Un premier pas vers la caractérisation de l'information véhiculée par les voix actées : dualité des informations personnage et locuteur

Mathias Quillot

ACADÉMIE D'AIX-MARSEILLE
AVIGNON UNIVERSITÉ

THÈSE

Présentée à Avignon Université en vue d'obtenir le diplôme de doctorat

SPÉCIALITÉ : INFORMATIQUE

École Doctorale 536 « Agrosciences & Sciences »
Laboratoire Informatique d'Avignon

Un premier pas vers la caractérisation de l'information véhiculée par les voix actées

Dualité des informations personnage et locuteur

par
Mathias Quillot

Soutenue publiquement le 27 septembre 2022 devant un jury composé de :

M^{me}	Chloé Clavel	Professeure à l'Institut Polytechnique de Paris, LTCI	Rapporteur
M.	Julien Pinquier	Maître de conférences (HDR) à l'Université de Toulouse, IRIT	Rapporteur
M^{me}	Corinne Fredouille	Professeure à Avignon Université, LIA	Présidente de jury
M^{me}	Marie Tahon	Maîtresse de conférences à Le Mans Université, LIUM	Examinateur
M.	Nicolas Obin	Maître de conférences à Sorbonne Université, IRCAM	Examinateur
M^{me}	Magalie Ochs	Maîtresse de conférences (HDR) à Aix-Marseille Université, LIS,ILCB	Examinatrice
M.	Richard Dufour	Professeur à Nantes Université, LSNN	Co-Encadrant
M.	Jean-François Bonastre	Professeur à Avignon Université, LIA	Directeur de thèse

Remerciements

Le doctorat est une épreuve marathonienne difficile que le ou la doctorante vit en solitaire. Pourtant, de nombreuses personnes l'accompagnent : notre encadrement, notre famille, nos amis et nos collègues. Il est donc normal et naturel pour moi de remercier toutes ces personnes qui m'ont soutenu et qui ont contribué à la réalisation de cette thèse de manière plus ou moins directe.

Tout d'abord, je remercie mon directeur de thèse Jean-François Bonastre d'avoir cru en moi et de ne pas m'avoir abandonné lorsque je recherchais un financement de thèse alors que de nombreux sujets m'étaient passés « sous le nez » pour différentes raisons. Je suis heureux que Jean-François m'ait offert cette opportunité. Il m'a notamment appris à formaliser des expériences scientifiques et à bien séparer une problématique technique d'une problématique de recherche. Cette rigueur intellectuelle qu'il m'a apporté, j'espère la mettre à contribution dans mes prochains travaux de recherche.

Mon co-encadrant Richard Dufour m'a aussi beaucoup apporté. Il a d'abord été un soutien moral indéniable pendant les périodes les plus compliquées. Il a toujours pris le temps de discuter avec moi et il savait me confronter avec la vérité pour me débloquer de certaines situations. Plus encore, il m'a aidé de nombreuses fois à préparer mes prises de parole et à écrire mes articles scientifiques. Pour tout cela, je ne le remercierai jamais assez.

Je souhaite bien entendu remercier tous les participants du projet ANR « The Voice » et tout particulièrement Nicolas Obin qui a été très présent au début de ma thèse. J'ai beaucoup apprécié les discussions que nous avons eues et les quelques verres que nous avons dégustés ensemble. Sa culture et son amour de la musique et de la littérature le rendent extrêmement intéressant. Grâce à lui, j'ai aussi pu rencontrer Rafael Ferro qui débutait comme moi sa thèse en 2018. Nous n'avons pas pu énormément échanger mais ces discussions que nous avons eues resteront gravées dans ma mémoire.

Pour moi, l'informatique est un outil transversal qui doit être mis à dispo-

sition de toutes les disciplines pour espérer un jour rendre le monde meilleur. Pour cette raison, la pluridisciplinarité me semble être un point essentiel à mettre en valeur dans les travaux de recherche appliquée. J'ai eu la chance durant mon master de travailler sur un projet ANR pluridisciplinaire nommé « GAFES ». Avec une équipe du LIA composée de Aurélien Pinet, Richard Dufour, Vincent Labatut, Georges Linares et Eric Sanjuan, nous travaillions avec des sociologues du Laboratoire Culture et Communication d'Avignon. Les échanges que nous avions eu étaient autant stimulants intellectuellement qu'humainement et m'avaient, à l'époque, conforté dans mon choix de devenir enseignant-chercheur et de travailler sur des thématiques pluridisciplinaires. Je souhaite donc remercier toute l'équipe du LIA/CERI qui m'a formé durant le master mais aussi les sociologues cités ci-après : Alexandre Delorme, Lauriane Guillou, Damien Malinas, Raphaël Roth, Emmanuel Éthis et Quentin Amalou.

Pour finir, je souhaite remercier ma famille et mes amis qui ont été présents tout le long de mon doctorat et qui ont su parfois accepter que je me mette en retrait pour me plonger dans mes études. En écrivant ces lignes, je pense bien sûr à ma mère, à mon père et à mon frère. Je pense aussi à mes deux meilleurs amis et au groupe d'amis que nous avons en commun. Je pense aussi aux collègues de travail que j'ai rencontrés au laboratoire et qui sont finalement devenus ce que je pourrais nommer « une famille ». Je remercie toutes les personnes qui ont participé et suivi de près comme de loin mon doctorat durant ces (presque) 5 longues années.

Résumé

Avant d'être distribuée dans différents pays, une œuvre telle qu'un jeu vidéo ou un film doit être changée et adaptée. Le sous-titrage et le doublage sont deux options pour adapter une œuvre. Si le sous-titrage est moins coûteux à réaliser, le doublage convient mieux à certains spectateurs qui préfèrent écouter les paroles, généralement dans leur langue natale, plutôt que de lire des sous-titres tout en écoutant les paroles d'une autre langue. Pour réaliser le doublage d'une œuvre, la première étape consiste à sélectionner les comédiens, parmi un ensemble de candidats, dont les voix vont remplacer celles d'origine. Cette sélection est le Casting Vocal. Elle est réalisée par le Directeur Artistique (DA), parfois nommé le directeur de casting.

Avec l'apparition de nouvelles plateformes de streaming telles que Disney+ et Amazon Prime et l'accroissement fulgurant de l'industrie du jeu vidéo, le nombre d'œuvres à distribuer à l'international augmente fortement. En réponse à cette demande, de plus en plus de comédiens sont disponibles dans le marché des voix. Le DA peut passer à côté de talents qui lui sont inconnus puisqu'il lui est impossible d'auditionner tous les candidats. Des outils de recommandation et de recherche de comédiens, basés sur le traitement automatique de la parole, aideraient les DA à trouver de nouveaux talents qui enrichiraient la diversité vocale des œuvres pour une meilleure immersion du public.

S'intéresser à la recommandation de comédiens implique d'étudier le concept de « voix actée ». Dans les œuvres multimédia, la voix actée est exprimée par des acteurs professionnels ; son but est de produire chez le spectateur l'effet désiré en donnant un comportement particulier au personnage. Son étude implique une double complexité en terme de production et de perception qui explique pourquoi la voix actée est si peu présente dans la littérature du traitement de la parole.

Des travaux précédents ont abordé le problème du casting vocal en se focalisant sur les voix de personnages de jeux vidéos. Dans ces travaux, la similarité de voix est centrale. Des systèmes exploitent les associations entre comédien

d'origine et comédien doubleur pour modéliser une partie du processus de décision de l'opérateur (le DA). La tâche consiste à prédire si les deux voix fournies au systèmes jouent le même personnage sous la forme d'une mesure de similarité personnage.

Dans ce manuscrit, nous nous intéressons à l'information personnage : l'ensemble des signes acoustiques dans un enregistrement vocal qui caractérisent le personnage joué. Bien que de précédents travaux aient montré l'existence d'une telle information dans la voix actée, la nature de cette information reste encore en grande partie inconnue. Nous cherchons dans ce manuscrit à éclairer des zones d'ombres en étudiant deux questions :

— Quel lien entretient l'information personnage avec son comédien ?
— Quels sont les marqueurs vocaux qui donnent forme au personnage ?

Dans un premier temps, nous construisons un protocole pour évaluer la présence d'information personnage dite « indépendante du locuteur ». Dans nos expériences, nous montrons que cette information existe mais qu'elle est très peu exprimée dans nos données.

Dans un second temps, nous montrons dans une expérience que l'information locuteur est utile à la construction de systèmes dédiés à la caractérisation du personnage joué.

Enfin, nous proposons une expérience qui consiste, depuis des étiquettes personnage et des enregistrements, à extraire des marqueurs vocaux dédiés à la caractérisation du personnage joué.

Abstract

Before being distributed in different countries, a media content such as a video game or a movie must be changed and adapted. Subtitling and dubbing are two options for adapting a media content. While subtitling is less expensive to produce, dubbing is often more suitable to some viewers who prefer to listen to the speech, usually in their native language, rather than reading subtitles while listening to the speech in another language. The first step in dubbing a media content is to select the voice actors, from a pool of candidates, whose voices will replace the original ones. This selection is known as voice casting. It is carried out by the Artistic Directors (AD), sometimes referred to as casting directors.

With the emergence of new streaming platforms such as Disney+ and Amazon Prime and the meteoric rise of the video game industry, the amount of media content to be distributed internationally is increasing significantly. In response to this demand, more and more voice actors are available in the voice market and the AD may miss out on unknown talent since it is impossible to audition all candidates. Voice actor recommandation and search tools, based on speech processing can help AD discover new talent with the potential to enrich the vocal diversity in media content which in turn can result in better audience immersion.

Work focused on the recommendation of voice actors implies the study of « acted voice » concept. In multimedia works, the acted voice is produced by professional actors ; its purpose is to produce in the spectator the desired effect by giving a particular behavior to the character. Its study implies a double complexity in terms of production and perception that explains why the acted voice is so little present in the literature of speech processing.

Previous work has addressed the problem of voice casting by focusing on the voices of video game characters. In this work, voice similarity is central. Systems exploit the associations between the original actor and the dubbing one to model part of the operator's decision process (the AD). The task is to

predict whether the two voices provided to the system play the same character in the form of a character similarity measure.

In this manuscript, we are interested in character information : the set of acoustic signs in a voice recording that characterize the played character. Although previous work has shown the existence of such information in the acted voice, the nature of this information is still largely unknown. In this thesis, we seek to shed light on some of these grey areas by investigating two questions :

— How does the character information relate to its actor ?
— What are the vocal markers that shape a character ?

First, we build a protocol to evaluate the presence of speaker-independent character information. In our experiments, we show that this information exists but has limited expressiveness in our data.

Then, we show that speaker information is useful to build systems dedicated to the characterization of the character played in a recording.

Finally, we propose an experiment which consists in extracting, from character labels and recordings, vocal markers dedicated to the characterization of the character played in a recording.

Sommaire

Chapitre 1

Introduction

L'intelligence, qu'elle soit humaine, animale ou végétale, ne peut émerger sans communication avec l'environnement. L'humain, par exemple, acquiert des informations du monde qui l'entoure et décide d'y réagir suite à un processus cognitif. Fruit de longues années d'évolution, l'humain a développé différentes modalités d'acquisition de l'information qui lui permettent de peindre continuellement une toile de son environnement. Ces modalités sont communément appelées des "sens". Outre les cinq sens de base (le goût, l'odorat, le toucher, l'ouïe et la vue), la proprioception, l'équilibrioception, la thermoception et la nociception forment le paquet sensoriel de l'être humain. Les informations ainsi acquises ne concernent pas seulement notre environnement mais aussi l'état de notre corps dans cet environnement. Nous prenons ainsi conscience de l'impact que nous avons sur le monde et vice-versa. À titre d'illustration, lorsque nous bougeons un bras, nous sommes conscients de sa position dans l'espace grâce à la proprioception et, si nous rencontrons un obstacle, nous le sentons grâce au toucher.

Tandis que l'ouïe est le sens dédié à la captation des sons de notre environnement, la voix, elle, est l'« ensemble des sons produits par les vibrations périodiques des cordes vocales » selon le Larousse. Contrairement à certaines espèces animales qui échangent des cris symboliques, l'homme est le seul à user de la parole – sa faculté à s'exprimer par un langage articulé – pour communiquer ([Boe+17]). Par son entremise, nous pouvons donner forme à nos idées, à nos demandes mais nous pouvons également communiquer des sentiments et des impressions. Bien que d'autres modalités d'expression complexes existent, telles que l'écriture ou le langage des signes, l'expression orale montre une efficacité toute particulière qui en fait aujourd'hui l'un des outils principaux utilisés pour la communication inter-humain quotidienne. Sa présence est telle

dans la société que les troubles comme la surdité, le bégaiement ou le mutisme sélectif sont un handicap non négligeable pour celles et ceux qui les subissent.

Les productions multimédia ne dérogent pas à la règle et la voix y joue généralement un rôle central. La voix a fait son apparition au cinéma au début du XXe siècle avec le tournage des premiers films non muets. Ce changement de paradigme cinématographique a bousculé toute l'industrie. Certains acteurs et actrices ont ainsi perdu leur notoriété parce que leur voix ne convenait pas à leur physique et engendrait une sensation de discordance chez le spectateur montrant ainsi que la voix occupe une place centrale au cinéma et est généralement indissociable d'une œuvre cinématographique. La voix engendre cependant de nombreuses contraintes lorsque les producteurs souhaitent distribuer leurs œuvres dans des pays étrangers.

Parmi celles-ci, la localisation d'une œuvre multimédia consiste à changer et adapter l'œuvre d'origine de manière à ce que les distributeurs internationaux puissent la distribuer dans un pays cible. Plusieurs options sont disponibles pour cela. Le sous-titrage est l'option la plus simple et la moins chère, mais n'est pas toujours la plus appréciée par les spectateurs. Certains spectateurs préfèrent écouter des paroles, généralement dans leur langue natale, plutôt que de lire des sous-titres tout en écoutant les paroles dans une autre langue. Le doublage est plus coûteux et demande plus de temps à mettre en place, mais il conserve mieux l'immersion du public en remplaçant la voix d'origine par la voix d'un comédien de doublage. Il nécessite un pré-processus de sélection de voix où un opérateur humain sélectionne un candidat parmi plusieurs pour jouer le personnage cible. Cette sélection est le *casting vocal*. Plus généralement, le casting vocal permet d'assurer la concordance des voix avec le produit multimédia. Il est aussi employé dans les documentaires, les livres audios ou des tutoriaux en ligne où les comédiens sélectionnés tâchent de "poser leurs voix". Cependant, nous nous intéressons particulièrement dans ce manuscrit à son usage pour le doublage de films et de jeux vidéos.

Avec l'émergence des plateformes de streaming telles que Netflix, Disney + ou Amazon Prime et l'accroissement fulgurant de l'industrie du jeu vidéo, le nombre d'œuvres à distribuer à l'international augmente fortement. En réponse à cette demande, de plus en plus de comédiens sont disponibles dans le marché des voix. Le casting vocal est cependant un processus artisanal qui requiert beaucoup de précision et qui est effectué par des opérateurs humains. Tel quel, il ne peut être appliqué sur un très grand nombre de comédiens (plusieurs centaines). Les opérateurs se fient souvent à leur mémoire, qui est limitée à un panel de comédiens avec lesquels ils ont travaillé et auxquels ils accordent du

crédit. Leurs voix sont alors surreprésentées et deviennent un élément qui peut potentiellement biaiser la perception du personnage par le public. Des outils de recommandation et de recherche d'acteurs, basés sur le traitement automatique de la parole, aideraient les opérateurs à trouver de nouveaux comédiens qui enrichiraient la diversité vocale des œuvres pour une meilleure immersion du public.

S'intéresser à la recommandation de comédiens implique d'étudier le concept de "voix actées". Dans les œuvres multimédia, la voix actée est exprimée par des acteurs professionnels, son but est de produire chez le spectateur l'effet désiré en donnant un comportement particulier au personnage. Cette rencontre entre le personnage et son comédien donne naissance à un rôle, terme défini par le Larousse comme étant le « personnage représenté par l'acteur, le danseur ». La voix actée est souvent sur-jouée dans le but de rendre l'effet expressif désiré mieux perceptible. Les aspects du jeu de l'acteur sont sujets à la perception de l'auditeur. L'interprétation par ce dernier dépend non seulement de lui-même et de son histoire, mais aussi du contexte de son écoute. Cette double complexité en terme de production et perception, explique pourquoi la voix actée est si peu présente dans la littérature du traitement de la parole.

Des précédents travaux ont abordé le problème du casting vocal en se focalisant sur les voix de personnages de jeux vidéos. Les auteurs de [ORB14 ; OR16 ; Gre+17 ; GDL18 ; Gre+19 ; Gre+20] proposent d'utiliser des enregistrements de jeux vidéos – en versions originale et doublée – où les comédiens doubleurs et doublés sont déjà connus. La notion de similarité de voix est alors centrale. Communément employée dans le domaine de la vérification du locuteur, elle est ici adaptée au casting vocal. Cette tâche consiste à prédire si les deux voix fournies au système jouent le même personnage sous la forme d'une mesure de similarité personnage.

Nous posons dans ce manuscrit « l'information personnage » comme l'ensemble des signes acoustiques dans un enregistrement vocal qui caractérisent le personnage joué. ([Gre+17]) et ([Gre+19]) ont démontré l'existence de cette information dans la voix, du moins, dans le contexte du jeu vidéo Mass Effect. Cependant, la nature de cette information reste encore en grande partie inconnue. Nous cherchons dans ce manuscrit à éclairer des zones d'ombres en étudiant deux questions :

— Quel lien entretient l'information personnage avec son comédien ? Il s'agit alors d'étudier l'impact qu'a le comédien sur l'information personnage.

— Quels sont les marqueurs vocaux qui donnent forme au personnage ? Il s'agit ici d'étudier les sources de variation qui véhiculent les spécificités du personnage.

Dans un premier temps, nous construisons un protocole pour évaluer la présence d'information personnage dite « indépendante du locuteur ». Dans nos expériences, nous montrons que cette information existe mais qu'elle est très peu exprimée dans nos données.

Dans un second temps, nous étudions l'influence que peut avoir l'information locuteur sur la construction de systèmes dédiés à la caractérisation du personnage joué.

Dans un troisième et dernier temps, nous proposons une expérience qui consiste, depuis des étiquettes personnage et des enregistrements, à extraire des marqueurs vocaux dédiés à la caractérisation du personnage joué.

Pour toutes les expériences, les données proviennent du jeu vidéo Mass Effect.

Ce manuscrit s'organise en trois grandes parties. La première introduit le contexte applicatif du doublage. La seconde introduit l'état de l'art en reconnaissance du locuteur et en automatisation du casting vocal. Le lecteur y trouvera notamment les informations nécessaires sur les systèmes de similarité de voix et d'extraction de représentation personnage. La troisième partie expose les deux contributions précédemment présentées qui ont été respectivement publiées dans [QDB21] et [Qui+21]. Tous les protocoles expérimentaux ainsi que les résultats d'expériences et leurs discussions y sont détaillés. Nous y présentons aussi une troisième contribution non liée à la partie de l'identité du locuteur, mais complémentaire aux deux précédentes. Celle-ci traite de l'extraction automatique d'étiquettes personnage sans vérité terrain. Pour finir, nous concluons et discutons des perspectives de recherche.

Première partie

Du casting vocal à la voix actée

Chapitre 2

Le casting vocal, un processus exécuté par des opérateurs humains

Sommaire

2.1 Le doublage vocal

Doubler une œuvre consiste à remplacer les voix originales par de nouvelles voix s'exprimant dans une langue cible. Le doublage sert à la localisation d'une œuvre lorsque celle-ci doit être diffusée dans différents pays.

Le terme « doublage » peut également prendre d'autres sens. Par exemple, il peut désigner la postsynchronisation lorsque les comédiens enregistrent en studio les dialogues qui n'ont pas été enregistrés en direct pour des raisons

techniques ou artistiques. Il peut également désigner, par convention, les prestations vocales enregistrées en amont d'un tournage, notamment dans le domaine de l'animation ou du jeu vidéo.

Ci-dessous, nous décrivons le doublage vocal.

2.1.1 Le processus de doublage

Pour doubler une œuvre, toute une chaine de production doit être suivie par différents opérateurs humains tels que le détecteur ou l'adaptateur. Nous présentons chacun de ses maillons ci-dessous.

La détection

Le détecteur inscrit sur une bande les indications dont l'auteur aura besoin. Parmi ces indications figurent le texte (dans la langue originale du programme à adapter), les respirations, rires et réactions des comédiens, les signes de détection permettant le « lipsync » (synchronisme labial) du texte de la version doublée.

L'adaptation

L'adaptateur est chargé de traduire le texte original sans en déformer le sens. Ce travail doit être réalisé en respectant le rythme propre de chaque langue. La langue anglaise est, par exemple, beaucoup plus synthétique que la langue française. L'adapteur doit alors respecter la longueur des dialogues d'origine pour éviter que les segments de prise de parole ne se chevauchent.

La vérification et la calligraphie

Lorsque l'adaptation est terminée, une vérification a lieu. Elle consiste à s'assurer que l'adaptation convient au Directeur Artistique (DA) – aussi appelé directeur de plateau – qui choisira les comédiens et les dirigera sur le plateau d'enregistrement. Le DA vérifie que le texte est synchrone et si l'adaptation est faite dans un français naturel.

Une fois que l'adaptation a reçu l'aval du DA, le calligraphe réalise deux tâches. D'abord, il recopie le texte de l'auteur sur une bande rythmo. Puis, il effectue le travail de « frappe ». Il s'agit de retranscrire la version doublée qu'il vient de recopier sur la bande rythmo, une bande horizontale défilant au bas de l'écran et comportant le texte que doivent prononcer les comédiens ainsi que les sons qu'ils doivent reproduire. La frappe est soumise à l'auteur qui peut ainsi vérifier que le calligraphe a correctement retranscrit le texte.

Enregistrement et mixage

L'ingénieur du son réalise les prises de son, il enregistre les comédiens de la version doublée. Il veille à ce que tous les comédiens parlent au même niveau sonore, sauf dans les cas où le personnage chuchote ou crie. Il s'assure également que le texte est intelligible et qu'aucun bruit de bouche ne vient parasiter le texte.

Une fois l'enregistrement terminé, l'ingénieur du son peut être amené à déplacer les répliques si ces dernières ont été prononcées en retard ou en avance par rapport à la Version originale. Puis, l'ingénieur du son réalise le mixage. Il s'agit pour lui de tout mettre en œuvre pour donner l'impression au spectateur que le texte dans la langue cible (e.g. français) est bien exprimé par le personnage.

Pour intégrer les voix de la version doublée à l'œuvre originale, une Version Internationale (VI) a été élaborée parallèlement au mixage de la VO. La VI ne comporte que les ambiances, musiques, effets sonores et bruitages sans les voix.

2.1.2 Deux stratégies de doublage

Deux stratégies de doublage sont identifiables et sont présentées ci-après. La première stratégie fait référence aux comédiens qui modifient très peu leur voix neutre lorsqu'ils jouent un personnage. La seconde stratégie, contrairement à la première, consiste pour le comédien à changer la personnalité de sa voix, rendant son identification difficile. Les personnages enfants joués par des comédiens adultes illustrent parfaitement cette seconde stratégie. L'identité vocale du comédien n'est pas entièrement effacée mais sa voix est modulée de manière à donner l'illusion au public qu'elle est produite par un enfant.

Les deux comédiens de doublage présentés ci-dessous sont de très bons exemples pour illustrer les deux stratégies précédemment discutées. Le premier, Adrien Antoine, est un doubleur très réputé pour avoir interprété différents personnages comme Superman ou Batman, tous deux héros principaux de leurs films respectifs. Par son historique, il semble évident de proposer ce comédien pour jouer Thor (voir Figure 2.1a) qu'il a en effet interprété plusieurs fois. Remarquons tout de même qu'il a aussi doublé le personnage "La Binocle", un personnage enfant de la série des RazMoket. Dans les rôles de super héros, nous pouvons dire que sa voix est « surjouée » mais qu'il suit la première stratégie de doublage où elle reste proche de sa voix propre. A contrario, pour jouer "La Binocle", il suit la seconde stratégie et déforme suffisamment sa voix

(a) Adrien Antoine (b) Christophe Le Moine

FIGURE 2.1 – Exemples de doublage réalisés par les deux comédiens, Adrien Antoine et Christophe Le Moine.

pour donner l'illusion de l'enfance.

Un deuxième exemple est celui de Christophe Le Moine. Ce doubleur-ci est habitué à jouer des personnages très différents et plus généralement des personnages secondaires. Il a doublé Cartman de *South Park*, Sam Gamji du *Seigneur des anneaux*, des personnages secondaires de jeux vidéos, T-Bone (un épicier) de *The Amazing Spider-Man*, une drag-queen dans *A Star is Born* etc. Ce comédien suit généralement la seconde stratégie de doublage. Le reconnaitre dans ses interprétations nécessite parfois une oreille aguerrie tant il joue avec sa voix pour l'adapter à ses personnages.

Pour ces deux acteurs, les critères de décision des DA sont différents. Adrien Antoine est un comédien dont la voix neutre – sa voix de tous les jours – exprime un caractère proche de ceux des super héros qu'il joue. Cependant, il est aussi capable de moduler sa voix et de jouer des personnages enfants de dessins animés. À l'opposé, Christophe Le Moine est un comédien qui a joué un nombre impressionnant de personnages secondaires et dont la diversité et la variabilité vocale semblent très élevés. Les DA vont donc plutôt associer des personnages guerriers, héroïques et séducteurs à Adrien Antoine alors qu'ils associeront plutôt des personnages secondaires avec des caractères atypiques à Christophe Le Moine. Pour comprendre pourquoi les DA associent ces comédiens à différents types de personnages, un ensemble de critères de décision sont discutés dans la sous-section 2.2.2.

2.1.3 Le rôle de la perception et des stéréotypes dans le doublage

Avant de définir les critères qui sont impliqués dans les décisions des DA, il est important de prendre conscience du rôle clef que joue la perception humaine. Tout comme au théâtre, les voix de films et de jeux vidéos jouent ou surjouent des émotions et des caractères de personnages pour véhiculer des messages clairs au public. Il est alors nécessaire de maîtriser le contenu du message transmis par la voix du comédien. Celle-ci ne diffuse pas seulement le contenu linguistique – comme expliqué dans le prochain chapitre présentant la voix actée – elle apporte aussi des informations sur les émotions et la personnalité du personnage. Si la voix réceptionnée par le public s'éloigne trop du caractère du personnage, l'auditeur peut ressentir un effet de dissonance qui l'empêchera de s'immerger dans l'univers de l'œuvre.

Le terme « stéréotype » est un concept clef dans la compréhension de la perception du public. Dans le Larousse, il est définit comme la « caractérisation symbolique et schématique d'un groupe qui s'appuie sur des attentes et des jugements de routine ». En l'introduisant comme étant un « cliché obtenu par reproduction » ([Rey05]), [Paq19] présente la stéréotypie comme un « cliché » qui « prend l'aspect d'un personnage ou d'une situation type, s'approchant alors de la "caricature" ». Ces caricatures s'inscrivent dans un contexte sociétal et peuvent évoluer. Le type de personnage « patron » en est un parfait exemple. Son image a accompagné la refonte politique de la France durant la période 1976-1997. Le patron était d'abord largement dépeint en « salaud » avec comme acteur emblématique du patron sûr de lui, arrogant et cynique, Michel Piccoli. Ce cliché a ensuite évolué et les personnages qui en ont dérivé ont été présentés comme des êtres soucieux de leurs salariés notamment dans *Ma petite entreprise* (1999) de Pierre Jolivet avec Vincent Lindon et *Le goût des autres* (2000) d'Agnès Jaoui avec Jean-Pierre Bacri. Les stéréotypes prennent aussi vie dans les accentuations vocales. [Lan19] présente un stéréotype apprécié par les films hollywoodiens où les « méchants » sont très souvent joués par des acteurs britanniques. Dans cet article, les auteurs affirment que Walt Disney « semble avoir un penchant particulier pour les rôles de méchantes incarnées par des femmes à l'accent britannique ». Ils font notamment référence à la reine maléfique de Blanche-Neige et à Cruella de Vil des 101 Dalmatiens. Selon les auteurs, Hollywood et Walt Disney donnent à l'accent britannique une connotation de pouvoir qui sied bien aux personnages méchants. Dans ce manuscrit, l'auteur pose l'hypothèse qu'il existe d'autres stéréotypes de voix qui sont implicites, du moins en France. Ceux-ci sont des polarités formées

par l'imaginaire collectif – non formellement listés par des professionnels – qui regroupent des personnages partageant les mêmes traits de caractères ou des traits vocaux. Par exemple, la voix de héros est plutôt rassurante, grave et séduisante. Lorsque le personnage est plus timide, une voix moins assurée et moins grave lui est généralement associée. Les voix nasillardes, un peu plus dérangeantes, vont très souvent être jouées par des personnages secondaires ou des méchants moins appréciés du public.

Il existe un autre phénomène similaire à la stéréotypie, nommé la « commodification des voix » ([Pla14]). La commodification est la transformation des voix en marchandises, vendues comme produits de différenciation (et indirectement de stigmatisation) avant d'être éventuellement réappropriées par les communautés que ces voix contribuaient à discriminer. La sélection et le jeu des comédiens sont alors sujets aux attentes de l'industrie du cinéma pour répondre à l'offre et à la demande (i.e., la loi du marché). Comme expliqué dans [Pla14], la commodification peut être un levier de stigmatisation. Un exemple y est donné avec Yasmine Modestine à qui le rôle d'un personnage noir a été refusé parce que sa voix ne correspondait pas à l'attente du public, autrement dit, parce qu'elle n'avait pas une voix de « Noire ». Elle explique aussi dans une citation de [Mod] que les acteurs étaient soumis à des effets de discrimination. Il convenait effectivement à une époque de considérer que «"les comédiens noirs ont des voix graves de Noirs" et les comédiens asiatiques ont une "voix aiguë d'asiatique". Les comédiens blancs ont la chance d'avoir une tessiture suffisamment étendue qui permet de doubler et les Noirs et les asiatiques et les Blancs.» — comme en témoigne Yasmine. Yasmine n'a pas été la seule à dénoncer cette prédominance de représentations ethniques figées véhiculées par le cinéma français, d'autres comédiens l'ont aussi fait. Même si des instituts comme la Haute Autorité de Lutte contre les Discriminations et pour l'Égalité (HALDE) ont finalement traité ce dossier et conclu « l'existence de préjugés persistants » dans les métiers du doublage [1], ce phénomène de discrimination reste encore d'actualité et prend parfois d'autres formes comme le suggère le tweet de Greg Germain traitant de la mesure « vraiment nulle » prise par la production des *Simpsons* qui refuse que des acteurs « blancs » doublent des « non-blancs » ; l'avis de Greg Germain étant que la bonne manière pour lutter contre le racisme culturel en France « serait que les acteurs non-blancs puissent comme les acteurs blancs doubler n'importe quel rôle » ; à partir du moment où leurs voix correspondent. Ce problème sociétal est à

1. Une note datée du 29 décembre 2009 rappelait que « le choix d'un comédien-doubleur devait se faire en fonction de sa qualité de voix et de sa compétence, et non en raison de sa couleur de peau ou de son origine » ([Mon09])

prendre en compte lors de la création d'outils de recommandation qui auront potentiellement un impact sur la discrimination des choix des DA. Il est en effet du devoir du chercheur de prévenir ces usages discriminatoires lorsqu'il fournit un système de recommandation.

2.2 Le Casting Vocal

Le Casting Vocal est le processus qui consiste à demander à différents candidats de jouer un personnage et à n'en sélectionner qu'un qui remplacera la voix d'origine. Ce processus est une tâche de haut niveau intellectuel réalisée par un Directeur Artistique. Ci-dessous, nous décrivons le processus du Casting Vocal. Les protagonistes de ce processus sont d'abord présentés. Puis, nous listons les critères intervenant dans les décisions. Pour finir, nous identifions deux manières de sélectionner une voix : avec, ou sans audition.

2.2.1 Les protagonistes du casting vocal

Trois protagonistes participent à la sélection de la voix qui doublera un personnage d'origine.

Le premier protagoniste est le comédien qui remplacera la voix d'origine. Il est appelé à jouer le personnage lors d'un casting vocal pour vérifier s'il est apte à jouer le rôle.

Le second protagoniste est le Directeur Artistique, ou le Voice Caster. Il est celle ou celui qui sélectionne le comédien qui remplacera la voix d'origine.

Pour finir, le client (le distributeur) est le troisième protagoniste intervenant dans ce processus. Le client paie une entreprise de prestation pour réaliser le doublage d'une œuvre, dont fait partie le DA. Il suit le déroulement du doublage et est en constante discussion avec le DA. Il influence parfois les décisions du DA, notamment pour des raisons économiques ou de marketing.

2.2.2 Les critères de sélection

Le casting vocal est un tâche de haut niveau intellectuel difficile à automatiser. Identifier les critères qui interviennent dans ce processus de décision est un défi, non seulement pour la discipline informatique, mais aussi pour d'autres disciplines telles que la sociologie ou les neurosciences. Dans la précédente partie, il a en effet été mis en avant que les décisions artistiques, qu'elles soient pour la sélection des comédiens ou la direction des jeux d'acteur, sont dépendantes

du contexte culturel et sociétal – et indirectement de la perception du public. Ce niveau de complexité est mis en exergue dans l'article [Bon19] où la tâche de casting vocal est qualifiée de « sous-définie » car les opérateurs eux-mêmes éprouvent des difficultés à décrire précisément les buts et les fondamentaux du casting vocal, même s'ils font ce travail tous les jours. Travailler sur ce type de problème est complexe parce que la seule connaissance sûre et disponible du processus de casting vocal concerne les choix réalisés précédemment par les opérateurs, soit les productions déjà doublées. Quand bien même il est impossible aujourd'hui de lister précisément les critères de décisions des DA, une première proposition d'identification de ces critères est présentée ci-dessous.

