Traitement automatique de la parole expressive : retour vers des systèmes interprétables?

Marie Tahon

LIUM
Laboratoire d'Informatique
Université du Mans

Le Mans Université

HABILITATION À DIRIGER DES RECHERCHES

Présentée à Le Mans Université

École Doctorale N° 601
Spécialité : Informatique
Unité de recherche : LIUM (anc. EA 4023)

Traitement automatique de la parole expressive : retour vers des systèmes interprétables ?

Marie TAHON

Soutenue publiquement le 23 janvier 2023 devant un jury composé de :

Président :	Yannick ESTÈVE	Professeur des Universités à l'Université d'Avignon
Rapporteurs :	Corinne FREDOUILLE	Professeure des Universités à l'Université d'Avignon
	Damien LOLIVE	Professeur des Universités à l'ENSSAT/Université de Rennes 1
	Emmanuel VINCENT	Directeur de Recherche INRIA, LORIA
Examinateurs	Anthony LARCHER	Professeur des Universités à l'Université du Mans
	Sylvain MEIGNIER	Professeur des Universités à l'Université du Mans

SOMMAIRE

Introduction **7**

I Définitions et fondamentaux 11

1 Production acoustique de la parole expressive 13

1.1 Cadre de la parole expressive . . . 13
1.1.1 Définitions . . . 13
1.1.2 Parole émotionnelle . . . 14
1.1.3 Parole journalistique . . . 17
1.2 Acoustique de la parole . . . 18
1.2.1 Anatomie . . . 18
1.2.2 Modèle source filtre . . . 20
1.2.3 Production des sons de la parole . . . 21
1.2.4 Effet de l'expressivité sur la production acoustique . . . 22
1.3 Conclusion . . . 23

2 Traitement du signal de parole expressive 25

2.1 Descripteurs audios . . . 25
2.1.1 Descripteurs prosodiques . . . 26
2.1.2 Descripteurs spectraux . . . 29
2.1.3 Représentations classiques pour le traitement de la parole . . . 32
2.1.4 Variations acoustiques en contexte expressif . . . 33
2.2 Parole expressive et interactions . . . 34
2.2.1 Prendre la parole . . . 35
2.2.2 Disfluences . . . 36
2.2.3 Identité, genre, âge . . . 37
2.3 Conclusion . . . 38

II Segmentation, caractérisation et synthèse 39

3 Collecte et pré-traitement du signal audio 41

3.1 Collecte des données de parole . . . 41
3.1.1 Styles de parole . . . 41
3.1.2 Conditions d'enregistrement et de collecte . . . 44
3.1.3 Corpus de parole pour le traitement automatique . . . 46
3.2 Segmentation du signal audio : parole, silence, bruit, parole superposée . . . 47
3.2.1 Modélisation acoustique du contenu audio . . . 47
3.2.2 Détection d'activité vocale (VAD) . . . 55
3.2.3 Détection des zones de parole superposée (OSD) . . . 56
3.3 Caractérisation du locuteur . . . 59
3.3.1 Modélisation acoustique du locuteur . . . 59
3.3.2 Détection du genre (GD) . . . 64
3.3.3 Cas de la parole expressive . . . 65
3.4 Segmentation et regroupement en locuteurs . . . 69
3.4.1 Approches pour la segmentation et le regroupement en locuteur . . . 69
3.4.2 Choisir le nombre de classes . . . 76
3.4.3 SRL et parole expressive . . . 77
3.5 Conclusion et discussions . . . 78

4 Caractérisation automatique de l'expression vocale 79

4.1 Annotation des données expressives . . . 79
4.1.1 Considérations générales . . . 79
4.1.2 Annotations en émotion . . . 82
4.1.3 Annotations expressives . . . 84
4.2 Reconnaissance automatique des émotions . . . 88
4.2.1 Modélisation acoustique des émotions . . . 88
4.2.2 Représentation multi-modale . . . 96
4.3 Apprentissage actif pour l'annotation semi-automatique . . . 100
4.3.1 Le principe de l'apprentissage actif . . . 100
4.3.2 Applications . . . 101
4.4 Conclusion et discussions . . . 107

5 Synthèse de parole expressive **111**
5.1 Systèmes de synthèse de la parole à partir du texte . . . 111
5.1.1 Vue générale . . . 111
5.1.2 Synthèse par concaténation . . . 112
5.1.3 Synthèse neuronale . . . 115
5.1.4 Les vocodeurs . . . 117
5.1.5 Évaluation des systèmes de synthèse . . . 119
5.2 Modélisation de la prononciation expressive . . . 120
5.2.1 Méthodologie générale . . . 120
5.2.2 Choix et impact des paramètres . . . 123
5.2.3 Adaptation à une voix expressive . . . 127
5.3 Génération d'une voix expressive . . . 131
5.3.1 Contrôle implicite de l'expressivité par les données . . . 131
5.3.2 Contrôle explicite de l'expressivité . . . 134
5.4 Conclusion et discussions . . . 137

III Perspectives **140**

6 Bilan et perspectives de recherche **141**
6.1 Modélisation de l'expressivité . . . 141
6.2 Algorithmes et société . . . 145

Bibliographie **149**

Curriculum Vitae **181**

INTRODUCTION

La difficulté la plus importante dans la conduction c'est l'intonation. Tu dois savoir ce qui est faux et comment le corriger. [Boulez]

Ce qui est intéressant en musique, c'est la coupure, c'est-à-dire la notion d'intervalle, de subdivision, de cassure, de trajectoire (non pas la notion de trajectoire directe, mais de trajectoire mêlée à d'autres paramètres comme ceux du temps par exemple). [Boulez]

Le plus nécessaire, le plus difficile et l'essentiel dans la musique, c'est le tempo. [Mozart]

La parole est un moyen de communication fondamental pour les êtres humains. Elle est le fruit d'une action cognitive qui traduit notre pensée en un signal sonore structuré dans le temps et compréhensible par les autres. En plus du contenu sémantique, le signal de parole nous informe sur des caractéristiques personnelles du locuteur comme son âge, son genre ou son état émotionnel. En écoutant quelqu'un parler, on peut aussi avoir connaissance d'un trouble de la parole ou d'autres types de pathologie physiologiques, cognitives ou motrices. L'étude des voix pathologiques est en champ de recherche à part entière que nous n'aborderons pas ici. Par contre, nous traiterons des caractéristiques du locuteur obtenues à partir d'une capture audio de sa voix, comme son identité, et les moyens expressifs qu'il met en œuvre lors de l'enregistrement.

L'étude de la parole expressive est un domaine de recherche pluridisciplinaire : de la production acoustique de la parole, aux mécanismes cognitifs mis en jeu par le locuteur pendant l'interaction pour exprimer sa pensée. Le style expressif est révélateur d'un contexte socio-culturel ou d'un type d'interaction et nous permet ainsi d'étudier les comportements vocaux des humains. Plus précisément, le traitement automatique de la parole expressive a pour objectif de généraliser les observations des spécialistes à des grandes quantités de données et des contextes variés.

Depuis les années 2000, les techniques d'apprentissage automatique se sont largement développées dans ce domaine, atteignant des performances toujours plus élevées. En particulier ces dix dernières années, l'utilisation massive et quasi systématique des réseaux de neurones, a permis des gains impressionnants sur des quantités de données toujours plus importantes.

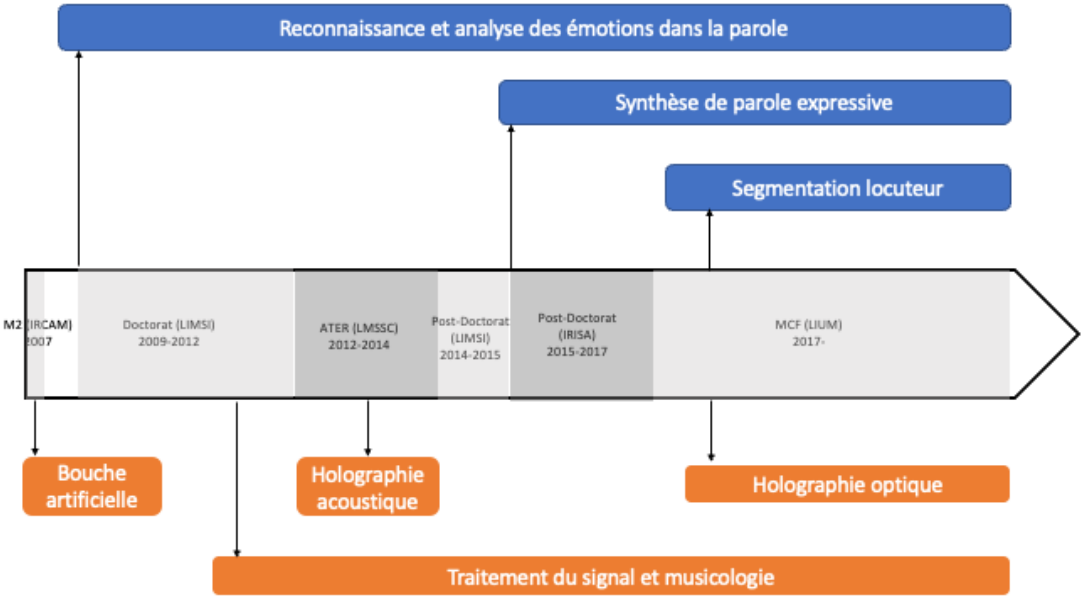


FIGURE 1 – Chronologie de mes domaines de recherche en fonction des postes occupés.

L'intérêt des grands groupes industriels pour les technologies vocales (assistants vocaux) a également contribué à l'essor de ce domaine. La communauté scientifique a bénéficié du renforcement des protocoles d'évaluation par l'organisation de concours, et de la tendance actuelle à la reproductibilité des expériences dans une volonté de transparence des méthodologies employées.

Depuis le début de mes travaux de recherche en 2009, j'ai cherché à expliciter ce qu'on appelle parole expressive, entre états émotionnels et caractéristiques du locuteur. J'ai constamment réalisé des aller-retours entre les méthodes d'apprentissage automatique considérées comme des boîtes noires peu interprétables mais performantes, et l'analyse du phénomène expressif à l'aide d'éléments acoustiques et linguistiques. Mon objectif est d'étudier comment et en quoi les systèmes de traitement automatique peuvent apporter des connaissances sur les différents mécanismes acoustique, cognitif et d'interaction qui induisent la production d'une parole expressive, et ce afin de faciliter aux experts des autres disciplines, leur analyse du signal et d'enrichir leur compréhension du phénomène expressif.

Mes thématiques de recherche sont guidées par une approche pluridisciplinaire du signal de parole comme illustré sur la chronologie de la figure 1. Pendant ma thèse (2009-2012), j'ai étudié la reconnaissance d'émotion dans des interactions humains-robots en combinant

des techniques d'apprentissage automatique avec une analyse acoustique de la voix émotionnelle [Tahon and Devillers, 2010][Tahon et al., 2011] [Tahon et al., 2012a], [Tahon, 2012]. Ma thèse a été l'occasion de collecter plusieurs corpus réalistes qui n'ont malheureusement pas pu être distribués à la communauté scientifique [Tahon et al., 2010] [Delaborde et al., 2009]. Lors de mon ATER (2012-2014), en plus de continuer mes travaux sur les émotions, j'ai mené des recherches en holographie acoustique sur le rayonnement de la flûte à bec [Tahon et al., 2014]. Un post-doctorat au LIMSI (2014-2015) m'a permis d'approfondir l'analyse acoustique des émotions, notamment en proposant des modèles de détection de rire [Tahon and Devillers, 2015], et en proposant des approches multi-corpus pour identifier les descripteurs audio les plus pertinents pour la classification des émotions [Devillers et al., 2015], [Tahon et al., 2015], [Tahon and Devillers, 2016a]. J'ai eu l'opportunité de réaliser des travaux dans le domaine de la synthèse de parole expressive pendant mon post-doctorat à l'IRISA (2015-2017) où j'ai étudié la modélisation de la prononciation et de la voix expressive dans un contexte de synthèse de parole à partir du texte [Tahon et al., 2016a], [Tahon et al., 2016b], [Qader et al., 2017], [Tahon et al., 2017], [Tahon et al., 2018]. Ces travaux ont aussi permis la collecte d'un corpus pour la synthèse en français SynPaFlex [Sini et al., 2018], [Tahon and Lolive, 2018] et la participation au challenge Blizzard en 2018 [Alain et al., 2017].

Enfin, depuis mon intégration au LIUM en 2017, mes thématiques se sont élargies et renforcées. J'ai eu l'opportunité de me rapprocher du traitement du signal bas-niveau, notamment avec des travaux en identification du locuteur (projet EXTEnSor porté par Anthony Larcher, LIUM, co-encadrement du post-doc de Ambuj Merish) [Larcher et al., 2021], en détection de parole superposée (projet GEM porté par David Doukhan, INA, co-encadrement de la thèse de Martin Lebourdais)[Lebourdais et al., 2022b][Lebourdais et al., 2022a] et en segmentation et regroupement en locuteur via de l'apprentissage actif (projet ALLIES, porté par Anthony Larcher, co-encadrement du post-doc de Meysam Shamsi)[Shamsi et al., 2022]. J'ai renforcé mon expertise en caractérisation du signal expressif avec la thèse de Manon Macary (CIFFRE avec AlloMedia) autour de la reconnaissance de la satisfaction dans les corpus de centres d'appel [Macary et al., 2020a][Macary et al., 2021], incluant la constitution du corpus AlloSat [Macary et al., 2020b], et enfin avec la caractérisation de l'hésitation dans la parole spontanée (collaboration avec le LPP) [Wottawa et al., 2020].

Mon intérêt pour la modélisation acoustique du signal audio, m'a également incité à mettre en place des collaborations pluridisciplinaires dans le domaine de l'analyse musicale assistée par ordinateur. J'ai réalisé des travaux autour de la qualité vocale dans la voix chantée traditionnelle [Sitchet and Tahon, 2016] [Tahon and Sitchet, 2016][Tahon and Sitchet, 2017]. Et

plus récemment, j'ai travaillé sur la segmentation en motifs musicaux [Tahon et al., 2019]. Depuis 2018, je collabore également avec le LAUM sur la mise en place de réseaux de neurones pour le débruitage d'images holographiques [Tahon et al., 2021][Tahon et al., 2022], [Montrésor et al., 2020]. Les travaux issus de ces collaborations ne seront pas développées dans ce document.

Nous nous intéresserons ici à plusieurs domaines qui font partie du traitement automatique de la parole expressive. Plus précisément, j'exposerai les innovations qui me semblent majeures dans les domaines de la segmentation du signal pour la collecte et l'annotation des données, la reconnaissance automatique de styles expressifs et leur synthèse.

Le document est organisé en trois parties. La première partie précise les définitions que j'emploie dans le cadre de la parole expressive (chapitre 1) et expose les fondamentaux de la production acoustique de la parole expressive (chapitre 2). Au travers de cette partie, je mets en évidence les problématiques spécifiques liées à l'expressivité telles que la subjectivité, la faible quantité de données disponibles et la variabilité des contextes. La deuxième partie expose les recherches que j'ai menées autour de la collecte et du pré-traitement des données audio (chapitre 3), la caractérisation de l'expressivité (chapitre 4) et la synthèse de parole expressive (chapitre 5). Cette partie développe la problématique de la caractérisation implicite ou explicite d'un style expressif. La troisième et dernière partie est dévolue aux perspectives de recherche à court et moyen terme, ainsi que ma vision de la place de ces travaux dans l'ensemble de la société.

Première partie

Définitions et fondamentaux

Chapitre 1

Production acoustique de la parole expressive

Ce chapitre commence par définir ce que j'entends par parole expressive dans mes recherches. Ce terme aux contours flous car extrêmement subjectif lorsqu'il est utilisé hors contexte, ne peut prendre sens que dans le cas d'une application précise. La deuxième section s'attache à présenter les modes de production acoustique du signal de parole en général, et expressive en particulier, notamment lorsque le locuteur exprime une émotion. Je détaille le modèle source-filtre, car c'est un élément fondamental dans la caractérisation et la synthèse de la parole par le biais des descripteurs audio (chapitre 2). C'est donc un élément clé pour la compréhension et l'analyse du phénomène expressif qui sera nécessaire pour interpréter et expliquer les résultats des modèles appris sur des données de parole.

1.1 Cadre de la parole expressive

1.1.1 Définitions

L'expression peut se définir comme "l'action permettant de communiquer ses pensées à autrui par le moyen d'un medium : parole, geste, physionomie, etc." [1].

L'expressivité est généralement associée aux émotions, la parole étant alors un moyen d'exprimer ses émotions suivant un degré d'intensité variable [Ståhl et al., 2005]. De manière similaire, une œuvre d'art est dite expressive lorsqu'elle induit chez l'observateur une émotion ou un sentiment particulier [Matravers, 2001]. La parole devient expressive lorsqu'elle exprime quelque chose au delà du contenu purement linguistique et sémantique. Ce quelque chose peut être un sentiment, une émotion, un style d'élocution ou une appartenance sociale. Avec une telle définition, tout acte de parole humaine est expressif suivant un degré d'intensité variable, qui est adapté de façon pragmatique au contenu linguistique et sémantique.

1. Définition Larousse - consulté le 2/08/2022

Dans l'ensemble de ce document, la définition de la parole expressive restera un compromis, et ne se précisera que dans un cadre applicatif restreint. Dans la plupart des études, l'expressivité se restreint à l'expression d'une émotion. En effet, les théories émotionnelles permettent de définir précisément, à l'aide d'étiquettes explicites (des mots) ce qu'est une émotion. Attention toutefois, même si les étiquettes –des mots donc– sont communes, la perception d'un état émotionnel d'un locuteur reste subjectif car étroitement lié à notre expérience personnelle et notre personnalité. Dans le domaine de la synthèse de parole, on cherche généralement à produire une voix expressive, c'est-à-dire qu'on s'attachera au fait que la voix n'exprime non pas une émotion particulière, mais qu'elle s'exprime en adéquation avec le contexte : assistant virtuel, livre audio, aide à l'apprentissage, aide aux personnes mal-voyantes, etc. Suivant le degré souhaité, l'expressivité peut donc être définie comme le fait d'utiliser une prononciation adaptée, ajouter de l'emphase, des marqueurs sociaux (accent, dialogue, style) et affectifs (rires, hésitations) et éventuellement exprimer une émotion.

Dans la suite, nous allons présenter deux domaines applicatifs de la parole expressive : la parole émotionnelle et la parole journalistique.

1.1.2 Parole émotionnelle

L'expression des émotions prendra plusieurs formes qui se divisent en différentes catégories en fonction du temps, mettant ainsi en évidence l'importance d'une structure temporelle dans notre vie affective comme le montre la figure 1.1. Ainsi les changements d'expression auront lieu sur des temps très courts de l'ordre de la seconde, une émotion (fullblown emotion) sera généralement assez courte (quelques heures) et intense, alors qu'une humeur (mood) est un état émotionnel sous-jacent prolongé. Par conséquence, un mot tel que "joie" peut décrire à la fois une émotion, une humeur ou une attitude qui ont des durées dans le temps très variables. Dans le domaine de la reconnaissance des émotions, on cherchera les changements d'expression liés à un état émotionnel. On traitera donc des signaux de parole entre quelques secondes et quelques minutes. Ce constat limite dès à présent la portée des outils de reconnaissance automatique des émotions qui seront développés.

Théories émotionnelles

Il existe un grand nombre de théories émotionnelles développées dans plusieurs domaines en parallèle : psychologie, physiologie, linguistique, sciences cognitives, informatique, etc. Nous venons d'évoquer le fait que les états émotionnels peuvent avoir des durées de vie très variables.

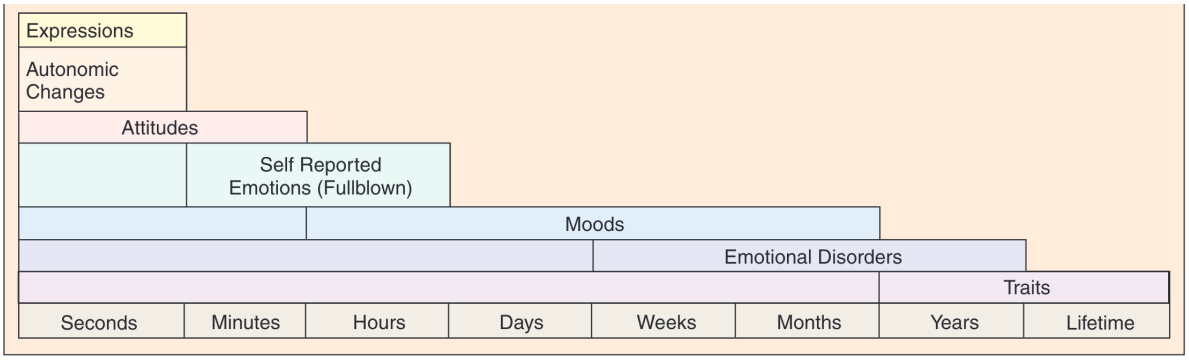


FIGURE 1.1 – Caractéristiques temporelles des catégories émotionnelles. Figure extraite de [Cowie et al., 2001].

De plus, leur perception est fortement subjective. Par conséquent, il est difficile, voire impossible, d'établir un réel consensus sur une définition des états émotionnels. Pour le traitement automatique de la parole expressive, nous devons cependant nous tenir à des théories algorithmiquement descriptives. C'est pourquoi le domaine de l'*affective computing* [Picard, 1997] repose essentiellement sur deux courants.

Le premier courant est issu d'une recherche autour d'un nombre fini d'émotions basiques qui seraient universelles. Ekman [Ekman, 1999] propose 6 catégories primaires à partir de l'étude de la physionomie (les mouvements du visage) : colère, joie, tristesse, surprise, dégoût et peur. Plutchik [Plutchik and Kellerman, 1980] en propose 8 en ajoutant la confiance et l'anticipation, et les place sur un cercle afin de pouvoir combiner ces émotions primaires avec d'autres catégories secondaires. Cette approche a l'avantage de correspondre à la description verbale des émotions ressenties par une personne. Cependant, elle ne permet pas facilement de caractériser des états émotionnels qui seraient à la frontière entre deux catégories, ou bien une graduation dans la force de l'émotion ressentie. Cette description est bien pratique pour annoter les corpus de parole émotionnelle à l'aide d'un nombre d'étiquettes fini. Cependant le contenu émotionnel caché derrière ces étiquettes est très spécifique aux conditions de collecte des données et leurs manifestations acoustiques varient fortement d'un corpus à un autre [Tahon and Devillers, 2016b]. De plus, ces étiquettes sont généralement annotées au niveau d'un segment de parole de durée variable (quelques secondes) correspondant à un groupe de souffle émotionnellement homogène [Tahon et al., 2011]. Ainsi l'évolution de l'émotion au cours du temps est difficile à représenter.

Le deuxième courant permet de s'affranchir de catégories préalablement définies, en pro-

posant de décrire les états affectifs avec des dimensions émotionnelles [Wundt, 1897]. Dans le modèle Circumplex proposé par Russel [Russell, 1980], on distingue trois dimensions indépendantes : agréable-désagréable (valence), tendu-détendu (activation), agité-calme (contrôle). Scherer [Scherer, 2005] propose d'ajouter la dimension conductif-obstructif. Cette description des émotions a l'avantage d'être compatible avec des états qui évoluent au cours du temps, notamment lorsque les dimensions sont des fonctions continues du temps, comme c'est le cas dans les corpus récents SEWA [Kossaifi et al., 2021] et AlloSat [Macary et al., 2020b]. Il sera alors possible de représenter un état affectif dans un espace à plusieurs dimensions incluant la dimension temporelle, par contre, il restera difficile de placer les étiquettes discrètes dans un espace commun aux différents corpus.

Il faut noter que ces théories prennent en compte l'ensemble des moyens d'expression : gestes, parole, physionomie, etc. De plus, elles définissent principalement l'émotion ressentie par l'humain, et très peu celle perçue par une personne extérieure. C'est pourquoi, une partie des recherches autour de la caractérisation des émotions dans la parole, intègre l'interaction entre l'émetteur et le (ou les) récepteur(s) comme c'est le cas du modèle de lentille Brunswikian [Scherer, 2003]. Malheureusement, ces modèles sont assez complexes, et leur intégration dans un système de traitement automatique reste un défi important du domaine.

Le signal social

Lorsqu'une personne parle sous l'effet d'une émotion, elle est en situation de communication et interagit avec une ou d'autres personnes. Par conséquent, l'étude de la parole émotionnelle, inclut aussi l'analyse du signal social [Vinciarelli et al., 2009] enregistré lors d'une interaction entre des locuteurs. L'étude du signal social se base souvent sur les états émotionnels des interactants, mais également sur tout un tas d'éléments générés pour l'interaction. Par exemple, Scherer définit *affect bursts* comme étant une partie de ces éléments : *very brief, discrete, nonverbal expressions of affect in both face and voice as triggered by clearly identifiable events* ([Scherer, 1994], p170). Une étude confirme que les émotions peuvent être facilement identifiables dans des enregistrements d'*affect bursts* tels que des rires, interjections, hésitations et respirations [Schröder, 2003]. Les éléments du signal social qui ont été les plus étudiés notamment grâce au challenge Interspeech 2013 [Schuller et al., 2013], car plus simples à caractériser automatiquement, sont les rires [Tahon et al., 2011] et les hésitations [Wottawa et al., 2020].

Il semble donc évident que l'étude de la parole émotionnelle ne peut se faire sans avoir une connaissance précise des circonstances sociales et interactionnelles qui ont mené à sa

production. L'annotation du signal de parole en étiquettes ou en dimensions émotionnelles est une tâche subjective. Ainsi la modélisation des émotions ne pourra pas se faire sans l'économie d'une analyse robuste du sens de l'étiquette ou de la dimension affective dans le contexte précis du corpus étudié. La deuxième partie de ce document présente plusieurs études sur la parole émotionnelle qui mettent en évidence la subjectivité de la tâche. Dans le contexte du traitement automatique de la parole, nous présenterons les limites et les besoins liées à l'annotation des données et à l'évaluation des modèles, en particulier en ce qui concerne la prise en compte du caractère subjectif.

1.1.3 Parole journalistique

L'étude de la parole journalistique est un domaine intéressant car elle ne se focalise pas particulièrement sur les émotions, mais plutôt sur un style expressif. En effet, dans les émissions de radio et de TV, les styles peuvent être très diverses suivant que le locuteur est un.e journaliste, un homme ou une femme politique, une personne interviewée. De plus, les émissions peuvent être des reportages ou des journaux où un seul locuteur parle, des dialogues, ou des débats incluant plusieurs participants. Ces deux aspects font de la parole journalistique un domaine d'étude riche et varié dans lequel l'expressivité peut prendre des degrés très différents.

L'étude des styles de parole se focalise généralement autour de la parole spontanée, préparée, conversationnelle ou formelle [Hirschberg, 2000]. La parole lue est également étudiée, notamment pour des applications en synthèse de parole [Montaño and Alías, 2016]. Dans le cadre de la parole journalistique, on définit ainsi un style de discours qui est fonction d'une situation de communication particulière, celle du journaliste. En effet, selon Charaudeau [Charaudeau, 2006], *à force d'échanges langagiers, les comportements des partenaires se stabilisent en instaurant des normes communicationnelles [...]. Le journaliste doit raconter, expliquer, capter, mais ce n'est point en historien, en savant, en politique*. Afin de remplir ses missions, le journaliste va avoir tendance à surjouer et à dramatiser. Par exemple, plutôt que d'interroger les experts en se mettant ainsi en position d'infériorité, il aura tendance à les interpeller, voire à les provoquer afin de capter l'attention de son auditoire. Pour raconter et capter, le style journalistique se caractérise par une intonation emphatique et un débit de parole rapide [Rodero Antón and Cores-Sarría, 2021].

Dans les contenus journalistiques, plusieurs rôles interagissent : journalistes, experts (politiques, scientifiques, etc.) et autres personnes interviewées [Adda-Decker et al., 2008]. L'analyse de ces rôles permet de caractériser les interactions entre les différents participants en utilisant les méthodes développées dans le domaine du traitement du signal social.

Dans la deuxième partie de ce document, nous étudierons la parole journalistique comme un ensemble de styles expressifs, éventuellement émotionnels, associés à des rôles. Cette étude mettra en évidence les besoins en traitement automatique de la parole depuis la segmentation du signal audio à la caractérisation des actes de dialogues dans différentes situations d'interactions.

1.2 Acoustique de la parole

Cette section aborde la production de la parole d'un point de vue acoustique. Cet aspect a longtemps été fondamental pour extraire du signal sonore des descripteurs audio qui représentent les éléments constitutifs de la parole afin d'alimenter ensuite les modèles. Aujourd'hui, avec les réseaux de neurones les connaissances liées à la production acoustique de la parole semblent superflues pour apprendre des modèles performants de traitement automatique, puisque le signal est traité directement via des réseaux de neurones sans traitement du signal en amont.

On pourrait donc conclure que les informaticiens n'ont pas besoin d'étudier l'acoustique de la parole. Cependant, comme nous le verrons plus tard dans les perspectives (chapitre 6), les besoins en explicabilité et interprétabilité des résultats fournis par les modèles sont de plus en plus pris en compte par la société. Il semble impossible de réaliser cette tâche sans un minimum de compréhension du phénomène acoustique à l'origine du signal sonore. En effet, cette connaissance permet d'identifier des éléments interprétables caractéristiques de la parole tels que l'intonation, les formants ou les phonèmes. Cela permet également d'identifier des éléments acoustiques qui peuvent modifier le signal et perturber le comportement des modèles, comme la présence de bruit de fond, l'influence de la capture du son, les types de voix (expressives ou pathologiques). En résumé, étudier la production de la parole permet d'anticiper les difficultés de modélisation et de savoir ce qu'on fait.

1.2.1 Anatomie

L'appareil vocal humain est constitué de trois parties principales (figure 1.2) : la structure sub-glottique (poumons, bronches et trachée), le larynx (cordes vocales et glotte) et le conduit vocal (pharynx, cavité buccale, nasale, joue et langue). Les cordes vocales peuvent être fermées (déglutition), en vibration (production d'un son voisé) ou ouvertes (respiration ou production d'un son non-voisé).

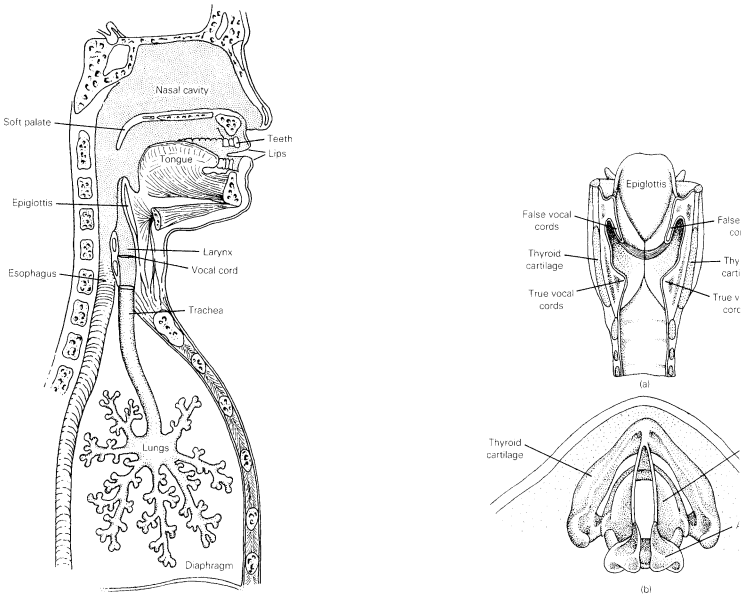


FIGURE 1.2 – Appareil vocal vue d'ensemble (gauche) et vue du larynx (droite) [Sundberg, 1995].

La production de sons voisés est caractéristique de notre appareil vocal. Ces sons sont produits grâce à une mise en vibration des cordes vocales, c'est ce qu'on appelle généralement "la voix". Le relâchement du diaphragme entraîne une expulsion de l'air (expiration passive), éventuellement aidé par les muscles expirateurs (expiration forcée). La forte pression ainsi générée sous les cordes vocales les forcent à s'ouvrir. L'air en passant crée une force de Bernoulli (augmentation de la vitesse et baisse de la pression) schématisée sur la figure 1.3. Ainsi de suite, un train de bouffées d'air est envoyé dans le conduit vocal qui entre alors en vibration. Le débit d'air passant dans le larynx est bien un signal périodique (voir figure 1.4) dont la fréquence fondamentale correspondant à la période d'ouverture et de fermeture des cordes vocales.

Les mécanismes de vibration des cordes vocales peuvent être différents suivant l'effort musculaire fournit par le locuteur. L'appellation "cordes vocales" est d'ailleurs abusive, puisque ces "cordes" sont en réalité des tissus d'épaisseur modifiable. La voix de poitrine correspond à un registre grave où les cordes vocales vont s'ouvrir et se fermer sur toute leur épaisseur. Ce mécanisme, dit "lourd", permet d'atteindre des fréquences fondamentales basses. La voix de tête correspond à un registre plus aigu où les cordes vocales sont fines et vont s'accoler sur une petite surface, la fréquence de vibration pourra alors être plus élevée. On notera également le mécanisme "fry" où la vibration est si lente qu'on entend les battements. C'est

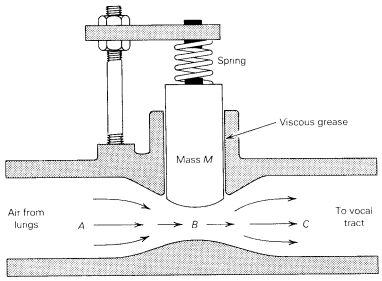


FIGURE 1.3 – Schéma de la vibration des cordes vocales sous l'effet d'une force de Bernoulli. Les cordes vocales sont modélisées par un système masse-ressort [Hall, 1998].

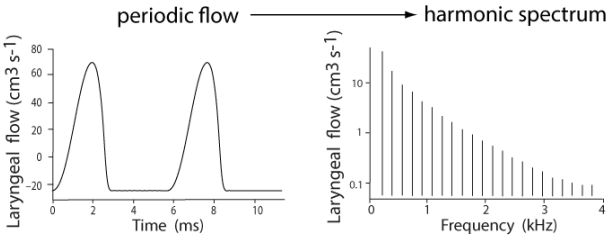


FIGURE 1.4 – Forme d'onde créée au niveau du larynx. Débit d'air à travers les cordes vocales (gauche) et spectre harmonique de ce débit périodique (droite) [Wolfe et al., 2016].

un mécanisme très utilisé en anglais américain dans les fins de phrase. Et le mécanisme de sifflet, qui correspond à la voix criée.

1.2.2 Modèle source filtre

En acoustique musicale, on distingue deux types d'instruments suivant le système de couplage existant entre la source (ici les cordes vocales) et le filtre (ici le conduit vocal). Dans un système dit à "anche faible", le filtre (la trompette ou la clarinette) va imposer une fréquence de vibration à la source (les lèvres du trompettistes ou l'anche de la clarinette). Dans un système dit à "anche forte", la source impose sa fréquence de vibration sans être influencée par le filtre. L'appareil vocal humain fait partie de la seconde catégorie, la fréquence de vibration des cordes vocales est indépendante des fréquences de résonance du conduit vocal (les formants).

Au niveau du larynx se produit alors un son de fréquence fondamentale F_0 celle de vibration des cordes vocales qui est contrôlée uniquement par les muscles du larynx. Les résonateurs sont principalement les cavités buccales et nasales. Mais le son émis par l'appareil vocal est également modulé par les articulateurs : lèvres, dents et langue. La position conjointe des résonateurs et des articulateurs va permettre de modifier les formants (sans changer la fréquence

fondamentale) et ainsi produire une grande quantité de sons voisés.

La forme du conduit vocal définit aussi le timbre de la voix. Chaque être humain est différent physiquement, et la forme du conduit vocal sera donc unique. De plus, les voix de chacun sont également diverses et variées en fonction de l'âge, du sexe, de la morphologie, etc. Le timbre d'une voix est principalement déterminé par les composantes morphologiques et l'utilisation des résonateurs. C'est grâce au timbre que l'on peut différencier deux personnes chantant la même mélodie ou parlant en même temps. C'est également grâce à un formant spécifique (autour de 4kHz) obtenu après des années de travail, que l'auditeur peut distinguer la voix du chanteur lyrique parmi l'orchestre tout entier. Les locuteurs femmes ont généralement un conduit vocal plus petit que celui des hommes (celui des enfants étant plus petit que celui des femmes). La longueur du conduit vocal va influencer directement le train de bouffée d'air et va donc modifier la fréquence fondamentale. Ainsi, les hommes ont une voix généralement plus grave que celle des femmes. Attention cependant, la physiologie explique en partie certains paramètres physiques du modèle source-filtre, mais l'effort musculaire conscient ou non, du locuteur influencera également la production vocale. De plus, malgré les efforts autour de la définition de la notion de timbre, cette notion étant fortement subjective, elle reste difficile à cerner.

Le lecteur pourra se référer à [Degottex, 2010] pour une revue détaillée des modèles source-filtre existant ainsi que sur les paramètres extraits du signal temporel capté à l'extérieur du conduit vocal.

1.2.3 Production des sons de la parole

Il existe principalement trois modes d'excitation de la source, *i.e.* des cordes vocales. Suivant ces modes, les sons de la parole sont classés et un symbole répertorié par l'alphabet phonétique international (IPA), leur est attribué. La mise en vibration des cordes vocales permet de produire les sons voisés. Les sons produits sont alors quasi-périodiques. L'énergie spectrale est concentrée autour de la F_0 et des premiers formants. C'est le cas des voyelles principales du français : /a, e, ɛ, i, o, ɔ, u/. L'ouverture brutale des cordes vocales permet de faire passer une unique bouffée d'air suivie ou non d'une mise en vibration. Ce mécanisme génère des sons impulsionnels, non périodiques qui contiennent une large bande de fréquence. On peut donc ainsi produire des consonnes plosives non-voisées (/p, t, k/) ou bien voisées (/b, d, g/). Les sons fricatifs sont obtenus en réalisant une constriction avec les dents ou les lèvres. Ainsi une turbulence va se créer lors du passage du flux d'air dans le conduit étroit. Ce mode d'excitation permet d'obtenir un signal permanent, non périodique contenant une large

bande de fréquence plutôt vers les aigüs. Les sons du français sont classés en 36 phonèmes : 16 voyelles, 3 semi-consonnes (/j, w, ɥ/) et 17 consonnes.

Les sons canoniques correspondant aux symboles phonétiques de l'IPA ne sont généralement pas réalisés tels quels lors d'une interaction réelle. En effet, la production vocale s'organise dans le temps, et des contraintes dynamiques et motrices vont limiter les mouvements des articulateurs et résonateurs lors de la production. Les sons réels ne se représentent donc pas suivant des catégories discrètes de phonèmes, mais plutôt suivant des continuum. La variabilité de la prononciation, étudiée dans le chapitre 5, dépendra du locuteur lui-même (de son état émotionnel et de son origine), mais également du contenu de l'interaction (style, imitation).

1.2.4 Effet de l'expressivité sur la production acoustique

Les émotions que l'ont ressent ont des conséquences sur le fonctionnement global de notre organisme : cerveau, muscles, amygdales, température, etc. Dans ce document, nous nous limitons uniquement aux effets sur l'appareil vocal qui peuvent être liés plus ou moins fortement aux autres organes.

Alors que la littérature autour de la reconnaissance des émotions exprimées vocalement par un auditeur est très fournie, il y a très peu d'analyses des différences physiologiques et acoustiques de l'appareil vocal en situation émotionnelle. Scherer [Scherer, 1986] a montré qu'une émotion ressentie avait des effets considérables sur la respiration, la phonation et l'articulation. Par exemple, une personne qui évalue une situation comme étant agréable aura tendance à relâcher son conduit vocal et donc l'agrandir. Par conséquent, la fréquence de vibration des cordes vocales aura tendance à diminuer. Dans le cas d'une situation désagréable, le conduit vocal aura tendance à se tendre, et donc à se rétrécir. De même, une personne qui considère qu'une situation va nécessiter une réaction puissante aura tendance à utiliser un registre de poitrine (un mécanisme de vibration des cordes vocales lourd) qui favorise une fréquence de vibration basse et une forte énergie, plutôt qu'un registre de tête.

On sait aussi que les situations de communication impliquent un degré de contrôle plus ou moins important sur la production acoustique. Pour une revue détaillée de la production acoustique depuis la parole lue (très contrôlée) vers la parole spontanée (peu contrôlée), le lecteur pourra se référer à [Meunier, 2014].

1.3 Conclusion

Ce premier chapitre nous permet de définir le cadre de l'étude de l'expressivité entre parole émotionnelle et parole journalistique. Dans ces deux cadres d'application, nous étudions les variations de l'expressivité. La connaissance des modèles de production acoustique du type source-filtre est pertinente dès lors que l'on souhaite étudier et analyser l'acte de parole. En effet, en situation de communication, l'état émotionnel du locuteur, ainsi que la situation elle-même, vont avoir un impact direct sur la production sonore.

Notre étude ne porte pas sur la production physique de la parole, mais sur la modélisation de l'expressivité dans des signaux de parole enregistrée. Ainsi, le prochain chapitre développe les éléments capturés par le signal audio et leur extraction automatique.

Chapitre 2

Traitement du signal de parole expressive

Ce chapitre s'attache à présenter les descripteurs audio utilisés pour modéliser le signal de parole en général, et expressive en particulier, notamment lorsque le locuteur exprime une émotion. Il met en évidence l'importance du modèle acoustique source-filtre dans la construction des représentations du signal qui seront utilisées par la suite dans l'ensemble des chaînes de traitement automatique. La connaissance fine de l'ensemble de ces paramètres, ainsi que des éléments linguistiques et interactionnels qui sous tendent l'acte de parole me semble indispensable pour ensuite interpréter et expliquer précisément les résultats obtenus à l'aide de modèles appris sur des grandes quantités de données. Nous verrons dans la deuxième partie, que les modèles neuronaux de type end to end se passent de ces paramètres explicites et apprennent une représentation latente qui leur est propre. Cependant, la nécessité d'un contrôle fin lorsqu'on sort d'un contexte général pour aller vers une application spécifique demande de ré-introduire des paramètres explicites. Enfin, je replace l'acte de parole dans un contexte d'interaction avec un ou plusieurs autres locuteurs.

2.1 Descripteurs audios

Aujourd'hui, les bases de données de signaux de parole sont de plus en plus grandes, et contiennent des enregistrements acquis dans des conditions de moins en moins contrôlées. Les corpus actuels cherchent à augmenter la diversité des locuteurs, des environnements acoustiques (capture audio et bruits de fond) et des scénarios de collecte, et ceci afin que les modèles appris sur ces données généralisent au mieux et restent performants même sur des conditions non vues lors de l'apprentissage. De plus, avec l'augmentation des ressources de calcul, l'extraction des descripteurs audio peut se faire de façon exhaustive, moyennant un coût matériel et énergétique. Ainsi, on va préférer multiplier les descripteurs audio, et sélectionner automatiquement les plus pertinents pour la tâche visée.

2.1.1 Descripteurs prosodiques

La prosodie est un phénomène complexe lié à la production de la parole. Il n'existe pas de définition consensuelle de la prosodie, mais on pourra néanmoins retenir celle-ci [1] :

Prosodie

La prosodie s'attarde plus précisément à l'impression musicale que fournit l'énoncé. On y observe des phénomènes prosodiques tels que : l'intonation, l'accentuation, le rythme, le débit et les pauses. Chacun de ces phénomènes prosodiques se manifeste par des variations au niveau de la fréquence, de la hauteur, de l'intensité et/ou de la durée.

La fréquence fondamentale

La fréquence fondamentale, ou F_0, est un des descripteurs de parole principaux. Les outils d'extraction de F_0 (méthode d'autocorrélation [Boersma and Weenink, 2018], algorithme YIN [de Cheveigné and Kawahara, 2002]) permettent d'obtenir la fréquence physique, celle qui correspond à la vibration des cordes vocales. Lorsque les modèles reposent en partie au moins, sur la F_0, il faut toujours prendre en compte l'incertitude liée aux erreurs faites par les algorithmes. Ces erreurs sont d'autant plus importantes que la qualité audio est dégradée ou que la voix est atypique. Dans ce cas, les paramètres par défaut des algorithmes ne sont généralement plus adaptés. L'intonation, c'est-à-dire la forme du contour donné par la fréquence fondamentale, est porteuse d'information concernant la structure temporelle du discours mais aussi de l'émotion du locuteur [Bänziger and Scherer, 2005]. Alors que la F_0 est classiquement donnée tous les 10 ms, c'est plutôt la composante segmentale (au niveau du phonème, voire du mot) de la prosodie qui est étudiée (macro-prosodie). La micro-prosodie permet d'étudier les variations sur des fenêtres temporelles inférieures à une centaine de ms. En particulier, les micro-variations temporelles de la F_0 (jitter) et de l'énergie (shimmer) peuvent être utiles dans le cas où une émotion forte ou une pathologie altèrent directement la production acoustique en modifiant le comportement vibratoire des cordes vocales.

Plusieurs modélisations de l'intonation au niveau macro-prosodie ont été proposées comme ToBI [Silverman et al., 1992] ou le prosogramme [Mertens, 2004]. Elles permettent de styliser

1. `http://www.latl.unige.ch/safran/data/phono/mod9/1_def/index.htm`

la F_0 en un ou plusieurs segments de droite qui peuvent être horizontaux ou avec une pente mélodique selon des seuils perceptifs paramétrables.

L'énergie

L'énergie (ou intensité sonore) est toujours difficile à manipuler. En effet, la distance du locuteur avec le microphone n'est généralement pas contrôlée pendant l'enregistrement, ni la calibration du matériel de captation, ce qui induit des différences d'intensité liées plutôt à l'environnement acoustique qu'au phénomène de parole étudié. Les proéminences ont lieu lorsque le locuteur accentue un élément de son discours. Elles se caractérisent par une variation rapide de la F_0, une augmentation de l'intensité sonore et un allongement de la durée du noyau syllabique. Il est possible de détecter automatiquement le noyau syllabique accentué grâce à ses caractéristiques prosodiques (forte intensité, maximum local de F_0) [Lacheret et al., 2013].

Le rythme

Le rythme peut se définir comme une séquence alternant des unités fortes et faibles [Lerdahl and Jackendoff, 1982], l'unité choisie est généralement le phonème ou la syllabe, mais peut être défini par d'autres éléments structurants de la parole. Alors que la distinction fort/faible renvoie à une mise en valeur d'une unité par rapport aux autres qui peut être réalisée avec des moyens variés (accentuation en intensité, allongement de la durée, timbre particulier, etc.), d'autres définitions utilisent explicitement la notion de durée. Dans [Vulliamy et al., 1982], le rythme est défini comme "une séquence ordonnée de durées relativement indépendantes de la mesure ou de la structure de la phrase". La "façon dont un ou plusieurs temps non accentués sont groupés ensemble en relation avec un temps accentué" [Magne et al., 2005], permet également de caractériser comment le rythme est perçu en fonction des mécanismes de groupement. Cette notion de regroupement d'unités reste difficilement mesurable même lorsque les unités sont segmentées. Ainsi, la plupart des études vont résumer le rythme à une succession de syllabes que l'on peut segmenter manuellement ou automatiquement. On définit alors un débit syllabique moyen. Il semble évident que la détection du rythme de la parole ne peut pas se limiter à un débit syllabique. Tout comme celle de la musique, l'organisation temporelle de la parole est un processus multi-échelle [Campbell, 2000b] comme représenté sur la figure 2.1 où la prosodie et le timbre sont des éléments structurants.

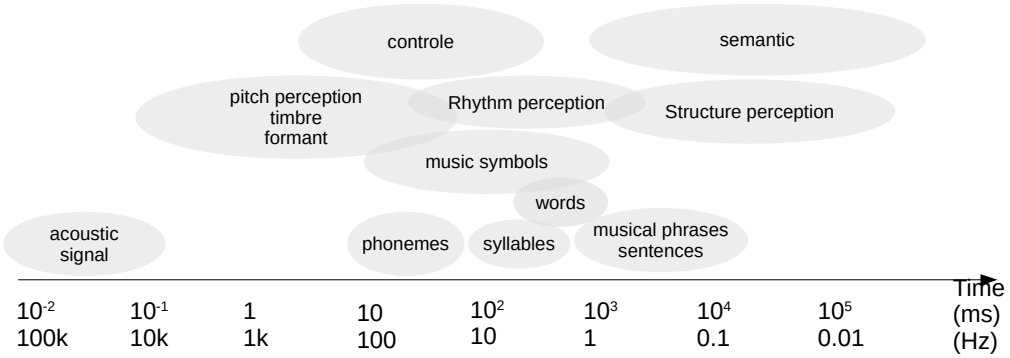


FIGURE 2.1 – Représentation des différentes unités rythmiques pour les signaux audio (parole et musique).

Variable physique, vs. variable perçue

La fréquence perçue par l'oreille humaine - *pitch* peut être différente de la F_0. En effet, l'auditeur perçoit un ensemble de paramètres physiques : énergie perçue - *loudness*, qualité de voix, durée, modulation de fréquence et d'amplitude [d'Alessandro and Mertens, 1995]. Un exemple simple est le tremolo utilisé par le chanteur. L'auditeur n'entend pas précisément cette modulation de la fréquence fondamentale mais une fréquence moyenne. La mesure de F_0 sera donc à manipuler avec précaution lors de l'interprétation des résultats. La F_0 étant perçue par l'oreille humaine sur une échelle logarithmique, on préfèrera transposer la fréquence en semiton avec une référence par exemple, à $La_2 = 110$ Hz (eq. 2.1).

$$f_{st} = 12 \cdot \log\left(\frac{f_{Hz}}{110}\right) \tag{2.1}$$

Toutes les variations de F_0 ne sont pas perçues. T'Hart [t'Hart, 1981] estime à 3 semitons l'intervalle de fréquence perçu en situation de communication. Suivant la durée considérée, un changement de fréquence pourra être plus ou moins perçu.

Ces deux exemples illustrent le fait que les variations d'un descripteur audio peuvent tout à fait ne pas être perçues par l'auditeur. Ainsi les machines seront capables de détecter des phénomènes que nos oreilles ne perçoivent pas forcément.

2.1.2 Descripteurs spectraux

Le signal de parole varie rapidement au cours du temps, c'est un signal non stationnaire. Néanmoins, on considère généralement que sur une fenêtre temporelle de l'ordre de la dizaine de ms (typiquement 30 ms), ce signal est stationnaire. Cette hypothèse se vérifie pour les voyelles et les consonnes fricatives. Elle est remise en question dans le cas des signaux impulsionnels de type plosives.

Représentations temps/fréquence

À partir de cette hypothèse, le signal temporel sera découpé en trames d'une durée typique de 30 ms par pas de 10 ms. Le signal étant considéré comme stationnaire sur une trame, il est pertinent de calculer une Transformée de Fourier (TF) de cette trame et ainsi analyser son contenu spectral. La représentation temps/fréquence de type spectrogramme s'est donc imposée pour le traitement de la parole. Le nom scientifique de l'opération mathématique associée à cet outil est la Transformée de Fourier à Court Terme (STFT en anglais). Ce nom provient de l'analyse effectuée sur des fenêtres de support temporel fini. Le signal temporel est alors représenté sous la forme d'une image où les abscisses représentent le temps, les ordonnées la fréquence et les couleurs l'amplitude du module du spectre. Cette représentation temporelle est un compromis entre la précision temporelle Δt et la précision fréquentielle Δf. En effet, à fréquence d'échantillonnage constante, $\Delta f \times \Delta t = c^{te}$.

