



Dimensions de variation de la parole spontanée pour l'étude inter-corpus des performances de systèmes de reconnaissance automatique de la parole

Solène Evain

► To cite this version:

Solène Evain. Dimensions de variation de la parole spontanée pour l'étude inter-corpus des performances de systèmes de reconnaissance automatique de la parole. Traitement du signal et de l'image [eess.SP]. Université Grenoble Alpes [2020-..], 2024. Français. NNT : 2024GRALM037 . tel-04984659

HAL Id: tel-04984659

<https://theses.hal.science/tel-04984659v1>

Submitted on 22 May 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES

École doctorale : MSTII - Mathématiques, Sciences et technologies de l'information, Informatique

Spécialité : Informatique

Unité de recherche : Laboratoire d'Informatique de Grenoble

Dimensions de variation de la parole spontanée pour l'étude inter-corpus des performances de systèmes de reconnaissance automatique de la parole

Variation dimensions of spontaneous speech for the inter-corpus study of automatic speech recognition system performance

Présentée par :

Solène EVAIN

Direction de thèse :

François PORTET

PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES

Solange ROSSATO

MAITRESSE DE CONFERENCES, UNIVERSITE GRENOBLE ALPES

Directeur de thèse

Co-encadrante de
thèse

Rapporteurs :

MARTINE ADDA-DECKER

DIRECTRICE DE RECHERCHE, CNRS ILE-DEFRANCE VILLEJUIF

RICHARD DUFOUR

MAITRE DE CONFERENCES HDR, UNIVERSITE DE NANTES

Thèse soutenue publiquement le **8 octobre 2024**, devant le jury composé de :

DOMINIQUE VAUFREYDAZ,

PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES

FRANÇOIS PORTET,

PROFESSEUR DES UNIVERSITES, UNIVERSITE GRENOBLE ALPES

MARTINE ADDA-DECKER,

DIRECTRICE DE RECHERCHE, CNRS ILE-DEFRANCE VILLEJUIF

RICHARD DUFOUR,

MAITRE DE CONFERENCES HDR, UNIVERSITE DE NANTES

CHRISTINE MEUNIER,

DIRECTRICE DE RECHERCHE, CNRS DELEGATION PROVENCE ET CORSE

JULIEN PINQUIER,

PROFESSEUR DES UNIVERSITES, UNIVERSITE TOULOUSE III - PAUL SABATIER

Président

Directeur de thèse

Rapporteure

Rapporteur

Examinateuse

Examinateur

Invités :

SOLANGE ROSSATO

MAITRESSE DE CONFERENCES, UNIVERSITE GRENOBLE ALPES

Résumé

Ces dernières années, la reconnaissance automatique de la parole a beaucoup progressé grâce au développement du *deep learning* et des modèles pré-appris. Toutefois, les performances sur la parole spontanée restent très variables, notamment en fonction des niveaux de spontanéité. Notre recherche vise tout d'abord à déterminer une méthode pour la capture d'ensembles représentatifs de différents niveaux de spontanéité, et pouvoir les situer les uns par rapports aux autres, en fonction des facteurs influençant leur spontanéité relative. Cette méthode ouvrirait la voie à des analyses de données inter-corpus. Nous étudions ensuite l'apport de la détermination de différents sous-types de parole spontanée sur les performances de systèmes de reconnaissance automatique de la parole.

La littérature montre que la spontanéité est influencée par de multiples facteurs tels que la relation entre les locuteurs, leur état émotionnel, la situation dans laquelle l'interaction a lieu, *etc...* Afin de déterminer ces facteurs parmi ceux pris en compte par les linguistes lors de la constitution de corpus oraux, nous avons collecté des données de parole spontanée représentatives de différentes situations, que nous avons analysées à travers le prisme des études en linguistique sur la parole spontanée. Nous avons ainsi pu identifier quatre dimensions majeures de variation de la parole spontanée : la *situation de communication*, le *degré d'intimité entre les locuteurs*, le *canal de communication* et le *type de communication*. Ces dimensions auront permis la création de trois cas d'étude, regroupant des données supposées homogènes et représentatives de situations plus ou moins propices à l'apparition de différents niveaux de spontanéité.