Les critères impliqués dans la sélection d'une voix ne sont pas toujours liés à la voix elle-même. D'autres critères, tels que les critères économiques, peuvent intervenir. Par exemple, il est possible qu'un DA demande à un comédien de jouer plusieurs personnages pour économiser de l'argent. Ce phénomène est fréquemment observé dans les jeux vidéos où beaucoup de personnages secondaires interviennent. Ils prennent généralement très peu la parole et les faire jouer par le comédien d'un personnage principal n'altère que très peu l'immersion du public. À l'opposé, le client peut exiger une voix star pour des raisons marketing. Le jeu de l'acteur s'adapte alors au maximum à son personnage dont l'identité est généralement modulée. Parfois aussi, le public associe la voix d'un comédien de doublage à un acteur étranger. Dans ce cas, le même comédien de doublage est souvent rappelé pour jouer les personnages interprétés par cet acteur. Les comédiens qui doublent régulièrement les personnages d'un même acteur sont appelés des « voix régulières ». Par exemple, le comédien Patrick Poivey était la voix régulière française de Bruce Willis, et nous pouvons considérer que Greg Germain est celle de Will Smith (bien qu'il ne l'ait pas doublé dans tous ses films).

Certains DA se basent parfois sur la physiologie des comédiens pour prendre leurs décisions. La similarité physiologique passe alors par la ressemblance du comédien de doublage avec l'acteur d'origine. Cette ressemblance peut aussi être intimement liée à l'âge. Il arrive que le doubleur vieillisse et que les maisons de production décident de changer de doubleur récurrent parce qu'elles considèrent sa voix trop vieillissante. Cela a apparemment été le cas récemment pour Greg Germain à qui il avait été communiqué que les maisons de productions américaines avaient lu sa fiche de comédien et l'avaient trouvé trop âgé pour interpréter le rôle de Will Smith. [2]

Comme vu précédemment, la localisation d'une voix est dépendante des

2. Interview de Greg Germain : https ://www.youtube.com/watch ?v=ywVDoC-0jn8

codes culturels du pays ciblé. Ces codes changeant d'un pays à l'autre, il est alors à supposer qu'un même personnage est joué de différentes manières en fonction de la culture cible. Par exemple, l'aspect séduisant d'une voix ne s'exprime pas par les mêmes propriétés acoustiques en France qu'au Japon. Les cultures de ces deux pays sont éloignées mais certaines différences vocales sont aussi observables entre la France et les États-Unis où les voix d'homme et de femme ont tendance à avoir une tonalité plus basse. Sachant toutes ces informations, il est supposé dans ce manuscrit que les critères vocaux de la sélection d'un comédien ne sont pas l'aboutissement d'une simple similarité acoustique de sa voix avec celle de l'acteur d'origine. Prendre en compte une translation culturelle des codes vocaux pour adapter son choix à la culture cible s'avère important.

La palette vocale – développée dans le prochain chapitre – est l'espace dans lequel le comédien peut naviguer pour moduler sa voix à sa guise. En jouant avec sa voix, le comédien peut prendre un accent, jouer une émotion particulière, ou adapter sa voix au caractère d'un personnage. Des professionnels de l'entreprise Dubbing Brothers nous ont appris lors de certains entretiens que l'amplitude de la palette vocale est un critère clef de la sélection de voix, parfois plus important que le timbre, et que le DA s'intéresse particulièrement à l'expérience théâtrale du candidat qui est hautement corrélée à la qualité et à la palette de sa voix.

Un dernier aspect du casting vocal concerne la sélection plurielle de voix. Il est à supposer que choisir des voix trop similaires pour des personnages différents peut dégrader l'immersion du public. Le casting vocal devient alors un problème de sélection de plusieurs voix. Cette approche présente plusieurs avantages. D'abord, elle permet de sélectionner des voix distinctes les unes des autres pour éviter de les confondre. Il permet aussi l'accentuation d'un trait de caractère d'un personnage. Une voix atypique, distincte des autres personnages, peut éventuellement donner un caractère singulier à un protagoniste.

2.2.3 Sélection de comédien avec ou sans audition

Pour sélectionner un comédien, un DA (ou directeur de casting) peut auditionner différents candidats. L'audition consiste à demander à chaque candidat de se présenter et de jouer le script qui lui a été fourni. Le DA se base ensuite sur l'interprétation de chaque candidat pour sélectionner celui ou celle qui remplacera la voix du personnage à doubler. Dans certaines auditions, le rôle du script fourni aux candidats n'est pas en lien avec le personnage à doubler.

Pour des raison de coût et d'efficacité, les auditions sont de plus en plus réalisées en ligne. Les comédiens disposent d'un home studio d'où ils enregistrent leurs essais et les font parvenir à la personne en charge du recrutement, e.g. le directeur de casting.

Il est aussi possible pour le DA de sélectionner des comédiens sans audition. Le DA se fit à l'historique des personnages joués par le comédien. Le DA se crée ainsi des représentations des palettes vocales des candidats et sélectionne le plus approprié. Son choix est réalisé en amont et indépendamment de l'interprétation du personnage par le candidat.

Pour aider les experts à sélectionner des candidats avec ou sans audition, de nombreuses plateformes de recherche de comédiens ont émergé ces dernières années [3]. Les comédiens s'y présentent en mettant à jour leur profil et les « *voice casters* » y recherchent des comédiens parmi ceux disponibles. Ces plateformes ne proposent pas de systèmes de recommandation mais seulement des systèmes de recherche basés sur différents critères tels que le genre, la tessiture et l'âge. Il existe une réelle demande de systèmes de recommandation de voix pour aider les DA à découvrir de nouveaux talents qui sont masqués par la masse d'acteurs présents sur le marché.

Conclusion

Dans ce chapitre, nous avons d'abord introduit le doublage vocal. Nous avons vu que le doublage vocal est un processus complexe faisant notamment intervenir différents opérateurs. Il consiste à remplacer la voix d'origine d'un personnage par la voix d'un comédien dans une langue cible. Dans ce but, le comédien joue avec sa voix pour correspondre aux attentes perceptives du public et pour mettre en exergue le personnage, intégrant notamment différents stéréotypes (culturels, linguistiques...).

Nous avons ensuite introduit le casting vocal. Nous avons vu que les critères de décision du Directeur Artistique (DA) ne sont pas seulement liés à la voix, mais aussi à des aspects économiques ou physiologiques. Aussi, nous avons vu qu'il est possible de sélectionner un comédien de doublage avec ou sans audition. Dans le cas de l'audition, les candidats jouent le script qui leur est donné et l'un d'entre-eux est sélectionné par le DA pour jouer le personnage à doubler. Dans le cas contraire, le DA se base uniquement sur l'historique du candidat.

3. Des plateformes américaines voicecasting.com ou voices.com comme des plateformes françaises voxingpro.fr ou voicematch.fr

Pour mieux comprendre comment automatiser le processus de casting vocal, il est nécessaire de définir plus précisément le sujet central de nos expérimentations, à savoir, la voix actée. Dans le prochain chapitre, nous présentons les aspects qui interviennent dans la production de la voix actée ainsi que les informations qu'elle véhicule.

Chapitre 3

La voix actée, ce que l'acteur joue et ce que la voix exprime

Sommaire

Selon le Larousse, l'une des définitions de la voix est la « faculté d'émettre des sons », en de l'humain. La parole, quant à elle, représente la façon dont nous utilisons notre voix et le langage pour nous exprimer. La voix est étudiée dans de nombreuses disciplines telles que la psychologie, la linguistique, la phonétique, la sociologie, les neurosciences, ainsi que tous les domaines qui en dérivent (e.g., la psycho-acoustique). À travers elle, l'informatique s'inté-

resse à réaliser, avec une machine, différentes tâches telles que la reconnaissance de la parole, la reconnaissance du locuteur, l'identification du langage, la reconnaissance des émotions, etc. Le développement des applications sur le traitement de la parole, notamment la parole spontanée, a explosé avec l'apparition des pratiques d'apprentissage automatique, particulièrement depuis l'essor des réseaux de neurones. Ce chapitre propose de définir la voix actée et de la positionner dans la littérature parmi les autres qualificatifs de voix.

Dans notre société, nous utilisons quotidiennement notre voix pour échanger des informations structurées sous forme de phrases. Lorsque nous produisons de la parole, des informations dites « non-verbales » sont aussi échangées. Elles peuvent être des émotions, des intentions, un indice sur le statut social du locuteur ou son identité. Tous les jours, la voix est employée sous différentes formes. Elle est dite *ordinaire* lorsqu'une personne s'exprime naturellement. Elle prend le qualificatif de *conversationnelle* lorsque deux individus discutent. Elle est *pathologique* lorsqu'une personne présente des troubles de la parole. Elle est *expressive* lorsqu'elle transmet une émotion. Elle prend finalement le qualificatif de *voix actée* lorsqu'elle est produite pour être réceptionnée par un public (e.g., spectateur d'une œuvre multimédia telle qu'un film projeté au cinéma). Par contraste avec la parole spontanée, l'étude de la parole actée demeure aujourd'hui marginale dans la communauté scientifique.

La voix actée se distingue de la voix spontanée : le livre [CC10] la définit comme étant « la voix qui agit sur ceux qui l'écoutent, à travers les mots, la musique qu'elle produit, les gestes et le corps qu'elle engage. » La voix actée est une construction fruit d'une interprétation maîtrisée et planifiée de sa voix par un acteur afin de produire un effet désiré chez un spectateur. Par exemple rendre manifeste le comportement d'un personnage fictif et faciliter la crédibilité et l'immersion du public dans une situation ou une trame. Dans ce contexte, la voix est souvent modulée, voire caricaturée, pour rendre audibles et sensibles les effets expressifs produits par le comédien.

3.1 Les aspects intervenant dans la production de la voix actée

La voix actée est un objet d'étude particulièrement complexe car sa réalisation fait intervenir un grand nombre de facteurs.

3.1.1 Le comédien

La modulation de la voix actée est directement dépendante du comédien qui la joue. Ce dernier a sa propre personnalité. Il a un physique et une voix qui lui sont propres. Son expérience lui a permis de développer ses compétences de comédien, dont la maîtrise de sa voix. Tous ces aspects influencent la voix actée lorsque le comédien joue un personnage.

3.1.2 Le personnage

Le comédien produit de la parole actée pour rendre manifeste le comportement d'un personnage. Les signes vocaux caractéristiques de ce personnage peuvent dépendre des époques et des œuvres dans lesquelles il apparaît. Par exemple, le personnage Batman est apparu dans différents contenus multimédia tels que des films, des séries animées et des jeux vidéos. Il est joué par Adam West dans le film « Batman : The Movie » en 1966 [1]. Le Batman dépeint dans cette œuvre est drôle et se rapproche du style des bandes dessinées américaines – les comics – de l'époque. Il se différencie du Batman moderne que nous pouvons trouver dans le film « The Dark Knight Rises » de 2014 [2] où il est joué par Christian Bale. Dans ce film, Batman, ainsi que tous les autres personnages, sont bien plus sombres et le caractère humoristique de ses origines n'est pas mis en avant. Certains traits restent cependant fidèles à ce personnage. Quelque soit l'œuvre, il est un super-héros (i.e., type) masculin (i.e., genre) et sans super-pouvoir, riche, fort physiquement (i.e., physiologie) et il se bat pour la justice (i.e., psychologie). Il porte un costume de chauve-souris et joue avec ses deux identités, à savoir l'homme riche et le super-héros. Batman se caractérise aussi par ses *punch lines* (i.e., langage) violentes et sa complicité avec son majordome Alfred. Dans les œuvres modernes, la voix de Batman a pris une place toute particulière. Ce dernier modifie sa voix pour ne pas être reconnu et prend un ton très grave. La phrase « I am Batman », prononcée avec un ton et une fréquence très grave est devenue culte et est couramment reprise dans d'autres œuvres telles que la série « Community » [3]. Elle est devenue la signature de Batman sans laquelle ce personnage ne serait plus le même, quelque soit la langue [4].

1. Bande annonce disponible sur YouTube
2. Bande annonce disponible sur YouTube
3. Scène du personnage Hamed dans Community qui se déguise en Batman disponible sur Youtube
4. Une vidéo YouTube est disponible avec « I am Batman » joué dans différentes langues

3.1.3 Les aspects culturels

Les codes culturels ne sont pas figés dans le temps. Les stéréotypes évoluent au cours des années et peuvent être injectés par les productions cinématographiques tout comme elles peuvent être l'exagération d'un trait sociétal, presque éthologique (e.g., voix d'homme ou de femme aux aspects séduisants). La complexité des codes culturels semble cependant ne pas s'arrêter à une dynamique temporelle. En effet, il est à supposer qu'à un même instant t deux individus différents ne perçoivent pas de la même manière la voix qu'ils écoutent, même s'ils sont originaires d'un même pays ou d'une même communauté. Pour s'assurer de la bonne réception du caractère du personnage par le public, il est nécessaire de prendre en compte ces aspects culturels lorsqu'un comédien joue un personnage.

3.1.4 La Direction Artistique

Le dernier aspect intervenant dans la production de la voix actée est la Direction Artistique. Celle-ci guide le comédien lors du processus de doublage pour correspondre au mieux aux codes du public et ainsi assurer la cohérence de sa voix dans l'œuvre. Le DA est l'expert qui guide le comédien pour atteindre cet objectif. Une bonne expérience cinématographique ainsi qu'une bonne connaissance du public et de ses attentes sont des qualités qu'elle ou il doit avoir développées, notamment pour réaliser ses choix artistiques. Le DA travaille en étroite collaboration avec son client, souvent une maison de production, pour être au plus près de ses attentes. Par simplification, nous considérons que les décisions de la maison de production (le client de l'entreprise de doublage) relèvent de la Direction Artistique même si ce n'est pas réaliste : il est par exemple envisageable de ne pas sélectionner un comédien pour des raisons financières. Les maisons de productions peuvent aussi refuser un comédien parce que, selon elles, ce dernier ne correspond pas aux attentes du public. Comme nous l'avons expliqué plus tôt, d'après ses propres propos, Greg Germain s'est vu refusé le doublage des derniers films de Will Smith parce que les maisons américaines considéraient qu'il était trop âgé.

3.2 Les informations présentes dans la voix

Lorsqu'une ou un comédien joue, elle ou il use de sa voix comme support pour produire de la parole. Cette voix véhicule différentes informations dont une partie a déjà été identifiée dans la littérature scientifique telle que les émotions, l'intention ou encore la langue. Ces informations sont exprimées

consciemment ou inconsciemment par le locuteur et sont détectables et analysables pour un être humain ou pour certains systèmes automatiques. Certaines de ces informations peuvent aider à caractériser le personnage joué tandis que d'autres peuvent, au contraire, nous induire en erreur. Dans cette section, nous présentons une liste de types d'informations exprimées par la voix du comédien que nous avons identifiée.

3.2.1 Les émotions et les attitudes

Les émotions sont un aspect fondamental du quotidien. Elles sont intégrées dans nos réactions psycho-physiologiques et sont l'une des premières sources d'information en terme de communication et d'interaction entre les humains. Chaque individu et chaque culture présentent des caractéristiques singulières qui tendent à rendre les expressions émotionnelles et leurs interprétations dépendantes de l'individu et de sa provenance.

Selon les psychologues américains, Robert Plutchik a créé les théories de classification émotionnelles avec une taxonomie des émotions humaines qui est divisée en huit émotions primaires ([Plu01]). Dans un plan trigonométrique, ces émotions sont placées proches lorsqu'elles sont similaires et sont écartées de 180 degrés lorsqu'elles sont opposées. D'autres émotions sont des mélanges d'émotions primaires, à l'instar des couleurs composées de couleurs primaires. Une troisième dimension représente l'intensité des émotions, de telle sorte que le modèle total dit « structurel » des émotions prenne la forme d'un cône. Dans la figure 3.2, les manifestations extrêmes des émotions primaires sont dépeintes au centre de la roue.

L'attitude est définie par [Wic00] comme un prédicteur de comportement social et diffère de l'émotion primaire. Selon [MO20], il existe différents types d'attitudes. Les attitudes sociales permettent de représenter l'attitude d'un locuteur envers son interlocuteur pendant une interaction inter-humain. Les auteurs expliquent aussi qu'elles se différencient des émotions qui sont des états internes d'un locuteur et des attitudes propositionnelles qui sont des attitudes d'un locuteur vis-à-vis d'un énoncé. Par exemple, l'article s'est intéressé aux attitudes sociales suivantes pour construire un corpus de données pour réaliser des études de prosodie : la dominance, la séduction, l'attitude amicale et l'attitude distante. Le choix de ces attitudes fut notamment inspiré de [AC17] qui a proposé récemment une catégorisation originale des attitudes de jeu des musiciens inspiré par le modèle de rose de Leary ([Lea57]). Les attitudes y sont décrites dans un espace bi-dimensionnel, la première dimension reflète l'hostilité et l'amicalité tandis que le second reflète la position du musicien dans une

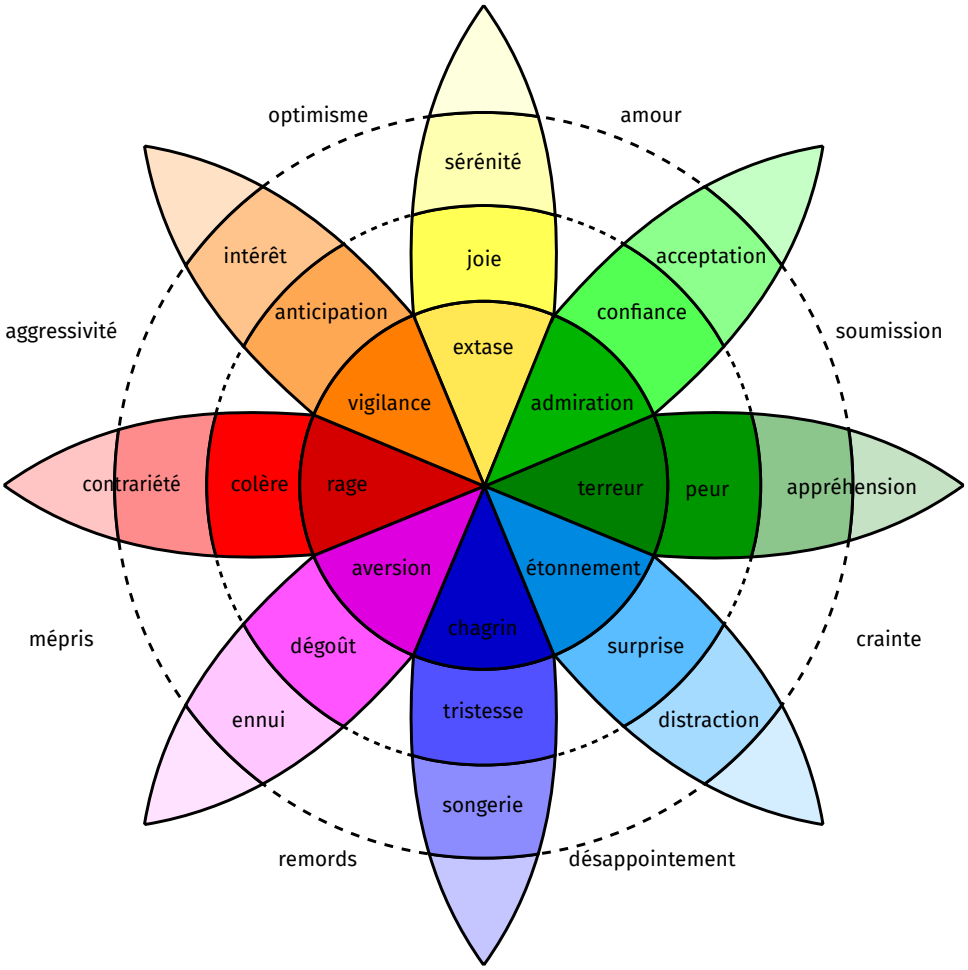


FIGURE 3.1 – Roue des émotions selon le modèle de la perception humaine de Plutchik. Les dimensions verticales des cônes représentent l'intensité, et le cercle représente les degrés des similarité entre les émotions. Les huit secteurs sont pensés pour indiquer qu'il y a huit dimensions d'émotions primaires définies par la théorie comme quatre paires d'oppositions. Dans ce modèle les émotions dans l'espace vide sont des diades primaires – émotions qui sont des mélanges de deux émotions primaires. ([Plu01])

hiérarchie sociale.

Le champs de recherche de l'informatique émotionnelle a pour objectif de donner aux machines l'intelligence de comprendre les émotions humaines et de synthétiser les émotions et les comportements émotifs ([Pic03]). Elle trouve sa place dans différentes applications telles que les jeux vidéos, les *serious game* pour les jeunes ayant un trouble du spectre de l'autisme ou encore l'interaction humain-machine. Du côté de la voix, cette intelligence se traduit par deux tâches distinctes :

— La reconnaissance vocales des émotions ([Sch18]) qui consiste à recon-

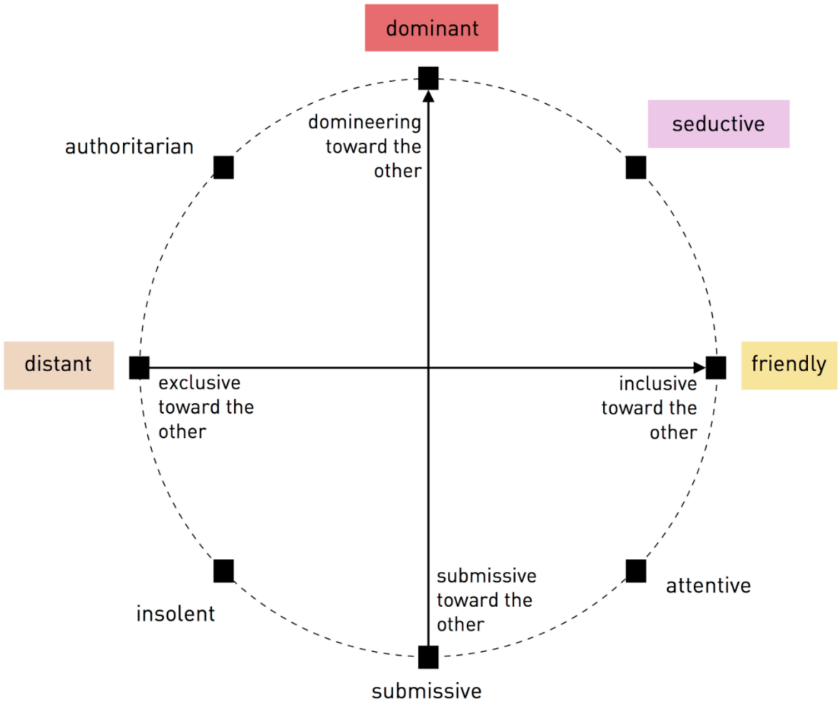


FIGURE 3.2 – Attitudes sociales représentées dans un espace à 2 dimensions proposées par [MO20] et inspirées des catégorisations de musiciens jouant des attitudes de [AC17]

naitre l'émotion d'un discours indépendamment de son contenu sémantique.

— La synthèse vocale ([GBC02 ; Nin+19]) qui consiste à produire de la parole en contrôlant l'émotion qu'elle véhicule.

Lorsqu'un comédien joue un personnage, il est à supposer que certaines émotions et attitudes sont véhiculées au travers de ses paroles. D'une part, le comédien doit simuler des émotions pour les rendre audibles aux spectateurs, et d'autre part il est probable que le comédien exprime des émotions indépendamment de sa volonté, bien que ces effets soient limités par ses compétences d'acteur. Ces informations sur les émotions et les attitudes sont probablement des indices qui aident à caractériser le jeu du comédien et le personnage joué.

3.2.2 La variabilité locuteur

Tous les jours nous identifions plus ou moins facilement des locuteurs familiers en écoutant leurs voix. Cette capacité humaine est rendue possible grâce à la variation entre les voix de différents locuteurs, connue sous le nom de « variabilité inter-locuteur ». Bien qu'il est supposé que différents locuteurs aient des voix différentes, la voix d'un locuteur n'est pas constante et varie toujours. Un locuteur ne prononce jamais un même mot exactement de la même ma-

nière. Cette propriété est connue sous le nom de « changement de style » ou « variabilité intra-locuteur » ([ER02]).

Lorsqu'un locuteur s'exprime, différentes variabilités peuvent être observées dans des enregistrements. Certaines sont liées au locuteur lui-même telles que le stress relatif à la tâche ([Han96]), l'effort vocal ([FH11 ; ZH11 ; ZH07 ; HNS17 ; HV09 ; MH13]), les émotions ([Vlo+00]) et la physiologie ([Lan10]). D'autres variabilités sont liées à l'interlocuteur qui peut être soit un humain, soit une machine. Elles peuvent aussi être liées à la technique et aux perturbations externes avec le canal de transmission ([Rey+95 ; Ken+07 ; ACLT00]), l'environnement externe ([RHR94 ; JSW07 ; Gre+10]) et la qualité des données ([KB97 ; Bes+00]).

À l'instar de la variabilité inter- et intra-locuteur, nous pouvons supposer qu'il existe dans la voix actée des variabilités inter- et intra-personnage. Ce qui nous intéresse dans ce manuscrit est de comprendre quels sont les signes caractéristiques d'un personnage et en quoi ils sont différents des signes caractéristiques du locuteur lorsque ce dernier ne joue pas ce personnage. Lorsqu'un comédien joue un personnage, nous supposons que certains aspects de la voix peuvent être partagés, ajoutés ou supprimés :

— **Aspects partagés :** certains aspects restent présents et aident à identifier le comédien. Ces signes peuvent provenir de la physiologie de son organe vocal dont il n'a pas la maitrise.
— **Aspects ajoutés :** le comédien ajoute des aspects propres au personnage qui ne sont pas observables lorsqu'il parle spontanément ou lorsqu'il joue d'autres personnages. Cela peut aussi être l'exagération d'un trait acoustique pour insister sur le caractère du personnage.
— **Aspects supprimés :** le comédien cache certains de ses traits vocaux spontanés, i.e. un accent.

D'autres questions, que nous ne traitons pas dans ce manuscrit, peuvent être posées. La capacité du locuteur à modifier sa voix est ce que nous pouvons appeler la *palette vocale*. Nous pouvons ainsi poser des questions autour de la capacité d'un individu à modifier sa voix et tromper son interlocuteur en se faisant passer pour autrui. Dans le contexte du doublage, si nous connaissons la voix « neutre » d'un comédien – i.e. lorsqu'il ne joue pas de personnage – est-il possible de le reconnaitre lorsqu'il joue un personnage ? Est-ce qu'un individu peut jouer la comédie pour se faire passer pour un autre individu, sans être aidé par un automate qui modifie sa voix ? La question de la palette vocale telle que nous la définissons pourrait aider à répondre non seulement

aux enjeux de la caractérisation de la voix actée mais aussi à des enjeux liés à la voix en générale et à l'identification du locuteur.

3.2.3 Le contenu linguistique

Chaque personnage peut employer des champs lexicaux différents et avoir un langage plus ou moins soutenu. Il est à supposer que le rapport à la langue d'un personnage dépend des attentes qui lui sont liées. Nous pouvons imaginer par exemple qu'un soldat dans un jeu vidéo emploiera des termes liés à la guerre. À l'instar de l'information locuteur, le contenu linguistique est un biais et ne doit pas servir à caractériser la dimension personnage dans la voix actée.

3.2.4 La langue et les accents

En fonction de notre degré de maîtrise, il est plus ou moins facile de reconnaitre la langue dans laquelle parle notre interlocuteur. Comme indiqué dans [Mar+11], chaque individu construit son propre style de parole, notamment à partir de son dialecte et de son accent. [Fle84] a par exemple observé dans ses expériences que l'accent d'un français qui parle anglais est facilement détectable pour un natif américain. Certains comédiens peuvent employer des accents pour insister sur certains traits de personnalité du personnage qu'ils jouent, e.g. l'accent britannique pour les vilains. L'accent peut donc être un indice vocal intéressant pour caractériser le personnage.

3.2.5 Les informations non liées à la voix

Pour finir, d'autres informations non liées à la voix sont présentes dans les enregistrements vocaux et sont importantes car elles peuvent servir à tort à la caractérisation d'un personnage. Nous listons les principales ci-dessous :

— **Le microphone employé :** chaque microphone a sa signature sonore et laisse une trace dans les enregistrements audio.
— **La distance du microphone**
— **Le volume sonore**
— **Ambiance :** dans certains enregistrements, la parole actée peut être accompagnée de différents sons d'ambiance tels qu'une musique, des bruits (e.g., une chaise déplacée ou le son d'une voiture), des voix (e.g. des individus discutant en arrière-plan dans un caffé) etc.
— **Effets sur les voix :** certains effets sont appliqués aux enregistrements. Par exemple, des filtres peuvent être appliqués donner l'impression au spectateur qu'un enregistrement vocal sort d'une radio.

Conclusion

Dans ce chapitre, nous avons introduit la notion de voix actée. Nous avons d'abord abordé la complexité de la voix actée en présentant les nombreux aspects intervenant dans sa production, à savoir les aspects liés au comédien, au personnage joué, à la culture ou à la direction artistique.

Nous avons ensuite présenté les différentes informations présentes dans la voix actée. Ces informations sont exprimées lorsque le comédien joue un personnage. Elles sont liées aux émotions du comédien, à son attitude, au comédien lui-même (la variabilité locuteur), au contenu linguistique, à la langue et aux accents.

À l'image de la variabilité locuteur, nous introduisons aussi le concept de variabilité personnage que nous nommons parfois dans ce manuscrit « information personnage » ou « dimension personnage ». Nous supposons que les variabilités locuteur et personnage sont différentes et cherchons à déterminer les signes caractéristiques d'un personnage.

Afin de mieux comprendre comment modéliser les variabilités locuteur et personnage, nous présentons, dans la prochaine partie, l'état de l'art en Reconnaissance Automatique du Locuteur (RAL) et en Casting Vocal Automatique (CVA).

Deuxième partie

Caractériser la dimension personnage véhiculée par la voix actée

Chapitre 4

La Reconnaissance Automatique du Locuteur (RAL)

Sommaire

Pour nous, êtres humains, reconnaitre la personne qui parle à partir du signal qu'elle émet est généralement facilité lorsqu'elle nous est familière ([VLKE85]),

e.g. la voix d'un proche ou d'une célébrité. Selon Hansen ([HH15]), la familiarité avec une voix dépend du temps passé à l'écouter. La capacité à reconnaitre une voix dépend aussi du contexte d'écoute. Dans la rue, certains bruits peuvent gêner cette identification. Au téléphone, la limitation de la bande passante rend les voix plus difficiles à reconnaître. Reconnaitre automatiquement celui ou celle qui parle se nomme la Reconnaissance Automatique du Locuteur (RAL). Ce champs d'étude a permis l'émergence et l'évolution de technologies telles que des applications de sécurité dans les banques, des applications d'assistance à la comparaison criminalistique de voix, ou encore la personnalisation de conversations pour les assistants vocaux (e.g., Siri, Google home).

Les travaux présentés dans ce manuscrit s'inscrivent dans la continuité de [Gre20] qui a proposé d'adapter les techniques employées en RAL pour construire et évaluer des modèles de similarité de voix actées. Dans ce chapitre, nous présentons en quoi consiste la reconnaissance du locuteur et détaillons les méthodes développées à cet effet.

4.1 Les fondements de la Reconnaissance Automatique du Locuteur

La Reconnaissance Automatique du Locuteur est généralement décrite comme l'authentification d'une personne par sa voix. Ci-après, nous présentons les différentes tâches qui composent la RAL et en décrivons la structure d'un système.

4.1.1 Dépendance ou non au texte

Il est possible de diviser les systèmes RAL en deux types : les systèmes dépendants du texte, et ceux qui y sont indépendants. Dans les systèmes dépendants, le texte prononcé par le locuteur en phase d'utilisation est le même que celui qui a été prononcé lors de la phase d'apprentissage. A contrario, le locuteur peut prononcer n'importe quelle phrase pour être reconnu par un système indépendant du texte. Néanmoins, il existe plusieurs niveaux de dépendance du texte suivant les applications ([BCP94]). Les systèmes à texte libre (*free text*), à texte suggéré (*text-prompted*), dépendants de traits phonétiques (*speech event dependent*), personnalisés dépendants du texte (*user-specific text dependent*). Leurs spécificités sont détaillées dans [BK17].

(a) Identification du locuteur (b) Vérification du locuteur

FIGURE 4.1 – Tâches d'identification et de vérification du locuteur

4.1.2 Les tâches de la RAL

Dans le domaine de la RAL, deux tâches se distinguent : l'identification et la vérification du locuteur.

L'identification consiste à reconnaitre la personne qui parle dans un enregistrement audio parmi un ensemble de locuteurs. En RAL, cette tâche est réalisée automatiquement par une machine. Comme illustré dans la figure 4.1a, elle suit un scénario dit d'identification en « espace clos » (*in-set* en anglais) où le locuteur à identifier est nécessairement présent dans le panel de locuteurs enregistrés par le système.

La vérification consiste à accepter ou refuser l'identité proclamée par un locuteur, en se basant sur un modèle qui lui est associé. Comme illustré dans la figure 4.1b, un système de vérification prend en entrée un enregistrement de test ainsi qu'une identité proclamée. La tâche consiste alors à prendre une décision binaire qui va confirmer ou infirmer le fait que l'enregistrement de test soit effectivement prononcé par le locuteur proclamé.

4.1.3 Structure d'un système de RAL

La construction d'un système de RAL – d'identification ou de vérification – se déroule en trois étapes illustrées dans la figure 4.2 :

① **Paramétrisation.** Cette étape vise à capturer les paramètres caractéristiques de la parole d'un enregistrement donné. Les paramètres basés sur la représentation spectrale de la parole sont corrélés à la forme du conduit vocal et sont les plus répandus dans les systèmes de RAL modernes

② **Modélisation.** Les paramètres acoustiques extraits d'un enregistrement ou d'un ensemble d'enregistrements donné sont utilisés pour construire un modèle pour chaque locuteur. Ce modèle résume l'information acoustique importante pour caractériser le locuteur ou la locutrice.