Représentation cepstrale

La représentation cepstrale a été conçue pour représenter les modèles source-filtre comme celui de la parole. Soit f la fréquence, on considèrera la Transformée de Fourier (TF) du signal source $G(f)$, la TF du filtre $H(f)$. Le signal filtré se définit alors comme $X(f) = G(f) \times H(f)$. Le cepstre réel est la Transformée de Fourier Inverse (TF^{-1}) du signal fréquentiel $X(f)$ suivant l'équation 2.2 où τ est la quéfrence (homogène à un temps). Grâce à l'introduction du logarithme, le cepstre réel peut s'écrire comme la somme de deux termes : l'un correspond à la source, l'autre au filtre. La représentation cesptrale permet donc théoriquement de séparer la contribution du conduit vocal et de la source à l'aide d'un filtre défini dans le domaine des quéfrences (le liftre ! !).

$$c(\tau) = TF^{-1} \log |X(f)| = TF^{-1} \log |G(f)| + TF^{-1} \log |H(f)| \tag{2.2}$$

Dans le cas des signaux numériques, on définira le spectre du signal filtré par X_k avec

$k \in [1, N]$, N étant le nombre d'échantillons de la TF Discrète (DFT). Les valeurs discrètes du cepstre c_n sont alors données par la transformée discrète inverse (IDFT) du logarithme du spectre X_k (eq. 2.3).

$$c_n = \frac{1}{N} \sum_{k=0}^{N-1} \log |X_k| e^{2j\pi kn/N} = \frac{1}{N} \sum_{k=0}^{N-1} \log |\sum_{m=0}^{N-1} x_m e^{-2j\pi km/N}| e^{2j\pi kn/N} \tag{2.3}$$

La figure 2.2 illustre les étapes de transformation qui permettent d'obtenir le cepstre. À gauche en haut, le spectre d'amplitude X_k du signal discret x_m échantillonné à $Fs = 8$ kHz, est représenté en fonction de la fréquence donnée en Hertz. Seul le spectre positif est représenté ici pour des fréquences allant de 0 à $Fs/2 = 4000$ Hz. Ici, la DFT est calculée sur 1024 échantillons, ce qui correspond à une trame de $\Delta t = 512/F_s = 64$ ms. On peut remarquer sur cette représentation une composante harmonique dont la fréquence fondamentale est $F_0 \simeq 128$ Hz. Au milieu, est représenté le spectre d'amplitude logarithmique $\hat{X}_k$ sur la même échelle de fréquence. Le cepstre c_n est représenté en bas en fonction de la quéfrence τ comprise entre 0 et 256. Sur cette représentation, on peut distinguer un pic pour $n = 61$. Ce pic permet de retrouver la contribution de la source. En effet, on a $F_0 = F_s/61 = 131$ Hz. Le cepstre a été largement exploité pour détecter la F_0 [Rabiner and Schafer, 2007].

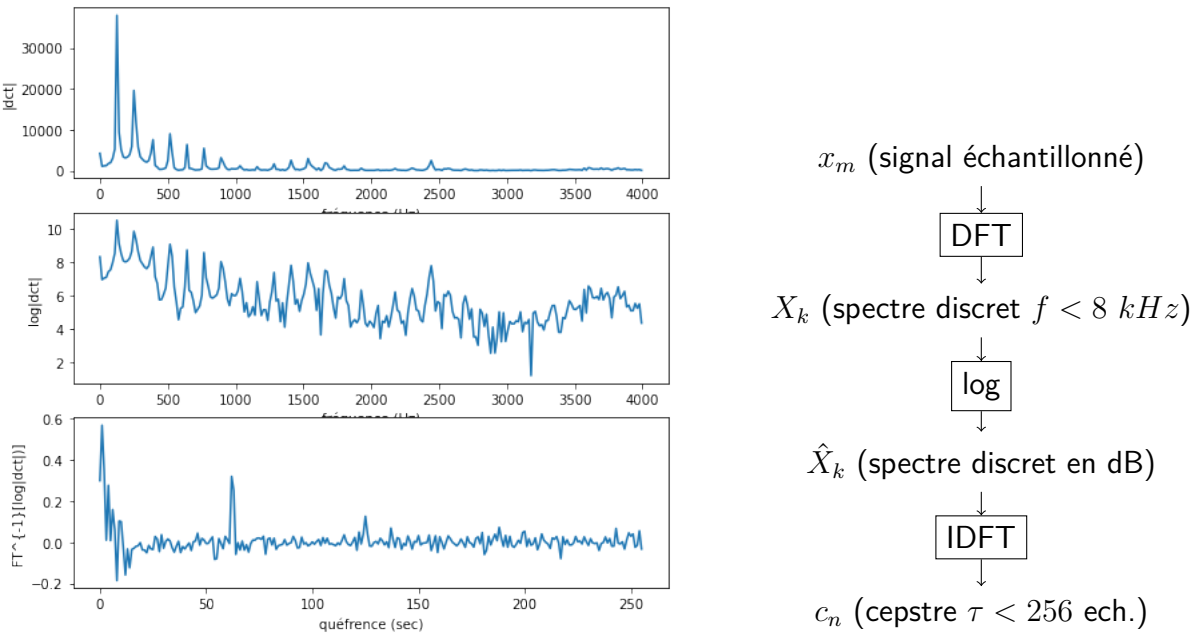


FIGURE 2.2 – De haut en bas (gauche) : spectre linéaire X_k, spectre logarithmique $\hat{X}_k$, cepstre c_n [Taylor] et chaîne de traitement associée (droite)

Les Mel-Frequency Cepstral Coefficients Les MFCCs [Davis and Mermelstein, 1980] constituent sans aucun doute la paramétrisation non probabiliste du spectre la plus répandue pour le traitement des signaux audio. Le signal temporel est préalablement filtré afin de rehausser les basses fréquences (*pre-emphasis filter*), puis segmenté en trames à l'aide d'une fenêtre de Hanning ou Hamming (qui réduit les lobes spectraux résiduels de la fenêtre rectangulaire). Le spectre de chaque trame est obtenu avec une DFT. Les fréquences de Mel modélisent la perception humaine des hauteurs et sont définies par la formule 2.4 où les coefficients peuvent prendre des valeurs plus ou moins différentes suivant les auteurs.

$$Mel(f) = \begin{cases} 1000 \log_2 \left(1 + \frac{f}{1000}\right) & f \geq 1000 \\ f & f < 1000 \end{cases} \tag{2.4}$$

Le spectre par trame sera ensuite multiplié par plusieurs filtres de type passe-bande centrés sur des fréquences de Mel. Ce banc de filtre contient R filtres triangulaires entre f_{min} et f_{max}. Les fréquences centrales des filtres sont calculées de sorte à obtenir R intervalles égaux sur une échelle Mel. Sur une bande passante de 4000 Hz, on pourra utiliser typiquement $R = 20$ filtres [Rabiner and Schafer, 2007].

Le module du spectre obtenu par un banc de filtres permet de réduire la dimension en calculant la valeur moyenne Y_r de l'amplitude pour chaque filtre (eq. 2.5). On multiplie ensuite par 20 le logarithme de cette valeur moyenne. Les valeurs $\hat{Y_r}$ obtenues à cette étape peuvent être utilisées telles qu'elles en entrée des systèmes de traitement automatique, on parle alors de représentation en banc de filtres mel logarithmiques (*log mel filter banks* (FB)). Enfin, la Transformée Cosinus Discrète (DCT) permet d'obtenir R coefficients MFCCs. Cette dernière opération a pour objectif de décorréler les coefficients, elle s'approche d'une décomposition en valeurs propres. Une dernière étape de normalisation des coefficients peut être réalisée.

$$Y_r = \sum_{k=0}^{N} Mel_{r,k} \times |X_k| \tag{2.5}$$

$$\hat{Y_r} = 20 \log(Y_r) \tag{2.6}$$

$$MFCC_m = \frac{1}{R} \sum_{r=1}^{R} \hat{Y_r} \cdot \cos\left(\frac{2\pi}{R}\left(r + \frac{1}{2}\right) m\right) \tag{2.7}$$

Plusieurs remarques découlent de cette description :

1. Si les MFCCs sont construits pour être décorrélés ce n'est pas le cas des représentations

temps/fréquence, ni même des filter banks.

2. Le coefficient 0 (MFCC0) est proportionnel à l'énergie du signal et est donc souvent supprimé pour les raisons citées plus haut.
3. Les mel-filter banks permettent de reconstruire l'enveloppe spectrale avec une précision d'autant meilleure que les fréquences sont basses (échelle Mel) [Rabiner and Schafer, 2007]

C'est principalement pour ces raisons que les coefficients MFCCs ont été très largement utilisés en traitement automatique de la parole et le sont toujours aujourd'hui.

2.1.3 Représentations classiques pour le traitement de la parole

Dans une grande majorité, les systèmes de traitement automatique de la parole vont utiliser une représentation temps/fréquence à court terme du signal audio. Par exemple, pour la transcription de la parole, on cherche à récupérer la signature acoustique des phonèmes. Le contenu spectral est privilégié afin de récupérer les formants des sons voisés, et la distribution spectrale de l'énergie pour les consonnes. Traditionnellement ce domaine pionnier utilise les MFCCs.

Pour l'identification du locuteur, on cherche à être indépendant du contenu linguistique. Par contre, les paramètres acoustiques de la source (F_0) sont importants. On prend alors plus des coefficients cepstraux (entre 12 et 15), auxquels on peut ajouter la F_0 extraite indépendamment et éventuellement l'énergie [Kinnunen and Li, 2010].

Pour la synthèse de parole, on utilise l'enveloppe spectrale afin de générer les sons de la parole, mais également la F_0 et la présence ou non de voisement afin de commander l'intonation. L'enveloppe spectrale doit ici modéliser à la fois les sons de la parole et la voix de synthèse. On cherche alors à avoir une précision spectrale importante, comme dans le cas du vocodeur WORLD, un mel-cepstre de dimension 60 [Morise et al., 2016].

En ce qui concerne la parole expressive, plusieurs types de descripteurs spectraux sont utilisés. En plus des représentations temps/fréquence, on va s'attacher à décrire la distribution d'énergie spectrale à l'aide de barycentres spectraux, ou/et de fréquences roll-off à x% [Clavel et al., 2008]. Ces types de descripteurs initialement conçus pour étudier les timbres instrumentaux [Peeters, 2004], sont maintenant aussi utilisés pour l'analyse de la parole émotionnelle [Eyben et al., 2010]. Ils sont supposés capturer la qualité vocale du locuteur, mais au vu du grand nombre de paramètres spectraux et de la diversité des modifications de timbre, il reste difficile d'évaluer lesquels sont les plus pertinents. Beaucoup d'auteurs utilisent les MFCCs pour les avantages décrits ci-dessus. Plus que les valeurs statiques des coefficients de

la trame courante, c'est l'évolution temporelle de ces descripteurs entre deux trames consécutives (Δ et $\Delta\Delta$) qui semble pertinente. Malgré toutes les initiatives pour proposer des représentations spectrales à court terme, les MFCCs introduits dès les années 80' semblent difficiles à remplacer en pratique.

Ensembles de descripteurs audio Lorsque l'on souhaite représenter le signal audio sous forme de descripteurs, il y a un compromis à trouver entre les descripteurs de spectre à court terme, l'énergie et la F_0, qui sont faciles à extraire, relativement indépendants du contenu linguistique et de la langue et peuvent être extraits en temps réel, et les descripteurs prosodiques perceptifs tels que le rythme ou débits, le pitch (F_0 perçue) ou des descripteurs de qualité vocale qui ont l'avantage d'être très pertinents pour certaines tâches spécifiques mais généralisent peu et sont souvent coûteux à extraire.

A l'heure actuelle, plusieurs ensembles de descripteurs audio (spectre et prosodie) ont été proposés comme compromis. Ceux-ci sont conçus comme des outils intégrés dans un langage commun et représentent la trame ou un segment temporel sous forme d'un vecteur de dimension élevée. Les ensembles de descripteurs ont été développés pour le traitement automatique de la musique et de la parole, ces deux types de signaux présentant un certain nombre de points communs.

Parmi les différents ensembles de descripteurs proposés, nous pouvons citer :

— IrcamDescriptor [Peeters, 2004] conçu pour la classification des sons musicaux,

— OpenSmile [Schuller et al., 2009][Eyben et al., 2010] conçu initialement pour la reconnaissance des émotions dans la parole,

— Librosa [McFee et al., 2015] conçu initialement pour le traitement du signal audio et musical,

— Torchaudio [Yang et al., 2022] conçu pour le traitement de l'audio et de la parole intégré avec les réseaux de neurones,

— ...

2.1.4 Variations acoustiques en contexte expressif

Comme nous l'avons évoqué dans le chapitre précédent, l'expression d'une émotion va impliquer des changements sur la respiration, la phonation et l'articulation. L'ensemble de l'appareil vocal va être modifié, ce qui va entraîner des variations au niveau des descripteurs

acoustiques. Selon [Banse and Scherer, 1996] les variables acoustiques suivantes sont clairement impliquées dans l'expression vocale des émotions : a) le niveau, l'étendue de la F_0 et son contenu intonatif ; l'énergie vocale ; la répartition du contenu fréquentiel, (incluant donc timbre et qualité vocale) ; la position des formants (incluant l'articulation) ; l'organisation temporelle (incluant le tempo et les pauses). Il est indéniable que la qualité de voix, et le timbre seront influencés par l'expressivité présente dans la voix. Dans le cas du chant lyrique, des descripteurs qualitatifs ont été proposés [Garnier et al., 2004] afin de mieux caractériser les différences de timbre entre plusieurs chanteurs, suivant les émotions qu'ils expriment. Par exemple, on parle de voix éraillée, sombre/claire, sourde/brillante, profonde, dure, nasillarde, stridente, grinçante, avec du souffle... Des caractéristiques acoustiques ont été associées à certains de ces qualificatifs. Par exemple, la brillance se définit par la position du centre de gravité des composantes du spectre. L'aspect détimbré correspond à une atténuation spectrale entre 2 et 4 kHz ou au delà de 4 kHz. Autre exemple, le phénomène de tension des muscles vocaux se mesure à partir d'un coefficient de relaxation proposé par Degottex [Degottex et al., 2011]. [Tahon, 2012] et [Busso and Rahman, 2012] confirment la pertinence des paramètres glottiques (coefficient de relaxation et F_0) pour la détection de la valence émotionnelle.

Plusieurs auteurs [Scherer, 2003], [Cowie et al., 2001] ont mis en relation un ensemble de descripteurs audio prosodiques et spectraux et différentes émotions à l'aide de corrélations statistiques ou bien en utilisant une sélection automatique avec des techniques d'apprentissage sur plusieurs corpus de parole émotionnelle [Tahon, 2012], [Eyben et al., 2016]. Cependant, les descripteurs les plus pertinents pour une tâche avec un corpus donné, ne sont pas nécessairement généralisables à d'autres données. Dans ce domaine, les MFCCs restent les descripteurs les plus robustes au changement de données [Tahon and Devillers, 2016b].

2.2 Parole expressive et interactions

Que ce soit dans le cas de la parole émotionnelle ou dans des informations radiophoniques ou télévisées, les locuteurs sont dans des situations de communication en interaction avec d'autres personnes. Dans cette section, nous discutons des phénomènes interactionnels qui ont un impact sur le signal de parole expressive.

2.2.1 Prendre la parole

Dans le cas d'une discussion entre plusieurs locuteurs, le signal de parole expressive pourra se découper en segments monolocuteurs, segments de parole superposée, ou bien du bruit de fond (incluant le brouhaha). L'analyse de la parole conversationnelle a pour objectif de caractériser le phénomène communicatif et d'étudier les interactions entre les différents intervenants, c'est-à-dire caractériser les règles d'enchainement de ces différents segments.

L'action de prendre la parole est un phénomène associé à une situation de communication (compétition, collaboration et charge cognitive). En cela, elle dépend de la situation elle-même (rôle, engagement, domination et état émotionnel), et implique généralement des indices non-verbaux, tels que les *affect bursts* [Scherer, 1994], comme par exemple, le fait de se racler la gorge, des silences, des interruptions et de la parole superposée [Heldner and Edlund, 2010].

Nous avons vu précédemment que les rôles pris par les intervenants avaient une influence sur leur position de domination et leur contrôle de la situation de communication. Ainsi, un intervenant dominant aura tendance à interrompre le discours des autres, par exemple lorsque le journaliste va chercher à contredire un.e interviewé.e. Les interruptions peuvent être vues comme des indicateurs de la puissance, du contrôle ou de la domination d'un locuteur sur ses interlocuteurs. Selon Goldberg [Goldberg, 1990], les interruptions ne sont pas nécessairement compétitives, mais peuvent être neutres (demande de clarification) ou collaboratives (ajout d'information complémentaire). Cette taxonomie a été reprise pour l'annotation des interruptions dans un corpus de parole journalistique [Adda-Decker et al., 2008]. D'autres auteurs [Gravano and Hirschberg, 2012] précisent que les interruptions ont plus de chance d'avoir lieu après l'apparition de certaines propriétés prosodiques. Il y a donc bien des éléments dans le signal sonore qui nous renseignent sur la présence possible d'une interruption. Et celle-ci favorise la parole superposée [Adda-Decker et al., 2008].

Parole superposée

On considère qu'il y a parole superposée lorsqu'au moins deux locuteurs parlent simultanément à des intensités sonores comparables. Dans le cas où un locuteur parle au milieu d'un brouhaha, même si celui-ci est constitué de la parole de plusieurs locuteurs simultané, on ne parlera pas de parole superposée. Dans le cas où un locuteur est traduit simultanément dans une langue, on pourra considérer de la parole superposée.

2.2.2 Disfluences

La présence d'hésitations dans le discours est d'autant plus importante que celui-ci est spontané. Dans la parole lue et préparée, les hésitations sont généralement actées afin d'exprimer un trait caractéristique. Dans la littérature scientifique, les hésitations dans la parole se caractérisent par les durées des pauses silencieuses et remplies (*fillers*) telles que décrit dans Maclay and Osgood [Maclay and Osgood, 1959].

Dans l'oral spontané, des phénomènes d'énonciation particuliers, appellés *disfluences* se caractérisent par l'"endroit où le déroulement syntagmatique est brisé" [Blanche-Benveniste, 1990]. Ces *disfluences* sont donc des ruptures dans l'énoncé. Elles peuvent faciliter la synchronisation entre les différents interlocuteurs [Clark, 1999], ou mettre en évidence une forte hésitation. On peut identifier plusieurs types de *disfluences* dans le discours suivant la catégorisation proposée par [Dutrey, 2014] :

— hésitations vocaliques, ou encore pauses remplies (*fillers*) (mmh, euh),

— marqueurs discoursifs [Bove, 2008] : connecteurs (mais, donc), particules discoursives (bon, voilà, ben), phatiques (tu sais, hein), régulateurs (oui, d'accord, je vois), locutions particulières (de toute façon), verbes parenthétiques (je veux dire, je crois que, je pense que),

— amorces, ou fragments de mots commencées et soit complétés, réparés ou inachevés [Pallaud and Henry, 2004],

— les disfluences d'édition telles que proposées par Shriberg [Shriberg, 1999] : les répétitions, révisions et faux départs.

Pour une revue détaillée des phénomènes de disfluences dans la parole spontanée conversationnelle, le lecture pourra se référer aux travaux de Dutrey [Dutrey, 2014]. Plusieurs études ont analysé les manifestations acoustico-prosodiques (durée des pauses, et des syllabes, variations de F_0) [Campione and Véronis, 2005] [Carlson et al., 2006], [Pon-Barry and Shieber, 2011], [Bell et al., 2003] et phonétiques [Shriberg, 1999] des disfluences dans la parole. Une étude spécifique aux informations télévisées et radiophoniques montre que la présence de répétitions, révisions et pauses chez les interviewé.e.s est proche statistiquement de celle contenue dans des conversations téléphoniques spontanées [Boula de Mareüil et al., 2005]. L'état émotionnel ainsi que la situation de communication (spontanée, préparée) va impliquer plus ou moins de *disfluences*. Ces perturbations du discours vont avoir un effet à la fois sur le contenu linguistique et sur le contenu acoustique et phonétique. Dans un contexte interactionnel, les disfluences structurent les prises de parole, notamment par le biais des interruptions. Les lieux de prise de

parole sont donc très fortement corrélées à la présence de *disfluences*.

Nous pouvons donc affirmer que l'étude de la parole expressive ne peut pas se limiter au seul contenu acoustique et spectral, mais doit prendre en compte la structure temporelle de la phrase au sens large.

2.2.3 Identité, genre, âge

La parole expressive en contexte d'interaction implique plusieurs locuteurs. Ces locuteurs se caractérisent acoustiquement par des informations para-linguistiques (au-delà du contenu linguistique). En effet, à partir du signal audio, et particulièrement du timbre de la voix, nous pouvons identifier les locuteurs, ainsi que leur genre. Il est évident que l'âge modifie la physiologie du conduit vocal, principalement entre l'enfance et l'âge adulte. Les personnes âgées peuvent voir leur appareil vocal s'altérer plus ou moins fortement au cours du temps à cause de difficultés respiratoires, de changements au niveau du larynx, ou de modifications des tissus vocaux entrainant une vibration atypique des cordes vocales [Vaca et al., 2015]. Ainsi, en dehors des enfants, l'identification de l'âge reste une tâche difficile, car la physiologie de chaque individu évolue différemment au cours du temps. Cela implique une difficulté supplémentaire pour l'identification des locuteurs lorsqu'on travaille avec des personnalités médiatiques comme des hommes ou femmes politiques, des journalistes de renom, sur un temps long. En effet, les données journalistiques fournies par des archivistes comme l'Institut National de l'Audiovisuel (INA) peuvent s'étaler sur des périodes temporelles de plusieurs années : l'identité reste la même, tandis que le timbre se modifie graduellement.

Il existe tout un champ d'étude autour de la caractérisation des voix pathologiques, dont le lecteur trouvera une revue détaillée dans [Fredouille, 2016]. Les challenges Interspeech successifs [Schuller et al., 2009] [Schuller et al., 2013] ont eu pour objectif de proposer des standards d'évaluation pour différentes informations para-linguistiques : émotions, signal social, conflit, personnalité, dépression, etc... Ces challenges ont également permis de développer des corpus de référence pour ce type de tâche.

Le choix des données dans ce domaine est crucial, car la subjectivité du phénomène nous limite dans les possibilités de généralisation des résultats obtenus. Les expressions humaines sont variées et diverses, elles dépendent des individus, de leurs expériences de vie, mais aussi du contexte socio-culturel. Afin de généraliser l'expression d'une émotion ou d'un trait para-linguistique particulier, il convient d'être particulièrement vigilant à l'échantillonnage statistique. Par exemple, les modèles appris sur des données prototypiques auront du mal à généraliser leurs décisions sur des données plus variables.

Dans la mesure du possible, on cherche donc à tirer des conclusions sur un grand nombre de locuteurs, éventuellement parlant des langues différentes dont les données sont collectées dans des contextes proches de ceux visés par la tâche finale. Dans ces conditions, on peut démêler les différentes informations para-linguistiques, afin de proposer des conclusions pour des groupes spécifiques d'individus.

Pour conclure, il est compliqué d'obtenir des représentations statistiquement comparables de chaque groupe d'individus. Par exemple, les femmes journalistes sont clairement minoritaires dans les médias. Il faudra donc être conscient des difficultés de représentation et des biais ainsi engendrés dans les modèles.

2.3 Conclusion

Malgré une évolution des modèles de production acoustique dans le domaine de la physique, les descripteurs audio utilisés pour représentation un signal de parole n'ont pas beaucoup évolués depuis le modèle source-filtre et les paramètres prosodiques. Nous avons vu dans ce chapitre qu'un changement d'état émotionnel impliquait des changements importants aux niveaux physiologique et cognitif. Ces changements sont capturés par des descripteurs audio de type spectraux (au niveau du phonème) ou prosodique (au niveau du groupe de souffle). Ils provoquent également des variations au niveau de l'organisation temporelle de la parole. Ces variations sont capturées à un niveau plus large et dépendent fortement du contexte de l'interaction. On peut les considérer comme des informations linguistiques ou phonétiques, ou bien les voir comme une variation rythmique au niveau de la structure. Nous avons également évoqué des limitations spécifiques au domaine liée à la variabilité des expressions humaines et à l'incertitude induite par la subjectivité de la perception. Ces limitations peuvent vite engendrer des biais de représentation qui fausseront les modèles appris sur ce type de données. Il convient donc d'être vigilant quant à la généralisation des résultats obtenus sur des données spécifiques.

Deuxième partie

Segmentation, caractérisation et synthèse

Chapitre 3

Collecte et pré-traitement du signal audio

Ce chapitre aborde les aspects bas-niveau du traitement automatique de la parole expressive. Dans le cadre de la reconnaissance d'émotions, les modèles sont généralement appris sur des segments audio mono locuteur sans bruit autre que la parole. En synthèse de parole, on cherche à générer un signal de parole d'un (ou plusieurs) locuteur(s) cible(s). Dans les deux cas, il est donc nécessaire de segmenter nos données en amont afin de les rendre compatibles avec les différentes tâches. La collecte de données pertinentes, suffisamment variées et utilisables pour apprendre un système est une étape primordiale qui demande à la fois une bonne connaissance de l'objet d'étude (ici parole expressive) et des mécanismes d'apprentissage des systèmes (section 3.1). Trois tâches de segmentation sont abordées dans ce chapitre : la détection de parole et de parole superposée (section 3.2), puis la caractérisation du locuteur (section 3.3) de son genre et de son identité, et enfin la segmentation et regroupement en locuteurs (SRL) dans la section 3.4. Les spécificités de ces trois tâches dans un contexte de parole expressive sont exposées et discutées.

Les résultats présentés dans cette partie ont été réalisés au LIUM. Les travaux autour de la détection de parole et de parole superposée sont menés dans le cadre de la thèse de Martin Lebourdais (ANR GEM). Mes premières recherches sur l'identification du locuteur ont été réalisées pendant ma thèse au LIMSI, tandis que les techniques et résultats récents ont été obtenus dans le cadre du projet ExTENSoR (post-doctorat d'Ambuj Merhish).

3.1 Collecte des données de parole

3.1.1 Styles de parole

Avant de discuter des protocoles de segmentation du signal de parole, il est important d'analyser les conditions de collecte et d'enregistrement de ces signaux. La voix en tant que

donnée personnelle est soumise à la Réglementation Générale de Protection des Données (RGPD) ce qui impose un cadre précis de collecte. Nous proposons de décrire les différents styles de parole rencontrés dans les corpus existants.

Parole spontanée

La parole spontanée est celle qui a lieu dans des conditions réelles où le participant est libre de s'exprimer comme il le souhaite tout en restant dans le contexte de l'interaction dans lequel il communique. Ces conditions réelles permettent de générer des données dites "écologiques", dont une définition est donné par [Douglas-Cowie et al., 2003] et [Campbell, 2000a] :

— les sujets doivent montrer des expressions d'émotions ressenties ;

— les sujets doivent être en situation d'interaction ;

— les expressions des émotions doivent être représentatives de la subtilité des expressions humaines ;

— l'expression des émotions doit être multimodale.

Cette définition rejoint celle proposée par [Devillers et al., 2010] pour les données "*real-life*".

Contexte téléphonique Au fur et à mesure des développements technologiques des moyens de communication, nos actes de parole passent de plus en plus par une captation audio, par exemple pour les conversations téléphoniques, et, plus récemment, les visio-conférences. Ces données sont soumises à la RGPD et leur collecte pose souvent des questions éthiques fortes en rapport avec le respect de la vie privée. Les conversations téléphoniques sont clairement "écologiques", à cela près qu'elles ne sont portées que par la modalité audio. Les données pouvant être collectées sont généralement celles traitées par les centres d'appels téléphoniques, où la diversité des expressions humaines est donc très relative. Depuis quelques années, les applications mobiles (comme par exemple les messages vocaux) collectent également des données de parole dont le contenu reste spécifique au contexte d'utilisation de l'application et au public qui l'utilise.

Contexte des médias Les médias radio/TV impliquent à la fois des orateurs professionnels comme les journalistes et les personnalités politiques, mais aussi des non-professionnels. Dans les débats d'opinion plusieurs participants interagissent les uns avec les autres suivant leur rôle. Même si ces locuteurs sont parfois amenés à sur-exprimer leurs points de vue, nous pouvons considérer qu'il s'agit de parole spontanée car ce ne sont pas des orateurs professionnels.

De même, certaines émissions ou journaux d'information incluent des interviews (plateau ou téléphone) que l'on peut considérer comme une interaction plus ou moins préparée entre le présentateur et l'interviewé.e [Adda-Decker et al., 2008]. Suivant les situations, on peut considérer que ces données sont "écologiques".

Contexte du laboratoire L'enregistrement de données spontanées en laboratoire suivant des protocoles établis au préalable a toujours été réalisé par les chercheurs en parole afin d'étudier des phénomènes précis. Les scénarios possibles sont variés : conversation avec un.e ami.e [Torreira et al., 2010], interview autour d'une thématique particulière [Tahon et al., 2010], etc.. Pour obtenir des contenus linguistiques expressifs plus authentiques que ceux obtenus par un jeu d'acteur, la technique par induction a été proposée [Douglas-Cowie et al., 2003]. Cette technique consiste à stimuler les participants par des éléments contrôlés (musique, photo, tâches à accomplir) [Kossaifi et al., 2021], ou éventuellement à l'aide d'un Magicien d'Oz, c'est-à-dire un opérateur caché derrière une machine qui simulera une interaction humain-robot [Delaborde et al., 2009], [Tahon et al., 2011], [Aubergé et al., 2006]. Certaines équipes de recherche vont mettre en place un environnement écologique pour la collecte multimodale de conversations spontanées comme dans D-ANS [Hennig et al., 2014]. Mais ce type de corpus reste très rare car coûteux en temps et en moyen d'annotation, et la diffusion des données se heurte à des questions éthiques fortes.

Parole actée

Contexte cinématographique Les films sont également une source d'enregistrements de parole. Le fait que les acteurs et les actrices soient des professionnel.les de la parole a l'avantage de fournir des expressions travaillées pour se conformer à un style particulier. Par exemple, jouer des émotions, forcer l'accent de Marseille, etc. De plus, les conditions d'enregistrement sont généralement de bonne qualité, avec peu de bruit de fond et un signal qui n'est pas dégradé par la chaîne de captation [Clavel et al., 2008]. Les données collectées dans ce type de contexte sont généralement prototypiques, les expressions, et émotions sont facilement reconnaissables et peu ambigües en comparaison avec les données naturelles [Erickson et al., 2006], [Laukka et al., 2007].

Contexte du laboratoire Le seul usage des films ne permet pas toujours d'obtenir les expressions souhaitées, et surtout elles ne sont pas forcément contrôlées. Il est alors possible de définir des scénarios dans lesquels les participant.e.s (acteurs et actrices profession-

nel.les ou non) vont devoir acter les expressions demandées, par exemple en doublant des films [Schuller et al., 2010] ou en modifiant le contenu expressif à partir d'un contenu linguistique fixe [Burkhardt et al., 2005]. Ce type de collecte a été largement utilisé pour le traitement automatique de la parole expressive ou non, à partir de la production d'un mot, d'une phrase ou d'un texte de manière actée. Il nécessite de recruter des acteurs et des actrices et de prendre le temps d'enregistrer des signaux sonores suivant un scénario préparé en amont. Par conséquent, ces corpus ne contiennent pas un nombre de locuteurs élevé (un dizaine). Malheureusement, ces données, même si elles ont un contenu linguistique et expressif contrôlé ne sont généralement pas en interaction.

Autres types

Parole lue La parole lue n'implique qu'un locuteur, mais reste néanmoins ancrée dans une situation de communication. L'auditeur ne parle pas, mais son attention doit être captée et maintenue par le lecteur. Il existe donc une interaction entre le lecteur et ses auditeurs, même si ceux-ci ne sont pas forcément présents physiquement (dans les cas des livres audio par exemple). De plus, le lecteur acte les dialogues en incarnant les différents personnages, et exprime des émotions même dans les parties narratives. En cela, plusieurs auteurs ont montré l'intérêt des livres audio pour l'étude de la parole expressive [Montaño and Alías, 2016][Tahon and Lolive, 2018]. Pour une revue complète autour de la caractérisation de l'expressivité en fonction des styles de parole, le lecteur pourra se référer à [Sini, 2020].

Parole préparée Même si le discours du locuteur n'est généralement pas écrit à l'avance, celui-ci l'a préparé. Le locuteur sait globalement ce qu'il doit dire et son discours a été structuré en amont. La parole préparée est largement utilisée dans le contexte des médias par les présentateurs ou présentatrices C'est aussi le cas des conférences TED [Hernandez et al., 2018], des vidéos en ligne et des podcasts [Lotfian and Busso, 2019].

3.1.2 Conditions d'enregistrement et de collecte

Les conditions d'enregistrement et le protocole de collecte des données de parole définissent les traitements nécessaires afin d'obtenir des données exploitables par la suite.

Données issues de centres d'appels téléphoniques Seules les conversations téléphoniques issues de centres d'appels où le consentement est demandé en début de conversation

(de manière plus ou moins implicite) peuvent être collectées. Les inconvénients des données téléphoniques sont leur mauvaise qualité sonore (dont un faible échantillonnage $F_s = 8$ kHz) et le manque de contrôle sur l'origine et le profil des locuteurs, ainsi que sur le contenu linguistique. Les avantages restent la spontanéité du discours et le nombre important de locuteurs.

Les enregistrements conversationnels provenant des centres d'appels contiennent deux canaux séparés dont un pour l'opérateur. Les conversations impliquent rarement plus de deux locuteurs (conversation dyadique), qui sont d'autant plus simples à identifier grâce à la présence des deux pistes. Sur ce type de données, il sera nécessaire de segmenter les zones de parole, par contre une identification des locuteurs ne sera pas utile.

Données issues des archivistes nationaux radio/TV L'archivage des enregistrements de parole a commencé dès la fin du XIXè siècle avec l'apparition du phonographe et du gramophone pour les études linguistiques et phonétiques [Joseph et al., 2022]. Ainsi en France, l'INA archive l'ensemble des contenus radiophoniques et télévisuels des médias français. Les données de parole correspondant aux journaux d'information, aux débats, interviews, etc. sont récupérées en quantité massive. Les contextes et les styles d'énonciation sont variés : parole lue, préparée, conversationnelle, etc. Étant donné l'énorme quantité de donnée stockée par les archivistes, le signal audio est nécessairement compressé (mp3, mp4) et conservé à des faibles fréquences d'échantillonnage (typiquement $F_s = 16$ kHz).

Les enregistrements fournis sont riches et variés en terme de contenus sonores (bruits, parole, musique) et de locuteurs. Afin de récupérer des zones de parole mono-locuteur, il sera donc nécessaire de 1) identifier les zones de parole, 2) segmenter et regrouper les locuteurs y compris sur les zones de parole superposée, 3) identifier les locuteurs quand c'est possible.

Données issues des laboratoires Les données de parole collectées en laboratoire sont généralement bien contrôlées et enregistrées avec des systèmes de captation de bonne qualité. Par contre, la mise en place du scénario, le recrutement des participants et l'enregistrement est très coûteux. C'est pourquoi ces données contiennent généralement un nombre de locuteurs assez limité (une dizaine).

Le protocole d'enregistrement peut impliquer un microphone par locuteur, ou bien des microphones d'ambiance, ou encore des antennes microphoniques. Dans le cas où chaque locuteur à son propre système de captation, il est possible moyennant un seuil d'énergie de récupérer sur chaque piste les zones de parole de chacun. Dans le cas où l'enregistrement d'un seul microphone est disponible pour plusieurs locuteurs, il sera nécessaire de procéder

à une segmentation similaire à celle appliquée aux données radio/TV. La présence d'une antenne microphonique favorisera la segmentation et le regroupement en locuteur (dans le cas où ceux-ci ne se déplacent pas) en indiquant de façon explicite la position du locuteur courant [Mariotte et al., 2022]

Livres audio Le livre audio est une source importante de donnée de parole. De même que pour la parole actée, les enregistrements sont généralement de très bonne qualité. Ce type de collecte permet d'obtenir une quantité importante d'enregistrements d'un unique locuteur. Les initiatives récentes de collecte des données du site LibriVox[1] où des lecteurs anonymes pas forcément professionnels, mettent à disposition leurs enregistrements de livres, ont permis de collecter des données sur un large panel de locuteurs (plus de 1000 pour LibriSpeech [Panayotov et al., 2015]), éventuellement dans plusieurs langues (15 langues différentes pour plus de 100 000h dans MLS [Pratap et al., 2020]).

3.1.3 Corpus de parole pour le traitement automatique

Je souhaite récapituler ici les bases de données existantes que j'ai utilisé lors de mes recherches. Ces bases de données sont utiles à la fois pour apprendre des modèles de traitement automatique de la parole, mais également pour l'analyse et la compréhension du phénomène expressif. Ainsi le tableau 3.1 synthétise les corpus de données conversationnelles, d'archive, de laboratoire et de livres audio et les classe suivant plusieurs critères : durée, langue, nombre de locuteur, expressivité.

Lorsque j'ai commencé ma thèse en 2009, les corpus de parole étaient une ressource encore assez rare. De plus, la RGPD n'étant pas encore en vigueur, il n'était pas évident de savoir dans quelle mesure il était possible de diffuser des données. En particulier, plus l'annotation était longue et coûteuse, moins les données étaient nombreuses, et plus la tendance à monnayer le corpus était forte. C'est le cas des corpus annotés en émotion, seuls quelques corpus actés par une dizaine de locuteurs étaient accessibles. Les corpus annotés ont été majoritairement diffusés via des challenges internationaux pour la transcription, l'identification du locuteur ou la segmentation et regroupement en locuteur.

Cette tendance a été depuis, complètement renversée. Les concepteurs de corpus sont incités à diffuser leurs données et leurs annotations afin qu'elles puissent servir à l'ensemble de la communauté. Cela permet de réduire les différences d'accès aux ressources entre la recherche

1. `https://librivox.org`

publique et privée. Ainsi les corpus vont être téléchargeables sous licence libre (avec parfois des restrictions pour une utilisation marchande) et référencés par des institutions comme ELRA [2], LDC [3] ou sur le site de ressources en français Ortolang [4].

3.2 Segmentation du signal audio : parole, silence, bruit, parole superposée

La plupart des tâches de traitement automatique de la parole, consistent à réaliser une classification sur une échelle temporelle plus ou moins large. Les tâches de segmentation abordées dans cette section ont pour objectif de caractériser des segments de parole suivant leur contenu audio : parole, parole superposée, bruit, musique, etc. Une fois cette segmentation réalisée, nous pouvons par exemple récupérer des segments mono-locuteur non bruités afin d'entraîner des modèles d'identification du locuteur ou bien de reconnaissance d'émotions.

3.2.1 Modélisation acoustique du contenu audio

Je propose de décrire dans cette section deux modélisations largement utilisées par la communauté : la modélisation probabiliste avec des mélanges de Gaussiennes (GMM - *Gaussian Mixture Models*) et la prédiction de séquence d'étiquettes à l'aide d'un réseau de neurone. Cette présentation n'est évidemment pas exhaustive car il existe plusieurs autres méthodes pour la modélisation acoustique, notamment les chaînes de Markov cachées (HMM - *Hidden Markov Models*) les modèles à vecteur support (SVM - *Support Vector Models*) qui ne seront pas détaillées dans ce document. D'autres architectures neuronales seront décrites plus loin notamment pour l'identification du locuteur, la synthèse de parole et la reconnaissance des émotions.

Modélisation probabiliste

La modélisation probabiliste va chercher à décrire explicitement la distribution de probabilité d'une classe. Les GMM permettent d'apprendre une distribution d'une classe donnée à partir d'une représentation vectorielle. Cette représentation peut être extraite d'une

2. European Language Ressource Association
3. Linguistic Data Consortium
4. Outils et Ressources pour un Traitement Optimisé de la LANGue `https://www.ortolang.fr/fr/accueil/ortolang/`

TABLE 3.1 – Listes des différents corpus de parole en fonction des conditions d'enregistrement et de l'expressivité présente dans les données.48

Nom du corpus	Durée	Langue	# Loc. (# émissions)	Expressivité	Applications	Infos
Données conversationnelles (téléphone, réunion)						
AMI [McCowan et al., 2006]	100 h	EN	NA	valence/activation	ASR/SRL/OSD/SSP/	multi-channel
VoxCeleb1&2 [Nagrani et al., 2020]	2794 h	EN	7353		SV	vidéos youtube
DiHARD [Ryant et al., 2021]	67 h	EN/CH	NA		SRL/OSD	multi-domaine/adult/child
AlloSat [Macary et al., 2020b]	37 h	FR	308	satisfaction	ASR/SER	tel
Données d'archive - informations journalistiques						
ESTER [Galliano et al., 2006]	110 h	FR	3059 (157)		SRL	
ETAPE [Gravier et al., 2012]	30 h	FR	688 (73)		OSD/SRL/ASR	
ALLIES [Shamsi et al., 2022]	328 h	FR	5901 (1008)		SRL/SI	
REPERE [Giraudel et al., 2012]	52 h	FR	1518 (291)		SRL/SI	
EPAC [Esteve et al., 2010]	105 h	FR	1935(20505)		ASR/SRL/NER	tel/studio
MSP-podcast [Lotfian and Busso, 2019]	100 h	EN	$>$ 83 (403 podcasts)	va/act/dom	ER	studio, collecte en cours
Données issues des laboratoires						
NCCFr [Torreira et al., 2010]	36 h	FR	46		ASR/SSP/OSD	dyades spontanées
IEMOCAP[Busso et al., 2008]	12 h	EN	10		SER	dyades actées et spontanées
SEWA [Kossaifi et al., 2021]	33 h	Multi (6)	398	val/act/liking	SER	commentaires de publicités
RECOLA [Ringeval et al., 2013]	4 h	FR	46	val/act/social	SER/SSP	dyades via vidéos chat
IDV-HR [Tahon et al., 2011]	2 h	FR	22	val/act/émotionss	SER	Interactions humain-robot
Livres audio						
SynPaFlex [Sini et al., 2018]	87 h	FR	1	émotions/styles	TTS	
LibriSpeech [Panayotov et al., 2015]	982 h	EN	2484		ASR	
LibriTTS [Zen et al., 2019]	585 h	EN	2436		TTS	
LJSpeech [Ito, 2017]	25 h	EN	1		TTS	
M-AILABS [Solak, 2019]	75 h	Multi	2		ASR/TTS	
VCTK [Yamagishi et al., 2012]	44 h	EN	109		TTS	

TTS (*Text to Speech*) : synthèse de parole à partir du texte.
ASR (*Automatic Speech Recognition*) : reconnaissance automatique de la parole.
OSD (*Overlap Speech Detection*) : détection automatique de parole superposée.
SV (*Speaker Verification*) : vérification du locuteur (est-ce que le locuteur est Mr X. ?)
SI (*Speaker Identification*) : identification du locuteur (qui est Mr X. ?)
SRL (*diarization*) : segmentation et regroupement en locuteur.
SER (*Speech Emotion Recognition*) : reconnaissance automatique des émotions dans la parole.
SSP (*Social Signal Processing*) : traitement du signal social.

image, d'une vidéo, d'un signal, etc... Classiquement, la trame de signal de parole est représenté par une observation $\mathbf{x}$ de dimension D contenant (par exemple) entre 10 et 15 coefficients MFCCs [Reynolds et al., 2000]. L'objectif est de modéliser chaque classe de signal (bruit, parole, etc...) par les paramètres de la modélisation probabiliste comme décrit ci-dessous [Reynolds, 1995].

Le mélange de Gaussiennes On considère une superposition de K densités Gaussiennes de la forme de l'équation 3.1 où $p(\mathbf{x})$ est appelée densité marginale de $\mathbf{x} \in \mathcal{R}^D$. Chaque densité Gaussienne $\mathcal{N}(x|\mu_k, \mathbf{\Sigma}_k)$ est appelée une composante du mélange, et possède sa propre moyenne μ_k et covariance $\mathbf{\Sigma}_k$. On notera les paramètres de la densité $\lambda = \{\mu_k, \mathbf{\Sigma}_k\}, k = 1, \ldots, K$.

$$p(\mathbf{x}|\lambda) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \quad \text{tel que} \quad \sum_{k=1}^{K} w_k = 1 \tag{3.1}$$

Étant donné que nous sommes avec des lois normales, la densité $\mathcal{N}(x|\mu_k, \mathbf{\Sigma}_k) = p_k(\mathbf{x})$ peut s'écrire explicitement suivant l'équation 3.2.

$$p_k(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\Sigma_k|^{1/2}} \times \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_k)^T(\mathbf{\Sigma}_k)^{-1}(\mathbf{x} - \mu_k)\right\} \tag{3.2}$$

En repartant de l'équation 3.1, pour une classe c, la probabilité que $\mathbf{x}$ appartienne à cette classe s'exprime suivant l'équation 3.3, où on notera λ_c les paramètres de la densité de la classe c.

$$p(\mathbf{x}|\lambda_c) = \sum_{k=1}^{K} w_{k,c} \cdot p_{k,c}(\mathbf{x}) \tag{3.3}$$

L'apprentissage consistera alors à apprendre les poids $w_{k,c}$ et les paramètres $(\mu_{k,c}, \sigma_{k,c})$ pour chaque classe. Les matrices de covariance devraient être pleines, cependant, plusieurs articles ont montré qu'une matrice diagonale permettait d'obtenir de meilleurs résultats tout en réduisant le nombre de paramètres [Reynolds et al., 2000]. En grande dimension, les paramètres ne peuvent pas être obtenus directement par des formulations analytiques. Leur estimation nécessite un algorithme itératif EM (*expectation, minimization*). À chaque itération, les paramètres du mélange (moyenne, variance, poids) de chacune des composantes sont estimés. Une description détaillée de cet algorithme est donnée dans [Bimbot et al., 2004], [Larcher, 2009].

Prise de décision L'objectif consiste à prédire la classe sachant l'observation $\mathbf{x}$ et non l'inverse. On utilisera alors la formule de Bayes (eq. 3.4).

$$p(c|\mathbf{x}) = \frac{p(c)p(\mathbf{x}|c)}{p(\mathbf{x})} \tag{3.4}$$

Après l'apprentissage, on a C classes chacune représentée par un modèle $\lambda_c, c = 1, \dots C$. Pour une observation acoustique $\mathbf{x}$ en entrée, l'objectif est de trouver quel modèle a la plus grande probabilité a posteriori. On cherche donc à résoudre, à l'aide de la formule de Bayes, le problème suivant donné par l'eq. 3.5.

Si on considère que toutes les classes ont la même chance d'apparaître (ce qui n'est pas forcément vrai), les termes de probabilité a priori $Pr(\lambda_c)$ et $p(\mathbf{x})$ sont constants pour toutes les classes et peuvent ainsi être ignorés dans le calcul du maximum. On ajoutera l'hypothèse comme quoi toutes les observations sont indépendantes du temps, ainsi $p(\mathbf{x}|\lambda_\mathbf{c}) = \Pi_t p(\mathbf{x_t}|\lambda_\mathbf{c})$ et on prendra le logarithme afin d'obtenir la log-vraisemblance donnée dans l'équation 3.7. Le terme $p(\mathbf{x_t}|\lambda_\mathbf{c})$ est défini par l'eq. 3.3. L'hypothèse d'indépendance des trames permet de n'obtenir qu'un seul score pour un segment de durée T trames.

$$\hat{c} = \underset{c=1\dots C}{\operatorname{argmax}} Pr(\lambda_c|\mathbf{x}) \tag{3.5}$$

$$= \underset{c=1\dots C}{\operatorname{argmax}} \frac{p(\mathbf{x}|\lambda_\mathbf{c})}{p(\mathbf{x})} Pr(\lambda_c) \tag{3.6}$$

$$= \underset{c=1\dots C}{\operatorname{argmax}} \sum_{t=1}^{T} \log p(\mathbf{x_t}|\lambda_\mathbf{c}) \tag{3.7}$$

Adaptation Ce type d'apprentissage nécessite d'avoir suffisamment de données par classe pour obtenir des modèles représentatifs pour chacune. Une des limitations majeures de cette approche, est que si les signaux d'une classe sont enregistrés dans des conditions acoustiques particulières, les modèles capturent aussi cet environnement (parfois même plus que l'acoustique de la classe elle-même). Le modèle représente alors la distribution acoustique de la classe et de son environnement et les résultats seront biaisés.

Pour palier au fait que les données peuvent être en très faible quantité pour certaines classes, et pour homogénéiser les paramètres du modèle GMM en fonction de l'environnement acoustique, on peut apprendre un modèle du monde (UBM - *Universal Background Model*) et ensuite l'adapter aux classes en particulier en utilisant l'adaptation MAP (*Maximum a Posteriori* [Barras et al., 2007].

Caractérisation neuronale de séquences audio

Plusieurs tâches de traitement automatique de l'audio bas-niveau consistent à caractériser chaque trame avec une étiquette, généralement binaire :

— Détection d'activité vocale (VAD) : présence ou non de parole (éventuellement avec de la musique, ou de la parole superposée)

— Détection de parole superposée (OSD) : présence ou non de parole superposée

— Détection de changement de locuteur (SCD) : présence ou non d'un changement de locuteur.

Position du problème Il s'agit donc de prédire une séquence de labels de durée équivalente à celle du signal d'entrée. L'objectif est de faire une classification à la trame d'une séquence (*sequence labeling*). Soit $\mathbf{x} \in \mathcal{X}$ une séquence extraite d'un segment audio : $\mathbf{x} = (x_1, x_2, \ldots, x_T)$ avec T le nombre total de trames. $\mathbf{x}$ peut être une séquence de vecteurs de MFCCs, ou bien des mel-fbanks, ou encore un spectrogramme. La tâche consiste alors à prédire la séquence d'étiquettes : $\mathbf{y} = (y_1, y_2, \ldots, y_T) \in \{0, 1\}^T$ telle que, par exemple pour un VAD $y_t = 1$ si il y a de la parole et $y_t = 0$ sinon.

Architecture d'un réseau de neurone récurrent Le modèle implémenté dans le toolkit PyAnnote [Bredin, 2020] pour la détection d'activité vocale (VAD) est un réseau de neurone récurrent [Gelly and Gauvain, 2018]. Nous présentons ici une version adaptée par Martin Lebourdais et Théo Mariotte au LIUM. Le réseau de neurone est constitué de deux couches récurrentes bi-dimensionnelles de type LSTM (*Long Short-term Memory*), puis deux couches linéaires de dimensions 128, et une couche linéaire de dimension 2 qui retourne deux valeurs w_k^t pour chaque trame t. Le réseau prend en entrée des segments audio de durée fixe, typiquement 2 s, soit $T = 200$ trames par pas 10 ms.

Une couche linéaire applique une transformation linéaire sur une entrée x telle que la sortie w est donnée par $w = xA^T + b$ où les poids de la matrice A et le biais b sont les paramètres appris lors de l'entrainement. Pour la dernière couche linéaire par exemple, $x \in \mathbb{R}^{200\times128}$ et $w \in \mathbb{R}^{200\times2}$. 200 est la taille de la séquence, c'est-à-dire le nombre de trames du segment audio fourni en entrée du réseau.

$$w'_k = \text{Softmax}(w_k) = \frac{e^{w_k}}{\sum_k e^{w_k}} \quad w_k = [w_k^1, w_k^2, \ldots, w_k^T] \tag{3.8}$$

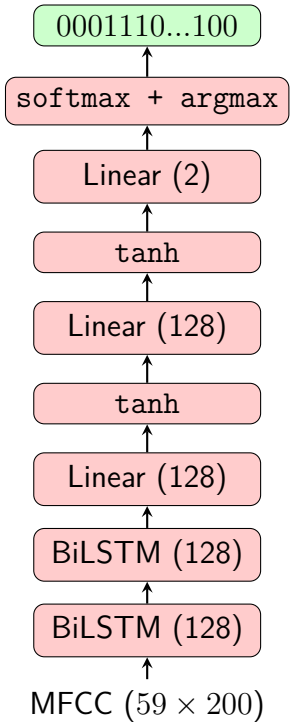


FIGURE 3.1 – Réseau récurrent pour la prédiction de parole superposée [Lebourdais et al., 2022a]

Chaque sortie de cette couche linéaire w, est normalisée entre 0 et 1 grâce à la fonction Softmax définie par l'équation 3.8. La somme des deux éléments obtenus pour chaque trame est unitaire : $w'_{t,k=0} + w'_{t,k=1} = 1$. Ainsi les deux valeurs obtenues par trame sont dépendantes. Cette opération permet d'obtenir une valeur qui peut être interprétée comme une probabilité (mais qui n'en est pas une). Les opérations probabilistes d'entropie pourront être appliquées dessus.