L'apport de ces différents cas d'étude pour l'étude des performances des systèmes est ensuite étudié au travers de différentes adaptations d'un modèle pré-appris sur le français. Tout d'abord, nous testons l'apport d'adaptations spécifiques du modèle à chacun de nos cas sur la reconnaissance automatique de données de ces mêmes cas. Ces différentes adaptations sont effectuées avec une petite quantité de données ($\approx 10h$ à chaque fois). Ensuite, nous étudions l'impact d'une adaptation au domaine spontané, avec une plus grande quantité de données moins contrôlées ($\approx 300h$), sur les mêmes ensembles de test que précédemment.

Notre étude montre, tout d'abord, que l'étiquetage de données en fonction des dimensions proposées permet de montrer la large plage de performance que peut avoir un système de reconnaissance automatique de la parole sur de la parole spontanée. Si le WER moyen obtenu sur nos différents cas avec notre meilleur système est de 29%, le score sur le cas de parole très spontanée (nommé *Usual_close*) est de 51%, sur de la parole moyennement spontanée (*Unusual_close*) de 23% et sur de la parole peu spontanée (*Unusual_distant*) de 13%. La catégorisation de différents

sous-types de parole spontanée dans les données d'évaluation présente donc une approche intéressante pour l'analyse fine des performances d'un système. Par les différents cas d'étude étudiés, nous montrons également que les systèmes semblent avoir plus de difficulté à traiter les situations très spontanées sur la dimension de la *situation de communication* (représentées par *usual/unusual*) que sur la dimension du *degré d'intimité entre les locuteurs* (représentées par *close/distant*). Enfin, nos différentes adaptations montrent un apport très significatif de l'adaptation d'un modèle pré-appris avec une large quantité de données variées, en comparaison à l'adaptation avec de petites quantités de données spécifiques à des sous-types de parole spontanée.

Abstract

In recent years, automatic speech recognition has made significant progress thanks to the development of deep learning and pre-trained models. However, performance on spontaneous speech remains highly variable, particularly depending on levels of spontaneity. Our research aims first to determine a method for capturing representative sets of different levels of spontaneity and to be able to situate them relative to each other, based on factors influencing their relative spontaneity. This method would pave the way for inter-corpus data analyses. We then study the contribution of determining different subtypes of spontaneous speech on the performance of automatic speech recognition systems.

The literature shows that spontaneity is influenced by multiple factors such as the relationship between speakers, their emotional state, the situation in which the interaction takes place, etc. To identify these factors among those considered by linguists when creating oral corpora, we collected representative spontaneous speech data from different situations and analyzed them through the lens of linguistic studies on spontaneous speech. We thus identified four major dimensions of variation in spontaneous speech : the communication situation, the degree of intimacy between speakers, the communication channel, and the type of communication. These dimensions allowed the creation of three case studies, grouping data presumed to be homogeneous and representative of situations more or less conducive to the emergence of different levels of spontaneity.

The contribution of these different case studies to the study of system performance is then examined through various adaptations of a pre-trained model on French. First, we test the contribution of specific model adaptations to each of our cases on the automatic recognition of data from these same cases. These different adaptations are carried out with a small amount of data (approximately 10 hours each time). Then, we study the impact of adapting to the spontaneous domain, with a larger amount of less controlled data (approximately 300 hours), on the same test sets as before.