③ **Décision.** La phase de décision désigne soit l'identification du locuteur, soit la vérification du locuteur. Dans le cas de la vérification, cette dé-

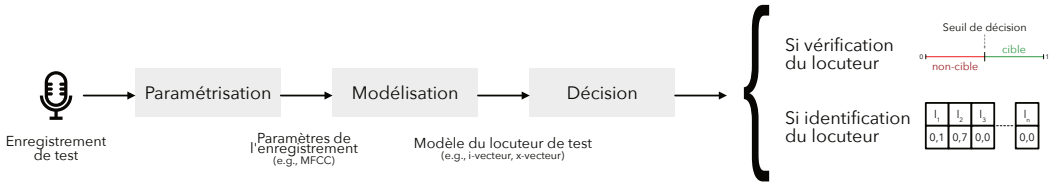


FIGURE 4.2 – Structure générale d'un système de vérification ou d'identification du locuteur.

cision est binaire et consiste à confirmer ou infirmer la correspondance de la session de test à une identité proclamée. Dans le cas de l'identification, le système compare la session de test avec les modèles de n locuteurs et choisit celui dont le score de comparaison est le plus élevé.

4.2 La paramétrisation du signal de parole

La paramétrisation du signal apparaît au tout début des processus des systèmes de RAL et consiste à extraire d'un signal les caractéristiques de la parole. Ci-après, nous présentons les principales approches de paramétrisation employées dans la Reconnaissance Automatique du Locuteur.

4.2.1 Extraire des caractéristiques acoustiques

Un signal est considéré comme stationnaire si ses composantes fréquentielles ne varient pas dans le temps. Bien que la parole soit un signal non stationnaire, l'hypothèse est posée selon laquelle elle peut être représentée approximativement par une succession d'états stationnaires. La paramétrisation d'un signal audio, présentée dans la figure 4.3, consiste à représenter le signal par une séquence de vecteurs numériques, de dimensions fixées. À chaque intervalle de temps Q, un vecteur x_i, appelé trame, est calculé à partir d'une fenêtre de signal de taille K.

Le calcul des MFCC (Mel Frequency Cepstral Coefficients) [1] est la méthode d'analyse cepstrale majoritairement employée en reconnaissance de la parole (haton et al. 2006) et sert de paramétrisation. La figure 4.4 présente ce processus de calcul. D'abord, une phase de pré-accentuation (optionnelle) a pour objectif de transformer le signal de manière à rendre les amplitudes des hautes et des basses fréquences égales. Ceci permet d'éviter une focalisation trop importante sur les basses-fréquences dont les amplitudes sont généralement plus fortes. La seconde étape consiste à réaliser un fenêtrage. Autrement dit, à chaque intervalle de temps Q, une fenêtre de la taille K est extraite du signal

1. Une implémentation en python est fournie par ShigekiKarita sur GitHub

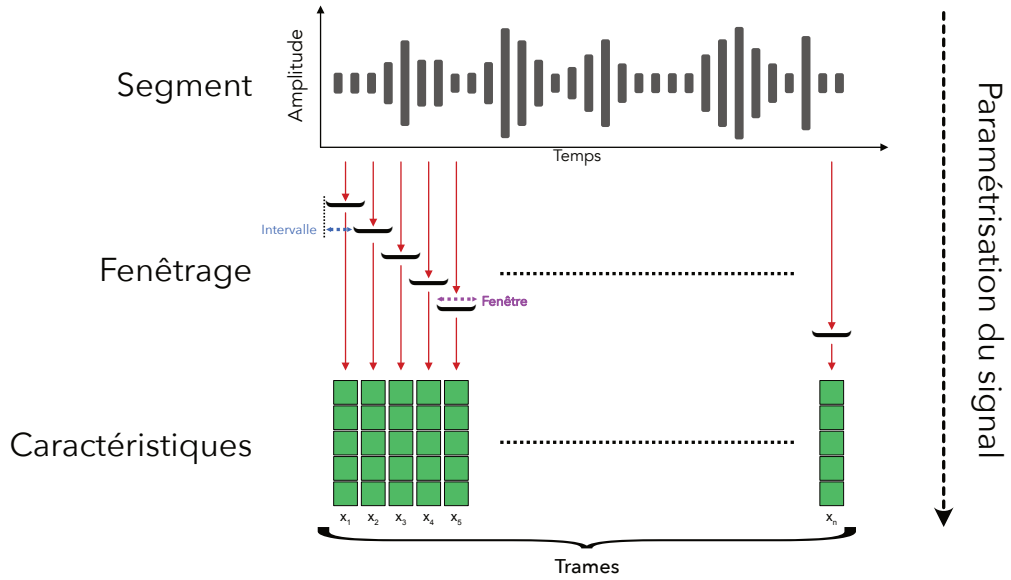


FIGURE 4.3 – Processus de paramétrisation du signal de parole

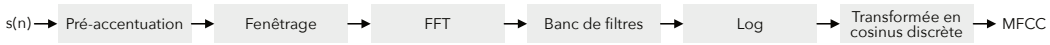


FIGURE 4.4 – Processus de calcul des MFCC (Mel Frequency Cepstral Coefficients)

d'origine. Ce fenêtrage peut induire certains effets de bord. Pour les limiter, une fonction de fenêtrage – telle que la fonction de Hamming – est généralement appliquée sur le signal. Une transformée de Fourier (i.e., FFT pour *Fast Fourier Transform*) est calculée sur chaque fenêtre pour passer le signal du domaine temporel vers le domaine fréquentiel. Cette représentation du signal est communément appelée le spectre. Pour coller à la perception humaine des fréquences qui est non linéaire, un banc de filtres de Mel est appliqué à ce spectre. Puis, une fonction logarithmique ainsi qu'une transformée par cosinus discrète (DCT) sont utilisées sur ce spectre pour extraire les coefficients cepstraux. Pour rappeler l'analogie avec le spectre, l'ensemble de ces coefficients est nommé le cepstre. Pour rappeler la fréquence, chaque coefficient est nommé la quéfrence. La concaténation de ces coefficients est ce que nous nommons les MFCC. Ces coefficients sont généralement complétés par leurs dérivées (notées Δ) et leurs dérivées seconde (notées $\Delta\Delta$). Pour en savoir plus sur les différents paramètres acoustiques, se référer à J.P.Campbell 1997 ; Hansen et al 2015.

4.2.2 La détection d'activité vocale

En RAL, toutes les trames ne sont pas prises en compte. La technique de *Vocal Activity Detection* (VAD) classifie chaque trame comme contenant

de la parole ou du silence. Grâce à cette technique, les systèmes s'appuient principalement sur les trames contenant de la parole et ignorent simplement les autres.

4.2.3 La paramétrisation prosodique

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 ([DDK07 ; SS08 ; FSS10 ; Sch+]). 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.

4.3 Évaluer un système de vérification du locuteur

La tâche de vérification est une tâche binaire qui compare un modèle décrivant l'identité proclamée avec un enregistrement test, i.e. deux modèles locuteur. Les comparaisons sont dites cibles lorsque le locuteur de l'enregistrement de test est celui de l'identité proclamée. Dans le cas contraire, elles sont dites non-cibles. À chaque comparaison, un score de similarité est calculé par le système de vérification. Ce score est ensuite comparé à un seuil de décision pour décider si oui ou non l'identité proclamée est celle du locuteur de l'enregistrement de test.

Ci-après, nous présentons comment évaluer un système de vérification du locuteur, de la méthode de score, jusqu'au calcul de l'erreur.

4.3.1 Les types d'erreurs

Dans le cadre de la vérification locuteur, deux types d'erreurs peuvent être observées :

— **Fausses acceptations (False Acceptance) :** le cas où le système accepte le locuteur alors que celui-ci n'est pas la personne qu'il prétend être.

— **Faux rejets (False Rejects) :** le cas où le système refuse l'accès à un locuteur alors qu'il correspond bien à l'identité proclamée.

Les taux d'erreurs correspondants ; FAR (taux de fausses acceptations) et FRR (taux de faux rejets) sont définis comme suit :

$$FAR = \frac{\#FA}{\#comparaisons\ non-cibles} \tag{4.1}$$

$$FRR = \frac{\#FR}{\#comparaisons\ cibles} \tag{4.2}$$

Pour pouvoir prendre en une décision, un seuil τ doit être fixé. Un seuil de valeur faible résulte en une augmentation des fausses acceptations (FA), alors qu'une valeur élevée donnera beaucoup de faux rejets (FR). Dans un contexte pratique, le réglage de ce seuil dépend de l'application cible et du niveau de sécurité désiré. Pour les applications de haute sécurité, un seuil élevé devrait être fixé de façon à minimiser les erreurs FA. Cependant, pour un système plus permissif, le seuil devrait alors avoir une valeur plus faible.

4.3.2 La mesure Equal Error Rate (EER)

Le taux d'égale erreur (EER pour *Equal Error Rate*) est l'une des mesures les plus populaires en vérification du locuteur et permet de comparer deux systèmes en se basant sur une seule mesure. Elle est définie comme le point opératoire où les valeurs FAR et FRR deviennent presque égales. Cette configuration est atteinte en faisant varier le seuil de décision τ de manière à rendre égales les valeurs FAR et FRR.

4.3.3 La courbe Detection Error Trade (DET)

Selon les objectifs de l'application cible, l'EER n'est pas nécessairement le point opératoire le plus adapté. Il est parfois nécessaire de trouver un compromis entre FA (Fausses Alarmes) et FR (Faux Rejets) différent de l'égalité. Dans ce cas, il est préférable de visualiser les performances d'un système sous la forme d'une courbe pour étudier plusieurs points opératoires à la fois. La courbe ROC est une solution employée dans les problèmes de décision binaire. Le taux de fausses acceptations (false acceptation rate) est tracé sur l'axe des abscisses tandis que le taux de détections correctes (correct detection rate) est tracé sur l'axe des ordonnées.

Une variante de ces courbes, appelée DET (Detection Error Tradeoff ou courbe de détection de compromis) est considérée plus adaptée aux applications de vérification du locuteur ([Mar+97]) et affiche deux types d'erreurs sur les deux axes. Le taux de fausses acceptations (false acceptation rate) est tracé sur l'axe des abscisses et le taux de faux rejets (false reject rate) est tracé sur l'axe des ordonnées.

4.4 Représenter la voix d'un locuteur

En apprentissage automatique, l'apprentissage de représentations est un ensemble de techniques qui permettent à un système d'automatiquement découvrir les représentations nécessaires pour détecter ou classifier des caractéristiques depuis des données brutes. Cela remplace l'ingénierie de caractéristiques manuelle et permet à des machines à la fois d'apprendre des représentations et de les utiliser pour réaliser une tâche spécifique.

L'apprentissage de représentations est motivée par le fait que l'apprentissage de tâches par des machines nécessite des entrées qui sont mathématiques et computationnellement pratiques à traiter. Cependant, les données du monde réel, telles que les images, les vidéos et les données de capteurs, sont difficiles à analyser telles quelles. Une alternative est de découvrir ces caractéristiques ou représentations à travers l'entraînement de systèmes d'apprentissage, sans s'appuyer sur des algorithmes explicites.

L'apprentissage de représentation peut être supervisé ou non-supervisé.

— En apprentissage de représentations supervisé, les représentations sont apprises à partir de données étiquetées.

— En apprentissage de représentations non-supervisée, les représentations sont apprises à partir de données non étiquetées.

En RAL, l'une des problématiques principales est la représentation de la voix présente dans un enregistrement. L'apprentissage de telles représentations doit aider à identifier un locuteur ou à vérifier son identité. Dans cette section, nous présentons les techniques employées en RAL pour représenter la voix d'un locuteur.

4.4.1 Les *i*-vecteurs, une approche traditionnelle

Les GMM

Les mélanges de gaussiennes, ou *Gaussian Mixture Model* (GMM)[2] , sont des modèles statistiques d'estimation de densité de probabilité. Ils sont des modèles génériques dans le sens où ils ne sont pas exclusifs à la RAL. En RAL, un locuteur peut être représenté par un modèle GMM qui estime les densités de probabilité des paramètres acoustiques de ses enregistrements de voix.

Pour que les GMM s'adaptent mieux aux enregistrements qui leur sont inconnus, ils sont généralement entrainés avec l'algorithme *Expectation Maxi-*

2. Article anglais qui détaille le fonctionnement des modèles GMM intitulé Build Better and Accurate Clusters with Gaussian Mixture Models sur www.analyticsvidhya.com

mization[3] (EM).

Les GMM-UBM

En utilisant un GMM et l'EM, le nombre d'enregistrements est généralement insuffisant pour construire un modèle locuteur fiable dans le cas de sessions courtes.

Pour répondre à cette limite, ([RQD00]) propose d'utiliser un modèle générique appelé *Universal Background Model*[4] (UBM).

Un modèle UBM est un GMM à grande échelle (généralement entre 512 et 2048 mélanges de lois gaussiennes avec 24 dimensions) entrainé sur une large quantité de parole, depuis une très grande population. Ce modèle est utilisé pour apprendre une distribution de caractéristiques indépendante du locuteur. Une fois l'UBM appris, les modèles des locuteurs d'apprentissage et de test en sont dérivés par adaptation *Maximum a posteriori* (MAP) ([GL94]) avec l'algorithme EM.

Les supervecteurs

Plus tard, le modèle GMM-SVM a été proposé. Celui-ci consiste à construire ce que nous nommons communément un supervecteur en RAL et à l'utiliser pour apprendre un système de classification avec la technique du *Support Vector Machine* (SVM). Après adaptation d'un UBM à un ensemble de trames, le supervecteur est calculé en concaténant l'ensemble des vecteurs de moyennes du GMM résultant.

Les i-vecteurs

La construction des supervecteurs pose quelques problèmes. L'adaptation MAP n'adapte pas seulement les paramètres aux caractéristiques spécifiques au locuteur, mais aussi au canal et à d'autres facteurs de bruit. Les supervecteurs générés ainsi ne sont donc pas idéaux.

Pour résoudre ce problème, différentes approches par analyse factorielles ont été proposées ([KMD03 ; KBD05 ; Ken+07]). L'approche *Joint Factor Analaysis* (JFA) ([Ken+07 ; Ken+08]) a été la plus plébiscitée. L'intuition derrière cette approche est qu'un supervecteur pour un locuteur donné peut être décomposé en différentes composantes dépendantes ou non au locuteur ainsi qu'au

3. Article de blog sur l'Expectation Maximization disponible sur machinelearningmaster.com
4. Article de blog sur l'UBM GMM disponible sur le site maelfabien.github.io

canal. Ainsi, le supervecteur idéal pour un locuteur peut être représenté de la manière suivante :

$$s = m + Vy + Ux + Dz \tag{4.3}$$

où m est la composante indépendante du locuteur, Vy la composante dépendante du locuteur, Ux la composante dépendante du canal et Dz la composante résiduelle. Ce modèle suppose cependant d'entrainer les matrices V (pour la voix), U (pour le canal) et D (pour les résidus) de valeurs propres. Ces matrices sont généralement appelées des sous-espaces latents.

Le paradigme i-vecteur a été proposé dans [Deh+11a] comme extension des modèles d'analyse factorielle et propose d'apprendre un seul sous espace latent qui contient à la fois la variabilité locuteur et session. Cet espace, appelé espace de variabilité totale, a permis d'avoir une représentation à faible dimension qui capture l'ensemble des variabilités acoustiques existant dans un enregistrement donné. Le modèle équivalent peut être représenté par :

$$m_{s,h} = m_0 + Tw_{s,h} \tag{4.4}$$

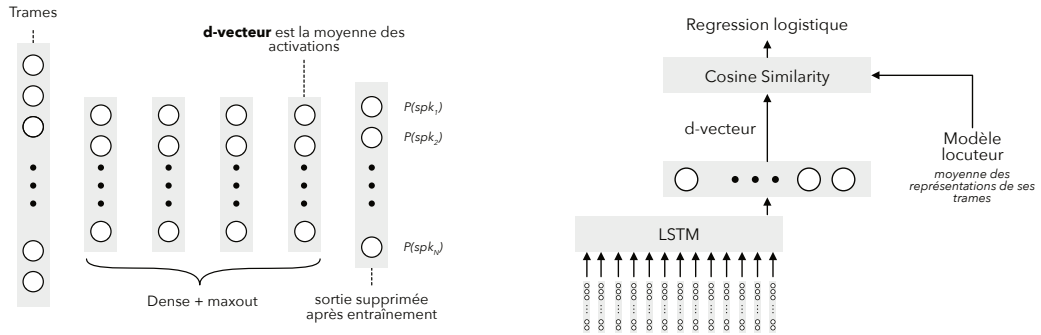
Où m_0 représente la moyenne de l'espace acoustique et correspond à l'empilement des moyennes de l'UBM, T est une matrice de projection de rang faible nommée la matrice de variabilité totale, et $w_{s,h}$ est un vecteur de facteurs de variabilité total à rang faible connu sous le nom de i-vecteur (pour vecteur d'identité). La modélisation i-vecteur peut être interprétée comme une compression appliquée sur la représentation d'une session dans l'espace des supervecteurs. Pour entrainer un tel modèle, la littérature emploie généralement une Linear Discriminent Analysis (LDA) [Kan+11].

4.4.2 Les *x*-vecteurs, une approche neuronale

Suite aux performances obtenues en reconnaissance de la parole avec les approches neuronales ([Hin+12]), certains chercheurs se sont demandés comment les adapter à la RAL. Ci-après, nous présentons les approches neuronales employées en RAL.

Les d-vecteurs

À notre connaissance, l'approche d-vecteur est apparue pour la première fois sous ce nom dans l'article [Var+14]. Dans cet article, un réseau est d'abord



(a) Architecture MLP (b) Architecture récurrente

FIGURE 4.5 – Architectures des *d*-vecteurs

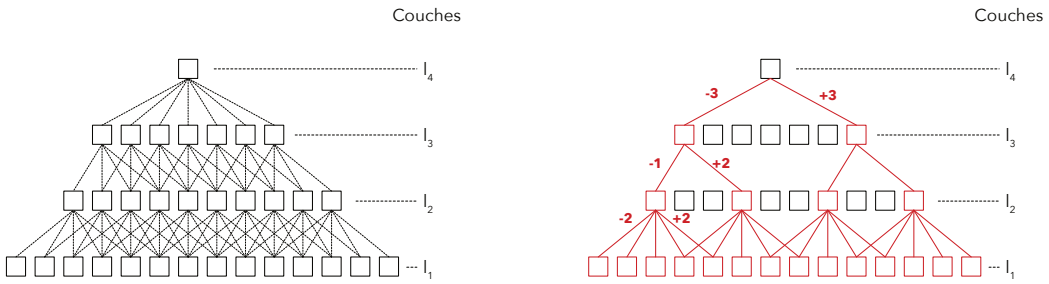
entrainé à identifier les locuteurs à partir d'une trame et de son contexte, i.e les trames qui l'entourent. Comme décrit dans la figure 4.5a, le *d*-vecteur d'une trame est extrait de la dernière couche cachée, avant la sortie du réseau. Le *d*-vecteur d'un segment audio est la somme ou la moyenne des *d*-vecteurs des trames.

Le problème du réseau précédent est qu'il ne permet pas de prendre en compte des aspects temporels à long terme puisque son inférence est calculée uniquement sur une trame ou sur un contexte de trames. [Hei+15] a donc proposé un réseau de neurones récurent comme nouvel extracteur de *d*-vecteur. Ce dernier est construit avec un LSTM. Il est évalué sur la même tâche de vérification du locuteur avec dépendance au texte. Son entraînement est réalisé de bout-en-bout puisque le réseau est entraîné directement sur la tâche de vérification du locuteur comme illustré sur la figure 4.5b.

Les TDNN

Pour modéliser les dépendances à long terme entre deux événements acoustiques, les RNN ont montré qu'ils étaient particulièrement efficaces ([SSB14]). Cependant, ces derniers sont plus difficiles à entraîner en parallélisation que les MLP, surtout lorsque des accélérations graphiques (GPU) sont utilisées. Pour pallier à ce problème, [PPK15b] propose d'utiliser une architecture TDNN pour construire un modèle acoustique employé pour résoudre des tâches de reconnaissance de la parole à large vocabulaire, en anglais *Large Vocabulary Continuous Speech Recognition task* (LVCSR). Il est à noter que ce terme est aussi référé à « Large Vocabulary Speech Recognition task » dans la littérature ([SSB14]).

Pour améliorer les performances, l'article [PPK15b] ne propose pas seulement d'employer un TDNN mais aussi de sous-échantillonner les entrées pour



(a) TDNN sans sous-échantillonnage (b) TDNN avec sous-échantillonnage

FIGURE 4.6 – Les TDNN avec et sans sous-échantillonnage

ne conserver qu'une partie du contexte. La figure 4.6b décrit le comportement d'un TDNN sur une étape temporelle t.

Les x-vecteurs

Pour modéliser les caractéristiques des locuteurs, [Sny+17] proposent de remplacer les i-vecteurs par un modèle neuronal. Ce modèle consiste à utiliser les TDNN pour modéliser des phénonèmes acoustiques à court et à long terme dans le signal. La durée des enregistrements étant variable, une couche de *Statistical Pooling* est employée dans le but de concentrer l'information dans un vecteurs de moyennes et d'écarts-types de taille fixe. Ainsi, tout ce qui est avant cette couche est calculé au niveau des trames tandis que tout ce qui s'ensuit est calculé au niveau du segment et peut servir à un système de classification standard (e.g. un MLP).

À l'instar des d-vecteurs, la dernière couche cachée du réseau sert à extraire la représentation x-vecteur d'un segment de parole.

[Sny+17] a utilisé la PLDA sur une tâche de vérification du locuteur pour montrer que cette approche neuronale, comparée aux i-vecteurs, améliorait les performances sur des phénomènes acoustiques à court terme et rivalisait sur des phénomènes à long terme. Les corpus utilisés étaient NISTS SRE 2010 et 2016. Leurs expériences ont aussi montré que les représentations x-vecteur et i-vecteur sont complémentaires et que leur fusion peut améliorer les performances des systèmes de vérification du locuteur à tous les niveaux. Il faudra cependant attendre l'article [Sny+18] pour voir apparaître le terme x-vecteur.

Notons que le terme x-vecteur peut prendre deux sens. Il peut faire référence au modèle de représentation locuteur tout comme il peut faire référence aux vecteurs de représentations extraits par inférence d'un de ces modèles après apprentissage.

Conclusion

Ce chapitre a introduit le domaine de la Reconnaissance Automatique du Locuteur. Les tâches de la RAL ont été introduites ainsi que la structure d'un système de RAL indépendant au texte. Les trois composantes ont été par la suite développées et nous avons mis l'accent sur les deux représentations locuteur qui sont très largement employées dans le domaine à savoir les i-vecteurs et les x-vecteurs.

Le prochain chapitre présente les systèmes état-de-l'art de la recommandation automatique de voix pour l'aide au casting vocal. La grande majorité de ces systèmes emploie les x-vecteurs ou les i-vecteurs comme extracteur de représentation pour nourrir leurs algorithmes ou comme *baseline* pour se comparer.

Chapitre 5

La similarité personnage, la base des systèmes de Casting Vocal Automatique (CVA)

Sommaire

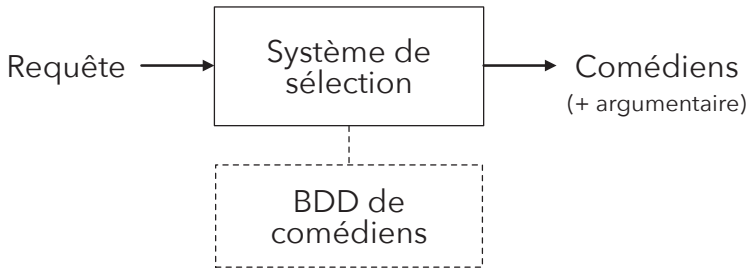


FIGURE 5.1 – Structure d'un système de Casting Vocal Automatique

5.1 Introduction au Casting Vocal Automatique

Le Casting Vocal consiste à sélectionner des comédiens de doublage. Ce processus est réalisé manuellement par des experts (e.g., DA, Voice Casters, Directeurs de Casting) qui se basent généralement sur une liste de comédiens avec qui ils ont déjà exercé et en qui ils ont confiance. Cependant, un comédien peut devenir très sollicité et augmenter le prix de ses prestations. De plus, une voix trop souvent rencontrée dans les œuvres peut mettre en danger l'immersion du public. Pour ces raisons, il est nécessaire pour l'opérateur expert de trouver de nouveaux comédiens mais le nombre de comédiens est en expansion et les tester humainement, notamment au travers d'auditions, devient trop coûteux en temps.

Les systèmes de Casting Vocal Automatique (CVA) consistent à sélectionner automatiquement une liste de comédiens parmi ceux présents dans une base de données. Leur objectif est de permettre à un utilisateur d'appliquer sa sélection à une plus grande échelle et de découvrir de nouveaux talents. Dans l'état de l'art, la manière de sélectionner des comédiens est basée sur la similarité de voix, et plus particulièrement la similarité personnage. Dans les sous-sections suivantes, nous présentons la structure des systèmes de CVA.

5.1.1 Structure d'un système de CVA

Comme décrit dans la figure 5.1, un système CVA se structure en 4 éléments, la requête, le système de sélection, la base de données de comédiens et le résultat.

La requête de l'utilisateur peut être exprimée sous différentes formes. Dans notre cas, la requête prend la forme d'un ensemble d'enregistrements d'un personnage en Version originale. Il est cependant possible d'imaginer des cas d'utilisation où l'utilisateur fournit au système des descriptions textuelles du

personnage cible ou des images le représentant.

Le système de sélection a pour objectif de comparer automatiquement la requête aux comédiens présents dans la base de données et de fournir à l'utilisateur un sous-ensemble de comédiens répondant à sa demande. Dans l'état de l'art, la sélection est assumée par un système de recherche de similarité de voix qui compare chaque voix de la base de données aux enregistrements fournis en entrée.

Les enregistrements résultant du système de sélection de voix peuvent être accompagnés d'un argumentaire présentant à l'utilisateur les critères étant intervenus dans la prise de décision. Pour les systèmes de CVA, la manière d'exprimer un argumentaire intelligible pour les experts reste à ce jour une question ouverte.

5.1.2 Simuler la perception humaine de la similarité de voix ou les décisions du Voice Caster

Dans le CVA, les voix sont sélectionnées par un système de recherche qui applique une similarité entre la voix d'origine, provenant de la requête utilisateur, et les voix présentes dans la base de données. Pour calculer un score de similarité entre deux voix, différentes approches sont possibles.

La première consiste à simuler la perception de la similarité de voix par le public. Pour la mettre en œuvre, [ORB14 ; OR16] proposent d'utiliser des données annotées manuellement pour entrainer un système à extraire des caractéristiques para-linguistiques dont l'agrégation forme ce que les auteurs nomment « la signature vocale ». L'entrainement est réalisé à partir d'annotations humaines. Pour calculer un score entre deux enregistrements, une distance cosinus est calculée entre leurs signatures vocales.

La seconde approche consiste à mimer les décisions d'un opérateur expert en s'appuyant sur les associations voix originale (e.g. version anglaise) et voix doublure (e.g. version française). [GDL18] est le premier travail à proposer de simuler les décisions d'un *voice caster* au travers d'un système de similarité personnage. Ce système est entrainé à calculer un score devant se rapprocher de 1 lorsque les deux enregistrements fournis en entrée appartiennent au même personnage et 0 dans le cas contraire. Cette approche se distingue aussi de la première par le fait qu'elle ne nécessite aucune annotation experte.

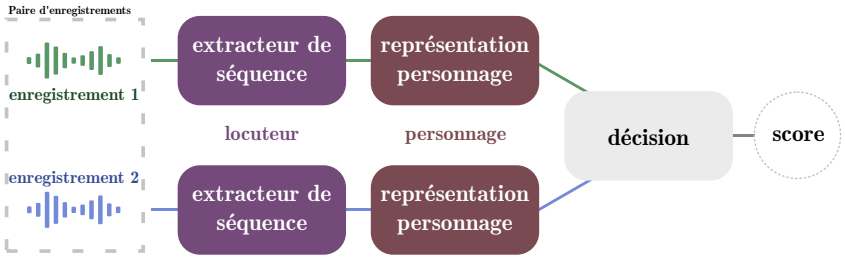


FIGURE 5.2 – Chaine de production de la similarité personnage.

5.1.3 La chaine de production de la similarité personnage

Dans ce manuscrit, nous portons notre intérêt sur la similarité personnage et présentons ci-après la chaine de calcul qui permet de calculer la similarité personnage entre deux enregistrements de voix.

Le chaine de production de la similarité personnage est présentée dans la figure 5.2. Cette dernière consiste à calculer un score de similarité entre deux enregistrements fournis en entrée, un dans la langue source, le second dans la langue cible. Ainsi, pour deux voix jouant le même personnage, le système doit fournir un score maximal qui se rapproche de 1 dans notre cas. Si les enregistrements appartiennent à des voix jouant des personnages différents, le score doit se rapprocher du score minimal, soit 0.

Pour calculer ce score, les enregistrements sont soumis à 3 modules mis bout à bout. Le premier est l'extracteur de séquence, transformant le signal, dont la durée est variable, en un vecteur de dimension fixe qui décrit ses variabilités acoustiques (e.g., modèles i-vecteur et x-vecteur). Ensuite, le module de représentation personnage, introduit dans [Gre+20], transforme cet espace vectoriel pour mettre en lumière l'information personnage présente dans l'enregistrement. Pour finir, un système de décision calcule un score de similarité qui sera soumis à une évaluation de tâche binaire.

5.2 Le *p*-vecteur : une représentation de l'information caractéristique du personnage joué

5.2.1 Définition de la représentation personnage

Pour entrainer des systèmes d'apprentissage automatique, utiliser les données brutes n'est généralement pas la solution optimale. Pour obtenir de meilleures performances, il est nécessaire de passer par des représentations intermédiaires, e.g. les x-vecteurs en reconnaissance du locuteur.

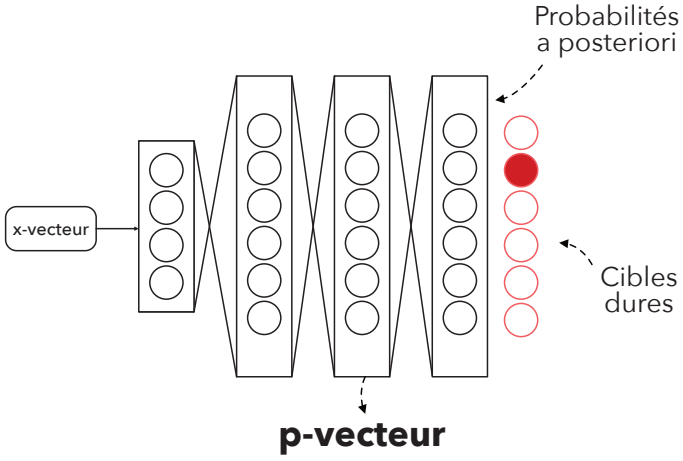


FIGURE 5.3 – Architecture neuronale des p-vecteurs

La représentation de voix est un élément clef des systèmes de similarité. [Gre+20] a proposé les p-vecteurs, une approche dédiée à la représentation des aspects personnage de la voix. Les enregistrements projetés dans l'espace des p-vecteurs doivent être proches lorsqu'ils appartiennent à un même personnage et éloignés dans le cas contraire. Cet espace doit être aussi invariant que possible au contenu linguistique, au comédien, à la langue parlée voire à la culture. Dans les sous-sections suivantes, nous présentons les architectures neuronales employées pour l'entrainement d'un espace p-vecteur.

5.2.2 Architecture neuronale des *p*-vecteurs

Pour représenter les aspects personnage d'un enregistrement vocal, une approche de projection neuronale (neural embedding) a été proposée dans [Gre+20] et illustrée dans la figure 5.3. L'approche consiste d'abord à entrainer un réseau de neurones à identifier, parmi n personnages connus (cibles dures), le personnage joué dans un enregistrement. Une fois le réseau entrainé, le p-vecteur d'un enregistrement est calculé par concaténation des valeurs de sortie de la dernière couche (avant le softmax) après avoir propagé l'enregistrement dans le réseau.

5.2.3 Homogénéisation de l'information personnage par distillation

Le volume du corpus d'entrainement joue un rôle clef dans la construction d'une représentation personnage. Cependant, les données de cinéma et de jeux vidéos sont difficiles d'accès. Les corpus d'extraits de voix actées sont la plus

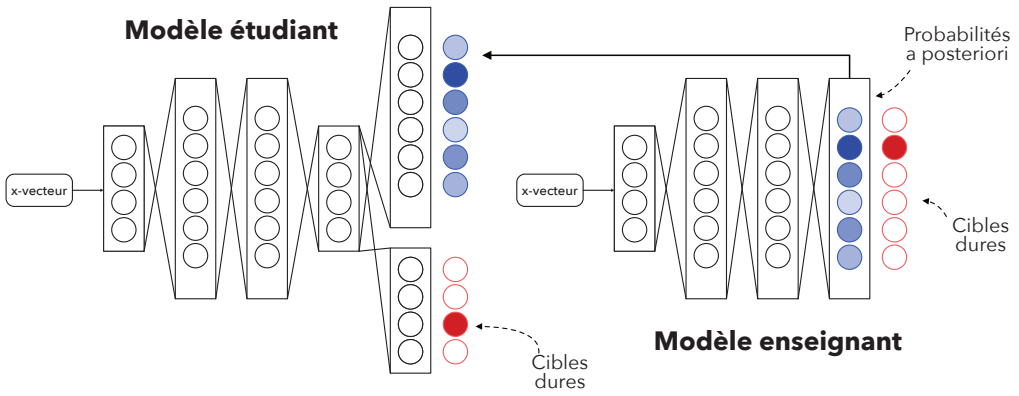


FIGURE 5.4 – L'approche par distillation de connaissance utilisée dans [Gre+20].

part du temps composés d'une quantité très limitée d'enregistrements (e.g. 7h pour Mass Effect) et les étiquettes personnage et langue ne sont parfois pas associées aux enregistrements. [Gre+20] propose d'utiliser l'approche par distillation de connaissances lors de l'apprentissage des p-vecteurs pour profiter des connaissances d'un second corpus, dont les enregistrements proviennent du jeu vidéo Skyrim. Les auteurs ont montré que l'approche p-vecteur est particulièrement intéressante lorsqu'elle est accompagnée d'un apprentissage par distillation de connaissance.

L'approche par distillation de connaissance fait intervenir deux modèles, un modèle enseignant et un modèle étudiant. Son objectif est de permettre à l'étudiant (i.e. un réseau de neurones) de résoudre des tâches compliquées en distillant la connaissance d'un modèle plus lourd, l'enseignant. Généralement, l'enseignant permet à l'étudiant d'apprendre des frontières de décision qui ne sont pas contenues dans les exemples d'entrainement.