Pour une tâche de classification, on utilise généralement une cross-entropie comme fonction de coût. Celle-ci est définie par l'équation 3.9 où p sera la prédiction et q la référence. Lors de l'apprentissage, la cross-entropie est ensuite retropropagée dans le réseau afin d'adapter les poids du réseau. La fonction de cross-entropie `CrossEntropyLoss` implémentée dans le toolkit PyTorch sera appliquée avant de calculer les softmax avec éventuellement un poids ω_k pour chacune des classes.

$$\mathcal{L} = -p \log q + (1-p) \log(1-q) \tag{3.9}$$

$$= -\sum_k y_k \log w'_k \quad \text{ici} \quad y_1 = y,\ y_2 = 1-y \tag{3.10}$$

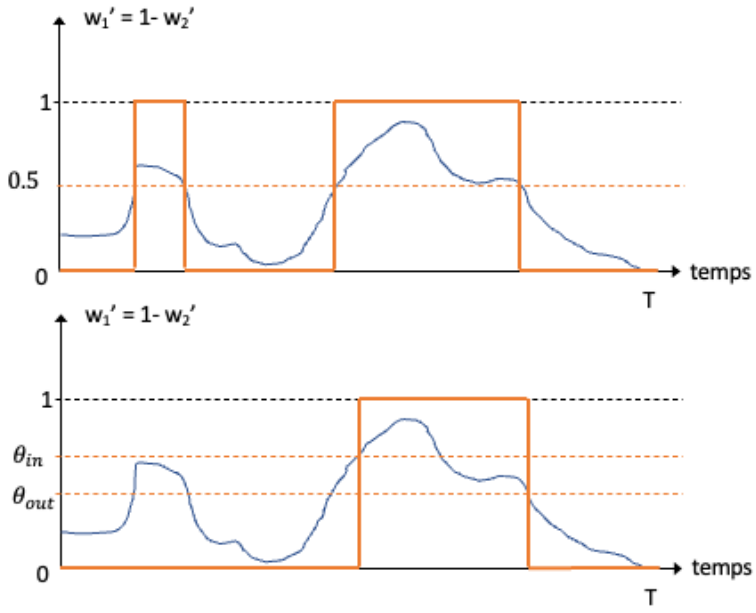


FIGURE 3.2 – Illustration de la prise de décision avec un argmax (haut) et des seuils (bas)

Prise de décision La classe correspondante à la trame t sera donnée par la fonction $y_t = \text{argmax}_{k=1,2}(w'_{t,k})$ qui retourne l'indice de la valeur la plus grande, soit 0 ou 1 pour notre exemple. L'avantage de cette approche est qu'elle permet de faire une classification à plusieurs classes. La classe prédite sera toujours donnée par la fonction argmax. Par contre, elle ne permet pas d'optimiser la sortie en prenant en compte les variations d'échelles pour chacune des sorties de la couche linéaire. Dans le cas binaire, le fait d'utiliser la fonction argmax équivaut à utiliser un seuil de décision à $\theta = 0.5$.

Une autre option, consiste à remplacer la fonction argmax par un seuil θ optimisé sur des données de développement. On peut même raffiner la décision en utilisant deux seuils : un pour la montée θ_{in}, l'autre pour la descente θ_{out} comme illustré sur la figure 3.2.1. On peut

remarquer que dans le cas d'une prise de décision avec un argmax (en haut) on obtient deux segments pour la classe 1, alors que si la décision se fait avec des seuils (en bas), un seul segment de classe 1 est prédit.

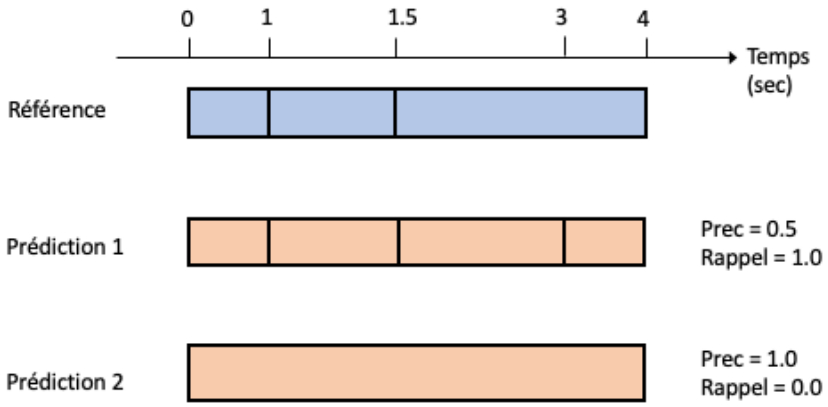


FIGURE 3.3 – Illustration de possibles biais dans les métriques précision et rappel en segmentation.

Évaluation Estimer les performances d'un système de segmentation automatique nécessite d'utiliser plusieurs métriques d'évaluation pour analyser à la fois les erreurs liées à la segmentation elle-même et celles dues aux classes prédites.

Les métriques les plus utilisées sont le taux d'erreur de détection (DetER - *Detection Error Rate*), les fausses alarmes (prédire un segment alors qu'il n'y en a pas), les segments manqués (rater un segment alors qu'il y en a un) et les confusions (prédire un label à la place d'un autre). Ces métriques sont calculées sur les durées des segments concernés. La campagne d'évaluation DiHard impose également l'utilisation des métriques de précision, rappel et f-score calculées soit à la trame, soit en pondérant par la durée du segment concerné. Le taux d'erreur de détection se calcule également en nombre de trames par rapport à la durée totale T du (des) fichiers(s) considéré(s) (voir eq. 3.11).

$$\text{DetER} = \frac{\text{FA} + \text{Miss} + \text{confusion}}{T} \tag{3.11}$$

Dans le cas de l'évaluation de la segmentation, le rappel pénalise les segments manquants et la précision pénalise les segments en trop. Ils apportent des informations complémentaires. Si on considère l'exemple illustré figure 3.3, dans le premier cas, un segment a été divisé en deux, il y a donc un segment en trop (compté en fausse alarme) (précision de $1 + 1.5/(5) = 0.5$), et aucun segment manqué (rappel de 1.0). Dans le second cas, l'ensemble des segments a été détecté, il n'y a donc aucune fausse alarme, mais que des segments manqués, la précision est alors de 1.0 et le rappel de 0.0. Suivant l'application visée, on favorisera le rappel ou la précision. Un compromis est d'utiliser le f-score qui est une pondération des deux métriques.

Limites Cette tâche est la brique de base de la plupart des applications de traitement automatique de la parole. Elle comporte cependant deux limitations majeures. La première est la difficulté à trouver une métrique d'évaluation unique. La bonne connaissance des différentes métriques permet à l'utilisateur d'évaluer honnêtement son système. La seconde difficulté réside dans l'obtention d'une référence de qualité. Malgré le fait que l'annotation parole / non-parole soit souvent considérée comme une tâche objective, finalement, dans un certain nombre de cas, la décision binaire est difficile. Par exemple lorsqu'il y a un bruit de fond important ou de la musique de forte intensité par dessus un signal de parole, à quel point considère-t-on que ce segment appartient à la classe parole ?

3.2.2 Détection d'activité vocale (VAD)

La tâche de détection d'activité vocale (VAD) consiste à identifier les zones de parole dans le signal audio. Suivant les auteurs, les zones de parole superposée sont exclues ou non par le VAD. Historiquement, on détermine les zones de parole grâce à des seuils définis sur des descripteurs acoustiques tels que les MFCCs [Haigh and Mason, 1993], le nombre de passages par 0 (Zero-Crossing Rate) ou encore l'énergie des hautes fréquences et l'intensité du signal. Cependant ces méthodes ne sont pas robustes à des bruits inconnus comme par exemple la musique. Une approche plus récente consiste à apprendre un modèle de type GMM pour chaque type de sons [Barras et al., 2006] : parole seule, parole bruitée, parole avec musique, musique seule et silence. Une fois que chaque segment de parole est caractérisé, on ne conserve que les signaux de parole seule (éventuellement multi-locuteur).

Aujourd'hui, les modèles neuronaux de type séquence vers séquence tels que le modèle

présenté dans la section précédente (section 3.2.1) sont largement utilisés pour le VAD. Ils ont l'avantage de fournir une segmentation précise et robuste à la trame et non plus au segment comme c'était le cas avec les GMM. Ce type de modèle sera d'autant plus performant que la diversité présente dans chacune des classes (speech et non-speech) est grande. Pour augmenter cette diversité, on utilisera de grandes quantités de données comme par exemple l'ensemble d'apprentissage du challenge DiHard III [Sahidullah et al., 2019] et on bruitera artificiellement les données en ajoutant un signal de bruit au signal de parole original, ou bien en ajoutant de la réverbération aux signaux à l'aide de réponses de salles encodées sous la forme de filtres (RIR - *Room Impulse Response*) [Dong and Lee, 2018] ou encore de la parole artificiellement superposée. À titre indicatif, le modèle de prédiction de séquence qui repose sur l'architecture proposée figure 3.1 évalué sur la partition de test du challenge DiHard, obtient de bons résultats en rappel (98.82%) ainsi qu'en precision (92.84%). Soit un f-score global de 95.74%.

3.2.3 Détection des zones de parole superposée (OSD)

Nous avons défini la parole superposée dans le chapitre 1 (section 2.2.1) comme une zone de signal audio où au moins deux locuteurs parlent simultanément à des intensités sonores comparables. La détection automatique de parole superposée a toujours été une tâche difficile qui a des conséquences sur les performances des systèmes de reconnaissance automatique de la parole [Bullock et al., 2020] et de regroupement et segmentation en locuteur [Garcia Perera et al., 2020]. De plus, l'identification de ce type de segment dans un signal audio permet d'avancer sur la caractérisation des interruptions et d'analyser plus précisément la structuration en tours de parole de la conversation. Un des objectifs de la thèse de Martin Lebourdais est de caractériser automatiquement les interactions hommes/femmes dans les médias français. La brique de départ de son travail est justement de détecter les zones de parole superposée.

Etude statistique des annotations en overlap Une des principales difficultés de cette tâche est le grand déséquilibre des classes. En effet, les zones de parole superposée peuvent être plus ou moins nombreuses suivant le type de situation (débat politique, parole spontanée, interview, etc.), mais elles sont de courte durée. Nous avons proposé une étude des annotations dans plusieurs corpus de parole disponibles pour notre tâche, à savoir les données AMI, DiHARD, ESTER1&2, REPERE, EPAC et ETAPE [Lebourdais et al., 2022b].

À partir de la segmentation et des annotations en locuteur, nous avons identifié les zones de parole superposée. La proportion de parole superposée est très variable suivant que nous avons

de la parole spontanée (AMI : 13.87%, DIHARD : 11.60%), des émissions journalistiques et des débats (EPAC : 5.29%, REPERE : 3.36%) ou des informations données par des présentateurs ou présentatrices (ESTER1&2 : 0.67%). Ainsi, suivant les corpus, nous pouvons considérer que la parole superposée est un événement rare (en durée cumulée). Comme le montre la figure 3.4 illustrant la distribution des durées des segments de parole superposée, plus les données sont spontanées (AMI et DiHARD) plus les segments de parole superposée sont courts < 1 s. Les corpus REPERE, ETAPE et EPAC contiennent des présentations d'informations mais aussi des émissions de débats, ce qui explique une part importante des segments de courte durée. Par contre, ESTER1&2 contiennent quasiment exclusivement des informations présentées par des journalistes, des experts et des politiques, la répartition des durées est alors très homogène.

Récemment, nous avons analysé une émission de téléréalité [5] où les zones d'overlap sont très longues. En effet, les participants parlent souvent en même temps mais généralement pas ensemble. Il y a donc très peu d'interruptions.

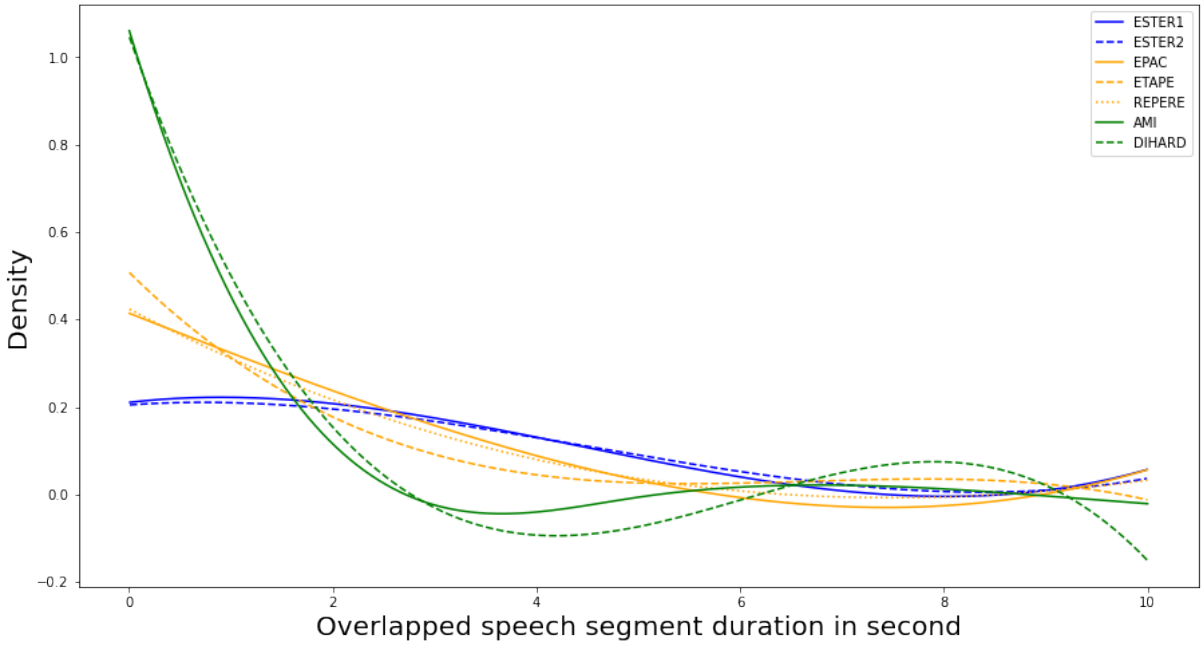


FIGURE 3.4 – Distribution normalisée des durées de segments de parole superposée dans différents corpus de parole. Figure extraite de [Lebourdais et al., 2022b]

5. *Les Marseillais*, émission diffusée sur W9 entre novembre 2012 et mai 2022.

Augmentation artificielle des données Pour entrainer un système à reconnaître les zones de parole superposée, il est nécessaire de remédier au déséquilibre des classes. Pour cela, la technique utilisée dans pyannote [Bredin et al., 2020] est de créer artificiellement des segments de parole superposée en additionnant deux signaux de parole de même durée. La référence s'obtient également en additionnant les références. L'inconvénient de cette méthode est que la superposition n'est pas réaliste : la superposition peut avoir lieu pour un même locuteur, sur des mots tronqués, à des instants qui ne sont a priori pas propices pour une superposition. De plus, elle ne permet pas de contrôler explicitement la répartition entre les segments de silence, de parole, de parole superposée (artificielle ou réelle) et de bruit. Nous avons donc mis en place un module de chargement de données (un `datasampler` intégré à la librairie PyTorch) qui permet de créer un sous-corpus (`batch`) dont les distributions mentionnées sont contrôlées par un pourcentage fourni par l'utilisateur. Des batchs différents seront donnés à chaque itération, chacun respectant la distribution souhaitée.

D'après nos expériences, il semble que le réseau de neurones apprend mieux lorsque les distributions ne sont pas contrôlées comme dans la méthode de pyannote. Nous expliquons ce phénomène contre-intuitif par le fait qu'un contrôle trop précis limite fortement la diversité des données d'apprentissage. Nous avons donc conservé l'augmentation de données par sélection aléatoire par la suite.

Features	System	# Param	Prec	Rappel	F1-score
Baseline [Bredin et al., 2020]			57.2	62.8	59.9
MFCC	LSTM	0.64	34.2	60.8	43.8
MFCC	TCN	0.27M	46.6	59.8	52.4
WavLM	LSTM	1.65M	61.0	63.6	62.3
WavLM	TCN	0.35M	60.1	67.1	63.4

TABLE 3.2 – Résultats issus de [Lebourdais et al., 2022a] obtenus sur la tâche de détection de la parole superposée évaluée en précision, rappel et f1-score (en %) sur la partition d'évaluation du challenge DiHard.

Détection automatique de la parole superposée Pour apprendre un détecteur de parole superposée, nous sommes partis de l'architecture présentée dans la section précédente (section 3.2.1). Pour un segment de parole de durée fixe en entrée, le système doit prédire une séquence d'étiquettes, avec ici $K = 2$: $y_t = 0$ si aucun ou un locuteur parle au temps t, $y_t = 1$ si au moins deux locuteurs parlent. Cette architecture est améliorée en utilisant des modèles de langue pré-entrainés tels que WavLM [Chen et al., 2022] en entrée à la place

des MFCCs, ou bien un réseau TCN (*Temporal Convolutional Network*) [Cornell et al., 2020] à la place des LSTM. Les résultats sur la détection de parole superposée reproduits dans le tableau 3.2, sont très encourageants et les performances obtenues dépassent l'état de l'art [Lebourdais et al., 2022a]. On peut noter toutefois que l'architecture récurrente présentée précédemment permet d'obtenir un bon rappel mais une mauvaise précision et que l'utilisation du TCN permet de réduire le nombre de paramètres du modèle, ce qui peut avoir un intérêt majeur lorsqu'on utilise une représentation acoustique du type WavLM.

Une analyse des erreurs de détection a montré que le système trouvait de la parole superposée en présence d'un traducteur humain en voix off, ou bien de brouhaha ou de musique. La difficulté est que, suivant l'intensité du bruit en arrière plan, les segments ont été annotés soit en bruit, soit en parole. Dans certains corpus, la présence d'un traducteur n'est pas considérée, les segments sont alors annotés soit avec un unique locuteur, soit comme n'étant pas de la parole. Dans d'autres corpus, la présence simultanée d'un traducteur et du locuteur original est annotée comme parole superposée par les annotateurs humains. Cette étude montre que la définition de la parole superposée inclut bien une dimension d'intensité sonore qui reste subjective. Cela confirme la définition proposée pour la parole superposée au chapitre **??**.

3.3 Caractérisation du locuteur

Je propose de décrire ici brièvement les grandes techniques utilisées pour l'identification du locuteur, ainsi que les limitations et contraintes qui peuvent exister en présence de parole expressive. Pour une revue détaillée, le lecteur pourra se référer à [Larcher, 2018].

3.3.1 Modélisation acoustique du locuteur

De même que pour la segmentation du contenu audio, je décris ci-dessous deux types de modélisation qui sont largement utilisées pour l'identification du locuteur, mais également l'identification du genre à partir du signal vocal. L'ensemble des modélisations présentées n'utilisent que le signal acoustique, cependant des approches utilisant le texte peuvent être utiles dans certains cas (pour le lecteur intéressé, voir [Larcher, 2018]).

Quelque soit le type de modélisation, la méthodologie utilisée aujourd'hui pour l'apprentissage et le test en identification du locuteur est la suivante :

1. Apprentissage : apprentissage des paramètres du modèle du monde avec une grande diversité de locuteurs et de conditions acoustiques (UBM).

2. Enrollement : adaptation des paramètres du modèle aux locuteurs connus que l'on cherchera à reconnaître spécifiquement.
3. Identification : identifier le modèle de locuteur qui a la représentation (ou distribution) la plus proche de celle du locuteur de test.

Modélisation probabiliste

L'approche GMM-UBM décrite dans la section précédente a été longtemps la plus utilisée par la communauté. Une classe est définie par l'identité d'un locuteur particulier et est modélisée par un mélange de Gaussiennes. La représentation d'un locuteur correspond alors à une distribution de probabilité définies par les paramètres des Gaussiennes apprises typiquement des séquences de coefficients MFCCs, auxquels on ajoute les dérivées première et seconde. Ce mélange peut être appris directement ou bien adapté à partir d'un modèle du monde UBM. Cette deuxième approche permet de compenser les faibles quantités de données disponibles pour certains locuteurs. Selon [Larcher, 2018], cette modélisation a plusieurs inconvénients :

1. Le nombre de distributions Gaussiennes pour modéliser un phénomène est difficile à estimer.
2. Le paradigme UBM-GMM et l'hypothèse d'indépendance implique que l'ensemble des phonèmes aient le même niveau de complexité, notamment le fait que la probabilité a priori d'un locuteur soit constante sur les locuteurs, ce qui n'est pas vrai si l'on considère des segments courts ayant un contenu linguistique spécifique.
3. Il n'y a pas de modélisation temporelle de la parole. Il sera alors nécessaire d'utiliser des modèles de Markov (non décrits ici).
4. La modélisation est très sensible aux différences d'environnements acoustiques.

Cette approche GMM-UBM a été améliorée en utilisant des super-vecteurs, appelés i-vecteurs qui contiennent l'ensemble des paramètres du mélange de Gaussiennes pour un locuteur [Dehak et al., 2011]. L'apprentissage d'une PLDA (*Probabilistic Linear Discriminant Analysis*) lors de la phase d'enrollement permet de calculer un score de similarité entre ces i-vecteurs et ainsi identifier le locuteur correspondant au segment de parole [Prince and Elder, 2007]. L'utilisation du Factor Anlayser permet de supprimer la variabilité liée à la session, et donc limiter l'effet de l'environnement acoustique sur les performances. Pour plus de détails, se référer aux travaux décrits dans [Prince and Elder, 2007] et [Larcher, 2018].

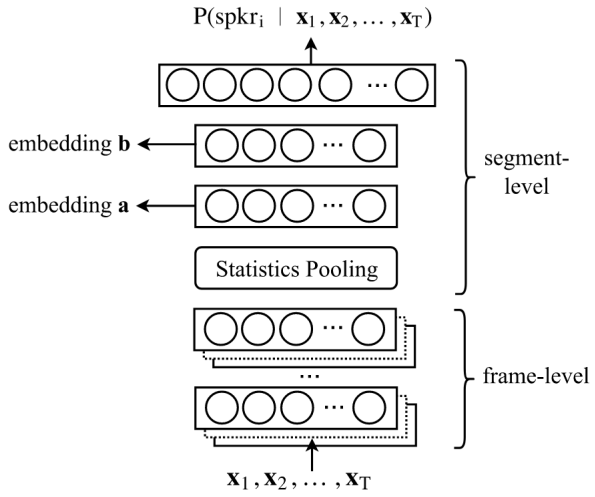


FIGURE 3.5 – Architecture du DNN de type x-vecteur. Figure extraite de [Snyder et al., 2017].

Modélisation par réseau de neurones

La solution actuelle pour la reconnaissance du locuteur consiste à entraîner un réseau de neurones pour une tâche d'identification en milieu fermé (pas de locuteurs inconnus) avec un grand nombre de locuteurs. Ce réseau est l'équivalent d'un modèle UBM. Une fois le réseau entraîné, on utilisera une représentation latente du réseau, appelée *embedding*, qui est supposée représenter le signal audio de façon spécifique à conserver essentiellement les informations nécessaires pour identifier un locuteur. Un vecteur d'*embeddings* peut être vu comme un vecteur de descripteurs acoustiques, sauf que la représentation est implicite et peu interprétable. Les x-vecteurs sont des *embeddings* [Snyder et al., 2017] qui ont montré des résultats très encourageants pour la reconnaissance du locuteur.

Architecture du réseau de neurones [Snyder et al., 2018] a proposé d'apprendre un réseau de neurones pour la reconnaissance du locuteur, en s'inspirant des travaux menés en Reconnaissance Automatique de la Parole (ASR - *Automatic Speech Recognition*). Le réseau représenté figure 3.5 calcule des *embeddings* de locuteurs à partir de segments audio de parole $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$ représentés sous-forme de MFCC de dimension 20. Les cinq premières couches (TDNN - *Time delay neural network*) travaillent au niveau trame, et sont conçues pour capturer un contexte temporel long. Chaque couche opère à une résolution temporelle

différente qui croît lorsqu'on monte dans le réseau [Peddinti et al., 2015].

La couche d'accumulation statistique (*statistics pooling layer*) reçoit les sorties de la dernière couche temporelle et calcule ses moyennes et écart-types. Ces valeurs sont concaténées et fournies en entrée de la couche suivante. Cette accumulation statistique rempli le même rôle que l'utilisation des paramètres (poids, moyenne, variance) du mélange de Gaussiennes vu auparavant. Les *embeddings* obtenus sont extraits des couches qui suivent l'accumulation statistique. Il n'y a donc plus d'information temporelle à proprement parler, mais des informations segmentales relevant de fenêtres temporelles de différentes durées. Les x-vecteurs peuvent donc capturer le timbre du locuteur, sa prononciation, mais également sur un temps plus long son style prosodique.

Le réseau est entrainé pour classifier les locuteurs et retourne la "probabilité" d'appartenance à la classe du locuteur i sachant la séquence d'observations acoustiques $\mathbf{x}$. Par conséquent, la fonction de coût est généralement une cross-entropie multiclasse. L'identification des locuteurs elle-même, est réalisée par une PLDA qui compare des paires d'*embeddings*. Ainsi le réseau et la mesure de similarité peuvent être entrainés sur des données différentes. Récemment, l'utilisation d'une nouvelle fonction de coût *Angular Softmax* utilisée plus classiquement en traitement des images a permis d'importants gains en performance [Garcia-Romero et al., 2019]. Cette fonction fait intervenir non plus une différence d'entropie, mais une différence sur l'angle entre les vecteurs de référence et prédits. Cette fonction de coût est adaptée au calcul d'une similarité cosinus entre deux locuteurs, ce qui a l'avantage d'éviter l'apprentissage d'une PLDA.

Prise de décision Une fois le réseau x-vecteur entraîné sur un grand nombre de locuteurs, on peut procéder à l'étape d'identification. On utilise le réseau pour extraire les *embeddings* des locuteurs connus de l'ensemble d'enrollement. On obtient alors une représentation pour chaque locuteur. Pour un locuteur inconnu de l'ensemble de test, on extrait ses *embeddings* à partir des signaux de parole disponibles. Ces derniers sont ensuite comparés aux *embeddings* des locuteurs connus grâce à une mesure de similarité de type PLDA ou similarité cosinus.

En parallèle de la modification de la fonction de coût permettant l'utilisation d'une similarité cosinus pour la prise de décision, d'autres améliorations du réseau ont été proposées, par exemple l'architecture ResNet [Chen et al., 2017], qui utilise des réseaux résiduels. Ce type de réseau permet d'utiliser un nombre de couches plus important tout en réduisant la complexité. En effet, les auteurs de ResNet [He et al., 2016] proposent d'ajouter des connexions avec des fonctions identités qui court-circuitent le réseau classique ce qui a pour effet, d'après les auteurs, de limiter le sur-apprentissage.

D'autres améliorations sur la représentation acoustique en entrée du réseau ont été proposées. Par exemple, SincNet [Ravanelli and Bengio, 2018] prend en entrée la forme d'onde du signal directement. Les couches basses du réseau sont basées sur des convolutions, ce qui permet de filtrer le signal temporel avec des paramètres de filtres adaptés aux données et à la tâche. Ainsi, le réseau se construit sa propre représentation temporelle de la parole et n'est pas contraint par les représentations temps-fréquences utilisées ailleurs.

Évaluation et intervalle de confiance Dans le cas de l'identification du locuteur, la tâche n'étant pas une classification binaire, il ne sera alors pas pertinent d'utiliser les métriques standards telles que rappel et précision. On préfèrera des métriques qui prennent en compte les proportions de fausses alarmes (FA : on croît reconnaître un locuteur) et les détections manquées (Miss : le locuteur n'a pas été reconnu comme tel).

Les métriques utilisées sont les suivantes :

— Fonction de coût (DCF - *Detection Cost Function*) : $DCF = \alpha P_{miss} + \beta P_{FA}$ où les paramètres α, β sont optimisés sur les données de développement ou bien choisis dans le cadre d'une tâche particulière.

— Courbe de détection d'erreurs (DET - *Detection error Tradoff*) : elle représente la proportion P_{miss} en fonction de P_{FA} paramétrées par le seuil de décision.

— Taux d'égales erreurs (EER : *Equal Error Rate*) : la valeur de faux positifs pour laquelle il y a autant de faux positifs que de faux négatifs.

Ces métriques sont des compromis entre le rappel et la précision. En fonction du contexte, on pourra choisir de pénaliser plus fortement les fausses alarmes ou les détections manquées. Les mesures de performances doivent être accompagnées d'une mesure de confiance qui dépend du nombre de segments de test N et de la proportion de reconnaissance obtenue p. Une approche simple consiste à utiliser un intervalle de confiance à 95% défini suivant l'équation 3.12. Dans le cas des jeux de données classiquement utilisés (VoxCeleb [Nagrani et al., 2020]), le test atteint plusieurs milliers de locuteurs. Ainsi pour un taux d'égale erreur de 2% et 1000 locuteurs, l'intervalle de confiance est de 0.87% ce qui reste très faible.

$$\text{confiance} = 1,96 \cdot \sqrt{\frac{p(1-p)}{N}} \cdot 100 \tag{3.12}$$

3.3.2 Détection du genre (GD)

Le genre peut être modélisé de façon simple par un modèle de type GMM appris sur des coefficients cepstraux. La détection du genre peut avoir des applications intéressantes, notamment pour étudier la représentation des femmes dans les contenus audio. Plus particulièrement, le projet GEM, en partenariat avec l'INA, cherche à analyser les interactions et étudier les représentations hommme/femme dans les médias français (radio/TV).

Une première étude réalisée par [Doukhan et al., 2018a] a proposé d'évaluer 3 détecteurs de genres[6] : GMM, i-vecteurs et un réseau convolutionnel (CNN). Ce dernier est constitué de 5 couches convolutionnelles et 4 couches denses et est alimenté par des bancs de filtres. Les performances obtenues sur le corpus français REPERE [Giraudel et al., 2012] montrent de meilleurs résultats avec le modèle CNN : f-measure$=$ 96.52%, avec un biais en faveur des hommes (rappel 98.04% contre 95.05 pour les femmes). Ce biais s'explique en partie par le déséquilibre des classes dans le corpus d'apprentissage : 46h pour 1129 hommes et 12h pour 557 femmes. Ces détecteurs de genre, ont été utilisés pour faire une caractérisation massive des représentations homme/femme dans les médias français [Doukhan et al., 2018b] et concluent à une sur-représentation des hommes dans les médias. Ce phénomène s'explique en partie par le biais sur le rôle des hommes (présentateur, journaliste, experts) et des femmes (interviewée, politique, etc...).

Dans le cadre du projet GEM, les travaux de Martin Lebourdais vise à identifier le genre dans la voix à partir du signal audio, sans connaître a priori l'identité du locuteur. Pour cela, nous avons développé un détecteur de genre neuronal [Lebourdais et al., 2022a], plus simple que celui proposé par [Doukhan et al., 2018a]. Il faut souligner ici, le manque de benchmark pour évaluer les différents modèles. A l'heure actuelle, chaque modèle est évalué sur une corpus différent, ce qui a pour conséquence une comparaison difficile avec l'état de l'art.

Architecture du modèle Le modèle neuronal choisi reste extrêmement simple et se base sur une couche récurrente suivit de deux couches linéaires. Le modèle peut-être alimenté soit par une représentation temps-fréquence de type banc de filtres, soit par des représentations pré-entrainées comme WavLM [Chen et al., 2022]. Ce sont ces dernières représentations qui ont montré les résultats les plus intéressants [Lebourdais et al., 2022a].

6. Les travaux sur la détection automatique du genre considère celui-ci comme étant binaire à partir du genre déclaré par le locuteur

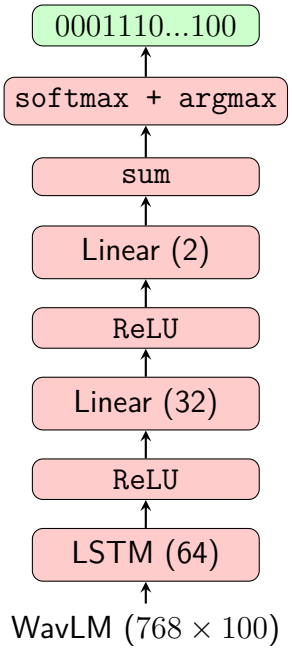


FIGURE 3.6 – Réseau récurrent pour la prédiction du genre [Lebourdais et al., 2022a]

Résultats Pour l'occasion, un sous-corpus d'ALLIES [Shamsi et al., 2022] a été extrait où chaque genre est représenté lors de l'entrainement par 30 000 segments d'1s d'audio et 40 locuteurs. À partir de ces données, le réseau décrit plus haut est testé sur 4000 segments d'1s. L'accuracy globale obtenue est 94.9%, avec un biais en faveur des hommes (97.8% d'accuracy contre 92.1% pour les femmes). Nous pouvons noter que ce biais en faveur des hommes se retrouvent dans la plupart des systèmes de traitement automatique de la parole (voir section 6.2).

3.3.3 Cas de la parole expressive

L'identification du locuteur dans un contexte de parole neutre ou journalistique est maintenant devenu très performante, avec des taux EER proche de 1% sur les corpus standards tels que VoxCeleb1&2. Cependant, dans un contexte de parole expressive, plusieurs sources de variabilités acoustiques vont venir perturber les modèles : le locuteur, l'environnement acoustique (ou la session) et l'état émotionnel du locuteur lors de l'enregistrement. Évidemment, d'autres facteurs peuvent jouer : le fait d'être enrhumé, de porter un masque, etc.

En 2010, au cours de mes travaux de thèse, j'ai étudié l'impact de la parole émotionnelle sur la reconnaissance du locuteur. Pour cela, nous avions à disposition un corpus collecté dans

un studio d'habitation auprès de personnes âgées et mal-voyantes ([Tahon, 2012], chap. 5). La principale difficulté est liée à la faible quantité de données par locuteur et au petit nombre de locuteurs. Les corpus utilisés dans la communauté pour les tâches d'identification contiennent plusieurs centaines de locuteurs, voire plusieurs milliers aujourd'hui dans VoxCeleb 1&2. Dans notre cas, le temps de parole par locuteur est très faible. Nous avons donc dû restreindre la durée des segments d'apprentissage entre 10 et 60 s par locuteur pour avoir une représentation homogène sur l'ensemble des participants.

Nous avons décidé de modéliser les locuteurs ou le genre à l'aide de GMM obtenus à partir de représentations acoustiques de type MFCCs (13 MFCC + 13ΔMFCC) et 256 Gaussiennes. Ici nous n'avons pas de modèle du monde, et donc pas d'adaptation de ce modèle aux locuteurs. Nous sommes donc en identification en milieu fermé.

Détection du genre sur des voix âgées, parole neutre IDV-HR [Tahon et al., 2011] est un corpus qui contient 22 locuteurs âgés enregistrés dans un appartement témoin, en interaction avec le robot Nao. Le protocole de collecte implique un magicien d'Oz afin d'induire des états émotionnels chez les participants. Dans un contexte d'interaction humain-robot, la machine doit être capable de reconnaître un locuteur (a minima son genre) sur un "bonjour", soit environ 1s. La tâche de détection du genre a également l'avantage d'avoir mécaniquement plus de segments par classe que celle d'identification du locuteur à quantité de donnée totale constante. Dans cette expérience, nous cherchons donc à évaluer l'impact de la durée de test sur la reconnaissance du genre. Vu le faible nombre de locuteurs, nous avons choisi de mettre les mêmes locuteurs dans les ensembles d'apprentissage et de test mais pas les mêmes instances sans phase d'enrollement donc.

Une première expérience réalisée sur de la parole neutre uniquement (60% du corpus) montre que l'apprentissage de GMMs pour modéliser le genre avec 60s par genre permet d'obtenir un taux d'erreur (pondéré par le nombre d'instances testées) d'environ 1% pour une durée de test de 20s et 7,8% pour une durée de 1s. On remarque que certains locuteurs voient leur genre systématiquement erronés. Par exemple, une des locutrices (femme de 80 ans) est toujours reconnue comme étant un homme, alors qu'un locuteur (homme de 70 ans) est souvent confondu avec une femme.

Détection du genre sur des voix âgées, parole émotionnelle Cette fois nous apprenons un modèle avec des données émotionnelles (respectivement neutres) et nous le testons sur des données neutres (resp. émotionnelles). La durée des segments d'apprentissage est de 30s

et la tâche est l'identification du genre des 22 locuteurs du corpus. La figure 3.7 montre une différence très claire entre les performances obtenues par les modèles appris sur de la parole neutre ou émotionnelle. Nous concluons que le fait d'utiliser uniquement de la parole émotionnelle perturbe énormément l'apprentissage des modèles d'identification du genre. Une étude plus approfondie sur la répartition des erreurs par émotion montre que la colère et la joie entraînent une hausse de l'erreur, particulièrement chez les hommes. En effet, la colère et la joie engendrent une restriction du conduit vocal et donc une hausse de la F_0. Or, on sait que la détection du genre se base en grande partie sur la F_0. Il est donc concevable qu'une F_0 plus élevée chez les hommes induise une prédiction du genre "femme"

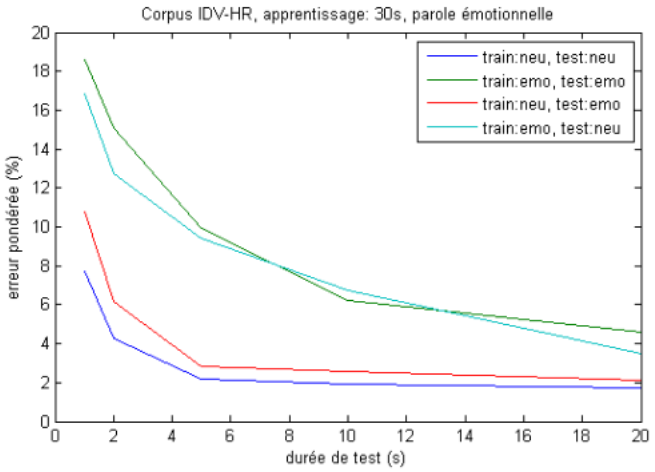


FIGURE 3.7 – Détection du genre sur de la parole émotionnelle en erreur pondérée

Identification du locuteur sur voix âgées, parole émotionnelle Nous avons proposé d'apprendre un modèle du monde sur le corpus de voix d'émotions prototypiques JEMO (62 locuteurs) et de l'adapter au corpus IDV-HR utilisé dans les expériences décrites ci-dessus. L'apprentissage a été fait avec 256 Gaussiennes sur de la parole neutre exprimée par les locuteurs de JEMO. Le modèle UBM est alors adapté (MAP) aux 22 locuteurs de IDV-HR sur 5s de parole neutre par locuteurs. Les modèles sont testés sur de la parole émotionnelle uniquement sur des segments de 1s. Le choix des 1s est motivé par le contexte applicatif. Le tableau 3.3 présente les résultats en taux d'erreur sur l'ensemble des fichiers de test. On remarque que la confusion entre locuteur la plus grande a lieu lorsque le locuteur exprime de la tristesse.

Test 1s	Confiance	Erreur
Colère	2,72	23,08
Joie	1,69	16,67
Tristesse	3,58	37,04

TABLE 3.3 – Effet des émotions sur les taux d'erreur pour l'identification du locuteur sur le corpus JEMO. Résultats issus de [Tahon, 2012].

Identification du locuteur et reconnaissance d'émotion Les x-vecteurs permettent d'obtenir des *embeddings* entrainés spécifiquement pour la tâche de reconnaissance du locuteur. On ne sait pas vraiment aujourd'hui quelles sont les autres informations capturées par cette représentation : phonétique, prosodiques, expressive, etc. [Pappagari et al., 2020] ont montré qu'un réseau de neurones de type x-vecteurs entrainé pour l'identification du locuteur, puis adapté pour la classification d'émotions permet de dépasser les performances état de l'art en reconnaissance des émotions sur 3 corpus émotionnels dont IEMOCAP [Busso et al., 2008] et MSP-Podcast [Lotfian and Busso, 2019].

Dans un contexte de parole émotionnelle, [Parthasarathy et al., 2017] a montré que les performances d'identification du locuteur se dégradaient fortement pour des valeurs extrêmes sur les dimensions valence, activation et dominance sur les données MSP-Podcast. Les auteurs ont vérifié que ce résultat n'était pas dû à un éventuel biais sur les durées des segments de parole émotionnelle. Ainsi, les conclusions obtenues en 2010 avec des modèles statistiques appris sur des quantités de données relativement petites, ont été confirmées par des expériences récentes réalisées avec des réseaux de neurones.

Enfin, dans un contexte où l'on souhaite anonymiser l'identité vocale du locuteur tout en conservant l'intelligibilité du signal de parole (projet ANR Deep Privacy), nous avons étudié l'impact d'une parole expressive sur les performances d'un système d'anonymisation développé dans le cadre du VoicePrivacyChallenge[7]. Dans le cadre du projet DeepPrivacy, Hubert Nourtel conclut que le système d'anonymisation dégrade les performances de reconnaissance des émotions, mais également la qualité de la transcription de parole [Nourtel et al., 2021]. Nous avons également montré qu'une simple modification de l'intonation via une transformation de la F_0 ne suffit pas à cacher les émotions du point de vue du système de reconnaissance automatique.

7. `https://www.voiceprivacychallenge.org/`

3.4 Segmentation et regroupement en locuteurs

La tâche de segmentation et regroupement en locuteurs (SRL) consiste à déterminer qui parle quand. Suite à cette tâche, le signal audio contient des segments mono-locuteurs sur lesquels on peut ensuite réaliser des traitements plus haut-niveau comme la caractérisation expressive ou l'étude des interactions entre les différents participants d'une conversation. Pour une revue exhaustive des systèmes et approches, le lecteur pourra se référer à [Moattar and Homayounpour, 2012].

Une première segmentation du signal de parole est obtenue en deux phases :

1. segmentation initiale : détection des changements de locuteurs,
2. regroupement : clustering des segments appartenant aux mêmes locuteurs.

Éventuellement, à partir du regroupement, une nouvelle segmentation peut être réalisée pour raffiner à nouveau le regroupement.

3.4.1 Approches pour la segmentation et le regroupement en locuteur

Approche probabiliste

Dans l'approche probabiliste, les distributions des représentations acoustiques du signal audio sont modélisées par des Gaussiennes. Cette approche prend appui sur la modélisation des locuteurs avec des GMMs décrite précédemment.

Segmentation La première phase de segmentation consiste à détecter une différence de locuteur entre deux trames consécutives. La segmentation initiale est réalisée de sorte à ce que chaque segment soit de durée suffisante pour identifier la voix du locuteur et contienne un unique locuteur. On préfère donc minimiser le nombre de changements manqués (*miss*) et maximiser la pureté, sachant qu'un changement de locuteur qui n'en est pas un pourra toujours être supprimé par la suite. La méthodologie décrite ci-dessous est adaptée de [Barras et al., 2006]. Elle a été implémentée dans le toolkit du LIUM S4D [Broux et al., 2018]. Cette segmentation est réalisée après avoir supprimé les segments de bruit, musique et parole superposée.

Le signal est segmenté en fenêtres temporelles glissantes s_t d'une durée fixe (quelques secondes). Chaque fenêtre est modélisée par une unique Gaussienne diagonale à partir d'une représentation cepstrale (typiquement 12 MFCCs + l'énergie). On définit une mesure de di-

vergence entre deux Gaussiennes représentant deux fenêtres adjacentes s_1, s_2 par $G(s_1, s_2)$, par exemple la distance de Mahalanobis définie par l'équation 3.13.

$$G(s_1, s_2) = (\mu_2 - \mu_1)^T \Sigma_1^{-1/2} \Sigma_2^{-1/2} (\mu_2 - \mu_1) \tag{3.13}$$

On optimise un seuil de détection G_s qui permet d'avoir des segments purs, d'une durée minimum (typiquement 2,5s). Grâce à l'hypothèse diagonale sur la matrice de covariance, le calcul de cette distance est simple et rapide. La distance de Mahalanobis peut être remplacée par une distance de Kullback-Leibler. Une autre approche compare deux modélisations : soit il vaut mieux modéliser la fenêtre $s = s_1 \oplus s_2$ avec une unique Gaussienne, soit il vaut mieux la modéliser par deux Gaussiennes représentant respectivement s_1 et s_2. On utilisera alors le rapport de vraisemblance global (GLR - *generalized lilelihood ratio*) défini par l'équation 3.14.

$$GLR(s_1, s_2) = -\log \frac{L(s|\mu, \Sigma)}{L(s_1|\mu_1, \Sigma_1) L(s_2|\mu_2, \Sigma_2)} \tag{3.14}$$

Cette segmentation initiale est difficile à évaluer car les frontières de référence sont généralement peu précises. On évalue plutôt l'application (identification des locuteurs) utilisant cette segmentation, en comparant les résultats obtenus à partir de la segmentation de référence. Des méthodes neuronales récentes basées sur les réseaux séquence vers séquence décrits précédemment (voir section 3.2.1) prédisent les changements de locuteur de façon plus précise [Yin et al., 2017] puisque les classes sont estimées à la trame.

Regroupement Une seconde segmentation permet de fusionner les segments d'un même locuteur (*agglomerative clustering*), *i.e.* supprimer les changements de locuteurs qui n'en sont pas. Pour cela, on utilisera une classification hiérarchique ascendante (HAC - *Hierarchical Agglomerative Clustering*) qui est un algorithmique itératif non supervisé permettant de regrouper les deux classes les plus proches. Le critère d'arrêt est BIC (*Bayesian Information Criterion*) détaillé ci-après.

Initialement chaque segment représente une classe c_i et est modélisé par une unique Gaussienne avec une matrice de covariance pleine. A chaque itération, les deux classes les plus proches sont fusionnées ou non suivant le critère BIC donné par l'équation 3.15 où n_i, n_j sont respectivement, le nombre de trames acoustiques des classes c_i, c_j, D la dimension de la représentation acoustique et $n = n_i + n_j$. Ce critère inclut une pénalité P pondérée par le coefficient λ. Il y a fusion des classes si $\Delta BIC < 0$, c'est-à-dire que les distributions sont suffisamment proches. En pratique, à chaque itération on ne fusionne que les classes les plus

proches, c'est-à-dire celles qui ont le ΔBIC minimum. Les itérations s'arrêtent lorsque le critère d'arrêt est atteint, c'est-à-dire que $\Delta BIC > 0$.

$$\Delta BIC = (n_i + n_j) \log |\Sigma| - n_i \log |\Sigma_i| - n_j \log |\Sigma_j| - \lambda P \tag{3.15}$$

$$= GLR(c_i, c_j) - \lambda P \tag{3.16}$$

$$P = \begin{cases} \frac{1}{2}\left(D + \frac{1}{2}D(D+1)\right) \log n & \text{si la matrice est pleine} \\ 2D & \text{si la matrice est diagonale} \end{cases} \tag{3.17}$$

Au début de l'algorithme d'agglomération itérative, les segments de parole sont courts, c'est pourquoi le choix de modèles simples avec peu de paramètres sont préférables. Au fur et à mesure, les segments sont plus longs, il est donc souhaitable de complexifier les modèles, notamment en utilisant des modèles d'identification du locuteur. C'est pourquoi il est nécessaire de faire une resegmentation à l'aide de modèles plus larges utilisant des représentations en locuteurs du type i or x-vecteurs *embeddings* [Larcher et al., 2021]. Une fois l'algorithme terminé, l'application d'un système d'identification du locuteur permet d'améliorer encore le regroupement en locuteur tout en donnant un nom à chaque classe de locuteur connu.

Evaluation La tâche de SRL s'évalue classiquement en DER (*Diarization Error Rate*, eq. 3.18) conformément aux recommandations du NIST [8]. Afin d'ajouter une tolérance aux frontières, le DER inclut un collar. Cela permet de prendre en compte la difficulté de la tâche de segmentation manuelle lors de l'annotation des corpus d'apprentissage. Le segment de référence est alors artificiellement augmenté de la valeur du collar au début et à la fin du segment. Le collar vaut typiquement 250 ms. Certaines campagnes d'évaluation évaluent sans collar, c'est le cas de la campagne DiHard.

$$DER = \frac{Conf + Miss + Fa}{T} \tag{3.18}$$

Dans le cas d'une tâche de SRL *cross-show* on cherche à regrouper les segments de parole d'un même locuteur au sein du même fichier audio, mais également entre les différents fichiers. Le choix de l'identifiant du locuteur est soumis à un problème de permutation. En effet, si sur un premier fichier audio, nous entendons les locuteurs 1 et 2, dont les segments de parole sont regroupés, et sur un second fichier nous identifions trois locuteurs différents 3, 4 et 5,

8. National Institute of Standards and Technology

comment savoir que les identifiants 3 et 5 correspondent au même locuteur ? Cette tâche nécessite de tester toutes les permutations possibles afin de s'assurer que le nouvel identifiant ne corresponde pas à un locuteur déjà identifié. On peut alors utiliser l'algorithme hongrois pour limiter le nombre de paires testées et permet donc de gagner en efficacité.

L'approche probabiliste reste encore une des plus utilisées. Récemment, d'autres approches fondées sur les réseaux de neurones ont été développées, comme par exemple EEND qui a pour objectif d'apprendre une représentation locale des locuteurs (appelée *attractors*) lors de la tâche de segmentation et regroupement [Fujita et al., 2019]. L'approche proposée intègre le problème de permutation directement dans la fonction de coût (*permutation loss*). Mano Brabant en stage dans l'équipe en 2022, a implémenté l'algorithme hongrois comme fonction de coût afin de remplacer la *permutation loss*.

Approche spectrale

Le clustering spectral permet de modéliser les représentations acoustiques et de fusionner des fenêtres temporelles similaires sans utiliser de représentation probabiliste. Cela permet de sortir de l'hypothèse de normalité et d'indépendance des descripteurs acoustiques. De plus, ce type d'approche a l'avantage de pouvoir combiner des représentations acoustiques locales et globales. L'approche décrite ci-dessous a été développée pour de la segmentation musicale [Tahon et al., 2019] mais est également utilisée pour le SRL [Park et al., 2020].

Dans le cas musical, la détection de répétitions sur un temps court (par exemple répétition d'accords) est une tâche relativement simple. Cependant, la combinaison de ces multiples répétitions sur une échelle temporelle plus large est beaucoup plus complexe. L'article [McFee and Ellis, 2014] propose une combinaison pondérée des éléments consistants au niveau local avec une représentation des répétitions à une échelle plus large.