Our study shows, first, that labeling data based on the proposed dimensions allows us to demonstrate the wide range of performance that an automatic speech recognition system can have on spontaneous speech. While the average WER (Word Error Rate) obtained on our different cases with our best system is 29%, the score on the very spontaneous speech case (named Usual_close) is 51%, on moderately spontaneous speech (Unusual_close) 23%, and on minimally spontaneous speech (Unusual_distant) 13%. Categorizing different subtypes of spontaneous speech in evaluation data thus presents an interesting approach for fine-grained performance analysis of a system. Through the different case studies examined, we also show

that systems seem to have more difficulty handling highly spontaneous situations on the dimension of the communication situation (represented by usual/unusual) than on the dimension of the degree of intimacy between speakers (represented by close/distant). Finally, our various adaptations show a very significant contribution of adapting a pre-trained model with a large amount of varied data, compared to adaptation with small amounts of data specific to subtypes of spontaneous speech.

Remerciements

Quelle aventure que le doctorat... Si quelques années en arrière quelqu'un m'avait dit que j'atteindrais un tel niveau universitaire, je ne l'aurais pas cru une seule seconde. Mais qu'est-ce que c'est passé vite et -surtout- qu'est-ce que c'était bien ! Ces années universitaires auront été marquées aussi bien par ma découverte de la linguistique et plus particulièrement de l'étude de la parole, que par ma découverte du traitement automatique, qui me triturent tous deux si bien l'esprit. Si l'on m'avait toujours dit que le doctorat était une aventure très solitaire, je ne me suis finalement jamais autant sentie entourée et soutenue que ces quatre dernières années. C'est pour cette raison que je suis heureuse de pouvoir réservier une petite partie de ce manuscrit aux remerciements.

Je commencerai par celles et ceux sans qui ce doctorat n'aurait pas été possible. Je remercie ainsi tout d'abord Martine Adda-Decker, Julien Pinquier, Richard Dufour, Christine Meunier et Dominique Vaufreydaz qui ont accepté de faire partie de mon jury. Vous avez très largement contribué à faire de ma soutenance un événement dont je me souviendrai toujours avec plaisir. Merci de m'avoir fait l'honneur de votre présence en personne pour ce grand jour, malgré les nombreuses heures de train que cela aura pu représenter pour les plus éloigné-e-s d'entre vous. Ensuite, et puisque le doctorat n'existerait pas sans des chercheurs et chercheuses prêt-e-s à donner de leur temps pour nous apprendre le métier, je remercie très sincèrement et non sans une certaine émotion Solange Rossato et François Portet. Travailler avec vous aura été un réel plaisir. Je vous remercie pour votre soutien, vos mots rassurants et toute la compréhension que vous avez pu montrer face à certaines de mes angoisses. Vous m'avez aidée à grandir de la meilleure des manières : en me faisant confiance, en me laissant expérimenter et parfois me tromper, et ce sans jamais me juger. Vous avez formé un binôme vraiment complémentaire, et j'ai pu apprendre beaucoup de chacun d'entre vous. Je tiens également à remercier Nathalie Henrich Bernardoni, Didier Schwab et Benjamin Lecouteux pour leur sujet de stage de M2 passionnant, qui m'aura fait mettre un premier pied dans la recherche. Sans vous, sans vos conseils et retours, je crois que je ne me serais jamais autorisée à penser que j'étais en capacité de faire un doctorat. Merci pour votre bienveillance, et je suis sincèrement contente que nous ayons pu rester en contact. Didier, je te serai éternellement reconnaissante de la confiance que tu m'as accordée pour mener de front la rédaction de mon manuscrit et ma mission de "pré-post-doc" sur le projet Pantagruel. Je suis ravie que nous ayons pu retravailler ensemble. Ça a été et ce sera d'ailleurs toujours un plaisir. Merci également à Gilles Sérasset pour son écoute, ses conseils sur le projet, et plus particulièrement sur la gestion des données, la gestion de projet et le versionnage. Merci aussi pour les précieux retours sur ma présentation de soutenance !