La figure 5.4 montre la structure globale proposée par [Gre+20] pour construire un modèle p-vecteur à partir d'une approche par distillation de connaissance. Deux modèles neuronaux sont utilisés, chacun prenant en entrée des x-vecteurs d'enregistrements de voix de personnages. Le modèle enseignant, à droite, apprend à reconnaitre les cibles dures, i.e. les labels de personnage en utilisant le softmax distillé. Une fois entrainé, les cibles douces calculées lors de l'inférence des données sur l'enseignant sont utilisées, en supplément des cibles dures, pour entrainer l'étudiant illustré à gauche dans la figure. Trois valeurs λ sont employées : 0 pour ne plus prendre en compte les douces, 0,5 pour donner le même poids aux douces qu'aux dures, 1 pour ignorer les dures. Finalement, les p-vecteurs sont extraits en concaténant les valeurs des sorties de la dernière couche du modèle étudiant sous la forme d'un vecteur.

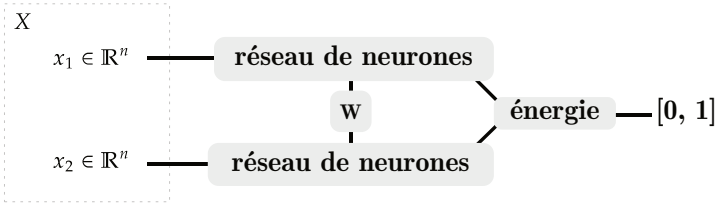


FIGURE 5.5 – Concept du réseau siamois

5.3 Évaluer la présence d'information personnage dans une représentation

5.3.1 Calculer des scores de similarité à l'aide des réseaux siamois

Évaluer la présence d'information personnage dans une représentation consiste à entrainer un système de similarité personnage et à l'évaluer sur une tâche binaire. Pour construire une mesure de similarité de voix, [Gre+19] propose d'entrainer un réseau de neurones. La figure 5.5 présente le réseau de neurones siamois utilisé à cet effet. Le réseau siamois est souvent considéré comme composé de deux réseaux de neurones partageant la même matrice de poids. Chaque vecteur d'entrée est fourni à l'un des deux réseaux. Autrement dit, les calculs appliqués aux deux entrées sont strictement les mêmes. Une fois projetées dans cet espace partagé, les deux entrées peuvent être comparées par le biais d'une mesure de distance ou de similarité nommée l'*énergie*. Dans le cas de [Gre+19], la distance euclidienne est utilisée en tant qu'énergie.

5.3.2 Une approche supervisée pour évaluer la représentation personnage

Pour évaluer de manière supervisée (tâche binaire) une mesure de similarité, il est nécessaire de transformer le score que retourne le système en une classe, à savoir la classe non-cible ou cible. La méthode employée consiste à comparer le score produit par le système à un seuil de décision. Si le score s est supérieur au seuil δ, alors la classe attribuée à la paire d'enregistrements est cible. Dans le cas contraire, la classe est non-cible. Différentes méthodes sont utilisées pour calculer le seuil. Dans notre cas, nous utilisons l'EER présentée dans la sous-section 4.3.2. Celui-ci consiste à sélectionner le seuil a posteriori de manière à équilibrer le taux de faux positif et de faux négatif sur des données.

Une autre méthode consiste à appliquer le t-test [Kim15], un test statistique

d'hypothèse, nommé le « le test student ». Appliqué à la similarité personnage, le t-test permet de comparer, intuitivement, les moyennes des scores obtenus dans les classes *cible* et *non-cible*. Il s'agit d'un test binomial où l'hypothèse nulle suppose que les moyennes des scores des deux groupes sont égales.

5.3.3 Une approche non-supervisée pour évaluer la représentation personnage

L'analyse par regroupement est une autre approche, non-supervisée, pour évaluer la représentation personnage. Elle consiste à appliquer un algorithme de regroupement (e.g. k-means [Mac67]) dans l'espace des p-vecteurs pour regrouper les enregistrements en n groupes et à vérifier si les enregistrements de chaque groupe appartiennent à un seul et même personnage, parmi n personnages.

Pour réaliser l'analyse, un algorithme regroupe d'abord les enregistrements projetés dans l'espace des p-vecteurs. Puis, les segments réunis dans un groupe sont assignés à l'étiquette du personnage le plus représenté dans ce même groupe. La F-mesure [ZZ09] peut finalement être estimée, le calcul étant réalisé sur l'hypothèse d'appartenance d'un enregistrement à une étiquette.

Il est à noter que la méthode présente cependant un inconvénient puisqu'il est possible que plus d'un groupe soit assigné à un même personnage. Toutefois, ce cas est la preuve que la représentation ne discrimine pas bien les personnages du test.

Conclusion

Dans ce chapitre, nous avons introduit le Casting Vocal Automatique (CVA) et avons présenté son état de l'art. La similarité personnage, qui est la base des systèmes de CVA, est composée de trois modules principaux : l'extracteur de séquence, le module personnage et le module de décision. L'extracteur de séquence transforme un signal audio donné en entrée en un vecteur à taille fixe. L'objectif du module personnage est de construire une représentation vectorielle dédiée à la caractérisation du personnage joué dans l'enregistrement audio fourni en entrée. Finalement, le système de décision prend deux représentations personnages en entrée et calcule un score de similarité qui s'approche de 1 lorsque les deux enregistrements jouent le même personnage et de 0 dans le cas contraire. Dans nos travaux, le système de décision est généralement utilisé sur une tâche binaire de similarité personnage dans le but d'évaluer la

qualité d'une représentation vectorielle, e.g. représentation personnage.

Dans la partie suivante, nous présentons les contributions de ce manuscrit. Dans le chapitre 6, nous présentons le cadre de travail et les problématiques. Le chapitre 7 décrit la première expérience qui consiste à mettre en lumière l'existence d'une Information Personnage Indépendante du Locuteur (IPIL). La seconde expérience présentée dans le chapitre 8 met en exergue l'influence du pré-entrainement locuteur (extracteur de séquence) sur l'apprentissage de la représentation personnage. Pour finir, la dernière expérience présentée dans le chapitre 9 propose une méthode de raffinage pour extraire du signal audio des descripteurs de voix dédiés à la caractérisation du personnage.

Troisième partie

Contributions

Chapitre 6

Le cadre de travail et les problématiques

Sommaire

6.1 Le projet ANR The Voice

Le projet The Voice est financé par l'Agence Nationale de la Recherche (ANR). Son coordinateur est Nicolas Obin, maître de conférences de l'Université Pierre et Marie Curie (Sorbonne Université) et chercheur dans l'équipe Analyse et Synthèse des Sons à l'IRCAM. L'objectif du projet The Voice est, à travers l'étude de la « palette vocale » d'acteurs professionnels, de produire des systèmes capables d'automatiser la création de voix dans un secteur de l'industrie créative, très important en termes de potentiel industriel, mais extrêmement exigeant en termes de qualité. Ce projet se veut pluri-disciplinaire

et intègre principalement dans ses collaborateurs et partenaires des informaticiens et des sociologues.

Ci-après, nous présentons les trois axes principaux du projet The Voice. Nous présentons également la composition du consortium ayant participé au projet.

6.1.1 Les trois axes de recherche du projet The Voice

La palette vocale

Le premier axe de recherche du projet The Voice s'oriente autour de la question de la «palette vocale» d'un acteur, vue sous les angles acoustique, perceptif, et réceptif, avec notamment la création d'une taxonomie de qualification de la voix en termes de « traits de personnalité » et d'« états expressifs » (ou d'« intentions de jeu » pour un acteur). La description des intentions de jeu visent à dépasser le paradigme classique « expression=émotion » pour proposer une description plus fine du jeu d'un acteur.

La recommandation de voix

Dans un second axe, le projet The Voice vise à construire des systèmes de recommandation de voix permettant d'automatiser la description d'une voix d'après la taxonomie définie précédemment, de comparer des voix d'acteurs sur la base de l'intégralité de leur palette vocale, et de modéliser les choix d'un opérateur humain - donc implicitement ses filtres acoustique, perceptif, et réceptif - pour la recommandation. Le projet souhaite aussi exploiter les retours des opérateurs humains pour enrichir les connaissances du systèmes au fur et à mesure de son utilisation, pour tendre vers la modélisation des filtres perceptifs et réceptifs des spectateurs eux-mêmes. Le travail présenté dans ce manuscrit s'inscrit dans cet axe de recherche.

La conversion de l'identité de voix

Le troisième et dernier axe du projet The Voice s'attaque au développement de technologies de conversion de l'identité de voix dans le but d'automatiser le processus de doublage. Au travers de cet objectif, le projet The Voice vise à produire des innovations scientifiques qui permettront une révolution des usages pour la production de voix, non seulement en accélérant et en améliorant les usages actuels à travers le passage à l'échelle avec des grandes bases d'acteurs pour le doublage (voix réelles), mais aussi en proposant des solutions

qui étendront les possibilités actuelles vers de nouvelles pratiques, aujourd'hui totalement inexistantes.

6.1.2 Le consortium

IRCAM

L'Institut de Recherche et Coordination Acoustique/Musique (IRCAM) est un institut de recherche spécialisé sur les sciences et technologies de la musique et du son. L'équipe Analyse et Synthèse (AS) des Sons coordonne le projet The Voice. Elle participe au développement de nombreuses technologies vocales, régulièrement utilisées en production artistique : en particulier, le casting vocal et la conversion de l'identité de la voix. Au-delà de la coordination, le rôle de l'équipe de l'IRCAM est de travailler sur l'axe de recherche de la conversion de l'identité de voix.

LIA

Le Laboratoire Informatique d'Avignon (LIA) est impliqué dans de nombreux projets de R&D nationaux, européens et internationaux (la thématique Langage du LIA concernée par ce projet a notamment participé ou coordonné plus de vingt projets ANR ou EU depuis 2008). Le LIA possède une expérience mondialement reconnue en traitement automatique de la parole et en caractérisation audio. Son rôle dans le projet The Voice est de de travailler sur l'axe de la recommandation de voix. Pour répondre à ces enjeux, trois chercheurs avignonais de l'équipe Culture et Communication (ECC) du Centre Norbert Elias (CNE, UMR 8562) viennent renforcer le LIA. Ils apportent à The Voice une spécialisation pointue sur la réception sociologique des oeuvres cinématographiques (cette association fait suite à plusieurs projets menés conjointement dont notamment l'ANR GAFES).

Dubbing Brothers

Dubbing Brothers est une société de production fondée en 1989, spécialisée dans le doublage de films et de séries. Leader sur le marché français, la majorité des films qui sortent dans les salles françaises sont doublés au sein de ce studio. C'est grâce à la confiance que leur accordent les plus grands studios français et américains (dont les géants Disney, Warner Bros, Sony Pictures ou encore Paramount) mais aussi les créateurs de séries comme HBO, AMC ou Netflix que cette société emploie maintenant plus de 200 personnes dans le monde entier, dont 150 permanents à La Plaine Saint-Denis, siège de la so-

ciété. L'effectif réel de la société est bien plus important car elle emploie aussi plusieurs centaines de comédiens par mois selon les demandes des studios pour effectuer les doublages. L'entreprise a doublé plus de 1500 films et a obtenu, il y a quelques années, une distinction rare sur le marché du doublage, à savoir le césar technique décerné à l'entreprise de doublage la plus performante et qualitative du marché. Dubbing Brothers dispose aussi d'un département informatique dont l'objectif est d'assurer la mise à disposition des plateformes de ressource et l'intégration d'un prototype industriel.

6.1.3 Le choix d'un corpus issu des jeux vidéos

L'un des objectifs du projet The Voice est de construire un corpus d'enregistrements de voix grâce à la collaboration du partenaire Dubbing Brothers. La construction d'un tel corpus s'est cependant confrontée à différentes limites. La limite principale de la récolte des données réside dans le processus d'archivage de l'entreprise. Une fois doublée, une œuvre est archivée dans des serveurs. Pour limiter la quantité d'espace qu'elle consomme, seules certaines pistes audio sont conservées. Pour des raisons évidentes de respect du RGPD (Règlement Général sur la Protection des Données), les données confidentielles telles que le nom et le prénom des comédiens sont supprimées. La bande rythmo n'est pas conservée et la Version originale est rarement disponible. Parmi les pistes conservées, les enregistrements sont généralement les résultats d'un mixage où les différents pistes (voix, ambiance etc) ont été fusionnées. Pour finir, les pistes ne sont pas segmentées. Pour une série par exemple, un enregistrement couvre l'entièreté d'un épisode. Il est alors complexe de retrouver qui parle quand.

Au vu des difficultés liées aux données, nous avons décidé dans le cadre de cette thèse de nous focaliser sur des données issues de jeux vidéos et plus particulièrement du 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.

6.2 Les contributions de ce manuscrit

Dans cette section, nous présentons les deux questions principales qui ont guidé nos expérimentations.

6.2.1 Quel lien entretient l'information personnage avec son comédien ?

Les aspects de la voix actée s'expriment au travers des informations locuteur et personnage. Dans ce manuscrit, nous posons la question suivante : Est-ce que les informations locuteur et personnage se distinguent l'une de l'autre ? Si tel est le cas, de quelle manière ?

Puisque la voix du personnage s'appuie sur celle du comédien qui le joue, nous supposons que les variabilités locuteur et personnage sont en partie liées et proposons une classification des informations véhiculées par la voix actée à un niveau de granularité moins élevé. Ainsi, nous listons les trois informations suivantes :

— **L'information locuteur indépendante du personnage :** fait référence à l'ensemble des variabilités propres au locuteur, dont il n'a pas nécessairement la maîtrise et qui apparaissent régulièrement dans les voix des personnages qu'il joue. Cette information ne joue pas un rôle clef dans la construction de l'identité du personnage.
— **L'information locuteur et personnage :** fait référence à l'ensemble des variabilités apportées par un jeu d'acteur propre au comédien ou par ses aspects vocaux qui appuient naturellement le caractère du personnage joué. Lorsqu'un personnage est joué par un nouveau comédien, ces aspects ne peuvent être conservés que si la voix du nouveau comédien est similaire à celle du précédent.
— **L'Information Personnage Indépendante du Locuteur (IPIL) :** est l'ensemble des variabilités apportées par le jeu de l'acteur pour appuyer des traits de caractère d'un personnage. Cette information est dite indépendante du locuteur car, lorsqu'un personnage est joué par un nouveau comédien, elle peut être conservée dans son entièreté.

Le chapitre 7 propose un protocole expérimental pour vérifier la présence d'IPIL (sous-ensemble C) dans des enregistrements de voix actées. Dans ce but, nous proposons d'entrainer un système de similarité de voix où l'information personnage des données a été neutralisée en modifiant les associations acteur d'origine et comédien de doublage. En comparant les résultats obtenus avec ceux d'un système entrainé sur des données originales – où les associations ne sont pas modifiées – nous mettons en valeur l'IPIL.

Le chapitre 8 propose de montrer l'influence du pré-entrainement locuteur sur la construction de représentations personnage. Dans ce but, nous entrainons des représentations personnage en utilisant différents configurations où

l'influence de l'extracteur de séquence sur l'apprentissage personnage est plus ou moins limité.

6.2.2 Comment identifier et nommer les descripteurs de la voix d'un personnage ?

Le système de similarité de voix de [ORB14 ; OR16] a été intégré dans un moteur de recommandation implémenté chez l'entreprise de doublage Dubbing Brothers avec l'aide de l'IRCAM. Ce système a montré sa capacité à mesurer une similarité humaine perceptive entre deux voix de personnage. Pourtant, un Directeur Artistique l'a essayé et, selon ses retours, la fréquence fondamentale semble être le seul critère employé à la sélection des comédiens de doublage à recommander. Le système serait donc biaisé. Il est aussi possible que des critères plus difficiles à percevoir par l'utilisateur interviennent dans les choix du système. En supposant cela, il semble nécessaire d'ajouter des argumentaires aux recommandations qui permettront à l'utilisateur de mieux percevoir les critères étant intervenus durant la décision du système.

Les argumentaires des systèmes de recommandations doivent être décrits dans un vocabulaire compris par des opérateurs experts. Cependant, il n'existe aucune nomenclature acceptée et standardisée par la communauté professionnelle pour décrire la voix actée. Les termes de couleur de voix existent mais la définition, par exemple, d'une voix « blanche » peut varier d'un individu à l'autre. Travaillant de concert, les acteurs et DA s'adaptent alors les uns aux autres pour communiquer au travers d'un vocabulaire qui semble évoluer avec le temps. Le manque de définition claire des différents descripteurs de la voix actée empêche la mise en place de protocoles d'annotation. Elle rend aussi difficile la mise en place d'argumentaires dans les systèmes de recommandations compréhensibles par ces experts.

Nous proposons dans ce manuscrit de participer à l'élaboration de la nomenclature de la voix actée et tentons d'identifier les différents descripteurs intervenant dans la caractérisation vocale du personnage. Dans ce but, le chapitre 9 propose d'appliquer des algorithmes de regroupement sur des représentations de voix dédiées à la caractérisation du personnage joué.

Chapitre 7

Mesure de présence d'information personnage indépendante du locuteur modélisée dans une représentation de voix

Sommaire

Comme présenté dans le chapitre 5, [Gre+19] a proposé d'utiliser un système de similarité de voix pour évaluer l'efficacité d'une représentation personnage. Ce système détecte si deux extraits de parole dans une langue source et cible appartiennent, ou non, au même personnage. Ces extraits sont toujours parlés dans deux langues différentes. Pour réaliser cette tâche, un classifieur binaire basé sur un réseau de neurones siamois ([CHL05 ; KK15]) est entraîné en amont des p-vecteurs. Les résultats expérimentaux ont montré l'habilité de la représentation p-vecteur à associer deux extraits de voix liés au même personnage, même lorsqu'un scénario particulièrement difficile est employé, où le personnage et les deux locuteurs d'une paire d'enregistrements donnée sont complètement inconnus pendant la phase d'entraînement. Cependant, il existe plusieurs biais potentiels dans cette expérience, comme le fait qu'un personnage donné est représenté par une unique paire d'acteurs, la longueur des extraits, leur contenu linguistique ou l'influence de l'information locuteur ([Gre+19]). Bien que le protocole proposé prenne en compte la plupart de ces biais, le dernier listé reste discutable car il y a toujours un enchevêtrement possible entre la dimension du locuteur et les aspects du personnage.

Notre objectif est de questionner la nature de l'information encodée par les systèmes de similarité personnage. En quoi se différencient les variabilités personnage et locuteur ? Existe-t-il une Information Personnage Indépendante du Locuteur (IPIL), « indépendante » dans le sens où elle permet de discriminer le personnage et ne donne pas, ou peu, d'information sur le locuteur (comédien). Pour répondre à ces questions, nous proposons dans ce chapitre une approche nouvelle permettant d'évaluer la présence d'IPIL. Cette évaluation aide à vérifier si les systèmes n'apprennent pas seulement à associer l'identité du locuteur mais aussi s'ils fondent leurs décisions sur des aspects spécifiques au personnage. L'idée principale est de commencer avec une solution proche de celle proposée dans [Gre+19] et [Gre+20]. Ensuite, nous construisons un modèle alternatif où l'information du personnage est neutralisée durant la phase d'entraînement en échangeant les étiquettes personnages des enregistrements de voix. Cet échange est réalisé au niveau acteur : tous les enregistrements prononcés par un acteur sont maintenant associés à tort à une étiquette personnage, choisie aléatoirement. Il convient de rappeler que, en raison du contexte de doublage vocal de ce travail, cela revient à rompre le lien entre un acteur original et le doubleur qui lui est associé. En comparant les performances du

système original à celles du second, nous nous attendons à observer une perte proportionnelle à la part d'information sur les personnages neutralisée (IPIL).

Ce chapitre est organisé de la manière suivante. Dans la section 7.1, nous présentons la représentation p-vector utilisée et comment nous l'évaluons. Dans la section 7.2, nous détaillons la partie centrale de ce travail, à savoir le protocole pour estimer la quantité relative d'information personnage. Dans la section 7.3, nous présentons les différentes expériences que nous avons réalisées et les résultats que nous avons obtenus. Pour finir, nous discutons de ces résultats dans la section 7.4. Ce travail a été publié dans l'article [QDB21]. Pour des raisons de reproductibilité, les scripts et les modèles sont disponibles sur GitHub [1].

7.1 Extraction et évaluation d'une représentation personnage

Dans cette section, nous décrivons les différentes briques – représentation personnage et similarité de voix – de notre système de similarité personnage basé sur la chaine de production présentée dans la sous-section 5.1.3. Une évaluation de ce système est aussi proposée en mode fermé où les locuteurs de test (et donc les personnages) sont connus durant l'entraînement.

7.1.1 Représentation orientée personnage

Cette section décrit la représentation de voix orientée personnage, dénommée p-vecteur, introduite dans [Gre+20]. Ces p-vecteurs sont construits depuis une représentation de signal de parole et sont destinés à mettre en exergue l'information personnage d'un enregistrement de voix donné.

Chaque enregistrement $r \in train$ est associé avec un personnage joué. Un perceptron multi-couche est entraîné à reconnaitre le personnage dans un espace fermé selon l'enregistrement donné, quelque soit la langue dans laquelle il est joué. En entrée du système d'extraction personnage, une représentation i-vecteur [Deh+11a] de 400 dimensions obtenue depuis le signal est fournie. Le système nous calcule, pour chaque classe, la probabilité que l'enregistrement appartienne à la classe (i.e., l'étiquette personnage). Une fois le réseau entraîné, [Gre+20] propose d'utiliser la dernière couche comme représentation, avant le softmax, que nous nommons p-vecteur.

1. https://blog.barracuda.com/2022/07/12/report-the-state-of-industrial-security-in-2022/

Le réseau de neurones est composé de quatre couches *Dense*, toutes avec 256 neurones, accompagnés par une fonction d'activation tangente hyperbolique et un dropout de 0.25, excepté pour la quatrième qui est utilisée comme représentation (embedding) et qui a seulement 64 neurones. Une dernière couche de softmax à la fin du réseau de neurones est ajoutée. L'algorithme que nous utilisons pour entrainer un réseau est l'*adadelta* avec une fonction de coût *categorical crossentropy*. Pour éviter l'overfitting, nous appliquons un arrêt prématuré avec un delta minimum de 0,1 et une patience de 10 époques pour l'algorithme d'entraînement.

7.1.2 Modèle de similarité de voix

Pour évaluer une représentation personnage, il est nécessaire, comme dit précédemment, de construire un système de similarité personnage. La tâche consiste à décider si deux enregistrements, un dans une langue source et le second dans une langue cible, appartiennent au même personnage ou non. Nous calculons, pour chaque paire de voix X, le score $H_f(X)$ et le comparons à un seuil défini avec l'EER a posteriori. Le module de décision emploie le réseau de neurones siamois présenté dans [Gre+19].

Évaluation des performances et intervalle de confiance

Le système de similarité de voix (i.e., module de décision) est utilisé dans ce travail pour vérifier l'efficacité des modules de représentation de voix orientée personnage (i.e., représentation personnage). Les performances sont estimées par un taux de réussite binaire du système de similarité de voix, qui est le ratio entre le nombre de paires correctement classifiées sur le nombre de total de paires classifiées.

Nous utilisons aussi un *test de proportion* pour mesurer la signification statistique des différents taux de réussite. Cette méthode prend deux proportions p_1 et p_2 et évalue l'hypothèse H_1 disant que les proportions sont équivalentes. Un intervalle de confiance est calculé en utilisant p_2 et l'hypothèse est confirmée si p_1 est dans cet intervalle. Autrement, l'hypothèse est rejetée. Nous comparons les taux de réussite de a_1 et a_2 de deux systèmes donnés en appliquant le test avec un niveau de signification de 5% et avec $p_1 = a_1$ et $p_2 = a_2$.

7.1.3 Description du corpus

Le corpus principal est composé des enregistrements de voix provenant du jeu vidéo *Mass Effect 3*. À l'origine publié en anglais, le jeu a ensuite été

FIGURE 7.1 – Personnages de Mass Effect 3 présents dans les corpus d'entrainement, de validation et de test.

traduit. Dans nos expériences, nous utilisons les versions anglaise et française des séquences audios, représentant environ 7,5 heures de parole dans chaque langue. Les segments (ou enregistrements) sont de 3,5 secondes en moyenne. Un personnage est défini par un unique couple Français-Anglais, composé de deux locuteurs distincts. Pour éviter des biais en terme de locuteur, nous ne prenons en compte qu'un sous-ensemble de données où nous sommes sûrs qu'aucun des acteurs ne joue plus qu'un personnage. Un seul segment audio correspond à une unique prise de parole provenant d'un acteur dans une langue particulière. Nous avons appliqué un filtre qui garde seulement les enregistrements pour lesquels la durée est supérieure à 1 seconde. Finalement, nous gardons seulement 16 personnages pour lesquels nous avons le plus d'enregistrements.

Contrairement à l'article [Gre+19] qui propose d'appliquer un protocole en 4-fold, nous avons pris la décision dans ces expériences de garder les 16 personnages, et leurs 32 locuteurs correspondant, pour à la fois les phases d'entrainement et de test. Nous séparons le corpus en trois sous-ensembles : l'entrainement (*train*), la validation (*val*) et le *test* en utilisant une règle de 80/10/10. Tous ces sous-ensembles sont composés de différents enregistrements provenant des mêmes 16 personnages (voir figure 7.1) et 32 locuteurs. Pour construire respectivement les sous-ensembles *train*, *val* and *test*, nous sélectionnons aléatoirement pour chaque personnage 144, 18 et 18 enregistrements, tout en équilibrant le nombre d'enregistrements français et anglais. Nous avons ensuite respectivement un total de 2 304, 288 et 288 enregistrements.

Pour chaque sous-ensemble, nous construisons des paires d'enregistrements où le premier élément est un segment de voix appartenant à l'acteur dans une langue source (anglais), et le second à un enregistrement prononcé par un autre comédien, le doubleur, dans la langue cible (français). Nous associons la classe *cible* aux paires de voix provenant des mêmes personnages, et *non-cible* dans le cas contraire. Les paires sont construites avec des segments sélectionnés aléatoirement tout en équilibrant le nombre de *cibles* et *non-cibles*. Ce processus d'appariement est nommé *données d'origine*. Nous avons, pour les paires

TABLE 7.1 – Résumé du nombre d'enregistrements et de paires d'enregistrements par sous-ensemble.

sous-ensemble	enregistrements	paires
train	2 304	165 888
val	288	2 592
test	288	2 592

respectivement construites depuis le *train*, *val* et *test*, 165 888, 2 592 et 2 592 paires.

Le tableau 7.1 résume le nombre d'enregistrements et de paires d'enregistrements présents dans les sous-ensembles *train*, *val* et *test*.

7.1.4 Performances de la représentation *p*-vector

Le tableau 7.2 montre les performances des systèmes de similarité personnage (*prot 2*) construits avec des *p*-vectors. Les performances des systèmes de comparaison de voix (*prot 1*) basés seulement sur des *i*-vecteurs (dans cette représentation, aucune information spécifique concernant le personnage joué n'est utilisée) sont données dans un but de comparaison.

Aussi, nous modélisons théoriquement un système dont les décisions sont aléatoires sur une tâche binaire. Son taux de réussite théorique est de 50% et son intervalle de confiance à 95% est [0,48; 0,52]. Nous le nommons le « système aléatoire ».

Une large différence de 7 points en taux de réussite (87% pour les *p*-vecteurs contre 80% pour les *i*-vecteurs) est observée dans le tableau 7.2. Au regard de l'intervalle de confiance, cette différence est significative. Cela confirme les résultats observés dans [Gre+20] (en utilisant un protocole différent) où il a été découvert que la représentation *p*-vecteur semble encoder des information spécifiques du personnage joué.

7.2 Estimation de la quantité d'information personnage dans la représentation *p*-vecteur

La précédente section prouve empiriquement que la représentation *p*-vecteur améliore les performances sur une tâche de similarité personnage. Néanmoins, au vu des données, il est possible que la stratégie du système soit d'associer des identités de locuteur, l'empêchant ainsi de modéliser une partie de l'information personnage, notamment indépendante du locuteur. Dans ce cas, les

TABLE 7.2 – Performances des i-vecteurs et des p-vecteurs sur les données d'origine ; « aléatoire » est la performance théorique d'un système aléatoire. Les limites de l'intervalle de confiance à 95% sont indiquées entre crochets.

	extracteur de séquence	couche personnage (p-v)	performance
aléatoire	×	×	0,50 [0,48 ; 0,52]
prot 1	i-v	×	0,80 [0,78 ; 0,82]
prot 2	i-v	origine	0,87 [0,86 ; 0,88]

systèmes seraient potentiellement biaisés et confondraient les informations dédiées à la caractérisation du locuteur de celles dédiées à la caractérisation du personnage. Dans le but de vérifier que cette amélioration observée en utilisant les p-vecteurs ne provienne pas seulement de son habileté à associer des identités de locuteurs, nous proposons comme principale contribution de ce chapître d'entraîner des systèmes de similarité de voix actée avec des associations erronées où l'information personnage est supposée être neutralisée.

7.2.1 Protocole aléatoire d'association

Le protocole d'association aléatoire que nous proposons consiste à entraîner un réseau de neurones sur des données intentionnellement erronées. Le modèle ainsi généré est nommé le *modèle modifié*.

Comme présenté dans la partie gauche de la figure 7.2, nous notons ρ_i le personnage i défini par une association de deux locuteurs, son acteur d'origine α_i et le comédien qui l'a doublé δ_i. Pour neutraliser l'information personnage, nous créons de nouvelles étiquettes ρ' où le lien personnage entre deux locuteurs est rompu. Ainsi, l'étiquette ρ'_i est associée à son acteur d'origine α_i et à un doubleur δ_j qui joue un personnage différent, i.e. $j \neq i$. Pour éviter des biais de genre trop évidents, deux locuteurs associés à la même étiquette ρ' partagent le même genre, à savoir « homme » ou « femme ». Les étiquettes ρ' remplacent les étiquettes ρ lors de l'entrainement du modèle modifié.

Puisqu'un personnage est représenté par une unique paire de locuteurs, changer les associations acteur/comédien revient à donner aléatoirement des étiquettes personnages à des fichiers, en respectant les contraintes précédemment décrites. Les locuteurs sont donc toujours associés par des paires (un en anglais et un en français) mais ils n'appartiennent plus au même personnage au sein d'une paire donnée.

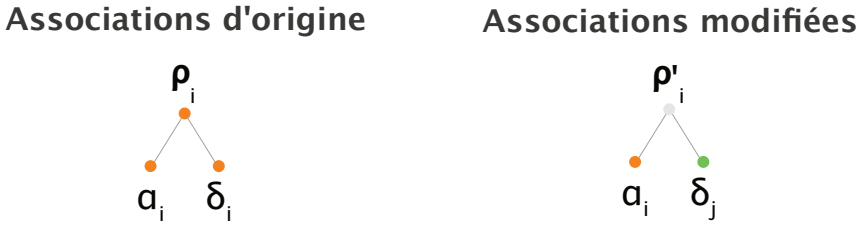


FIGURE 7.2 – Définition des étiquettes ρ et ρ'.

Les performances des systèmes modifiés devraient montrer le pouvoir locuteur des p-vecteurs. L'observation d'une différence de performance entre les modèles d'origine et modifié devrait en revanche indiquer la part du personnage indépendante du locuteur encodée par les p-vecteurs.

7.2.2 Sous-ensemble d'associations aléatoires

Comme expliqué dans la section 7.2.1, le protocole d'association aléatoire consiste à intentionnellement échanger le doubleur associé à l'acteur d'origine. Une fois fait, de nouvelles paires de voix sont générées et associées aux étiquettes *cibles* et *non-cibles* en suivant les mêmes étapes présentées dans la sous-section précédente. Cet ensemble de données généré grâce aux associations aléatoires est nommé *modifié*.

En utilisant ce nouveau protocole modifié, deux modules devraient être impactés puisqu'ils utilisent les étiquettes personnages d'origine ou modifiées. Ces modules sont la représentation p-vecteur (*représentation personnage*) et le système de similarité de voix siamois (*decision*). Nous entrainons donc un extracteur de p-vecteur, nommé *modifié*, avec les nouvelles étiquettes personnages aléatoires pour les voix de doublage. De la même manière, nous entrainons aussi une version de notre système de similarité de voix en utilisant les étiquettes modifiées. Bien sûr, quand le système de similarité de voix est entrainé en utilisant ces étiquettes modifiées, les mêmes appariements (modifiés) de locuteurs sont utilisés pour mesurer les performances dans les données de test.

7.3 Expériences et résultats

TABLE 7.3 – Performance (taux de réussite) des représentations *i*-vecteur et *p*-vecteur sur les données modifiées. Les limites de l'intervalle de confiance à 95% sont données entre crochets. Les lignes "prot 1" et "prot 2" sont reportées depuis le tableau 7.2 pour améliorer la lisibilité du lecteur.

	extracteur de séquence	paires		performance
		module personnage (*p*-v)	module de décision	
prot 1	*i*-v	×	origine	0,80 [0,78 ; 0,82]
prot 2	*i*-v	origine	origine	0,87 [0,86 ; 0,88]
prot 3	*i*-v	×	modifié	0,80 [0,78 ; 0,82]
prot 4	*i*-v	modifié	modifié	0,84 [0,83 ; 0,85]
prot 5	*i*-v	origine	modifié	0,75 [0,73 ; 0,77]

7.3.1 Protocole pour la mise en lumière de l'Information Personnage Indépendante du Locuteur (IPIL)

Notre première expérience consiste à mesurer la présence, ou non, d'IPIL dans une représentation de voix. Le protocole modifié est un moyen de neutraliser l'information personnage. Par conséquent, la différence absolue entre le score d'un système entrainé sur les données d'origine et d'un deuxième système entrainé sur les données modifiées constitue notre indice pour mesurer la présence d'IPIL. Dans cette optique, nous entraînons un extracteur de séquence *i*-vecteur et construisons un espace *p*-vecteur par-dessus en utilisant des données d'origine ou modifiées. La représentation personnage est ensuite évaluée avec le réseau de neurones siamois entrainé sur la similarité personnage comme expliqué précédemment. Le tableau 7.3 résume les performances des systèmes. Pour valider que les taux de réussite de deux systèmes sont significativement différents, les intervalles de confiance sont écrits en dessous des scores de performance. Dans ce tableau, *prot 3* et *4* correspondent respectivement aux versions modifiées de *prot 1* et *2*. Nous proposons aussi une version mixée nommée *prot 5* où les *p*-vecteurs (module personnage) sont entraînés en utilisant les données d'origine et sont évalués sur des données modifiées (module de décision) dans le but de mesurer la présence d'information locuteur dans le sous-espace vectoriel.