Construction de la matrice de similarité Soit une représentation temporelle d'un signal audio $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T] \in \mathcal{R}^{D \times T}$. On extrait une matrice de récurrence binaire $R \in \{0,1\}^{T \times T}$ définie par l'eq. 3.19 où $k > 0$ fixe le degré de connexion entre deux trames. Cette matrice capture les récurrences locales de $\mathbf{x}$.

$$R_{ij} = \begin{cases} 1 \text{ si } x_i, x_j \text{ sont des k plus proches voisins} \\ 0 \text{ sinon} \end{cases} \tag{3.19}$$

On définit ensuite une matrice de similarité S entre deux segments qui peuvent être éloignés

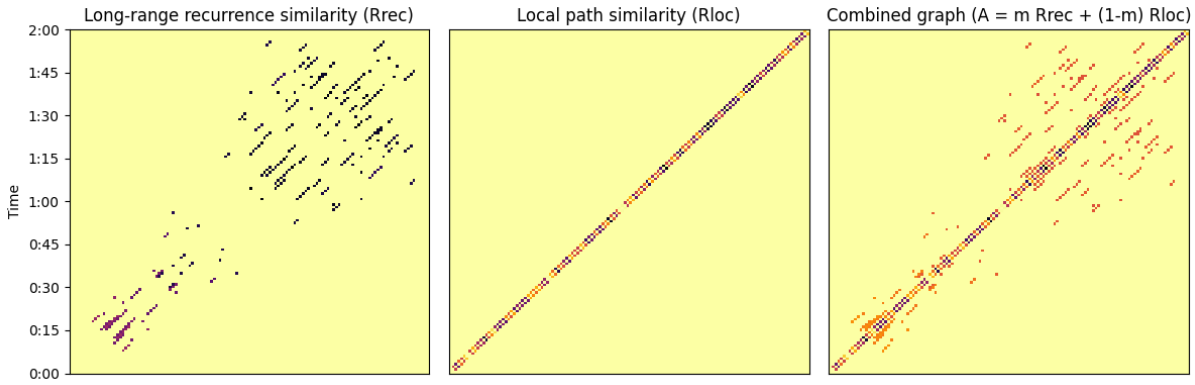


FIGURE 3.8 – Représentation des matrices de similarité à long terme R et à court terme S, ainsi que le graphe combiné A pour un extrait musical de Free Jazz.

dans le temps (eq. 3.20). Ces segments y_i peuvent avoir une durée plus longue qu'une trame.

$$S_{ij} = \exp\left(-\frac{1}{2\sigma^2|y_i - y_{i+1}|^2}\right) \tag{3.20}$$

On combine alors les deux matrices de similarité suivant l'équation 3.21 : l'une caractérise les récurrences locales S, l'autre les récurrences globales R. Le paramètre $0 < \mu < 1$ est une pondération. Pour plus de détails sur cette phase de construction se référer à [McFee and Ellis, 2014] et une implémentation python pour l'analyse musicale [9]. Une fois cette matrice A obtenue, on peut appliquer le principe du clustering spectral. Une représentation des matrices de similarité est illustrée figure 3.8.

$$A = \mu S + (1 - \mu)R \tag{3.21}$$

Application du clustering spectral Le clustering spectral exploite les valeurs et vecteurs propres de la matrice de similarité A. On suppose que A ne possède pas de valeurs négatives, alors l'algorithme est le suivant :

1. calculer le degré $d_i = \sum_j A_{ij}$ et la matrice diagonale $D = \text{diag}\{d_1, d_2, \ldots, d_T\}$
2. construire le Laplacien normalisé $L = D^{-1/2}AD^{-1/2}$
3. déterminer les vecteurs propres $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K$ correspondant aux K plus grandes valeurs

9. `https://git-lium.univ-lemans.fr/tahon/spectral-clustering-music`

propres, où K est le nombre final de clusters. On définit la matrice $V = [\mathbf{v}_1|\mathbf{v}_2|\dots|\mathbf{v}_K] \in \mathcal{R}^{T\times K}$

4. normaliser les lignes de V et construire W telle que $w_{ij} = x_{ij}/\sqrt{\sum_j x_i j^2}$
5. regrouper les T lignes de Y et associer le point i au cluster j si et seulement si la $i$ème ligne de Y est associée au cluster j.

L'étape 5 se fait, par exemple, avec un algorithme de K-means clustering. Les frontières des différents segments est l'ensemble $\{i|c_i \neq c_{i+1}\}$ où i est l'indice de la trame courante et c_i le cluster auquel appartient la trame i.

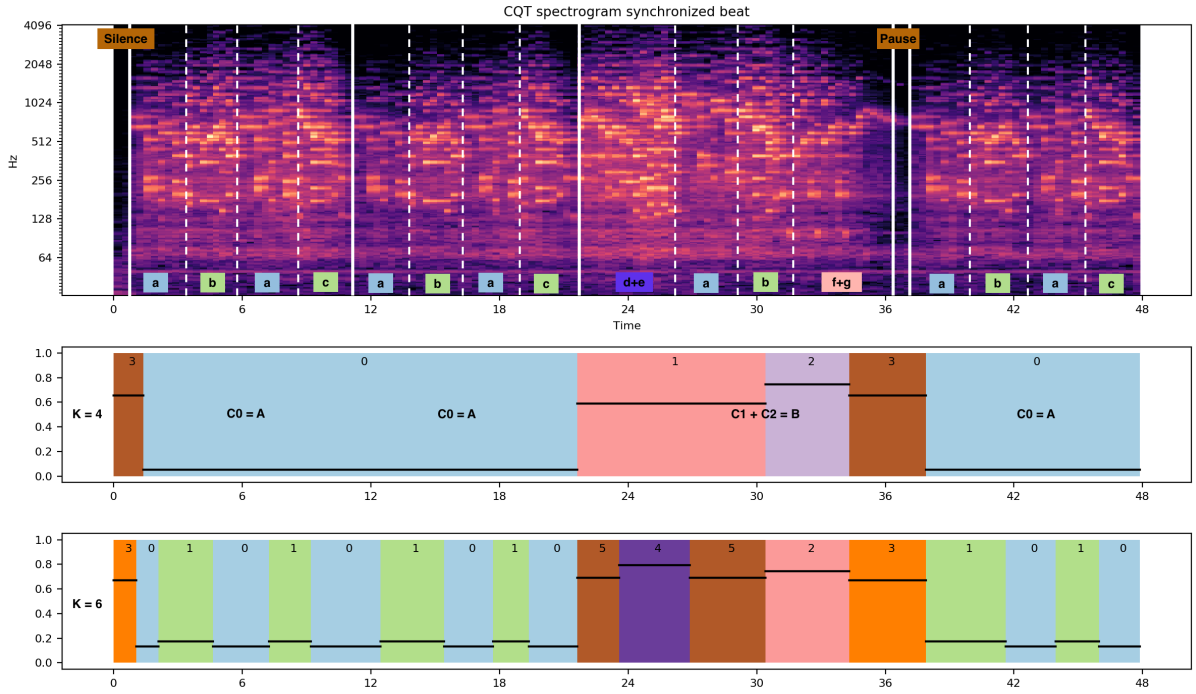


FIGURE 3.9 – De haut en bas : représentation du spectrogramme à Q constant, la segmentation obtenue avec $K = 4$ puis $K = 6$ pour un extrait de la sonate K545 pour piano de Mozart. La segmentation en motifs réalisée par un musicologue, est donnée en blanc sur la représentation temps-fréquence.

Faire varier K permet de contrôler directement la granularité de la segmentation obtenue. Dans le cas du SRL K sera le nombre de locuteurs présents dans le signal, dans le cas musical, K sera le nombre de motifs musicaux, comme illustré sur les figures 3.9 et 3.10. Dans l'illustration pour la musique, les différents clusters ont permis d'identifier différents motifs musicaux qui se répètent dans le morceau. Dans l'extrait de Mozart, ces motifs sont

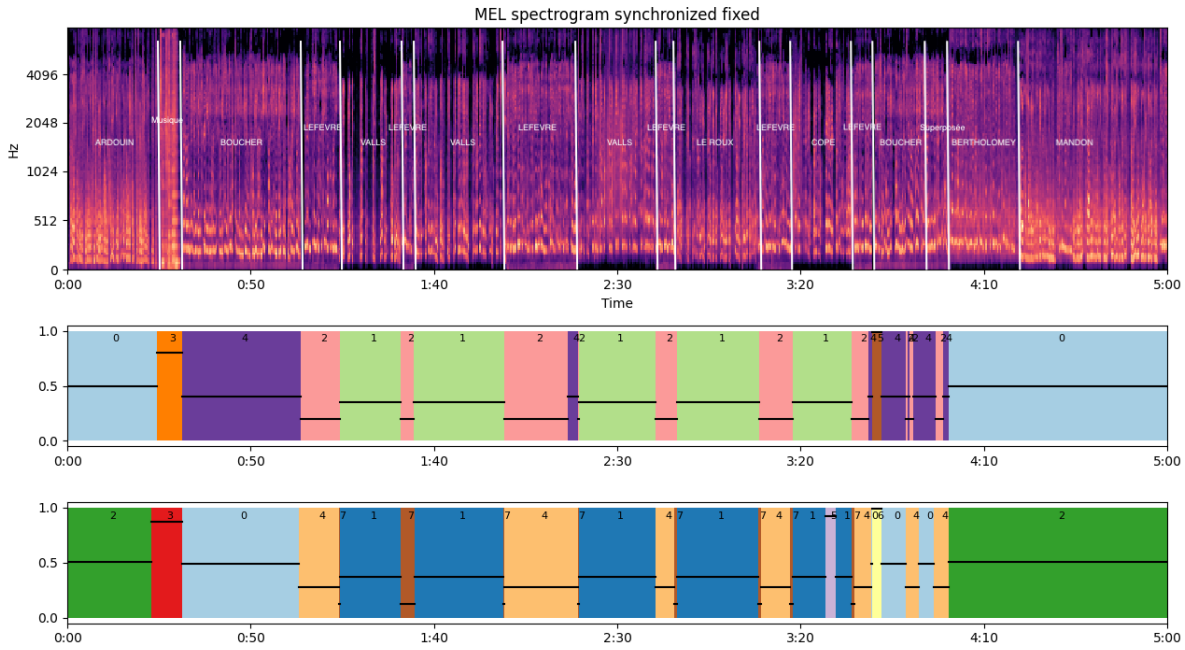


FIGURE 3.10 – De haut en bas : représentation du mel-spectrogramme, la segmentation obtenue avec $K = 6$ puis $K = 8$ pour un extrait d'une émission TV.

avant tout rythmiques, mais l'approche proposée permet d'identifier des motifs mélodiques également suivant les descripteurs audios fournis en entrée. Pour une segmentation à $K = 6$ clusters, nous retrouvons la structure AABA où le groupe 3 (orange) correspond à du silence. La partie A est identifiée par le groupe 0 (bleu, question) et le groupe 1 (vert, réponse). Les groupes 2,4,5 font partie de la partie B qui est différente à la fois au niveau de la tonalité et du rythme.

L'application de cette analyse automatique a permis d'initier l'analyse d'un morceau de Free Jazz (Globe Unity [Tahon, 2020]). Le Free Jazz est un style de musique improvisée cherchant à déconstruire une éventuelle structure. Théoriquement, le morceau n'est donc pas structuré, or l'utilisation de l'approche automatique a permis d'en mettre une en évidence. Dans le cas d'une analyse comme celles présentées dans le cas de la musique, le fait de pouvoir choisir le nombre de clusters K afin de faire varier la granularité des segments est extrêmement précieuse.

Nous avons également testé l'approche sur une tâche de SRL illustrée figure 3.10. Nous pouvons observer que la segmentation est tout à fait correcte dès lors que les différences spectrales sont fortes (parole ou musique, ou encore différences de canaux acoustiques). Par contre, cette approche seule aura du mal à regrouper correctement les locuteurs. Par exemple,

sur la segmentation à $K = 8$, les groupes 4 et 7 correspondent aux mêmes locuteurs, ce qui a bien été identifié lorsque $K = 6$. Aux alentours de 4 minutes, deux locutrices parlent en même temps (Boucher et Bertholomey), ce qui provoque une confusion dans l'alternance des groupes. Nous pouvons donc conclure que cette approche apporte des résultats sans doute moins bons en terme de DER que l'approche présentée précédemment. Par contre, elle semble très utile pour la réalisation d'une première segmentation, et éventuellement identifier des zones de musique, de silence, et éventuellement de parole superposée. Dans le cas du SRL, la matrice d'affinité peut être construite à l'aide d'une divergence de Kullback-Leibler (KL), distance de Mahanalobis (G) ou le rapport de vraisemblance généralisé (GLR) telles que définis précédemment dans l'approche probabiliste [Ning et al., 2006].

3.4.2 Choisir le nombre de classes

Déterminer le nombre correct de clusters K est un problème complexe qui est loin d'être résolu. Les méthodes de clustering proposent plusieurs approches que nous décrivons ci-dessous. L'utilisation de l'approche spectrale incite à utiliser les valeurs propres pour déterminer le nombre de clusters, mais ce n'est pas la seule technique possible.

L'indice de Silhouette [Rousseeuw, 1987] cherche à minimiser la distance moyenne entre l'échantillon et tous les autres points appartenant au même cluster a, en maximisant la distance moyenne entre l'échantillon et tous les autres points du groupe le plus proche b : $s = \frac{b-a}{\max(a,b)}$ s est borné par -1, une valeur de 0 indique que les clusters se recoupent.

L'indice de Calinski-Harabaz [Caliński and Harabasz, 1974] est défini par le rapport entre la dispersion moyenne intra-cluster et la dispersion inter-cluster. Un score est élevé si les clusters sont denses et correctement séparés. Cette approche a l'avantage d'être rapide à calculer.

L'indice de Davies-Bouldin [Davies and Bouldin, 1979] DB correspond à la similarité moyenne entre chaque cluster c_i et son plus proche cluster c_j. La similarité $r_{ij} = \frac{s_i + s_j}{d_{ij}}$ est obtenue à l'aide de s_i, la distance moyenne entre chaque point du cluster c_i et son barycentre, et de d_{ij} la distance entre les barycentres des clusters c_i et c_j : $DB = \frac{1}{K} \sum_i^K \max r_{ij}$.

Bien que l'ensemble de ces méthodes soient classiquement utilisées pour optimiser le nombre de clusters K, la validité de ces mesures a toujours été réalisée sur des jeux de

données simulées. Dans un cas de données réelles, ces indices n'apportent souvent pas de réponse satisfaisante [Lamirel et al., 2016].

La méthode des valeurs propres détermine K à partir d'une chute importante des amplitudes des valeurs propres (rangées par ordre décroissant) de la matrice du Laplacien. Cette méthode a été implémentée dans le système de SRL VBx fondé sur un modèle de Markov Bayésien [Diez et al., 2020] pour estimer le nombre de locuteurs.

Dans les tâches de SRL, le nombre de locuteurs présents sur un segment de parole est typiquement compris entre 2 et 7. Les indices qui ne sont pas basés sur les valeurs propres ont tendance à surestimer fortement le nombre de clusters. La méthode des valeurs propres permet de donner une première estimation qui doit être consolidée par la suite avec des systèmes d'identification du locuteur.

SRL en multi-canal L'utilisation d'antennes de microphones permet d'ajouter des informations spatiales en plus des représentations spectrales. L'information de la direction du locuteur qui parle permet d'améliorer les systèmes de SRL [Zheng et al., 2022]. Dans le cadre d'une collaboration entre deux doctorants du laboratoire, nous avons étudié la possibilité d'inclure l'information spatiale dans les systèmes de segmentation. Avec les systèmes actuels, l'utilisation de l'information spatiale semble apporter de l'information pour la détection de parole superposée, mais le manque de données collectées et annotées au niveau spatial ne permet pas de généraliser le résultat.

3.4.3 SRL et parole expressive

Selon [Shriberg, 2007], l'utilisation de caractéristiques à court terme, telle que la représentation cepstrale à la trame, ne permet pas de capturer certaines caractéristiques du locuteur du type prosodique, ou bien son style expressif. L'ajout de descripteurs segmentaux (*long term features*) comme des paramètres de qualité vocale [Zewoudie et al., 2018] ou des paramètres prosodiques [Friedland et al., 2009] a été montré comme bénéfique pour la tâche de SRL.

Dans la section précédente sur l'identification du locuteur, nous avons discuté le fait que l'état émotionnel du locuteur ajoutait de la confusion pour les systèmes d'identification. Il est possible que l'émotion exprimée par le locuteur fausse également la détermination du nombre de locuteurs K lors de l'étape de clustering. En effet, un locuteur triste pourrait être identifié différemment que ce même locuteur joyeux. Également, plus la parole est expressive, plus la structure des tours de parole au cours de la conversation peut devenir complexe,

avec une grande quantité de parole superposée et des voix déformées par rapport à leur style "neutre". Ces aspects ajoutent un défi supplémentaire pour les tâches de segmentation et de caractérisation dans un contexte expressif.

3.5 Conclusion et discussions

Segmentation audio Ce chapitre présente les aspects bas-niveau du traitement automatique de la parole expressive. Les systèmes qui y sont présentés permettent de segmenter le signal de parole et d'étiqueter certaines zones afin de pouvoir appliquer des traitements plus haut-niveau ensuite. Les étiquettes fournies à ce stade sont la présence de parole, de bruit, l'identité du locuteur, son genre et la présence de parole superposée. J'ai également résumé la tâche de SRL. Les systèmes de segmentation peuvent être également vus comme des briques de base pour l'annotation des données. En effet, à partir des étiquettes retournées, on peut sélectionner les données pertinentes afin de les enrichir d'une annotation manuelle.

Perspectives Ces tâches sont très rarement évaluées dans un contexte spécifique de parole expressive. Or il semble important de porter nos efforts vers ses données, car leur traitement entraîne généralement une forte dégradation des systèmes. De plus, les segmentations de référence ont longtemps été considérées comme "vraies" et objectives. En effet, on peut facilement imaginer qu'une tâche qui consiste à placer des frontières autour d'un segment de parole homogène suivant une caractéristique donnée, est objective. Cependant, la perception humaine est toujours subjective avec un degré plus ou moins important suivant la tâche. L'accord inter-annotateur sur le positionnement des frontières de segment n'a jamais été évalué à ma connaissance. De même, lorsqu'on identifie des locuteurs, l'oreille humaine peut être biaisée par le contenu linguistique, peut ne pas percevoir de différence, peut se tromper sur l'identité de la personne. Ou encore la présence de parole superposée reste à l'appréciation de l'annotateur qui peut considérer que le second locuteur parle moins fort que le premier. Ainsi, même si les systèmes de segmentation sont soumis à des protocoles établis depuis longtemps, notamment par le NIST et le LNE[10], il semble nécessaire d'évaluer également l'erreur induite par les aspects subjectifs de la chaine de traitement. Une perspective intéressante serait de pouvoir affecter plusieurs étiquettes de haut-niveau sur une même trame (par exemple musique et parole), dans l'idée que cela permette ensuite une interprétation plus fine du phénomène conversationnel.

10. Laboratoire National d'Essais

Chapitre 4

Caractérisation automatique de l'expression vocale

Les travaux présentés dans ce chapitre ont été réalisé pour partie pendant ma thèse (reconnaissance de catégories émotionnelles) et au LIUM. Les recherches menées sur la reconnaissance de satisfaction dans des données de centres d'appels ont été réalisées dans le cadre de la thèse de Manon Macary (CIFRE avec AlloMedia). D'autres travaux sur la reconnaissance de caractéristiques expressives ont été réalisés : sur l'hésitation en collaboration avec le Laboratoire de Phonétique et de Phonologie (LPP) dans le cadre du stage de M2 d'Appolline Marin, sur les interruptions dans les médias dans le cadre de la thèse de Martin Lebourdais (ANR GEM) et Rémi Uro (CIFRE avec l'INA), et sur les styles discoursifs dans la parole lue en collaboration avec l'IRISA (ANR SynPaFlex).

Dans l'ensemble des recherches présentées, on suppose qu'il existe une référence explicite de ce qu'on cherche. Dans tous les cas, on définira donc une ontologie qui permet ensuite de mettre en place un schéma d'annotation et une évaluation adéquate. L'apprentissage actif est abordé dans la dernière section comme une technique pour l'annotation semi-supervisée des corpus de parole expressive.

4.1 Annotation des données expressives

4.1.1 Considérations générales

Dans un premier temps, je souhaite préciser certaines notions qui sont spécifiques à la parole expressive. Ces notions sont inspirées la plupart du temps de théories psychologiques, qui ont été adaptées au fil du temps et des expériences aux contraintes de l'informatique.

Émotion perçue/ressentie Suivant le scénario de collecte des données, l'annotation sera réalisée par des personnes différentes :

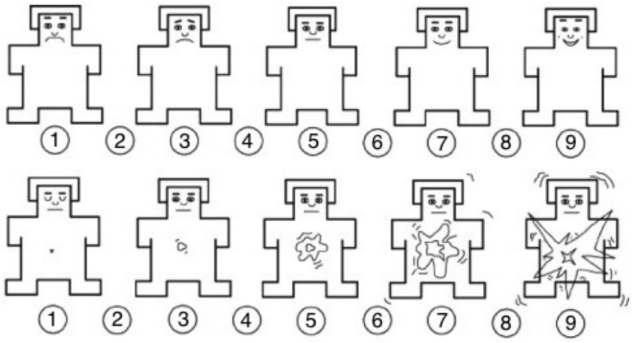


FIGURE 4.1 – Pictogrammes pour l'annotation de ses propres émotions. Figure extraite de [Bradley and Lang, 1994].

1. La personne qui a été enregistrée identifiera elle-même l'émotion qu'elle pense avoir exprimée. Par exemple, la personne choisi un pictogramme (par exemple ceux représentés figure 4.1) représentant selon elle, l'émotion qu'il a ressentie. Ce protocole a été utilisé pour annoter le corpus RECOLA [Ringeval et al., 2013].
2. Plusieurs annotateurs extérieurs à la conversation écoutent le segment de parole, et indiquent l'émotion qu'ils ont perçu en utilisant une étiquette parmi plusieurs.
3. Dans le cas où les émotions sont actées, l'étiquette est définie à l'avance et oriente l'acteur dans le choix de ses expressions. En ce cas, l'annotation par une personne tiers n'est pas nécessaire.

Émotion discrète/continue Dans un premier temps, le domaine de l'*affective computing* s'est focalisé principalement sur une modélisation en classes d'émotions telles que définies dans la section 1.1.2. La tendance actuelle est plutôt à l'utilisation de dimensions affectives, principalement valence et activation. Les catégories émotionnelles sont annotées généralement au niveau d'un segment de parole qui aura été segmenté préalablement afin qu'il ne contienne l'expression d'une unique émotion. Ces segments de parole de durées variables sont souvent associés à des groupes de souffles [Cowie et al., 2001], [Devillers et al., 2010], [Clavel et al., 2008], [Schuller, 2018].

Les dimensions affectives sont annotées soit au niveau d'un segment de parole préalablement segmenté [Tahon et al., 2011], soit en continu en fonction du temps [Ringeval et al., 2013], [Kossaifi et al., 2021], [Macary et al., 2020b]. Dans le second cas, l'annotateur utilise un ou-

til, par exemple FeelTrace [Cowie et al., 2000], qui permet de faire varier une des dimensions affectives avec un joystick ou un curseur tout en écoutant l'extrait sonore.

Évaluation des modèles Introduite dans les différents challenges du domaine (Interspeech [Schuller et al., 2009], ComPare [Schuller et al., 2013], AVEC [Ringeval et al., 2019], etc.), la modélisation des émotions s'est standardisée et la tâche se résume souvent à prédire une étiquette par segment de parole (tâche de classification), ou bien une courbe continue du temps (tâche de régression).

Le challenge Interspeech 2009 a standardisé l'évaluation de la classification en catégories émotionnelles avec l'utilisation des rappels moyens par classe pondérés (WAR qui correspond à l'*accuracy*) ou non pondérés (UAR). Étant donné que les classes émotionnelles sont généralement peu équilibrées avec une sur-représentation de la classe neutre, notamment pour les émotions *real-life*, on préfère utiliser le UAR. En effet, soit un système binaire où 70% des segments sont neutres, les autres correspondent à une émotion, si le modèle prédit uniquement la classe neutre, le WAR surestime le modèle (70%) alors que le UAR reflète plus la difficulté de la tâche (33%).

Le challenge AVEC 2019 a standardisé l'évaluation des prédictions continues avec l'utilisation du *Concordance Correlation Coefficient* (CCC) défini par l'équation 4.1 où x est la prédiction et y la référence. μ_x et μ_y sont les moyennes des deux variables sur la durée du segment et σ_x and σ_y leur écart-type. ρ est le coefficient de corrélation entre les deux variables x et y.

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{4.1}$$

Ce coefficient est égal à 1 lorsque les deux courbes sont confondues et tend vers 0 lorsqu'elles s'éloignent. Il a l'avantage de prendre en compte à la fois, les moyennes et les dispersions des données. Par contre, dans cette métrique, le temps n'a aucune importance car chacun des échantillons temporels (x_t, y_t) sont considérés comme étant indépendants les uns des autres.

Aspect subjectif La perception et le ressenti d'une émotion est par nature, subjectif. La tâche de modélisation nécessite donc de prendre en compte la diversité des perceptions humaines afin de construire un modèle pertinent.

Afin d'obtenir une étiquette unique à partir de plusieurs annotations subjectives, le vote majoritaire est la solution la plus utilisée. Éventuellement, il est possible de pondérer les annotations individuelles suivant la confiance dans l'annotateur (son niveau d'expertise, son

sérieux, etc.). Dans le cas où il y a un très grand nombre d'annotateurs d'origine non contrôlée (*crowdsourcing*), cette pondération peut être intéressante.

Pour les dimensions affectives, la courbe de référence est souvent une moyenne (éventuellement pondérée) des valeurs obtenues par chacun des annotateurs. Afin de prendre en compte les éventuels décalages temporels des annotateurs, il est possible de modéliser les temps de réaction individuels après l'écoute d'une émotion [Mariooryad and Busso, 2015].

La prise en compte de cette subjectivité dans les modèles est très étudiée aujourd'hui [Han et al., 2017], particulièrement dans le cas des dimensions affectives. En effet, celles-ci étant représentées par des valeurs numériques, il est plus simple d'effectuer des opérations statistiques dessus. Par exemple, il est possible de modéliser l'incertitude sur la valeur d'une dimension par une loi Gaussienne [Lotfian and Busso, 2017], un échantillonnage de Monte-Carlo [Sridhar et al., 2021], ou encore de réaliser un modèle par annotateur [Macary et al., 2022].

Vu que les données annotées représentent une très faible proportion de ce qui peut être exprimé dans les conditions d'enregistrement, il est intéressant d'évaluer la robustesse des modèles lorsqu'ils sont évalués sur des données différentes. De plus, les résultats obtenus sont généralement très spécifiques aux locuteurs participants, au scénario de collecte et à l'environnement acoustique. L'ajout d'intervalles de confiance pour encadrer le résultat final a l'avantage d'estimer l'effet de l'échantillonnage réalisé sur l'ensemble de test que ce soit pour une mesure d'UAR [Tahon et al., 2016b] ou de CCC [Macary et al., 2022]. En présence d'un corpus de test peu représentatif, un intervalle de confiance est absolument nécessaire pour estimer si un modèle obtient de meilleurs résultats ou non qu'un autre.

4.1.2 Annotations en émotion

Catégories émotionnelles

L'annotation en catégories émotionnelles a été longtemps l'approche privilégiée pour la construction de corpus d'apprentissage de modèle d'émotion dans la parole. Au cours de ma thèse, j'ai eu l'occasion de collecter plusieurs corpus *real-life* IDV-HH [Tahon et al., 2010], NAO-HR [Delaborde et al., 2009] [Tahon et al., 2012b], IDV-HR [Tahon et al., 2011], COMPARSE [Giraud et al., 2013] et acté JEMO [Brendel et al., 2010]. Je ne détaille pas ici les différents protocoles utilisés. Le lecteur intéressé pourra aller consulter les articles correspondant aux corpus mentionnés ci-dessus. Le fait que les situations conversationnelles soient très différentes entre les corpus, engendre une difficulté à mettre en place des expériences cross-corpus. En effet, la colère annotée dans un corpus sera différente d'un point de vue des

réalisations acoustiques et du contenu sémantique, dans une conversation téléphonique, dans une discussion entre amis, ou dans une interaction humain-robot où le robot joue un rôle de tricheur.

Il me semble que le choix en catégories a l'avantage d'être très explicite car porté par des mots dont le sens global est partagé par tous. Mais il a l'inconvénient majeur d'englober des contenus différents suivant les situations et les expériences de chacun, ce qui limite fortement le pouvoir de généralisation des modèles. De plus les étiquettes sont souvent affectées à des segments de durée très variables et ne prennent donc pas en compte l'aspect temporel. Or le fait de reconnaître une émotion ou une autre est fondé sur certains éléments très précis (une emphase, une inflection de voix, un timbre particulier) qui ont lieu autour d'un point d'ancrage temporel [Lin et al., 2014].

Dimensions affectives

Contrairement aux étiquettes discrètes, l'annotation continue de dimensions affectives est plus implicite. Notamment, les notions d'activation et de contrôle restent mal définie. Cependant, ces dimensions semblent mieux généraliser à des situations émotionnelles différentes à l'échelle près. De plus, elles ont l'avantage de prendre en compte l'aspect temporel de l'évolution de l'état affectif du locuteur.

Dans cette section, je présente le cas du corpus AlloSat qui a été collecté pendant la thèse de Manon Macary en partenariat avec l'entreprise Allo-Media [Macary et al., 2020b]. Il s'agit de conversations provenant de centres d'appels téléphoniques pour différents domaines (énergie, voyage, assurance, etc.). Nous avons choisi d'étudier l'évolution temporelle de la dimension affective satisfaction/frustration au cours de la conversation. Cette dimension affective a été motivée par un intérêt économique pour l'entreprise qui cherche à vendre des analyses haut-niveau des conversations en plus des transcriptions linguistiques.

Le coût de l'annotation continue étant très élevé, nous avons d'abord sélectionné des conversations où il est probable qu'il se passe quelque chose. Les conversations ont donc été sélectionnées sur des critères de variations prosodiques (F_0 et énergie) et sur la présence de mots-tokens ayant une polarité émotionnelle. À ces conversations, nous avons ajouté des enregistrements choisis aléatoirement, pour lesquels il est probable que le contenu émotionnel soit faible. 303 conversations d'une durée moyenne de 7 min 24 s ont été annotées par trois annotateurs experts qui ont été préalablement formés avec des conversations témoins. Un exemple de ces trois annotations est illustré sur la figure 4.2.

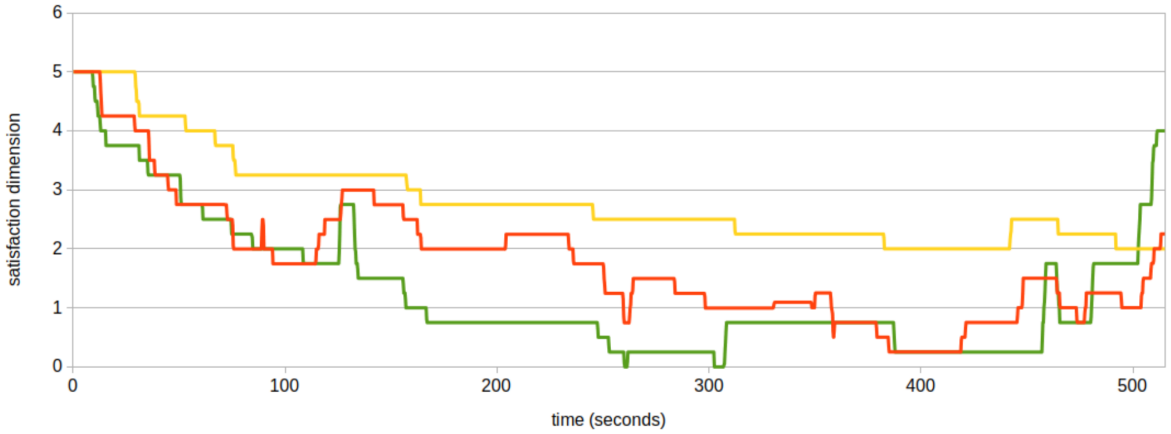


FIGURE 4.2 – Visualisation des trois annotations continues sur une conversation du corpus AlloSat. Figure extraite de [Macary et al., 2020b].

4.1.3 Annotations expressives

Comme nous l'avons déjà abordé au chapitre 1, la caractérisation des émotions est le domaine de recherche principal pour ce qui est de la parole expressive. Cependant, d'autres aspects expressifs ont également été étudiés et ont donc nécessité une annotation. Cette section détaille trois types d'annotations (hésitation, interruption, discours) réalisées sur de la parole expressive. Pour chaque type, l'objectif est de construire une modélisation à partir des annotations afin de caractériser automatiquement l'expressivité en fonction du contenu acoustique et phonetico-linguistique.

Annotation de l'hésitation

Ce projet s'inscrit dans une collaboration avec le LPP autour de l'hésitation dans la parole spontanée et a fait l'objet du travail de stage de M2 d'Appolline Marin. À partir du corpus NCCFr de conversations dyadiques spontanées [Torreira et al., 2010], nous avons proposé un protocole pour annoter la parole hésitante [Wottawa et al., 2020]. Alors que la plupart des études sur l'hésitation se concentrent sur les disfluences au niveau de l'organisation temporelle du discours (rythme, durée des silences, des pauses remplies, des syllabes) [Campione and Véronis, 2005], nous avons proposé d'annoter les conversations suivant un degré d'hésitation afin de prédire celui-ci à partir de représentations acoustiques (de la même manière qu'on prédirait une émotion). Nous avons fait le choix d'annoter également des dimensions affectives pour évaluer à quel point l'hésitation est corrélée ou non aux émotions.

L'objectif de l'étude porte sur la caractérisation de l'hésitation dans la parole, c'est pourquoi nous avons décidé également d'ajouter une transcription phonétique qui permet d'évaluer les corrélations entre les représentations acoustiques et l'hésitation au niveau des phonèmes. Le corpus NCCFr a donc été segmenté en groupes de souffle, et annoté sur les dimensions suivantes :

— Dimensions affectives : valence, activation et contrôle entre -2 et 2

— Degré d'hésitation : entre -3 (très confiant) et 3 (très hésitant)

— Transcription phonétique : obtenue automatiquement à partir d'un alignement forcé entre le signal audio et le texte.

Annotations des interruptions

Ce travail s'inscrit dans les travaux de thèse de Martin Lebourdais et Rémi Uro. L'objectif est de modéliser les interruptions entre les participants d'une émission radio ou TV. À partir des données issues des archives de l'INA, nous avons sélectionné les débats et interviews de chaînes d'information où plusieurs participants sont présents en même temps. Ces données sont issues du corpus ALLIES [Shamsi et al., 2022].

Un premier protocole proposé par [Adda-Decker et al., 2008] s'est concentré sur les corrélations entre la parole superposée et la production de *disfluences*. Après segmentation, chaque segment de parole superposée est annoté suivant trois étiquettes indépendantes :

— Contribution élaborée : liée principalement à la durée de l'overlap

— Accord : avec le premier locuteur

— Interruption

Dans l'article, les catégories contribution et interruption ont montré un accord inter-annotateur dans 80% des cas, alors que pour la catégorie accord, le consensus n'est pas évident. Cette première étude a permis de faire émerger plusieurs grands types de catégories de parole superposée :

— *back-channel* : non interruptif, non élaboré, généralement neutre, parfois ressenti comme un accord,

— ajout d'information complémentaire : non interruptif, élaboré, généralement neutre ou en accord,

— *turn stealing* (interruption volontaire) : volonté claire d'interrompre le premier locuteur, même si cette action ne réussit pas,

— *anticipated turn taking* (prise de parole anticipée) : le second locuteur semble avoir perçu des éléments comme quoi le premier locuteur a fini de parler.

En nous basant sur ces travaux antérieurs, nous avons considéré que les interruptions avaient lieu majoritairement dans les zones de parole superposée. En 2022, nous avons donc réalisé une campagne d'annotation de la parole superposée dans le cadre de la thèse de Martin Lebourdais. Les quatre catégories proposées ci-dessus, ainsi que les départs simultanés ont été annotées par trois annotateurs sur plus de 4000 segments contenant de la parole superposée. Il est évident que notre protocole implique qu'une interruption n'est pas définie en dehors du cadre de la parole superposée, ce qui reste à évaluer. Les taux d'accord par segment audio sont reportés figure 4.3, le kappa de fleiss sur les 5 catégories et 3 annotateurs est de 48.7, ce qui correspond à une concordance moyenne. Les résultats de cette campagne d'annotation montrent que la perception des départs anticipés et des backchannels atteint un relatif consensus. Au contraire, dès lors que l'on introduit la notion de "volonté d'interrompre", le consensus a du mal à émerger, ce qui nous amène à conclure que la perception d'une interruption est un phénomène subjectif.

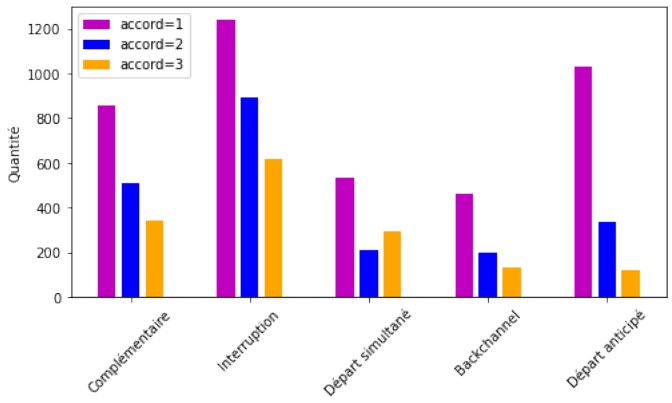


FIGURE 4.3 – Quantité de segments de parole superposée où aucun annotateur n'est d'accord, 2 annotateurs sont d'accord (bleu), tous les annotateurs sont d'accord (orange) pour les 5 catégories considérées. Le total des segments considérés est 4351.

Lieux de l'interruption

Dans le cadre de la thèse de Rémi Uro, nous avons cherché à savoir sur quels critères acoustiques ou linguistiques, un instant de la conversation est favorable pour prendre la parole, que ce soit une interruption ou non. À moyen terme, cela permettra de définir les lieux d'interruption qu'ils soient en présence de parole superposée ou non. Une étude perceptive a

été réalisée en 2022 sur 64 stimuli équilibrés en genre. La fin d'un segment se caractérise soit par une fin de phrase, soit par une pause sans que le locuteur n'ait conclu. Les segments sont ensuite tronqués de leurs derniers mots (entre 3 et 0). Les participants devaient soit lire le texte, soit écouter le signal de parole et choisir le plus rapidement possible (moins de 30 s) si la prise de parole pouvait avoir lieu "maintenant", "bientôt" ou "pas encore".

Dans le cas où le locuteur a fini sa phrase à la fin du segment, les évaluateurs ont clairement indiqué qu'il était possible de prendre la parole à la fin du segment alors qu'il faut attendre si le segment est tronqué (avec texte ou avec audio). Par contre, dans le cas où le locuteur ne conclut pas sa phrase à la fin du segment, les évaluateurs estiment majoritairement qu'il est possible de prendre la parole en fin de segment (40% "maintenant", par rapport à 30% "pas encore") avec le contenu linguistique. Dans le cas où seul l'audio est disponible, les évaluateurs ne sont pas capables de prédire la prise de parole. Ces résultats montrent également que le genre n'est pas un élément discriminant sur les interruptions. Ainsi, il semble très difficile de prédire perceptivement le lieu de l'interruption uniquement à partir de l'audio. Le contenu linguistique semble aider, mais cela reste à confirmer sur de plus grandes quantités de données.

Annotations du discours

Dans le cadre du projet SynPaFlex, des données mono-locuteur issues de livres audio ont été annotés en catégories émotionnelles, mais également en styles de discours, prosodie et prononciation [Sini et al., 2018]. Le corpus est constitué de plusieurs genres littéraires : romans, nouvelles, contes, fables et poèmes. L'objectif est d'explorer l'expressivité à partir de différentes perspectives littéraires et discoursives, dans le but de contrôler la synthèse de parole expressive.

À partir d'un schéma d'annotation très complet, un évaluateur a remplit manuellement les champs suivants :

- Patrons prosodiques : question, note, nuance (style de lecture particulier à la locutrice), suspense, résolution, chant et autre,
- Personnages : identifiant du personnage, personnalité vocale, narrateur
- Émotion : tristesse, colère, peur, joie, surprise et dégoût (inspiré de [Ekman, 1999]),
- Intensité émotionnelle : entre 1 et 3,
- Personnalité : introversion/extraversion (comment l'émotion est exprimée par la parole),
- Phonétique : substitutions, suppressions et insertions de phonèmes, élongations importantes, silences et pauses.

Attention, dans ces travaux, l'annotation émotionnelle reflète la perception d'un unique annotateur, ce qui a du sens dans un contexte de synthèse de parole, mais qui ne pourra *a priori* pas se généraliser à d'autres perceptions. Il est évident que pour la construction d'un système de reconnaissance, un plus grand nombre d'annotateurs est nécessaire afin de modéliser la subjectivité de la perception des émotions dans la parole. Une première étude de classification des émotions annotées [Sini et al., 2018] a permis de mettre en évidence deux groupes d'expressions vocales à partir de descripteurs audio uniquement : les émotions positives (joie, surprise, neutre) et les émotions négatives (tristesse, peur, dégoût). Dans une seconde expérience [Tahon and Lolive, 2018], nous avons cherché à classer les discours directs et indirects dans la parole lue d'une même locutrice, toujours à partir de descripteurs audio uniquement. Nous avons montré entre autres que la F_0 est très homogène dans les segments prédits comme étant de style direct, alors que sa dispersion est beaucoup plus élevée dans le cas du style indirect. Cela montre, que l'intonation est un élément clé du changement de style de discours, notamment lorsque le lecteur incarne un personnage.

4.2 Reconnaissance automatique des émotions

La première section de ce chapitre est dédiée aux annotations de la parole expressive. Cette seconde section présente différentes options pour l'apprentissage de modèles supervisés pour l'expressivité. Un changement d'état émotionnel implique des modifications physiologiques et en particulier des modifications au niveau de l'appareil vocal, comme nous l'avons vu au premier chapitre. Ces modifications vont entraîner mécaniquement des variations des descripteurs acoustiques liés à l'intonation, au timbre/qualité vocale, ou encore l'apparition d'*affect bursts*. De plus, les modifications d'ordre cognitif vont elles, avoir un impact sur la structuration du discours, c'est-à-dire le rythme, l'élocution (la prononciation), les *disfluences* et le choix des mots employés. C'est pourquoi, il est pertinent d'étudier deux types de représentations qui sont détaillées dans cette section : acoustique et linguistique. Avec l'apparition des modèles pour la parole pré-entrainés sur des grandes quantités de données, de nouvelles représentations émergent pour la reconnaissance des émotions.

4.2.1 Modélisation acoustique des émotions

La modélisation des expressions acoustiques des émotions nécessite de combiner deux types de représentations : les aspects prosodiques de la parole (intonation, intensité et rythme) et

les aspects spectraux (qualité vocale, contenu phonétique, etc.). Il s'agit de capturer l'émotion en tant que telle, mais également de quantifier l'axe temporel de façon appropriée.

Approche traditionnelle

Représentation du signal émotionnel Nous avons vu dans la section 2.1 les descripteurs audios pertinents pour représenter le signal acoustique de parole. Deux types de descripteurs ont été présentés : ceux qui tendent à décrire les aspects prosodiques de la parole (intonation, intensité et rythme) et ceux qui cherchent à représenter le contenu spectral du signal (formants, centre de gravité, fréquences roll-off, MFCC, Fbanks ou encore LPC (*linear prediction coefficients*)). La recherche de l'ensemble de descripteurs idéal pour la caractérisation des émotions a été et est toujours un "graal" [Eyben et al., 2016][Tahon and Devillers, 2016b].

Traditionnellement, les émotions sont représentées par des catégories annotées sur des segments de parole correspondant à un groupe de souffle. Ces segments ont des durées variables allant de quelques ms à plusieurs secondes. Il a été montré que les modèles sont capables de reconnaître automatiquement l'émotion dès la 1ère seconde [Schuller et al., 2010]. Les modèles traditionnels pour la reconnaissance d'émotion cherchent à représenter les distributions statistiques de différents paramètres spectraux et prosodiques à l'aide de fonctions (moyenne, écart-type, pentes, quartile) qui sont appliquées au niveau du segment complet. Cette combinaison entre des descripteurs prosodiques et spectraux extraits à la trame (30 ms) typiquement tous les 10 ms et des fonctions statistiques permettent de modéliser à la fois l'expression vocale de l'émotion et de quantifier l'axe temporel.

Les ensembles de descripteurs proposés pour les différents challenges Interspeech (OS384, eGeMAPS, etc.) ont permis d'avoir un premier standard des représentations acoustiques pour la reconnaissance des émotions et ainsi poser les bases pour une évaluation commune des systèmes de prédiction. Par exemple, l'ensemble OS384 proposé comme baseline pour le challenge Interspeech 2009 contient 16 descripteurs bas-niveau (*Low-Level Descriptors* - LLD) dont on prendra également les dérivées premières Δ, et 12 fonctions statistiques calculées sur le segment entier (voir tableau 4.1). La prosodie est capturé par la F_0, l'énergie, tandis que le contenu spectral est capturé par les MFCCs, le HNR (*Harmonic to Noise Ratio*, mesure du bruit) et le ZCR (*Zero Crossing Rate*, mesure du bruit et du contenu fréquentiel). Dans cet ensemble, le rythme est modélisé implicitement par l'utilisation des fonctions statistiques qui vont capturer l'évolution temporelle. Un segment de parole correspondant à une catégorie d'émotion sera alors représenté par un vecteur contenant tous les descripteurs.

LLD (16 × 2)	Fonctions (12)
(Δ) ZCR	mean
(Δ) RMS Energy	standard deviation
(Δ) $F0$	kurtosis, skewness
(Δ) HNR	extremes : value, relative position, range
(Δ) MFCC1-12	linear regression : offset, slope, MSE

TABLE 4.1 – Ensemble de descripteurs OS384 proposé pour le challenge Interspeech 2009 [Schuller et al., 2009].

Modèles Avant le paradigme neuronal, les approches statistiques ont été majoritairement utilisées pour modéliser les émotions dans la parole, principalement des SVM et dans une moindre mesure des GMM (vu au chapitre 2). Les vecteurs acoustiques sont normalisés par rapport au genre, au corpus ou bien au locuteur avant d'être fournis en entrée des modèles.

Les corpus de parole émotionnelle contiennent généralement peu de données par rapport à la dimension des vecteurs de représentation. Par exemple, l'ensemble de descripteurs du challenge ComParE contient 6373 descripteurs [Schuller et al., 2013] plus ou moins redondants entre eux. Il faut donc avant tout sélectionner automatiquement les descripteurs pertinents pour la tâche. Pour cela, plusieurs approches peuvent être envisagées : le classement des descripteurs par un calcul d'entropie ou de distance, ou bien l'apprentissage d'un modèle pour supprimer (ou ajouter) les descripteurs un à un et les classer suivant le score obtenu par le modèle (méthode avec wrapper). Pour une revue des méthodes existantes, le lecteur pourra se référer à [Tahon and Devillers, 2016a]. Les approches statistiques de type PCA (*Principal Component Analysis*) sont également efficaces pour réduire la dimension de l'entrée du modèle. La PCA permet d'obtenir un nouvel espace de représentation des données, chaque nouvelle dimension étant une combinaison linéaire des dimensions d'origine. Si on ne conserve que les dimensions principales, on assure des variables plus indépendantes et plus représentatives de la variance présente dans les données.

Un modèle SVM va chercher à discriminer deux classes en déterminant l'hyperplan séparateur entre ces deux classes. En phase de test, le modèle pourra alors prédire une classe à partir de la représentation acoustique fournie en entrée. Contrairement aux GMMs, il est difficile d'adapter un modèle SVM appris sur une tâche à une autre tâche, il vaut mieux les entraîner de zéro. Par contre, ils ont l'avantage de fournir une mesure de confiance explicite qui correspond à la distance entre la donnée et l'hyperplan.

Les inconvénients de l'approche traditionnelle est qu'elle ne permet pas de modéliser fi-

nement l'évolution temporelle de la prosodie. De plus, le calcul des descripteurs acoustiques (extraction à la trame et application de fonctions statistiques) ne permet pas de modéliser le contexte, ni de mettre en évidence des points d'ancrage temporels où l'émotion serait plus forte et plus caractéristique qu'ailleurs dans le signal. Et enfin, ces approches ne sont pas adaptées à une augmentation importante de la quantité de données. Par contre, leur simplicité et l'utilisation de descripteurs audio issus principalement de la modélisation acoustique du conduit vocal et des paramètres prosodiques, les rendent facilement interprétables.

Résultats Pendant ma thèse, j'ai étudié les descripteurs acoustiques afin de proposer des améliorations aux ensembles existants [Tahon, 2012]. En particulier, j'ai ajouté des descripteurs de rythme (débit de voisement) et d'articulation (estimation du triangle vocalique). Je me suis également attachée à adapter les descripteurs existants sur des échelles perceptives. Par exemple, exprimer les fréquences en semi-tons plutôt qu'en Hertz, utiliser des bandes de Bark, ou de Mel pour calculer l'énergie fréquentielle (ce qu'on appelle aujourd'hui les fbanks). Ces implémentations sont maintenant largement utilisées par la communauté.

Pendant mon post-doc [Tahon and Devillers, 2016b], j'ai analysé la robustesse de plusieurs ensembles de descripteurs en comparaison avec l'ensemble de référence OS384 par le biais d'expériences cross-corpus : le modèle est entraîné sur un corpus et testé sur un autre. Les 4 corpus utilisés sont en français, tous en interaction humain-robot avec des enfants ou des adultes, des données spontanées ou actées et des environnements acoustiques différents. Les résultats en termes de rappel non pondéré moyen sur 3 classes (positif, neutre, négatif) sont reportés dans le tableau 4.2.

Au LIMSI, j'ai développé un ensemble de descripteurs Li174 qui regroupe l'ensemble des améliorations fines apportées à l'existant. D'après les résultats, malgré un nombre de descripteurs plus faible, les performances sont aussi bonnes que la baseline OS384. Deux ensembles de "meilleurs" descripteurs sélectionnés automatiquement à partir d'un calcul d'entropie (descripteurs indépendants) obtiennent des performances légèrement supérieures à la baseline : LiB50 et LiB25 contenant respectivement 50 et 25 descripteurs de l'ensemble Li174. Ainsi, ces résultats montrent qu'en sélectionnant correctement les descripteurs, on gagne en performance ($+1.6$ points) avec seulement 25 descripteurs. Ce gain n'est pour autant pas significatif au vu de nos données. Enfin, trois ensembles de coefficients cepstraux ont été étudiés : LiCEP48 (incluant les Δ), LiCEP24 contenant les coefficients cepstraux de 1 à 12 extraits sur les parties voisées et non-voisées, et OSCEP24 contenant les coefficients cepstraux de 1 à 12 de la baseline OS384 (incluant les Δ). On remarque que l'ensemble LiCEP24 obtient les

Descripteurs	UAR (%)
Os-384	40.0
Li-174	40.0
Li-B50	42.0
Li-CEP48	42.8
Li-B25	41.6
Li-CEP24	43.4
Os-CEP24	35.2

TABLE 4.2 – Classification des émotions cross-corpus (apprentissage sur un corpus, test sur un autre). Résultats donnés en UAR moyen sur 4 corpus en français

meilleurs performances ($+13.4$ points, significatif cette fois). Ce résultat démontre la robustesse des coefficients cepstraux aux variations acoustiques liées à un changement de corpus, mais également la pertinence de ne pas considérer les parties voisées et non-voisées sur le même plan.

Cette expérience nous permet de conclure à l'importance de ne pas considérer le signal de parole émotionnel comme homogène. Les phonèmes ont des signatures acoustiques très différentes (principalement entre parties voisées et non voisées) qu'il faut prendre en compte lors de l'extraction des descripteurs. En effet, il semble difficile de comparer le contenu spectral de la voyelle $/\mathrm{a}/$ et du $/\int/$ à l'aide d'une moyenne sur l'ensemble du segment. La modélisation du signal émotionnel doit donc à la fois intégrer les changements acoustiques liés à la parole elle-même (au niveau phonémique), mais aussi les changements prosodiques (au niveau du groupe de souffle).

