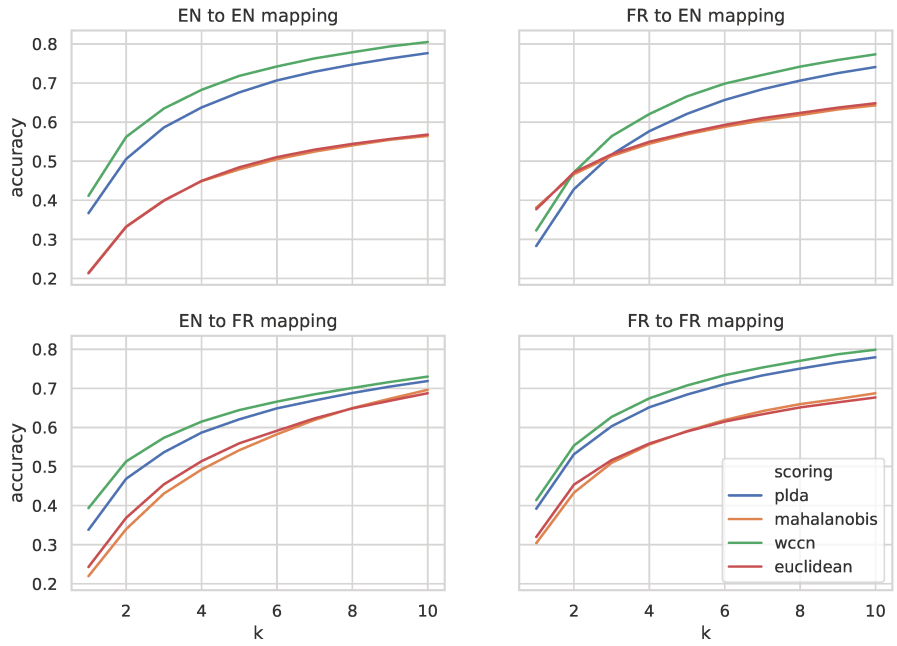

Figure A.1 – Exactitude des prédiction du système A en fonction de k selon différentes méthodes de comparaison.

Mapping	Scoring	System A				System B				System C			
		Fold 1	Fold 1	Fold 3	Mean	Fold 1	Fold 1	Fold 3	Mean	Fold 1	Fold 1	Fold 3	Mean
EN → EN	Euclidean	74,04	64,69	69,65	69,46	71,41	61,99	65,79	66,4	57,3	52,99	51,75	54,01
	Mahalanobis	74	64,38	69,62	69,33	71,20	62,03	65,68	66,30	57,19	52,99	52,41	54,20
	WCCN	26,08	35,5	30,32	**30,63**	28,8	38,04	34,27	**33,70**	42,25	48,48	49,79	**46,84**
	PLDA	29,14	39,16	34,56	34,29	31,10	40,17	37,06	36,11	50,51	57,59	50,46	52,85
EN → FR	Euclidean	53,06	48,74	51,33	51,04	57,26	49,61	53,57	53,48	63,43	55,44	56,90	58,59
	Mahalanobis	53,02	48,63	51,20	50,95	57,35	49,52	53,62	53,50	63,07	55,46	57,25	58,59
	WCCN	46,81	51,40	48,54	48,92	42,61	50,36	46,41	**46,46**	36,40	46,83	45,03	**42,75**
	PLDA	47,11	51,28	47,84	**48,74**	45,96	51,30	46,86	48,04	37,77	48,41	44,80	43,66
FR → EN	Euclidean	53,40	51,19	53,12	52,57	57,05	51,42	52,56	53,68	59,90	51,03	55,72	55,55
	Mahalanobis	53,23	51,07	53,04	52,45	57,14	51,33	52,72	53,73	59,73	51,16	54,94	55,28
	WCCN	46,94	48,91	47,15	47,67	42,86	48,56	47,20	**46,21**	40,70	52,07	41,99	**44,92**
	PLDA	46,56	48,06	46,06	**46,89**	43,88	47,34	47,50	46,24	40,87	48,00	46,44	45,10
FR → FR	Euclidean	71,20	62,13	69,25	67,53	69,75	59,78	64,70	64,74	47,69	41,87	53,46	**47,67**
	Mahalanobis	70,65	61,70	68,34	66,90	68,99	59,07	64,24	64,10	48,03	42,87	53,43	48,11
	WCCN	29,23	38,18	31,28	**32,90**	30,54	40,34	35,49	**35,46**	52,24	56,66	47,19	52,03
	PLDA	29,91	40,08	33,09	34,36	34,37	41,42	38,02	37,94	52,19	55,28	45,43	50,97

Table A.1 – % d'EER.

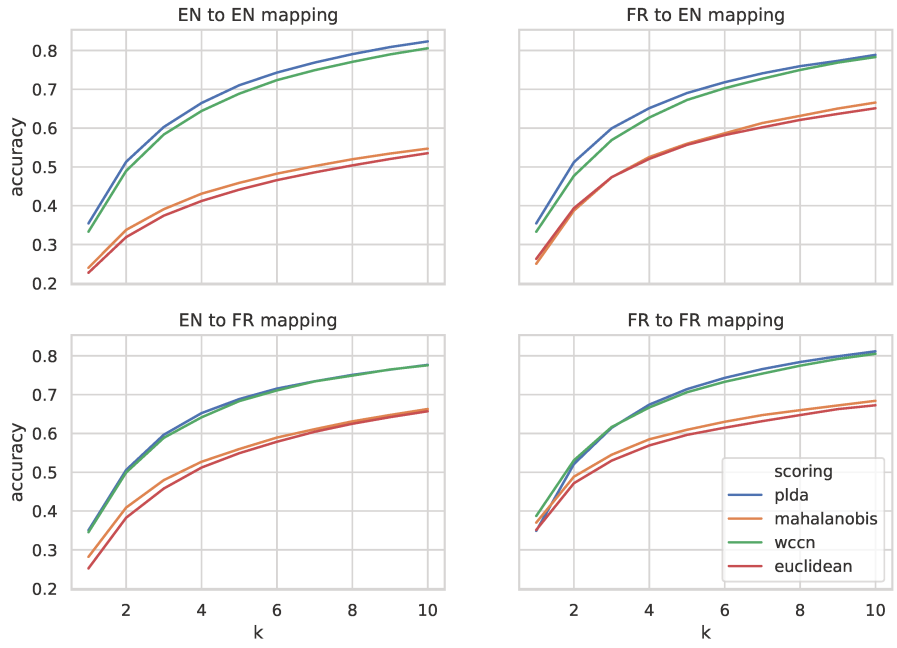

Figure A.2 – Exactitude des prédiction du système *B* en fonction de *k* selon différentes méthodes de comparaison.

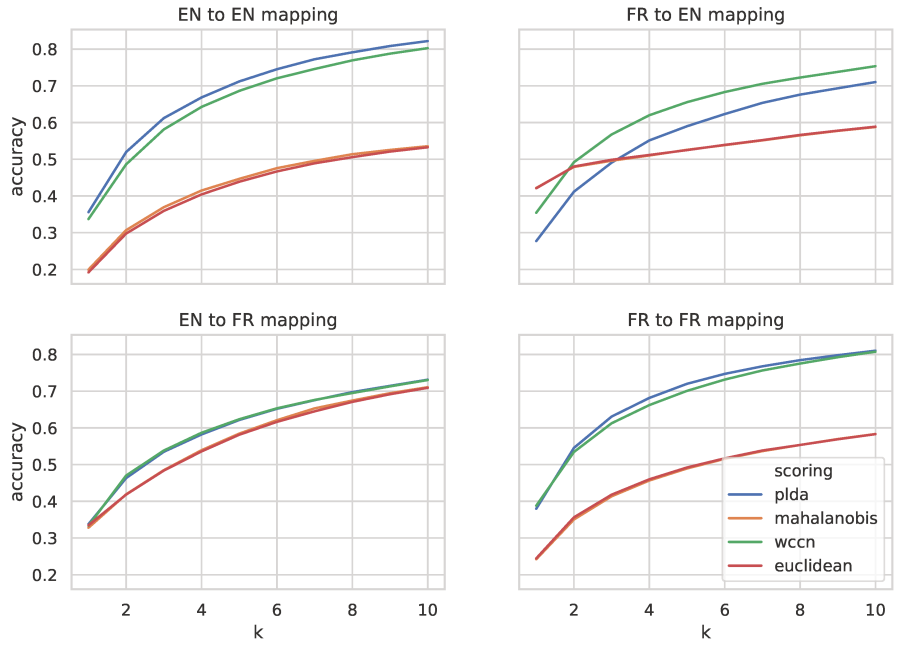

Figure A.3 – Exactitude des prédiction du système *C* en fonction de *k* selon différentes méthodes de comparaison.