TABLE 7.4 – Performance (taux de réussite) des représentations x-vecteur et p-vecteur sur les données modifiées. Les limites des intervalles de confiance à 95% sont données entre crochets.

	extracteur de séquence	paires		performance
		module personnage (p-v)	module de décision	
prot 6	x-v	×	origine	0,85 [0,83 ; 0.87]
prot 7	x-v	origine	origine	0,90 [0,89 ; 0,91]
prot 8	x-v	×	modifié	0,76 [0,74 ; 0,78]
prot 9	x-v	modifié	modifié	0,90 [0,89 ; 0,91]
prot 10	x-v	origine	modifié	0,77 [0,75 ; 0,79]

7.3.2 Comparaison avec un extracteur de séquences à base de réseaux de neurones

Puisque nous basons nos systèmes sur des représentations locuteurs i-vecteurs, nous souhaitons aussi comparer ces résultats avec un extracteur de séquence neuronal. Dans ce but, nous construisons un extracteur x-vecteur avec l'outil Kaldi ([Pov+11b]) en utilisant le corpus Voxceleb ([CNZ18]). Nous l'utilisons comme extracteur de séquence à la place des i-vecteurs. Ensuite, nous entraînons des p-vecteurs et les évaluons avec des réseaux de neurones siamois en suivant exactement les mêmes protocoles d'origine et modifié employés pour les i-vecteurs. Le tableau 7.4 présente les résultats : les protocoles de *prot 6* jusqu'à *prot 10* correspondent respectivement aux protocoles de *prot 1* jusqu'à *prot 5* où la seule différence est le remplacement des i-vecteurs par l'approche x-vecteur comme extracteur de séquence.

7.3.3 Variabilité des associations de locuteur (IPIL)

Lorsque nous modifions les données, l'information personnage est supposée neutralisée. Cependant, quel est l'impact du choix aléatoire des associations acteur/comédien sur le modèle modifié ? Est-ce que des modèles appris sur des associations aléatoires différentes encoderont la même information locuteur quel que soit les associations aléatoires générées, tant que l'acteur d'origine n'est pas associé à son doubleur et que le biais du genre est évité ? Pour répondre à ces questions, nous proposons de générer de nouvelles étiquettes erronées où les associations acteur/comédien sont différentes de ρ et ρ'. Ces

TABLE 7.5 – Performance (taux de réussite) des représentations x-vecteur et p-vecteur sur les données modifiées avec les étiquettes ρ_2. Les limites des intervalles de confiance à 95% sont données entre crochets.

	extracteur de séquence	paires		performance
		module personnage (p-v)	module de décision	
prot 6	x-v	×	origine	0,85 [0,83 ; 0.87]
prot 7	x-v	origine	origine	0,90 [0,89 ; 0,91]
prot 8_2	x-v	×	modifié 2	0,81 [0,79 ; 0,83]
prot 9_2	x-v	modifié 2	modifié 2	0,88 [0,86 ; 0,90]
prot 10_2	x-v	origine	modifié 2	0,71 [0,69 ; 0,73]

étiquettes sont nommées ρ_2'.

Le tableau 7.5 présente l'ensemble des résultats obtenus avec le protocole *modifié 2*. Pour faciliter la compréhension, les noms des protocoles sont identiquement étiquetés. Seulement, un indice 2 est ajouté lorsque les données modifiées utilisées emploient les étiquettes erronées ρ_2'. Le protocole n'a été appliqué que sur les x-vecteurs.

7.4 Discussion

La première partie de notre analyse se focalise sur les systèmes basés sur les i-vecteurs, listés dans le tableau 7.3. Puisque *prot 3* donne les mêmes résultats que *prot 1* (0,80), nous pouvons conclure que neutraliser l'information personnage n'a aucun effet sur le taux de réussite des systèmes. Cela confirme que l'information encodée par les i-vecteurs est principalement présentée depuis un angle locuteur, biaisant probablement le réseau siamois. Ce dernier a éventuellement des difficultés à modéliser l'Information Personnage Indépendante du Locuteur (IPIL). Cependant, puisque dans ce cas les performances du modèle modifié ne sont pas statistiquement supérieures à celles du modèle d'origine, il est impossible d'affirmer si l'IPIL est modélisée ou non.

Nous analysons ensuite la contribution des p-vecteurs sur l'information encodée. Comme nous le savons, la construction des p-vecteurs par dessus les i-vecteurs met en lumière l'information personnage. Cela a été démontré par le fait que le système p-vecteur entraîné sur les données d'origine, à la fois

pour les couches personnage et décision, surpassait les performances de celui entraîné sur les i-vecteurs ; avec respectivement *prot 2* (0,87) et *prot 1* (0,80). Nous pouvons aussi observer dans le tableau 7.3 qu'appliquer les p-vecteurs permet d'encoder de l'IPIL puisque nous observons une différence significative entre les performances de prot 2 (0,87) et prot 4 (0,84).

Le tableau 7.3 montre aussi le taux de réussite obtenu sur *prot 5* (0.75) où la dimension personnage est neutralisée lors de l'apprentissage du module de décision. Bien que les p-vecteurs apprennent à associer des locuteurs avec les associations d'origine, les réseaux siamois entrainés avec les données modifiées arrivent à trouver l'information qui permet d'associer des locuteurs. En conséquence, nous supposons que l'information locuteur est présente dans les p-vecteurs.

Puisque les réseaux de neurones sont l'état de l'art de la littérature en représentation locuteur, nous avons comparé l'usage des i-vecteurs avec un extracteur de séquence neuronal, les x-vecteurs. Dans le tableau 7.4, nous pouvons observer que tous les scores sont meilleurs que ceux obtenus avec les i-vecteurs. Cependant, nous observons des tendances différentes.

Dans le cas des x-vecteurs, l'apprentissage des p-vecteurs ne semble pas mettre en valeur l'IPIL puisque nous n'observons pas de différences entre *prot 7* et *prot 9* (0,90). Pourtant, l'usage des x-vecteurs, sans module de représentation personnage, semble plus enclin à encoder de l'information dédiée à la caractérisation du personnage puisque nous observons une différence absolue et significative de 9% entre *prot 6* (0,85) et *prot 8* (0,76). Ces résultats sont différents de ceux observés avec les i-vecteurs et s'expliquent certainement par la différence de nature entre les deux approches, les i-vecteurs étant une approche générative et les x-vecteurs une approche par tâche de classification.

En créant de nouvelles associations erronées, étiquetées ρ_2', et en les appliquant aux systèmes basés sur les x-vecteurs, nous observons que le changement d'association introduit une variabilité dans les résultats des systèmes de similarité. Nous remarquons notamment dans le tableau 7.5 une différence significative entre prot 7 (0,90) et prot 9_2 (0,88) suggérant que le système a encodé de l'IPIL. Cette différence n'était pas observée avec le prot 9 (0,90).

Pour finir, puisque nous observons qu'aucun système entraîné sur des données modifiées donne de meilleurs performances que ceux entrainés sur les données d'origine, nous supposons que nous atteignons l'objectif de neutraliser l'information personnage tel que prévu. Nous nommons l'information déduite des associations comédien/doubleur l'IPIL et nos expériences montrent que

cette information peut être encodée dans nos représentations de voix et prise en compte par les systèmes de similarité.

Conclusion

Dans ce chapitre, nous proposons de mettre en lumière la part indépendante du locuteur en rapport avec le personnage joué dans les voix actées. Nous nous appuyons sur les p-vecteurs, une approche d'apprentissage de représentation dédiée aux informations du personnage dans les voix actées. Pour les évaluer, nous avons utilisé des réseaux siamois capables de décider si deux voix sont reliées au même personnage ou non. Nous avons d'abord parcouru les expériences issues de [GDL18 ; Gre+19] qui montrent que les p-vecteurs aident à résoudre cette tâche. Ensuite, nous nous sommes focalisés sur le principal objectif de ce chapitre, à savoir évaluer si les p-vecteurs capturent réellement l'information du personnage joué et ne mémorisent pas seulement les voix des locuteurs. Pour cela, nous avons conçu une configuration spécifique capable de neutraliser l'information personnage dans les p-vecteurs tout en conservant intact sa capacité à mémoriser l'information relative au locuteur.

Nos expériences ont montré que cette configuration neutralise l'information personnage et donne un bon cadre de travail pour analyser le biais du locuteur sur un système basé sur le personnage. Grâce à cette méthode, nous avons aussi montré que les p-vecteurs peuvent mettre en exergue la partie personnage spécifique, ou non, à l'identité du locuteur. Cependant, nous avons aussi mis en avant que les performances ne sont pas des indicateurs de la qualité de la représentation. En effet, le système qui atteint les meilleurs performances n'encode pas, ou peu, d'IPIL, ce qui nous amène à conclure que le système se contente d'associer les identités des locuteurs alors qu'une autre source d'information peut aider à caractériser le personnage.

Sur la base de ces résultats, nous cherchons dans le prochain chapitre à montrer l'influence qu'a l'extracteur de séquence sur la construction de la représentation personnage, notamment sur l'IPIL qu'elle encode.

Chapitre 8

Mise en exergue de l'influence du pré-entraînement locuteur sur l'information personnage

Sommaire

Dans le chapitre précédent, nous avons proposé un protocole pour mettre en exergue la partie indépendante du locuteur de l'information personnage (IPIL). Les expériences que nous avons menées ont montré qu'une grande partie de l'information modélisée par les systèmes de similarité personnage semble relative à la dimension locuteur mais qu'une faible partie en est indépendante.

Cependant, les extracteurs de séquences utilisés dans ces expériences proviennent du domaine de la reconnaissance du locuteur. Ces extracteurs requièrent un très large volume de données d'entraînement issu de milliers de locuteurs et sont optimisés pour la tâche d'identification du locuteur. Leur usage dans une représentation personnage semble obligatoire car les corpus annotés disponibles dans le domaine du doublage sont de petite taille. Ces extracteurs de séquence peuvent être considérés comme un pré-entraînement ([HLM19 ; PPK15a ; Sny+18]), dans le sens où un transfert d'apprentissage est réalisé ([HLM19 ; HVD15 ; LP+16 ; PY10 ; VI15]).

Bien que l'usage d'un pré-entraînement locuteur aide clairement à la construction d'une représentation de voix orientée personnage, cela peut aussi créer des biais dans le système. Si certains travaux ([RF16 ; WQY17 ; Raj+19]) se sont déjà intéressés à l'information encodée dans ces représentations, il semble cependant légitime de se demander si ce pré-entraînement ne guide pas les p-vecteurs trop fortement vers l'information locuteur. Ce chapitre est dédié à cette question et a pour objectif de vérifier les deux hypothèses suivantes :

- Ⓐ Le plus un extracteur de séquence est dédié à la reconnaissance du locuteur, le plus il intègre des informations de haut niveau spécifiques au locuteur et risque de perdre de l'information concernant le personnage lui-même. Donc, prendre la représentation à un niveau plus bas dans l'extracteur de séquence devrait aider à capturer des informations moins spécifiques au locuteur et permettre de mieux caractériser la dimension personnage.
- Ⓑ Adapter des parties du modèle d'extraction de séquence à la tâche personnage devrait aider à construire une représentation personnage.

La section 8.1 présente le cadre de travail de la représentation personnage des voix. Le corpus et les détails du réseau de neurones sont présentés dans la section 8.2. Les résultats de notre expérience sont présentés pour la première et la seconde hypothèse dans les sections 8.3 et 8.4 respectivement. Pour des soucis de reproductibilité, les scripts et les modèles sont disponibles sur GitHub [1]. Enfin, nous concluons en présentant les conclusions et des orienta-

1. github.com/LIAvignon/specom2021-influence-of-speaker-pre-training-on-character-

tions possibles pour les travaux futurs. Ce travail a été publié dans l'article [Qui+21].

8.1 Représentation personnage de voix basée sur des réseaux de neurones

Dans cette section, nous décrivons les différentes briques – représentation personnage et similarité de voix – de notre système de similarité personnage basé sur la chaine de production présentée dans la sous-section 5.1.3. La représentation personnage et le module de décision sont présents dans les sections 8.1.1 et 8.1.2 respectivement. Enfin, la section 8.1.3 se focalise sur le module d'extraction de séquence et ses liens avec le système de représentation de voix.

8.1.1 Représentation orientée personnage

Les p-vecteurs ont d'abord été introduits dans [Gre+20]. Ils sont construits depuis une représentation du signal de parole et sont destinés à mettre en exergue l'information relative au personnage dans un enregistrement donné. Pour construire des p-vecteurs, un perceptron à plusieurs couches (MLP pour l'anglais *Multi-Layer Perceptron*) est entraîné à classifier le personnage joué dans l'enregistrement donné. Une fois le MLP entrainé, les auteurs proposent d'utiliser la dernière couche, avant le softmax, comme espace de représentation, formant le p-vecteur d'un extrait de voix. Dans cet article, les enregistrements sont classifiés par le MLP parmi 16 ou 12 personnages pour les protocoles décrits dans les sections 8.2.2 et 8.2.3.

8.1.2 La similarité personnage

Dans le but d'évaluer la représentation personnage, nous utilisons la tâche de similarité personnage (module de décision). La tâche consiste à décider si deux enregistrements de parole exprimés par deux locuteurs appartiennent au même personnage ou non. Dans ce travail, les deux enregistrements sont exprimés dans deux langues différentes. Nous utilisons le même système défini dans la sous-section 7.1.2.

voice-representation

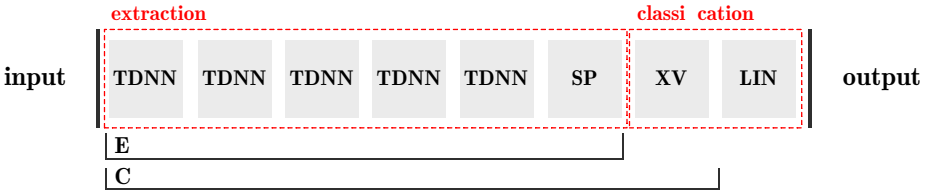


FIGURE 8.1 – Architecture de l'extracteur x-vecteur.

8.1.3 Extracteur de séquence

L'extracteur de séquence est basé sur l'approche de reconnaissance du locuteur x-vecteur ([PPK15a ; Sny+18]), décrit dans la figure 8.1. Il peut être décomposé en deux parties : 1) la partie extraction, où les caractéristiques acoustiques sont extraites depuis le signal, 2) la partie classification, dédiée à la tâche cible (ici, la reconnaissance du locuteur). Dans cette figure, *TDNN* signifie *Time Delay Neural Network*, comme référé dans [Sny+18], et *SP* signifie *Statistical Pooling*. La couche *XV* est une couche linéaire complètement connectée depuis laquelle nous extrayons classiquement les x-vecteurs, dans la littérature. Cette couche est suivie d'un *Leaky RELU*, une normalisation par lot (*batchnorm*), et un *dropout*. LIN est une couche linéaire complètement connectée qui précède la couche softmax.

Les auteurs de [RF16] montrent que les couches de ce type de réseau encodent de l'information à différents niveaux d'abstraction, dépendamment de la couche observée. En suivant l'hypothèse Ⓐ présentée dans l'introduction de ce chapitre, nous assumons que le plus proche nous sommes de la fonction objective (la sortie du réseau), plus présente est l'information locuteur dans le code de la couche et plus il est probable qu'utiliser cette information induise un biais locuteur dans l'entraînement d'une représentation personnage.

Dans le but de vérifier ces hypothèses, nous comparons les p-vecteurs issus des deux différents extracteurs de séquence. L'un, dénoté C, est une architecture x-vecteur classique contrairement au second, dénoté E, qui s'arrête une couche plus tôt, à la couche de *statistical pooling*. Comme résumé dans la figure 8.2, nous utilisons C ou E suivi d'un couche complètement connectée ($DENSE$) et d'une couche de représentation de laquelle nous extrayons les p-vecteurs (PV). E et C représentent la partie pré-entrainée à reconnaitre le locuteur en utilisant un corpus volumineux, à savoir *VoxCeleb2* ([CNZ18]). Les couches $DENSE$ et PV sont toujours entrainées en utilisant une fonction objective orientée personnage et un corpus dédié, à savoir *Mass Effect 3* (voir 8.2.1).

FIGURE 8.2 – L'architecture neuronale du module de représentation personnage (p-vecteur).

8.2 Protocole expérimental

Nous présentons les données dans la sous-section 8.2.1 et nous montrons comment nous les découpons dans les sous-sections 8.2.2 et 8.2.3. Nous détaillons ensuite le pré-entraînement x-vecteur et l'entraînement des p-vecteurs dans les sous-sections 8.2.4 et 8.2.5 respectivement. Enfin, nous décrivons l'évaluation dans la sous-section 8.2.6.

8.2.1 Description du corpus

Le corpus est composé d'enregistrements de voix provenant du jeu vidéo *Mass Effect 3*. Contrairement aux films, les voix de jeux vidéos présentent plusieurs particularités expliquées dans [Gre+19] (des effets radio sont présents dans les voix d'origine du corpus). Leurs enregistrements sont aussi plus faciles à collecter puisqu'ils sont séparés des ambiances même dans leur forme finale, contrairement aux enregistrements des archives de doublage de films. D'origine paru en anglais, le jeu a été traduit et les voix ont été jouées à nouveau dans plusieurs autres langues. Dans notre expérience, nous utilisons les versions anglaise et française des séquences audio, représentant un total d'environ 7,5 heures de parole dans chaque langue. Les segments (ou enregistrements) durent environ 3,5 secondes en moyenne. Un personnage est défini par une unique paire anglais-français de deux locuteurs distincts. Pour éviter des biais de l'identité du locuteur, nous considérons seulement un petit sous-ensemble où nous sommes certains qu'aucun des acteurs ne joue plus qu'un personnage. Un enregistrement audio correspond à une unique prise de parole d'un acteur dans une langue particulière. Nous appliquons ensuite un filtre qui garde seulement les enregistrements dont la durée est supérieure à une seconde. Enfin, nous ne gardons que 16 personnages pour lesquels nous avons le plus grand nombre d'enregistrements, comme [Gre+19] l'a fait.

8.2.2 Mise en exergue de l'Information Personnage Indépendante du Locuteur (IPIL)

Le jeu de données est composé seulement de personnages joués par des comédiens strictement différents. Les systèmes entrainés peuvent donc apprendre à associer des identités de locuteurs sans prendre en compte les particularités des personnages. Pour vérifier que ces systèmes ne sont pas trop faussés par cette configuration d'entraînement, nous proposons un protocole *modifié*. Ce protocole a pour objectif de neutraliser l'information personnage en modifiant les associations entre les acteurs, tout en respectant différentes contraintes pour éviter des biais comme le genre. En d'autres termes, les acteurs d'origine et doubleurs ne sont plus assignés au même personnage et sont associés à d'autres locuteurs parmi les 32 – en ayant toujours des paires de locuteurs français-anglais. En comparant la différence absolue des scores obtenus entre les protocole d'*origine* et *modifié*, nous pouvons mettre en lumière l'Information Personnage Indépendante du Locuteur (IPIL) capturée par les représentations. La présence ou l'absence de cette information est un indice pour identifier de potentiels biais dans un système. Nous notons S_m les ensembles *train*, *val* et *test* dérivant de ce protocole. Ce protocole est détaillé dans la sous-section 7.1.3.

8.2.3 Vérifier la capacité de généralisation de la représentation personnage

Dans le but de mesurer la capacité de nos systèmes à généraliser, nous proposons un protocole qui découpe les données différemment. Cette fois, nous ne gardons que 12 personnages, et leurs 24 locuteurs, pour la phase d'entraînement et nous utilisons les 4 personnages restant pour la phase de test. Nous continuons à découper le corpus en trois sous-ensembles, i.e. l'entraînement, la validation et le test. Le *train* et la *val* sont composés d'enregistrements différents mais provenant des mêmes 12 personnages. Pour construire les sous-ensembles *train*, *val* et *test*, nous sélectionnons aléatoirement pour chaque personnage 144, 36 and 180 enregistrements, tout en équilibrant le nombre d'enregistrements anglais et français. Nous avons ensuite un total de 1728 (*train*), 432 (*val*) et 720 (*test*) enregistrements. Nous notons ces sous-ensembles S_n. Les personnages présents dans les trois sous-ensembles (train, val et test) sont illustrés dans la figure 8.3.

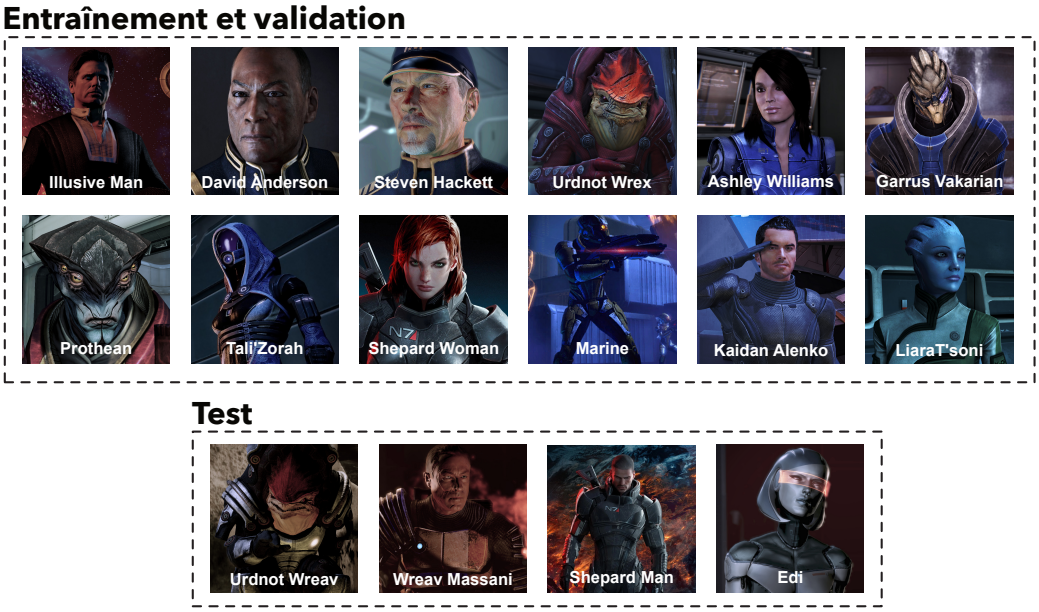


FIGURE 8.3 – Personnages de Mass Effect 3 présents dans les sous-ensembles d'entraînement, de validation et de test dans le découpage de données dédié à l'évaluation de la capacité à généraliser des systèmes de similarité de voix.

8.2.4 Pré-entraînement du modèle *x*-vecteur

Le modèle x-vecteur est entrainé sur le corpus Voxceleb 2. La couche XV de laquelle nous extrayons les x-vecteurs est accompagnée par une fonction d'activation LEAKY RELU et une normalisation de lot (*batchnorm*). La dernière couche linéaire (*i.e.* LIN dans la figure 8.1) est attachée à un softmax logarithmique, la sortie du réseau. La fonction de coût est la perte *cross-entropy*. Les MFCCs sont utilisés comme entrée, extraites en utilisant l'outil Kaldi ([Pov+11a]) avec les paramètres suivants : 30 coefficients cepstraux, 25 ms de largeur de fenêtre, $20 - 7{,}600$ Hz largeur de bande passante.

8.2.5 Entraînement du réseau de neurones *p*-vector

Comme dans l'article [Gre+20], la couche *DENSE* (i.e., voir figure 8.2) est composée d'une couche linéaire avec une fonction d'activation tangente hyperbolique et d'un *dropout* de 0,25. La couche p-vecteur (*i.e.* PV dans la figure 8.2) est composée des mêmes éléments, excepté pour le dropout dont la valeur est de 0,5. Nous calculons finalement un softmax logarithmique à la sortie du réseau. La fonction de coût que nous avons utilisée pour entrainer le réseau est la fonction de coût *cross-entropy*.

TABLE 8.1 – Résumé du nombre d'enregistrements et de paires d'enregistrements par sous-ensemble présents dans S_o, S_m et S_n.

	sous-ensemble	enregistrements	paires
S_o	*train*	2 304	165 888
S_o	*val*	288	2 592
S_o	*test*	288	2 592
S_m	*train*	2 304	165 888
S_m	*val*	288	2 592
S_m	*test*	288	2 592
S_n	*train*	1 728	124 416
S_n	*val*	432	7 776
S_n	*test*	720	64 800

8.2.6 Évaluation

Comme montré dans la section 8.1.2, nous évaluons les p-vecteurs sur une tâche de similarité personnage en utilisant une mesure basée sur les réseaux siamois. Nous générons des essais *cibles* composés de paires d'enregistrements appartenant au même personnage et *non-cibles* faites à partir d'enregistrements appartenant à deux personnages différents. Pour éviter des biais, les nombres de *cibles* et *non-cibles* sont équilibrés, comme le nombre de paires entre deux acteurs. Pour les ensembles S_o et S_m, nous générons 165 888 (*train*), 2 592 (*val*) et 2 592 (*test*) essais. Pour l'ensemble S_n, nous générons 124 416 (*train*), 7 776 (*val*) et 64 800 (*test*) essais. Le nombre d'enregistrements de paires par sous-ensemble de données est résumé dans le tableau 8.1.

Le seuil est décidé a posteriori au point d'*Equal Error Rate* (EER). Les performances du système sont ensuite exprimées en taux de réussite sur le test.

8.3 Réduire l'information discriminant le locuteur

Dans le but de vérifier l'hypothèse (A) décrite en introduction de ce chapitre, nous proposons de réduire le pouvoir discriminant du locuteur de l'extracteur de séquence en utilisant la configuration E (extraire un vecteur de séquence au niveau du *satistical pooling*) au lieu de la configuration C (extracteur x-vecteur classique).

D'abord, nous évaluons la perte en termes de pouvoir discriminant du locuteur lorsque nous utilisons E à la place de C (i.e., voir figure 8.1). Nous mesurons un EER de 22% (calculé sur la tâche de vérification du locuteur sur VoxCeleb1) en utilisant E, que nous pouvons comparer à l'EER de 6% que

nous obtenons pour la configuration C. Ce résultat valide la première partie de notre hypothèse : prendre une représentation vectorielle à un niveau plus bas dans le réseau réduit son pouvoir discriminant.

Nous vérifions ensuite l'impact de cette diminution sur la représentation personnage. Dans ce but, nous construisons deux représentations p-vecteurs, en utilisant les sorties des réseaux C (*config 1*) et E (*config 2*) respectivement. Dans les deux cas, nous gelons les couches de pré-entraînement locuteur (C ou E) pendant l'entraînement p-vecteur. Nous entrainons ensuite deux configurations de réseaux en utilisant le protocole à espace clos.

Nous effectuons une expérience comparative en utilisant trois configurations de données (*origine*, *modifié* et *mixte*) et deux extracteurs de séquences déjà présentés (*config 1* et *config 2*). Dans *origine*, les p-vecteurs (personnage) et la similarité de voix (décision) sont entrainés sur les données d'origine S_o, les vraies associations locuteurs des personnages. Dans *modifié*, les deux modules sont entrainés sur des données modifiées S_m, des associations artificielles mais consistantes entre les niveaux personnage et décision. Dans *mixte*, les p-vecteurs sont entrainés sur les données d'origine S_o tandis que le système de décision utilise les associations S_m.

Comme présenté dans le chapitre précédent, nous assumons que la différence en termes de taux de réussite entre *origine* et *modifié* est une évidence directe de la quantité d'information personnage indépendante du locuteur (IPIL) encodée dans la représentation p-vecteur. Nous supposons aussi que le taux de réussite obtenu en utilisant la configuration *mixte* est un moyen de mesurer la quantité d'information spécifique au locuteur dans les p-vecteurs.

Le tableau 8.2 présente les résultats de l'ensemble des expériences. Nous observons une différence de 3,2 points entre les configurations *origine* et *modifié* dans le cas de *config 1* et 1,7 points dans le cas de *config 2*. Cela montre que les p-vecteurs encodent de l'IPIL. La lecture de ce tableau montre également que la quantité de cette information personnage est plus faible lorsqu'un extracteur de séquence avec moins de pouvoir discriminant du locuteur (*config 2*) est utilisé.

Le protocole *mixte* montre un haut niveau général et une plus forte présence de l'information locuteur pour la *config 2* que pour la *config 1*.

TABLE 8.2 – Taux de réussite de *config 1* (en utilisant C) and *2* (en utilisant E) sur le test avec le protocole original, modifié, ou mixte.

	module personnage	module de décision	performance config 1	performance config 2
original	S_o	S_o	92.3	95.7
modifié	S_m	S_m	89.1	94.0
mixte	S_o	S_m	79.7	81.8

TABLE 8.3 – Configuration d'entraînement pour les systèmes config 3 and 4.

config	couches	pré-entraînement	gel
3	E	VoxCeleb 2	×
4	E	×	×

8.4 Donner plus de pouvoir à la classification personnage

Cette section met à l'épreuve l'hypothèse Ⓑ décrite en introduction de ce chapitre. Nous proposons de donner plus de pouvoir au classifieur personnage pendant la construction des p-vecteurs en réduisant l'influence du pré-entraînement de l'extracteur de séquence. Dans ce but, nous entraînons des p-vecteurs en suivant deux types de configurations présentées dans le tableau 8.3. Dans la configuration 3, nous ne gelons pas les couches E pour donner plus de pouvoir au classifieur p-vecteur. Dans la configuration 4, nous donnons tout le pouvoir aux p-vecteurs en supprimant complètement le pré-entraînement.

Le tableau 8.4 présente les résultats obtenus sur la tâche de similarité personnage sur les systèmes de *config 3* and *4*. Bien que les performances soient aussi élevées ($\geq 90\%$), nous observons moins de différences en termes de taux de réussite entre les configurations *origine* et *modifié*. Nous avons +0,5 point et −0,8 point de différence pour respectivement *config 3* et *4*. À partir de ces résultats, et compartivement à ceux obtenus avec *config 1*, nous observons une évaporation de l'information personnage indépendante du locuteur quand nous donnons plus de pouvoir au classifieur personnage.

Le tableau 8.5 montre le nombre de paramètres modifiés pendant l'étape d'entraînement (*paramètres apprenables*). Il montre aussi le nombre total de paramètres où les paramètres gelés sont pris en compte. Puisque *config 3* et *4* ont un plus grand nombre de paramètres à apprendre que *config 1* et *2*, nous proposons d'évaluer un système plus petit sans pré-entraînement constitué de 664 144 paramètres. Nous le nommons *petit*. Nous obtenons un taux de réussite

TABLE 8.4 – Résultats obtenus en taux de réussite calculé sur le test avec les protocoles d'origine, modifié et mixte en utilisant la *config 3* et *4*.

	module personnage	module de décision	performance config 3	performance config 4
original	S_o	S_o	95.6	93.6
modifié	S_m	S_m	95.1	94.4
mixte	S_o	S_m	81.5	82.1

TABLE 8.5 – Nombre de paramètres total et de paramètres d'apprentissage pour chaque configuration de système étudiée.

système	paramètres d'apprentissage	paramètres total
config 1	296 528	4 523 492
config 2	1 570 384	4 260 836
config 3	4 260 836	4 260 836
config 4	4 260 836	4 260 836
petit	664 144	664 144

de 92,5 and 89,5 respectivement sur les protocoles *origine* et *modifié*. Cette différence de 3 points suggère que réduire le nombre de paramètres a permis au système p-vecteur d'encoder plus d'information personnage.

Pour finir, nous souhaitons vérifier la capacité de généralisation de nos différentes configurations sur des personnages inconnus du sous-ensemble d'entraînement (le protocole S_n). Les résultats obtenus avec ce protocole sont reportés dans le tableau 8.6. Les systèmes configurés avec *config 1* et *2* obtiennent respectivement un taux de réussite de 68,5% et 68,6% en termes de similarité personnage. Ce résultat suggère fortement que décroitre le pouvoir discriminant du locuteur de l'extracteur de séquence ne dégrade pas la capacité de généralisation de la représentation personnage. Ces expériences tendent aussi à montrer que l'usage du pré-entraînement locuteur aide la représentation personnage (p-vecteur) à généraliser sur des personnages inconnus puisque la *config 3*, *config 4* et *petit* (où le pré-entraînement locuteur est partiellement ou totalement supprimé) obtiennent des performances plus faibles que *config 1* et *2* (avec pré-entraînement locuteur).

Conclusion

Ce chapitre investit l'influence du pré-entraînement locuteur sur la représentation personnage. Une première expérience montre que choisir une représentation locuteur moins discriminante pour l'extracteur de séquence réduit la présence d'IPIL. Une seconde expérience montre que réduire partiellement ou

TABLE 8.6 – Information Personnage Indépendante du Locuteur (IPIL), Information Locuteur (IL) et Pouvoir de Généralisation (PG) par système. L'IPIL est la différence des taux de réussite obtenus entre les protocoles *origine* et *modifié*, IL est le taux de réussite obtenu en utilisant le protocole *mixe* et GP est le taux de réussite calculé en utilisant le protocole de généralisation (S_n).

	GP	IL	IPIL
config 1	68,5	79,7	3,2
config 2	68,6	81,8	1,7
config 3	61,2	81,5	0,5
config 4	61,1	82,1	−0,8
petit	66,9	80,8	3,0

complètement le pré-entraînement locuteur ne rend pas possible l'encodage de plus d'information personnage, ou de mieux conserver sa part indépendante du locuteur. Nous pouvons conclure que le pré-entraînement de l'extracteur de séquence ne guide pas trop fortement les p-vecteurs vers l'information locuteur et apporte de l'information utile à la construction de la représentation personnage. Basée sur ces résultats, la notion de la voix d'un personnage apparaît essentiellement inséparable de celle de son locuteur. Dans la suite de ce manuscrit, nous proposons une dernière expérience qui consiste à extraire des descripteurs de voix en utilisant des algorithmes de regroupement dans l'espace des p-vecteurs.