Les expériences cross-corpus sont très informatives sur l'homogénéité des annotations entre les différents scénarios de collecte. Dans une expérience réalisée au LIMSI entre des corpus *real-life* avec des personnes âgées et des corpus actés par des jeunes adultes [Tahon et al., 2015], nous avons montré que les corpus avec des personnes âgées dont on induit des émotions sont très complexes et spécifiques par rapport aux émotions collectées en conditions de laboratoire avec des jeunes adultes. En effet, les distributions des descripteurs acoustiques ne se recoupent pas de manière évidente (notamment les variations de F_0 et le jitter). De manière étonnante, un modèle d'émotion appris avec un corpus de personnes âgées ne généralise pas bien sur un autre corpus collecté dans des conditions similaires avec des locuteurs différents. Par contre, les modèles appris sur le corpus d'émotions actées par des jeunes adultes généralise assez bien aux autres conditions et locuteurs. Notre hypothèse est que la parole actée étant plus prototypique, l'apprentissage des frontières de chaque classe est plus facile pour le modèle, et

également plus pertinente lors de l'inférence.

Approche neuronale

L'utilisation des réseaux de neurones pour la reconnaissance des émotions est arrivée assez tardivement comparativement aux autres domaines du traitement automatique de la parole. Un des premiers articles sur la reconnaissance des émotions avec un réseau de neurone date de 2008 [Wöllmer et al., 2008]. La motivation principale pour utiliser un réseau récurrent avec des couches de type LSTM était de pouvoir modéliser les dépendances à long terme pour capturer l'évolution temporelle de l'émotion. Cette première étude a établit l'intérêt des réseaux récurrents pour la reconnaissance de catégories émotionnelles et pour la prédiction continue de dimensions affectives.

Cette section reprend les travaux de thèse de Manon Macary en collaboration avec l'entreprise Allo-Media. L'objectif est de prédire automatiquement un degré de satisfaction à un instant donné (allant de frustré à satisfait) à partir de conversations téléphoniques de centres d'appels issues du corpus AlloSat. La satisfaction a été annotée de façon à avoir un point tous les 250 ms, et ce, sur l'ensemble de la durée de la conversation. Dans la suite, nous appellerons segment l'intervalle de temps correspondant a un point d'annotation.

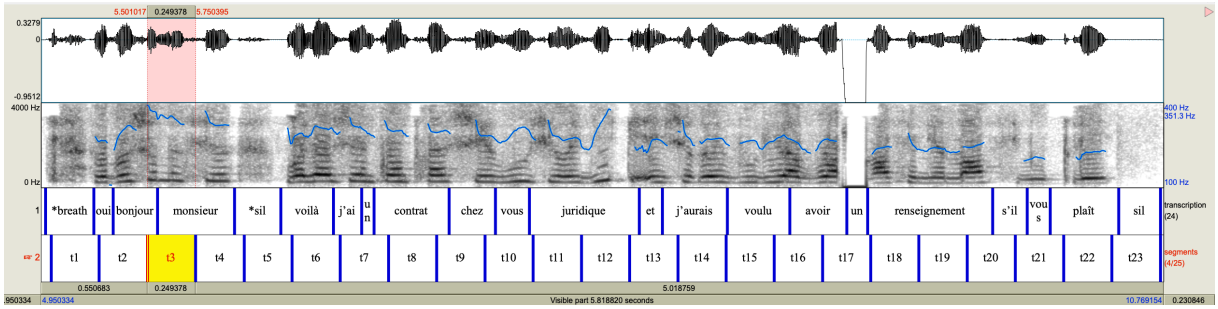


FIGURE 4.4 – Exemple d'une conversation du corpus AlloSat. La visualisation sous Praat permet de voir l'alignement entre les mots issus d'une transcription automatique et les segments émotionnels t_i d'une durée de 250 ms obtenus après l'annotation continue.

Représentation acoustique du signal émotionnel Dans un premier temps, ce réseau est alimenté par des descripteurs acoustiques (de type MFCCs ou eGeMAPs) qui prendront la forme d'une matrice de dimension $D \times T$ où D est la dimension de la représentation du signal et T est le nombre de trames.

Les descripteurs acoustiques $\mathbf{d}_n$ (MFCCs ou eGeMAPS) sont extraits à la trame n par pas de 10 ms. Pour un segment émotionnel t_i d'une durée de 250 ms (tel que représenté sur la figure 4.4), on calcule la moyenne des descripteurs sur l'ensemble du segment, soit sur $N = 25$ trames. Celui-ci est représenté par le vecteur $\mathbf{x}_i$ donné par l'équation 4.2. La séquence complète contenant T segments de 250 ms, est donnée par $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T]$.

$$\mathbf{x}_i = \frac{1}{N} \sum_{k=1}^{N} \mathbf{d}_{i+k} \tag{4.2}$$

Architecture neuronale Je présente ici l'architecture neuronale proposée par Manon Macary dans sa thèse pour la prédiction continue de la satisfaction. Cette architecture est inspirée des travaux présentés dans le challenge AVEC 2019. La tâche du challenge consiste à prédire des dimensions affectives continues telles que activation, valence et liking (corpus SEWA) [Ringeval et al., 2019]. A la suite de ces travaux, un réseau de neurone récurrent de type séquence vers séquence composé de 4 couches BiLSTM a été proposé [Schmitt et al., 2019] et a été adapté dans les travaux de Manon. Le réseau est représenté figure 4.5.

On peut noter que l'architecture est sensiblement identique à celle utilisée pour la détection du genre ou de la parole superposée, à cela près qu'elle restitue en sortie directement une valeur correspondant à la satisfaction et pas un score semblable à une probabilité d'appartenance à une classe. La différence principale réside dans la durée du champ réceptif : environ 1 à 2 s dans le cas de la parole superposée, plusieurs minutes dans le cas de la parole émotionnelle. Cette différence de traitement s'explique par le fait qu'une émotion a une évolution dans le temps qui est longue en comparaison de la précision souhaitée pour la segmentation audio.

Représentations pré-entrainées Sur le même principe que les x-vecteurs présentés section 3.3, des *embeddings* sont extraits d'un réseaux de neurone appris pour une tâche particulière. Ces réseaux de neurones sont très complexes et utilisent le principe des transformers qui ne sera pas détaillé dans ce document. La tâche est généralement une tâche de prédiction de la trame suivante, considérant qu'une partie des trames (ou mots) précédentes est masquée. Cela permet d'apprendre des modèles de parole non supervisés sur d'énormes quantités de données non annotées. Dans les travaux de Manon, nous avons expérimenté un modèle pré-entrainé qui prend un signal audio en entrée Wav2Vec1.0 [Schneider et al., 2019] et nous l'avons utilisé comme extracteur d'une représentation acoustique du signal de parole.

Le modèle Wav2Vec prend en entrée un signal audio et applique ensuite deux réseaux. L'encodeur transforme la forme d'onde en une représentation latente adaptée à la tâche. Le

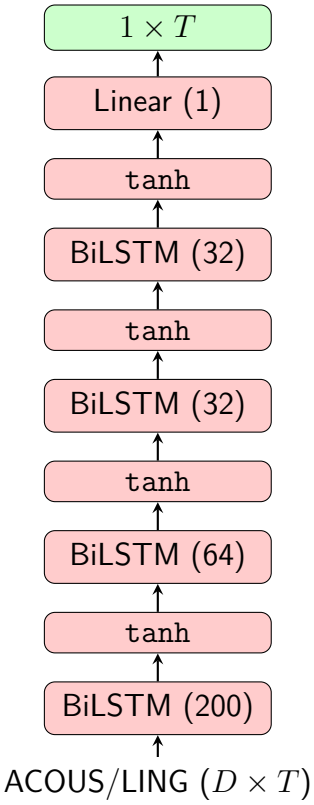


FIGURE 4.5 – Réseau de neurone récurrent pour la prédiction continue d'une dimension affective. Figure adaptée de [Macary et al., 2022]

réseau de contexte combine des fenêtres temporelles de tailles différentes afin d'obtenir des représentations contextualisées sur différents niveaux. Plus précisément, les deux réseaux sont constitués de couches convolutionnelles, de couches de normalisation et de fonction ReLU. Des connexions de type "court-circuit" (*skip connexions*) que nous avons déjà évoquées dans le cadre des réseaux résiduels, sont ajoutées pour aider le réseau à converger. Grâce au système de fenêtrage temporel, le champ de réception de la dernière couche est de 810ms. Le système est entraîné sur une tâche non supervisée, c'est-à-dire qu'il doit prédire la trame suivante en fonction du contexte précédent et courant. Ce modèle est entrainé sur 960 h de parole provenant du corpus de livres audio LibriSpeech.

Dans le cas où nous souhaitons utiliser les *embeddings* de type Wav2Vec avec le réseau récurrent décrit précédemment, deux possibilités s'offrent à nous : soit les *embeddings* sont extraits en ne donnant que le contexte du segment émotionnel de 250 ms en entrée du modèle Wav2Vec, soit ils sont extraits en donnant la séquence entière. Dans les deux cas, la représentation finale est une moyenne des *embeddings* obtenus sur le segment de 250 ms (voir

eq. 4.2). Sachant que le champ réceptif total de Wav2Vec est de l'ordre du segment émotionnel, l'extraction avec contexte n'a finalement qu'un impact négatif, comme nous le verrons dans les résultats.

Ces représentations pré-entraînées ont l'énorme avantage d'avoir "vu" plusieurs centaines d'heures de parole lors de l'apprentissage. Elles sont donc capables de représenter une large diversité des variations présentes dans les données de parole. Les corpus émotionnels étant généralement de petite taille (AlloSat : 30h), l'utilisation de ces représentations permet de compenser le manque de diversité. Lorsque ces modèles sont utilisés comme "extracteurs", les poids du réseau pré-entrainé sont alors gelés et l'apprentissage ne se fait que sur le réseau récurrent.

Il est également possible d'adapter les représentations à la tâche spécifique (*fine-tuning*) en intégrant le réseau pré-entrainé au modèle et en mettant à jour l'ensemble des poids. Cette approche nécessite plus de mémoire que la première option. Le tableau 4.3 présente les résultats obtenus en terme de régression sur la dimension de satisfaction sur les corpus de développement et de test d'AlloSat. Nous voyons que l'adaptation n'apporte pas de gain en performance sur l'ensemble de test.

		Satisfaction	
Descripteurs	**# entrée**	**Dev**	**Test**
wav2vec-EN (gelé)	512	0.851	**0.730**
wav2vec-FR (adapté)	512	**0.865**	0.635

TABLE 4.3 – Résultats CCC pour la tâche de reconnaissance de la satisfaction sur AlloSat. Figure extraite de [Macary et al., 2021]

Dans cette section, nous avons vu différentes approches neuronales ou non pour la classification ou la régression des émotions à partir d'un vecteur de représentation acoustique uniquement. Le choix d'une représentation acoustique pertinente a toujours été un objectif important dans ce domaine. L'utilisation de modèles de parole pré-entraînés couplés à des réseaux qui capturent la dynamique temporelle de l'émotion apporte un grand nombre de réponses dans ce domaine. L'inconvénient est la difficulté d'interpréter ces réponses au delà de la performance sur une tâche donnée.

4.2.2 Représentation multi-modale

L'ajout d'informations linguistiques permet d'avoir une représentation multi-modale de l'émotion exprimée qui prenne en compte à la fois l'impact des modifications physiologiques

sur la production acoustique et l'impact cognitif sur la structuration de la parole. L'aspect multi-modal de l'expression vocale des émotions a déjà été démontré. Par exemple, la valence semble être plutôt portée par le contenu linguistique que le contenu acoustico-prosodique [Gunes and Schuller, 2013]. Nous complétons dans cette section la description des travaux de thèse de Manon Macary portant plus spécifiquement sur l'utilisation jointe des représentations acoustiques et linguistiques.

Représentation pré-entraînée pour le texte Le modèle CamemBERT [Martin et al., 2020] est un modèle de type BERT [Devlin et al., 2019] adapté au français. BERT est un réseau de neurone bi-directionnel de type *transformer* conçu pour prendre du texte en entrée et qui a pour tâche de prédire le mot (ou token) suivant. Il utilise la technique de masquage qui cache aléatoirement des tokens en entrée. Le modèle doit prédire le token caché en n'utilisant que le contexte. Le modèle CamemBERT est entrainé sur la partie française du corpus OSCAR [Ortiz Suárez et al., 2019]. Word2vec [Mikolov et al., 2013] est une représentation de type *embeddings* largement utilisée dans le domaine du traitement automatique du langage et en particulier pour la détection automatique de sentiment [Pasti et al., 2020] qui a également été testée.

Synchronisation des modalités Pour extraire les *embeddings* de type CamemBERT à partir de la transcription linguistique, il suffit de fournir en entrée la séquence de tokens voulue afin d'en obtenir une représentation. Cependant, dans notre cas, la séquence de tokens doit être alignée avec le pas d'annotation, c'est-à-dire toutes les 250 ms. Dans un cas de classification au niveau d'un groupe de souffle, cette synchronisation ne serait pas nécessaire, car le segment émotionnel serait représenté par un vecteur d'*embeddings* auquel le modèle devra associer une catégorie.

De même qu'avec Wav2Vec, les *embeddings* de type CamemBERT sont extraits soit en fournissant uniquement la séquence de tokens prononcés sur le segment de 250 ms (par exemple pour le segment t_3 de l'exemple illustré figure 4.4, les mots "bonjour monsieur"), ou bien la séquence entière depuis "oui" jusqu'à "plait". La séquence d'*embeddings* extraite aura pour dimension le nombre de tokens fourni en entrée. Afin de se synchroniser avec les segments émotionnels, si le segment ne contient qu'un token, il ne sera représenté que par les *embeddings* de ce token, si le segment contient plusieurs tokens, il sera représenté par une moyenne (éventuellement pondérée par la durée) des *embeddings* de ces tokens. Le champ réceptif des modèles BERT est de plusieurs mots, en fait la séquence entière. Ainsi, nous verrons que la

prise en compte d'un contexte large est bénéfique au modèle.

Dans le cas des word2vec, la synchronisation est plus simple car la notion de contexte n'existe pas. La représentation linguistique est extraite au niveau mot et moyennée au niveau du segment de 250 ms.

Fusion des modalités Nous avons exploré différentes représentations classiques (word2vec et MFCCs) et pré-entrainées (CamemBERT et Wav2Vec), linguistique et acoustique pour la tâche de prédiction continue de la satisfaction sur le corpus AlloSat. Nous avons également chercher à fusionner les informations des deux modalités. L'avantage des réseaux de neurones pour la fusion est qu'il est assez simple de concaténer deux couches provenant d'entrées différentes, et ce à différentes profondeurs du réseau. De plus, l'entraînement du réseau grâce à cette architecture permet d'apprendre des représentations conjointes des deux modalités évoluant ainsi dans un même espace. Dans l'expérience décrite dans l'article [Macary et al., 2022], nous avons testé plusieurs types de fusion :

— Précoce : les matrices d'entrée des deux modalités sont concaténées. Cette approche a un inconvénient majeur : les matrices acoustique et linguistique sont distribuées dans des espaces différents ;

— Intermédiaire : deux couches intermédiaires venant de modalités différentes sont concaténées et le résultat est fourni en entrée de la suite du réseau.

— Tardive : deux réseaux indépendants prédisent chacun une valeur, la décision finale est la somme pondérée de ces deux valeurs. Les coefficients de pondération sont optimisés sur l'ensemble de développement. C'est ce type de fusion qui fonctionne le mieux dans notre cas.

Résultats Les résultats par modalité et avec une fusion des décisions issues des deux modalités sont présentés dans le tableau 4.4.

Les résultats de ces travaux montrent clairement l'avantage des représentations pré-entrainées sur les représentations traditionnelles (MFCC ou word2vec). Nous montrons également que les représentations linguistiques, surpassent largement les représentations acoustiques pour cette tâche. La modalité acoustique permet d'améliorer les résultats en fusion sur l'ensemble de développement uniquement.

Le fait que la modalité linguistique ait une si grande importance dans les performances nous a interpellé. En effet, la plupart des modèles de reconnaissance des émotions se basent quasi-exclusivement sur des représentations audio, et cela permet de se conformer à la notion

		Satisfaction	
Modalité	**# entrée**	**Dev**	**Test**
ACOUSTIQUE			
MFCC	48	0.851	0.651
wav2vec-EN (sans)	512	0.844	0.806
wav2vec-EN (avec)	512	0.823	0.656
LINGUISTIQUE			
Word2vec	40	0.883	0.881
CamemBERT (sans)	768	0.916	0.817
CamemBERT (avec)	768	0.917	**0.924**
FUSION TARDIVE			
0.34 MFCC + 0.66 word2vec	48 + 40	0.897	0.840
0.28 wav2vec-EN + 0.72 CamemBERT	512 + 768	**0.932**	0.920

TABLE 4.4 – Résultats CCC de fusion multi-modale pour la tâche de reconnaissance de la satisfaction sur AlloSat. Tableau extrait de [Macary et al., 2022]

de para-linguistique. C'est pourquoi nous avons voulu aller plus loin dans l'analyse afin de comprendre pourquoi le linguistique était si important.

Une analyse linguistique sur les conversations frustrées montrent que dans le cas des conversations téléphoniques spontanées, les *disfluences* sont très importantes, on note également la présence de marqueurs sémantiques. Les modèles d'aujourd'hui tels que CamemBERT sont capables de capturer de tels phénomènes ce qui explique les très bonnes performances obtenues avec le linguistique seul. Pour une conversation téléphonique, on peut supposer que le contrôle est très important, et que l'émotion passera plus facilement par les mots que par les aspects prosodiques et timbraux. De plus la qualité téléphonique peut dégrader la qualité des représentations acoustiques, par exemple sur la qualité vocale. Pour plus de détails, le lecteur pourra se référer à la thèse de Manon Macary [Macary, 2022].

Le principal inconvénient de ces représentations pré-entraînées est qu'elles sont difficilement interprétables. Au vu des performances obtenues, on suppose qu'elles permettent de modéliser à la fois le contenu spectral (comme les MFCC), les aspects prosodiques (incluant l'organisation temporelle) et les aspects linguistiques et phonétiques. Un énorme chantier s'ouvre alors pour analyser les phénomènes acoustiques et linguistiques pertinent pour la reconnaissance d'émotion capturés par les *embeddings*. Ce problème est assez complexe, car les données sur lesquelles sont entraînées les modèles ne contiennent pas de contenu émotionnel explicite.

Les modèles classiques pour la reconnaissance d'émotion cherchent à représenter les distributions statistiques de différents paramètres spectraux et prosodiques à l'aide de fonctions

(moyenne, écart-type, pentes, quartile) qui sont appliquées au niveau du groupe de souffle. Or le pas d'annotation utilisé ici (de 250 ms) est plus court qu'un groupe de souffle, plutôt de l'ordre du mot. Wav2Vec utilise un contexte court au sens de la phrase prosodique ($<$ 250 ms), alors que CamemBERT utilise un contexte très large de plusieurs mots pouvant se rapprocher du groupe de souffle. Au vu de leurs performances respectives, on peut supposer que le large contexte de CamemBERT est bénéfique pour la reconnaissance d'émotion et permet de mieux capturer les aspects prosodiques. Un élément fort de ces représentations réside donc dans l'utilisation du contexte par les modèles, ce qui désambigüise certains phénomènes. Par exemple, la tristesse se manifeste généralement par une baisse de la F_0, mais suivant le contexte, une baisse de F_0 peut aussi signifier la fin d'une phrase.

4.3 Apprentissage actif pour l'annotation semi-automatique

4.3.1 Le principe de l'apprentissage actif

Cette technique permet d'adapter les modèles à une situation réelle en fonction des retours d'un utilisateur humain [Settles, 2009]. L'annotation semi-supervisée à l'aide de l'apprentissage actif est un compromis intéressant pour réduire l'intervention manuelle et améliorer les performances des méthodes d'apprentissage existantes. Ces dernières années, l'utilisation de l'apprentissage actif pour l'annotation semi-supervisée s'est développé dans le domaine du traitement automatique de la parole et du langage [Klie et al., 2018], [Marinelli et al., 2019], [Long et al., 2018], [Thiam et al., 2016].

Le principe est le suivant. Supposons que nous ayons à notre disposition une grande quantité de données, mais qu'une faible partie de celle-ci est annotée. Nous pouvons apprendre un premier modèle à partir des données annotées, qui nous permettra alors d'annoter automatiquement les données qui n'ont pas encore d'étiquette. On demande alors à un humain de confirmer ou non l'étiquette prédite par le système sur un nombre limité de données. À partir de ces nouvelles données vérifiées, il est possible d'apprendre un nouveau modèle, ou d'adapter le modèle existant. Les étapes se résument aux suivantes :

1. Apprentissage d'un modèle initial,
2. Prédiction des étiquettes sur les données non annotées,
3. Sélection des données pertinentes et confirmation par un humain,
4. Adaptation/Apprentissage du modèle $\rightarrow$ on retourne à l'étape 2.

Pour réduire les coûts d'annotation des émotions ou d'autres traits expressifs, l'utilisation d'apprentissage actif va se développer.

Le point le plus important sera de choisir un critère de sélection des données à annoter pertinent pour la tâche. D'après [Abdelwahab and Busso, 2017], les stratégies les plus utilisées pour cette sélection, sont les suivantes :

— Sélection par confiance : exemples pour lesquels le classifieur est le moins confiant,

— Sélection par comité (*Query By Committee* - QBC) : exemples pour lesquels plusieurs classifieurs ne sont pas d'accord,

— Prédiction de l'erreur attendue : les exemples sélectionnés doivent minimiser l'erreur attendue si ils sont incorporés à l'ensemble d'apprentissage,

— Réduction de la variance : les exemples sélectionnés tendent à minimiser la variance de la distribution estimée sur les prédictions,

— Stratégies de pondération des densités : choisir les exemples en fonction de la distribution, en évitant les *outliers*.

Suivant les modèles utilisés, certains critères seront plus simples à mettre en œuvre que d'autres. Par exemple, la distance d'un échantillon à l'hyperplan obtenu avec un SVM peut être une mesure de confiance du modèle très simple à obtenir. Les données les plus informatives sont celles comprises entre l'hyperplan et les marges. Ces données sont aussi celles pour lesquelles le SVM a le moins confiance. C'est l'approche utilisée par [Abdelwahab and Busso, 2017]. La mesure de confiance peut être aussi obtenue en modélisant l'accord inter-annotateur [Zhang et al., 2013]. La méthode QBC peut être utile dans les cas où une mesure de confiance n'est pas facilement accessible. Lorsqu'on adapte le modèle existant ou qu'on apprend un nouveau modèle avec les nouvelles données, il est possible de pondérer les données déjà vues en fonction d'un facteur d'oubli [Prokopalo et al., 2020].

4.3.2 Applications

Apprentissage actif pour l'annotation d'émotion

Le corpus SynPaFlex [Sini et al., 2018] n'a été annoté que partiellement en émotion sur les 80h que contient la base de données. L'objectif du stage de M2 de Frédéric LeBellour en 2018, réalisé en co-encadrement avec l'IRISA était de proposer un protocole d'apprentissage actif pour l'annotation semi-supervisée du corpus [LeBellour, 2018].

Protocole La partie annotée du corpus contient 8750 fichiers audio correspondant à 15 heures de lecture. Pour cela, nous sommes partis d'un modèle simple de type SVM pour classifier 5 catégories émotionnelles (joie, tristesse, peur, colère et dégoût). La segmentation en groupes de souffle homogène sur le plan expressif n'a pas été étudiée, et l'approche retenue était de découper un fichier audio en segment de 5 s décalés par rapport au précédent de 500 ms.

Chaque segment est ensuite représenté à partir de l'extraction des 384 descripteurs de l'ensemble OS384 [Schuller et al., 2009]. Une sélection de type *forward* permet de ne conserver que 200 descripteurs utiles pour la tâche de classification des émotions.

Une première expérience de classification entre les segments considérés comme expressifs, et ceux qui ne le sont pas (RAS ou "neutre") montre un bon taux de reconnaissance de plus de 90% de f-score. Après équilibrage des classes, 667 segments par catégorie émotionnelle sont conservés. Un tiers des segments est conservé pour le test. Avec un modèle SVM, en cross-validation à 5 plis, on obtient un f-score de 70% sur les 5 classes, ce qui est assez consistant. On observe toutefois une forte confusion entre tristesse, dégoût et colère sur la matrice de confusion illustrée figure 4.6.

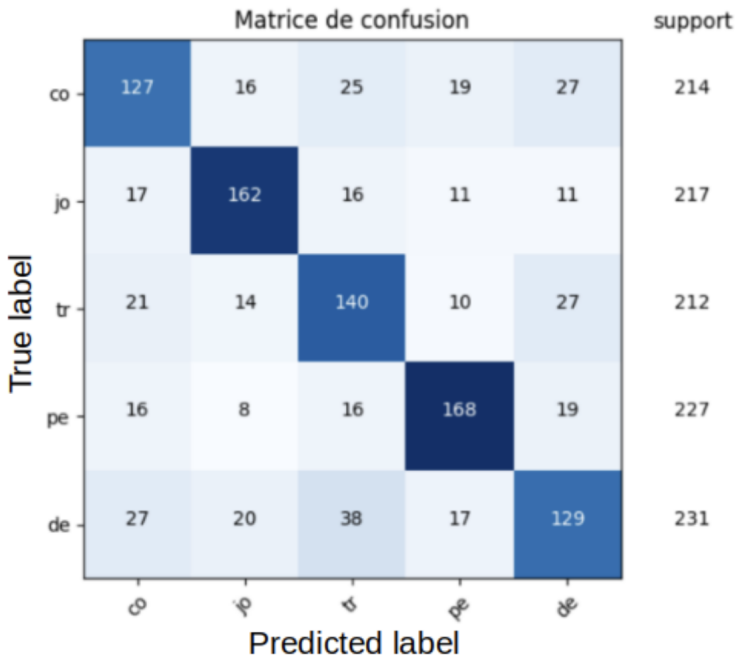


FIGURE 4.6 – Matrice de confusion obtenue sur l'ensemble de test des données annotées de SynPaFlex avec un modèle SVM. Figure extraite du rapport de F. Le Bellour [LeBellour, 2018]

La mise en place de l'apprentissage actif passe par la proposition d'une mesure d'incertitude. Pour cela, nous reprenons l'idée de la distance entre la donnée et l'hyperplan séparateur du SVM. Un problème cependant se pose pour une classification à plusieurs classes. En effet, une donnée proche d'un hyperplan entre deux classes n'est pas nécessairement incertaine car elle peut appartenir à une troisième classe. En effet, la fonction de décision retourne pour chaque donnée, la distance par rapport à chaque séparatrice. La décision du classifieur est la classe correspondant à la distance la plus élevée. Ainsi, une donnée peut se retrouver dans une classe sans être du bon côté de la séparatrice. A partir de ces distances, on peut construire les fonctions d'incertitude suivantes :

— Distance maximum par rapport à l'ensemble des séparatrices,

— Distance moyenne par rapport à l'ensemble des séparatrices,

— Tirage aléatoire,

— QBC à l'aide de modèles complémentaires (ici SVM à noyau polynomial, linéaire, random forests et k-nearest neighbours)

Pour les simulations, les données annotées sont séparées en deux ensembles : l'un pour l'apprentissage initial (20%), l'autre représentant l'ensemble des données non-annotées.

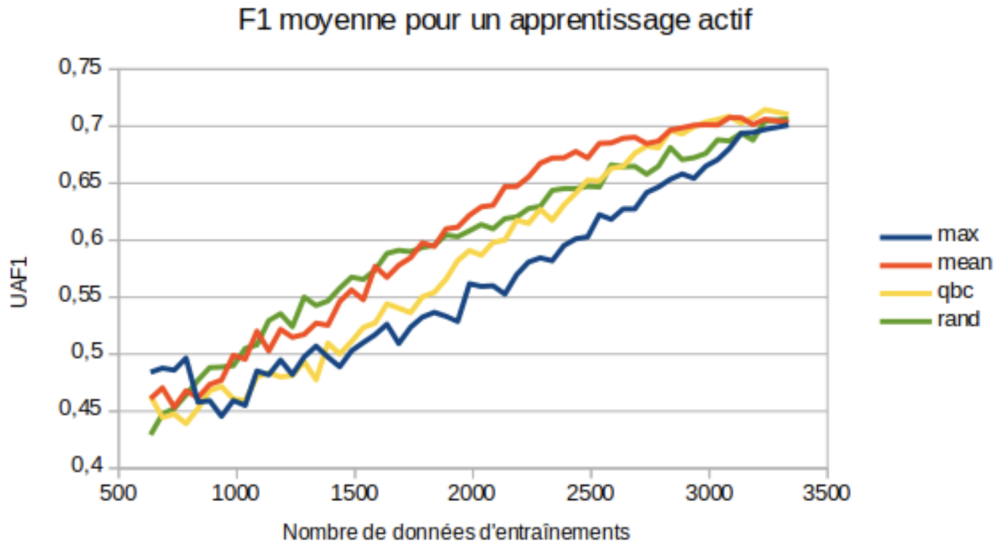


FIGURE 4.7 – Simulation d'un apprentissage actif évalué par un f-score avec différentes fonctions d'incertitude. Figure extraite du rapport de F. Le Bellour [LeBellour, 2018]

Résultats La figure 4.7 présente l'évolution du f-score en fonction des différentes fonctions d'incertitude et du nombre de données d'entrainement. L'évaluation est réalisée en cross-validation à 5 plis sur les données annotées. L'ensemble des données annotées est augmenté de 50 nouveaux segments à chaque itération. Toutes les approches permettent d'aboutir au même résultat de 70% avec la totalité des segments. La simulation présentée sur la figure ne représente qu'un choix de données initial, une autre évolution aurait pu avoir lieu si les données avaient été tirées dans un ordre différent. Cependant, les tendances sont semblables sur l'ensemble des tirages. Nous pouvons observer que pour quelques itérations (< 100), le tirage aléatoire est clairement moins performant que les autres fonctions de décisions, particulièrement le "max". Malheureusement les résultats obtenus ne sont pas en faveur de l'apprentissage actif car la sélection des données ambigües ne semble pas aider la classification. Un des inconvénients de ce protocole est le manque de données équilibrées pour construire un ensemble de test fixe pour l'évaluation complète du protocole. Le fait que le corpus SynPaFlex ne contienne qu'une seule locutrice a été motivé par des applications en synthèse de parole, mais n'est pas forcément pertinent si l'on souhaite généraliser le protocole d'apprentissage actif.

Apprentissage actif pour l'annotation de l'hésitation

Dans le cadre d'une collaboration avec le LPP, nous avons fait annoter une partie du corpus NCCFr avec un degré d'hésitation associé à un segment de parole (voir section 4.1.3 et l'article [Wottawa et al., 2020]). Dans la continuité des travaux de stage de Frédéric Le Bellour, l'objectif de cette étude était de développer un protocole pour l'annotation semi-supervisée de l'hésitation dans des conversations spontanées.

Protocole Le corpus de conversations contient 32 locuteurs et est découpé en trois ensembles de manière à ce que les locuteurs de l'ensemble *Test* soient différents de ceux présents dans les ensembles d'apprentissage. Cette répartition simule l'arrivée de nouvelles données venant de locuteurs inconnus.

— *Test* : 10 premières minutes de 2 locuteurs (1H, 1F) (360 segments)

— *Small* : 2 premières minutes de 30 locuteurs (1116 segments)

— *Large* : 10 premières minutes de 30 locuteurs (5474 segments)

Nous avons exploré six modèles de régression qui prédisent le degré d'hésitation d'un segment (compris entre -3 et 3) : SVR (SVM adapté pour la régression, avec des noyaux

polynomial, linéaire ou gaussien), régression linéaire ; et des méthodes incluant de la sélection automatique de descripteurs (lasso et ridge). Ces deux dernières techniques de régression sont généralement utilisées pour créer des modèles parcimonieux en présence d'un grand nombre de descripteurs. Elles sont censées évitées le sur-apprentissage. Les modèles sont évalués en terme de RMSE entre la valeur prédite et la valeur de référence. Un ensemble très complet de 232 descripteurs acoustiques incluant des informations spectrales, phonétiques et prosodiques a été proposé par le LPP et extrait pour chaque segment annoté.

Les modèles initiaux sont appris avec les 6 modèles m décrits ci-dessus sur l'ensemble de données *Small*. La fonction de décision est basée sur l'approche QBC où on considère que la dispersion des valeurs prédites par chacun des modèles est informative pour sélectionner les données à annoter.

1. Pour chaque modèle $m \in [1, \ldots, 6]$ on prédit les degrés d'hésitation $\hat{y}_{m,i}$ de chaque segment i. On obtient alors les scores par modèle $s_{o,m}$, resp. $s_{t,m}$, sur l'ensemble des données à annoter $i \in \mathcal{O}$, resp. sur les données de test $i \in \mathcal{T}$.
2. Pour chaque donnée à annoter $i \in \mathcal{O}$, on détermine la moyenne μ_i et l'écart-type σ_i sur les prédictions de chaque modèle $\sigma_i^2 = \frac{1}{6}\sum_{m=1}^{6}(\hat{y}_{m,i} - \mu_i)^2$
3. On trie les données à annoter sur σ_i croissant, et obtient la liste ordonnée $\mathcal{L}_o$ des données à annoter $i \in \mathcal{O}$
4. Si la sélection se fait sur une **incertitude maximum**, on prend les N premiers indices de $\mathcal{L}_o$.
   Si la sélection se fait sur une **incertitude minimum**, on prend les N derniers indices de $\mathcal{L}_o$.
5. 6 nouveaux modèles sont ré-entrainés (et optimisés, avec sélection de descripteurs) avec les N nouvelles données.
6. On itère les opérations ci-dessus tant que card$(\mathcal{O}) \leq N$

Résultats Pour cette simulation on prendra $N = 100$. L'évolution des RMSE sur l'ensemble de test et de développement sont représentées sur la figure 4.8. Attention, l'ensemble de développement correspond à l'ensemble $\mathcal{O}$ de données à annoter, il est donc variable au cours des itérations. Au cours des itérations, l'évaluation se fait donc sur un nombre de plus en plus petit de données. On peut observer que la sélection des données ayant la plus grande certitude (vert et bleu) semble bénéfique sur le résultat final sur $\mathcal{O}$ (traits pointillés). Les modèles ré-entrainés avec ces données sûres atteignent de bonnes performances sur $\mathcal{O}$ lorsque

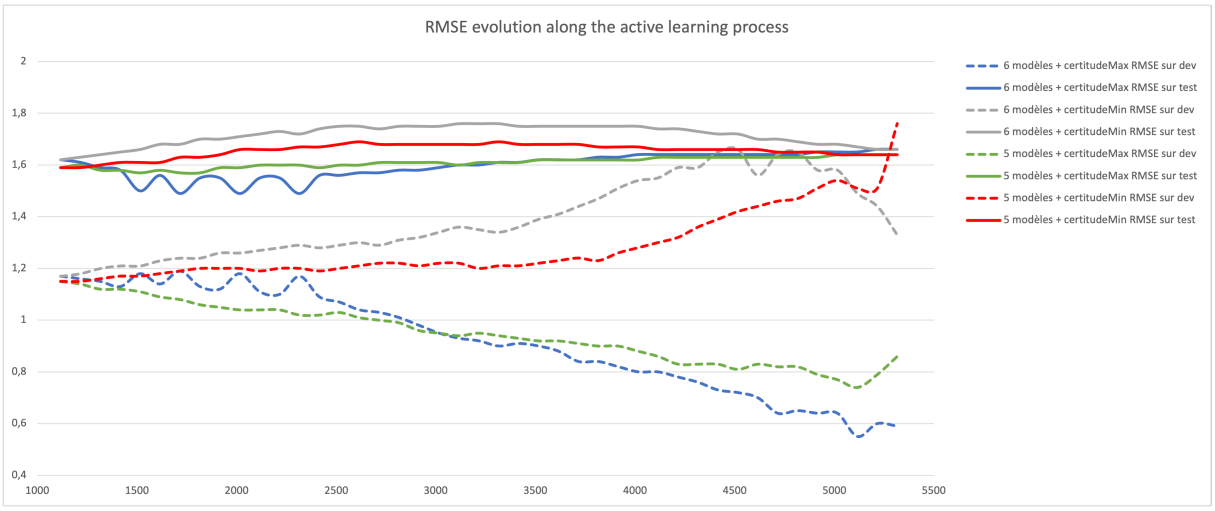


FIGURE 4.8 – RMSE en fonction du nombre de données annotées. Simulation d'un apprentissage actif évalué par une RMSE avec deux fonctions d'incertitude basée sur l'approche QBC sur 5 ou 6 modèles. L'évaluation se fait soit sur le dev dont le nombre de segment diminue avec les itérations (pointillés) soit sur le test (traits pleins) qui reste identique.

toutes les données sont utilisées et elles ne dégradent jamais les performances du modèle. Au contraire, la sélection des données ayant la plus grande incertitude (rouge et gris) ne semble pas pertinente : pour un nombre de données entre 3000 et 4000 elle dégrade clairement la RMSE sur les données de test. Ce résultat va à l'encontre des approches généralement utilisées en apprentissage actif : dans notre étude, il semble préférable de confirmer l'annotation des données pour lesquelles les modèles sont le plus d'accord.

Par contre, la conclusion va dans le même sens que l'observation illustrée section 4.2, où nous avions remarqué qu'il était préférable d'apprendre un modèle de reconnaissance d'émotion sur des données prototypiques plutôt que des données spontanées. Ces deux observations montrent que dans un contexte où la quantité de donnée est limitée et où la caractéristique est subjective, il est préférable que les contours des différentes classes soient clairement identifiés et qu'il vaut sans doute mieux avoir peu de données fiables que beaucoup de données bruitées par l'hétérogénéité des annotations.

Apprentissage actif pour le regroupement en locuteur

Dans le cadre du projet ALLIES, j'ai participé à la mise en place d'un protocole d'apprentissage actif pour la tâche de segmentation et regroupement en locuteur [Shamsi et al., 2022].

Je ne détaillerais pas ici l'ensemble de la démarche mais uniquement quelques éléments clés.

La segmentation n'est pas évaluée, seulement le regroupement en locuteurs avec la métrique DER vue précédemment. Comme vu dans la section 3.4, le regroupement consiste à fusionner deux clusters qui appartiendraient au même locuteur. Dans un contexte où l'humain est intégré dans la boucle de l'apprentissage, on peut lui demander de confirmer la pertinence d'une fusion entre deux clusters. L'annotateur devra alors écouter deux segments provenant de chacun des clusters et confirmer si ils proviennent bien du même locuteur ou non (cas 1). On peut également lui demander si deux segments considérés comme provenant d'un même locuteur appartiennent bien au même locuteur (cas 2).

Les fonctions de sélection des données à présenter à l'annotateur humain sont déduites des distances entre les clusters en question. Les distances entre les clusters sont calculées à partir d'une PLDA entraînée sur l'ensemble d'apprentissage, et d'un seuil défini sur l'ensemble de développement. Pour la correction manuellement assistée, on définit une mesure de confiance c définie comme étant le score PLDA pour le cas 1, et l'inverse du score de PLDA pour le cas 2. Les données sont triées suivant ces deux critères et présentées à l'humain pour validation. Un énorme travail a été fait dans les articles [Shamsi et al., 2022] et [Prokopalo et al., 2022] pour définir des critères d'évaluation des systèmes d'apprentissage actif. Ces critères mesurent à la fois la performance des systèmes (DER, Taux d'erreur de Jaccard, couverture, pureté des clusters) mais aussi le nombre d'intervention de l'annotateur humain qui doit rester le plus faible possible. Les résultats montrent que les premières questions sont celles qui apportent le plus de bénéfices. En effet, c'est dans ces premières questions que l'on va sélectionner les données (ou clusters) les plus informatifs.

4.4 Conclusion et discussions

Modélisation de la voix expressive L'ensemble des travaux présentés autour de l'analyse et la reconnaissance automatique de styles expressifs montrent la diversité de cette thématique et la difficulté à trouver une définition consensuelle. Au cours de ces travaux, j'ai à la fois développé des systèmes prédictifs, mais également proposé des méthodes pour interpréter ces systèmes, principalement à l'aide de descripteurs audio. À l'heure des réseaux de neurones, de nouvelles méthodes d'interprétabilité émergent, notamment avec l'utilisation des *embeddings*. Cependant, il semble nécessaire pour fournir une "explication" compréhensible des résultats du système de revenir aux modèles physiques de production de la parole et aux études linguistiques.

Nos travaux ont également l'intérêt d'utiliser des représentations pre-entraînées en entrée

des systèmes de caractérisation automatique. Nous avons montré que ce type d'approche permet d'économiser du temps d'apprentissage et d'atteindre de meilleures performances. Elles sont aujourd'hui largement utilisées pour la reconnaissance automatique des émotions. Par contre, ces approches ne permettent pas d'interpréter facilement les résultats obtenus. Nous avons également étudié l'utilisation de la modalité linguistique conjointement à la modalité acoustique à partir d'une transcription automatique et nous avons montré que dans la plupart des cas le linguistique apportait beaucoup d'information pour la caractérisation de l'expressivité. Cela tient à deux éléments : tout d'abord les bonnes performances des systèmes de transcription actuels, et aussi le fait que le linguistique capture une organisation structurelle sur un temps plus long que ne le font les représentations acoustiques. Avec ces deux modalités, nous pouvons représenter le signal de parole sur une structure rythmique multi-échelle comme présentée dans le chapitre 1.

Apprentissage actif Les résultats obtenus avec l'apprentissage actif ouvrent de nouvelles perspectives en ce qui concerne l'annotation semi-supervisée des corpus de parole expressive. À l'heure actuelle, les performances obtenues ne sont pas aussi élevées qu'attendues, mais il reste encore beaucoup à explorer dans le processus de sélection des "bons" exemples à confirmer par l'humain. Une limitation importante est la modélisation de l'annotation par l'humain et la prise en compte de l'incertitude sur l'incertitude sur le choix de l'étiquette. Les protocoles actuels ne permettent pas d'estimer si l'étiquette choisie par l'annotateur lors de l'annotation du corpus complet, n'aurait pas été différente si l'annotateur avait eu à confirmer une étiquette proposée par un système automatique, ou si un autre annotateur avait annoté le même segment.

Perspectives Enfin, la majorité des techniques que j'ai utilisées repose des apprentissages supervisés. Ceux-ci sont limités dans le sens où ils reposent sur une ontologie définie à l'avance par un groupe d'expert ayant une perception plus large du phénomène (contexte, modalités, expérience, etc.) que le modèle avec le signal audio uniquement. Ainsi la notion de catégories émotionnelles me semblent aujourd'hui obsolète. En effet, nous savons que ces catégories sont des mots qui renvoient à des concepts très généraux alors que dans un corpus collecté suivant un contexte donné, telle catégorie aura une implémentation très spécifique, différente de celle obtenue dans un autre corpus. Le développement de nouvelles approches semi-supervisées permettrait de trouver un compromis entre la vision "machine" et celle des experts humains.

Le concept d'émotion est de mon point de vue, assez artificiel. En effet, à quoi peut servir dans la vraie vie une application qui prédirait qu'une personne est en colère sans en savoir la

cause. De plus, cette ontologie par catégorie inclut généralement une classe neutre, qui est en réalité une classe "poubelle" désignant soit une absence d'émotion, soit que l'annotateur a reconnu une émotion qui ne fait pas partie des catégories disponibles. Cette classe neutre, n'a pas véritablement de sens en réalité, elle ne fait d'ailleurs jamais partie des modèles psychologiques (les *big six* ou le modèle *circumflex*). La présence de cette classe sous entend l'existence d'un état normal, ou neutre, et d'états "anormaux" guidés par les émotions. Ce glissement sémantique me semble risqué et peu utile dans le développement d'applications réelles. De plus ces catégories sont étiquettées sur un segment de parole entier, considérant ainsi que l'émotion est homogène sur l'ensemble du segment. Or il a été montré que des points d'ancrage existaient et correspondent au climax émotionnel.

Je propose donc de poursuivre ces recherches en utilisant systématiquement des dimensions continues qui varient au cours du temps comme nous l'avons fait dans la thèse de Manon Macary.

Chapitre 5

Synthèse de parole expressive

Ce chapitre débute par une brève introduction des systèmes de synthèse à partir du texte (section 5.1) au travers deux techniques : la synthèse par sélection et concaténation d'unités, et la synthèse paramétrique neuronale. À partir de cette introduction, je me focalise sur la synthèse expressive principalement suivant deux aspects : la prononciation et la voix, incluant le timbre et la prosodie.

Les travaux sur la prononciation expressive présentés dans la section 5.2 ont été conduits entièrement à l'IRISA dans l'équipe de recherche Expression pendant mon Post-doctorat (2015-2017) dans le cadre du projet ANR SynPaFlex coordonné par Damien Lolive. Plusieurs publications en présentent les résultats. La section 5.3 exposent des recherches sur la modélisation de la voix expressive menées à la fois pendant mon Post-Doc à l'IRISA et au LIUM après ma prise de poste en 2017. Les recherches au LIUM sont liées à la thèse de Thibault Gaudier qui a débuté en octobre 2021, elles sont très récentes et n'ont pas encore donné lieu à des publications.

5.1 Systèmes de synthèse de la parole à partir du texte

Dans cette section, je décris brièvement les systèmes de synthèse que j'ai utilisé pour mes travaux de post-doctorat et à l'Université du Mans : un système de synthèse par sélection d'unité et concaténation développé par l'équipe de l'IRISA et un système neuronal utilisé au LIUM.

5.1.1 Vue générale

Les premiers systèmes de synthèse de parole étaient mécaniques (machine parlante de Von Kempelen, 1791) et commandés manuellement pour modifier l'articulation des voyelles principalement et quelques consonnes. Puis, à partir des années 50 apparaissent les premiers synthétiseurs à filtres formantiques (*Parametric Artificial Talker* de Lawrence, 1952) qui modé-

lisent directement la voix suivant le modèle source-filtre présenté au chapitre 1. La modélisation du conduit vocal se fait grâce à des filtres disposés en cascade, tandis que la génération de la source se fait soit par un générateur d'impulsions à une fréquence F_0, soit un générateur de bruit, comme par exemple dans OVE-II [Fant et al., 1962]. Rapidement, la contrainte de synthétiser un signal de parole à partir d'une commande textuelle est apparue (*Text-to-Speech* synthesis, TTS). Le synthétiseur OVE-II, développé par les Bell Labs était commandé par 8 potentiomètres qui permettaient de modifier les fréquences des filtres. À titre informatif, il a fallu un hiver pour apprendre à réaliser une phrase !

D'autres types de modélisation ont été explorées, comme les modèles articulatoires qui vont chercher à reproduire la physiologie humaine (géométrie et mouvement). Puis sont apparues deux approches basées sur des données enregistrées plutôt que sur des modèles physiques : la synthèse par concaténation et la synthèse paramétrique. Ces approches font aujourd'hui partie du domaine du traitement automatique de la parole. Pour une revue historique des systèmes de synthèse de parole à partir du texte, le lecteur pour se référer à [Lemmetty, 1999] (chap. 2) et [Guennec, 2016] (chap. 2).

Il faut noter que les recherches autour de la synthèse de parole sont clairement pluridisciplinaires [Boeffard et al., 2012] (introduction) : considérations physiologiques, acoustiques, linguistiques. La synthèse est utilisée dans des contextes variés incluant la santé et la musique. Dans le cas de la musique, on pourra se référer par exemple aux travaux de C. d'Alessandro autour de la synthèse de chant performative contrôlée par le geste [d'Alessandro, 2011].

5.1.2 Synthèse par concaténation

Le principe de base de cette approche est de sélectionner des échantillons de signaux de parole existant, et de les concaténer afin de produire une séquence de parole. Ces échantillons correspondent à des unités phonétiques, généralement des diphones, c'est-à-dire la fin d'un phonème et le début d'un autre. Si en français on considère généralement 35 phonèmes, la base de données d'unités en contexte devra contenir une quantité importante de diphones $< 35^2$. L'intérêt des diphones est de pouvoir prendre en compte la co-articulation entre deux phonèmes consécutifs [Dixon and Maxey, 1968].

La principale problématique est qu'un diphone donné ne sera représentatif que d'un unique contexte (linguistique, syntaxique, sémantique, phonétique, prosodique et expressif). Afin de rendre le signal généré intelligible et adapté au contexte, il va donc falloir multiplier la quantité d'unités avec des contextes différents. Ces unités sont donc stockées dans une base de données avec leur contexte. Le processus de synthèse à proprement parler se résume alors à

une recherche de meilleur chemin dans une base de données très large.

L'outil ROOTS [Chevelu et al., 2014] développé à l'IRISA permet de représenter une séquence d'unités avec son contexte, comme illustré sur la figure 5.1. L'unité (représenté par un graphème sur la figure) contient toutes les informations contextuelles, il n'y a pas besoin d'aller regarder ce qui se passe au voisinage dans la séquence.

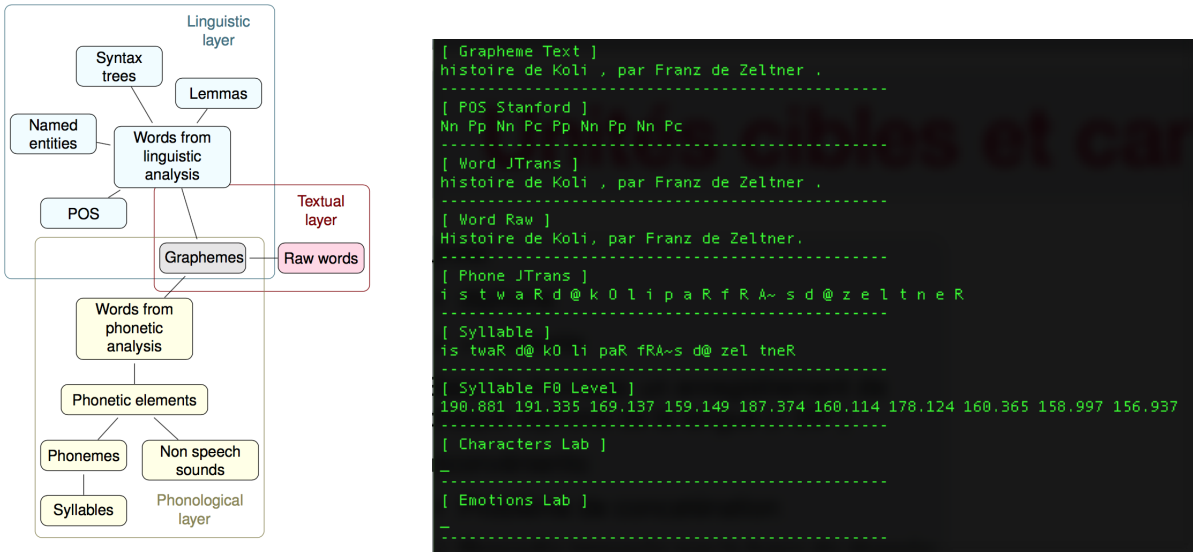


FIGURE 5.1 – Visualisation d'une séquence d'unités candidate avec l'outil ROOTS. La figure de droite présente les différents types de séquences qui peuvent être associées aux unités (ici des phonèmes) (extraite de [Chevelu et al., 2014]. La figure de gauche montre un exemple de séquence issue du corpus SynPaFlex.

Le texte à synthétiser est traité afin de déterminer la séquence d'unités cibles souhaitée ainsi que les contextes de chacune des ces unités. Il va ensuite falloir déterminer la séquence d'unités candidates permettant d'obtenir la séquence qui "sonne" le mieux. Deux coûts sont ainsi définis :

— Les contextes de chaque unité cible et candidates doivent être les plus proches possibles. Certains éléments doivent même être identiques comme le diphone lui-même → *coût cible*.

— Les unités candidates doivent se joindre sans artefact → *coût de jonction*.

Le meilleur parcours dans le treillis des unités candidates est retenu grâce à un algorithme de Viterbi. Pour quantifier la meilleure séquence (*coût cible)*, il s'agira de quantifier 1) la similarité linguistique entre l'unité cible et les unités candidates, par exemple en comptant combien de descripteurs sont différents (*independent feature formulation)* ; et 2) la similarité

acoustique entre l'unité cible et les unités candidates en estimant les propriétés acoustiques de la cible (inconnues car *a priori* pas de forme d'onde) et en calculant une distance entre ces différentes propriétés (*acoustic space formulation*). Cette similarité acoustique est réalisée dans des systèmes de synthèse hybride qui utilisent à la fois la concaténation et la modélisation paramétrique [Yan et al., 2010], [Merritt et al., 2016].