Annexe B

Fonction triplet et distance angulaire

Sommaire

B.1 La fonction triplet

Les réseaux triplets sont une variante des réseaux siamois qui est digne d'être mentionnée ici. L'architecture met en œuvre trois réseaux identiques et qui partagent leurs paramètres comme dans l'architecture classique. Néanmoins, au lieu d'utiliser des paires d'entrées, elle prend en entrée un triplet $(x_p\, x_a\, x_n)$ que l'on note τ où x_p et x_a sont deux exemples d'une même catégorie (similaires) et x_n est un contre-exemple (d'une catégorie différente). En considérant ces entrées on souhaite donc faire en sorte que la distance séparant x_p et x_a soit inférieure à la distance entre x_a et x_n :

$$D_W(x_p, x_a) < D_W(x_a, x_n). \tag{B.1}$$

Cette fonction triplet, proposée par Wang et al. 2014 est définie ainsi :

$$\Delta_\tau = \| G_W(x_a) - G_W(x_p) \|_2^2 - \| G_W(x_a) - G_W(x_n) \|_2^2 \tag{B.2}$$

En considérant $\mathcal{T}$ l'ensemble des triplets $\tau = (x_p\, x_a\, x_n)$ on cherche alors à

minimiser la fonction suivante :

$$L(\mathcal{T}) = \sum_{\tau \in \mathcal{T}} max(0, \Delta_\tau + \alpha) \tag{B.3}$$

où α est tel que $\Delta_\tau + \alpha < 0$.

B.2 Mesure de distance

Dans la littérature, il est souvent fait mention de l'utilisation des distances angulaires basées sur la similarité cosinus :

$$\text{similarité} = cos(x_1, x_2) = \frac{x_1 . x_2}{\|x_1\| . \|x_2\|} \tag{B.4}$$

qui est bornée et que l'on peut facilement mettre sous la forme d'une distance angulaire :

$$\text{distance angulaire} = \frac{\cos^{-1}(\text{similarité})}{\pi}. \tag{B.5}$$

Annexe C

Information privilégiée

Dans un monde idéal, l'apprentissage automatique peut se résumer assez simplement. Il s'agit de trouver parmi un ensemble de fonctions, celle qui approxime le mieux la bonne décision à partir d'un ensemble d'exemples d'apprentissage. Dans le monde des humains, la prise de décision se fait aussi à partir des différentes expériences passées qui permettent au fur et à mesure d'apprendre à choisir la décision qui sera la plus adaptée à un exemple de situation particulier. Dans un grand nombre de cas, l'apprentissage de l'*Élève* se fait à l'aide d'un référent, le *Maître*, qui maîtrise déjà le problème. C'est lui qui enseigne, guide, conseille, commente. On suppose alors que des mécanismes spécifiques d'apprentissage sont induits à partir de l'interaction entre le *Maître* et l'*Élève*. Le paradigme d'apprentissage automatique par information privilégiée, qui est proposé par Vapnik et al. 2015, consiste donc à fournir à l'étape d'entraînement, une information supplémentaire que l'on note $x^\star$ à propos d'un exemple d'apprentissage x. Cette information est donnée par un *Maître* intelligent.

De manière générale, un problème d'apprentissage automatique peut se résumer à apprendre à partir d'un ensemble de données :

$$(x_1,y_1),...,(x_i,y_i),\ x_i \in X,\ y_i \in \{0,1\} \tag{C.1}$$

qui sont générées à partir d'une fonction de probabilité $P(x,y)$, la fonction $y = f^\star_\theta(x)$ qui minimise la probabilité de mauvaise prédiction. Ici, les vecteurs $x_i \in X$ décrivent les exemples générés selon une probabilité $P(x)$ et $y_i \in \{0,1\}$ sa classification donnée par une probabilité conditionnelle $P(y|x)$.

Le cadre d'apprentissage avec information privilégiée, décrit un modèle

plus complexe qui, à partir d'un ensemble de triplets :

$$(x_1,x_1^\star,y_1),...,(x_i,x_i^\star,y_i),\ x_i \in X,\ x_i^\star \in X^\star,\ y_i \in \{0,1\} \tag{C.2}$$

générés à partir d'une fonction de probabilité $P(x,x^\star,y)$ cherche à trouver la fonction paramétrique $y = f_\theta^\star(x), \theta \in \Theta$ qui garantit la plus petite probabilité possible de fausses prédictions.

À la différence du cadre classique d'apprentissage, on se retrouve durant la phase d'entraînement avec plus d'information puisqu'à la place des simples paires exemple-label (x,y) on a des triplets $(x,x^\star,y)$. Ainsi, pour chaque exemple d'apprentissage (x_i,y_i), le *Maître* génère l'information privilégiée $x_i^\star$ à partir d'une probabilité conditionnelle inconnue $P(x_i^\star|x_i)$.

Le fait est que l'*Élève* n'a pas accès à l'information privilégiée du *Maître* au moment du test. Le cadre d'apprentissage proposé par Vapnik cherche donc à tirer parti de l'information $x^\star$ à l'étape d'apprentissage pour construire un modèle qui soit capable de surpasser au moment du test, celui appris seulement à partir des paires (x_i,y_i).

Glossaire

ADR Automated Dialogue Recording 5, 6

BF Big Five 46–48, 155

BNF Bottleneck Features 31

CMN Cepstral Mean Normalization 24, 78

CMVN Cepstral Mean Variance Normalization 24

CNN Réseau de Neurones Convolutifs 44, 48

DA Direction Artistique 5, 8, 9, 11, 66, 98, 118, 119

DET Detection Error Tradeoff 20

DNN Réseau de Neurones Profond (Deep Neural Network) 31, 32, 44, 117

EER Equal Error Rate 21, 83, 103, 144, 155

EFR Eigen Factor Radial 31

EM Espérance-Maximisation 27

FA Fausse Alarme 20, 21, 29

FPR False Positive Rate 83

FR Faux Rejet 20, 21

GMM Modèle de Mélange Gaussien 26–29, 43, 44, 153

HMM Modèle de Markov Caché (Hidden Markov Model) 44

HNR Rapport Signal sur Bruit (Harmonic to Noise Ratio) 43

IHM Interaction Homme-Machine 45

JFA Joint Factor Analysis 29, 30

LDA Analyse Discriminante Linéaire 30

LR Rapport de Vraisemblance 19, 73

LSTM Long Short-Term Memory 44

MAP Maximum A Posteriori 27, 28

MDS Positionnement Multidimensionnel (Multidimensional Scaling) 53, 56, 57

MFCC Mel Frequency Cepstral Coefficients 22, 23, 43, 78, 117

MLP Perceptron Multi-Couche 127, 128

NIST National Institute of Standard and Technology 21, 78, 79, 103

PLDA Analyse Discriminante Linéaire Probabiliste 30, 31, 73, 79, 82, 83, 92, 93

RAL Reconnaissance Automatique du Locuteur 18, 21, 23, 28, 29, 31, 33, 66, 70, 72, 73, 78, 82, 117

RNN Réseau de Neurones Récurrents 44

RPG Jeu de Rôle (Role Playing Game) 71

SDC Shifted Delta Cepstra 23

SNN Réseaux de Neurones Siamois (Siamese Neural Networks) 94, 95, 97, 105, 110–112, 155

SRE Speaker Recognition Evaluation 21, 78, 79, 103

SVM Machine à Vecteurs de Support 28, 44, 47, 59

TDNN Time Delay Neural Network 33

TPR True Positive Rate 83

UBM Modèle du Monde (Universal Background Model) 27–29, 75, 79, 153

VA Valence-Activation 40, 42

VAD Vocal Activity Detection 23, 78

VF Version Française 7, 8, 66, 76, 77, 79, 80, 83, 89, 98, 99, 101, 104, 110, 123, 124, 135, 136

VO Version Originale 6–8, 66, 76, 77, 79, 80, 83, 89, 98, 99, 101, 104, 110, 123, 124, 135, 136

WCCN Within-Class Covariance Normalization 28, 30, 73, 82

Liste des figures

Liste des tableaux

Bibliographie

Ouvrages de référence

[Abi05] Jean Abitbol. *L'odyssée de la voix*. Robert Laffont, 2005. isbn : 2081289571.

[Ada+08a] Yoshihiro Adachi, Shinichi Kawamoto, Shigeo Morishima et Satoshi Nakamura. « Perceptual similarity measurement of speech by combination of acoustic features ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2008, p. 4861-4864. url : https://doi.org/10.1109/ICASSP.2008.4518746.

[Ada+08b] Yoshihiro Adachi, Shinichi Kawamoto, Shigeo Morishima et Satoshi Nakamura. « Perceptual similarity measurement of speech by combination of acoustic features ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2008, p. 4861-4864. url : https://doi.org/10.1109/ICASSP.2008.4518746.

[AIG15] Christos-Nikolaos Anagnostopoulos, Theodoros Iliou et Ioannis Giannoukos. « Features and classifiers for emotion recognition from speech : a survey from 2000 to 2011 ». In : *Artificial Intelligence Review* 43.2 (2015), p. 155-177. url : https://doi.org/10.1007/s10462-012-9368-5.

[AJ01] A Adamson et V Jenson. *Schrek*. Sous la dir. de J Katzenberg. DreamWorks SKG, 2001.

[Asa+17] Taichi Asami, Ryo Masumura, Yoshikazu Yamaguchi, Hirokazu Masataki et Yushi Aono. « Domain adaptation of dnn acoustic models using knowledge distillation ». In : *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017. url : https://doi.org/10.1109/ICASSP.2017.7953145.