Chapitre 9

Le raffinage d'étiquettes : une méthode semi-supervisée d'extraction de descripteurs de voix sans vérité terrain

Sommaire

L'extraction de caractéristiques et de descripteurs de voix est un problème connu qui prend différentes formes dans la littérature. Des études se sont intéressées aux descriptions prosodique et acoustique des voix. L'article [CM04] propose de voir la prosodie comme 4 paramètres : la hauteur (*pitch*), le volume (*power*), le rythme (*duration*) et la qualité (*voice quality*). D'autres travaux ont étudié la classification supervisée de segments vocaux pour reconnaitre la langue parlée, l'accent ou le locuteur dans un segment de parole ([SRM03 ; Deh+11b]). Ces étiquettes ne sont cependant pas dédiées à la caractérisation du personnage joué.

Dans les articles [OR16 ; ORB14], un non-initié est entraîné par des experts à annoter des enregistrements vocaux. Cette compétence lui sert ensuite à annoter les enregistrements de voix françaises du jeu vidéo Mass Effect avec des étiquettes classifiées en différentes catégories : les étiquettes physiologiques (e.g. genre, âge), phonétiques (e.g. qualité de voix, tension, effort vocal), de timbre, articulatoires, prosodiques (e.g. fréquence fondamentale f0, taux de parole) et de jeu d'acteur (e.g. attitude, émotion, situation, archétype). Un classifieur neuronal est ensuite entraîné à reconnaitre les étiquettes associées aux enregistrements qui lui sont fournis. Une fois entraîné, ce système extrait la « signature vocale » d'un enregistrement. L'article compare l'usage de cette signature avec celle des *i*-vecteurs dans une étude perceptive et montre qu'utiliser des annotations humaines permettrait d'extraire des représentations de voix plus proches de la perception humaine de la similarité de voix.

Une telle annotation pose cependant plusieurs problèmes, notamment lorsque nous souhaitons entraîner des modèles neuronaux. Pour atteindre de bonnes performances, ces modèles nécessitent une quantité très élevée de données. Pour le moment, aucun corpus de voix de doublage ne propose d'annotations. Certains corpus sont dédiés aux émotions dans la voix actée, mais aucun ne décrit ce que nous pourrions nommer un profil vocal. Différents éléments expliquent cette absence d'annotations. En effet, il n'existe à ce jour aucune taxonomie standardisée pour décrire les voix actées. Nous entendons souvent parler de couleur de la voix (blanche, bleu, rouge), de types de voix (voix rauque), des stéréotypes (un guerrier, un scientifique fou), ou de directions

données à l'acteur (voix droite) mais tous ces termes sont sensiblement différents en fonction de la personne qui l'emploie. Ils sont fortement sujet à interprétation. L'inexistence d'une taxonomie commune rend alors impossible la construction d'un protocole d'annotation de données vocales. Ces annotations auraient pu permettre d'étudier la caractérisation de certains types de voix. Sans taxonomie disponible, nous nous intéressons au moyen de faire émerger automatiquement des caractéristiques de voix actées depuis un ensemble d'enregistrements. C'est pourquoi ce chapitre pose les questions suivantes : quelles sont les étiquettes personnage présentes dans la voix actée et comment classifier nos enregistrements par étiquette sans les connaitre a priori ?

La solution que nous proposons consiste à automatiser le processus de classification d'enregistrements vocaux par étiquettes personnage, sans connaissance a priori de la vérité terrain. Pour ce faire, nous posons dans nos expériences un premier postulat : il existe, pour chaque session d'enregistrement, une caractéristique vocale dominante que l'on peut associer à la voix du personnage joué. Autrement dit, à chaque enregistrement de voix de plusieurs secondes, nous pouvons associer une étiquette représentant un descripteur vocal. La seule vérité terrain que nous ayons pour chaque enregistrement est l'identité du personnage qui le joue. Dans ce chapitre, nous proposons une méthode que nous qualifions de « raffinage » pour décomposer les étiquettes personnages en sous-étiquettes (étiquettes raffinées) de caractéristiques vocales. Des méthodes d'évaluation sans connaissance a priori sont présentées et discutées.

Nous présentons d'abord la méthode de raffinage qui permet de trouver les étiquettes personnage recherchées. Nous présentons ensuite l'algorithme k-means ainsi que différentes distances pour appliquer la méthode de raffinage. Différentes métriques d'algorithmes de regroupement sont aussi présentées. Puis, nous proposons un protocole expérimental et discutons des résultats. Pour finir, nous concluons et nous discutons des perspectives.

9.1 Raffinage des étiquettes initiales

9.1.1 Les étiquettes raffinées

Nous supposons, dans nos travaux, que les étiquettes (des classes) sont décomposables en différentes sous-étiquettes de plus bas niveau. Dans notre cas, à partir de l'étiquette initiale du personnage associée à un enregistrement, nous souhaitons décomposer le personnage en descripteurs vocaux tels que l'émotion jouée (e.g. agressif, heureux), une couleur de voix (e.g. blanc, rouge) ou un

timbre (e.g. rauque, éclaircie). Cependant, la contrainte que nous posons dans cette expérience consiste à ne pas connaitre a priori la nature de ces étiquettes et à n'avoir aucun exemple dans nos données des étiquettes recherchées. Les définitions de ces étiquettes ne sont d'ailleurs potentiellement pas intuitives ou intelligibles pour des être humains, même experts.

L'équation 9.1 formalise la définition d'une étiquette raffinée. Nous notons R l'ensemble des étiquettes raffinées et $a_i^{(l)} \in \{r_1, r_2 ... r_{|R|}\}$ la i^{eme} étiquette raffinée associée à l'étiquette initiale l. Précisons aussi que le nombre n d'étiquettes raffinées est différent en fonction de l. Ainsi, d'une étiquette initiale l nous souhaitons trouver les labels raffinés $a_i^{(l)}$.

$$l \begin{cases} a_1^{(l)} \\ a_2^{(l)} \\ ... \\ a_n^{(l)} \end{cases} \tag{9.1}$$

Posons les fonctions $f : e \mapsto l$ et $g : e \mapsto r$ où e est une observation (e.g. enregistrement vocal). L'équation 9.2 fournit des exemples fictifs. Elle montre les images attendues des fonctions f et g en fonction de deux étiquettes initiales l_1 et l_2. Dans notre modélisation du problème, nous supposons qu'une étiquette raffinée r peut apparaître chez plusieurs observations dont les étiquettes initiales sont différentes. C'est le cas ici avec r_1 qui apparaît à la fois dans une observation associée à l_1, mais aussi dans une observation associée à l_2. Dans le cas de nos données, cela signifie qu'une étiquette peut apparaître chez plusieurs personnages.

$$l_1 \begin{cases} f : e_1^{(l_1)} \mapsto l_1 \\ f : e_2^{(l_1)} \mapsto l_1 \end{cases} , l_1 \begin{cases} g : e_1^{(l_1)} \mapsto r_1 \\ g : e_2^{(l_1)} \mapsto r_2 \end{cases} , l_2 \begin{cases} f : e_1^{(l_2)} \mapsto l_2 \\ f : e_2^{(l_2)} \mapsto l_2 \end{cases} , l_2 \begin{cases} g : e_1^{(l_2)} \mapsto r_1 \\ g : e_2^{(l_2)} \mapsto r_3 \end{cases} \tag{9.2}$$

9.1.2 Extraction des étiquettes raffinées

La figure 9.1 présente l'extraction des étiquettes raffinées en trois étapes.

— La première étape consiste à entraîner un système à extraire des représentations des observations $e \in E$ en utilisant les étiquettes initiales $l \in L$.

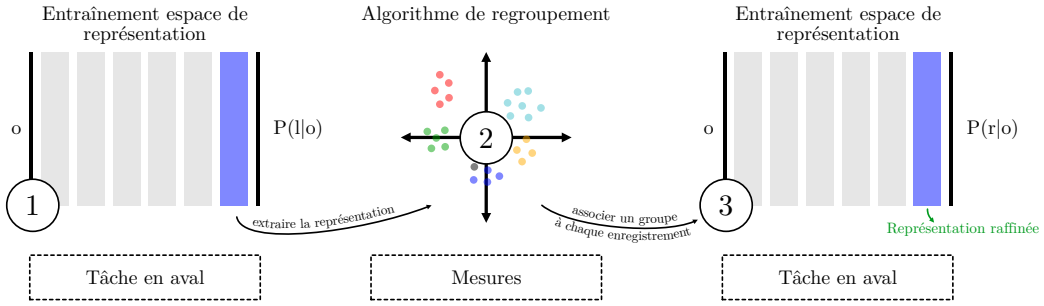


FIGURE 9.1 – Processus de raffinage d'un espace de représentation.

— La seconde consiste à utiliser un algorithme de regroupement pour associer des nouvelles étiquettes $r \in R$ aux observations e.
— La troisième consiste finalement à entraîner un nouveau système de classification pour en extraire une représentation raffinée.

Pour évaluer la pertinence des étiquettes r, nous employons deux méthodes. Appliquée à l'étape 2, la première méthode consiste à calculer différentes mesures de qualité de regroupements telles que la v-mesure, ou la pureté. Appliquée à l'étape 3, la seconde méthode consiste à extraire une représentation des observations e et de les soumettre à l'évaluation d'une tâche. Cette tâche peut être diverse : reconnaissance du locuteur, reconnaissance des mots ou, dans notre cas, similarité personnage. Grâce à cette seconde méthode, il est possible d'évaluer la représentation obtenue à l'étape 1 et de comparer avec les résultats obtenus à l'étape 3. Ainsi, nous pouvons mettre en exergue l'influence qu'a eu le raffinage sur l'encodage des informations des représentations personnage.

9.2 Algorithme *k*-means

9.2.1 Description de l'algorithme

Le regroupement est un type d'algorithme qui consiste à placer des éléments dans différents groupes dans le but que les éléments appartenant au même groupe soient similaires et que ceux associés à des groupes différents soient dissimilaires. K-means ([Mac67]) est un algorithme itératif très connu, convergeant vers un minimum local d'une partition donnée en recherchant le modèle qui minimisera l'expression suivante :

$$J(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||^2 \tag{9.3}$$

Où C_k est le groupe k dans l'ensemble des groupes C de taille K, x_i est le vecteur de représentation d'une entité et μ_k le centroïde du groupe k. Cet algorithme nécessite de connaitre a priori le nombre de regroupements k. Habituellement, l'algorithme k-means est employé en synergie avec la distance euclidienne. Il existe cependant d'autres distances dont deux sur lesquelles nous nous focaliserons dans ce travail de recherche.

— **la distance cosinus.** Elle est très employée dans la littérature de la reconnaissance du locuteur, notamment pour calculer un score de similarité entre des représentations vectorielles de voix.
— **la distance de Mahalanobis.** Elle nous intéresse car elle a été utilisée dans des travaux similaires qui consistaient à trouver des locuteurs à partir d'enregistrements plongés dans l'espace des x-vecteurs grâce à un algorithme de regroupement ([Lap18]).

9.2.2 Distances

Pour comparer deux vecteurs, il existe différentes distances dans la littérature telles que la distance *city-block*, la distance de corrélation, la L1-norm ou la L2-norm. Nous prenons le parti dans cette thèse de nous focaliser sur les distances euclidienne, cosinus et de Mahalanobis puisqu'elles apparaissent souvent dans la littérature de la similarité de voix. La distance euclidienne est triviale et très connue. Cependant, nous présentons ci-dessous les deux autres distances.

La distance cosinus est très souvent employée en RAL pour calculer la similarité entre deux vecteurs. Elle présente l'avantage de normaliser indirectement les vecteurs comparés en ne donnant aucune importance à leurs normes. Nous la retrouvons dans différents domaines de recherche de la littérature. Notamment en NLP (Natural Language Processing), où elle permet de comparer les contenus sémantiques de deux mots ([Mik+13]). Des travaux en vision par ordinateur s'en servent aussi pour comparer des représentations d'images ([ZL03]), notamment en la combinant avec un algorithme de regroupement ([Cha+17]). La distance cosinus est aussi utilisée dans le domaine de la vérification locuteur ([Kum+15]) où elle propose encore de très bonnes performances et ne nécessite pas de phase d'apprentissage, contrairement à la méthode PLDA ([AK+20]).

Tout comme la distance cosinus, la distance de Mahalanobis est largement utilisée dans la littérature. Nous la retrouvons dans la vision par ordinateur ([CM04 ; PT09]) ou dans le langage naturel ([Deu18]). La distance de Mahalanobis est aussi utilisée en synergie avec des algorithmes de regroupement tels

TABLE 9.1 – Étude de fonction des mesures des regroupements

	angle de vue	$\lim_{k\to 1} f(C;X)$	$\lim_{k\to \lvert X\rvert} f(C;X)$
entropie	groupe	$+\infty$	0
homogénéité	groupe	0	1
pureté de groupe	groupe	0	1
complétude	classe	1	0
pureté de classe	classe	1	0
v-mesure	hybride	$\times$	$\times$
pureté-K	hybride	$\times$	$\times$

qu'il est décrit dans [Lap18] où l'objectif est de retrouver les groupes d'enregistrements appartenant au même locuteur.

9.2.3 Métriques d'évaluation des regroupements

Pour évaluer un modèle de regroupement, différentes mesures sont utilisées. Ces mesures nécessitent de connaitre pour chaque observation de la base de données le groupe auquel l'algorithme l'a associé ainsi que son étiquette d'origine. Dans cette sous-section, nous présentons les mesures de regroupement suivantes : entropie, homogénéité, complétude, v-mesure, pureté de classe, pureté de groupe et pureté-K.

Ces mesures se calculent sous deux angles de vue différents. Les mesures d'entropie, d'homogénéité et de pureté de groupe vérifient si chaque groupe est associé à des observations appartenant à la même classe. Elles offrent un angle de vue de groupe. Les mesures de complétude et de pureté de classe vérifient si chaque classe est associée à des observations appartenant au même groupe. Elles offrent un angle de vue de classe.

Pour analyser le comportement de ces mesures, posons C un modèle de regroupement appris sur les entités X avec comme paramètre k, le nombre a priori de groupes. Posons f la fonction qui mesure la qualité du modèle C (e.g. l'entropie) sur X. Le tableau 9.1 montre le comportement des mesures en calculant pour chacune d'entre elles la limite de f en $k \to +\infty$ et $k \to 0$, en supposant qu'aucun groupe n'est vide.

L'entropie est une mesure similaire à l'homogénéité. Bien qu'elle aide à comparer deux systèmes, elle présente un inconvénient majeur. Elle est bornée entre 0 et $+\infty$. Interpréter la mesure d'un seul et unique modèle est alors difficile contrairement à l'homogénéité qui est bornée entre 0 et 1. Notons que la pureté de groupe est une mesure très similaire à l'homogénéité.

Les mesures de complétude et de pureté de classe, en plus d'être très simi-

laires, sont à l'antipode des précédentes mesures. Elles ne mesurent pas l'homogénéité des groupes mais l'homogénéité des classes. Plus la valeur se rapproche de 1, plus la classe concernée n'est observée que dans un seul groupe. Plus la valeur se rapproche de 0, plus la classe concernée est observée dans un grand nombre de groupes à des quantités similaires.

Les mesures hybrides sont des moyennes pondérées des mesures de classes et de groupe. Leurs limites en $k \to +\infty$ et $k \to 0$ ne sont pas claires tant elles dépendent du rapport de force entre homogénéité (pureté de groupe) et complétude (pureté de classe).

Démontrer les limites de chaque fonction est assez simple. Pour l'homogénéité et la pureté de groupe, elle est maximale lorsque les éléments de chaque groupe appartiennent à la même classe. Lorsque nous atteignons $k = |X|$, chaque élément est alors assigné à un groupe et chaque groupe est assigné à un élément, en supposant qu'il n'existe pas de groupe vide. Il n'existe alors qu'une seule classe pour chaque groupe. L'homogénéité est donc maximale, la pureté de groupe aussi. Dans le cas contraire, lorsque $k = 1$, toutes les entités appartiennent au même groupe. Le groupe est donc hétérogène au maximum et les valeurs d'homogénéité et de pureté de groupe tendent alors vers 0. Pour la complétude et la pureté de classe, le raisonnement est similaire et trivial.

Entropie appliquée à un modèle de regroupement

La mesure d'entropie en théorie de l'information représente l'espérance de l'incertitude a priori qu'une variable aléatoire X puisse prendre la valeur x. Elle est parfois illustrée de la manière suivante : l'entropie de X est le nombre moyen de questions (binaires) auxquelles il est nécessaire de répondre avant d'être sûr de la réalisation de X. Appliquée aux modèles des algorithmes de regroupement, elle ressemble fortement au calcul de l'homogénéité d'un groupe de points. Plus une classe est présente dans un groupe, plus la mesure d'entropie est faible. Si le groupe ne possède qu'une classe, cette mesure est égale à 0, son minimum. Au contraire, si toutes les classes son équilibrées et ont la même probabilité d'apparition, alors la valeur de l'entropie est maximale. L'entropie a tendance à diminuer lorsque le nombre de groupes (la valeur k du k-means) est augmenté jusqu'à atteindre la valeur 0 lorsque le nombre de groupes est égal au nombre de d'entités $|X|$.

Pour calculer l'entropie d'un groupe, posons d'abord la variable aléatoire $Y^{(k)}$ associée à chaque groupe C_k d'un modèle où les entités (les lignes) des données peuvent être associées aux classes $\{l_1, l_2, ..., l_J\}$ avec les probabilités

$\{P_{y^{(k)}}(l_1|C_k), P_{y^{(k)}}(l_2|C_k), ..., P_{y^{(k)}}(l_J|C_k)\}$. L'entropie d'un groupe k se mesure de la manière suivante :

$$E(Y^{(k)}) = -\sum_{j}^{J} P(l_j|C_k) * log_2(P(l_j|C_k)) \tag{9.4}$$

avec $log_2(0) = 0$. L'entropie d'un modèle de regroupement se calcule de la manière suivante :

$$\frac{\sum_{k=1}^{K} E(Y^{(k)}) \times n_k}{n} \tag{9.5}$$

Ici, nos classes sont les personnages, n est le nombre total d'entités et n_k est le nombre d'entités du groupe k.

Homogénéïté

Le calcul d'homogénéité ([RH07]) est très proche de celui de l'entropie. Sa principale différence est son caractère normalisé qui borne la mesure entre 0 et 1. Cette mesure est aussi comparable à la précision puisque sa valeur est maximale lorsque chaque groupe contient seulement les membres d'une seule et même classe. Ci-dessous, la formule de l'homogénéité :

$$H(L|C) = -\sum_{k=1}^{|K|}\sum_{j=1}^{|J|} \frac{a_{k,j}}{n} log \frac{a_{k,j}}{\sum_{j=1}^{|J|} a_{k,j}} \tag{9.6}$$

$$H(C) = -\sum_{k=1}^{|K|}\sum_{j=1}^{|J|} \frac{a_{k,j}}{J} log \frac{\sum_{k=1}^{|K|} a_{k,j}}{J} \tag{9.7}$$

Avec $a_{k,j}$ le nombre d'occurences associées à la fois au groupe C_k et au label l_j.

Complétude

La complétude ([RH07]) est complémentaire à l'homogénéité. Elle est comparable avec le rappel classiquement utilisé pour mesurer les performances d'un système de classification puisqu'elle est maximale lorsqu'une classe n'apparaît que dans un seul groupe. Elle est aussi bornée entre 0 et 1. La formule ci-dessous décrit le calcul de la complétude :

$$H(C|L) = -\sum_{j=1}^{|J|}\sum_{K=1}^{|K|}\frac{a_{j,k}}{n}log\frac{a_{j,k}}{\sum_{k=1}^{|K|}a_{j,k}} \tag{9.8}$$

$$H(C) = -\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{a_{j,k}}{|J|}log\frac{\sum_{j=1}^{J}a_{j,k}}{J} \tag{9.9}$$

V-Mesure

Notons h l'homogénéité et c la complétude. La V-mesure ([RH07]) est la moyenne harmonique pondérée de l'homogénéité et de la complétude.

$$V_\beta = \frac{(1+\beta)\times h \times c}{\beta \times h + c} \tag{9.10}$$

Cette mesure est assez similaire à la F-mesure, si β est supérieur à 1, la complétude est pondérée plus fortement dans le calcul. Si β est inférieur à 1, l'homogénéité est pondérée plus fortement. L'avantage que présentent les calculs de l'homogénéité, la complétude et la f-mesure est qu'elles ne dépendent pas du nombre de classes, du nombre de regroupements, de la taille des données et de l'algorithme de regroupement utilisé. Ces mesures peuvent donc être appliquées pour comparer différents regroupements indépendamment du nombre de points (n-invariance), du nombre de classes ou du nombre de regroupements. De plus, calculer l'homogénéité et la complétude permet d'avoir une meilleure précision pour évaluer la qualité d'un regroupement.

9.3 Protocole expérimental

9.3.1 Corpus et découpage des données

Dans les expériences de ce chapitre, nous utilisons des enregistrements de voix actées provenant de deux jeux vidéos connus, Mass Effect et Skyrim. Pour les données Mass Effect, nous reprenons le découpage de données S_n présenté dans la section 8.2.3 où les personnages présentés au système durant la phase de test ne font pas partie du sous-ensemble d'entraînement.

Durant la phase de validation, un algorithme de regroupement permet d'extraire des étiquettes de voix que nous appelons des descripteurs. Utiliser les données Mass Effect pour réaliser ce regroupement pose cependant certains problèmes. L'ensemble de validation de Mass Effect souffre d'un manque de variabilité puisqu'il ne contient que des personnages connus de l'entrainement.

Le système ayant construit une représentation de voix servant à caractériser le personnage joué, il est difficile d'en faire émerger des étiquettes qui s'éloignent de la définition de personnage. Malheureusement, nous n'avons pas, dans Mass Effect, des enregistrements de personnages inconnus que nous pourrions ajouter à l'ensemble de validation puisque ces derniers sont utilisés pour le test. Nous assumons donc que l'usage de Skyrim comme un second ensemble de validation apportera plus de variabilité et aidera à mettre en lumière des étiquettes de descripteurs de voix dans les représentations personnage. Les données Skyrim sont composées de 18 250 enregistrements, provenant de 32 personnages différents. Tout comme sur Mass Effect, chaque enregistrement joué en anglais est associé à un enregistrement français et vice-versa. Cependant, dans cette expérience, nous n'équilibrons pas le nombre d'enregistrement par personnage des données Skyrim.

Par ailleurs, le corpus Skyrim présente un défaut et n'est en conséquence pas utilisé comme corpus d'entrainement. En effet, l'information du comédien jouant le personnage n'est pas fiable et utiliser le corpus en entraînement peut introduire des biais locuteur.

9.3.2 P-vecteurs

Tout comme dans l'article [Gre+20], les p-vecteurs sont extraits depuis un réseau de neurones entraîné à classifier le personnage joué dans l'enregistrement fourni. Ce réseau est composé des couches suivantes : une couche linéaire suivie d'une fonction d'activation tangente hyperbolique et d'un *dropout* de 0,25. Une seconde couche est composée des mêmes éléments où seul le *dropout* est modifié à 0,5. De cette couche, nous extrayons les p-vecteurs. Finalement, nous calculons un softmax logarithmique à la sortie du réseau. La fonction de coût que nous avons utilisée est l'entropie croisée.

9.3.3 Regroupements avec k-means

Nous souhaitons dans cette expérience valider la méthode proposée. Pour ce faire, nous nous focalisons sur l'algorithme k-means, bien que d'autres algorithmes existent. Celui-ci est le plus simple à mettre en place et est largement employé pour le regroupement de données. L'algorithme de regroupement employé est le *k-means* du *toolkit sklearn* avec les paramètres par défaut. Pour converger plus rapidement vers une solution, nous initialisons les barycentres avec l'option k-means++. L'algorithme s'arrête lorsqu'il converge. Le nombre maximum d'itérations est de 300 en cas de non convergence. L'algorithme

est lancé 10 fois, puis nous sélectionnons le modèle de regroupement ayant la meilleure inertie.

9.3.4 Raffinage des *p*-vecteurs

Une fois l'algorithme de regroupement terminé, nous construisons, pour chaque valeur k, un nouveau réseau de neurones en utilisant la même architecture neuronale que celle employée pour les p-vecteurs. Au lieu de prédire les étiquettes initiales, ce nouveau réseau est entraîné à prédire les étiquettes raffinées grâce à l'algorithme de regroupement. Après entraînement, nous pouvons extraire de sa dernière couche une représentation de l'enregistrement fourni que nous nommons p-vecteur raffiné.

9.3.5 Évaluation

Nous évaluons nos représentations de voix (p-vecteurs et p-vecteurs raffinés) sur une tâche de similarité personnage en utilisant la mesure de similarité d'un réseau de neurones siamois. Nous générons des essais *Cibles* composés de paires d'enregistrements appartenant au même personnage et des essais *Non-cibles* composés de paires d'enregistrements appartenant à deux personnages différents. Pour éviter des biais, nous équilibrons le nombre de paires *Cibles* et *Non-cibles*. Nous équilibrons également le nombre de paires entre deux acteurs. En conséquence, pour l'ensemble S_n, nous générons 124 416 (*train*), 7 776 (*val*) and 64 800 (*test*) essais. Le seuil de décision est défini *a posteriori* au point de ratio d'erreur égal (Equal Error Rate). La performance du système est ensuite exprimée en taux de réussite sur l'ensemble de test.

9.4 Résultats et discussion

Tester toutes les valeurs possibles de k étant trop coûteux, les expériences présentées ci-après proposent de ne tester que certaines valeurs de k.

Plusieurs valeurs nous paraissaient évidentes à tester comme 4, 8, 12 et 24 qui sont le nombre de personnages (4 et 12) et de locuteurs (12 et 24) dans le *train* et le *test*. Nous posons pour hypothèse aussi que le nombre de descripteurs vocaux que nous recherchons est supérieur au nombre de personnages que notre système doit apprendre à modéliser. Nous proposons donc aussi de nous intéresser aux valeurs 18, 32, 48 et 64 qui sont des valeurs arbitraires mais supérieures à 12. En contre-hypothèse, nous essayons aussi des valeurs inférieurs à 4 et à 12 telles que $k = 2$ et $k = 6$.

9.4.1 Sélection du nombre de classes *k*

Nous avons d'abord appliqué l'algorithme de regroupement k-means sur des p-vecteurs extraits des données S_n de Mass Effect. Les mesures v-mesure et pureté calculées sur ce modèle sont reportées dans le tableau 9.2. Nous observons que le regroupement appliqué aux données Mass Effect met bien en valeur les étiquettes initiales des 12 personnages présents dans le corpus. En effet, les valeurs des métriques v-mesure (0,91) et pureté (0,90) sont plus élevées lorsque $k = 12$.

Sur le test, nous observons que les valeurs les plus élevées de la v-mesure sont à $k = 6$ (0,58), et celles de la pureté à $k = 2$ (0,70). Ces deux valeurs de k sont les plus proches du nombre de personnages (4) présents dans le corpus de test. Les groupes représentent certainement des étiquettes personnages dégradées par la difficulté des p-vecteurs à généraliser sur des personnages inconnus à la phase de test.

Nous avons ensuite remplacé les étiquettes personnages par les étiquettes des groupes pour entraîner ce que nous nommons des p-vecteurs raffinés. Le réseau de classification des p-vecteurs raffinés (i.e. représentation raffinées dans la figure 9.1) ne doit pas reconnaitre le personnage joué, mais le groupe auquel l'enregistrement appartient. Ces p-vecteurs raffinés sont utilisés pour entrainer 10 réseaux de neurones siamois. Une fois entrainés, nous sélectionnons le meilleur réseau sur la validation. Les taux de réussite (*accuracy*) de ce système sur la validation et le test sont reportés dans les deux dernières lignes du tableau 9.2.

Nous observons un phénomène très intéressant. D'abord, sans surprise, le meilleur taux de réussite sur la validation est obtenu à $k = 12$, soit le nombre de personnages. Cependant, contrairement aux métriques précédentes, nous n'observons pas dans les résultats une loi normale. En effet, pour la v-mesure de la validation, plus k s'éloigne de 12, plus les valeurs diminuent. De même pour la pureté. Par contre, alors que nous observons une diminution du taux de réussite sur la validation à $k = 18$ (0,92), à $k = 64$ sa valeur est légèrement plus élevée (0,93). Plus intéressant encore, si nous ignorons $k = 12$ parce qu'il représente des étiquettes personnage dont nous avons déjà connaissance et qui, par conséquent, ne nous intéressent pas, le système à $k = 64$ devient le nouveau meilleur système avec un taux de réussite de 0,70 sur le test. Ce résultat est 4 points supérieur à celui obtenu à $k = 12$ et dépasse les résultats que nous avions eu dans nos précédentes expériences.

En résumé, les mesures v-mesure et pureté ne permettent pas de sélectionner

TABLE 9.2 – Résultats obtenus en utilisant les enregistrements de Mass Effect pour l'algorithme de regroupement k-means. Les mesures v-mesure et pureté sont calculées sur les regroupements extraits des p-vecteurs. Les taux de réussite (lignes "siamois val" et "test") sont obtenus sur les réseaux siamois entraînés avec les p-vecteurs raffinés.

	2	4	6	8	12	18	24	32	48	64
v-mesure val	0,40	0,62	0,73	0,80	**0,91**	0,86	0,83	0,80	0,75	0,74
v-mesure test	0,54	0,56	**0,58**	0,50	0,48	0,45	0,44	0,44	0,41	0,41
pureté-K val	0,41	0,53	0,64	0,73	**0,90**	0,78	0,68	0,61	0,52	0,48
pureté-K test	**0,70**	0,65	0,64	0,54	0,50	0,45	0,43	0,40	0,32	0,31
siamois val	0,67	0,73	0,80	0,87	**0,94**	0,92	0,93	0,93	0,92	**0,93**
siamois test	0,55	0,65	0,62	0,66	0,66	0,69	0,63	0,66	0,63	**0,70**

le meilleur k. En effet, les points se rassemblent naturellement par personnage. Le meilleur k sur la validation est donc $k = 12$, soit, le nombre de personnages présents dans le train et la validation. Le meilleur k sur le test est une valeur proche de $k = 4$, soit le nombre de personnages présents dans le test. Ces mesures ne permettent donc pas de mettre en lumière des descripteurs vocaux différents des étiquettes personnages. A contrario, entraîner un réseau de neurones à reconnaître ces caractéristiques pour en extraire une nouvelle représentation de voix (p-vecteurs raffinés) et évaluer cette représentation sur la tâche de similarité personnage semble mettre en avant des regroupements plus intéressants. Pour cela, il faut cependant faire attention à la manière de sélectionner le meilleur système par la validation puisque celle-ci a tendance à avoir un pic en termes de performances lorsque k est égal au nombre de classes initiales, ici, les personnages joués.

La question de la sélection du nombre k reste encore ouverte puisque la méthode qui consiste à apprendre une nouvelle représentation de voix et à l'évaluer est très coûteuse. Contrairement aux mesures v-mesure et pureté, l'évaluation par les réseaux siamois semble faire émerger des étiquettes raffinées qui s'écartent de la définition des étiquettes personnages. Cependant, sélectionner k en entraînant 10 réseaux siamois à chaque valeur de k testée a un coût. Dans la prochaine expérience, nous proposons d'utiliser un second corpus pour la phase de regroupement, composé de personnages inconnus à l'entraînement. Nous espérons ainsi que les étiquettes raffinées s'écarteront de la définition des étiquettes personnages.

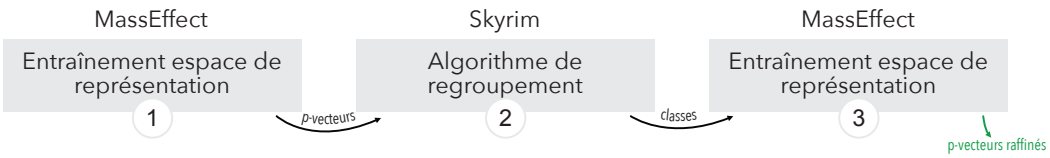


FIGURE 9.2 – Corpus utilisés à chaque étape du processus de raffinage. Les étapes font référence à celles présentées dans la figure 9.1.

9.4.2 Utiliser un corpus secondaire pour l'algorithme de regroupement

Dans une seconde expérience, nous souhaitons vérifier s'il est possible d'utiliser un corpus secondaire pour extraire des étiquettes raffinées, différentes des étiquettes personnage. Pour cela, nous proposons cette fois-ci d'appliquer l'algorithme de regroupement sur les représentations p-vecteurs extraites du corpus Skyrim. Comme présenté dans la figure 9.2, ces p-vecteurs sont appris sur Mass Effect et mettent en exergue le personnage joué dans un enregistrement donné. En utilisant un corpus secondaire composé de personnages inconnus aux p-vecteurs, nous espérons y faire émerger des étiquettes différentes de celles des personnages lors de la phase de regroupement.

Pour l'étape 1, l'extracteur des p-vecteurs est exactement le même que l'expérience précédente. Il est donc entraîné sur Mass Effect avec S_n et les données de Skyrim lui sont complètement inconnues. L'étape de regroupement (i.e., étape 2 dans la figure 9.2) est appliquée sur les p-vecteurs extraits des enregistrements de Skyrim. Le reste du processus d'entraînement et d'évaluation reste le même à savoir : nous extrayons les étiquettes des groupes associées à chaque enregistrement du corpus Mass Effect et entraînons un système de classification (p-vecteurs raffinés avec réseaux siamois) avec le sous-ensemble d'entraînement et de validation de S_n. Nous extrayons ensuite les p-vecteurs raffinés puis les évaluons sur la tâche de similarité personnage sur l'ensemble S_n.

Cette fois-ci, les mesures v-mesure et pureté sont calculées sur les données de Skyrim.