J'aimerais ensuite remercier tou-te-s celles et ceux qui ont contribué (ou qui contribuent encore) à faire de mon ancien lieu de travail ce qu'il était. Quel plaisir ça aura été de travailler au sein du LIG, et plus particulièrement au sein de l'équipe GETALP. S'il est difficile de remercier tout le monde, j'aimerais tout de même faire un petit clin d'oeil aux chercheurs et chercheuses permanent-e-s avec qui j'interagissais régulièrement et qui me manqueront particulièrement : Emmanuelle, Marco, Maximim, Didier, François, Gilles. Je garderai de très bons souvenirs des cafés GETALP et des DNN Meetup (mention spéciale au DNN meetup Karaoké). Merci aussi à l'ensemble des doctorants/post-doctorants et ingénieurs de l'équipe pour leur bonne humeur, leur soutien, leur ouverture d'esprit et leur tolérance. Une mention spéciale à Mahault, William, Eric, Hâ, Marcely, Damien, Ange, Yagmur et Fabien, toutes et tous habitant-e-s du bureau 325, LE meilleur bureau ! Rien n'aurait été pareil sans nos fous-rires, nos discussions à rallonge, nos pauses repas/café et nos sessions de nettoyage ou de décoration du bureau. Ensemble on a tout aussi bien pu célébrer nos petites victoires, que se soutenir mutuellement lors de coups durs, et je dois dire que je me sens vraiment chanceuse d'avoir pu être entourée de personnes aussi formidables. Qui sait si on ne se retrouvera pas dans un autre bureau 325 un jour... The rest is still unwritten ! Mention spéciale également à Adrien : merci encore une fois d'avoir toujours répondu avec patience et de façon didactique à mes questions, mais aussi de m'avoir accompagnée dans mes folies Pint of Science et RJCP ! Ce sera toujours avec grand plaisir que je partagerai une bière avec vous tous.

J'aimerais également dire quelques mots sur la communauté parole, que j'ai appris à connaître grâce au CA de l'AFCP, au groupe des jeunes chercheurs en parole et au comité d'organisation des RJCP. Merci à tous, et notamment à Martine Adda-Decker, Nathalie Henrich Bernardoni, Delphine Charuau, Mathias Quillot, Vincent Martin, Andrés Lara et Thibault Bañeras-Roux pour leurs encouragements, leur confiance, et leur motivation à toute épreuve. Je suis très fière de ce que nous avons pu accomplir et je sais déjà que nous n'en resterons pas là ! Et puisque l'associatif aura occupé une grande place dans ma vie doctorale, merci aussi aux copains de Pint of Science pour les super éditions grenobloises que l'on a pu organiser ces dernières années !

Mon parcours scolaire n'aura pas toujours tout à fait suivi une ligne droite, mais c'est la finalité qui compte (il paraît). Je tiens ainsi à remercier très sincèrement les trois enseignantes (Gaëlle Ferré, Corine Astesano et Solange Rossato) grâce auxquelles j'ai pu découvrir la linguistique et plus spécifiquement l'étude de l'interaction et de la parole qui m'ont immédiatement passionné et me passionnent encore. Vous m'avez permis de construire mon parcours scolaire et ensuite professionnel et d'en faire ce qu'il est aujourd'hui, et je ne suis pas sûre que j'aurais

aimé que ça se passe différemment. Avec les années de retard viennent aussi une meilleure connaissance de soi et de ses envies...

Enfin, je ne pourrais pas conclure ces remerciements sans en adresser un tout spécial à celui qui partage ma vie depuis quelques années. Clément, merci de m'avoir écouté râler, m'exalter, douter, stresser, et j'en passe... Tu m'as souvent aidé à prendre du recul et tu as toujours su me rappeler combien il est tout aussi important de savoir prendre soin de soi et se reposer. Bon, j'ai pas toujours appliqué ces conseils, mais je les ai entendus et, promis, je ferai mieux un jour. Merci aussi à ma famille, mes périodes d'exil sous le soleil breton auront toujours été réparateurs. Vous n'avez pas pu être présents pour ma soutenance (quelle idée d'habiter si loin aussi!), mais merci pour toute la force que vous m'avez envoyée les jours qui l'ont précédée. Ça a beaucoup compté pour moi.