[ASA06] Kanae Amino, Tsutomu Sugawara et Takayuki Arai. « Speaker similarities in human perception and their spectral properties ». In : *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WESPAC)*. T. 9. 2006.

[AV15] Namrata Anand et Prateek Verma. « Convoluted feelings convolutional and recurrent nets for detecting emotion from audio data ». In : *Technical Report*. Stanford University, 2015.

[BAK17] Gautam Bhattacharya, Md Jahangir Alam et Patrick Kenny. « Deep Speaker Embeddings for Short-Duration Speaker Verification ». In : *18th Annual Conference of the International Speech Communication Association*. 2017, p. 1517-1521.

[Bat07] C.M Bateman. *Game Writing : Narrative Skills for Videogames*. Applied English Series. Charles River Media, 2007. isbn : 9781584504900.

[BB03] Mathieu Ben et Frédéric Bimbot. « D-MAP : A distance-normalized MAP estimation of speaker models for automatic speaker verification ». In : *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. T. 2. IEEE. 2003, p. II-69.

[BB10] Oliver Baumann et Pascal Belin. « Perceptual scaling of voice identity : common dimensions for different vowels and speakers ». In : *Psychological Research PRPF* 74.1 (2010), p. 110. url : https://doi.org/10.1007/s00426-008-0185-z.

[BCV13] Yoshua Bengio, Aaron Courville et Pascal Vincent. « Representation learning : A review and new perspectives ». In : *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), p. 1798-1828. url : https://doi.org/10.1109/TPAMI.2013.50.

[Bel+00] Pascal Belin, Robert J Zatorre, Philippe Lafaille, Pierre Ahad et Bruce Pike. « Voice-selective areas in human auditory cortex ». In : *Nature* 403.6767 (2000), p. 309. url : https://doi.org/10.1038/35002078.

[Ben+09] Yoshua Bengio, Jérôme Louradour, Ronan Collobert et Jason Weston. « Curriculum learning ». In : *26th Annual International Conference on Machine Learning*. ACM. 2009, p. 41-48. url : https://doi.org/10.1145/1553374.1553380.

[Ber92] Diane S Berry. « Vocal types and stereotypes : Joint effects of vocal attractiveness and vocal maturity on person perception ». In : *Journal of Nonverbal Behavior* 16.1 (1992), p. 41-54. url : https://doi.org/10.1007/BF00986878.

[BFB04] Pascal Belin, Shirley Fecteau et Catherine Bedard. « Thinking the voice : Neural correlates of voice perception ». In : *Trends in Cognitive Sciences* 8.3 (2004), p. 129-135. url : https://doi.org/10.1016/j.tics.2004.01.008.

[Bim+04] Frédéric Bimbot, Jean-François Bonastre, Corinne Fredouille, Guillaume Gravier, Ivan Magrin-Chagnolleau, Sylvain Meignier, Teva Merlin, Javier Ortega-García, Dijana Petrovska-Delacrétaz et Douglas A Reynolds. « A tutorial on text-independent speaker verification ». In : *EURASIP Journal on Advances in Signal Processing* 2004.4 (2004), p. 101962. url : https://doi.org/10.1155/S1110865704310024.

[BL13] Martin Barnier et Isabelle Le Corff. « Introduction-Le cinéma européen et les langues ». In : *Mise au point. Cahiers de l'association française des enseignants et chercheurs en cinéma et audiovisuel* 5 (2013). url : https://doi.org/10.4000/map.1372.

[BMB11] Pierre-Michel Bousquet, Driss Matrouf et Jean-François Bonastre. « Intersession compensation and scoring methods in the i-vectors space for speaker recognition ». In : *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[BMK14] Molly Babel, Grant McGuire et Joseph King. « Towards a more nuanced view of vocal attractiveness ». In : *PloS one* 9.2 (2014), e88616. url : https://doi.org/10.1371/journal.pone.0088616.

[BMV+12] Mohamad Hasan Bahari, M.L McLaren, D.A Van Leeuwen et al. « Age estimation from telephone speech using i-vectors ». In : *13th Annual Conference of the International Speech Communication Association*. 2012.

[BN08] Carlos Busso et Shrikanth S Narayanan. « The expression and perception of emotions : Comparing assessments of self versus others ». In : 2008, p. 257-260.

[Boë00] Louis-Jean Boë. « Forensic voice identification in France ». In : *Speech Communication* 31.2-3 (2000), p. 205-224. url : https://doi.org/10.1016/S0167-6393(99)00079-5.

[Bon+03] Jean-François Bonastre, Frédéric Bimbot, Louis-Jean Boë, Joseph P Campbell, Douglas A Reynolds et Ivan Magrin-Chagnolleau. « Person authentication by voice : A need for caution ». In : *Eighth European Conference on Speech Communication and Technology*. 2003.

[Bon+15] Jean-François Bonastre, Juliette Kahn, Solange Rossato et Moez Ajili. « Forensic speaker recognition : Mirages and reality ». In : (2015). Sous la dir. de S. Fuchs/D, p. 255. url : https://doi.org/10.3726/978-3-653-05777-5.

[Bon14] Bérénice Bonhomme. « Les stars et le cinéma d'animation ». In : *Mise au point. Cahiers de l'association française des enseignants et chercheurs en cinéma et audiovisuel* 6 (2014). url : https://doi.org/10.4000/map.1736.

[Bou+12] Pierre-Michel Bousquet, Anthony Larcher, Driss Matrouf, Jean-François Bonastre et Oldřich Plchot. « Variance-spectra based normalization for i-vector standard and probabilistic linear discriminant analysis ». In : *Odyssey Proceedings*. 2012.

[Bou82] Pierre Bourdieu. *Ce que parler veut dire : l'économie des échanges linguistiques*. Fayard, 1982. isbn : 2213012164.

[Bro+94] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger et Roopak Shah. « Signature verification using a "siamese" time delay neural network ». In : *Advances in Neural Information Processing Systems* (1994), p. 737-744. URL : https://doi.org/10.1142/S0218001493000339.

[Bru+10] Laetitia Bruckert, Patricia Bestelmeyer, Marianne Latinus, Julien Rouger, Ian Charest, Guillaume A Rousselet, Hideki Kawahara et Pascal Belin. « Vocal attractiveness increases by averaging ». In : *Current Biology* 20.2 (2010), p. 116-120. URL : https://doi.org/10.1016/j.cub.2009.11.034.

[Bru56] Egon Brunswik. *Perception and the representative design of psychological experiments*. University of California Press, 1956.

[Bur17] Neil Burger. *The Upside*. Sous la dir. de T. Black, J Blumenthal et S. Tisch. The Weinstein Company, 2017.

[Bus+08] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee et Shrikanth S Narayanan. « IEMOCAP : Interactive emotional dyadic motion capture database ». In : *Language resources and evaluation* 42.4 (2008), p. 335. URL : https://doi.org/10.1007/s10579-008-9076-6.

[Cam+09] Joseph P Campbell, Wade Shen, William M Campbell, Reva Schwartz, Jean-François Bonastre et Driss Matrouf. « Forensic speaker recognition ». In : IEEE. 2009. URL : https://doi.org/10.1109/MSP.2008.931100.

[Cam97] Joseph P Campbell. « Speaker recognition : A tutorial ». In : *Proceedings of the IEEE* 85.9 (1997), p. 1437-1462. URL : https://doi.org/10.1109/5.628714.

[CC14] Sandra Cornaz et Diane Caussade. « Musique, voix chantée et apprentissage : une revue de littérature et quelques propositions d'exploitation en didactique de la phonétique des langues ». In : (2014).

[Chi82] Michel Chion. *La voix au cinéma*. T. 1. Editions de l'Etoile, 1982. ISBN : 978-2-8664-2004-8.

[CHL05] Sumit Chopra, Raia Hadsell et Yann LeCun. « Learning a similarity metric discriminatively, with application to face verification ». In : *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2005, p. 539-546. URL : https://doi.org/10.1109/CVPR.2005.202.

[Cho+15] François Chollet et al. *Keras*. 2015.

[CNZ18] Joon Son Chung, Arsha Nagrani et Andrew Zisserman. « VoxCeleb2 : Deep Speaker Recognition ». In : *19th Annual Conference of the International Speech Communication Association*. 2018.

[Cou01] Pierre Couprie. *Le vocabulaire de l'objet sonore*. 2001.

[CSR06] William M Campbell, Douglas E Sturim et Douglas A Reynolds. « Support vector machines using GMM supervectors for speaker verification ». In : *IEEE signal processing letters* 13.5 (2006), p. 308-311. url : https://doi.org/10.1109/LSP.2006.870086.

[DDK07] Najim Dehak, Pierre Dumouchel et Patrick Kenny. « Modeling prosodic features with joint factor analysis for speaker verification ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 15.7 (2007), p. 2095-2103. url : https://doi.org/10.1109/TASL.2007.902758.

[DDP93] Peter B Denes, Peter Denes et Elliot Pinson. *The speech chain*. Macmillan, 1993.

[Deh+09] Najim Dehak, Reda Dehak, Patrick Kenny, Niko Brümmer, Pierre Ouellet et Pierre Dumouchel. « Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification ». In : *Tenth Annual Conference of the International Speech Communication Association*. 2009.

[Deh+10] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel et Pierre Ouellet. « Front-end factor analysis for speaker verification ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2010), p. 788-798.