Le tableau 9.3 présente les résultats obtenus. Les scores de v-mesure et de pureté sont bien inférieurs à ceux observés avec Mass Effect car la représentation personnage n'a pas été entrainée sur Skyrim. La pureté est maximale à $k = 2$. Dans ce cas, les groupes séparent les personnages hommes et femmes. Ces étiquettes représentent donc le genre des personnages. La v-mesure corrèle avec le nombre k. Plus k est élevé, plus la v-mesure est élevée. Nous

TABLE 9.3 – Résultats obtenus en utilisant les enregistrements de Skyrim pour l'algorithme de regroupement k-means. Les mesures v-mesure et pureté-K sont calculées sur les regroupements extraits des p-vecteurs, eux-mêmes extraits des enregistrements de Skyrim. Les taux de réussite (lignes "siamois val" et "test") sont obtenus sur les réseaux siamois entraînés avec les p-vecteurs raffinés.

	2	4	6	8	12	18	24	32	48	64
v-mesure	0,23	0,27	0,30	0,33	0,36	0,37	0,38	0,40	0,41	**0,42**
pureté-K	**0,26**	0,23	0,23	0,24	0,24	0,24	0,23	0,23	0,22	0,22
siamese val	0,67	0,72	0,78	0,84	0,87	0,90	0,90	0,90	**0,91**	×
siamese test	0,50	**0,72**	0,65	0,63	0,67	0,62	0,66	0,66	0,62	×

observons donc la valeur maximale de la v-mesure à $k = 64$. Contrairement à $k = 2$, il est difficile de mettre des mots sur ce que représentent ces 64 étiquettes. Nous pouvons noter dans les deux cas, pour la v-mesure et la pureté, que le nombre d'étiquettes s'éloigne du nombre de personnages. Utiliser un second corpus, composé de personnages inconnus à la phase d'entrainement, semble permettre de mettre en lumière des étiquettes raffinées différentes des étiquettes personnages.

9.4.3 Impact de la distance sur l'algorithme k-means

Dans les expériences précédentes, nous avons utilisé la distance euclidienne pour l'algorithme k-means. Ici, nous proposons d'essayer d'autres distances, à savoir la distance cosinus et la distance de Mahalanobis. Ainsi, la distance euclidienne est remplacée par l'une de ces distances en suivant exactement le même protocole que celui présenté dans la sous-section 9.4.2 où les données de Skyrim sont utilisées pour la phase de regroupement.

Le tableau 9.4 présente les résultats obtenus en utilisant la distance cosinus et la distance de Mahalanobis. Pour la distance cosinus, nous observons des valeurs pour les métriques v-mesure et pureté similaires à celles observées avec la distance euclidienne. En effet, la valeur maximale de la v-mesure est observée à $k = 64$ tandis que la valeur maximale de la pureté est observée à $k = 2$. Dans le cas de la distance de Mahalanobis, les valeurs de la v-mesure et de la pureté sont en moyenne supérieures à celles observées pour les distances cosinus et euclidienne.

La valeur maximale de la v-mesure dans le cas de Mahalanobis est aussi observée à $k = 64$. Cependant, la valeur maximale de la pureté est observée à $k = 32$, un résultat différent de ceux obtenus avec les distances euclidiennes et cosinus.

TABLE 9.4 – Résultats obtenus en utilisant les enregistrements de Skyrim pour l'algorithme de regroupement k-means et les distances cosinus et de Mahalanobis. Les mesures v-mesure et pureté-K sont calculées sur les regroupements extraits des p-vecteurs, eux-mêmes extraits des enregistrements de Skyrim. Les taux de réussite (lignes "siamois val" et "test") sont obtenus sur les réseaux siamois entraînés avec les p-vecteurs raffinés.

	2	4	6	8	12	18	24	32	48	64
distance cosine										
v-mesure	0,23	0,27	0,30	0,32	0,35	0,37	0,38	0,39	0,41	**0,41**
pureté-K	**0,25**	0,24	0,23	0,24	0,24	0,24	0,23	0,23	0,23	0,21
siamese val	0,70	0,71	0,78	0,84	0,88	**0,91**	0,90	0,90	×	×
siamese test	0,56	**0,70**	0,63	0,63	0,66	0,67	0,69	0,66	×	×
distance de Mahalanobis										
v-mesure	0,22	0,30	0,31	0,36	0,40	0,43	0,47	0,52	0,54	**0,55**
pureté	0,26	0,25	0,24	0,26	0,29	0,30	0,32	**0,36**	0,36	0,35
siamese val	0,63	0,69	0,74	0,77	0,77	0,78	0,80	**0,80**	×	×
siamese test	0,59	0,59	0,56	0,61	0,54	0,45	0,55	**0,64**	×	×

En n'observant que les métriques v-mesure et pureté, la distance de Mahalanobis semble mieux conserver l'information personnage puisque les valeurs des métriques sont supérieures. Pourtant, lorsque nous évaluons la représentation raffinée, les taux de réussite des systèmes de similarité personnage sont bien plus faibles. La distance euclidienne et la distance cosinus semblent être les métriques les mieux placées pour extraire des étiquettes qui conservent au mieux l'information personnage.

Conclusion

Nous avons présenté le processus de raffinage des p-vecteurs pour extraire des étiquettes de descripteurs de voix. Nous nous sommes intéressés à trois éléments principaux du processus de raffinage : la métrique d'évaluation pour sélectionner la meilleure valeur k, l'importance du corpus utilisé pour l'algorithme de regroupement, l'impact de la distance sur l'algorithme de regroupement.

Les résultats nous ont d'abord montré qu'utiliser un corpus pour l'algorithme de regroupement où les personnages sont connus de la phase d'apprentissage de la représentation empêche la mise en exergue de groupes différents des étiquettes personnages, du moins lorsque nous nous intéressons aux mesures de regroupement v-mesure et pureté. L'évaluation par la similarité personnage semble trouver un nombre d'étiquettes différent du nombre de personnages. Cependant, cette évaluation induit un coût puisqu'elle nécessite

d'entraîner un réseau de neurones.

La seconde expérience a montré qu'utiliser un second corpus pour extraire des étiquettes raffinées permet de mettre en avant des descripteurs vocaux s'écartant de la définition des étiquettes personnages. La pureté a d'ailleurs permis de trouver des étiquettes définissant le genre (i.e. homme et femme). La v-mesure et les réseaux siamois ont cependant mis en exergue des valeurs très élevées de k (i.e. $k > 32$) dont il est difficile de trouver une signification intelligible pour l'humain.

Une dernière expérience a montré que les distances euclidiennes et cosinus permettent à nos réseaux de neurones de mieux conserver l'information personnage en comparaison avec la distance de Mahalanobis.

Pour finir, dans nos expérimentations, l'usage d'un raffinage des étiquettes initiales améliore la représentation personnage. Nous avons en effet observé un taux de réussite à 70% sur le test en sélectionnant le meilleur système sur la validation, en ignorant le système $k = 12$. Nous avons aussi observé un taux de réussite à 72% en sélectionnant le meilleur système sur le test. Ces résultats n'avaient jamais été atteints dans nos précédentes expériences.

Cette première expérience a permis de mettre en lumière la limite de nos données Mass Effect pour l'extraction automatique d'étiquettes raffinées et l'importance du sous-ensemble de validation dont la variabilité personnage est trop proche de l'ensemble d'entrainement. Pour améliorer notre approche, nous proposons de nous intéresser, dans de futurs travaux, à la définition d'un alphabet personnage où chaque personnage est un mot, et chaque descripteur est une lettre. Ainsi, en assemblant des descripteurs, nous pouvons reconnaître des personnages. Nous proposons aussi d'améliorer notre approche en prennant en compte la possibilité d'observer plusieurs étiquettes dans un même enregistrement en travaillant au niveau du segment, voire au niveau de la trame. Une dernière proposition serait l'usage de techniques de Traitement Automatique du Langage Naturel (TALN) pour associer automatiquement des termes à des descripteurs vocaux.

Chapitre 10

Conclusion et perspectives

Sommaire

10.1 Conclusion

Ce manuscrit s'est intéressé à deux questions. La première concerne le lien qu'entretient l'information personnage avec la voix du comédien. Présentée dans la sous-section 3.2.2, la dimension personnage fait référence aux variabilités intra- et inter-personnage que nous pouvons observer dans différents enregistrements de voix. La seconde question concerne la définition de marqueurs vocaux dédiés à la caractérisation du personnage joué.

Pour répondre à ces questions, les notions de Casting Vocal et de voix actée ont d'abord été introduites dans les chapitres 2 et 3. Nous avons vu que le Casting Vocal est un processus complexe qui fait intervenir des critères de décision non seulement liés à la voix, mais aussi à des aspects économiques et physiologiques. Nous avons aussi abordé la complexité de la voix actée en décrivant les nombreux aspects intervenant dans sa production et les nombreuses informations qu'elle véhicule.

Nous avons, dans les chapitres 4 et 5, présenté l'état de l'art de la Recon-

naissance Automatique du Locuteur (RAL) et des systèmes de Casting Vocal Automatique (CVA). Pour les systèmes RAL, nous avons détaillé les trois composants principaux qui les constituent : la paramétrisation, la modélisation, et la décision. Nous avons ensuite vu que les systèmes CVA sont principalement basés sur les techniques de RAL et que la similarité personnage est une brique centrale de leur architecture.

Dans le chapitre 6, nous avons présenté le contexte dans lequel cette thèse s'est déroulée, notamment le projet ANR « The Voice ». Nous avons mis en lumière les difficultés que nous avons rencontrées et qui nous ont empêché de récolter des données exploitables et de mettre en place des protocoles d'annotation par les DA (DA). Enfin, les deux questions principales soulevées par cette thèse ont été détaillées.

Dans le chapitre 7, nous avons proposé un protocole de neutralisation de l'information personnage dans le but de mettre en exergue la partie de cette information qui est indépendante du locuteur (IPIL). Pour cela, nous avons proposé une expérience avec un protocole intitulé « original », dans lequel les associations locuteur-personnage sont celles du doubleur et de son comédien d'origine, et un protocole intitulé « modifié », dans lequel les associations locuteur-personnage sont erronées. En trompant son apprentissage, le système de similarité personnage (ou de représentation personnage) n'est plus amené à modéliser une information personnage mais à baser ses décisions sur sa seule capacité à associer deux locuteurs. En comparant les résultats obtenus avec un système original et un système modifié, nous avons ainsi pu mettre en exergue l'IPIL. Nos résultats ont montré que, bien qu'il existe une part d'IPIL, la voix d'un personnage semble essentiellement se caractériser par celle de son ou sa comédienne.

Puisque la voix du comédien semble être une brique importante de la construction vocale d'un personnage, nous nous sommes intéressés dans le chapitre 8 à l'influence du pré-entraînement locuteur (extracteur de séquence) sur la construction d'une représentation personnage. Nous avons réalisé une expérience où différentes configurations de l'extracteur de séquence étaient utilisées pour entraîner des représentations personnage. Certaines donnaient plus de pouvoir au pré-entrainement locuteur tandis que d'autres limitaient ce pouvoir. Les résultats ont montré que l'usage d'un extracteur de séquence orienté locuteur ne semble pas introduire de biais et contribue même à améliorer les capacités de généralisation de la représentation personnage.

Les expériences que nous avons menées souffrent de plusieurs limites. Le

contexte de la caractérisation du locuteur est restreint au personnage qu'il ou elle joue puisque nous ne travaillons que dans le domaine des enregistrements de voix actées. Pour mieux caractériser la voix d'un locuteur et étudier plus précisément le lien locuteur/personnage, il serait utile d'élargir ce contexte et de travailler aussi avec des enregistrements de parole non actée produits par les comédiens présents dans le corpus. Une autre limite est la taille des corpus de données avec lesquels nous travaillons. La variabilité personnage du corpus Mass Effect est relativement faible puisque nous n'avons que 16 personnages exploitables. De plus, ces personnages ne sont joués que par un comédien dans chaque langue. Ces échantillons de voix étant rares, la construction d'un corpus de test où les personnages sont joués par différents comédiens dans une même langue serait une bonne base pour mieux comprendre le comportement des modèles de représentation personnage. Malgré tout, ces expériences nous amènent à conclure que la dimension personnage de la voix actée semble intrinsèquement liée au locuteur qui interprète le personnage. En perspectives, nous pouvons chercher à mesurer l'IPIL présent dans un enregistrement pour caractériser différents styles de jeux d'acteur. Pour cela, nous pourrions notamment utiliser des techniques de désenchevêtrement. Nous détaillons ces propositions dans la section 10.2.

Dans le chapitre 9, nous nous sommes attaqués à la problématique de l'extraction non supervisée d'étiquettes (descripteurs) caractérisant le personnage joué, sans vérité terrain. L'approche proposée consiste à utiliser un algorithme non-supervisé dans un espace personnage pour y faire émerger des facteurs de variabilité que nous exprimons finalement sous la forme d'étiquettes. Les expériences ont montré qu'il est nécessaire, dans le contexte de nos données, d'exploiter un second corpus (non annoté) pour que des étiquettes qui diffèrent de la définition de personnage puissent émerger. Dans ce manuscrit, nous avons identifié différentes limites inhérentes à notre approche. Il est par exemple difficile d'expliquer le sens de ces étiquettes et de les associer à des termes intelligibles pour un expert. La mise en place d'expériences subjectives pourrait donc être une piste de recherche pour associer des mots à ces étiquettes. Ensuite, nous supposons dans nos travaux que chaque enregistrement peut correspondre à une étiquette. Il semble cependant plus approprié de supposer que différentes étiquettes peuvent correspondre à différents sous-segments d'un enregistrement. L'une des perspectives de ce travail pourrait donc être la définition d'un alphabet personnage que nous détaillons dans les perspectives de ce chapitre.

Ci-après, nous présentons différentes perspectives au travail présenté dans

ce manuscrit. D'abord, nous présentons les perspectives à court-terme. Ces perspectives dérivent directement des expériences que nous avons réalisées et toutes les données nécessaires à leur mise en œuvre sont disponibles. Puis, nous présentons les perspectives à long-terme. Ces perspectives nécessitent des données encore non présentes ou font référence à des modèles difficiles à formaliser et à évaluer à ce jour. Dans une troisième section, nous décrivons des perspectives de collaboration avec l'industrie du cinéma, notamment en détaillant des solutions pour récolter de nouvelles données exploitables.

10.2 Perspectives à court-terme

Identifier le style de jeu en mesurant les taux d'information personnage et d'IPIL

En perspectives au travail que nous avons présenté, il semble intéressant d'explorer la quantification de l'information personnage et de l'IPIL pour identifier le style de jeu du comédien. Dans le chapitre 7, nous avons vérifié la présence d'IPIL sur un ensemble d'enregistrements. Réussir à quantifier le volume de cette information dans un segment audio permettrait éventuellement d'identifier différents styles de jeux de comédiens. Ainsi, nous pourrions poser deux hypothèses :

— **Le comédien ne transforme pas sa voix :** un volume faible d'IPIL signifie que le comédien a peu modifié sa voix et qu'il est reconnaissable par rapport à sa voix « neutre ».
— **Le comédien transforme sa voix :** un volume élevé d'IPIL signifie que le comédien a modifié sa voix « neutre » et caché certains de ses marqueurs caractéristiques pour interpréter son personnage.

L'alphabet personnage

Les travaux présentés dans ce manuscrit se sont attaqués au problème de la découverte de descripteurs de voix caractéristiques du personnage dans les enregistrements audio. Dans le cadre de nos recherches, ces travaux répondaient à deux objectifs :

— Participer à l'élaboration de la taxonomie de la voix actée
— Découvrir un nouveau vocabulaire pour construire des argumentaires de recommandation

La modélisation d'un alphabet personnage est une perspective à explorer. Chaque lettre de cet alphabet est une caractéristique personnage et la combi-

naison de plusieurs de ces lettres constitue un mot, i.e. un personnage.

Dans ce travail, différentes hypothèses sont à vérifier :

— L'apparition de certaines marqueurs vocaux d'un personnage dépend du contexte de production de sa voix. Ainsi, d'un enregistrement ou d'un ensemble d'enregistrements à un autre, certaines lettres (marqueurs) personnage peuvent apparaître ou disparaître. Il existerait alors une grammaire qui permettrait d'aider à constituer le personnage.
— Certaines lettres sont partagées par différents personnages.
— Certaines lettres que nous pourrions intituler de « neutres », ne caractérisent pas le personnage joué mais aident les règles de grammaire à contextualiser l'enregistrement.
— L'ordre des lettres dans un mot n'a pas d'importance.

Caractériser le jeu d'acteur par l'analyse de la variabilité des émotions et des attitudes chez les comédiens

Le jeu d'acteur est la façon dont un acteur (ou comédien) interprète un rôle. Le jeu d'acteur semble naturellement varier d'un individu à l'autre et dépendre de nombreux facteurs liés au comédien, e.g. expérience ou physiologie. Nous supposons qu'une partie du jeu de l'acteur se caractérise par la manière dont le comédien joue des émotions et des attitudes pour les rendre manifestes chez le spectateur. Pour vérifier cette hypothèse, nous pourrions mettre en place une expérience qui consiste à entraîner un système à reconnaitre le comédien ou le personnage joué à partir d'une représentation dédiée à la caractérisation de l'émotion ou de l'attitude jouée que nous nommons « représentation émotionnelle de la voix ». La validation de cette hypothèse donnerait des pistes d'amélioration pour l'espace personnage. Elle indiquerait par exemple si l'usage d'une représentation émotionnelle de la voix est utile en complément de la représentation locuteur (extracteur de séquence), et si elle aiderait à caractériser les aspects émotionnels du jeu d'acteur.

10.3 Perspectives à long-terme

La palette vocale

Un acteur (comédien) joue différents personnages, chacun ayant éventuellement un caractère différent. Pour rendre manifeste le caractère de chacun de ses personnages, le comédien adapte sa voix à son rôle. Le domaine de réalisation de la voix d'un comédien est ce que nous nommons la palette vocale.

En perspective de nos travaux, nous pouvons nous intéresser à la modélisation de la palette vocale d'un comédien. À un niveau acoustique, elle représente l'ensemble des valeurs acoustiques que la voix peut atteindre, e.g. la tessiture de voix. À un niveau personnage, elle représente l'ensemble des personnages que la voix d'un comédien peut interpréter.

En appliquant le concept de palette vocale à nos espaces personnages, nous pouvons espérer créer une carte de chaleur représentant le domaine de réalisation de la voix d'un comédien – d'un point de vue personnage. Cependant, une limite clairement identifiée à ce travail est la suivante : les dimensions locuteur et personnage se confondent dans la mesure où, dans nos données, un comédien ne joue généralement qu'un personnage.

Pour espérer construire un modèle de la palette vocale – personnage – d'un comédien, il est nécessaire d'employer des données de comédiens jouant différents personnages. Ainsi, nous pourrions entraîner un système à prédire la palette vocale d'un comédien – une distribution dans un espace personnage par exemple – à partir d'un ou plusieurs exemples, i.e. des enregistrements.

Nous pourrions aussi utiliser un principe simple pour aider à la construction de cette palette vocale. Si deux comédiens A et B jouent deux personnages différents mais vocalement similaires, il est probable que le comédien A puisse jouer les mêmes personnages que B et vice versa. La distance, dans un espace locuteur, entre les différentes voix de personnages du corpus peut ainsi éventuellement contribuer au calcule de la distribution de probabilité de la palette vocale d'un comédien.

Pour finir, combiné à l'alphabet personnage, la palette vocale nous permettrait de connaître tous les marqueurs vocaux – personnages – qu'un comédien peut jouer. Avec un tel modèle, nous pourrions espérer, à partir d'un ensemble d'enregistrements, connaître les différents mots personnages qu'un comédien peut jouer.

La subjectivité dans un choix artistique

Une autre question à poser concerne le caractère subjectif de la tâche de sélection des comédiens de doublage. La sélection de comédiens est un choix artistique et est intrinsèquement subjectif. Il est donc légitime de supposer que tous les DA ne font pas les mêmes choix face aux mêmes candidats. Un système de recommandation évalué humainement peut être évalué positivement par un DA et négativement par un autre.

Partant de ce postulat, il semble évident qu'utiliser un modèle généraliste,

appris à partir des décisions et des retours utilisateurs de différents DA, n'est certainement pas la bonne solution. Cependant, apprendre *from scratch* un modèle différent pour chaque DA ne semble pas non-plus la solution la plus appropriée car le processus d'apprentissage peut nécessiter un grand nombre d'annotations et donc coûter un temps précieux au DA. De plus, bien que chaque DA ait ses spécificités, il est raisonnable de penser qu'ils partagent toutes et tous un socle commun lié à des aspects culturels et professionnels et à leur métier. Sous cette supposition, entraîner un modèle généraliste et l'adapter à chaque DA semble la solution optimale.

Les modèles multimodaux

La thèse que nous avons présentée tourne autour du traitement de la voix et plus particulièrement de la voix actée. Cependant, de nombreuses autres modalités peuvent être exploitées. Certains DA affirment savoir si un comédien correspond à ses attentes juste en le voyant. Il est alors à supposer qu'il existe un fort a priori physiologique qui intervient dans sa décision. Un système de similarité personnage pourrait baser ses décisions sur les données physiologiques du candidat : âge, taille, etc. Le système pourrait aussi se baser sur un modèle du candidat composé à la fois des données physiologiques et vocales.

Entraîner ce type de système pose notamment la question de la construction d'un espace multimodal où, par exemple, des enregistrements audio et des images – voire du texte – sont projetés dans un même espace vectoriel. L'usage d'un modèle pré-entraîné tel que *data2vec* ([Bae+22]) peut notamment être envisagé. Ce modèle apprend de manière auto-supervisée (self-supervised) à projeter dans un même espace du texte, de l'audio et des images.

En partant de l'hypothèse que les directeurs de casting se basent sur les aspects physiologiques du comédiens pour le sélectionner, nous pouvons nous demander si les critères physiologiques ont une place plus importante que les critères vocaux dans les décisions du DA. À l'instar du modèle du candidat, nous pouvons imaginer modéliser le personnage à partir à la fois de ses enregistrements vocaux mais aussi de ses caractéristiques physiologiques. Ce travail pose cependant différentes questions notamment la mise en place de critères physiologiques standards qui couvrent à la fois le cas des personnages humains mais aussi le cas des personnages non-humains.

10.4 Perspectives de collaboration avec l'industrie du cinéma

Récolte de données en phase critique d'un projet de doublages

En collaborant avec Dubbing Brothers, nous avons appris que les les données enregistrées lors du processus de doublage suivent le cycle de vie suivant :

1. **Phase critique :** toutes les données sont d'abord enregistrées et conservées sur un serveur. Ces données comprennent des informations sur les acteurs, la bande rythmo et les projets de logiciels d'édition audio propriétaires contenant les voix séparées des sons d'ambiance et des autres voix. De plus, la Version Internationale (VI) et la Version originale (VO) sont disponibles durant cette phase.
2. **Archivage :** les données sont ensuite archivées et doivent respecter deux contraintes. La première consiste à supprimer les informations personnelles des comédiens pour respecter le RGPD qui contraint l'entreprise à conserver ces données durant maximum 6 mois. Les noms des comédiens, leur âge ou toute autre information personnelle est supprimée. La seconde contrainte est l'espace disponible sur le serveur d'archives. Pour optimiser cet espace, certaines pistes audio sont supprimées et d'autres sont fusionnées entre elles. Les bandes rythmo sont aussi supprimées. La Version originale est aussi supprimée.

Les données archivées sont disponibles en grandes quantités, plus de cent Téraoctets, mais les pistes des enregistrements sont généralement fusionnées en une seule. Ainsi, il est impossible de dissocier facilement les enregistrements de voix des musiques et autres sons d'ambiance. De plus, ces données ne sont presque pas annotées et il est impossible de récupérer des informations telles que l'identité du doubleur, le numéro d'épisode etc.

Pour espérer accéder des données exploitables, il est donc nécessaire de mettre en place un protocole avec une entreprise partenaire pour récolter les données durant la phase critique – l'enregistrement – du processus du doublage.

Annotations des DA pour évaluer ou améliorer l'espace de projection des enregistrements audio

L'évaluation (objective) actuelle de nos systèmes ne permet pas d'assurer leur capacité à modéliser les décisions des DA dans la mesures où la tâche est difficile à formaliser. Comme nous l'avons montré dans ce manuscrit, les DA eux-mêmes ne savent pas expliquer quels critères sont intervenus lors de leurs décisions. De plus, les données sont uniquement constituées de candidats sélectionnés pour un rôle. Pourtant, un candidat non sélectionné peut tout de même

TABLE 10.1 – Tableau des modèles à entraîner pour les perspectives.

ID	Résumé
1	similarité personnage par représentation émotionnelle
2	similarité personnage par physiologie
3	système de recommandation par retours utilisateurs
4	évaluation avancée de l'information personnage
5	modélisation de la palette vocale
6	alphabet personnage

correspondre aux attentes du DA. Nous avons alors deux axes d'améliorations :

1. L'annotation de données où il est spécifié si la voix d'un comédien semble correspondre au rôle à doubler, même si ce comédien joue un autre personnage dans l'enregistrement écouté.
2. La récolte de retours des utilisateurs d'un système de recommandation pour améliorer l'espace de projection des enregistrements audio.

Un personnage, plusieurs rôles

Dans nos expériences, les notions de personnage et de rôle se confondent puisque, pour chaque personnage, nous possédons uniquement les enregistrements d'un rôle. Il arrive pourtant que certains personnages soient joués par différents comédiens en fonction des rôles, e.g. la voix de Will Smith en français n'est pas toujours interprétée par le même comédien. Nous pouvons supposer qu'utiliser des enregistrements d'un personnage joué par plusieurs comédiens peut contribuer à la réduction de l'influence du locuteur pour ne conserver que les informations dédiées à la caractérisation du personnage. Ces données étant rares, elles peuvent aussi être utilisées pour améliorer l'évaluation de la similarité personnage en vérifiant si les voix de deux comédiens jouant le même personnage dans la même langue sont proches dans l'espace personnage.

10.5 Planification des perspectives

Pour finir, nous présentons ci-après une planification à moyen-terme d'expériences pouvant être menées à la suite de nos recherches. Pour réaliser ces expériences, les modèles à entraîner sont listés dans le tableau 10.1. Les données nécessaires à leur entraînement sont listées dans le tableau 10.2. Les expériences et les données sont liées par un identifiant(ID).

Dans le tableau 10.2, nous associons chaque modèle (et ses données) à un coût pour chaque entité pouvant intervenir dans la récole des données. Le laboratoire (« coût labo. ») est l'entité ayant les compétences de recherche tandis que le partenaire (« coût part. ») est une entreprise ou un organisme

TABLE 10.2 – Tableau des données à récolter pour les perspectives.

ID	Intitulé	Provenance	Coût labo.	Coût part.
1	Émotions	Corpus	Faible	Zéro
2	Photo de visage d'acteurs	Internet	Moyen	Zéro
3	Retours utilisateurs	Partenaire	Fort	Fort
4	Voix personnages multi-comédien	Partenaire	Fort	Fort
5	Voix comédiens multi-personnage	Partenaire	Fort	Fort
6	Annotations humaines	Partenaire	Fort	Fort

réalisant le doublage ou ayant les données recherchées. Par le biais d'un code couleur, nous associons chaque modèle à un risque. Ci-dessous, les échelles de coût et risque sont détaillées :

L'échelle des coûts est la suivante :

— **Zéro :** pas de coût pour l'entité (laboratoire ou partenaire).

— **Faible :** les données existent déjà.

— **Moyen :** les données existent mais nécessitent des traitements ou des annotations.

— **Fort :** les données n'existent pas et leur récole nécessite un fort investissement.

L'échelle des risques est la suivante :

— **faible (vert) :** vision presque parfaite des tâches à réaliser pour récolter les données et très peu de risques.

— **moyen (orange) :** une bonne vision des tâches à réaliser pour récolter les données avec peu d'inconnues.

— **fort (rouge) :** beaucoup d'inconnues et une forte probabilité pour que la récolte des données ne réussisse pas, e.g. données inexistantes ou contraintes par la RGPD.

Glossaire

ANR Agence Nationale de la Recherche ix, 57, 59

CVA Casting Vocal Automatique viii, 28, 45–48, 50, 52, 108, 119

DA Directeur Artistique iii, iv, 8, 10, 11, 13, 14, 16, 22, 46, 62, 108, 113, 114

DCT Discrete Cosine Transform 35

DET Detection Error Tradeoff 37

EER Equal Error Rate viii, 31, 37, 51, 66, 84, 100

EM Expectation Maximization 38, 39

FA Fausses Alarmes 37

FAR False Acceptance Rate 36, 37

FFT Fast Fourier Transform 35

FR Faux Rejets 37

FRR False Rejection Rate 36, 37

GMM Gaussian Mixture Model 38, 39

HALDE Haute Autorité de Lutte contre les Discriminations et pour l'Égalité 12

IPIL Information Personnage Indépendante du Locuteur ix, x, 53, 61, 63–65, 71–75, 77, 78, 82, 85, 87, 88, 108–110, 121

IRCAM Institut de Recherche et Coordination Acoustique/Musique 57, 59, 62

LDA Linear Discriminent Analysis 40

LIA Laboratoire Informatique d'Avignon 59

LSTM Long Short-Term Memory 41

MAP Maximum a posteriori 39

MFCC Mel Frequency Cepstral Coefficients 34, 35, 83, 119

PLDA Probabilistic Linear Discriminant Analysis 42, 94

RAL Reconnaissance Automatique du Locuteur viii, 28, 31–36, 38–40, 42, 43, 94, 107, 108

RGPD Règlement Général sur la Protection des Données 60, 114, 116

SVM Support Vector Machine 39

TDNN Time Delay Neural Network 41, 42, 80, 119

UBM Universal Background Model 39, 40

VAD Vocal Activity Detection 35

VI Version Internationale 9, 114

VO Version originale 9, 46, 60, 114

Liste des figures

Liste des tableaux

Bibliographie

Ouvrages de référence

[AC17] Jean-Julien Aucouturier et Clément Canonne. « Musical friends and foes : The social cognition of affiliation and control in improvised interactions ». In : *Cognition* 161 (avr. 2017), p. 94-108. ISSN : 00100277. DOI : 10.1016/j.cognition.2017.01.019.

[ACLT00] Roland Auckenthaler, Michael Carey et Harvey Lloyd-Thomas. « Score Normalization for Text-Independent Speaker Verification Systems ». In : *Digital Signal Processing* 10.1-3 (jan. 2000), p. 42-54. ISSN : 10512004. DOI : 10.1006/dspr.1999.0360.

[Aga19] Abien Fred Agarap. « Deep Learning using Rectified Linear Units (ReLU) ». In : *arXiv :1803.08375 [cs, stat]* (fév. 2019).

[AK+20] Musab T. S. Al-Kaltakchi, Raid Rafi Omar Al-Nima, Mahmood Alfathe et Mohammed A. M. Abdullah. « Speaker Verification Using Cosine Distance Scoring with i-vector Approach ». In : *2020 International Conference on Computer Science and Software Engineering (CSASE)*. 2020, p. 157-161. DOI : 10.1109/CSASE48920.2020.9142088.

[AO03] Hazem Munawer Al-Otum. « Morphological operators for color image processing based on Mahalanobis distance measure ». In : *Optical Engineering* 42.9 (2003), p. 2595 -2606. DOI : 10.1117/1.1594727.

[Api+21] Andrea Apicella, Francesco Donnarumma, Francesco Isgrò et Roberto Prevete. « A survey on modern trainable activation functions ». en. In : *Neural Networks* 138 (juin 2021), p. 14-32. ISSN : 08936080. DOI : 10.1016/j.neunet.2021.01.026.

[Bae+22] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu et Michael Auli. « data2vec : A General Framework for Self-supervised Learning in Speech, Vision and Language ». In : *CoRR* abs/2202.03555 (2022). arXiv : 2202.03555.

[BCP94] Frédéric Bimbot, Gérard Chollet et Andrea Paoloni. « Assessment methodology for speaker identification and verification systems - an overview of SAM-a esprit project 6819 - task 2500 ». In : *Proc. ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*. 1994, p. 75-82.

[BD15] Ben Barsties et Marc De Bodt. « Assessment of voice quality : Current state-of-the-art ». In : *Auris Nasus Larynx* 42.3 (2015), p. 183-188. issn : 0385-8146. doi : https://doi.org/10.1016/j.anl.2014.11.001.

[Bes+00] L. Besacier, S. Grassi, A. Dufaux, M. Ansorge et F. Pellandini. « GSM speech coding and speaker recognition ». In : *2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100)*. T. 2. Istanbul, Turkey : IEEE, 2000, p. II1085-II1088. isbn : 978-0-7803-6293-2. doi : 10.1109/ICASSP.2000.859152.

[BK17] Waad Ben Kheder. « Reconnaissance du locuteur en milieux difficiles ». Theses. Université d'Avignon, juill. 2017.

[Boe+17] Louis-Jean Boe, Frédéric berthommier, Thierry Legou, Guillaume Captier, Caralyn Kemp, Thomas R. Sawallis, Yannick Becker, Arnaud Rey et Joel Fargot. « Evidence of a Vocalic Proto-System in the Baboon (Papio papio) Suggests Pre-Hominin Speech Precursors ». In : *PLOS ONE* 12 (1 jan. 2017), e0169321. issn : 1932-6203. doi : 10.1371/journal.pone.0169321.

[Bon+03] Jean-Francois Bonastre, Frederic Bimbot, Louis-Jean Boe, Joseph P Campbell, Douglas A Reynolds et Ivan Magrin-Chagnolleau. « Person Authentication by Voice : A Need for Caution ». In : (2003), p. 4.

[Bon19] Jean-François Bonastre. « Representation Learning for Under-defined Tasks ». In : *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*. Sous la dir. d'Ingela Nyström, Yanio Hernández Heredia et Vladimir Milián Núñez. Cham : Springer International Publishing, 2019, p. 42-47.

[BP06] Matthew M. Botvinick et David C. Plaut. « Short-term memory for serial order : A recurrent neural network model. » In : *Psychological Review* 113.2 (2006), p. 201-233. issn : 1939-1471, 0033-295X. doi : 10.1037/0033-295X.113.2.201.

[Bre17] Hervé Bredin. « TristouNet : Triplet loss for speaker turn embedding ». In : *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mars 2017, p. 5430-5434. doi : 10.1109/ICASSP.2017.7953194.

[BW19] Sebastian Bock et Martin Weis. « A Proof of Local Convergence for the Adam Optimizer ». In : *2019 International Joint Conference on Neural Networks (IJCNN)*. Budapest, Hungary : IEEE, juill. 2019, p. 1-8. isbn : 978-1-72811-985-4. doi : 10.1109/IJCNN.2019.8852239.