Sommaire

Liste des figures	xv
Liste des tableaux	xix
Introduction	xxiii
A Etat de l'art	
I Les systèmes de reconnaissance automatique de la parole	1
1 L'approche statistique	2
1.1 L'extraction de paramètres acoustiques	4
1.2 Le modèle acoustique	5
1.3 Le dictionnaire de prononciation	7
1.4 Le modèle de langue	8
2 Les approches neuronales	8
2.1 Les types de réseaux neuronaux utilisés en reconnaissance automatique de la parole	9
2.1.1 Le fonctionnement général d'un réseau de neurones et réseau de neurones à propagation avant (FNN)	10
2.1.2 Les réseaux de neurones récurrents (RNN)	12
2.1.3 Réseaux à mémoire à court et long terme (LSTM)	13
2.1.4 Réseaux de neurones convolutionnels (CNN)	14
2.2 Les systèmes hybrides HMM-DNN	15
2.3 Les systèmes à base de classification temporelle connectionniste (CTC)	17
2.4 Les systèmes à base de transducteur (RNN-T)	19
2.5 Les systèmes à base d'attention	20
2.5.1 L'architecture encodeur-décodeur	20
2.5.2 L'attention	21

2.6	Les modèles pré-appris	24
2.6.1	Le fonctionnement général des modèles pré-appris	24
2.6.2	Les modèles pré-appris pour le français	26
2.6.3	Les modèles multilingues Whisper	27
2.6.4	L'architecture Wav2Vec 2.0	29
3	L'évaluation des systèmes de reconnaissance automatique de la parole	30
3.1	Le WER et ses limites	30
3.2	Les principaux corpus pour la reconnaissance automatique de la parole en français	32
3.2.1	Les corpus monolingues	32
3.2.2	Les corpus multilingues	34
3.3	Les performances de reconnaissance automatique	35
3.3.1	Quelques résultats généraux sur différents types de parole	35
3.3.2	Résultats sur la parole spontanée	38
4	Synthèse	41
II	La parole spontanée : un objet d'étude complexe	43
1	Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole	46
1.1	Les caractéristiques de la parole spontanée	47
1.1.1	Les caractéristiques acoustiques	47
1.1.2	Les caractéristiques linguistiques	53
1.1.3	Les difficultés induites pour la reconnaissance automatique de la parole	54
1.1.4	Déterminer des niveaux de spontanéité en fonction de ces caractéristiques	57
2	La parole spontanée : caractériser ses variations	60
2.1	Définir la parole spontanée	60
2.2	Les facteurs de variation de la spontanéité	62
2.2.1	Les facteurs interpersonnels	62
2.2.2	Les facteurs environnementaux	64
2.2.3	Les facteurs plus personnels	65
2.3	Revue des modèles à une ou plusieurs dimensions relatifs à la variation stylistique	65
2.3.1	Le niveau de formalité	66
2.3.2	L'attention portée au discours	66

2.3.3	Le niveau de contrôle	67
2.3.4	Modèles prenant en compte d'autres dimensions	68
3	Synthèse	70
III	Questions de recherche	73
B	Méthodologie	
IV	Cadre méthodologique	79
1	Les données de parole spontanée	80
1.1	Le processus de sélection des corpus de parole spontanée	81
1.2	Les dimensions de variation de la parole spontanée	82
1.3	L'étiquetage des corpus et ensembles de données pour les expérimentations	83
2	Les systèmes de reconnaissance automatique de la parole de référence	84
2.1	Les modèles pré-appris sur le français : le projet LeBenchmark . .	84
2.2	Deux systèmes état de l'art comme systèmes de référence	86
2.2.1	Le système Speechbrain	86
2.2.2	Le système Whisper en <i>zero-shot</i>	89
2.3	Métriques d'évaluation	89
3	Plans d'expérimentations	89
3.1	Mesure des performances des systèmes état de l'art selon les sous-types de parole spontanée	90
3.2	Les adaptations aux différents sous-types de parole spontanée . .	91
3.3	L'utilisation des données de parole spontanée non catégorisées . .	91
4	Synthèse	92
C	Contributions	
V	Modèles pré-appris pour le français : le projet LeBenchmark	99
1	Collecte et pré-traitement des données	100
1.1	Méthode de sélection des données	100
1.2	Présentation des corpus	101