[Deh+11a] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel et Pierre Ouellet. « Front end factor analysis for speaker verification ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (2011), p. 788-798. url : https://doi.org/10.1109/TASL.2010.2064307.

[Deh+11b] Najim Dehak, Pedro A Torres-Carrasquillo, Douglas Reynolds et Reda Dehak. « Language recognition via i-vectors and dimensionality reduction ». In : *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[Des13] Troy G Deskins. « Stereotypes in video games and how they perpetuate prejudice ». In : *McNair Scholars Research Journal* 6.1 (2013), p. 5.

[DO13] Gilles Degottex et Nicolas Obin. « Phase distortion statistics as a representation of the glottal source : Application to the classification of voice qualities ». In : *14th Annual Conference of the International Speech Communication Association*. 2013.

[DP72] Charles Darwin et Phillip Prodger. *The expression of the emotions in man and animals*. Oxford University Press, USA, 1872.

[DT07] Karen E Dill et Kathryn P Thill. « Video game characters and the socialization of gender roles : Young people's perceptions mirror sexist media depictions ». In : *Sex roles* 57.11-12 (2007), p. 851-864. url : https://doi.org/10.1007/s11199-007-9278-1.

[EKK11a] Moataz El Ayadi, Mohamed S. Kamel et Fakhri Karray. « Survey on speech emotion recognition : Features, classification schemes, and databases ». In : *Pattern Recognition* 44.3 (2011), p. 572-587. URL : https://doi.org/10.1016/j.patcog.2010.09.020.

[EKK11b] Moataz El Ayadi, Mohamed S Kamel et Fakhri Karray. « Survey on speech emotion recognition : Features, classification schemes, and databases ». In : *Pattern Recognition* 44.3 (2011), p. 572-587. URL : https://doi.org/10.1016/j.patcog.2010.09.020.

[Ekm99] Paul Ekman. « Basic emotions ». In : *Handbook of cognition and emotion* 98.45-60 (1999), p. 16.

[FK12] Hanna S Feiser et Felicitas Kleber. « Voice similarity among brothers : evidence from a perception experiment ». In : *21st Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*. 2012.

[FMC08] Liqin Fu, Xia Mao et Lijiang Chen. « Speaker independent emotion recognition based on SVM/HMMs fusion system ». In : *International Conference on Audio, Language and Image Processing (ICASSP)*. IEEE. 2008, p. 61-65. URL : https://doi.org/10.1109/ICALIP.2008.4590144.

[Fon+07] Johnny RJ Fontaine, Klaus R Scherer, Etienne B Roesch et Phoebe C Ellsworth. « The world of emotions is not two-dimensional ». In : *Psychological science* 18.12 (2007), p. 1050-1057. URL : https://doi.org/10.1111/j.1467-9280.2007.02024.x.

[FSS10] Luciana Ferrer, Nicolas Scheffer et Elizabeth Shriberg. « A comparison of approaches for modeling prosodic features in speaker recognition ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2010, p. 4414-4417. URL : https://doi.org/10.1109/ICASSP.2010.5495632.

[Fur81] Sadaoki Furui. « Cepstral analysis technique for automatic speaker verification ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 29.2 (1981), p. 254-272. URL : https://doi.org/10.1109/TASSP.1981.1163530.

[Gam04] Yves Gambier. « La traduction audiovisuelle : un genre en expansion ». In : *Meta : journal des traducteurs/Meta : Translators' Journal* 49.1 (2004), p. 1-11. URL : https://doi.org/10.7202/009015ar.

[GB10] Xavier Glorot et Yoshua Bengio. « Understanding the difficulty of training deep feedforward neural networks ». In : *Thirteenth International Conference on Artificial Intelligence and Statistics*. Sous la dir. d'Yee Whye Teh et Mike Titterington. T. 9. PMLR, 2010, p. 249-256. URL : http://proceedings.mlr.press/v9/glorot10a.html.

[GC09] Dayana Ribas Gonzalez et José R Calvo de Lara. « Speaker verification with shifted delta cepstral features : Its Pseudo-Prosodic Behaviour ». In : *First Iberian SLTech*. 2009.

[GE11] Daniel Garcia-Romero et Carol Y Espy-Wilson. « Analysis of i-vector length normalization in speaker recognition systems ». In : *Twelfth Annual Conference of the International Speech Communication Association*. 2011.

[GK17] Rolf Giesen et Anna Khan. *Acting and Character Animation : The Art of Animated Films, Acting and Visualizing*. CRC Press, 2017. isbn : 1498778631.

[GL94] Jean-Luc Gauvain et Chin-Hui Lee. « Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains ». In : *IEEE transactions on speech and audio processing* 2.2 (1994), p. 291-298. url : https://doi.org/10.1109/89.279278.

[Gol05] B Golse. « Les précurseurs corporels et comportementaux du langage verbal ». In : *Au commencement était la voix* (2005). url : https://doi.org/10.1016/j.neurenf.2005.09.022.

[Gun+11] Hatice Gunes, Björn Schuller, Maja Pantic et Roddy Cowie. « Emotion representation, analysis and synthesis in continuous space : A survey ». In : *Face and Gesture 2011*. IEEE. 2011, p. 827-834.

[Hat+06] Jean-Paul Haton, Christophe Cerisara, Dominique Fohr, Yves Laprie et Kamel Smaïli. *Reconnaissance automatique de la parole : Du Signal à son Interprétation*. Dunod, 2006. isbn : 2-10-052803-3.

[HCL06] Raia Hadsell, Sumit Chopra et Yann LeCun. « Dimensionality reduction by learning an invariant mapping ». In : *Computer vision and pattern recognition, 2006 IEEE computer society conference on*. T. 2. IEEE. 2006, p. 1735-1742. url : https://doi.org/10.1109/CVPR.2006.100.

[HD10] Ali Hassan et Robert I Damper. « Multi-class and hierarchical SVMs for emotion recognition ». In : *11th Annual Conference of the International Speech Communication Association*. 2010.

[HDG04] Susan M. Hughes, Franco Dispenza et Gordon G. Gallup. « Ratings of voice attractiveness predict sexual behavior and body configuration ». In : *Evolution and Human Behavior* 25.5 (2004), p. 295-304. url : https://doi.org/10.1016/j.evolhumbehav.2004.06.001.

[Hen01] Nathalie Henrich. « Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception ». Thèse de doct. Paris 6, 2001.

[Hev36] Kate Hevner. « Experimental studies of the elements of expression in music ». In : *American journal of Psychology* 48.2 (1936), p. 246-268.

[HGP11] Carolyn R Hodges-Simeon, Steven JC Gaulin et David A Puts. « Voice correlates of mating success in men : examining "contests" versus "mate choice" modes of sexual selection ». In : *Archives of Sexual Behavior* 40.3 (2011), p. 551-557. url : https://doi.org/10.1007/s10508-010-9625-0.

[HH15] John H.L Hansen et Taufiq Hasan. « Speaker recognition by machines and humans : A tutorial review ». In : *IEEE Signal processing magazine* 32.6 (2015), p. 74-99. url : https://doi.org/10.1109/MSP.2015.2462851.

[HKS06] Andrew O Hatch, Sachin Kajarekar et Andreas Stolcke. « Within-class Covariance Normalization for SVM-based Speaker Recognition ». In : *Proceedings of ICSLP*. 2006, p. 1471-1474.

[HP16] Alexander K Hill et David A Puts. « Vocal attractiveness ». In : *Encyclopedia of Evolutionary Psychological Science, eds V. Weekes-Shackelford, TK Shackelford, and VA Weekes-Shackelford (Cham : Springer International Publishing)* (2016), p. 1-5. doi : 10.1007/978-3-319-16999-6_1880-1.

[Hua+19] Kun-Yi Huang, Chung-Hsien Wu, Qian-Bei Hong, Ming-Hsiang Su et Yi-Hsuan Chen. « Speech Emotion Recognition Using Deep Neural Network Considering Verbal and Nonverbal Speech Sounds ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, p. 5866-5870.

[HVD14] Geoffrey Hinton, Oriol Vinyals et Jeffrey Dean. « Distilling the Knowledge in a Neural Network ». In : (2014).

[IM15] Yusuke Ijima et Hideyuki Mizuno. « Similar speaker selection technique based on distance metric learning using highly correlated acoustic features with perceptual voice quality similarity ». In : *IEICE TRANSACTIONS on Information and Systems* 98.1 (2015), p. 157-165. url : https://doi.org/10.1587/transinf.2014EDP7183.

[JM07] Jeroen Jansz et Raynel G Martis. « The Lara phenomenon : Powerful female characters in video games ». In : *Sex roles* 56.3-4 (2007), p. 141-148. url : https://doi.org/10.1007/s11199-006-9158-0.

[Joy+17] Neethu Mariam Joy, Sandeep Reddy Kothinti, Srinivasan Umesh et Basil Abraham. « Generalized distillation framework for speaker normalization ». In : *18th Annual Conference of the International Speech Communication Association*. 2017.

[JS99] Tom Johnstone et Klaus R Scherer. « The effects of emotions on voice quality ». In : *14th International Congress of Phonetic Sciences*. 1999, p. 2029-2032.