[CC10] Bernard Lortat-Jacob Maria Manca Claude Calame Florence Dupont. 2010.

[CG18] Yu-An Chung et James Glass. *Speech2Vec : A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech*. 2018. arXiv : 1803.08976 [cs.CL].

[Cha+17] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang et Chunhong Pan. « Deep Adaptive Image Clustering ». In : *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2017.

[Chi21] Davide Chicco. « Siamese Neural Networks : An Overview ». In : *Artificial Neural Networks*. Sous la dir. d'Hugh Cartwright. T. 2190. Series Title : Methods in Molecular Biology. New York, NY : Springer US, 2021, p. 73-94. isbn : 978-1-07-160825-8 978-1-07-160826-5. doi : 10.1007/978-1-0716-0826-5_3.

[CHL05] Sumit Chopra, Raia Hadsell et Yann LeCun. « Learning a Similarity Metric Discriminatively, with Application to Face Verification ». In : *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. T. 1. IEEE, 2005, p. 539-546. isbn : 0-7695-2372-2.

[CM04] N. Campbell et Parham Mokhtari. « Voice Quality : the 4 th Prosodic Dimension ». In : 2004.

[CNZ18] J. S. Chung, A. Nagrani et A. Zisserman. « VoxCeleb2 : Deep Speaker Recognition ». In : *Interspeech*. 2018.

[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 (sept. 2007), p. 2095-2103. issn : 1558-7916. doi : 10.1109/TASL.2007.902758.

[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 (mai 2011), p. 788-798. issn : 1558-7916, 1558-7924. doi : 10.1109/TASL.2010.2064307.

[Deh+11b] Najim Dehak, Pedro Torres-Carrasquillo, Douglas Reynolds et R. Dehak. « Language Recognition via I-Vectors and Dimensionality Reduction ». In : *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (jan. 2011), p. 857-860.

[Deu18] Michel Deudon. « Learning semantic similarity in a continuous space ». In : *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. 2018, p. 994-1005.

[Dow04] Laura J. Downing. « What African languages tell us about accent typology ». In : *ZAS Papers in Linguistics* 37 (jan. 2004), p. 101-136. issn : 1435-9588. doi : 10.21248/zaspil.37.2004.247.

[DP18] Rohan Kumar Das et S. R.Mahadeva Prasanna. « Speaker Verification from Short Utterance Perspective : A Review ». In : *IETE Technical Review (Institution of Electronics and Telecommunication Engineers, India)* 35.6 (2018), p. 599-617. ISSN : 09745971.

[EF17] K. Ezzine et M. Frikha. « A comparative study of voice conversion techniques : A review ». In : *International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*. 2017, p. 1-6.

[ER02] Penelope Eckert et John R. Rickford, éd. *Style and Sociolinguistic Variation*. 1re éd. Cambridge University Press, jan. 2002. ISBN : 978-0-521-59191-1 978-0-521-59789-0 978-0-511-61325-8. DOI : 10.1017/CBO9780511613258.

[FH11] Xing Fan et John H. L. Hansen. « Speaker Identification Within Whispered Speech Audio Streams ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.5 (juill. 2011), p. 1408-1421. ISSN : 1558-7916, 1558-7924. DOI : 10.1109/TASL.2010.2091631.

[Fle84] James Emil Flege. « The detection of French accent by American listeners ». In : *The Journal of the Acoustical Society of America* 76.3 (sept. 1984), p. 692-707. ISSN : 0001-4966. DOI : 10.1121/1.391256.

[Fra+17] Eduard Frant¸I, Frant¸ Frant¸I, Ioan Ispas, Voichita Dragomir, Monica Dasc, ˘ Alu, Elteto Zoltan et Ioan Cristian Stoica. *Voice Based Emotion Recognition with Convolutional Neural Networks for Companion Robots*. 2017, p. 222-240.

[FSS10] Luciana Ferrer, Nicolas Scheffer et Elizabeth Shriberg. « A comparison of approaches for modeling prosodic features in speaker recognition ». In : *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. Dallas, TX, USA : IEEE, 2010, p. 4414-4417. ISBN : 978-1-4244-4295-9. DOI : 10.1109/ICASSP.2010.5495632.

[GBC02] Christer Gobl, Eva Bennett et Ailbhe Ni Chasaide. « EXPRESSIVE SYNTHESIS : HOW CRUCIAL IS VOICE QUALITY ? » In : sept. 2002. DOI : 10.1109/WSS.2002.1224380.

[GDL18] Adrien Gresse, Richard Dufour et Vincent Labatut. « Mesure de similarité fondée sur des réseaux de neurones siamois pour le doublage de voix ». In : *XXXIIe Journées d'Études sur la Parole*. 2018, p. 10-18. DOI : http://dx.doi.org/10.21437/JEP.2018-2.

[Ger] Greg Germain. *Tweet de Greg Germain*.

[Gid+17] John Gideon, Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis et Emily Mower Provost. « Progressive neural networks for transfer learning in emotion recognition ». In : *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2017, p. 1098-1102. arXiv : 1706.03256.

[GL94] J.-L. 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. doi : 10.1109/89.279278.

[Gre+10] Craig S. Greenberg, Alvin F. Martin, Linda Brandschain, Joseph P. Campbell, Christopher Cieri, George R. Doddington et John J. Godfrey. « Human Assisted Speaker Recognition In NIST SRE10 ». In : *Odyssey*. 2010.

[Gre+17] Adrien Gresse, Mickael Rouvier, Richard Dufour, Vincent Labatut et Jean francois Bonastre. « Acoustic Pairing of Original and Dubbed Voices in the Context of Video Game Localization ». In : *Interspeech*. 2017, p. 2839-2843. doi : 10.21437/Interspeech.2017-1311.

[Gre+19] Adrien Gresse, Mathias Quillot, Richard Dufour, Vincent Labatut et Jean francois Bonastre. « Similarity metric based on siamese neural networks for voice casting ». In : *ICASSP*. 2019, p. 6585-6589. isbn : 978-1-4799-8131-1. doi : 10.1109/ICASSP.2019.8683178.

[Gre20] Adrien Gresse. « L'Art de la Voix : Caractériser l'information vocale dans un choix artistique ». Thèse de doct. Avignon : Avignon Université, fév. 2020.

[Gre+20] Adrien Gresse, Mathias Quillot, Richard Dufour et Jean francçois Bonastre. « Learning Voice Representation Using Knowledge Distillation For Automatic Voice Casting ». In : *Interspeech*. 2020, p. 160-164. doi : doi:10.21437/Interspeech.2020-2236.

[Han96] John H.L. Hansen. « Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition ». In : *Speech Communication* 20.1-2 (nov. 1996), p. 151-173. issn : 01676393. doi : 10.1016/S0167-6393(96)00050-7.

[HCL06a] R. Hadsell, S. Chopra et Y. LeCun. « Dimensionality Reduction by Learning an Invariant Mapping ». In : *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. T. 2. 2006, p. 1735-1742.

[HCL06b] R. Hadsell, S. Chopra et Y. LeCun. « Dimensionality Reduction by Learning an Invariant Mapping ». In : *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. T. 2. 2006, p. 1735-1742.

[Héb08] Matthieu HÉBERT. « Text-Dependent Speaker Recognition ». In : *Springer Handbook of Speech Processing*. Sous la dir. de Jacob BENESTY, M. Mohan SONDHI et Yiteng Arden HUANG. Berlin, Heidelberg : Springer Berlin Heidelberg, 2008, p. 743-762. ISBN : 978-3-540-49127-9. DOI : 10.1007/978-3-540-49127-9_37.

[Hei+15] Georg HEIGOLD, Ignacio MORENO, Samy BENGIO et Noam SHAZEER. « End-to-End Text-Dependent Speaker Verification ». In : *arXiv :1509.08062 [cs]* (sept. 2015).

[HH15] John H.L. HANSEN et Taufiq HASAN. « Speaker Recognition by Machines and Humans : A tutorial review ». In : *IEEE Signal Processing Magazine* 32.6 (nov. 2015), p. 74-99. ISSN : 1053-5888. DOI : 10.1109/MSP.2015.2462851.

[Hin+12] Geoffrey HINTON, Li DENG, Dong YU, George E. DAHL, Abdelrahman MOHAMED, Navdeep JAITLY, Andrew SENIOR, Vincent VANHOUCKE, Patrick NGUYEN, Tara N. SAINATH et Brian KINGSBURY. « Deep Neural Networks for Acoustic Modeling in Speech Recognition : The Shared Views of Four Research Groups ». In : *IEEE Signal Processing Magazine* 29.6 (2012), p. 82-97. DOI : 10.1109/MSP.2012.2205597.

[HLM19] Dan HENDRYCKS, Kimin LEE et Mantas MAZEIKA. « Using Pre-Training Can Improve Model Robustness and Uncertainty ». In : *Proceedings of the 36th International Conference on Machine Learning*. Sous la dir. de Kamalika CHAUDHURI et Ruslan SALAKHUTDINOV. T. 97. PMLR, 2019, p. 2712-2721.

[HNS17] John H. L. HANSEN, Mahesh Kumar NANDWANA et Navid SHOKOUHI. « Analysis of human scream and its impact on text-independent speaker verification ». In : *The Journal of the Acoustical Society of America* 141.4 (avr. 2017), p. 2957-2967. ISSN : 0001-4966. DOI : 10.1121/1.4979337.

[HS63] Wm. A. HARGREAVES et John A. STARKWEATHER. « Recognition of Speaker Identity ». In : *Language and Speech* 6.2 (avr. 1963), p. 63-67. ISSN : 0023-8309, 1756-6053. DOI : 10.1177/002383096300600202.

[HV09] J. HANSEN et V. VARADARAJAN. « Analysis and Compensation of Lombard Speech Across Noise Type and Levels With Application to In-Set/Out-of-Set Speaker Recognition ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 17.2 (fév. 2009), p. 366-378. ISSN : 1558-7916. DOI : 10.1109/TASL.2008.2009019.

[HVD15] Geoffrey HINTON, Oriol VINYALS et Jeff DEAN. *Distilling the Knowledge in a Neural Network*. 2015. arXiv : 1503.02531 [stat.ML].

[Iid+03] Akemi IIDA, Nick CAMPBELL, Fumito HIGUCHI et Michiaki YASUMURA. « A corpus-based speech synthesis system with emotion ». In : *Speech communication* 40.1-2 (2003), p. 161-187.

[Ios+18] Elias Iosif, Ioannis Klasinas, Georgia Athanasopoulou, Elisavet Palogiannidi, Spiros Georgiladakis, Katerina Louka et Alexandros Potamianos. « Speech understanding for spoken dialogue systems : From corpus harvesting to grammar rule induction ». In : *Computer Speech and Language* 47 (2018), p. 272-297. issn : 10958363.

[Jou21] Pierre Jourlin. *La boîte translucide : un éclairage sur l'intelligence artificielle*. Editions universitaires d'Avignon, 2021.

[JSW07] Qin Jin, Tanja Schultz et Alex Waibel. « Far-Field Speaker Recognition ». In : *IEEE Transactions on Audio, Speech and Language Processing* 15.7 (sept. 2007), p. 2023-2032. issn : 1558-7916. doi : 10.1109/TASL.2007.902876.

[Kan+11] Ahilan Kanagasundaram, Robbie Vogt, David Dean, Sridha Sridharan et Michael Mason. « i-vector based speaker recognition on short utterances ». In : *Interspeech 2011*. ISCA, août 2011, p. 2341-2344. doi : 10.21437/Interspeech.2011-58.

[KB97] Mark Kuitert et Lou Boves. « Speaker verification with GSM coded telephone speech ». In : *EUROSPEECH*. 1997.

[KBD05] P. Kenny, G. Boulianne et P. Dumouchel. « Eigenvoice modeling with sparse training data ». In : *IEEE Transactions on Speech and Audio Processing* 13.3 (mai 2005), p. 345-354. issn : 1063-6676. doi : 10.1109/TSA.2004.840940.

[Ken+07] Patrick Kenny, Gilles Boulianne, Pierre Ouellet et Pierre Dumouchel. « Joint Factor Analysis Versus Eigenchannels in Speaker Recognition ». In : *IEEE Transactions on Audio, Speech and Language Processing* 15.4 (mai 2007), p. 1435-1447. issn : 1558-7916. doi : 10.1109/TASL.2006.881693.

[Ken+08] P. Kenny, P. Ouellet, N. Dehak, V. Gupta et P. Dumouchel. « A Study of Interspeaker Variability in Speaker Verification ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 16.5 (juill. 2008), p. 980-988. issn : 1558-7916. doi : 10.1109/TASL.2008.925147.

[KGP90] Jody Kreiman, Bruce R. Gerratt et Kristin Precoda. « Listener Experience and Perception of Voice Quality ». In : *Journal of Speech, Language, and Hearing Research* 33.1 (1990), p. 103-115. doi : 10.1044/jshr.3301.103. eprint : https://pubs.asha.org/doi/pdf/10.1044/jshr.3301.103.

[Kim+10] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck et Douglas Turnbull. « MUSIC EMOTION RECOGNITION : A STATE OF THE ART REVIEW ». In : 2010.

[Kim15] Tae Kyun Kim. « T test as a parametric statistic ». In : *Korean Journal of Anesthesiology* 68.6 (2015), p. 540. issn : 2005-6419, 2005-7563. doi : 10.4097/kjae.2015.68.6.540.

[KK15] Gregory Koch et Gregory Koch. « Siamese Neural Networks for One-Shot Image Recognition ». In : *Cs.Toronto.Edu* 2 (2015).

[KMD03] P. Kenny, M. Mihoubi et Pierre Dumouchel. « New MAP estimators for speaker recognition ». In : *Proc. 8th European Conference on Speech Communication and Technology (Eurospeech 2003)*. 2003, p. 2961-2964.

[Ko18] Byoung Chul Ko. « A brief review of facial emotion recognition based on visual information ». In : *Sensors (Switzerland)* 18 (2 fév. 2018). issn : 14248220. doi : 10.3390/s18020401.

[Kum+15] C. Santhosh Kumar, Kuruvachan K. George, K. I. Ramachandran et Ashish Panda. « Weighted cosine distance features for speaker verification ». In : *2015 Annual IEEE India Conference (INDICON)*. 2015, p. 1-5. doi : 10.1109/INDICON.2015.7443376.

[Kun71] Shirou Kunihira. *Effects of the Expressive Voice on Phonetic Symbolism*. 1971, p. 427-429.

[KZS15] Gregory Koch, Richard Zemel et Ruslan Salakhutdinov. « Siamese neural networks for one-shot image recognition ». In : *ICML deep learning workshop*. T. 2. Lille. 2015.

[Lan10] Andreas Lanitis. « A survey of the effects of aging on biometric identity verification ». In : *International Journal of Biometrics* 2.1 (2010), p. 34. issn : 1755-8301, 1755-831X. doi : 10.1504/IJBM.2010.030415.

[Lan19] James Lane. *Pourquoi Hollywood adore autant les accents stéréotypés ?* Déc. 2019.

[Lap18] Itshak Lapidot. « Convergence problems of Mahalanobis distance-based k-means clustering ». In : *2018 IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)*. 2018, p. 1-5. doi : 10.1109/ICSEE.2018.8646138.

[Lar+20] « Language Models are Few-Shot Learners ». In : *Advances in Neural Information Processing Systems*. Sous la dir. de H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan et H. Lin. T. 33. Curran Associates, Inc., 2020, p. 1877-1901.

[LC17] Julie Loison-Charles. « Traduire les accents de l'anglais vers le français en doublage audiovisuel ». In : *Palimpsestes. Revue de traduction* 30 (sept. 2017), p. 82-98. doi : 10.4000/palimpsestes.2439.

[Lea57] Timothy Leary. *Interpersonal diagnosis of personality ; a functional theory and methodology for personality evaluation.* Interpersonal diagnosis of personality ; a functional theory and methodology for personality evaluation. Pages : xv, 518. Oxford, England : Ronald Press, 1957.

[LG14] Omer Levy et Yoav Goldberg. « Neural word embedding as implicit matrix factorization ». In : *Advances in neural information processing systems* 27 (2014), p. 2177-2185.

[LLF20] Xugang Lu, Sheng Li et Masakiyo Fujimoto. « Automatic speech recognition ». In : *SpringerBriefs in Computer Science* (2020), p. 21-38. issn : 21915776.

[LML13] Haizhou Li, Bin Ma et Kong Aik Lee. « Spoken language recognition : From fundamentals to practice ». In : *Proceedings of the IEEE* 101.5 (2013), p. 1136-1159. issn : 00189219.

[LP+16] David Lopez-Paz, Léon Bottou, Bernhard Schölkopf et Vladimir Vapnik. *Unifying distillation and privileged information*. 2016. arXiv : 1511.03643 [stat.ML].

[Mac67] J. Macqueen. « Some methods for classification and analysis of multivariate observations ». In : *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*. 1967, p. 281-297.

[Mar+11] David Martínez, Oldřich Plchot, Lukáš Burget, Ondřej Glembek et Pavel Matějka. « Language recognition in ivectors space ». In : *Interspeech 2011*. ISCA, août 2011, p. 861-864. doi : 10.21437/Interspeech.2011-329.

[Mar+97] Alvin F. Martin, George R. Doddington, Teresa M. Kamm, Mark Ordowski et Mark A. Przybocki. « The DET curve in assessment of detection task performance ». In : *EUROSPEECH*. 1997.

[MH13] Mahnoosh Mehrabani et John H.L. Hansen. « Singing speaker clustering based on subspace learning in the GMM mean supervector space ». In : *Speech Communication* 55.5 (juin 2013), p. 653-666. issn : 01676393. doi : 10.1016/j.specom.2012.11.001.

[Mik+13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado et Jeff Dean. « Distributed representations of words and phrases and their compositionality ». In : *Advances in neural information processing systems*. 2013, p. 3111-3119.

[MN21] Aansh Malik et Ha Nguyen. « Exploring automated voice casting for content localization using deep learning ». In : *SMPTE Motion Imaging Journal* 130 (3 avr. 2021), p. 12-18. issn : 21602492. doi : 10.5594/JMI.2021.3057695.

[MO20] Clément Le Moine et Nicolas Obin. « Att-HACK : An Expressive Speech Database with Social Attitudes ». In : *Speech Prosody 2020*. ISCA, mai 2020, p. 744-748. doi : 10.21437/SpeechProsody.2020-152.

[Mod] *Cinéma : «Le métier du doublage a un problème avec la couleur»*. Accessed : 2021-11-2. 2008.

[Mon09] Le Monde. *Doublage : la Halde accuse le cinéma français de discrimination raciale*. Jan. 2009.

[Mun+20] Sung Hwan Mun, Woo Hyun Kang, Min Hyun Han et Nam Soo Kim. *Robust Text-Dependent Speaker Verification via Character-Level Information Preservation for the SdSV Challenge 2020*. 2020. arXiv : 2010.11408 [eess.AS].

[Nar97] Sridhar Narayan. « The generalized sigmoid activation function : Competitive supervised learning ». en. In : *Information Sciences* 99.1-2 (juin 1997), p. 69-82. issn : 00200255. doi : 10.1016/S0020-0255(96)00200-9.

[Nin+19] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing et Liang-Jie Zhang. « A Review of Deep Learning Based Speech Synthesis ». In : *Applied Sciences* 9.19 (sept. 2019), p. 4050. issn : 2076-3417. doi : 10.3390/app9194050.

[OR16] Nicolas Obin et Axel Roebel. « Similarity search of acted voices for automatic voice casting ». In : t. 24. Institute of Electrical et Electronics Engineers Inc., sept. 2016, p. 1642-1651. doi : 10.1109/TASLP.2016.2580302.

[ORB14] Nicolas Obin, Axel Roebel et Gregoire Bachman. « On automatic voice casting for expressive speech : Speaker recognition vs. speech classification ». In : *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2014, p. 950-954. isbn : 9781479928927. doi : 10.1109/ICASSP.2014.6853737.

[Paq19] Thierry Paquot. « Le cinéma, petite fabrique de stéréotypes ». In : *Hermès* n° 83 (1 2019), p. 119. issn : 0767-9513. doi : 10.3917/herm.083.0119.

[Par+16] Titouan Parcollet, Mohamed Morchid, Pierre-Michel Bousquet, Richard Dufour, Georges Linares et Renato De Mori. « Quaternion Neural Networks for Spoken Language Understanding ». In : *2016 IEEE Spoken Language Technology Workshop (SLT)*. San Diego, CA : IEEE, déc. 2016, p. 362-368. isbn : 978-1-5090-4903-5. doi : 10.1109/SLT.2016.7846290.

[Pas+19] Santiago Pascual, Mirco Ravanelli, Joan Serrà, Antonio Bonafonte et Yoshua Bengio. *Learning Problem-agnostic Speech Representations from Multiple Self-supervised Tasks*. 2019. arXiv : 1904.03416 [cs.LG].

[Pic03] Rosalind W. Picard. « Affective computing : challenges ». In : *International Journal of Human-Computer Studies* 59.1-2 (juill. 2003), p. 55-64. issn : 10715819. doi : 10.1016/S1071-5819(03)00052-1.

[PK11] Marc D. Pell et Sonja A. Kotz. « On the time course of vocal emotion recognition ». In : *PLoS ONE* 6 (11 nov. 2011). issn : 19326203. doi : 10.1371/journal.pone.0027256.

[Pla14] Gaëlle Planchenault. « La commodification des voix au cinéma : un outil de différentiation et de stigmatisation langagière ». In : *Entrelacs* (11 déc. 2014). issn : 1266-7188. doi : 10.4000/entrelacs.1566.

[Plu01] Robert Plutchik. « The Nature of Emotions ». In : *American Scientist* 89.4 (2001), p. 344. issn : 0003-0996, 1545-2786. doi : 10.1511/2001.4.344.

[Pon+18] Emmanuel Ponsot, Juan José Burred, Pascal Belin et Jean-Julien Aucouturier. « Cracking the social code of speech prosody using reverse correlation ». In : *Proceedings of the National Academy of Sciences* 115.15 (avr. 2018), p. 3972-3977. issn : 0027-8424, 1091-6490. doi : 10.1073/pnas.1716090115.

[Pov+11a] 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 Catalog No. : CFP11SRW-USB. Hilton Waikoloa Village, Big Island, Hawaii, US : IEEE Signal Processing Society, 2011. isbn : 978-1-4673-0366-8.

[Pov+11b] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek et al. « The Kaldi speech recognition toolkit ». In : *IEEE 2011 workshop on automatic speech recognition and understanding*. 2011.

[PPK15a] Vijayaditya Peddinti, D. Povey et S. Khudanpur. « A time delay neural network architecture for efficient modeling of long temporal contexts ». In : *Interspeech*. 2015.

[PPK15b] Vijayaditya Peddinti, Daniel Povey et Sanjeev Khudanpur. « A time delay neural network architecture for efficient modeling of long temporal contexts ». In : *Interspeech 2015*. ISCA, sept. 2015, p. 3214-3218. doi : 10.21437/Interspeech.2015-647.

[PT09] Raquel R Pinho et João Manuel RS Tavares. « Tracking features in image sequences with kalman filtering, global optimization, mahalanobis distance and a management model ». In : (2009).

[PY10] S. J. Pan et Q. Yang. « A Survey on Transfer Learning ». In : *IEEE Transactions on Knowledge and Data Engineering* 22.10 (2010), p. 1345-1359.

[QDB21] Mathias Quillot, Richard Dufour et Jean-François Bonastre. « Assessing speaker-independent character information for acted voices ». In : *SPECOM*. Sous la dir. de Springer. 2021. isbn : 978-3-030-87802-3. doi : 10.1007/978-3-030-87802-3_51.

[Qui+20] Mathias Quillot, Lauriane Guillou, Adrien Gresse, Rafaël Ferro, Raphaël Röth, Damien Malinas, Richard Dufour, Axel Roebel, Nicolas Obin, Jean-François Bonastre, al La, Raphaël Roth et Emmanuel Ethis. « La voix actée : pratiques, enjeux, applications ». In : *XXXIIIe Journées d'Études sur la Parole*. ATALA et AFCP, juin 2020, p. 525-533.

[Qui+21] Mathias Quillot, Jarod Duret, Richard Dufour, Mickael Rouvier et Jean-François Bonastre. « Influence of speaker pre-training on character voice representation ». In : *SPECOM*. 2021. isbn : 978-3-030-87802-3. doi : 10.1007/978-3-030-87802-3_52.

[Raj+19] Desh Raj, David Snyder, Daniel Povey et Sanjeev Khudanpur. « Probing the information encoded in x-vectors ». In : *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE. 2019, p. 726-733.

[Rey05] Alain Rey. *Dictionnaire culturel en langue française, Paris*. Sous la dir. de Le Robert. 2005.

[Rey92] D.A. Reynolds. *A Gaussian Mixture Modeling Approach to Text-independent Speaker Identification*. College of Engineering, Georgia Institute of Technology, 1992.

[Rey+95] D.A. Reynolds, M.A. Zissman, T.F. Quatieri, G.C. O'Leary et B.A. Carlson. « The effects of telephone transmission degradations on speaker recognition performance ». In : *1995 International Conference on Acoustics, Speech, and Signal Processing*. T. 1. Detroit, MI, USA : IEEE, 1995, p. 329-332. isbn : 978-0-7803-2431-2. doi : 10.1109/ICASSP.1995.479540.

[RF16] Michael Rouvier et Benoit Favre. « Investigation of speaker embeddings for cross-show speaker diarization ». In : *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2016, p. 5585-5589.

[RH07] Andrew Rosenberg et Julia Hirschberg. « V-Measure : A Conditional Entropy-Based External Cluster Evaluation Measure. » In : jan. 2007, p. 410-420.

[RHR94] R.C. Rose, E.M. Hofstetter et D.A. Reynolds. « Integrated models of signal and background with application to speaker identification in noise ». In : *IEEE Transactions on Speech and Audio Processing* 2.2 (avr. 1994), p. 245-257. issn : 10636676. doi : 10.1109/89.279273.

[RM87] David E. Rumelhart et James L. McClelland. « Learning Internal Representations by Error Propagation ». In : *Parallel Distributed Processing : Explorations in the Microstructure of Cognition : Foundations*. 1987, p. 318-362.

[RQD00] Douglas A. Reynolds, Thomas F. Quatieri et Robert B. Dunn. « Speaker Verification Using Adapted Gaussian Mixture Models ». In : *Digital Signal Processing* 10.1 (2000), p. 19-41. issn : 1051-2004. doi : https://doi.org/10.1006/dspr.1999.0361.

[RR95] D.A. Reynolds et R.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. doi : 10.1109/89.365379.

[Sch+] Nicolas Scheffer, Luciana Ferrer, Martin Graciarena, Sachin Kajarekar, Elizabeth Shriberg et Andreas Stolcke. « The SRI NIST 2010 speaker recognition evaluation system ». In : (), p. 4. doi : 10.1109/ICASSP.2010.5495632.

[Sch01] Marc Schröder. « Emotional speech synthesis : A review ». In : *European Conference on Speech Communication and Technology (EUROSPEECH)*. 2001, p. 561-564. isbn : 8790834100.

[Sch+13] Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi et al. « The INTERSPEECH 2013 computational paralinguistics challenge : Social signals, conflict, emotion, autism ». In : *Annual Conference of the International Speech Communication Association (INTERSPEECH)*. 2013.

[Sch18] Björn W. Schuller. « Speech emotion recognition : two decades in a nutshell, benchmarks, and ongoing trends ». In : *Communications of the ACM* 61.5 (avr. 2018), p. 90-99. issn : 0001-0782, 1557-7317. doi : 10.1145/3129340.

[Sny+17] David Snyder, Daniel Garcia-Romero, Daniel Povey et Sanjeev Khudanpur. « Deep Neural Network Embeddings for Text-Independent Speaker Verification ». In : *Interspeech 2017*. ISCA, août 2017, p. 999-1003. doi : 10.21437/Interspeech.2017-620.

[Sny+18] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey et S. Khudanpur. « X-Vectors : Robust DNN Embeddings for Speaker Recognition ». In : *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, p. 5329-5333.

[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)*. Calgary, AB : IEEE, avr. 2018, p. 5329-5333. isbn : 978-1-5386-4658-8. doi : 10.1109/ICASSP.2018.8461375.

[SRK18] Monorama Swain, Aurobinda Routray et P. Kabisatpathy. « Databases, features and classifiers for speech emotion recognition : a review ». In : *International Journal of Speech Technology* 21.1 (2018), p. 93-120. issn : 15728110.

[SRM03] I. Shafran, M. Riley et M. Mohri. « Voice signatures ». In : *2003 IEEE Workshop on Automatic Speech Recognition and Understanding (IEEE Cat. No.03EX721)*. 2003, p. 31-36. doi : 10.1109/ASRU.2003.1318399.

[SS08] Elizabeth Shriberg et Andreas Stolcke. « The case for automatic higher-level features in forensic speaker recognition ». en. In : *Interspeech 2008*. ISCA, sept. 2008, p. 1509-1512. doi : 10.21437/Interspeech.2008-433.

[SSB14] Haşim Sak, Andrew Senior et Françoise Beaufays. « Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition ». In : *arXiv :1402.1128 [cs, stat]* (fév. 2014).

[Sto21] Brad Stone. *Amazon Unbound : Jeff Bezos and the Invention of a Global Empire*. 2021.

[Sty09] Yannis Stylianou. « Voice transformation : a survey ». In : *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2009, p. 3585-3588.

[TT05] Jianhua Tao et Tieniu Tan. « Affective Computing : A Review ». In : *Affective Computing and Intelligent Interaction*. Sous la dir. de David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Dough Tygar, Moshe Y. Vardi, Gerhard Weikum, Jianhua Tao, Tieniu Tan et Rosalind W. Picard. T. 3784. Series Title : Lecture Notes in Computer Science. Berlin, Heidelberg : Springer Berlin Heidelberg, 2005, p. 981-995. isbn : 978-3-540-29621-8 978-3-540-32273-3. doi : 10.1007/11573548_125.

[Val+20] Rafael Valle, Jason Li, Ryan Prenger et Bryan Catanzaro. « ICASSP 2020 - ». In : sous la dir. de Speech et Signal Processing (ICASSP) 2020 IEEE International Conference on Acoustics. IEEE, 2020. isbn : 9781509066315.

[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 : *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Florence, Italy : IEEE, mai 2014, p. 4052-4056. isbn : 978-1-4799-2893-4. doi : 10.1109/ICASSP.2014.6854363.

[VI15] Vladimir Vapnik et Rauf Izmailov. « Learning Using Privileged Information : Similarity Control and Knowledge Transfer ». In : *J. Mach. Learn. Res.* 16.1 (2015), 2023–2049.

[VLKE85] Diana Van Lancker, Jody Kreiman et Karen Emmorey. « Familiar voice recognition : Patterns and parameters : I. Recognition of backward voices. » In : *Journal of Phonetics* 13.1 (1985), p. 19-38. issn : 1095-8576(Electronic),0095-4470(Print).

[Vlo+00] Claude Vloeberghs, Patrick Verlinde, Carl Swail, Herman Steeneken et Allan South. « The Impact of Speech Under "Stress" on Military Speech Technology. (l'Impact de la parole en condition de "stress" sur less technologies vocales militaries) ». In : (mars 2000), p. 112.

[Wic00] A Wichmann. « The attitudinal effects of prosody, and how they relate to emotion ». In : *ITRW on Speech and Emotion* (2000), p. 5.

[WL21] Feng Wang et Huaping Liu. « Understanding the Behaviour of Contrastive Loss ». In : *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Nashville, TN, USA : IEEE, juin 2021, p. 2495-2504. ISBN : 978-1-66544-509-2. DOI : 10.1109/CVPR46437.2021.00252.

[WQY17] Shuai Wang, Yanmin Qian et Kai Yu. « What does the speaker embedding encode ? » In : *Interspeech*. 2017, p. 1497-1501.

[Yu+19] Yong Yu, Xiaosheng Si, Changhua Hu et Jianxun Zhang. « A Review of Recurrent Neural Networks : LSTM Cells and Network Architectures ». In : *Neural Computation* 31.7 (juill. 2019), p. 1235-1270. ISSN : 0899-7667, 1530-888X. DOI : 10.1162/neco_a_01199.

[Zey+18] Albert Zeyer, Kazuki Irie, Ralf Schlüter et Hermann Ney. « Improved Training of End-to-end Attention Models for Speech Recognition ». In : *Interspeech 2018*. ISCA, sept. 2018, p. 7-11. DOI : 10.21437/Interspeech.2018-1616.

[ZH07] Chi Zhang et John H. L. Hansen. « Analysis and classification of speech mode : whispered through shouted ». In : *Interspeech 2007*. ISCA, août 2007, p. 2289-2292. DOI : 10.21437/Interspeech.2007-621.

[ZH11] Chi Zhang et John H. L. Hansen. « Whisper-Island Detection Based on Unsupervised Segmentation With Entropy-Based Speech Feature Processing ». In : *IEEE Transactions on Audio, Speech, and Language Processing* 19.4 (mai 2011), p. 883-894. ISSN : 1558-7916, 1558-7924. DOI : 10.1109/TASL.2010.2066967.

[Zha18] Zijun Zhang. « Improved Adam Optimizer for Deep Neural Networks ». In : *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. Banff, AB, Canada : IEEE, juin 2018, p. 1-2. ISBN : 978-1-5386-2542-2. DOI : 10.1109/IWQoS.2018.8624183.

[ZL03] Dengsheng Zhang et Guojun Lu. « Evaluation of similarity measurement for image retrieval ». In : *International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003*. T. 2. 2003, 928-931 Vol.2. DOI : 10.1109/ICNNSP.2003.1280752.

[ZZ09] Ethan Zhang et Yi Zhang. « F-Measure ». In : *Encyclopedia of Database Systems*. Sous la dir. de Ling Liu et M. Tamer Özsu. Boston, MA : Springer US, 2009, p. 1147-1147. ISBN : 978-0-387-35544-3 978-0-387-39940-9. DOI : 10.1007/978-0-387-39940-9_483.

AVIGNON UNIVERSITÉ

Un premier pas vers la caractérisation de l'information véhiculée par les voix actées

Dualité des informations personnage et locuteur

par
Mathias Quillot

LABORATOIRE INFORMATIQUE D'AVIGNON