1	Collecte et pré-traitement des données	100
1.1	Méthode de sélection des données	100
1.2	Présentation des corpus	101

1.2.1	Augmenter la quantité de données disponible : le corpus audiocite.net	102
1.3	Préparation des données	103
1.4	Création d'ensembles d'apprentissage	104
2	Modèles disponibles	107
3	Performances des modèles pour une tâche de reconnaissance automatique de la parole	109
3.1	Modèles pré-appris utilisés comme extracteurs de paramètres : Interspeech 2021	109
3.2	Modèles pré-appris adaptés à la tâche : NeurIPS 2021	110
3.3	LeBenchmark 2.0 : <i>Computer Speech & Language</i>	112
4	Synthèse	114
VI	Dimensions de variation de la parole spontanée et étiquetage de données	117
1	La collection de corpus de parole spontanée	118
1.1	Les données issues du Corpus d'Étude pour le Français Contemporain (CEFC)	119
1.2	Autres corpus rassemblés	121
2	Dimensions de variation de la parole spontanée	123
2.1	Analyse des <i>data papers</i>	124
2.1.1	Le rôle des locuteurs	125
2.1.2	Le degré d'intimité entre les locuteurs	126
2.1.3	Le thème de la discussion	127
2.1.4	Les émotions	128
2.1.5	Le lieu	129
2.1.6	Le nombre d'interlocuteurs	130
2.1.7	Le canal de communication et la présence physique des locuteurs	130
2.1.8	La présence d'un auditoire	130
2.1.9	Discussion	131
2.2	Détermination de niveaux de parole spontanée pour la reconnaissance automatique de la parole : une méthode en quatre dimensions	132
3	Préparation des données pour la reconnaissance automatique de la parole	135
3.1	Tri gros grain des données finales	135
3.2	Normalisation des transcriptions et fichiers audio	138

3.3	Étiquetage des données pour la création d'ensembles représentatifs de sous-types de parole spontanée	139
4	Synthèse	143
VII	Résultats d'expérimentations	147
1	Performances de reconnaissance avec des systèmes état de l'art	149
1.1	Analyse des résultats	151
2	Adaptation spécifique d'un modèle pré-appris	152
2.1	Analyse des résultats	153
2.2	Validation croisée sur chacun des cas d'étude	156
2.2.1	Validation croisée pour le cas <i>Usual_close</i>	158
2.2.2	Validation croisée pour le cas <i>Unusual_close</i>	160
2.2.3	Validation croisée pour le cas <i>Unusual_distant</i>	162
2.3	Analyse générale de la validation croisée	163
3	Apports d'une adaptation au domaine	165
3.1	Un système adapté au domaine spontané	165
3.2	Adaptation spécifique d'un système adapté au domaine	167
3.2.1	Analyse des résultats	168
4	Synthèse	168
Conclusion, limites et perspectives		171
Références		204
D	Annexes	
A	Correspondance API-SAMPA	207
B	Performances de validation croisée (WER) sur les différents cas d'étude	209