Le coût de jonction quantifie la qualité de la concaténation entre deux unités candidates consécutives dans une séquence possible. Il estime la perception d'un artefact au niveau acoustique à partir d'une distance entre les coefficients cepstraux (éventuellement énergie et F_0) et leur évolution dynamique (Δ) des trames avant et après le point de jonction. L'ajout de règles issues d'études perceptives autorise la pondération des artefacts en fonction des caractéristiques de l'unité : certaines différences acoustiques se perçoivent plus sur des liquides ou des voyelles que sur des consonnes non-voisées.

Pour une séquence temporelle d'unités cibles $\mathbf{u} = [\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_n]$ où $\mathbf{u}_i$ est un élément contenant le diphone et son contexte, de dimension D, et un ensemble d'unités candidates $\mathbf{t}$, le coût cible est donnée par l'équation 5.1 où ω_j est le poids donné à la caractéristique j. Le coût de jonction est défini par l'équation 5.2 où d est une distance entre les paramètres acoustiques. u_0 et u_n seront généralement des silences. Chercher le meilleur chemin $\mathcal{U}^*$, c'est minimiser le coût total sur l'ensemble des unités candidates de la base $\mathcal{U}$ (eq. 5.3).

$$C^t(t_i, u_i) = \sum_{j=1}^{D} \omega_j C_j^t(t_{ij}, u_{ij}) \tag{5.1}$$

$$C^c(u_i, u_{i+1}) = d(u_i, u_{i+1}) \tag{5.2}$$

$$\mathcal{U}^* = \underset{u_i \in \mathcal{U}}{\operatorname{argmin}} \left[\sum_{i=0}^{n} C^t(t_i, u_i) + \sum_{i=0}^{n-1} C^c(u_i, u_{i+1}) \right] \tag{5.3}$$

La synthèse par concaténation est souvent décriée car peu intelligible. En effet, ce type d'approche permet d'avoir des signaux de bonne qualité puisqu'on utilise des signaux pré-enregistrés sans générer la forme d'onde, mais au détriment d'artefacts liés à la concaténation de ces signaux. Son manque de flexibilité est également pointé du doigt : si l'on souhaite changer la voix de synthèse, il faut ré-enregistrer l'ensemble de la base. De même, si on souhaite un style expressif qui n'est pas présent dans la base, il faudra ré-enregistrer des échantillons sachant qu'il faudra retrouver le locuteur ou la locutrice et que sa voix n'ait pas trop changé. Par contre, ce type d'approche possède des avantages importants : le contrôle de l'expressivité passe par des commandes prosodiques ou symboliques explicites qui sont intégrées dans les

coûts cible et de jonction. La synthèse par sélection et concaténation d'unités reste donc un bon compromis entre contrôle et qualité comme souligné dans [Lolive, 2017].

5.1.3 Synthèse neuronale

La synthèse paramétrique a toujours utilisé des approches similaires à celles développées pour la reconnaissance automatique de parole. L'objectif du système paramétrique est de générer les paramètres nécessaires à la reconstruction de la forme d'onde. Pour cela, on utilisera un vocodeur qui prendra en entrée des paramètres temporels, typiquement l'enveloppe spectrale, la F_0 et un descripteur d'apériodicité (pour les sons de type "bruit"), et fournira en sortie le signal de parole correspondant. Ces paramètres doivent être prédits à partir d'une entrée de phonèmes (éventuellement enrichis d'éléments contextuels). Comme le temps des phonèmes n'est pas le même que celui du signal de parole, le système TTS paramétrique doit donc prédire les bonnes valeurs des paramètres mais également leur durée.

Le système de synthèse HTS [Zen et al., 2007] est basé sur une modélisation à l'aide de chaînes de Markov cachées. Ce système consiste en deux modèles : le modèle de séquence est un transducteur pondéré à états finis, le modèle d'observation est un mélange de Gaussiennes qui modélise les distributions de chaque état. Un phonème, ou un diphone, est modélisé par 5 états, chacun de ces états sera alors représenté par un modèle de type GMM. L'enveloppe spectrale (cepstre de mel), la F_0 et l'apériodicité sont alors modélisées par des GMM. La durée d'un phonème en contexte est modélisée par le HMM qui devient ainsi un modèle de durée.

Aujourd'hui l'approche HMM-GMM a été remplacée par des architectures neuronales. Le problème TTS peut être vu comme une régression entre deux séquences : une séquence de phonème (ou éventuellement de mots) vers une séquences de représentations acoustiques. Un réseau de neurone est donc un modèle adapté pour réaliser cette régression.

L'architecture neuronale Tacotron développée par Google [Shen et al., 2018] est à la base de la plupart des modèles récents. Cette architecture auto-régressive est construite à partir de réseaux récurrents de type LSTM qui sont adaptés pour la modélisation de séquences temporelles. Une schématisation du système complet est donné sur la figure 5.2. L'encodeur transforme l'entrée linguistique (textuelle ou phonétique) en une représentation latente (paramétrique) qui contient l'information nécessaire pour reconstruire le signal. Un mécanisme d'attention fait office de modèle de durée et va dilater la représentation linguistique pour l'adapter au temps du signal. Le décodeur prend en entrée cette représentation temporelle pour la convertir en un spectogramme (ici un mel-spectrogramme). Cette représentation temps-fréquence est ensuite fournie en entrée d'un vocodeur (WavNet [Oord et al., 2016] est

utilisé dans cette publication).

La fonction de coût est une MSE (*mean squared error*) calculée entre la représentation temps-fréquence prédite et celle de référence. Une des difficultés de ce type d'approche est de prédire la fin du signal. Pour cela, la sortie du mécanisme d'attention est concaténée avec la sortie du LSTM du décodeur et projeté sous forme de scalaire afin de retourner la "probabilité" que la séquence soit complète.

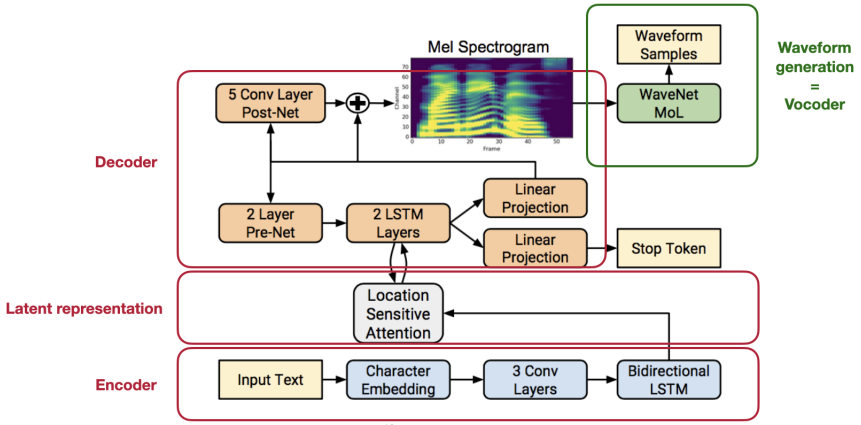


FIGURE 5.2 – Schéma de l'architecture du système de TTS Tacotron. Figure adaptée de [Shen et al., 2018].

Cette première version d'un système TTS neuronal ne permet de synthétiser qu'une seule voix, des adaptations multi-locuteurs ont été proposées en ajoutant en plus de la représentation textuelle une représentation du locuteur [Ping et al., 2018b]. Ces systèmes ont atteints des niveaux de qualité impressionnants, moyennant quelques difficultés pour une application réelle. Plus précisément, le système de critère d'arrêt ne fonctionne pas toujours, produisant des signaux qui "bouclent". Lorsque le texte d'entrée contient des mots répétés, les systèmes auto-régressifs produisent de sérieuses erreurs d'attention. De plus, lors de l'inférence, le temps de génération d'une phrase longue peut devenir prohibitif. Des modèles utilisant des parallélisations améliorent certaines de ces limitations comme ClariNet [Ping et al., 2018a] ou FastSpeech [Ren et al., 2019]. D'autres types d'architectures auto-régressives ont été utilisées, comme les transformers dans FastSpeech qui utilisent des alignements externes, ou internes Glow-TTS [Kim et al., 2020] un modèle génératif de type *flow-based*.

Au delà de cette profusion de systèmes, l'intérêt des approches neuronales est de pouvoir fournir des représentations externes au réseau afin de l'adapter à un locuteur, un style, un type de prosodie, etc. C'est cet aspect qui m'intéresse particulièrement et qui sera présenté dans la

section 5.3 de ce chapitre.

5.1.4 Les vocodeurs

Les systèmes TTS paramétriques ne sont pas conçus pour générer directement une forme d'onde, à quelques exceptions près comme WavNet [Oord et al., 2016]. En effet, ces systèmes prédisent des paramètres qui sont ensuite fournis en entrée d'un vocodeur.

Les vocodeurs tels que Straight [Kawahara et al., 1999] ou World [Morise et al., 2016] prennent en entrée les paramètres explicites du modèle source-filtre : F_0, enveloppe spectrale et apériodicité comme illustré sur la figure 5.3.

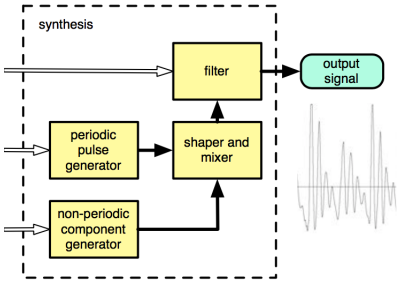


FIGURE 5.3 – Schéma de principe d'un vocodeur basé sur un modèle source filtre.

Cependant, la plupart des systèmes TTS neuronaux prédisent une représentation temps-fréquence de type mel-spectrogramme, en ignorant totalement de prédire la phase du signal final. Les vocodeurs fonctionnant avec une représentation explicite du modèle source-filtre ne sont plus adaptés. Une manière très efficace et rapide, moyennant une qualité imparfaite, est de faire une transformée de Fourrier inverse en estimant la phase à partir de l'amplitude. C'est basiquement le principe de l'algorithme de Griffin-Lim [Griffin and Lim, 1984] et son amélioration permettant d'interpoler la phase à partir du spectre d'amplitude [Perraudin et al., 2013]. Même si la phase du signal audio est généralement mise de côté car difficile à traiter, elle n'en reste pas moins un élément fondamental pour la génération de la forme d'onde et a donc une influence importante sur la qualité d'un signal [Koutsogiannaki et al., 2014].

Les vocodeurs neuronaux réalisent cette même opération de transformée de Fourier inverse en apprenant la procédure à partir de données. Pour cela, il faut un réseau capable de prédire une séquence d'amplitude à partir d'une représentation temps-fréquence.

WavNet [Oord et al., 2016], est une méthode auto-régressive qui prédit une forme d'onde

à partir d'elle-même tout en conditionnant la valeur de l'amplitude sur le spectrogramme. Le calcul de la "probabilité" conditionnelle de l'amplitude x_t à l'instant t sachant la représentation spectrale de la trame courante h et les valeurs des échantillons aux instants précédents x_{t-i} se fait suivant une approche largement utilisée en traitement automatique du langage (eq. 5.4). Pour cela WavNet utilise une architecture convolutionnelle, en incluant des dilatations entre les différentes couches afin de capturer des dépendances temporelles de durées variées.

$$p(x|h) = \prod_{t=1}^{T} p(x_t|x_1, \ldots, x_{t-1}, h) \tag{5.4}$$

En pratique, le conditionnement sur le spectre se fait grâce à une "porte" (Gated Activation Unit, GAU), insérée après chaque filtre convolutionnel. Une fonction de rééchantillonnage est préalablement appliquée au spectre h : $y = f(h)$ afin de recaler les trames sur la fréquence d'échantillonnage du signal en sortie. Le conditionnement sur le spectre est donné par l'équation 5.5 où k est l'indice de la couche, f, l'indice du filtre de convolution, g l'indice de la porte, W est le filtre convolutionnel et V une projection linéaire. Ainsi on adapte les poids des filtres à apprendre W et V en fonction du spectre h.

$$z = \tanh\left(W_{f,k} * x + V_{f,k} * y\right) \times \sigma\left(W_{g,k} * x + V_{g,k} * y\right) \tag{5.5}$$

WavGAN [Donahue et al., 2019] et WavGlow [Prenger et al., 2019] sont des variantes de ce vocodeur qui utilisent respectivement les techniques des *Generative Adversarial Networks* (GAN) (incluant un générateur et un discriminateur) et les approches *flow-based*. WavGlow génère la forme d'onde à partir d'une distribution apprise sur des variables latentes à partir de mélanges de Gaussiennes. L'avantage de représenter les variables latentes par des distributions statistiques est que celles-ci sauront prédire des échantillons cohérents même si on s'éloigne des données d'apprentissage. Ce type d'approche permet de mieux généraliser les modèles et par conséquent limite la dépendance du vocodeur à la (ou les) voix du locuteur (des locuteurs) sur lequel(s) il est entraîné.

Le modèle Hi-Fi GAN [Kong et al., 2020] repose sur le principe de la décomposition en série de Fourier. Un discriminateur va recevoir en parallèle plusieurs périodes spécifiques du signal audio, l'autre va chercher à discriminer les patterns sur différentes échelles temporelles. Ce vocodeur a permis de générer des signaux de parole de très bonne qualité.

Généralement ces vocodeurs sont évalués perceptivement en termes de qualité absolue (*Mean Opinion Score* - MOS entre 0 et 5) pour des apprentissages sur LJSpeech (un unique locuteur anglais) [Ito, 2017] ou VCTK (109 locuteurs anglais) [Yamagishi et al., 2012]. Ce-

pendant, il est possible que les modèles capturent une partie de la voix des locuteurs utilisés pour l'entrainement. Effectivement, lorsqu'on cherche à généraliser sur des locuteurs inconnus, les scores de qualités sont dégradés. Dans HiFi-GAN, le score MOS sur des locuteurs connus est de 4.36 (± 0.07) et tombe à 3.77 sur des locuteurs inconnus. Et cette expérience ne nous dit pas quel locuteur est reconnu.

5.1.5 Évaluation des systèmes de synthèse

La synthèse de parole est évaluée à la fois sur des critères "objectifs", et perceptifs. Le descriptif des méthodes d'évaluation pour la synthèse détaillées ci-dessous s'inspire de [Gaudier, 2020].

Les critères objectifs Des métriques sont extraites soit directement à partir des paramètres générés par un système de TTS paramétrique (enveloppe spectrale, F_0, ou représentation temps-fréquence suivant le modèle utilisé) soit sur la forme d'onde directement.

— Distance mel-cepstrale (*mel-cepstral distortion MCD)* : distance euclidienne entre deux représentations cepstrales alignées temporellement,

— Distance sur les distributions de probabilité : log-vraisemblance entre deux distributions de probabilité correspondant aux représentations acoustiques de deux signaux,

— Taux d'erreurs : calcul de taux d'erreur en phonème ou en mots entre le signal de référence et le signal à évaluer. Cela nécessite d'avoir un système de transcription. Et il faudra prendre en compte les erreurs liées au système de transcription lui-même.

Les critères subjectifs Les tests perceptifs sont utilisés dans plusieurs domaines de la linguistique à la synthèse de parole. Ces tests sont généralement couteux en temps de mise en place. Un nombre minimum de participants est nécessaire et il faut trouver un compromis entre ce nombre et la longueur du test [1].

— *Mean Opinion Score* (MOS) est une mesure de la qualité absolue d'un signal de parole. L'échelle est généralement compris entre 1 (mauvais) à 5 (excellent). L'inconvénient de cette évaluation est qu'elle ne permet pas de comparer deux systèmes entre eux. Et il est également difficile de comparer deux évaluateurs entre eux car les participants adaptent leur jugement et l'échelle qu'ils utilisent à la qualité globale des signaux évalués.

1. Voir pour cela la page de G. Degottex : `http://gillesdegottex.eu/?page_id=38`

— *Multi Stimulus with Hidden Reference and Anchors* (MUSHRA) mesure la qualité de plusieurs signaux. Les participants doivent évaluer un ou plusieurs signaux par rapport à un signal de référence. La note donnée est comprise entre 0 et 100.

— Test de préférence : c'est de loin le test le plus simple à réaliser. Le test AB mesure la qualité relative d'un signal A par rapport à B et l'évaluateur vote pour son signal préféré. Pour un test ABX, les participants doivent choisir parmi les signaux A et B, lequel est le plus similaire à la référence X.

— Test d'intelligibilité : les participants doivent retranscrire ce qu'ils ont entendu. Cette transcription est ensuite comparée au texte original avant synthèse. Pour éviter que les participants ne s'aident du contexte, les phrases à évaluer n'ont pas de sens (*Semantically Unpredictable Sentence* SUS).

Le challenge Blizzard organisé chaque année par le groupe d'intérêt SynSig [2] depuis 2005, a permis la mise en place de standards d'évaluation suivant les différentes tâches proposées chaque année. Notamment en 2021 [Zhou et al., 2020], les tests perceptifs cherchent à évaluer la similarité, le naturel, l'intelligibilité, et l'acceptabilité (par exemple de phrases en espagnol incluant des mots d'anglais).

5.2 Modélisation de la prononciation expressive

Très peu d'études ont cherché à analyser les relations entre l'expression d'une émotion et la prononciation. En effet, les études sur la prononciation se concentrent principalement sur l'analyse des *disfluences* qui dé-structure le discours par rapport à une "normalité". Les travaux présentés dans cette section ont été conduits dans le cadre de mon post-doctorat à l'IRISA au sein du projet ANR SynPaFlex.

5.2.1 Méthodologie générale

Les systèmes de synthèse à partir du texte peuvent prendre en entrée une séquence de phonèmes correspondant à la prononciation de ce texte. L'approche classique consiste à utiliser un dictionnaire de prononciation qui contient les séquences de phonèmes associés aux mots, incluant éventuellement des variantes et les liaisons. En complément, un ensemble de règles donne la prononciation des mots hors vocabulaire. Ces règles sont soit construites manuellement, soit apprises automatiquement à l'aide de modèles statistiques sur

2. `https://www.synsig.org/index.php/Blizzard_Challenge`

un corpus de parole (convertisseur *grapheme-to-phoneme* G2P) comme c'est le cas en ASR [Karanasou et al., 2013]. L'avantage des approches statistiques est qu'elles peuvent prendre en entrée un contexte plus large que le mot lui-même et ainsi désambigüiser la prononciation. Par exemple, dans [Lecorvé and Lolive, 2015], un modèle G2P basé sur des CRFs (*Conditional Random Fields*) et des transducteurs à états finis (*Weighted Finite State Transducers* - WFST) prédit la séquence de phonème d'une phrase entière en entrée. Je ne rentrerai pas dans le détail de ces modèles dans ce document, mais nous utilisons ces systèmes afin d'obtenir une représentation phonétique du texte. Il faut donc avoir conscience que cette séquence de phonèmes, appelée *canonique*, sera dépendante des règles utilisées pour la générer.

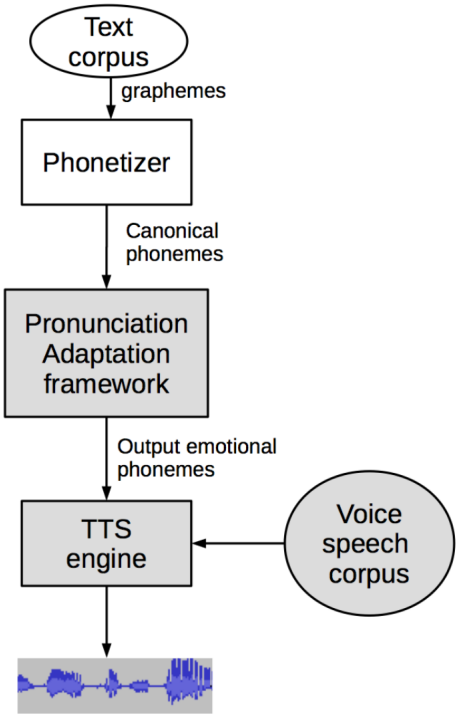


FIGURE 5.4 – Schéma pour l'adaptation de la prononciation à la voix de synthèse.

Cependant, il est souhaitable que la prononciation fournie en entrée du système de synthèse soit adaptée au style recherché et à la voix utilisée pour la synthèse elle-même. En effet, nous savons que la prononciation est influencée par le type d'interaction, le style de parole, l'expressivité ou l'origine du locuteur. Il y a donc des différences entre la prononciation de la voix de synthèse et celle des locuteurs utilisés pour apprendre les règles ou modèles générant la séquence de phonème à partir du texte. La méthodologie employée repose sur les travaux

de thèse de Raheel Qader, notamment ceux présentés dans [Qader et al., 2015].

Les CRFs Les CRFs sont très répandus pour le traitement de séquences d'événements discrets [Camelin et al., 2010], la phonétisation de mots [Lecorvé and Lolive, 2015], ou d'énoncés [Karanasou et al., 2013]. Cette modélisation est particulièrement adaptée pour l'apprentissage de données séquentielles discrètes et permet d'intégrer et de combiner simplement plusieurs caractéristiques. Les CRFs permettent de modéliser plusieurs types de dépendances [Lafferty et al., 2001] comme illustré sur la figure 5.5.

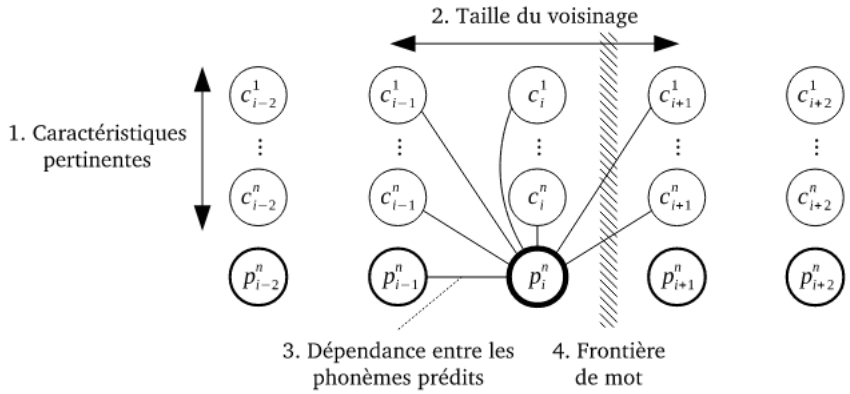


FIGURE 5.5 – Dépendances et paramètres à traiter pour l'apprentissage des CRFs. Figure extraite de [Lolive, 2017].

Soit une séquence temporelle de phonèmes à prédire $p = \{p_1, \ldots, p_N\}$ où chaque phonème p_i est représenté par une série de caractéristiques observables c_i. Les caractéristiques appartiennent à un ensemble fini de labels. Les variables aléatoires $\mathbf{P}$ et $\mathbf{C}$ sont dépendantes. L'objectif du modèle sera de déterminer $p(\mathbf{C}|\mathbf{P})$ sans modéliser la probabilité a priori $p(\mathbf{P})$. Pour cela le modèle doit chercher quelles caractéristiques parmi celles considérées dégradent ou sont pertinentes pour prédire la séquence p.

L'ensemble des caractéristiques est localisé sur le phonème courant p_i, mais peut être étendu sur son voisinage en considérant les caractéristiques des phonèmes voisins c_{i-2} ou c_{i+1} et en les incorporant aux caractéristiques c_i. Ce voisinage est définit par une fenêtre W centrée autour de p_i. Nous avons étudié l'impact de cette fenêtre sur l'adaptation. On peut également modéliser les dépendances entre deux phonèmes consécutifs, par exemple pour éviter les enchaînements impossibles entre deux phonèmes.

Données Les phonèmes *canoniques* sont générés avec l'outil LIA-Phon [Béchet, 2001], largement utilisé pour la phonétisation de phrases en français. On souhaite adapter cette prononciation *canonique* à celle de la voix de synthèse. Dans l'expérience ci-dessous, la voix de synthèse est générée à partir d'un corpus en Français de voix de femme qui prononce des phrases sur un ton neutre pour des interactions vocales humain-machine dans le domaine de la télécommunication. On appellera ce corpus *Telecom*. La prononciation réalisée par cette locutrice sera appelée prononciation *cible*. Le corpus a été divisé aléatoirement en deux : 70% pour l'apprentissage et 30% pour la validation.

Modèles Les séquences de phonèmes sont modélisées par des CRFs entraînés à prédire une séquence la plus proche possible de la *cible* à partir de la séquence *canonique*. Les séquences de phonèmes prédites sont comparées à la *cible* avec une mesure d'erreur sur les phonèmes (*Phoneme Error Rate* - PER) définie par l'équation 5.6 où N est le nombre de phonèmes dans la séquence de référence. Une sélection des descripteurs les plus pertinents est réalisée en cross-validation sur l'ensemble d'apprentissage avec une approche de type *forward* décrite dans le paragraphe suivant.

$$PER = \frac{\text{substitutions} + \text{suppression} + \text{insertions}}{N} \tag{5.6}$$

5.2.2 Choix et impact des paramètres

A minima les CRFs utilisent les phonèmes *canoniques*, et nous avons enrichi cette représentation à l'aide d'un ensemble de 52 caractéristiques linguistiques, phonologiques, articulatoires et prosodiques décrites dans le tableau 5.1. Les fréquences de mots dans le français ont été extraites des Google ngrams [Lin et al., 2012]. Les descripteurs prosodiques sont basés sur l'énergie, la F_0, la durée et le débit syllabique. Le contour de F_0 est déterminé à partir d'une valeur de glissando définie dans [t'Hart, 1981]. Il faut noter ici que les descripteurs prosodiques tels que la F_0 ont été extraits directement de la référence audio. En effet, dans cette étude, nous ne disposons pas de prédicteur de prosodie à partir du texte. Si la prosodie semble pertinente, il faudra poursuivre dans cette direction.

Un processus de sélection des meilleurs descripteurs est mis en place par groupe (linguistique, phonologique, articulatoire et prosodique) suivant un protocole de cross-validation à 7 plis. Puis les groupes sont combinés afin de trouver la meilleure configuration. L'approche de sélection *forward* commence avec les phonèmes *canoniques* uniquement, les descripteurs sont ajoutés un à un jusqu'à que l'ensemble optimal soit atteint suivant un critère d'arrêt défini

Caractéristiques linguistiques (18)
Word [7] ♦ **Stem** [7] ♦ Lemma [0] ♦ POS [2] ♦ Stop word [0] ♦ Word [0], stem [2], lemma [1] freq. in French (common, normal, rare) ♦ Word [1], stem [1], lemma [2] freq. in corpus ♦ Word freq. knowing previous word in French [2], in corpus [1] ♦ Word freq. knowing next word in French [2] in corpus [3] ♦ Number of word occurence in corpus [0] (numerical) ♦ Word position [3], reverse position [0] in utterance (numerical)
Caractéristiques phonologiques (17)
Canonical syllables [7] ♦ Phoneme in syllable position [0] ♦ Phoneme in word position [0] (begin, middle, end) ♦ **Syllable in word position** [6] ♦ Phoneme position [0] and **reverse position** [4] in syllable (numerical) ♦ **Phoneme position** [5] and **reverse position** [5] in word (numercial) ♦ Syllable position [3] and reverse position [1] in word (numercial) ♦ **Word length in phoneme** [4] (numerical) ♦ Word length in syllable [2] (numerical) ♦ Syllable short [1] and long [0] structure (CVC, CCVCC) ♦ Syllable type [1] (open, closed) ♦ Phoneme in syllable part [0] (onset, nucleus, coda) ♦ **Pause per Syllable** [4] (low, normal, high)
Caractéristiques articulatoires (9)
Phoneme type [2] (vowel, consonant) ♦ Phoneme aperture [3], shape [1], place [1] and manner [2] (open, close, front, central, undef, etc.) ♦ Phoneme is affricate [0], rounded [3], doubled [0] or voiced [3] ? (boolean)
Caractéristiques prosodiques (7)
Syllable Energy [7] (low, normal, high) ♦ **Syllable [4] and phoneme [7] tone** (from 1 to 5) ♦ F_0 **phoneme contour** [7] (decreasing, flat, increasing) ♦ **Speech rate** [7] (low, normal, high) ♦ Distance to next [3] and **previous pause** [7] (from 1 to 3)

TABLE 5.1 – Groupes de descripteurs utilisés pour modéliser la prononciation. Entre parenthèse, le nombre de vote obtenus sur les 7 plis, en gras, les descripteurs sélectionnés. Tableau extrait de [Tahon et al., 2016a].

par l'équation 5.7 où $\epsilon = 0.1$ est fixé empiriquement et n le nombre de descripteurs ajoutés. Cette sélection est réalisée sur chacun des 7 plis de la cross-validation, puis un vote majoritaire permet de classer les descripteurs. Seuls ceux qui reçoivent un nombre de votes supérieur ou égal à 4 (sur 7) sont inclus dans l'ensemble final.

$$PER(n+1) > PER(n) - \epsilon \tag{5.7}$$

Nous pouvons remarquer que dans le groupe linguistique, les mots et leur stem ont été sélectionnés malgré la redondance entre ces deux caractéristiques, alors que les fréquences de mots dans le corpus et dans la langue ont reçu peu de votes. Les descripteurs articulatoires ne semblent pas pertinents pour cette tâche, contrairement aux descripteurs phonologiques, qui sont eux, redondants avec les phonèmes *canoniques*. Les caractéristiques syllabiques n'ont

Baseline (no adaptation)		10.7 [0.0]
Canonical phoneme only (with adaptation)		6.6 [-4.1]
C + L + Ph	selected (9)	3.9 [-6.8]
C + L + Ph + Pr	selected (15)	3.3 [-7.2]

TABLE 5.2 – PER obtenu sur l'ensemble de validation suivant différentes configurations. Entre parenthèses le gain en points de pourcentage.

pas été sélectionnés, sans doute la syllabe n'est pas une échelle pertinente. Par contre, on remarque que l'ensemble des descripteurs prosodiques sont sélectionnés soulignant encore une fois l'importance de la prosodie sur la prononciation.

Résultats objectifs et perceptifs Trois modèles CRFs ont été appris sur trois groupes de descripteurs différents : phonèmes *canoniques* uniquement (C), phonèmes et descripteurs linguistiques et phonologiques sélectionnés (C+L+Ph) et phonèmes et descripteurs linguistiques, phonologiques et prosodiques (C+L+Ph+Pr). Les modèles ont été évalués de façon objective en terme de PER (tableau 5.2) et de façon subjective sur les signaux de synthèse générés à partir des prononciations adaptées (figure 5.6).

Les résultats objectifs montrent que la combinaison des trois groupes de descripteurs est bénéfique pour diminuer le PER par rapport à l'utilisation des phonèmes *canoniques* seuls. L'apport des descripteurs prosodiques est cependant peu important par rapport aux descripteurs phonologiques et linguistiques. En effet, la plupart des confusions entre la séquence *canonique* et la séquence *cible* concernent des allophones : $\mathrm{o} \leftrightharpoons \mathrm{ɔ}$, $\mathrm{e} \leftrightharpoons \mathrm{ɛ}$ et $\tilde{\mathrm{ɛ}} \leftrightharpoons \tilde{\mathrm{œ}}$. Ces confusions ne peuvent pas vraiment être considérées comme des erreurs en français, mais plutôt comme une modification du style de parole. D'autres substitutions existent à cause de stratégies d'annotation différentes entre LIA-Phon et l'annotation du corpus *Telecom* comme par exemple $\mathrm{ɲ} \leftrightharpoons \mathrm{nj}$, $\mathrm{ə} \leftrightharpoons \mathrm{ø}$. Parmi les insertions, on peut noter le cas du shwa $/\mathrm{ə}/$ qui est connu pour être parfois omis lors de l'annotation et/ou de la réalisation. Les suppressions concernent principalement les liaisons entre les mots, comme par exemple les $/\mathrm{t}, \mathrm{z}/$ qui ne sont pas systématiquement générées par le phonétiseur mais souvent prononcées par la locutrice du corpus *Telecom*.

Forts de ces résultats, nous espérons une amélioration de la qualité de la synthèse générée à partir de cette prononciation adaptée. Pour évaluer cette qualité, nous avons sélectionné 40 phrases telles que l'échantillonnage respecte la distribution des PER entre les séquences *canonique* et *cible*. 14 participants de langue française devaient répondre à la question :

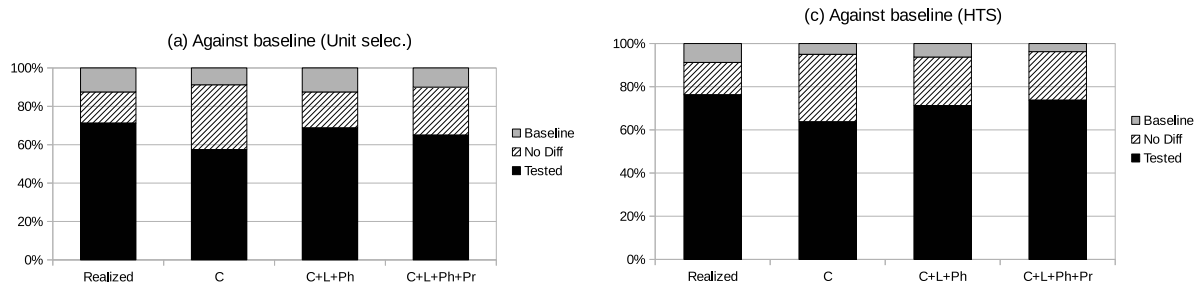


FIGURE 5.6 – Résultats des tests perceptifs de comparaison AB avec un système de sélection d'unité (a) et HTS (b). Avec un modèle de prononciation incluant des caractéristiques linguistiques et phonologiques, le système adapté (noir) est préféré au système sans adaptation (gris).

Entre A et B, quel exemple est de meilleure qualité ?. Deux systèmes de synthèse ont été utilisés : le système par concaténation développé par l'équipe [Guennec and Lolive, 2014] et le système paramétrique HTS-v2 [Zen et al., 2007] afin que nos conclusions ne dépendent pas des artefacts d'une approche ou d'une autre. Le système baseline est obtenu à partir des phonèmes *canoniques* directement sans phase d'adaptation.

Les résultats perceptifs montrent clairement que les adaptations proposées sont jugées meilleures que la baseline quelque soit le système de synthèse utilisé. L'ajout de descripteurs linguistiques et phonologiques augmente le nombre de systèmes adaptés préférés. Par contre, l'ajout d'informations prosodiques ne semble pas améliorer significativement la qualité de la synthèse. Cela nous arrange, car la problématique de prédiction de la prosodie à partir du texte est très complexe.

Ainsi, nous avons proposé une méthodologie pour adapter la prononciation issue d'un phonétiseur à celle réalisée par la voix utilisée pour la génération des signaux de synthèse. Nous avons montré que cette adaptation permettait à la fois de corriger des confusions liés à des stratégies d'annotation différentes, mais également d'adapter la séquence de phonème à la prononciation d'un locuteur en particulier.

Importance du contexte Dans l'étude décrite ci-dessus, le contexte phonétique utilisé pour la modélisation de la prononciation par le CRF était le phonème courant lui-même. Une étude complémentaire [Tahon et al., 2016b] a montré qu'un contexte allant de -2 à +2 phonèmes permettait d'obtenir de meilleurs résultats tant en terme de réduction du PER, qu'en terme de qualité de synthèse perçue. Nous avons également estimé l'impact de la quantité de donnée

nécessaire à l'adaptation. À partir de 5 min de parole l'addition de nouvelles données à un coût important pour un gain en terme de PER relativement faible : un facteur 10 sur la durée d'apprentissage améliore le PER de 0.5 points de pourcentage. Idéalement, un $PER = 0$ pourrait être atteint avec $3 \cdot 10^8$ h de données.

5.2.3 Adaptation à une voix expressive

À partir de l'approche mise en place pour adapter la prononciation, nous avons étudié la possibilité d'adapter la séquence de phonème à un style expressif particulier. Ici le style expressif sera explicité par six catégories émotionnelles. Le principe général est illustré figure 5.7 et se décompose en deux protocoles distincts.

1. Vo+Exp : les phonèmes *canoniques* sont adaptés une première fois à la voix de synthèse (phonèmes *Vo*), puis une deuxième fois à une émotion (phonèmes *Vo+Exp*)
2. Exp : les phonèmes *canoniques* sont directement adaptés à une émotion (phonèmes *Exp*)

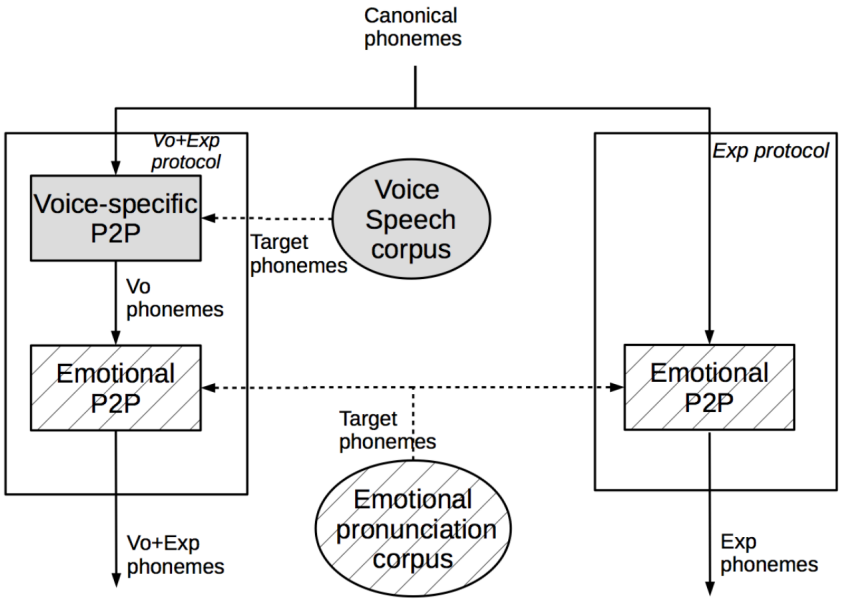


FIGURE 5.7 – Schéma pour l'adaptation de la prononciation à la voix de synthèse.

Données La voix de synthèse sera toujours générée à partir du corpus *Telecom*. Les prononciations expressives seront modélisées à partir d'un corpus d'émotions actées, noté *EmoPron* [Bartkova et al., 2016]. 50 phrases émotionnellement marquées (par exemple *Mais t'es con ou quoi ?*) ont été jouées pour chacune des 6 émotions (colère, peur, tristesse, dégoût, surprise et joie) ainsi que sur un ton neutre. Dans notre protocole, il est important de pouvoir comparer les modifications de prononciations liées à l'émotion sur des phrases identiques. Ce corpus de parole émotionnel a été transcrit en phonèmes automatiquement avec une vérification manuelle. Le corpus contient environ 8 min de données par catégorie d'émotion, ce qui nous semble suffisant au regard des 5 min nécessaires pour atteindre une adaptation satisfaisante [Tahon et al., 2016b]. Une limitation de ces données est que le contenu linguistique est déjà coloré par une émotion, et le fait d'acter une autre émotion sur ce type de phrase induit des situations peu réalistes.

Une prononciation expressive ? Il n'est pas évident d'affirmer qu'il existe une prononciation particulièrement liée à une catégorie émotionnelle. Nous nous sommes donc intéressés à deux questions : est-ce qu'il y a une prononciation expressive différente de la prononciation neutre pour un style de parole spontanée ? et est-ce qu'il y a des différences de prononciation entre les différentes catégories émotionnelles ? Pour cela, nous mesurons les différences entre les séquences de phonèmes correspondant aux phrases prononcées sur un ton neutre, et les phrases prononcées avec un style émotionnel avec la métrique PER. Ainsi, il y a bien des différences plus ou moins grandes suivant les émotions, avec un PER moyen de 6.4% entre le style neutre et les styles d'expression de toutes les émotions confondues. Nous notons également plusieurs confusions entre les styles neutres et émotionnels :

- Suppression du schwa canonique /ə/,
- Ré-insertion du /ə/ : plus de schwas sont prononcés (notamment pour la colère),
- Substitution du canonique /ʒ/ par /ʃ/ : dévoisement lié à une assimilation, phénomène important en parole spontanée, par exemple pour *je ne sais pas* est typiquement exprimé dévoisé /ʃ s ɛ p a/ avec une suppression du schwa et de la négation,
- Suppression des liquides /ʁ/ et /l/ : suppression du /l/ dans *il y a* qui devient alors /i y a/ phénomène reporté par [Brognaux et al., 2014].

La forte part des suppressions par rapport aux substitutions s'exprime également dans une augmentation du débit de parole pour la parole expressive (5.7 syllabes/sec) par rapport à la parole neutre (5.0 syllabes/sec).

Expressive utterance (surprise) :	
Canonical	k i p ø b j ɛ̃ m a v w a ʁ l ɛ s e s ø m e s a ʒ ə ʒ ø n ø v w a v ʁ ɛ m ɑ̃ p a
Realized	k i p o b j ɛ̃ m a v w a ʁ l e s e s - m e s a ʒ ə ʒ - v w a v ʁ e m ɑ̃ p a
Vo (Ling+Phon)	k i p ø b j ɛ̃ m a v w a ʁ l e s e s ø m e s a ʒ - ʒ ø n ø v w a v ʁ ɛ m ɑ̃ p a
Vo (Ling+Phon+Pros)	k i p ø b j ɛ̃ m a v w a ʁ l e s e s - m e s a ʒ ə ʒ - v w a v ʁ ɛ m ɑ̃ p a
Vo+Exp (Ling+Phon)	k i p ə b j ɛ̃ m a v w a ʁ l e s e s - m e s a ɛ - ʒ - v w a v ʁ e m ɑ̃ p a
Vo+Exp (Ling+Phon+Pros)	k i p ə b j ɛ̃ m a v w a ʁ l e s e s - m e s a ʒ ə ʒ - v w a v ʁ ɛ m ɑ̃ p a
Exp (Ling+Phon+Pros)	k i p ə b j ɛ̃ m a v w a ʁ l ɛ s e s - m e s a ʃ ə ʒ - v w a v ʁ e m ɑ̃ p a

TABLE 5.3 – Prononciation expressive de la phrase surprise *“Qui peut bien m’avoir laissé ce message ? Je ne vois vraiment pas.”*. En gris les changements par rapport à la séquence canonique.

Adaptation de la prononciation expressive Le tableau 5.3 présente un exemple d’adaptations de la prononciation obtenue avec les trois modèles Vo, Vo+Exp et Exp. On remarque que la transformation du /ʒ/ en /ʃ/ représentative de la prononciation spontanée n’est modélisé que par le système Exp.

Nous avons également proposé d’utiliser la similarité cosinus pour évaluer la distance entre les prononciations expressives et neutres générées par les différents modèles. Pour cela, le modèle TF-IDF a été adapté en remplaçant les termes par les confusions de phonèmes (substitutions, insertions et suppressions) obtenues entre la prononciation canonique p_c et une prononciation adaptée p_e suivant une émotion particulière e. On comptera par exemple combien de fois la confusion /e/ $\rightarrow$ /ɛ/ apparaît entre p_c et p_e. La paire (p_c, p_e) est représentée par un vecteur contenant les TF-IDF calculées sur l’ensemble des confusions présentes dans les données. La similarité cosinus est donnée par l’équation 5.8 où $\overrightarrow{C}_{p_c,p_e}$ est le vecteur de confusion entre p_c et p_e. Les valeurs de similarité par paires sont représentées sur la figure 5.8 uniquement pour le protocole Exp.

$$\cos\theta_{1,2} = \frac{\overrightarrow{C}_{p_c,p_1} \cdot \overrightarrow{C}_{p_c,p_2}}{\left\|\overrightarrow{C}_{p_c,p_1}\right\| \cdot \left\|\overrightarrow{C}_{p_c,p_2}\right\|} \tag{5.8}$$

Nous montrons que globalement les similarités obtenus entre deux séquences de même émotions sont comprises entre 0.93 (tristesse) et 0.98 (colère et peur). Nous montrons également que tristesse et dégoût sont très différents (0.66) alors que peur et colère sont très proches (0.97). De cette étude, nous concluons donc qu’il existe une prononciation expressive qui se démarque d’un ton neutre, en contexte de parole spontanée. Il existe aussi des différences de prononciation entre les émotions qui dépendent des catégories elle-mêmes, mais ce résultat reste à confirmer sur un corpus de parole émotionnelle plus diversifié et moins acté.

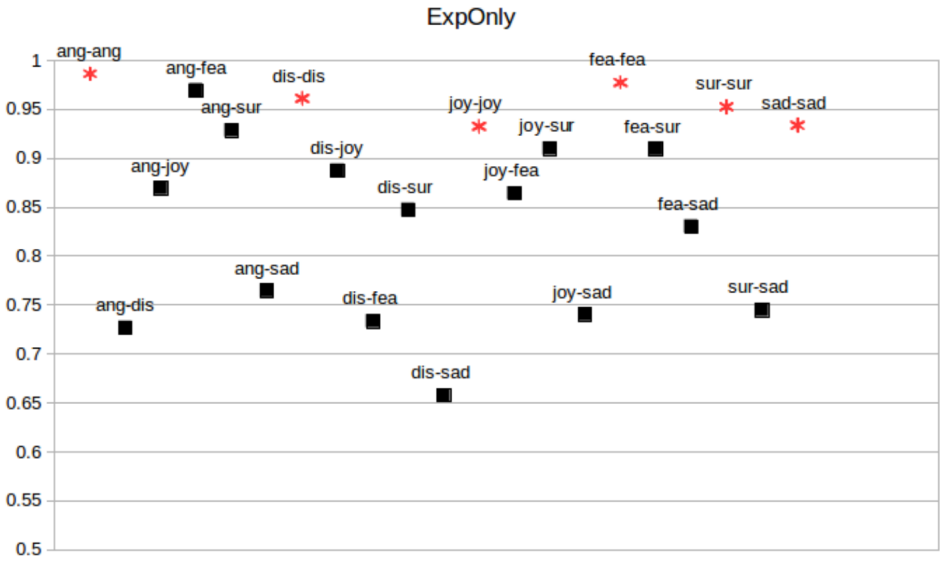


FIGURE 5.8 – Similarité cosinus entre les catégories émotionnelles e. La similarité est calculée sur la base des confusions entre deux paires (p_c, p_e) avec p_c la séquence de phonèmes canoniques.

Résultats perceptifs Pour finir, il reste à évaluer la qualité sonore et expressive des signaux de parole générés à partir de la prononciation expressive prédite. Pour ce test perceptif, nous avons demandé à 11 participants d'évaluer 60 phrases dont le contenu linguistique était neutre ou émotionnellement coloré, sur des tests comparatifs AB. Les questions posées étaient les suivantes :

— *Entre A et B, quel exemple est de meilleur qualité ?*

— *Entre A et B, quel exemple est le plus expressif ?*

— *Pour l'exemple le plus expressif, quelle émotion percevez-vous ?*

Les exemples synthétisés à partir d'une prononciation adaptée à la voix de synthèse uniquement (protocole Vo) a été préférée en terme de qualité. Les qualités pour les protocoles Exp et Vo+Exp ont été jugées similaires, avec une préférence pour Vo+Exp dans le cas d'un contenu linguistique neutre. Cela confirme l'intérêt de la double adaptation, et principalement de l'adaptation à la voix de synthèse pour la qualité.

Les participants ont jugé les exemples générés à partir d'un texte émotionnellement coloré avec l'adaptation Vo+Exp plus expressifs que sans adaptation (43% contre 27% avec les phonèmes *canoniques*) ou une simple adaptation à la voix (41% contre 24% avec Vo seul). Nous

confirmons donc la capacité de la double adaptation à générer une prononciation plus expressive qu'une simple adaptation, à condition que le contenu linguistique soit également expressif. Si les participants ont bien été capables de retrouver des marques expressives dans les exemples synthétisés, il n'a pas été possible pour eux de déterminer la catégorie émotionnelle. Cela confirme le fait que les similarités de prononciation entre les différentes catégories émotionnelles sont très fortes, sauf dans certains cas. Pour le lecteur intéressé, l'ensemble des résultats ainsi que des analyses complémentaires sont détaillées dans la publication [Tahon et al., 2018].

5.3 Génération d'une voix expressive

L'objectif de la synthèse de parole expressive est de générer des signaux de parole qui soit de bonne qualité et dont l'expressivité est adaptée au contexte de l'utilisation du système. Clairement, pour une application de type GPS ou commande vocale, l'expressivité n'aura pas d'intérêt en soi. On cherchera alors à générer une voix de bonne qualité, et la plus "naturelle" possible. Par contre, dans une application de type lecture de livres audio, nous chercherons à obtenir une voix qui change son style en fonction du type littéraire, qui modifie son timbre de voix suivant les différents personnages et qui adapte l'expression vocale à l'émotion que l'on souhaite faire ressentir à l'auditeur.

Nous avons vu dans le premier chapitre que l'expressivité se définit principalement par le contexte de l'interaction. Cela implique l'utilisation d'une prononciation adaptée, l'ajout d'emphase, de marqueurs sociaux et affectifs, ou encore l'expression d'une émotion. Dans le cas de la synthèse, se pose la question du contrôle de l'expressivité. Comment l'utilisateur peut-il choisir, modifier, contrôler l'expressivité qu'il souhaite ? Pour répondre à cette question, plusieurs pistes se dégagent :

1. L'expressivité est celle d'un (ou plusieurs) locuteur(s) dans un corpus de parole donné $\rightarrow$ définition *implicite*,
2. L'expressivité correspond à l'expression d'une émotion connue $\rightarrow$ définition *implicite*,
3. L'expressivité est définie par un ensemble de paramètres prosodiques, phonétiques $\rightarrow$ définition *explicite*.

5.3.1 Contrôle implicite de l'expressivité par les données

Dans le cadre de mon post-doctorat à l'IRISA, j'ai analysé les influences respectives de différents aspects de la chaîne de synthèse sur la perception de l'expressivité [Tahon et al., 2017].

Représentation et perception de l'expressivité apprise sur des corpus différents Pour cela, nous avons repris le protocole d'adaptation de la prononciation à celle de la voix de synthèse et à une émotion cible décrit dans la section précédente en le généralisant à plusieurs voix de synthèse. Les trois caractéristiques que nous étudions sont les suivantes :

— L'expressivité de la voix liée au locuteur et au contexte induit par le scénario de collecte. Pour cela, nous avons utilisé 3 corpus de parole mono-locuteur : *Telecom* (interactions vocales, voix neutre), *Audiobook* (lecture de livres audio, voix modérée) et *Commentary* (commentaires précédant des séries TV, voix expressive).

— L'expressivité liée à la prononciation. Nous avons utilisé le corpus émotionnel présenté précédemment *EmoPron*

— L'expressivité liée au contenu linguistique en entrée. Nous avons utilisé des phrases issues des corpus neutre *Telecom* et émotionnel *EmoPron*. Les phrases utilisées pour les tests ont été écartées des corpus d'apprentissage pour les modèles de prononciation (cross-validation) et la voix de synthèse neutre (apprentissage 70%).

Le choix des voix de synthèse a été motivé par les scénarios de collecte de ces données qui ont induit plus ou moins d'expressivité dans la voix. Nous avons validé ce choix par l'extraction de paramètres prosodiques. Le tableau 5.4 montre qu'effectivement les variations de fréquence fondamentale (en semiton pour normaliser par rapport au locuteur) et la moyenne du débit de parole sont plus élevés pour de la parole expressive que de la parole neutre.