[Kah+11] Juliette Kahn, Nicolas Audibert, Solange Rossato et Jean-François Bonastre. « Speaker verification by inexperienced and experienced listeners vs. speaker verification system ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2011, p. 5912-5915.

[KB05] Neil T Kleynhans et Etienne Barnard. « Language dependence in multilingual speaker verification ». In : *Sixteenth Annual Symposium of the Pattern Recognition Association of South Africa*. PRASA, 2005.

[KD04] Patrick Kenny et Pierre Dumouchel. « Disentangling speaker and channel effects in speaker verification ». In : *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE. 2004, p. 37-40.

[Kel+16] Finnian Kelly, Anil Alexander, Oscar Forth, Samuel Kent, Jonas Lindh et Joel Åkesson. « Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features ». In : *17th Annual Conference of the International Speech Communication Association*. 2016, p. 1567-1568.

[Ken+14] Patrick Kenny, Themos Stafylakis, Pierre Ouellet, Vishwa Gupta et Md Jahangir Alam. « Deep Neural Networks for extracting Baum-Welch statistics for Speaker Recognition ». In : *Odyssey Proceedings*. 2014, p. 293-298.

[Ken05] Patrick Kenny. « Joint factor analysis of speaker and session variability : Theory and algorithms-technical report CRIM-06/08-13 ». In : *Montreal, CRIM* 2005 (2005).

[Ken10] Patrick Kenny. « Bayesian speaker verification with heavy-tailed priors ». In : *Odyssey Proceedings*. 2010.

[Ker62] Lawrence George Kersta. « Voiceprint identification ». In : *The Journal of the Acoustical Society of America* 34.5 (1962), p. 725-725. url : https://doi.org/10.1038/1961253a0.

[KFM02] Robert M Krauss, Robin Freyberg et Ezequiel Morsella. « Inferring speakers' physical attributes from their voices ». In : *Journal of Experimental Social Psychology* 38 (2002), p. 618-625. url : https://doi.org/10.1016/S0022-1031(02)00510-3.

[KH10] Andreas M Kaplan et Michael Haenlein. « Users of the world, unite ! The challenges and opportunities of Social Media ». In : *Business horizons* 53.1 (2010), p. 59-68. url : https://doi.org/10.1016/j.bushor.2009.09.003.

[Kin+06] Tomi Kinnunen, Chin Wei, Eugene Koh, Lei Wang, Haizhou Li et Eng Siong Chng. « Temporal discrete cosine transform : Towards longer term temporal features for speaker verification ». In : *Fifth International Symposium on Chinese Spoken Language Processing (ISCSLP)*. 2006, p. 547-558.

[KK01] H Kido et H Kasuya. « Everyday expressions associated with voice quality of normal utterance—Extraction by perceptual evaluation ». In : *Acoustic Society of Japan* 57.5 (2001), p. 337-344.

[KMD03] Patrick Kenny, Mohamed Mihoubi et Pierre Dumouchel. « New MAP estimators for speaker recognition ». In : *Eighth European Conference on Speech Communication and Technology*. 2003.

[KPC13] Vera Kempe, David A Puts et Rodrigo A Cárdenas. « Masculine men articulate less clearly ». In : *Human Nature* 24.4 (2013), p. 461-475. url : https://doi.org/10.1007/s12110-013-9183-y.

[KR12] Shashidhar G Koolagudi et K Sreenivasa Rao. « Emotion recognition from speech : a review ». In : *International journal of speech technology* 15.2 (2012), p. 99-117. url : https://doi.org/10.1007/s10772-011-9125-1.

[Kre18] Jody E Kreiman. « Reconsidering the nature of voice ». In : *The Journal of the Acoustical Society of America* 144.3 (2018), p. 1765-1765. url : https://doi.org/10.1121/1.5067809.

[Kul+13] Brian Kulis et al. *Metric learning : A survey*. T. 5. 4. Now Publishers, Inc., 2013, p. 287-364. url : https://doi.org/10.1561/2200000019.

[KZS15] Gregory Koch, Richard Zemel et Ruslan Salakhutdinov. « Siamese neural networks for one-shot image recognition ». Mém. de mast. University of Toronto, 2015.

[Lav80] John Laver. *The phonetic description of voice quality : Cambridge Studies in Linguistics*. Cambridge University Press, Cambridge, 1980. isbn : 0521231760.

[LB11] Marianne Latinus et Pascal Belin. « Human voice perception ». In : *Current Biology* 21.4 (2011), R143-R145. url : https://doi.org/10.1016/j.cub.2010.12.033.

[Le 11] David Le Breton. *Eclats de voix. Une anthropologie des voix*. Métailié, 2011. isbn : 2864248425.

[LE10] Jonas Lindh et Anders Eriksson. « Voice similarity-a comparison between judgements by human listeners and automatic voice comparison ». In : *Proceedings from FONETIK*. 2010, p. 63-69.

[Leg05a] Adrien M Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

[Leg05b] Adrien Marie Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.

[Lei+14] Yun Lei, Nicolas Scheffer, Luciana Ferrer et Mitchell McLaren. « A novel scheme for speaker recognition using a phonetically-aware deep neural network ». In : *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2014, p. 1695-1699.

[Li+17a] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan et Zhenyao Zhu. « Deep Speaker : an End-to-End Neural Speaker Embedding System ». In : (2017). url : https://arxiv.org/abs/1705.02304.

[Li+17b] Jinyu Li, Michael L Seltzer, Xi Wang, Rui Zhao et Yifan Gong. « Large-scale domain adaptation via teacher-student learning ». In : 2017.

[LJY09] Marko Lugger, Marie-Elise Janoir et Bin Yang. « Combining classifiers with diverse feature sets for robust speaker independent emotion recognition ». In : *17th European Signal Processing Conference*. IEEE. 2009, p. 1225-1229.

[LNH10] Iker Luengo, Eva Navas et Inmaculada Hernáez. « Feature analysis and evaluation for automatic emotion identification in speech ». In : *IEEE Transactions on Multimedia* 12.6 (2010), p. 490-501. url : https://doi.org/10.1109/TMM.2010.2051872.

[Loa08] Deborah Loakes. « A forensic phonetic investigation into the speech patterns of identical and non-identical twins ». In : *International Journal of Speech, Language and the Law* 15.1 (2008), p. 97-100. url : http://dx.doi.org/10.1558/ijsll.v15i1.97.

[Lop+16] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf et Vladimir Vapnik. « Unifying distillation and privileged information ». In : *International Conference on Learning Representations*. 2016.

[Mar+12] David Martinez, Lukáš Burget, Luciana Ferrer et Nicolas Scheffer. « iVector-based prosodic system for language identification ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2012, p. 4861-4864. url : https://doi.org/10.1109/ICASSP.2012.6289008.

[MBM08] Driss Matrouf, Jean-François Bonastre et Salah Eddine Mezaache. « Factor analysis multi-session training constraint in session compensation for speaker verification ». In : *Ninth Annual Conference of the International Speech Communication Association*. 2008.

[MBZ17] Seyedmahdad Mirsamadi, Emad Barsoum et Cha Zhang. « Automatic speech emotion recognition using recurrent neural networks with local attention ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, p. 2227-2231. url : https://doi.org/10.1109/ICASSP.2017.7952552.

[McC09] Robert R McCrae. *The Five-Factor Model of personality traits : consensus and controversy*. Sous la dir. de Philip J. Corr et Gerald Matthews. Cambridge University Press, 2009, p. 148-161. isbn : 978-0-521-86218-9.

[McD13] Kirsty McDougall. « Assessing perceived voice similarity using Multidimensional Scaling for the construction of voice parades ». In : *International Journal of Speech, Language & the Law* 20.2 (2013). URL : http://dx.doi.org/10.1558/ijsll.v20i2.163.

[McD14] Kirsty McDougall. « Listeners' perception of voice similarity in Standard Southern British English versus York English ». In : *23rd Annual Conference of the International Association for Forensic Phonetics and Acoustics (IAFPA)*. 2014.

[MDW09] Gerald Matthews, Ian J Deary et Martha C Whiteman. *Personality Traits*. Cambridge University Press, 2009. URL : https://doi.org/10.1017/CBO9780511812743.

[MLF15] Mitchell McLaren, Yun Lei et Luciana Ferrer. « Advances in deep neural network approaches to speaker recognition ». In : *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2015, p. 4814-4818.

[MLS13] Tomas Mikolov, Quoc V Le et Ilya Sutskever. « Exploiting similarities among languages for machine translation ». In : *arXiv preprint arXiv :1309.4168* (2013).

[MM16] Konstantin Markov et Tomoko Matsui. « Robust Speech Recognition Using Generalized Distillation Framework ». In : *17th Annual Conference of the International Speech Communication Association*. 2016.

[MOG77] J Markel, B Oshika et A Gray. « Long-term feature averaging for speaker recognition ». In : *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25.4 (1977), p. 330-337. URL : https://doi.org/10.1121/1.2003182.

[Moh+12] Gelareh Mohammadi, Antonio Origlia, Maurizio Filippone et Alessandro Vinciarelli. « From speech to personality : Mapping voice quality and intonation into personality differences ». In : *20th ACM International Conference on Multimedia*. ACM. 2012, p. 789-792. URL : https://doi.org/10.1145/2393347.2396313.