					$F0$		SR	
Corpus	Expressivité	# utt.	Dur.	# phon.	moy	σ	moy	σ
Corpus de voix								
Telecom - train 70%	Neutral	5044	4h51'	151,945	89	2.7	4.7	2.1
Audiobook	Moderate	3339	10h45'	379,897	77	3.2	6.3	1.2
Commentary	Expressive	1631	5h25'	173,858	85	5.0	6.0	1.7
Corpus de pronunciation								
Expressive	Expressive	6×47	0h41'	16,248	84	7.1	6.3	1.8
Corpus de texte								
Telecom - eval 30%	Neutral	2162	2h04'	64,960				
Expressive	Expressive	6×47	0h41'	16,248				

TABLE 5.4 – Analyse des caractéristiques prosodiques en fonction du corpus. La F_0 est donnée en semiton, et le débit syllabique (SR) en nombre de syllabe par seconde.

À partir de ces données nous avons généré plusieurs signaux synthétiques en utilisant 1) les trois voix de synthèse, 2) l'adaptation de la prononciation à ces voix et à une émotion cible

parmi 6, et 3) un texte cohérent avec cette émotion ou neutre. Les signaux ainsi obtenus sont évalués objectivement avec la mesure du PER et subjectivement avec des tests de perception. Je ne détaille pas ici l'ensemble des résultats obtenus dans [Tahon et al., 2017]. Cependant, les deux évaluations objective et perceptive confirment l'intérêt de l'adaptation de la prononciation à la voix de synthèse en généralisant sur différents corpus. Les tests perceptifs montrent que l'expressivité est d'autant mieux perçue que les signaux synthétisés sont de bonne qualité. Alors que la perception de l'expressivité repose principalement sur l'adéquation entre la prononciation et la prosodie, la perception des émotions semble être plus liée au contenu linguistique. En ce sens, ces derniers résultats rejoignent les conclusions de la thèse de Manon Macary, où nous avons montré que la perception de la frustration semble plutôt liée au contenu linguistique qu'aux aspects acoustiques.

Représentation implicite de l'expressivité avec les réseaux de neurones Avec l'utilisation des réseaux de neurones, le choix du style expressif se fait généralement via une représentation latente du style encodé dans des *embeddings*. Ceux-ci sont ensuite fournis au système de TTS, de la même façon que les *embeddings* de texte et ceux du locuteur [Skerry-Ryan et al., 2018]. Ces *embeddings* sont appris en utilisant un modèle auto-génératif de type Tacotron conditionné sur le locuteur et le contenu phonétique. Cette approche permet de modéliser la prosodie de façon très fine, par contre un changement de contenu linguistique va générer une prosodie inadaptée. De plus, cette représentation latente étant apprise de façon non supervisée sur un corpus de livre audio, il est difficile d'interpréter les *embeddings* prosodiques générés par le système à partir d'une séquence de phonème.

Le modèle multi-locuteur GST proposé dans [Wang et al., 2018] repose sur des *embeddings* de style appris de façon non supervisée, l'avantage est de ne pas avoir à étiqueter en amont des catégories explicites de styles contenus dans le corpus. De plus cette approche apporte des catégories qui sont *a priori* plus interprétables que des *embeddings*. Ainsi le modèle identifiera de lui-même un certain nombre de styles prosodiques. Malgré tout l'avantage d'obtenir des groupes de styles de manière non supervisée, il ne semble pas aisé d'identifier à quel type d'expression appartient ces groupes. À partir d'un corpus de livres audio, les auteurs ont pu identifier deux styles : "débit de parole rapide" et "parole animée". Pour une revue détaillée des systèmes existants, le lecteur pourra se référer à [Kulkarni, 2022].

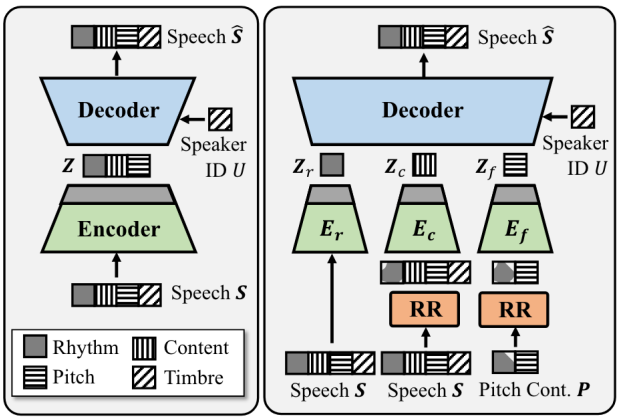
5.3.2 Contrôle explicite de l'expressivité

Le fait d'interpréter le contenu expressif de façon implicite, uniquement à partir des données utilisées pour l'apprentissage des modèles est une approche qui produit de la synthèse de qualité mais n'est pas interprétable. Dans un premier temps, il est intéressant de revenir au modèle source-filtre et de proposer des paramètres de type contour intonatif ($F_0(t)$), rythmique, accentuation, et des catégories expressives explicites telles que les émotions.

Le modèle Mellotron [Valle et al., 2020] repose sur l'architecture GST, en ajoutant un conditionnement sur des commandes rythmiques et intonatives explicites. Le rythme et l'intonation sont obtenus à partir d'un signal audio ou d'une partition musicale. Le rythme est encodé à partir d'un alignement forcé, tandis que la F_0 est extraite du signal audio ou de la partition. Une des difficultés est de modéliser les différentes caractéristiques du corpus de façon indépendante. Ainsi plusieurs travaux proposent des techniques pour démêler les représentations comme par exemple [Lu et al., 2021].

Dans l'approche [Qian et al., 2020], le signal de parole est décomposé suivant quatre composantes : contenu linguistique, timbre, intonation et rythme en s'inspirant des systèmes de conversion de voix. L'objectif du système SpeechFlow représenté sur la figure 5.9 est de réaliser une conversion de la prosodie d'un signal source en fournissant au système des commandes prosodiques au système. La différence avec le Mellotron est que les encodeurs sont entrainés avec des entrées multiples de façon à assurer une indépendance des paramètres. L'intonation sera modélisée par le contour normalisé de la F_0, qui contient également une information rythmique. Le timbre sera modélisé par les fréquences formantiques, et le contenu linguistique par la séquence de phonème. Par exemple, pour apprendre une représentation du contour indépendamment du rythme, chaque segment est étiré ou rétréci.

Le rythme est souvent contrôlé par une commande explicite de type débit de parole [Tånnander and Edlund, 2021]. Une des difficultés lorsqu'on ralentit le débit de parole, est qu'il ne s'agit pas uniquement d'allonger certains types de phonèmes. En effet, la parole lente contient typiquement plus de pauses silencieuses et remplies, comme dans le cas de la parole hésitante (voir section 4.1.3). L'utilisation du débit de parole comme unique commande rythmique ne permet pas de faire des transformations non linéaires. Ainsi, lorsque la conception de commandes explicites ne correspond pas à une application réelle, il est intéressant de laisser à l'utilisateur la possibilité de corriger lui-même la sortie du système. Ainsi, le système EditSpeech [Tan et al., 2021] propose de modifier le signal de parole généré automatiquement à l'aide de trois opérations : suppression d'une zone du signal qui correspond à des mots choisis, l'insertion et le remplacement de mots. Dans le cas du replacement, il s'agit de créer un nou-



(a) AUTOVC (b) SPEECHFLOW

FIGURE 5.9 – Schéma de principe du système SpeechSplit inspiré d'AutoVC où E_r est l'encodeur du rythme, E_c l'encodeur du contenu linguistique et E_f l'encodeur de la $F0$. Figure extraite de [Qian et al., 2020].

veau contenu, tout en conservant ce qui existe déjà. Pour cela, le système proposé inclut un modèle de durée des phones dans la séquence phonétique. Cette approche peut être étendue à d'autres paramètres que les aspects phonétiques, par exemple la prosodie ou l'expressivité.

Dans un contexte applicatif, pour que l'utilisateur puisse générer un échantillon de voix ayant l'expressivité voulue, il est vital de construire des commandes suffisamment explicites et adaptées pour que l'humain puisse s'en saisir. Le choix de commandes adaptées à l'utilisateur est un sujet assez complexe qui fait l'objet de la thèse de Thibault Gaudier débutée en octobre 2021. Dans cette thèse, le signal généré par un système de synthèse à partir du texte (TTS + Vocodeur) pourra être modifié par un utilisateur expert, par exemple pour de l'édition journalistique. L'expert pourra corriger la prononciation de certains mots (séquence de phonèmes et leur durée), ou corriger le style expressif afin qu'il soit adapté au contenu linguistique (intonation, accentuation de certains mots ou groupes de mots, style expressif) et modifier la voix en informant le système de l'identité du locuteur cible (par exemple un journaliste célèbre).

À l'heure actuelle, le système développé est illustré sur la figure 5.10. Le premier module (rouge) mis en place permet de générer une représentation de type phonétique (*Phonetic Posterior Gram* - PPG) issue de modèles pour la transcription. La représentation en PPG [Hazen et al., 2009] a été beaucoup utilisée pour la conversion de voix, notamment dans le *Voice Conversion Challenge 2020* [Zheng et al., 2020]. Cette représentation permet d'aligner

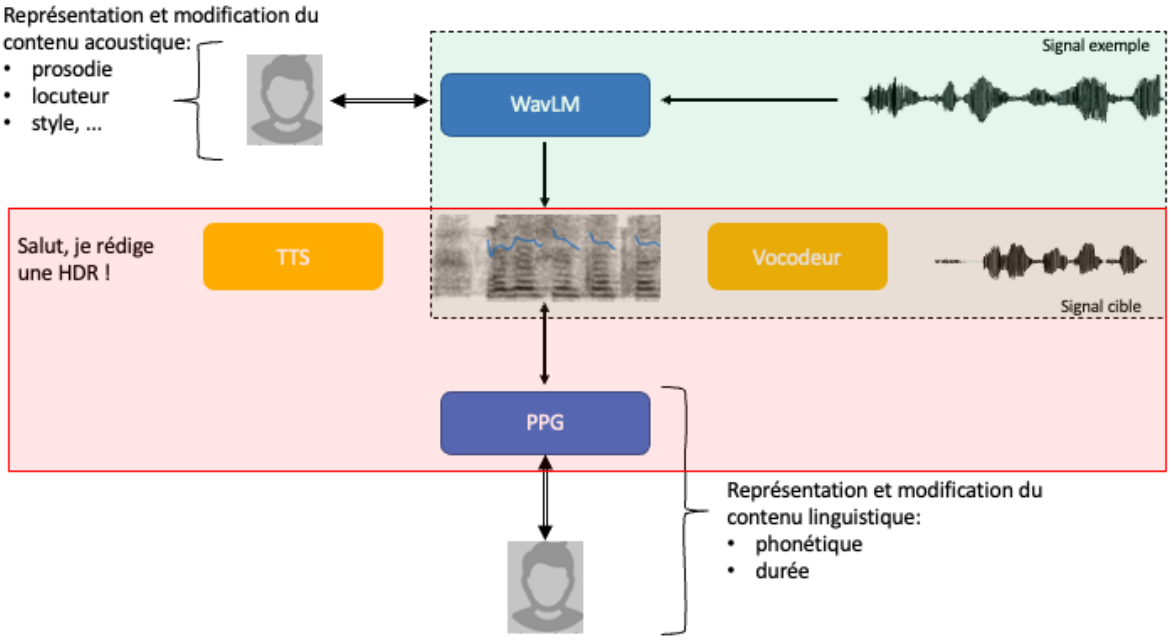


FIGURE 5.10 – Schéma de principe du module de correction de la synthèse (thèse de Thibault Gaudier). En rouge le module de correction linguistique et phonétique, en vert le module de correction prosodique et expressif.

une séquence d'unités phonétiques avec le temps. Elle contient donc au moins des informations phonétiques et de durée. Un de nos objectifs actuels est d'estimer la quantité d'information locuteur et expressive contenue dans les PPG. *In fine*, l'utilisateur pourra modifier une représentation ergonomique et contextualisée de type PPG afin de corriger le contenu phonétique et rythmique du signal synthétisé.

Le second module (vert) a été proposé dans le cadre du workshop JSALT [3] auquel Thibault Gaudier a participé. Dans un premier temps, il s'agit de reproduire le signal *exemple* à partir d'*embeddings* de signal. Nous avons choisi pour commencer d'utiliser la représentation pré-entraînée WavLM car elle a montré de gros bénéfices dans les tâches d'identification du locuteur et de segmentation en parole superposée. Nous avons montré que nous pouvions générer un signal cible intelligible à partir de cette représentation WavLM. Le vocodeur est un modèle WavGlow appris sur un corpus de voix anglaises. Le signal *exemple* est prononcé par un locuteur tamasheq issu du corpus [Boito et al., 2022]. À partir des *embeddings* WavLM extraits sur le signal *exemple*, un signal cible est généré. Ces travaux préliminaires montrent qu'il y a un

3. https://www.clsp.jhu.edu/speech-translation-for-under-resourced-languages/

effet locuteur sur le vocodeur neuronal puisque on perçoit le timbre de voix et l'accent anglais de la locutrice du corpus utilisé pour l'apprentissage de Wav-Glow sur le contenu linguistique en tamasheq du signal cible. L'effet locuteur sur les vocodeurs neuronaux n'a pas encore été très exploré, alors que c'est une limitation majeure lorsqu'on cherche à synthétiser de nouvelles voix. Dans [Maiti and Mandel, 2020], les auteurs montrent qu'il faut une quantité de locuteurs différents suffisamment importante (> 56) pour que l'effet locuteur du vocodeur ne soit plus perçu.

À partir de ces travaux préliminaires, nous chercherons, dans un second temps, à expliciter les aspects expressifs (prosodie et prononciation) via les représentations pré-entraînées (comme WavLM). Une commande utilisateur pourra alors modifier les valeurs des *embeddings* afin qu'ils correspondent au style qu'il souhaite. Cette étape nécessitera le développement de protocoles pour interpréter le contenu des *embeddings* et les modifier selon une condition extérieure. La thèse de Mano Brabant qui débute en octobre 2022 porte sur cette problématique.

5.4 Conclusion et discussions

Après une rapide présentation des systèmes de synthèse de la parole à partir du texte, les travaux présentés dans ce chapitre montrent que la synthèse de parole expressive demande de modéliser à la fois une prononciation et une prosodie adaptées au style voulu.

Modélisation de la prononciation expressive Les recherches que j'ai menées pendant mon post-doctorat à l'IRISA dans le cadre du projet ANR SynPaFlex autour de la modélisation de la prononciation pour de la synthèse expressive, ont donné lieu à plusieurs publications (3 conférences internationales et 1 article de revue). Dans ce même contexte, le corpus SynPaFlex mono-locuteur de parole lue a été richement annoté, incluant une annotation émotionnelle. Ce corpus a été mis à disposition de la communauté à la fois pour la caractérisation de la parole expressive (incluant la reconnaissance des émotions) et pour la synthèse de parole expressive. Comme nous l'avons déjà évoqué précédemment, la diffusion de corpus dans des langues variées, autre que l'anglais, est un point essentiel pour supporter la recherche dans un domaine où les technologies demandent des quantités de données de plus en plus importantes.

Modélisation de la voix expressive J'ai initié des recherches autour de la modélisation explicite d'une voix expressive au LIUM avec la thèse de Thibault Gaudier. L'objectif n'est pas de proposer un nouveau système de synthèse car cela impliquerait de se retrouver en concurrence

avec les GAFAM. Nos travaux sont ancrés dans un contexte précis d'édition journalistique où un utilisateur expert doit interagir avec le module de génération de voix. En cela, nous cherchons à modéliser explicitement la prosodie au travers des paramètres classiques (F_0, rythme, intensité, accentuation) mais également de proposer des modèles de styles adaptés à la parole journalistique. Ces modèles serviront ensuite à corriger la sortie d'un système de synthèse. Ainsi nous souhaitons tirer profit des représentations pré-entraînées développées dans d'autres domaines du traitement automatique de la parole (identification du locuteur, ASR principalement) afin de gagner en performance, tout en cherchant à modifier ces représentations et de les adapter aux conditions fournies par l'utilisateur. Ces travaux n'ont pas encore été publiés, cependant la participation de Thibault Gaudier au workshop JSALT a permis de confronter et de généraliser notre protocole à un contexte multilingue.

Perspectives Les premiers systèmes de synthèse de parole ont été conçus en se basant directement sur le modèle source-filtre de la production acoustique de la parole, en y adjoignant un modèle de durée. Les vocodeurs associés utilisent en entrée des paramètres de source (F_0 et voisement) et de filtre (enveloppe spectrale). Avec l'utilisation des modèles neuronaux du type end-to-end (Tacotron), les paramètres de ces modèles ont disparu à la fois pour l'utilisateur et pour le développeur. Cependant, on remarque que dans les publications récentes autour de la synthèse de voix expressive, ces paramètres tendent à re-devenir explicites, notamment le rythme et l'intonation.

En effet, dès lors qu'un utilisateur souhaite contrôler son outil et le paramétrer à sa convenance, il est nécessaire de lui fournir des paramètres de contrôle. Et ceux-ci sont tout naturellement issus du modèle source-filtre. Dans un avenir proche, il est important de trouver un compromis entre les très bonnes performances en qualité des systèmes de synthèse neuronaux et la possibilité de les contrôler. La caractérisation d'un style expressif étant une tâche très difficile et subjective, l'utilisation de méthodes non-supervisées pour regrouper des styles semblent une excellente piste [Qian et al., 2020]. Le fait de ce placer dans un contexte d'édition implique de déterminer des paramètres de contrôle ergonomiques pour l'utilisateur. En ce sens, l'approche ajout/suppression/substitution proposée sur le contenu linguistique par [Tan et al., 2021] me semble favorable pour rendre l'application utilisable.

Nous avons montré que dans plusieurs cas, incluant la reconnaissance des émotions dans la parole et la détection de parole superposée, l'utilisation d'*embeddings* est une approche performante. Cependant ces représentations ne sont pas interprétables car apprises automatiquement de façon non-supervisée à partir d'un corpus de données. Une des perspectives à

court terme est d'analyser ces représentations et de proposer des protocoles qui permettent de projeter ces vecteurs de grande dimension peu interprétables vers des vecteurs de petite dimension représentant des paramètres connus, par exemple ceux du modèle source-filtre, et du modèle de durée.

À long terme, l'interprétation des représentations pré-entraînées pourra permettre de mieux comprendre l'organisation temporelle de la parole. En effet, les aspects rythmiques sont complexes à modéliser car ils sont multi-échelle comme évoqué dans le chapitre 1. Or les modèles conçus pour apprendre les *embeddings* sont justement multi-échelle dans ce sens où ils capturent des informations sur des fenêtres temporelles de tailles variables. Des parallèles entre les signaux de parole et les signaux musicaux nous permettront de mieux analyser cette structure rythmique.

Troisième partie

Perspectives

Chapitre 6

Bilan et perspectives de recherche

Ce document présente les avancées qui me paraissent importantes dans le domaine du traitement automatique de la parole expressive sur les quinze dernières années. Ces années marquent également l'arrivée massive des réseaux de neurones qui ouvrent de nouvelles perspectives. Ces avancées sont mises en regard des travaux de recherche dans lesquels je me suis impliquée. En accord avec mon parcours de chercheuse pluridisciplinaire, je me suis attachée à décrire les travaux réalisés dans plusieurs domaines tels que la segmentation du signal, sa caractérisation incluant l'identification du locuteur et des phénomènes expressifs (hésitations, interruptions, émotions, etc.) et sa génération. Cette diversité est également un des points forts des études sur le signal de parole. L'ensemble des travaux a été mené dans l'objectif d'être diffusé le plus largement possible dans la communauté scientifique et dans la société, à travers une participation active à des logiciels libres, ou des bases de données publiques.

Je propose quatre axes de recherche pour les prochaines années que je présente dans la section 6.1. Des perspectives ou cadres de travail plus généraux sont exposés dans la section 6.2.

6.1 Modélisation de l'expressivité

Mes travaux de recherche couvre l'analyse de la parole expressive sur plusieurs niveaux : depuis la segmentation du signal audio (zones de parole, de silence, de parole superposée, locuteur, etc.) à la caractérisation haut niveau (interruption, hésitations, émotion, etc.), et à la génération d'un signal de parole expressif. Le fait d'étudier les deux facettes (analyse et synthèse) permet de définir finement un phénomène expressif par des caractéristiques acoustiques, prosodiques, phonétiques et linguistiques, mais également de valider ces caractéristiques par la synthèse de signaux et leur évaluation perceptive. Ce double point de vue est, à mon sens, très important pour appréhender les comportements oraux des êtres humains dans toute leur diversité et leur complexité.

Segmentation automatique du signal audio Les corpus de parole actuels contiennent une très grande quantité de données (plus de 1000 h) dont la segmentation et l'annotation manuelle est devenue impossible. Même sur des quantités de données "raisonables" comme dans le corpus ALLIES [Shamsi et al., 2022], le coût d'annotation (par exemple en locuteur) est très élevé principalement du fait que la segmentation manuelle est une tâche laborieuse et les segments obtenus sont d'une précision variable intra et inter annotateurs. Il y a donc un véritable enjeu à développer des outils de segmentation précis et fiables. La détection automatique des zones de parole, de musique, de parole superposée et les changements de locuteur est un domaine en plein essor. Les travaux de thèse de Martin ont permis de proposer des modèles robustes pour la parole superposée, j'envisage d'aller plus loin, notamment en incluant les zones de segments de musique et la possibilité de prédire plusieurs catégories pondérées sur un même segment. Dans le projet GEM, l'objectif principal est d'étudier les interactions entre les hommes et les femmes dans les médias français, l'identification automatique du genre permet de proposer des analyses à très grande échelle. Les techniques actuelles ne permettent pas d'identifier les genres sur une zone de parole superposée, or c'est justement en caractérisant plus finement ces zones là que nous pourrons analyser avec l'aide de sociologues les interactions entre les différents genres.

Pour que ces modèles, notamment les détecteurs de genre, soit équitables, il est nécessaire d'évaluer les biais induits par les données, leurs représentations et par les modèles. Dans la continuité des travaux de post-doctorat d'Ambuj Merhish, j'envisage d'évaluer les biais de genre et d'origine dans les tâches touchant à l'identité des locuteurs puis de développer des techniques afin de les réduire.

Modélisation faiblement / non supervisée des styles expressifs L'utilisation de méthodes supervisées implique de définir une ontologie des phénomènes étudiés soit en utilisant des dimensions affectives, soit des catégories discrètes. Cette ontologie a l'avantage d'être le fruit d'une réflexion humaine et donc interprétable facilement. De plus, c'est généralement sur cette définition que les différents acteurs en jeu se mettent d'accord. Par exemple, dans le projet GEM, la définition de ce qu'est une "interruption" est différente suivant que l'on vient des sciences sociales, politiques, militantes ou de l'informatique. L'intérêt d'un projet pluridisciplinaire est justement de permettre aux différents domaines d'échanger autour des définitions. Ce point est crucial car il permet d'estimer les potentiels biais et les limitations induites par les simplifications nécessaires au traitement automatique de la parole. Cependant, alors que l'annotation manuelle se fait sur la perception de différents aspects plus ou moins implicites

et parfois sur plusieurs modalités, la prédiction par un système n'utilise qu'une représentation restreinte du phénomène en entrée (par exemple le signal uniquement). À moyen terme, je souhaite explorer différentes approches pour la caractérisation des phénomènes expressifs à l'aide de méthodes non supervisées. La confrontation de l'ontologie obtenue à partir du système (et donc de sa représentation restreinte) avec celle définie par un groupe pluridisciplinaire aidera mieux identifier les caractéristiques manquantes au modèle. En particulier, j'envisage d'étendre l'approche de *clustering spectral* développée pour de la segmentation en motifs musicaux à de la segmentation et regroupement en styles expressifs dans des corpus de parole journalistique ou de livres audio.

Caractérisation de l'expressivité en interaction : prononciation, prosodie, accentuation, disfluences Dans mes travaux de recherche, j'ai montré qu'un style expressif est dépendant à la fois du contenu phonetico-linguistique et de la prosodie. En collaboration avec le LPP, j'ai également initié une étude autour de l'hésitation dans la parole spontanée qui a permis de mettre en évidence certaines caractéristiques acoustiques et phonétiques. J'ai beaucoup utilisé des approches de type "sélection de descripteurs" qui ont l'avantage d'être automatisables, mais qui ne reflètent pas toujours les perceptions des auditeurs (travaux réalisés au LIMSI en thèse et à l'IRISA en post-doctorat). En effet, ces méthodes sélectionnent les meilleurs descripteurs sur la base d'un critère de performance sur un ensemble de données restreint. Ce qui fonctionne sur un corpus, ne généralisera pas nécessairement à d'autres. Or la perception humaine tend à généraliser le phénomène à des contextes beaucoup plus variés que celui lié au corpus. Je souhaite continuer à explorer ces différentes modalités pour la caractérisation d'un style expressif en prenant mieux en compte le contexte.

Dans le cadre de la parole spontanée, la présence de *disfluences* et d'*affect bursts* modifie l'organisation temporelle du signal audio par rapport à une parole préparée ou lue. Même si beaucoup de chercheurs ont exploré les aspects rythmiques de la parole, il reste encore beaucoup d'éléments à étudier. Par exemple, dans la thèse de Rémi Uro, nous cherchons à définir les zones du signal de parole où il est plus probable d'avoir une interruption. L'extraction d'une structure rythmique de la parole, c'est-à-dire l'organisation (répétitions, variations) temporelle des différentes unités qui la composent, semble caractéristique d'un style ou de l'expression d'une émotion particulière. L'utilisation de modèle de prédiction prenant en entrée des séquences temporelles, tels que les réseaux récurrents, ouvrent des perspectives intéressantes pour l'exploration du rythme dans les signaux musicaux. J'envisage d'explorer l'extraction automatique de représentations visuelles d'une structure temporelle multi-échelle dans la continuité

de mes travaux en collaboration avec l'IReMus.

Développement d'approches *human in the loop* pour la synthèse et l'annotation Les technologies développées avec les approches *end-to-end* ont apporté de forts gains de performances que ce soit dans le domaine de la caractérisation ou la synthèse automatique de la parole. Ces modèles sont généralement appris sur des données peu contrôlées (livres audio, journaux d'information, etc.) qui sont disponibles en grandes quantités. L'intervention humaine dans la collecte et l'annotation de ces données est limitée à son strict minimum. Cela explique pourquoi les modèles *end-to-end* reposent si peu sur des modèles de connaissances (modèles de production acoustique, modèles linguistique, cognitifs, etc.). L'injection de connaissance dans les réseaux de neurones peut se faire lors de l'apprentissage directement (conditionnement sur des paramètres extérieurs) ou lors de l'inférence (contrôle). Elle nécessite une intervention humaine, mais permet d'accéder à des éléments de connaissance interprétables.

J'envisage de continuer le développement d'approches pour l'apprentissage actif, principalement pour l'annotation semi-supervisée des données (dans la suite du projet ALLIES). Dans cette approche, l'intervention humaine se limite à valider ou non la prédiction fournie par le modèle sur un segment de parole. Le modèle continuera son apprentissage en prenant en compte les annotations humaines. L'analyse de l'évolution des représentations latentes du modèle en fonction du nombre d'interventions permettra de trouver un compromis entre un regroupement complètement non-supervisé réalisé automatiquement, la perception humaine, ou encore une ontologie définie par un groupe d'experts.

Dans le contexte de la synthèse de parole, l'intervention humaine passera par la modification de paramètres de contrôle afin d'obtenir un signal synthétique de style adapté au contexte. La collaboration avec le LIA, à travers le projet Européen SELMA piloté par DeutscheWelle, consortium journalistique allemand, finance la thèse de Thibault Gaudier dont l'objectif est de proposer un système de synthèse de parole neuronal permettant à un expert de corriger/modifier le signal prédit à partir du texte. Au delà des aspects purement informatique de conception et d'apprentissage du système, un gros travail en collaboration avec des éditeurs journalistes sera nécessaire pour identifier les paramètres de contrôle pertinent pour cette tâche.

6.2 Algorithmes et société

Aujourd'hui les applications en Intelligence Artificielle ont un impact extrêmement important pour les sociétés. Ces applications sont commercialisées, utilisées à des fins diverses et variées, dont les objectifs sont parfois peu louables. Il est donc de notre devoir de chercheur.e de proposer des outils performants avec un exposé des limites et des risques liés à ces outils.

Interprétabilité et explicabilité L'explosion des performances obtenues par les approches neuronales a fait que les applications les utilisant sont aujourd'hui largement utilisées (traduction automatique, synthèse de parole, transcription, identification du locuteur, etc...). On peut notamment citer l'outil controversé COMPARSE [1] utilisé par la justice des USA pour aider les juges à prédire le risque de récidive de criminels. Cependant, si les utilisateurs sont satisfaits lorsque les prédictions sont conformes aux attentes, ou bien qu'elles ne semblent pas (ou peu) porter préjudice aux personnes, le risque peut devenir important pour des domaines sensibles tels que la santé et la justice. Par exemple, la délivrance par une machine d'un diagnostic erroné à partir de symptômes catégorisés, ou d'un signal audio, entraînera un traitement inadapté ou pas de traitement du tout. Une erreur sur l'identification d'un locuteur lors d'une expertise criminalistique aura une forte influence sur la décision de justice, pouvant entraîner une condamnation injuste ou au contraire la non condamnation d'un criminel. À ce stade de développement des technologies vocales, il semble vital que les systèmes fournissent à la fois le résultat, une confiance dans cette prédiction, ainsi qu'un certain nombre d'explications [Campbell et al., 2009]. Ces explications à l'interface entre les humains et le processus de décision, aideront l'expert à prendre sa décision finale, par exemple : "le système a identifié une pathologie des cordes vocales parce que le jitter calculé sur la F_0 de certains phonèmes est très élevé".

En tant que chercheur.e.s en Intelligence Artificielle, certes, nous développons des systèmes performants, mais nous devons également les analyser et les comprendre. Tout un nouveau champ de recherche s'est développé ces dernières années autour de l'explicabilité et l'interprétabilité des modèles appris automatiquement comme l'atteste l'apparition de nombreuses conférences scientifiques sur ce sujet [2].

L'approche post-hoc est celle que j'ai poursuivie jusqu'alors. Elle requiert l'utilisation d'un modèle entraîné pour déterminer des liens entre des paramètres interprétables, les données

1. `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`
2. IJCAI workshop for Interpretability for AI in 2017, la conférence annuelle ACM *Fairness, Accountability and Transparency* (FaccT) en 2018, etc.

et les résultats des modèles : réduction de dimension pour la visualisation de relations sémantiques [Liu et al., 2018], sélection du meilleur ensemble de descripteurs pour une tâche donnée [Tahon and Devillers, 2016a], analyser les corrélations entre les descripteurs acoustiques et les résultats d'un système [González Hautamäki et al., 2019]. Cependant aucune de ces techniques n'intègrent une explication dans le système lui-même. La technique LIME (*Local intepretable model-agnostic explanations*) pour la classification d'images [Ribeiro et al., 2016] identifie les zones sur lesquelles le modèle s'est appuyé pour prendre sa décision. Cette technique pourra être adapté à des représentations temps-fréquences du signal audio. Les explications *counterfactual* ont l'avantage de comparer le comportement d'un modèle entre deux exemples [Strobelt et al., 2019]. Le principe est de modifier légèrement les valeurs d'un descripteur et d'observer l'effet sur la prédiction du modèle. Ces deux approches sont intéressantes mais elles nécessitent une forte compréhension de l'outil et sont difficiles à exploiter à grande échelle.

La thèse de Mano Brabant explorera plutôt les représentations vectorielles utilisées pour le traitement de la parole (WavLM, x-vector, etc.) pour être capable de les interpréter avec des descripteurs experts issus des modèles linguistiques et acoustiques sans nécessairement avoir à ré-entraîner des modèles consommateurs de ressources tels que WavLM. La projection des *embeddings* extraits sur différentes fenêtres temporelles dans des espaces parcimonieux de grande dimension devrait permettre de mettre à jour des dimensions latentes interprétables et ainsi de reconnecter les espaces abstraits à des attributs du signal de parole humainement compréhensibles [Noé et al., 2022].

Je souhaite que ces travaux de recherche soient généralisables à d'autres domaines que celui du traitement automatique de la parole. C'est pourquoi j'envisage de collaborer avec le LAUM, entre autres, sur l'étude des relations entre les propriétés mécaniques de vibration d'anches de saxophone et la qualité perçue par un instrumentiste.

Biais de genre, mixité et données privées Je souhaite que mes travaux de recherche autour de la parole permettent à des chercheur.e.s en sciences humaines d'étudier à large échelle les représentations de genre dans les interactions vocales. Dans le cadre du projet GEM, la caractérisation des interruptions en fonction du genre permettra de mettre en évidence les biais de représentation des femmes et des hommes dans les medias français.

Ces travaux vont de paire avec l'évaluation des biais dans les données et modèles que nous construisons. En cela, les travaux de post-doctorat d'Ambuj Merish dans le cadre du projet ExTenSOR ont permis de poser les fondations d'un travail sur le long terme. Dans ces travaux

que je n'ai pas détaillé dans le document, nous avons remarqué dans plusieurs expériences que malgré le fait que les modèles soient appris sur des quantités de données parfaitement équilibrées entre locuteurs homme et femme, les performances sont systématiquement biaisées en faveur des hommes. Cela implique un moins bon fonctionnement des outils de technologie vocale sur les voix féminines (notamment transcription, et identification). Malgré une exploration assez large du biais dans les données et à différents niveaux des réseaux de neurones, nous n'avons pas à l'heure actuelle d'explication à ce biais. La plupart des systèmes utilisent des représentations temps-fréquences basées sur l'échelle de Mel qui est linéaire en dessous de 1000 Hz. Ainsi les F_0 des hommes et des femmes seraient mécaniquement distribuées différemment. Une piste serait d'explorer des représentations fréquentielles logarithmiques sur l'échelle de la F_0, ou bien d'utiliser des réseaux de neurones convolutionnels de type SincNet [Ravanelli and Bengio, 2018] qui prennent en entrée la forme d'onde directement. Ce travail va de paire avec un engagement sur la mixité homme-femme en informatique au sein de l'Université au niveau des personnels et des étudiants.

À l'heure où les grands groupes industriels et les institutions publiques dépensent des moyens colossaux pour développer des algorithmes d'IA (par exemple le *health data hub*), il est du devoir des chercheur.e.s académiques de vulgariser leurs connaissances afin d'apporter une expertise citoyenne aux législateur.e.s, aux juristes, aux acteurs politiques pour l'encadrement de l'utilisation de ces outils. Le Règlement Général sur la Protection des Données (RGPD) adopté au niveau national et européen apporte un premier cadre pour la définition et l'utilisation des données personnelles. Il est important que les expert.e.s et chercheur.e.s en IA participent à la définition de ce cadre, notamment pour caractériser les aspects "privés". Par exemple dans le projet DeepPrivacy, les travaux de Hubert Nourtel ont posé la question du caractère privé de l'état émotionnel d'un locuteur. La diffusion de données de parole à l'ensemble de la communauté, la mise à disposition des codes et des logiciels libres, ainsi que l'accessibilité des articles présentant des travaux sur lesquels je collabore, sont des points essentiels pour moi. La participation aux échanges avec d'autres disciplines telles que les sciences humaines, mais aussi des juristes ou des institutions civiles et politiques, font partie de mon travail quotidien de chercheuse.

Prise en compte du contexte environnemental On sait aujourd'hui que le numérique a un impact environnemental important et évoluant très rapidement. En particulier, les technologies utilisant des réseaux de neurones profonds sont très coûteuses en énergie et en ressources premières de type métaux rares. De mon point de vue, les chercheur.e.s ne peuvent plus

continuer à entraîner des modèles gigantesques consommateurs de ressources sans avoir au préalable étudié son impact environnemental au regard des bénéfices pour la société.

Dans le domaine du traitement automatique de la parole, nous avons la chance d'avoir à disposition un grand nombre de modèles pré-entraînés qui fournissent des représentations fines du signal audio. L'utilisation de ces représentations (sans les adapter à la tâche) permet d'alléger le nombre de paramètres des modèles spécifiques et de réduire le nombre de calculs nécessaires pour atteindre de bonnes performances. C'est effectivement ce que nous avons montré que ce soit pour la détection de parole superposée ou la reconnaissance d'émotions. La prise en compte systématique du coût carbone est un point qui me semble important pour que les décideur.euse.s et les utilisateur.rice.s puissent se rendre compte de l'impact environnemental des technologies de type *deep learning*. Cela pourrait également renforcer les échanges avec nos tutelles afin de conditionner certains financements à leur impact environnemental, notamment dans le domaine de l'IA.

Si l'on doit envisager un avenir encore plus contraint pour la recherche en informatique, notamment dans le cas probable d'une pénurie de ressources pour la construction des machines ou des restrictions énergétiques, alors la place des technologies en traitement automatique de la parole devra être complètement repensée. Les thématiques liées à l'analyse des contenus médiatiques (que ce soit TV, radio, tweeter, ou autres réseaux sociaux) me semble prioritaire afin de continuer à mettre en évidence les biais, les absurdités, les fausses nouvelles.

BIBLIOGRAPHIE

[Abdelwahab and Busso, 2017] Abdelwahab, M. and Busso, C. (2017). Incremental adaptation using active learning for acoustic emotion recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5160–5164, New Orleans, LA. IEEE.

[Adda-Decker et al., 2008] Adda-Decker, M., Barras, C., Adda, G., Paroubek, P., de Mareüil, P. B., and Habert, B. (2008). Annotation and analysis of overlapping speech in political interviews. In *LREC*, page 7, Marrakech, Marroco.

[Alain et al., 2017] Alain, P., Barbot, N., Chevelu, J., Lecorvé, G., Lolive, D., Simon, C., and Tahon, M. (2017). The IRISA Text-To-Speech System for the Blizzard Challenge 2017. In *Blizzard Challenge (satellite of Interspeech)*, Stockholm, Sweden.

[Aubergé et al., 2006] Aubergé, V., Audibert, N., and Rilliard, A. (2006). De E-Wiz à C-Clone : recueil, modélisation et synthèse d'expressions authentiques. *Revue des Sciences et Technologies de l'Information - Série RIA : Revue d'Intelligence Artificielle*, 20 :499–527. Publisher : Lavoisier.

[Banse and Scherer, 1996] Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, 70(3) :614–636.

[Barras et al., 2007] Barras, C., Zhu, X., Gauvain, J.-L., and Lomel, L. (2007). The CLEAR'06 LIMSI Acoustic Speaker Identification System for CHIL Seminars | SpringerLink. In *Multimodal Technologies for Perception of Humans. CLEAR 2006*, LNCS 4122.

[Barras et al., 2006] Barras, C., Zhu, X., Meignier, S., and Gauvain, J.-L. (2006). Multistage speaker diarization of broadcast news. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5) :1505–1512. Conference Name : IEEE Transactions on Audio, Speech, and Language Processing.

[Bartkova et al., 2016] Bartkova, K., Jouvet, D., and Delais-Roussarie, E. (2016). Prosodic parameters and prosodic structures of French emotional data. In *Speech Prosody 2016*, pages 644–648. ISCA.

[Bell et al., 2003] Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *J. Acoust. Soc. Am.*, 113 :1001–1024.

[Bimbot et al., 2004] Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., and Reynolds, D. A. (2004). A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Advances in Signal Processing*, 2004(4) :101962.

[Blanche-Benveniste, 1990] Blanche-Benveniste, C. (1990). *Le français parlé. Etudes grammaticales*. Sciences du langage. Paris, cnrs éditions edition.

[Boeffard et al., 2012] Boeffard, O., Charonnat, L., Maguer, S. L., Lolive, D., and Vidal, G. (2012). Towards Fully Automatic Annotation of Audiobooks for TTS. In *LREC*, Istanbul, Turkey.

[Boersma and Weenink, 2018] Boersma, P. and Weenink, D. (2018). Praat : doing phonetics by computer [Computer program]. Technical report, Version 6.0.37, retrieved 14 March 2018 from http ://www.praat.org/.

[Boito et al., 2022] Boito, M. Z., Bougares, F., Barbier, F., Gahbiche, S., Barrault, L., Rouvier, M., and Estève, Y. (2022). Speech Resources in the Tamasheq Language. In *LREC*, page 6, Marseille, France.

[Boula de Mareüil et al., 2005] Boula de Mareüil, P., Habert, B., Bénard, F., Adda-Decker, M., Barras, C., Adda, G., and Paroubek, P. (2005). A quantitative study of disfluencies in French broadcast interviews. In *Disfluency in Spontaneous Speech Workshop*.

[Bove, 2008] Bove, R. (2008). *Analyse syntaxique automatique de l'oral : étude des disfluences*. phdthesis, Université de Provence - Aix-Marseille I.

[Bradley and Lang, 1994] Bradley, M. M. and Lang, P. J. (1994). Measuring emotion : The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1) :49–59.

[Bredin, 2020] Bredin, H. (2020). pyannote.audio 2.0a1 : Neural building blocks for speaker diarization.

[Bredin et al., 2020] Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020). Pyannote.Audio : Neural Building Blocks for Speaker Diarization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128. ISSN : 2379-190X.

[Brendel et al., 2010] Brendel, M., Zaccarelli, R., and Devillers, L. (2010). Building a system for emotions detection from speech to control an affective avatar. In *LREC*, page 6, Valleta, Malta.

[Brognaux et al., 2014] Brognaux, S., Picart, B., and Drugman, T. (2014). Speech synthesis in various communicative situations : Impact of pronunciation variations. In *Interspeech*, pages 1524–1528.

[Broux et al., 2018] Broux, P.-A., Desnous, F., Larcher, A., Petitrenaud, S., Carrive, J., and Meignier, S. (2018). S4D : Speaker Diarization Toolkit in Python. In *Interspeech 2018*, pages 1368–1372. ISCA.

[Bullock et al., 2020] Bullock, L., Bredin, H., and Garcia-Perera, L. P. (2020). Overlap-Aware Diarization : Resegmentation Using Neural End-to-End Overlapped Speech Detection. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7114–7118. ISSN : 2379-190X.

[Burkhardt et al., 2005] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., and Weiss, B. (2005). A Database of German Emotional Speech. In *Interspeech*, pages 1517–1520, Lissabon, Portugal.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP : interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4) :335–359.

[Busso and Rahman, 2012] Busso, C. and Rahman, T. (2012). Unveiling the acoustic properties that describe the valence dimension. In *Interspeech 2012*, pages 1179–1182. ISCA.

[Bänziger and Scherer, 2005] Bänziger, T. and Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, 46(3) :252–267.

[Béchet, 2001] Béchet, F. (2001). LIA-PHON : un système complet de phonétisation de texte. *Traitement Automatique des Langues (TAL)*, 42(1) :47–67.

[Caliński and Harabasz, 1974] Caliński, T. and Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics - Theory and Methods*, 3 :1–27.

[Camelin et al., 2010] Camelin, N., Bechet, F., Damnati, G., and de Mori, R. (2010). Detection and Interpretation of Opinion Expressions in Spoken Surveys. *IEEE Transactions on Audio, Speech and Language Processing*. Publisher : Institute of Electrical and Electronics Engineers.

[Campbell et al., 2009] Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J.-F., and Matrouf, D. (2009). Forensic speaker recognition. *IEEE*. Accepted : 2010-03-04T20 :58 :22Z Publisher : Institute of Electrical and Electronics Engineers.

[Campbell, 2000a] Campbell, N. (2000a). DATABASES OF EMOTIONAL SPEECH. In *ITRW on Speech and Emotion*, page 5, Newcastle, Nothern Ireland, UK.

[Campbell, 2000b] Campbell, N. (2000b). Timing in Speech : A Multi-Level Process. In Horne, M., editor, *Prosody : Theory and Experiment : Studies Presented to Gösta Bruce*, Text, Speech and Language Technology, pages 281–334. Springer Netherlands, Dordrecht.

[Campione and Véronis, 2005] Campione, E. and Véronis, J. (2005). Pauses and hesitations in French spontaneous speech. In *Disfluency in Spontenaous Speech Workshop*, pages 43–46, Aix-en-Provence, France.

[Carlson et al., 2006] Carlson, R., Gustafson, K., and Strangert, E. (2006). Cues for Hesitation in Speech Synthesis. In *9th International Conference on Spoken Language Processing*, pages 1300–1303, Pittsburgh, PA. ISCA.

[Charaudeau, 2006] Charaudeau, P. (2006). Discours journalistique et positionnements énonciatifs. Frontières et dérives. *Semen. Revue de sémio-linguistique des textes et discours*, (22). Number : 22 Publisher : Presses universitaires de Franche-Comté.

[Chen et al., 2022] Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., Wu, J., Zhou, L., Ren, S., Qian, Y., Qian, Y., Wu, J., Zeng, M., Yu, X., and Wei, F. (2022). WavLM : Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, pages 1–14. Conference Name : IEEE Journal of Selected Topics in Signal Processing.

[Chen et al., 2017] Chen, Z., Xie, Z., Zhang, W., and Xu, X. (2017). ResNet and Model Fusion for Automatic Spoofing Detection. In *Interspeech 2017*, pages 102–106. ISCA.

[Chevelu et al., 2014] Chevelu, J., Lecorvé, G., and Lolive, D. (2014). ROOTS : A toolkit for easy, fast and consistent processing of large sequential annotated data collections. In *9th International Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

[Clark, 1999] Clark, H. H. (1999). Speaking in time. In *ETRW on Dialogue and Prosody*, pages 1–6, Veldhoven, The Netherlands. ISCA.

[Clavel et al., 2008] Clavel, C., Vasilescu, I., Devillers, L., Richard, G., and Ehrette, T. (2008). Fear-type emotion recognition for future audio-based surveillance systems. *Speech Communication*, 50(6) :487–503.

[Cornell et al., 2020] Cornell, S., Omologo, M., Squartini, S., and Vincent, E. (2020). Detecting and Counting Overlapping Speakers in Distant Speech Scenarios. In *Interspeech 2020*, pages 3107–3111. ISCA.

[Cowie et al., 2000] Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., and Schroeder, M. (2000). 'FEELTRACE' : AN INSTRUMENT FOR RECORDING PERCEIVED EMOTION IN REAL TIME. In *ITRW on Speech and Emotion*, page 6, Newcastle, Northern Ireland, UK.

[Cowie et al., 2001] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. (2001). Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1) :32–80. Conference Name : IEEE Signal Processing Magazine.

[d'Alessandro, 2011] d'Alessandro, C. (2011). Computerized chironomy : Five years of gesture - controlled voice and speech synthesis at LIMSI. In *International Workshop on Performative Speech and Singing Synthesis (P3S 2011)*, page 4p, Vancouver, Canada.

[d'Alessandro and Mertens, 1995] d'Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer Speech & Language*, 9(3) :257–288.

[Davies and Bouldin, 1979] Davies, D. L. and Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2) :224–227. Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence.

[Davis and Mermelstein, 1980] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4) :357–366. Conference Name : IEEE Transactions on Acoustics, Speech, and Signal Processing.

[de Cheveigné and Kawahara, 2002] de Cheveigné, A. and Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4) :1917–1930. Publisher : Acoustical Society of America.

[Degottex, 2010] Degottex, G. (2010). *Glottal source and vocal-tract separation*. PhD thesis, Université Pierre et Marie Curie - Paris VI, Paris, France.

[Degottex et al., 2011] Degottex, G., Roebel, A., and Rodet, X. (2011). Phase Minimization for Glottal Model Estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5) :1080–1090.

[Dehak et al., 2011] Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front-End Factor Analysis for Speaker Verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4) :788–798. Conference Name : IEEE Transactions on Audio, Speech, and Language Processing.

[Delaborde et al., 2009] Delaborde, A., Tahon, M., Barras, C., and Devillers, L. (2009). A Wizard-of-Oz game for collecting emotional audio data in a children-robot interaction. In *International Workshop on Affective-Aware Virtual Agents and Social Robots*, Boston, United States. ACM Press.

[Devillers et al., 2015] Devillers, L., Tahon, M., Sehili, M. E. A., and Delaborde, A. (2015). Inference of Human Beings' Emotional States from Speech in Human-Robot Interactions. *International Journal of Social Robotics*. Publisher : Springer.

[Devillers et al., 2010] Devillers, L., Vaudable, C., and Chasatgnol, C. (2010). Real-life emotion-related states detection in call centers : a cross-corpora study. In *Interspeech*, pages 2350–2355, Makuhari, Chiba, Japan.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Diez et al., 2020] Diez, M., Burget, L., Landini, F., Wang, S., and Cernocky, H. (2020). Optimizing Bayesian Hmm Based X-Vector Clustering for the Second Dihard Speech Diarization Challenge. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6519–6523, Barcelona, Spain. IEEE.

[Dixon and Maxey, 1968] Dixon, N. and Maxey, H. (1968). Terminal analog synthesis of continuous speech using the diphone method of segment assembly. *IEEE Transactions on Audio and Electroacoustics*, AU-16(1) :39–50.

[Donahue et al., 2019] Donahue, C., McAuley, J., and Puckette, M. (2019). Adversarial Audio Synthesis. arXiv :1802.04208 [cs].

[Dong and Lee, 2018] Dong, H.-Y. and Lee, C.-M. (2018). Speech intelligibility improvement in noisy reverberant environments based on speech enhancement and inverse filtering. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018(1) :3.

[Douglas-Cowie et al., 2003] Douglas-Cowie, E., Campbell, N., Cowie, R., and Roach, P. (2003). Emotional speech : Towards a new generation of databases. *Speech Communication*, 40(1) :33–60.

[Doukhan et al., 2018a] Doukhan, D., Carrive, J., Vallet, F., Larcher, A., and Meignier, S. (2018a). An Open-Source Speaker Gender Detection Framework for Monitoring Gender Equality. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5214–5218. ISSN : 2379-190X.

[Doukhan et al., 2018b] Doukhan, D., Poels, G., Rezgui, Z., and Carrive, J. (2018b). Describing Gender Equality in French Audiovisual Streams with a Deep Learning Approach. *VIEW Journal of European Television History and Culture*, 7(14) :103–122. Number : 14 Publisher : Sound & Vision.

[Dutrey, 2014] Dutrey, C. (2014). *Analyse et détection automatique de disfluences dans la parole spontanée conversationnelle*. PhD thesis, Université Paris Sud - Paris XI.

[Ekman, 1999] Ekman, P. (1999). Basic Emotions. In Dagleish, T. and Power, M., editors, *Handbook of Cognition and Emotion*, pages 301–320. Wiley, New-York.

[Erickson et al., 2006] Erickson, D., Yoshida, K., Menezes, C., Fujino, A., Mochida, T., and Shibuya, Y. (2006). Exploratory study of some acoustic and articulatory characteristics of sad speech. *Phonetica*, 63(1) :1–25.

[Esteve et al., 2010] Esteve, Y., Bazillon, T., Antoine, J.-Y., Bechet, F., and Farinas, J. (2010). The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news. In *LREC*, page 4, Valleta, Malta.

[Eyben et al., 2016] Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., and Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2) :190–202. Conference Name : IEEE Transactions on Affective Computing.

[Eyben et al., 2010] Eyben, F., Wöllmer, M., and Schuller, B. (2010). Opensmile : the munich versatile and fast open-source audio feature extractor. In *Proceedings of the international conference on Multimedia - MM '10*, page 1459, Firenze, Italy. ACM Press.

[Fant et al., 1962] Fant, G., Martony, J., Rengman, U., and Risberg, A. (1962). OVE II synthesis strategy. In *Speech Communication*, Stockholm, Sweden.

[Fredouille, 2016] Fredouille, C. (2016). *Traitement Automatique et Troubles de la Voix et de la Parole : champs d'application, contraintes et limites*. Habilitation à diriger des recherches, Université d'Avignon et du Pays de Vaucluse.

[Friedland et al., 2009] Friedland, A. G., Oriol Vinyals, B., Yan Huang, C., and Muller, D. C. (2009). Fusing short term and long term features for improved speaker diarization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4077–4080. ISSN : 2379-190X.

[Fujita et al., 2019] Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., and Watanabe, S. (2019). End-to-End Neural Speaker Diarization with Permutation-Free Objectives. In *Interspeech 2019*, pages 4300–4304. ISCA.

[Galliano et al., 2006] Galliano, S., Geoffrois, E., and Gravier, G. (2006). Corpus description of the ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. In *LREC*, page 4, Genoa, Italy.

[Garcia Perera et al., 2020] Garcia Perera, L. P., Villalba, J., Bredin, H., Du, J., Castan, D., Cristia, A., Bullock, L., Guo, L., Okabe, K., Nidadavolu, P. S., Kataria, S., Chen, S., Galmant, L., Lavechin, M., Sun, L., Gill, M.-P., Ben-Yair, B., Abdoli, S., Wang, X., Bouaziz, W., Titeux, H., Dupoux, E., Lee, K. A., and Dehak, N. (2020). Speaker Detection in the Wild : Lessons Learned from JSALT 2019. In *The Speaker and Language Recognition Workshop (Odyssey 2020)*, pages 415–422. ISCA.

[Garcia-Romero et al., 2019] Garcia-Romero, D., Snyder, D., Sell, G., McCree, A., Povey, D., and Khudanpur, S. (2019). x-Vector DNN Refinement with Full-Length Recordings for Speaker Recognition. In *Interspeech 2019*, pages 1493–1496. ISCA.

[Garnier et al., 2004] Garnier, M., Dubois, D., Poitevineau, J., Henrich, N., and Castellengo, M. (2004). Perception et description verbale de la qualité vocale dans le chant lyrique : une approche cognitive. In *XXVème Journées d'Études sur la Parole*, Fez, Marocco.

[Gaudier, 2020] Gaudier, T. (2020). *Conversion de voix : application au changement de personnages*. PhD thesis, Université de Rennes 1, Rennes, France.

[Gelly and Gauvain, 2018] Gelly, G. and Gauvain, J.-L. (2018). Optimization of RNN-Based Speech Activity Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(3) :646–656. Conference Name : IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[Giraud et al., 2013] Giraud, T., Soury, M., Hua, J., Delaborde, A., Tahon, M., Gómez Jáuregui, D. A., Eyharabide, V., Filaire, E., Le Scanff, C., Devillers, L., Isableu, B., and Martin,

J.-C. (2013). Multimodal Expressions of Stress during a Public Speaking Task : Collection, Annotation and Global Analyses. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013)*, 2013 Humaine Association Conference on effective Computing and Intelligent Interaction (ACII)„ Genève, Switzerland.

[Giraudel et al., 2012] Giraudel, A., Carre, M., Mapelli, V., Kahn, J., Galibert, O., and Quintard, L. (2012). The REPERE Corpus : a multimodal corpus for person recognition. In *LREC*, page 7, Istanbul, Turkey.

[Goldberg, 1990] Goldberg, J. A. (1990). Interrupting the discourse on interruptions : An analysis in terms of relationally neutral, power- and rapport-oriented acts. *Journal of Pragmatics*, 14(6) :883–903.

[González Hautamäki et al., 2019] González Hautamäki, R., Hautamäki, V., and Kinnunen, T. (2019). On the limits of automatic speaker verification : Explaining degraded recognizer scores through acoustic changes resulting from voice disguise. *The Journal of the Acoustical Society of America*, 146(1) :693–704.

[Gravano and Hirschberg, 2012] Gravano, A. and Hirschberg, J. (2012). A corpus-based study of interruptions in spoken dialogue. In *Interspeech 2012*, pages 855–858. ISCA.

[Gravier et al., 2012] Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *LREC - Eighth international conference on Language Resources and Evaluation*, page na, Turkey.

[Griffin and Lim, 1984] Griffin, D. W. and Lim, J. S. (1984). Signal estimation from modified short-time Fourrier transform. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 32(2) :236–243.

[Guennec, 2016] Guennec, D. (2016). *Study of unit selection text-to-speech synthesis algorithms*. Theses, Université Rennes 1. Issue : 2016REN1S055.

[Guennec and Lolive, 2014] Guennec, D. and Lolive, D. (2014). Unit Selection Cost Function Exploration Using an A* based Text-to-Speech System. In *International Conference on Text, Speech and Dialogue (TSD)*.

[Gunes and Schuller, 2013] Gunes, H. and Schuller, B. (2013). Categorical and dimensional affect analysis in continuous input : Current trends and future directions. *Image and Vision Computing*, 31(2) :120–136.

[Haigh and Mason, 1993] Haigh, J. and Mason, J. (1993). Robust voice activity detection using cepstral features. In *Proceedings of TENCON '93. IEEE Region 10 International Confe-*

rence on Computers, Communications and Automation, pages 321–324, Beijing, China. IEEE.

[Hall, 1998] Hall, D. E. (1998). Musical Acoustics : An Introduction. *The Journal of the Acoustical Society of America*, 69(4) :1232–1233. Publisher : Acoustical Society of America.

[Han et al., 2017] Han, J., Zhang, Z., Schmitt, M., Pantic, M., and Schuller, B. (2017). From Hard to Soft : Towards more Human-like Emotion Recognition by Modelling the Perception Uncertainty. In *Proceedings of the 25th ACM international conference on Multimedia*, MM '17, pages 890–897, New York, NY, USA. Association for Computing Machinery.

[Hazen et al., 2009] Hazen, T. J., Shen, W., and White, C. (2009). Query-by-example spoken term detection using phonetic posteriorgram templates. In *2009 IEEE Workshop on Automatic Speech Recognition Understanding*, pages 421–426.

[He et al., 2016] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. ISSN : 1063-6919.

[Heldner and Edlund, 2010] Heldner, M. and Edlund, J. (2010). Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4) :555–568.

[Hennig et al., 2014] Hennig, S., Chellali, R., and Campbell, N. (2014). The D-ANS Corpus : the Dublin-Autonomous Nervous System Corpus of Biosignal and Multimodal Recordings of Conversational Speech. In *LREC*, page 6, Reykjavick, Iceland.

[Hernandez et al., 2018] Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., and Estève, Y. (2018). TED-LIUM 3 : Twice as Much Data and Corpus Repartition for Experiments on Speaker Adaptation | SpringerLink. In *International Conference on Speech and Computer (SPECOM)*.

[Hirschberg, 2000] Hirschberg, J. (2000). A Corpus-Based Approach to the Study of Speaking Style. In Horne, M., editor, *Prosody : Theory and Experiment : Studies Presented to Gösta Bruce*, Text, Speech and Language Technology, pages 335–350. Springer Netherlands, Dordrecht.

[Ito, 2017] Ito, K. (2017). The LJ Speech Dataset.

[Joseph et al., 2022] Joseph, C., Didirkova, I., Dodane, C., and Schweitzer, C. (2022). Les premières archives de la parole (1890-1940). In *Journées d'Etudes sur la Parole*, Noirmoutiers, France.

[Karanasou et al., 2013] Karanasou, P., Yvon, F., Lavergne, T., and Lamel, L. (2013). Discriminative training of a phoneme confusion model for a dynamic lexicon in ASR. In *Interspeech 2013*, pages 1966–1970. ISCA.

[Kawahara et al., 1999] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction : Possible role of a repetitive structure in sounds. *Speech Communication*, 27(3-4) :187–207.

[Kim et al., 2020] Kim, J., Kong, J., Kim, S., and Yoon, S. (2020). Glow-TTS : A Generative Flow for Text-to-Speech via Monotonic Alignment Search. *NeurIPS*, page 11.

[Kinnunen and Li, 2010] Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition : From features to supervectors. *Speech Communication*, 52(1) :12–40.

[Klie et al., 2018] Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E., and Gurevych, I. (2018). The INCEpTION Platform : Machine-Assisted and Knowledge-Oriented Interactive Annotation. In *CoLing*, pages 5–9, Santa Fe, New Mexico, USA.

[Kong et al., 2020] Kong, J., Kim, J., and Bae, J. (2020). HiFi-GAN : Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. *arXiv :2010.05646 [cs, eess]*. arXiv : 2010.05646.

[Kossaifi et al., 2021] Kossaifi, J., Walecki, R., Panagakis, Y., Shen, J., Schmitt, M., Ringeval, F., Han, J., Pandit, V., Toisoul, A., Schuller, B., Star, K., Hajiyev, E., and Pantic, M. (2021). SEWA DB : A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3) :1022–1040. Conference Name : IEEE Transactions on Pattern Analysis and Machine Intelligence.

[Koutsogiannaki et al., 2014] Koutsogiannaki, M., Simantiraki, O., Degottex, G., and Stylianou, Y. (2014). The importance of phase on voice quality assessment. In *Interspeech 2014*, pages 1653–1657. ISCA.

[Kulkarni, 2022] Kulkarni, A. (2022). *Expressivity transfer in deep learning based text-to-speech synthesis*. PhD thesis, Université de Lorraine, Nancy, France.

[Lacheret et al., 2013] Lacheret, A., Simon, A. C., Goldman, J.-P., and Avanzi, M. (2013). Prominence perception and accent detection in French : from phonetic processing to grammatical analysis. *Language Sciences*, 39 :95–106.

[Lafferty et al., 2001] Lafferty, J., McCallum, A., and Pereira, F. (2001). Conditional random fields : probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289.

[Lamirel et al., 2016] Lamirel, J.-C., Dugué, N., and Cuxac, P. (2016). New efficient clustering quality indexes. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 3649–3657. ISSN : 2161-4407.

[Larcher, 2009] Larcher, A. (2009). *Modèles acoustiques à structure temporelle renforcée pour la vérification du locuteur embarquée*. Theses, Université d'Avignon. Issue : 2009AVIG0170.

[Larcher, 2018] Larcher, A. (2018). *Modèles acoustiques pour la reconnaissance du locuteur*. Habilitation à diriger des recherches, Université du Mans.

[Larcher et al., 2021] Larcher, A., Mehrish, A., Tahon, M., Meignier, S., Carrive, J., Doukhan, D., Galibert, O., and Evans, N. (2021). SPEAKER EMBEDDINGS FOR DIARIZATION OF BROADCAST DATA IN THE ALLIES CHALLENGE. In *ICASSP*, Toronto, Canada.

[Laukka et al., 2007] Laukka, P., Audibert, N., and Aubergé, V. (2007). Graded structure in vocal expression of emotion : What is meant by "prototypical expressions" ? In *International Workshop on Paralinguistic Speech - between models and data, ParaLing'07*, pages 1–4, Saarbrücken, Germany.

[LeBellour, 2018] LeBellour, F. (2018). Annotation (semi-)automatique de l'expressivité et de l'émotion dans la parole. Rapport de stage Recherche, Université de Rennes 1 / IRISA.

[Lebourdais et al., 2022a] Lebourdais, M., Tahon, M., Laurent, A., and Meignier, S. (2022a). Overlapped speech and gender detection with WavLM pre-trained features. In *Interspeech 2022*, pages 5010–5014. ISCA.

[Lebourdais et al., 2022b] Lebourdais, M., Tahon, M., Laurent, A., Meignier, S., and Larcher, A. (2022b). Overlaps and Gender Analysis in the Context of Broadcast Media. In *LREC 2022*, Marseille, France.

[Lecorvé and Lolive, 2015] Lecorvé, G. and Lolive, D. (2015). Adaptive Statistical Utterance Phonetization for French. In *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5 p., 2 columns.

[Lemmetty, 1999] Lemmetty, S. (1999). History and Development of Speech Synthesis. Master's thesis, Helsinki University of Technology, Helsinki, Finland.

[Lerdahl and Jackendoff, 1982] Lerdahl, F. and Jackendoff, R. S. (1982). *A Generative Theory of Tonal Music*. MIT Press, Cambridge, MA, USA.

[Lin et al., 2014] Lin, J.-C., Wei, W.-L., Wu, C.-H., and Wang, H.-M. (2014). Emotion recognition of conversational affective speech using temporal course modeling-based error weighted cross-correlation model. In *Signal and Information Processing Association Annual*

Summit and Conference (APSIPA), 2014 Asia-Pacific, pages 1–7, Chiang Mai, Thailand. IEEE.

[Lin et al., 2012] Lin, Y., Michel, J.-B., Aiden, E. L., Orwant, J., Brockman, W., and Petrov, S. (2012). Syntactic Annotations for the Google Books Ngram Corpus. In *50th Annual Meeting of the Association for Computational Linguistics*, pages 169–174, Jeju, Republic of Korea.

[Liu et al., 2018] Liu, S., Bremer, P., Thiagarajan, J. J., Srikumar, V., Wang, B., Livnat, Y., and Pascucci, V. (2018). Visual Exploration of Semantic Relationships in Neural Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 24(1) :553–562. Conference Name : IEEE Transactions on Visualization and Computer Graphics.

[Lolive, 2017] Lolive, D. (2017). *Vers plus de contrôle pour la synthèse de parole expressive*. Habilitation à diriger des recherches, Université de Rennes 1, IRISA.

[Long et al., 2018] Long, Y., Ye, H., Li, Y., and Liang, J. (2018). Active Learning for LF-MMI Trained Neural Networks in ASR. In *Proc. Interspeech 2018*, pages 2898–2902.

[Lotfian and Busso, 2017] Lotfian, R. and Busso, C. (2017). Formulating emotion perception as a probabilistic model with application to categorical emotion classification. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 415–420, San Antonio, TX. IEEE.

[Lotfian and Busso, 2019] Lotfian, R. and Busso, C. (2019). Building Naturalistic Emotionally Balanced Speech Corpus by Retrieving Emotional Speech from Existing Podcast Recordings. *IEEE Transactions on Affective Computing*, 10(4) :471–483.

[Lu et al., 2021] Lu, C., Wen, X., Liu, R., and Chen, X. (2021). Multi-Speaker Emotional Speech Synthesis with Fine-Grained Prosody Modeling. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5729–5733. ISSN : 2379-190X.

[Macary, 2022] Macary, M. (2022). *Analyse de données massives en temps réel pour l'extraction d'informations sémantiques et émotionnelles de la parole*. PhD thesis, Le Mans Université.

[Macary et al., 2020a] Macary, M., Lebourdais, M., Tahon, M., Estève, Y., and Rousseau, A. (2020a). Multi-corpus Experiment on Continuous Speech Emotion Recognition : Convolution or Recurrence ? In Karpov, A. and Potapova, R., editors, *Speech and Computer*, Lecture Notes in Computer Science, pages 304–314, Cham. Springer International Publishing.

[Macary et al., 2022] Macary, M., Tahon, M., Estève, Y., and Luzzati, D. (2022). Mutual impact of acoustic and linguistic representations for continuous emotion recognition in call-center conversations.

[Macary et al., 2020b] Macary, M., Tahon, M., Estève, Y., and Rousseau, A. (2020b). AlloSat : A New Call Center French Corpus for Satisfaction and Frustration Analysis. In *Language Resources and Evaluation Conference, LREC 2020*, Marseille, France.

[Macary et al., 2021] Macary, M., Tahon, M., Estève, Y., and Rousseau, A. (2021). On the use of Self-supervised Pre-trained Acoustic and Linguistic Features for Continuous Speech Emotion Recognition. In *IEEE Spoken Language Technology Workshop*, Virtual, China.

[Maclay and Osgood, 1959] Maclay, H. and Osgood, C. E. (1959). Hesitation Phenomena in Spontaneous English Speech. *WORD*, 15(1) :19–44. Publisher : Routledge _eprint : https ://doi.org/10.1080/00437956.1959.11659682.

[Magne et al., 2005] Magne, C., ARAMAKI, M., Astesano, C., Gordon, R. L., Ystad, S., Farner, S., Kronland-Martinet, R., and Besson, M. (2005). Comparison of Rhythmic processing in Language and Music : An interdisciplinary approach. *Journal of Music and Meaning*, 3. Publisher : University of Southern Denmark.

[Maiti and Mandel, 2020] Maiti, S. and Mandel, M. I. (2020). Speaker Independence of Neural Vocoders and Their Effect on Parametric Resynthesis Speech Enhancement. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 206–210, Barcelona, Spain. IEEE.

[Marinelli et al., 2019] Marinelli, F., Cervone, A., Tortoreto, G., Stepanov, E. A., Fabbrizio, G. D., and Riccardi, G. (2019). Active Annotation : Bootstrapping Annotation Lexicon and Guidelines for Supervised NLU Learning. In *Interspeech 2019*, pages 574–578. ISCA.

[Mariooryad and Busso, 2015] Mariooryad, S. and Busso, C. (2015). Correcting Time-Continuous Emotional Labels by Modeling the Reaction Lag of Evaluators. *IEEE Transactions in Affective Computing*, 6(2) :97–108.

[Mariotte et al., 2022] Mariotte, T., Larcher, A., Thomas, J.-H., and Montrésor, S. (2022). Traitement multi-microphone pour la segmentation automatique de la parole en réunion. In *Congrès Français d'Acoustique*, page 8, Marseille, France.

[Martin et al., 2020] Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, E., Seddah, D., and Sagot, B. (2020). CamemBERT : a Tasty French Language Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

[Matravers, 2001] Matravers, D. (2001). *Art and Emotion*. Oxford University Press, Oxford, New York.

[McCowan et al., 2006] McCowan, I., Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kraaij, W., Kadlec, J., Karaiskos, V., Kronenthal, M., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2006). The AMI Meeting Corpus : A Pre-announcement. In Renals, S. and Bengio, S., editors, *Machine Learning for Multimodal Interaction*, Lecture Notes in Computer Science, pages 28–39, Berlin, Heidelberg. Springer.

[McFee and Ellis, 2014] McFee, B. and Ellis, D. P. W. (2014). ANALYZING SONG STRUCTURE WITH SPECTRAL CLUSTERING. In *International Society for Music Information Retrieval*, page 6.

[McFee et al., 2015] McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O. (2015). LibROSA : Audio and Music Signal Analysis in Python. In *Proc. of the 14th Python in Science Conference*.

[Merritt et al., 2016] Merritt, T., Clark, R. A. J., Wu, Z., Yamagishi, J., and King, S. (2016). Deep neural network-guided unit selection synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5145–5149, Shanghai. IEEE.

[Mertens, 2004] Mertens, P. (2004). The Prosogram : Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model. In *Speech Prosody Workshop*, page 4, Nara, Japan.

[Meunier, 2014] Meunier, C. (2014). *VARIATION DE LA PAROLE : CONTRAINTES LINGUISTIQUES ET MECANISMES D'ADAPTATION*. Habilitation à diriger des recherches, Université Lumière - lyon2.

[Mikolov et al., 2013] Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv :1301.3781 [cs].

[Moattar and Homayounpour, 2012] Moattar, M. H. and Homayounpour, M. M. (2012). A review on speaker diarization systems and approaches. *Speech Communication*, 54(10) :1065–1103.

[Montaño and Alías, 2016] Montaño, R. and Alías, F. (2016). The role of prosody and voice quality in indirect storytelling speech : Annotation methodology and expressive categories. *Speech Communication*, 85 :8–18.

[Montrésor et al., 2020] Montrésor, S., Tahon, M., Laurent, A., and Picart, P. (2020). Computational de-noising based on deep learning for phase data in digital holographic interferometry. *APL Photonics*, 5(3). Publisher : AIP Publishing LLC.

[Morise et al., 2016] Morise, M., Yokomori, F., and Ozawa, K. (2016). WORLD : A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99.D(7) :1877–1884.

[Nagrani et al., 2020] Nagrani, A., Chung, J. S., Xie, W., and Zisserman, A. (2020). Voxceleb : Large-scale speaker verification in the wild. *Computer Speech & Language*, 60 :101027.

[Ning et al., 2006] Ning, H., Liu, M., Tang, H., and Huang, T. (2006). A Spectral Clustering Approach to Speaker Diarization. In *ICSLP (Satellite of Interspeech)*, Pittsburgh, USA.

[Nourtel et al., 2021] Nourtel, H., Champion, P., Jouvet, D., Larcher, A., and Tahon, M. (2021). Evaluation of Speaker Anonymization on Emotional Speech. In *SPSC 2021 - 1st ISCA Symposium on Security and Privacy in Speech Communication*.

[Noé et al., 2022] Noé, P.-G., Nautsch, A., Matrouf, D., Bousquet, P.-M., and Bonastre, J.-F. (2022). A bridge between features and evidence for binary attribute-driven perfect privacy. In *ICASSP 2022*, Singapore, Singapore.

[Oord et al., 2016] Oord, A. v. d., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., and Kavukcuoglu, K. (2016). WaveNet : A Generative Model for Raw Audio. *arXiv :1609.03499 [cs]*. arXiv : 1609.03499.

[Ortiz Suárez et al., 2019] Ortiz Suárez, P. J., Sagot, B., and Romary, L. (2019). Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Challenges in the Management of Large Corpora (CMLC-7)*.

[Pallaud and Henry, 2004] Pallaud, B. and Henry, S. (2004). Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé. In *7es Journées Internationales d'Analyse Statistique des Données Textuelles*, pages 848–858, Louvain-la-Neuve, Belgique.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech : An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, South Brisbane, Queensland, Australia. IEEE.

[Pappagari et al., 2020] Pappagari, R., Wang, T., Villalba, J., Chen, N., and Dehak, N. (2020). X-Vectors Meet Emotions : A Study On Dependencies Between Emotion and Speaker Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7169–7173. ISSN : 2379-190X.

[Park et al., 2020] Park, T. J., Han, K. J., Kumar, M., and Narayanan, S. (2020). Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap. *IEEE Signal Processing Letters*, 27 :381–385. Conference Name : IEEE Signal Processing Letters.

[Parthasarathy et al., 2017] Parthasarathy, S., Zhang, C., Hansen, J. H., and Busso, C. (2017). A study of speaker verification performance with expressive speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5540–5544, New Orleans, LA. IEEE.

[Pasti et al., 2020] Pasti, R., Vilasbôas, F. G., Roque, I. R., and de Castro, L. N. (2020). A Sensitivity and Performance Analysis of Word2Vec Applied to Emotion State Classification Using a Deep Neural Architecture. In Herrera, F., Matsui, K., and Rodríguez-González, S., editors, *Distributed Computing and Artificial Intelligence, 16th International Conference*, Advances in Intelligent Systems and Computing, pages 199–206, Cham. Springer International Publishing.

[Peddinti et al., 2015] Peddinti, V., Povey, D., and Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech 2015*, pages 3214–3218. ISCA.

[Peeters, 2004] Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. Technical Report 1.0, IRCAM.

[Perraudin et al., 2013] Perraudin, N., Balazs, P., and ndergraard, P. L. S. (2013). A fast Griffin-Lim algorithm. In IEEE, editor, *Workshop on Applications of Signal Processing to Audio and Acoustics*, New Platz, NY, USA.

[Picard, 1997] Picard, R. W. (1997). *Affective Computing*. MIT Press, Cambridge, MA, USA.

[Ping et al., 2018a] Ping, W., Peng, K., and Chen, J. (2018a). ClariNet : Parallel Wave Generation in End-to-End Text-to-Speech. In *arXiv :1807.07281*.

[Ping et al., 2018b] Ping, W., Peng, K., Gibiansky, A., Arik, S. O., Kannan, A., Narang, S., Raiman, J., and Miller, J. (2018b). Deep Voice 3 : Scaling Text-to-Speech with Convolutional Sequence Learning. In *International Confernce on Learning Representations (ICLR)*.

[Plutchik and Kellerman, 1980] Plutchik, R. and Kellerman, H. (1980). *Theories of Emotion - 1st Edition*, volume 1. Academic Press.

[Pon-Barry and Shieber, 2011] Pon-Barry, H. and Shieber, S. M. (2011). Recognizing Uncertainty in Speech. *EURASIP Journal on Advances in Signal Processing*, 2011(1) :251753.

[Pratap et al., 2020] Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). MLS : A Large-Scale Multilingual Dataset for Speech Research. In *Interspeech 2020*, pages 2757–2761. ISCA.

[Prenger et al., 2019] Prenger, R., Valle, R., and Catanzaro, B. (2019). Waveglow : A Flow-based Generative Network for Speech Synthesis. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621. ISSN : 2379-190X.

[Prince and Elder, 2007] Prince, S. J. and Elder, J. H. (2007). Probabilistic Linear Discriminant Analysis for Inferences About Identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil. IEEE.

[Prokopalo et al., 2020] Prokopalo, Y., Meignier, S., Galibert, O., Barrault, L., and Larcher, A. (2020). Évaluation de systèmes apprenant tout au long de la vie. In Benzitoun, C., Braud, C., Huber, L., Langlois, D., Ouni, S., Pogodalla, S., and Schneider, S., editors, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, pages 516–524, Nancy, France. ATALA.

[Prokopalo et al., 2022] Prokopalo, Y., Shamsi, M., Barrault, L., Meignier, S., and Larcher, A. (2022). Active Correction for Incremental Speaker Diarization of a Collection with Human in the Loop. *Applied Sciences*. Publisher : MDPI.

[Qader et al., 2015] Qader, R., Lecorvé, G., Lolive, D., and Sébillot, P. (2015). Probabilistic Speaker Pronunciation Adaptation for Spontaneous Speech Synthesis Using Linguistic Features. In *International Conference on Statistical Language and Speech Processing (SLSP)*, Budapest, Hungary.

[Qader et al., 2017] Qader, R., Lecorvé, G., Lolive, D., Tahon, M., and Sébillot, P. (2017). Statistical Pronunciation Adaptation for Spontaneous Speech Synthesis. In *TSD*, Pragua, Czech Republic.

[Qian et al., 2020] Qian, K., Zhang, Y., Chang, S., Hasegawa-Johnson, M., and Cox, D. (2020). Unsupervised Speech Decomposition via Triple Information Bottleneck. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7836–7846. PMLR. ISSN : 2640-3498.

[Rabiner and Schafer, 2007] Rabiner, L. R. and Schafer, R. W. (2007). Introduction to Digital Speech Processing. *Foundations and Trends® in Signal Processing*, 1(1–2) :1–194.

[Ravanelli and Bengio, 2018] Ravanelli, M. and Bengio, Y. (2018). Speaker Recognition from Raw Waveform with SincNet. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 1021–1028.

[Ren et al., 2019] Ren, Y., Ruan, Y., Tan, X., Qin, T., Zhao, S., Zhao, Z., and Liu, T.-Y. (2019). FastSpeech : Fast, Robust and Controllable Text to Speech. *NeurIPS*, page 10.

[Reynolds, 1995] Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 17(1) :91–108.

[Reynolds et al., 2000] Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. *Digital Signal Processing*, 10(1-3) :19–41.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why Should I Trust You ?" : Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

[Ringeval et al., 2019] Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E.-M., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., and Pantic, M. (2019). AVEC 2019 Workshop and Challenge : State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. arXiv : 1907.11510.

[Ringeval et al., 2013] Ringeval, F., Sonderegger, A., Sauer, J., and Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, Shanghai, China. IEEE.

[Rodero Antón and Cores-Sarría, 2021] Rodero Antón, E. and Cores-Sarría, L. (2021). Best prosody for news : a psychophysiological study comparing a broadcast to a narrative speaking style. *Communication Research*, page 34p. Accepted : 2022-06-03T06 :28 :14Z Publisher : SAGE Publications.

[Rousseeuw, 1987] Rousseeuw, P. J. (1987). Silhouettes : A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20 :53–65.

[Russell, 1980] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6) :1161–1178. Place : US Publisher : American Psychological Association.

[Ryant et al., 2021] Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., Du, J., Ganapathy, S., and Liberman, M. (2021). The Third DIHARD Diarization Challenge. In *Interspeech 2021*, pages 3570–3574. ISCA.

[Sahidullah et al., 2019] Sahidullah, M., Patino, J., Cornell, S., Yin, R., Sivasankaran, S., Bredin, H., Korshunov, P., Brutti, A., Serizel, R., Vincent, E., Evans, N., Marcel, S., Squartini, S., and Barras, C. (2019). The Speed Submission to DIHARD II : Contributions & Lessons Learned. arXiv :1911.02388 [cs, eess].

[Scherer, 1986] Scherer, K. R. (1986). Vocal affect expression : A review and a model for future research. *Psychological Bulletin*, 99(2) :143–165. Place : US Publisher : American Psychological Association.

[Scherer, 1994] Scherer, K. R. (1994). Affect Bursts. In *Emotions*. Psychology Press. Num Pages : 34.

[Scherer, 2003] Scherer, K. R. (2003). Vocal communication of emotion : A review of research paradigms. *Speech Communication*, 40(1) :227–256.

[Scherer, 2005] Scherer, K. R. (2005). What are emotions ? and how can they be measured ? *Social Science Information*, 44(4) :695–729.

[Schmitt et al., 2019] Schmitt, M., Cummins, N., and Schuller, B. W. (2019). Continuous Emotion Recognition in Speech — Do We Need Recurrence ? In *Interspeech 2019*, pages 2808–2812. ISCA.

[Schneider et al., 2019] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec : Unsupervised Pre-Training for Speech Recognition. In *Interspeech 2019*, pages 3465–3469. ISCA.

[Schröder, 2003] Schröder, M. (2003). Experimental study of affect bursts. *Speech Communication*, 40(1-2) :99–116.

[Schuller et al., 2009] Schuller, B., Steidl, S., and Batliner, A. (2009). The interspeech 2009 emotion challenge. In *Interpseech*, pages 312–315, Brighton, UK.

[Schuller et al., 2013] Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., Chetouani, M., Weninger, F., Eyben, F., Marchi, E., Mortillaro, M., Salamin, H., Polychroniou, A., Valente, F., and Kim, S. (2013). The INTERSPEECH 2013 computational paralinguistics challenge : social signals, conflict, emotion, autism. In *Interspeech 2013*, pages 148–152. ISCA.

[Schuller et al., 2010] Schuller, B., Zaccarelli, R., Rollet, N., and Devillers, L. (2010). CINEMO - A French Spoken Language Resource for Complex Emotions : Facts and Baselines. In *LREC*, Valetta, Malta.

[Schuller, 2018] Schuller, B. W. (2018). Speech emotion recognition : two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5) :90–99.

[Settles, 2009] Settles, B. (2009). Active Learning Literature Survey. Technical report, University of Wisconsin-Madison.

[Shamsi et al., 2022] Shamsi, M., Larcher, A., Barrault, L., Meignier, S., Prokopalo, Y., Tahon, M., Mehrish, A., Petitrenaud, S., Galibert, O., Gaist, S., Anjos, A., Marcel, S., and Costa-Jussà, M. R. (2022). Towards Lifelong Human Assisted Speaker Diarization. *Computer Speech and Language*. Publisher : Elsevier.

[Shen et al., 2018] Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., and Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *arXiv :1712.05884 [cs]*. arXiv : 1712.05884.

[Shriberg, 2007] Shriberg, E. (2007). Higher-Level Features in Speaker Recognition | SpringerLink. In *Speaker Classification*, number 4343 in Lecture Notes in Computer Science. Springer edition.

[Shriberg, 1999] Shriberg, E. E. (1999). PHONETIC CONSEQUENCES OF SPEECH DISFLUENCY. In *ICPhS*, pages 619–622, San Fransisco, USA.

[Silverman et al., 1992] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Pirce, P., Pierrehumbert, J., and Hirschberg, J. (1992). TOBI : a standard for labeling English prosody. In *Proc. 2nd International Conference on Spoken Language Processing (ICSLP)*, pages 867–870.

[Sini, 2020] Sini, A. (2020). *Characterisation and generation of expressivity in function of speaking styles for audiobook synthesis*. phdthesis, Université Rennes 1.

[Sini et al., 2018] Sini, A., Lolive, D., Vidal, G., Tahon, M., and Delais-Roussarie, E. (2018). SynPaFlex-Corpus : An Expressive French Audiobooks Corpus Dedicated to Expressive Speech Synthesis. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

[Sitchet and Tahon, 2016] Sitchet, P.-E. and Tahon, M. (2016). LE VOISEMENT DANS LE GWOKA : ENTRE LE PARLÉ ET LE CHANTÉ. In *Journées d'Informatique Musicale (JIM)*, Albi, France.

[Skerry-Ryan et al., 2018] Skerry-Ryan, R., Battenberg, E., Xiao, Y., Wang, Y., Stanton, D., Shor, J., Weiss, R., Clark, R., and Saurous, R. A. (2018). Towards End-to-End Prosody Transfer for Expressive Speech Synthesis with Tacotron. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4693–4702, Stockholmsmässan, Stockholm Sweden. PMLR.

[Snyder et al., 2017] Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). Deep Neural Network Embeddings for Text-Independent Speaker Verification. In *Interspeech 2017*, pages 999–1003. ISCA.

[Snyder et al., 2018] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018). X-VECTORS : ROBUST DNN EMBEDDINGS FOR SPEAKER RECOGNITION. In *ICASSP*.

[Solak, 2019] Solak, I. (2019). The M-AILABS Speech Dataset – caito.

[Sridhar et al., 2021] Sridhar, K., Lin, W.-C., and Busso, C. (2021). Generative Approach Using Soft-Labels to Learn Uncertainty in Predicting Emotional Attributes. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Nara, Japan. IEEE.

[Strobelt et al., 2019] Strobelt, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H., and Rush, A. M. (2019). Seq2seq-Vis : A Visual Debugging Tool for Sequence-to-Sequence Models. *IEEE Transactions on Visualization and Computer Graphics*, 25(1) :353–363. Conference Name : IEEE Transactions on Visualization and Computer Graphics.

[Ståhl et al., 2005] Ståhl, A., Sundström, P., and Höök, K. (2005). A Foundation for Emotional Expressivity. In *DUX, Designing for User Experience*, page 21.

[Sundberg, 1995] Sundberg, J. (1995). Le chant, les instruments de l'orchestre (Préfacé par J.C. Risset). *Pour la Science*.

[Tahon, 2012] Tahon, M. (2012). *Analyse acoustique de la voix émotionnelle de locuteurs lors d'une interaction humain-robot*. phdthesis, Université Paris Sud - Paris XI.

[Tahon, 2020] Tahon, M. (2020). spectral-clustering-music.

[Tahon et al., 2014] Tahon, M., Bavu, E., Melon, M., and Garcia, A. (2014). Attack transient exploration on sopranino recorder with time-domain Near-Field Acoustic Holography method. In *International Symposium on Musical Acoustics (ISMA)*, Le Mans, France.

[Tahon et al., 2019] Tahon, M., Belghith, Z., Chouvel, J.-M., and Michel, P. (2019). SEGMENTATION AUTOMATIQUE DE CATÉGORIES MUSICALES : ÉTUDE EXPLORATOIRE SUR LE FREE JAZZ. In *Journées d'Informatique Musicale*, Bayonne, France.

[Tahon et al., 2012a] Tahon, M., Degottex, G., and Devillers, L. (2012a). Usual voice quality features and glottal features for emotional valence detection. In *Speech Prosody*, page 4, Shanghai, China.

[Tahon et al., 2010] Tahon, M., Delaborde, A., Barras, C., and Devillers, L. (2010). A corpus for identification of speakers and their emotions. In *Language Ressources and Evaluation Conference (LREC)*, Valleta, Malta.

[Tahon et al., 2011] Tahon, M., Delaborde, A., and Devillers, L. (2011). Real-life Emotion Detection from Speech in Human-Robot Interaction : Experiments across Diverse Corpora with Child and Adult Voices. In *Interspeech*, Firenze, Italy.

[Tahon et al., 2012b] Tahon, M., Delaborde, A., and Devillers, L. (2012b). Corpus of Children Voices for Mid-level Markers and Affect Bursts Analysis. In *Language Ressource and Evaluation Conference (LREC)*, Istanbul, Turkey.

[Tahon and Devillers, 2010] Tahon, M. and Devillers, L. (2010). Acoustic measures characterizing anger across corpora collected in artificial or natural context. In *Speech Prosody*, Chicago, United States.

[Tahon and Devillers, 2015] Tahon, M. and Devillers, L. (2015). LAUGHTER DETECTION FOR ON-LINE HUMAN-ROBOT INTERACTION. In *Interdisciplinary Workshop on Laughter and Non-verbal Vocalisations in Speech*, Enschede, Netherlands.

[Tahon and Devillers, 2016a] Tahon, M. and Devillers, L. (2016a). Towards a Small Set of Robust Acoustic Features for Emotion Recognition : Challenges. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(1) :16–28. Conference Name : IEEE/ACM Transactions on Audio, Speech, and Language Processing.

[Tahon and Devillers, 2016b] Tahon, M. and Devillers, L. (2016b). Towards a Small Set of Robust Acoustic Features for Emotion Recognition : Challenges. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 24 :16.

[Tahon et al., 2018] Tahon, M., Lecorvé, G., and Lolive, D. (2018). Can we Generate Emotional Pronunciations for Expressive Speech Synthesis ? *IEEE Transactions on Affective Computing*, 11(4) :684–695. Publisher : Institute of Electrical and Electronics Engineers.

[Tahon et al., 2017] Tahon, M., Lecorvé, G., Lolive, D., and Qader, R. (2017). Perception of expressivity in TTS : linguistics, phonetics or prosody ? In *Statistical Language and Speech Processing (SLSP)*, volume 10583, page 262.

[Tahon and Lolive, 2018] Tahon, M. and Lolive, D. (2018). Discourse phrases classification : direct vs. narrative audio speech. In *Speech Prosody*, Poznan, Poland.

[Tahon et al., 2022] Tahon, M., Montrésor, S., and Picart, P. (2022). Deep Learning Network for Speckle De-Noising in Severe Conditions. *Journal of Imaging*, 8(6) :165. Publisher : MDPI.

[Tahon et al., 2021] Tahon, M., Picart, P., and Montresor, S. (2021). Towards Reduced CNNs for De-Noising Phase Images Corrupted with Speckle Noise. *Photonics*, 8(7). Publisher : MDPI.

[Tahon et al., 2016a] Tahon, M., Qader, R., Lecorvé, G., and Lolive, D. (2016a). Improving TTS with corpus-specific pronunciation adaptation. In *Interspeech*, San Fransisco, USA.

[Tahon et al., 2016b] Tahon, M., Qader, R., Lecorvé, G., and Lolive, D. (2016b). Optimal feature set and minimal training size for pronunciation adaptation in TTS. In *International Conference on Statistical Language and Speech Processing (SLSP)*, Pilzen, Czech Republic.

[Tahon et al., 2015] Tahon, M., Sehili, M. A., and Devillers, L. (2015). Cross-Corpus Experiments on Laughter and Emotion Detection in HRI with Elderly People. In Tapus, A., André, E., Martin, J.-C., Ferland, F., and Ammi, M., editors, *Social Robotics*, Lecture Notes in Computer Science, pages 633–642, Cham. Springer International Publishing.

[Tahon and Sitchet, 2016] Tahon, M. and Sitchet, P.-E. (2016). La nasalité dans le répertoire Gwoka de la Guadeloupe. In *Congrès Français d'Acoustique (CFA)*, Le Mans, France.

[Tahon and Sitchet, 2017] Tahon, M. and Sitchet, P.-E. (2017). The transmission of voicing in traditional Gwoka : Between identity and memory. *Journal of Interdisciplinary Voice Studies*, 2(2) :157–175.

[Tan et al., 2021] Tan, D., Deng, L., Yeung, Y. T., Jiang, X., Chen, X., and Lee, T. (2021). EditSpeech : A Text Based Speech Editing System Using Partial Inference and Bidirectional Fusion. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 626–633.

[t'Hart, 1981] t'Hart, J. (1981). Differential sensitivity to pitch distance, particularly in speech. *The Journal of the Acoustical Society of America*, 69(3) :811–821.

[Thiam et al., 2016] Thiam, P., Meudt, S., Schwenker, F., and Palm, G. (2016). Active Learning for Speech Event Detection in HCI. In *Artificial Neural Networks in Pattern Recognition*, pages 285–297.

[Torreira et al., 2010] Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52(3) :201–212.

[Tånnander and Edlund, 2021] Tånnander, C. and Edlund, J. (2021). Methods of slowing down speech. In *11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 43–47. ISCA.

[Vaca et al., 2015] Vaca, M., Mora, E., and Cobeta, I. (2015). The Aging Voice : Influence of Respiratory and Laryngeal Changes. *Otolaryngology–Head and Neck Surgery*, 153(3) :409–413. Publisher : SAGE Publications Inc.

[Valle et al., 2020] Valle, R., Li, J., Prenger, R., and Catanzaro, B. (2020). Mellotron : Multispeaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6189–6193. ISSN : 2379-190X.

[Vinciarelli et al., 2009] Vinciarelli, A., Salamin, H., and Pantic, M. (2009). Social Signal Processing : Understanding social interactions through nonverbal behavior analysis. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 42–49. ISSN : 2160-7516.

[Vulliamy et al., 1982] Vulliamy, G., Josephs, N. A., Holt, G., and Horn, D. (1982). The New Grove Dictionary of Music and Musicians. *Popular Music*, 2 :245–258.

[Wang et al., 2018] Wang, Y., Stanton, D., Zhang, Y., Skerry-Ryan, R. J., Battenberg, E., Shor, J., Xiao, Y., Ren, F., Jia, Y., and Saurous, R. A. (2018). Style Tokens : Unsupervised Style Modeling, Control and Transfer in End-to-End Speech Synthesis. *arXiv :1803.09017 [cs, eess]*. arXiv : 1803.09017.

[Wolfe et al., 2016] Wolfe, J., Tze Wei Chu, D., Chen, J.-M., and Smith, J. (2016). An Experimentally Measured Source–Filter Model : Glottal Flow, Vocal Tract Gain and Output Sound from a Physical Model. *Acoustics Australia*, 44(1) :187–191.

[Wottawa et al., 2020] Wottawa, J., Tahon, M., Marin, A., and Audibert, N. (2020). Towards Interactive Annotation for Hesitation in Conversational Speech. In *LREC 2020*, Marseille, France.

[Wundt, 1897] Wundt, W. M. (1897). *Outlines of psychology*. Leipzig, W. Engelmann ; New York, G.E. Stechert.

[Wöllmer et al., 2008] Wöllmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., and Cowie, R. (2008). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. In *Interspeech 2008*, pages 597–600. ISCA.

[Yamagishi et al., 2012] Yamagishi, J., Veaux, C., and MacDonald, K. (2012). CSTR VCTK Corpus : English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92). *The Rainbow Passage which the speakers read out can be found in the International Dialects of English Archive : (http ://web.ku.edu/~idea/readings/rainbow.htm).* Accepted : 2019-11-13T17 :09 :33Z Publisher : University of Edinburgh. The Centre for Speech Technology Research (CSTR).

[Yan et al., 2010] Yan, Z.-J., Qian, Y., and Soong, F. K. (2010). RIch-context Unit Selection (RUS) approach to high quality TTS. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4798–4801. ISSN : 2379-190X.

[Yang et al., 2022] Yang, Y.-Y., Hira, M., Ni, Z., Astafurov, A., Chen, C., Puhrsch, C., Pollack, D., Genzel, D., Greenberg, D., Yang, E. Z., Lian, J., Hwang, J., Chen, J., Goldsborough, P., Narenthiran, S., Watanabe, S., Chintala, S., and Quenneville-Bélair, V. (2022). Torchaudio : Building Blocks for Audio and Speech Processing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6982–6986. ISSN : 2379-190X.

[Yin et al., 2017] Yin, R., Bredin, H., and Barras, C. (2017). Speaker Change Detection in Broadcast TV Using Bidirectional Long Short-Term Memory Networks. In *Interspeech 2017*, pages 3827–3831. ISCA.

[Zen et al., 2019] Zen, H., Dang, V., Clark, R., Zhang, Y., Weiss, R. J., Jia, Y., Chen, Z., and Wu, Y. (2019). LibriTTS : A Corpus Derived from LibriSpeech for Text-to-Speech. In *Interspeech 2019*, pages 1526–1530. ISCA.

[Zen et al., 2007] Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. W., and Tokuda, K. (2007). The HMM-based speech synthesis system (HTS version 2.0). In *Speech Synthesis Workshop (SSW)*, pages 294–299.

[Zewoudie et al., 2018] Zewoudie, A. W., Luque, J., and Hernando, J. (2018). The use of long-term features for GMM- and i-vector-based speaker diarization systems. *EURASIP Journal on Audio, Speech, and Music Processing*, 2018(1) :14.

[Zhang et al., 2013] Zhang, Z., Deng, J., Marchi, E., Schuller, B., and München, T. U. (2013). Active Learning by Label Uncertainty for Acoustic Emotion Recognition. In *Interspeech*.

[Zheng et al., 2020] Zheng, L., Tao, J., Wen, Z., and Zhong, R. (2020). CASIA Voice Conversion System for the Voice Conversion Challenge 2020. In *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pages 136–139. ISCA.

[Zheng et al., 2022] Zheng, N., Li, N., Yu, J., Weng, C., Su, D., Liu, X., and Meng, H. (2022). Multi-Channel Speaker Diarization Using Spatial Features for Meetings. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7337–7341. ISSN : 2379-190X.

[Zhou et al., 2020] Zhou, X., Ling, Z.-H., and King, S. (2020). The Blizzard Challenge 2020. In *Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, pages 1–18. ISCA.

TABLE DES FIGURES

1 Chronologie de mes domaines de recherche en fonction des postes occupés. . . 8

1.1 Caractéristiques temporelles des catégories émotionnelles. Figure extraite de [Cowie et al., 2001]. 15
1.2 Appareil vocal vue d'ensemble (gauche) et vue du larynx (droite) [Sundberg, 1995]. 19
1.3 Schéma de la vibration des cordes vocales sous l'effet d'une force de Bernoulli. Les cordes vocales sont modélisées par un système masse-ressort [Hall, 1998]. . 20
1.4 Forme d'onde créée au niveau du larynx. Débit d'air à travers les cordes vocales (gauche) et spectre harmonique de ce débit périodique (droite) [Wolfe et al., 2016]. 20

2.1 Représentation des différentes unités rythmiques pour les signaux audio (parole et musique). 28
2.2 De haut en bas (gauche) : spectre linéaire X_k, spectre logarithmique $\hat{X}_k$, cepstre c_n [Taylor] et chaîne de traitement associée (droite) 30

3.1 Réseau récurrent pour la prédiction de parole superposée [Lebourdais et al., 2022a] 52
3.2 Illustration de la prise de décision avec un argmax (haut) et des seuils (bas) . 53
3.3 Illustration de possibles biais dans les métriques précision et rappel en segmentation. 54
3.4 Distribution normalisée des durées de segments de parole superposée dans différents corpus de parole. Figure extraite de [Lebourdais et al., 2022b] 57
3.5 Architecture du DNN de type x-vecteur. Figure extraite de [Snyder et al., 2017]. 61
3.6 Réseau récurrent pour la prédiction du genre [Lebourdais et al., 2022a] 65
3.7 Détection du genre sur de la parole émotionnelle en erreur pondérée 67
3.8 Représentation des matrices de similarité à long terme R et à court terme S, ainsi que le graphe combiné A pour un extrait musical de Free Jazz. 73
3.9 De haut en bas : représentation du spectrogramme à Q constant, la segmentation obtenue avec $K = 4$ puis $K = 6$ pour un extrait de la sonate K545 pour piano de Mozart. La segmentation en motifs réalisée par un musicologue, est donnée en blanc sur la représentation temps-fréquence. 74

3.10 De haut en bas : représentation du mel-spectrogramme, la segmentation obtenue avec $K = 6$ puis $K = 8$ pour un extrait d'une émission TV. 75

4.1 Pictogrammes pour l'annotation de ses propres émotions. Figure extraite de [Bradley and Lang, 19

4.2 Visualisation des trois annotations continues sur une conversation du corpus AlloSat. Figure extraite de [Macary et al., 2020b]. 84

4.3 Quantité de segments de parole superposée où aucun annotateur n'est d'accord, 2 annotateurs sont d'accord (bleu), tous les annotateurs sont d'accord (orange) pour les 5 catégories considérées. Le total des segments considérés est 4351. . 86

4.4 Exemple d'une conversation du corpus AlloSat. La visualisation sous Praat permet de voir l'alignement entre les mots issus d'une transcription automatique et les segments émotionnels t_i d'une durée de 250 ms obtenus après l'annotation continue. 93

4.5 Réseau de neurone récurrent pour la prédiction continue d'une dimension affective. Figure adaptée de [Macary et al., 2022] 95

4.6 Matrice de confusion obtenue sur l'ensemble de test des données annotées de SynPaFlex avec un modèle SVM. Figure extraite du rapport de F. Le Bellour [LeBellour, 2018] 102

4.7 Simulation d'un apprentissage actif évalué par un f-score avec différentes fonctions d'incertitude. Figure extraite du rapport de F. Le Bellour [LeBellour, 2018]103

4.8 RMSE en fonction du nombre de données annotées. Simulation d'un apprentissage actif évalué par une RMSE avec deux fonctions d'incertitude basée sur l'approche QBC sur 5 ou 6 modèles. L'évaluation se fait soit sur le dev dont le nombre de segment diminue avec les itérations (pointillés) soit sur le test (traits pleins) qui reste identique. 106

5.1 Visualisation d'une séquence d'unités candidate avec l'outil ROOTS. La figure de droite présente les différents types de séquences qui peuvent être associées aux unités (ici des phonèmes) (extraite de [Chevelu et al., 2014]. La figure de gauche montre un exemple de séquence issue du corpus SynPaFlex. 113

5.2 Schéma de l'architecture du système de TTS Tacotron. Figure adaptée de [Shen et al., 2018]. 116

5.3 Schéma de principe d'un vocodeur basé sur un modèle source filtre. 117

5.4 Schéma pour l'adaptation de la prononciation à la voix de synthèse. 121

5.5 Dépendances et paramètres à traiter pour l'apprentissage des CRFs. Figure extraite de [Lolive, 2017]. 122
5.6 Résultats des tests perceptifs de comparaison AB avec un système de sélection d'unité (a) et HTS (b). Avec un modèle de prononciation incluant des caractéristiques linguistiques et phonologiques, le système adapté (noir) est préféré au système sans adaptation (gris). 126
5.7 Schéma pour l'adaptation de la prononciation à la voix de synthèse. 127
5.8 Similarité cosinus entre les catégories émotionnelles e. La similarité est calculée sur la base des confusions entre deux paires (p_c, p_e) avec p_c la séquence de phonèmes canonique. 130
5.9 Schéma de principe du système SpeechSplit inspiré d'AutoVC où E_r est l'encodeur du rythme, E_c l'encodeur du contenu linguistique et E_f l'encodeur de la $F0$. Figure extraite de [Qian et al., 2020]. 135
5.10 Schéma de principe du module de correction de la synthèse (thèse de Thibault Gaudier). En rouge le module de correction linguistique et phonétique, en vert le module de correction prosodique et expressif. 136

LISTE DES TABLEAUX

3.1 Listes des différents corpus de parole en fonction des conditions d'enregistrement et de l'expressivité présente dans les données. 48
3.2 Résultats issus de [Lebourdais et al., 2022a] obtenus sur la tâche de détection de la parole superposée évaluée en précision, rappel et f1-score (en %) sur la partition d'évaluation du challenge DiHard. 58
3.3 Effet des émotions sur les taux d'erreur pour l'identification du locuteur sur le corpus JEMO. Résultats issus de [Tahon, 2012]. 68

4.1 Ensemble de descripteurs OS384 proposé pour le challenge Interspeech 2009 [Schuller et al., 2009]. 90
4.2 Classification des émotions cross-corpus (apprentissage sur un corpus, test sur un autre). Résultats donnés en UAR moyen sur 4 corpus en français 92
4.3 Résultats CCC pour la tâche de reconnaissance de la satisfaction sur AlloSat. Figure extraite de [Macary et al., 2021] 96
4.4 Résultats CCC de fusion multi-modale pour la tâche de reconnaissance de la satisfaction sur AlloSat. Tableau extrait de [Macary et al., 2022] 99

5.1 Groupes de descripteurs utilisés pour modéliser la prononciation. Entre parenthèse, le nombre de vote obtenus sur les 7 plis, en gras, les descripteurs sélectionnés. Tableau extrait de [Tahon et al., 2016a]. 124
5.2 PER obtenu sur l'ensemble de validation suivant différentes configurations. Entre parenthèses le gain en points de pourcentage. 125
5.3 Prononciation expressive de la phrase surprise *"Qui peut bien m'avoir laissé ce message ? Je ne vois vraiment pas."*. En gris les changements par rapport à la séquence canonique. 129
5.4 Analyse des caractéristiques prosodiques en fonction du corpus. La F_0 est donnée en semiton, et le débit syllabique (SR) en nombre de syllabe par seconde. 132