[Mor+05] Shigeo Morishima, Akinobu Maejima, Shuhei Wemler, Tamotsu Machida et Masao Takebayashi. « Future cast system ». In : *ACM SIGGRAPH 2005 Sketches*. ACM. 2005, p. 20.

[MS79] R McCrae Cochrane et Jacqueline Sachs. « Phonological learning by children and adults in a laboratory setting ». In : *Language and Speech* 22.2 (1979), p. 145-149. URL : https://doi.org/10.1177/002383097902200204.

[MV] Gelareh Mohammadi et Alessandro Vinciarelli. « Automatic personality perception : Prediction of trait attribution based on prosodic features extended abstract ». In : t. 3. 3, p. 273-284. URL : https://doi.org/10.1109/T-AFFC.2012.5.

[MV12] Gelareh Mohammadi et Alessandro Vinciarelli. « Automatic attribution of personality traits based on prosodic features ». In : (2012).

[MV15] Gelareh Mohammadi et Alessandro Vinciarelli. « Automatic Attribution of Personality Traits Based on Prosodic Features ». In : *ACII 2015 Affective Computing and Intelligent Interaction* 3 (2015), p. 29-32. url : https://doi.org/10.1109/T-AFFC.2012.5.

[MVM10] Gelareh Mohammadi, Alessandro Vinciarelli et Marcello Mortillaro. « The voice of personality : Mapping nonverbal vocal behavior into trait attributions ». In : *Proceedings of the 2nd International Workshop on Social Signal Processing*. ACM. 2010, p. 17-20.

[NEL06] Daniel Neiberg, Kjell Elenius et Kornel Laskowski. « Emotion recognition in spontaneous speech using GMMs ». In : *Ninth International Conference on Spoken Language Processing*. 2006.

[NFD03] Tin Lay Nwe, Say Wei Foo et Liyanage C De Silva. « Speech emotion recognition using hidden Markov models ». In : *Speech communication* 41.4 (2003), p. 603-623. url : https://doi.org/10.1016/S0167-6393(03)00099-2.

[NL01] Clifford Nass et Kwan Min Lee. « Does computer-synthesized speech manifest personality ? Experimental tests of recognition, similarity-attraction, and consistency-attraction ». In : *Journal of experimental psychology : applied* 7.3 (2001), p. 171.

[NMH11] Francis Nolan, Kirsty McDougall et Toby Hudson. « Some Acoustic Correlates of Perceived (Dis) Similarity between Same-accent Voices ». In : *ICPhS*. 2011, p. 1506-1509.

[NO96] Francis Nolan et Tomasina Oh. « Identical twins, different voices ». In : *International Journal of Speech, Language and the Law* 3.1 (1996), p. 39-49.

[Nol+09] Francis Nolan, Kirsty McDougall, Gea De Jong et Toby Hudson. « The DyViS database : style-controlled recordings of 100 homogeneous speakers for forensic phonetic research ». In : *International Journal of Speech, Language & the Law* 16.1 (2009). url : http://dx.doi.org/10.1558/ijsll.v16i1.31.

[Nol+11] Francis Nolan, Peter French, Kirsty McDougall, Louisa Stevens et Toby Hudson. « The role of voice quality 'settings' in perceived voice similarity ». In : *International Association for Forensic Phonetics and Acoustics* (2011).

[OR16] Nicolas Obin et Axel Roebel. « Similarity search of acted voices for automatic voice casting ». In : *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24.9 (2016), p. 1642-1651.

[ORB14a] Nicolas Obin, Axel Roebel et Gregoire Bachman. « On automatic voice casting for expressive speech : Speaker recognition vs. speech classification ». In : 2014, p. 950-954.

[ORB14b] Nicolas Obin, Axel Roebel et Grégoire Bachman. « On automatic voice casting for expressive speech : Speaker recognition vs. speech classification ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, p. 950-954.

[Oya76] Susan Oyama. « A sensitive period for the acquisition of a nonnative phonological system ». In : *Journal of psycholinguistic research* 5.3 (1976), p. 261-283. url : https://doi.org/10.1007/BF01067377.

[PD03] Aniruddh D Patel et Joseph R Daniele. « An empirical comparison of rhythm in language and music ». In : *Cognition* 87.1 (2003), B35-B45. url : https://doi.org/10.1016/S0010-0277(02)00187-7.

[PE07] Simon J.D Prince et James H Elder. « Probabilistic linear discriminant analysis for inferences about identity ». In : *2007 IEEE 11th International Conference on Computer Vision*. IEEE. 2007, p. 1-8. url : https://doi.org/10.1109/ICCV.2007.4409052.

[Pic00] Rosalind W Picard. *Affective computing*. MIT press, 2000.

[Pis+14] Katarzyna Pisanski, Paul J Fraccaro, Cara C Tigue, Jillian JM O'Connor, Susanne Röder, Paul W Andrews, Bernhard Fink, Lisa M DeBruine, Benedict C Jones et David R Feinberg. « Vocal indicators of body size in men and women : a meta-analysis ». In : *Animal Behaviour* 95 (2014), p. 89-99. url : https://doi.org/10.1016/j.anbehav.2014.06.011.

[PIS16] Ryan Price, Ken-ichi Iso et Koichi Shinoda. « Wise teachers train better DNN acoustic models ». In : *EURASIP Journal on Audio, Speech, and Music Processing* 2016 (2016). url : https://doi.org/10.1186/s13636-016-0088-7.

[PKD89] George Papcun, Jody Kreiman et Anthony Davis. « Long-term memory for unfamiliar voices ». In : *The Journal of the Acoustical Society of America* 85.2 (1989), p. 913-925. url : https://doi.org/10.1121/1.397564.

[Pla14] Gaëlle Planchenault. « La commodification des voix au cinéma : un outil de différentiation et de stigmatisation langagière ». In : *Entrelacs. Cinéma et audiovisuel* 11 (2014). url : https://doi.org/10.4000/entrelacs.1566.

[Plu84] Robert Plutchik. *Emotions : A General Psychoevolutionary Theory*. Lawrence Erlbaum Associates, 1984, p. 197-219. url : https://doi.org/10.1016/B978-0-12-558701-3.50007-7.

[Pov+11a] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget et Ondrej Glembek. « The Kaldi speech recognition toolkit ». In : *IEEE Workshop on Automatic Speech Recognition and Understanding*. 2011.

[Pov+11b] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer et Karel Vesely. « The Kaldi Speech Recognition Toolkit ». In : *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, 2011.

[PPC11] Paolo Petta, Catherine Pelachaud et Roddy Cowie. *Emotion-oriented systems : the HUMAINE handbook*. Springer, 2011. isbn : 978-3-642-15184-2.

[Put+11] David A Puts, Julia L Barndt, Lisa LM Welling, Khytam Dawood et Robert P Burriss. « Intrasexual competition among women : Vocal femininity affects perceptions of attractiveness and flirtatiousness ». In : *Personality and Individual Differences* 50.1 (2011), p. 111-115. url : https://doi.org/10.1016/j.paid.2010.09.011.

[PXW10] Aniruddh D Patel, Yi Xu et Bei Wang. « The role of F0 variation in the intelligibility of Mandarin sentences ». In : *Speech Prosody Fifth International Conference*. 2010.

[Rév13] Joana Révis. *La voix et soi : Ce que notre voix dit de nous*. De Boeck Superieur, 2013. isbn : 2353272312.

[Rey97] Douglas A Reynolds. « Comparison of background normalization methods for text-independent speaker verification ». In : *Fifth European Conference on Speech Communication and Technology*. 1997.

[RFN07] Robert E Remez, Jennifer M Fellowes et Dalia S Nagel. « On the perception of similarity among talkers ». In : *The Journal of the Acoustical Society of America* 122.6 (2007), p. 3688-3696. url : https://doi.org/10.1121/1.2799903.

[Ria+18] Rachid Riad, Corentin Dancette, Julien Karadayi, Neil Zeghidour, Thomas Schatz et Emmanuel Dupoux. « Sampling strategies in Siamese Networks for unsupervised speech representation learning ». In : *19th Annual Conference of the International Speech Communication Association*. 2018.

[Ril+09] Albert Rilliard, Takaaki Shochi, Jean-Claude Martin, Donna Erickson et Véronique Aubergé. « Multimodal indices to Japanese and French prosodically expressed social affects ». In : *Language and speech* 52.2-3 (2009), p. 223-243.

[Ril+13] Albert Rilliard, João Antônio de Moraes, Donna Erickson et Takaaki Shochi. « Social affect production and perception across languages and cultures–the role of prosody ». In : *Leitura* 2.52 (2013), p. 15-41.

[Ril+16] Albert Rilliard, Donna Erickson, João A De Moraes et Takaaki Shochi. « On the varying reception of speakers expressivity across gender and cultures, and inference in their personalities ». In : *Sonorities : speech, singing and reciting expressivity* (2016), p. 149-163.

[ROE09] Mika Raento, Antti Oulasvirta et Nathan Eagle. « Smartphones : An emerging tool for social scientists ». In : *Sociological methods & research* 37.3 (2009), p. 426-454. url : https://doi.org/10.1177%2F0049124108330005.

[Ros99] Phil Rose. « Differences and distinguishability in the acoustic characteristics of Hello in voices of similar-sounding speakers ». In : *Australian Review of Applied Linguistics* 22.1 (1999), p. 1-42. url : https://doi.org/10.1075/aral.22.1.01ros.

[Rou+16] Mickael Rouvier, Pierre-Michel Bousquet, Moez Ajili, Waad Ben Kheder, Driss Matrouf et Jean-François Bonastre. « LIA system description for NIST SRE 2016 ». In : *arXiv preprint arXiv :1612.05168* (2016).

[RQD00] Douglas A Reynolds, Thomas F Quatieri et Robert B Dunn. « Speaker verification using adapted Gaussian mixture models ». In : *Digital signal processing* 10.1-3 (2000), p. 19-41. url : https://doi.org/10.1006/dspr.1999.0361.

[RR95] Douglas A Reynolds et Richard C Rose. « Robust text-independent speaker identification using Gaussian mixture speaker models ». In : *IEEE Transactions on Speech and Audio Processing* 3.1 (1995), p. 72-83. url : https://doi.org/10.1109/89.365379.

[Rus80] James A Russell. « A circumplex model of affect ». In : *Journal of personality and social psychology* 39.6 (1980), p. 1161. url : https://dx.doi.org/10.1017%2FS0954579405050340.

[Sáe+06] Nicolás Sáenz-Lechón, Juan I Godino-Llorente, Víctor Osma-Ruiz, Manuel Blanco-Velasco et Fernando Cruz-Roldán. « Automatic assessment of voice quality according to the GRBAS scale ». In : *International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE. 2006, p. 2478-2481. url : https://doi.org/10.1109/IEMBS.2006.260603.

[San+18] Eugenia San Segundo, Paul Foulkes, Peter French, Philip Harrison, Vincent Hughes et Colleen Kavanagh. « Cluster analysis of voice quality ratings : Identifying groups of perceptually similar speakers ». In : 2018, p. 173-176.

[Sar+18] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma et Najim Dehak. « Emotion Identification from Raw Speech Signals Using DNNs ». In : *19th Annual Conference of the International Speech Communication Association*. 2018, p. 3097-3101.

[Sau+10] Disa A Sauter, Frank Eisner, Paul Ekman et Sophie K Scott. « Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations ». In : *Proceedings of the National Academy of Sciences* 107.6 (2010), p. 2408-2412. url : https://doi.org/10.1073/pnas.0908239106.

[SBM16] Achintya Kumar Sarkar, Jean-François Bonastre et Driss Matrouf. « A study on the roles of total variability space and session variability modeling in speaker recognition ». In : *International Journal of Speech Technology* 19.1 (2016), p. 111-120. url : https://doi.org/10.1007/s10772-015-9324-2.

[SBW01] Klaus R Scherer, Rainer Banse et Harald G Wallbott. « Emotion inferences from vocal expression correlate across languages and cultures ». In : *Journal of Cross-cultural psychology* 32.1 (2001), p. 76-92. url : https://doi.org/10.1177%2F0022022101032001009.

[Sch+10] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller et Shrikanth S Narayanan. « The interspeech 2010 paralinguistic challenge ». In : *Eleven Annual Conference of the International Speech Communication Association*. 2010.

[Sch+11a] Nicolas Scheffer, Luciana Ferrer, Martin Graciarena, Sachin Kajarekar, Elizabeth Shriberg et Andreas Stolcke. « The SRI NIST 2010 speaker recognition evaluation system ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2011, p. 5292-5295. url : https://doi.org/10.1109/ICASSP.2011.5947552.

[Sch+11b] Björn Schuller, Anton Batliner, Stefan Steidl et Dino Seppi. « Recognising realistic emotions and affect in speech : State of the art and lessons learnt from the first challenge ». In : *Speech Communication* 53.9-10 (2011), p. 1062-1087. url : https://doi.org/10.1016/j.specom.2011.01.011.

[Sch+11c] Björn Schuller, Stefan Steidl, Anton Batliner, Florian Schiel et Jarek Krajewski. « The insterspeech 2011 speaker state challenge ». In : *Twelve Annual Conference of the International Speech Communication Association*. 2011.

[Sch+12] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet et al. « The interspeech 2012 speaker trait challenge ». In : *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[Sch+13] Stefan Scherer, John Kane, Christer Gobl et Friedhelm Schwenker. « Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification ». In : *Computer Speech & Language* 27.1 (2013), p. 263-287. url : https://doi.org/10.1016/j.csl.2012.06.001.

[Sch+14] Stefan R Schweinberger, Hideki Kawahara, Adrian P Simpson, Verena G Skuk et Romi Zäske. « Speaker perception ». In : *Wiley Interdisciplinary Reviews : Cognitive Science* 5.1 (2014), p. 15-25.

[Sch+91] Klaus R Scherer, Rainer Banse, Harald G Wallbott et Thomas Goldbeck. « Vocal cues in emotion encoding and decoding ». In : *Motivation and emotion* 15.2 (1991), p. 123-148. url : https://doi.org/10.1007/BF00995674.

[Sch03a] Klaus R Scherer. « Vocal communication of emotion : A review of research paradigms ». In : *Speech Communication* 40.1-2 (2003), p. 227-256. url : https://doi.org/10.1016/S0167-6393(02)00084-5.

[Sch03b] Klaus R Scherer. « Vocal communication of emotion : A review of research paradigms ». In : *Speech Communication* 40.1-2 (2003), p. 227-256. url : https://doi.org/10.1016/S0167-6393(02)00084-5.

[Sch09] Klaus R Scherer. « The dynamic architecture of emotion : Evidence for the component process model ». In : *Cognition and emotion* 23.7 (2009), p. 1307-1351. url : https://doi.org/10.1080/02699930902928969.

[Sch74] Klaus R Scherer. « Voice quality analysis of American and German speakers ». In : *Journal of Psycholinguistic Research* 3.3 (1974), p. 281-298.

[Sch78] Klaus R Scherer. « Personality inference from voice quality : The loud voice of extroversion ». In : *European Journal of Social Psychology* 8.4 (1978), p. 467-487. url : https://doi.org/10.1002/ejsp.2420080405.

[Sch84] Klaus R Scherer. « Les émotions : fonctions et composantes ». In : *Cahiers De Psychologie Cognitive/Current Psychology Of Cognition* (1984).

[Sel+10] Aaron Sell, Gregory A Bryant, Leda Cosmides, John Tooby, Daniel Sznycer, Christopher Von Rueden, Andre Krauss et Michael Gurven. « Adaptations in humans for assessing physical strength from the voice ». In : *Proceedings of the Royal Society B : Biological Sciences* 277.1699 (2010), p. 3509-3518. url : https://dx.doi.org/10.1098%2Frspb.2010.0769.

[SG00] Ulrich Schimmack et Alexander Grob. « Dimensional models of core affect : A quantitative comparison by means of structural equation modeling ». In : *European Journal of Personality* 14.4 (2000), p. 325-345. url : https://doi.org/10.1002/1099-0984(200007/08)14:4%3C325::AID-PER380%3E3.0.CO;2-I.

[Sho05] Eric Shouse. « Feeling, emotion, affect ». In : *M/c journal* 8.6 (2005), p. 26.

[SHS97] Stefan R Schweinberger, Anja Herholz et Werner Sommer. « Recognizing famous voices : Influence of stimulus duration and different types of retrieval cues ». In : *Journal of Speech, Language, and Hearing Research* 40.2 (1997), p. 453-463. url : https://doi.org/10.1044/jslhr.4002.453.

[SM17] Eugenia San Segundo et Jose A Mompean. « A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity ». In : *Journal of Voice* 31.5 (2017), 644-e11. url : https://doi.org/10.1016/j.jvoice.2017.01.005.

[Sny+16] David Snyder, Pegah Ghahremani, Daniel Povey, Daniel Garcia-Romero, Yishay Carmiel et Sanjeev Khudanpur. « Deep neural network-based speaker embeddings for end-to-end speaker verification ». In : *2016 IEEE Spoken Language Technology Workshop (SLT)*. IEEE. 2016, p. 165-170.

[Sny+17] David Snyder, Daniel Garcia-Romero, Daniel Povey et Sanjeev Khudanpur. « Deep Neural Network Embeddings for Text-Independent Speaker Verification ». In : *18th Annual Conference of the International Speech Communication Association*. 2017.

[Sny+18] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey et Sanjeev Khudanpur. « X-vectors : Robust dnn embeddings for speaker recognition ». In : *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, p. 5329-5333. url : https://doi.org/10.1109/SLT.2016.7846260.

[Soo+87] Frank K Soong, Aaron E Rosenberg, Bling-Hwang Juang et Lawrence R Rabiner. « Report : A vector quantization approach to speaker recognition ». In : *AT&T technical journal* 66.2 (1987), p. 14-26. url : https://doi.org/10.1002/j.1538-7305.1987.tb00198.x.

[Sri+14] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever et Ruslan Salakhutdinov. « Dropout : A Simple Way to Prevent Neural Networks from Overfitting ». In : *Journal of Machine Learning Research* 15 (2014), p. 1929-1958. url : http://jmlr.org/papers/v15/srivastava14a.html.

[SS08] Elizabeth Shriberg et Andreas Stolcke. « The case for automatic higher-level features in forensic speaker recognition ». In : *Ninth Annual Conference of the International Speech Communication Association*. 2008.

[ST14] Felix Schröter et Jan-Noël Thon. « Video game characters. Theory and analysis ». In : *Diegesis* 3.1 (2014).

[Su+17] Ming-Hsiang Su, Chung-Hsien Wu, Kun-Yi Huang, Qian-Bei Hong et Hsin-Min Wang. « Personality trait perception from speech signals using multiresolution analysis and convolutional neural networks ». In : *9th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE. 2017, p. 1532-1536. URL : https://doi.org/10.1109/APSIPA.2017.8282287.

[SU03] A Stanton et L Unkrich. *Le Monde de Némo*. Sous la dir. de G Walters. Pixar Animation Studios, 2003.

[SW13] Thapanee Seehapoch et Sartra Wongthanavasu. « Speech emotion recognition using support vector machine ». In : *5th International Conference on Knowledge and Smart Technology (KST)*. IEEE, 2013, p. 86-91. URL : https://doi.org/10.1109/KST.2013.6512793.

[Tah12] Marie Tahon. « Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot ». Thèse de doct. Paris 11, 2012.

[Tes03] Mihoko Teshigawara. « Voices in Japanese Animation : How People Perceive the Voices of Good Guys and Bad Guys ». In : *Working Papers of the Linguistics Circle* 17 (2003), p. 149-158.

[Tes04a] Mihoko Teshigawara. « Random splicing : A method of investigating the effects of voice quality on impression formation ». In : *Speech Prosody 2004, International Conference*. 2004.

[Tes04b] Bernard Teston. *L'évaluation instrumentale des dysphonies. Etat actuel et perspectives*. 2004.

[Tim+12] Polzehl Tim, Schoenenberg Katrin, Moller Sebastian, Metze Florian, Gelareh Mohammadi et Alessandro Vinciarelli. « On speaker-independent personality perception and prediction from speech ». In : *13th Annual Conference of the International Speech Communication Association*. 2012.

[TN11] E Toledano et O Nakache. *Intouchable*. Sous la dir. de N Duval Adassovsky, Y Zenou et L Zeitoun. TF1 Film Production, 2011.

[Tri+16] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Bjorn Schuller et Stefanos Zafeiriou. « Adieu features ? End-to-end speech emotion recognition using a deep convolutional recurrent network ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. T. 2016-May. 2016, p. 5200-5204. URL : https://doi.org/10.1109/ICASSP.2016.7472669.

[TVD12] Khiet P Truong, David A Van Leeuwen et Franciska M.G De Jong. « Speech-based recognition of self-reported and observed emotion in a dimensional space ». In : *Speech Communication* 54.9 (2012), p. 1049-1063. URL : https://doi.org/10.1016/j.specom.2012.04.006.

[TZS18] Panagiotis Tzirakis, Jiehao Zhang et Bjorn W Schuller. « End-to-end speech emotion recognition using deep neural networks ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, p. 5089-5093. URL : https://doi.org/10.1109/JSTSP.2017.2764438.

[Var+14] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno et Javier Gonzalez-Dominguez. « Deep neural networks for small footprint text-dependent speaker verification ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2014, p. 4052-4056. URL : https://doi.org/10.1109/ICASSP.2014.6854363.

[VI15] Vladimir Vapnik et Rauf Izmailov. « Learning Using Privileged Information : Similarity Control and Knowledge Transfer ». In : *Journal of Machine Learning Research* 16.61 (2015), p. 2023-2049. URL : http://jmlr.org/papers/v16/vapnik15b.html.

[VKE85] Diana Van Lancker, Jody Kreiman et Karen Emmorey. « Familiar voice recognition : Patterns and parameters : I. Recognition of backward voices ». In : *Journal of phonetics* (1985). URL : https://doi.org/10.1016/S0095-4470(19)30723-5.

[VM14] Alessandro Vinciarelli et Gelareh Mohammadi. « A survey of personality computing ». In : *IEEE Transactions on Affective Computing* 5.3 (2014), p. 273-291. URL : https://doi.org/10.1109/TAFFC.2014.2330816.

[Wan+14] Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen et Ying Wu. « Learning fine-grained image similarity with deep ranking ». In : *IEEE Conference on Computer Vision and Pattern Recognition*. 2014, p. 1386-1393. URL : https://doi.org/10.1109/CVPR.2014.180.

[Wat+17] Shinji Watanabe, Takaaki Hori, Jonathan Le Roux et John R Hershey. « Student-teacher network learning with enhanced features ». In : *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017. URL : https://doi.org/10.1109/ICASSP.2017.7953163.

[Whe+14] John R Wheatley, Coren A Apicella, Robert P Burriss, Rodrigo A Cárdenas, Drew H Bailey, Lisa LM Welling et David A Puts. « Women's faces and voices are cues to reproductive potential in industrial and forager societies ». In : *Evolution and Human Behavior* 35.4 (2014), p. 264-271. URL : https://doi.org/10.1016/j.evolhumbehav.2014.02.006.

[Wol72] Jared J Wolf. « Efficient acoustic parameters for speaker recognition ». In : *The Journal of the Acoustical Society of America* 51.6B (1972), p. 2044-2056. URL : https://doi.org/10.1121/1.1913065.

[WS02] C Wedge et C Saldanha. *L'Âge de Glace*. Sous la dir. de L Forte. 20th Century Fox, 2002.

[XL12] Rui Xia et Yang Liu. « Using i-vector space model for emotion recognition ». In : *Thirteenth Annual Conference of the International Speech Communication Association*. 2012.

[YPS12] Sibel Yaman, Jason Pelecanos et Ruhi Sarikaya. « Bottleneck features for speaker recognition ». In : *Odyssey Proceedings*. 2012.

[YWK13] Takanori Yamada, Longbiao Wang et Atsuhiko Kai. « Improvement of distant-talking speaker identification using bottleneck features of DNN ». In : *14th Annual Conference of the International Speech Communication Association*. 2013, p. 3661-3664.

[ZBE04] Elisabeth Zetterholm, Mats Blomberg et Daniel Elenius. « A comparison between human perception and a speaker verification system score of a voice imitation ». In : *Tenth Australian International Conference on Speech Science & Technology*. 2004, p. 393-397.

[ZD89a] Miron Zuckerman et Robert E Driver. « What sounds beautiful is good : The vocal attractiveness stereotype ». In : *Journal of Nonverbal Behavior* 13.2 (1989), p. 67-82. url : https://doi.org/10.1007/BF00990791.

[ZD89b] Miron Zuckerman et Robert E. Driver. « What sounds beautiful is good : The vocal attractiveness stereotype ». In : *Journal of Nonverbal Behavior* 13.2 (1989), p. 67-82.

[Zeg+16a] Neil Zeghidour, Gabriel Synnaeve, Nicolas Usunier et Emmanuel Dupoux. « Joint learning of speaker and phonetic similarities with siamese networks ». In : *17th Annual Conference of the International Speech Communication Association*. 2016, p. 1295-1299.

[Zeg+16b] Neil Zeghidour, Gabriel Synnaeve, Maarten Versteegh et Emmanuel Dupoux. « A deep scattering spectrum—Deep Siamese network pipeline for unsupervised acoustic modeling ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, p. 4965-4969. url : https://doi.org/10.1109/ICASSP.2016.7472622.

[Zei12] Matthew D Zeiler. « ADADELTA : an adaptive learning rate method ». In : *arXiv preprint arXiv :1212.5701* (2012).

[ZK93] Miron Zuckerman et Miyake Kunitate. « The Attractive Voice : What Makes It So? » In : *Health (San Francisco)* 17.2 (1993). url : https://doi.org/10.1007/BF01001960.

[ZT08] Cuiling Zhang et Tiejun Tan. « Voice disguise and automatic speaker recognition ». In : *Forensic science international* 175.2 (2008), p. 118-122. url : https://doi.org/10.1016/j.forsciint.2007.05.019.

[ZYH17] Chunlei Zhang, Chengzhu Yu et John HL Hansen. « An investigation of deep-learning frameworks for speaker verification antispoofing ». In : *IEEE Journal of Selected Topics in Signal Processing* 11.4 (2017), p. 684-694. URL : https://doi.org/10.1109/JSTSP.2016.2647199.

Publications personnelles

[Gre+17] Adrien Gresse, Mickael Rouvier, Richard Dufour, Vincent Labatut et Jean-François Bonastre. « Acoustic Pairing of Original and Dubbed Voices in the Context of Video Game Localization ». In : *18th Annual Conference of the International Speech Communication Association*. 2017. URL : https://dx.doi.org/10.21437/Interspeech.2017-1311.

[Gre+19] Adrien Gresse, Mathias Quillot, Richard Dufour, Vincent Labatut et Jean-François Bonastre. « Similarity Metric Based on Siamese Neural Netorks for Voice Casting ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019. URL : https://dx.doi.org/10.1109/ICASSP.2019.8683178.

AVIGNON UNIVERSITÉ

L'Art de la Voix

Caractériser l'information vocale dans un choix artistique

par
Adrien Gresse

LABORATOIRE INFORMATIQUE D'AVIGNON