Liste des figures

I.1	Principe général d'un système de reconnaissance automatique de la parole	4
I.2	Chaîne de Markov cachée (HMM)	6
I.3	Exemple de décomposition HMM-GMM (triphones) pour le segment [alOr] où le [O] serait prolongé	7
I.4	Architecture générale d'un système de reconnaissance automatique de la parole (<i>recognizer</i> sur la figure) <i>end-to-end</i> (Vazhe-nina et Markov, 2020)	9
I.5	Le Perceptron de Rosenblatt (1958) (image tirée de Alves (2006))	10
I.6	Réseau de neurones entièrement connectés à propagation avant (sur la base d'une image de (Pelletier <i>et al.</i> , 2019))	11
I.7	Différents types de RNN (Li, 2023)	12
I.8	Architecture d'un réseau de neurones récurrents (RNN)	13
I.9	Réseau de neurones LSTM (<i>Long-Short Term Memory</i>) (Navarro <i>et al.</i> , 2020)	14
I.10	Réseau de neurones convolutionnels. Figure inspirée de (Phung et Rhee, 2019).	15
I.11	Dates-clés des avancées en reconnaissance automatique de la parole neuronale	17
I.12	Classification temporelle connectionniste (CTC) (image tirée de (Prabhavalkar <i>et al.</i> , 2017))	18
I.13	Exemple de traitement de la sortie d'un réseau de neurones par la CTC : (1) Les caractères qui sont répétés et qui se suivent sont fusionnés. (2) Le caractère blanc est supprimé.	18
I.14	Schéma d'un RNN-T (image tirée de (Prabhavalkar <i>et al.</i> , 2017))	19
I.15	L'architecture encodeur-décodeur de Cho <i>et al.</i> (2014)	20
I.16	Schéma explicatif du mécanisme d'attention (adapté de Bahda-nau <i>et al.</i> (2014))	21
I.17	Attentions globales et locales par (Luong <i>et al.</i> , 2015)	22
I.18	L'architecture Transformer de Vaswani <i>et al.</i> (2017)	23
I.19	Classification de tâches prétextes pour l'apprentissage de modèles pré-appris pour l'audio selon (Mohamed <i>et al.</i> , 2022)	25

LISTE DES FIGURES

I.20	L'approche Whisper de Radford <i>et al.</i> (2022)	28
I.21	Architecture autosupervisée Wav2Vec 2.0 (avec l'aimable autorisation de Jonathan Boigne)	29
I.22	Résultats de Tancoigne <i>et al.</i> (2020) sur quatre ensembles de test, obtenus via 8 systèmes commerciaux de reconnaissance automatique de la parole (WER%).	39
II.1	Exemple de réduction tiré de Wu et Adda-Decker (2020). “par exemple” est prononcé [paRa~p]	49
II.2	Exemple de fenêtrage syntaxique de la séquence du corpus CaFE (Bazillon <i>et al.</i> , 2008b)	54
II.3	Différence de niveau de formalité en fonction du contexte plus ou moins formel entre locuteurs introvertis et extravertis.	66
II.4	Dimensions de variation des styles de parole selon Labov (1973) .	67
II.5	Variation des styles de parole selon le contrôle (Wagner <i>et al.</i> , 2015)	67
II.6	Catégorisation des genres/styles de parole de (Goldman <i>et al.</i> , 2009)	68
II.7	Les dimensions le long desquelles les styles de parole peuvent être situés selon Eskenazi (1993)	69
II.8	Le continuum communicatif de Koch et Oesterreicher (2001) .	69
II.9	Valeurs paramétriques du langage parlé (à gauche) et du langage écrit (à droite) selon Koch et Oesterreicher (2001)	70
II.10	Relief conceptionnel de l'entretien professionnel par Koch et Oesterreicher (2001)	70
IV.1	Détail des données d'apprentissage, d'optimisation et de test utilisées pour l'apprentissage et/ou l'évaluation des systèmes état de l'art Speechbrain et Whisper.	91
IV.2	Détail des données d'apprentissage, d'optimisation et de test utilisées pour l'entraînement et l'évaluation des systèmes adaptés aux sous-types de parole spontanée capturés dans les cas d'étude	92
IV.3	Détail des données d'apprentissage, d'optimisation et de test utilisées pour l'entraînement et l'évaluation du système adapté au domaine puis adapté aux sous-types de parole spontanée	93
IV.4	Ensembles de données d'apprentissage, de développement et de test qui seront utilisés pour l'apprentissage et l'évaluation de nos systèmes	94
V.1	Représentation des différents types de parole dans les ensembles de données <i>small</i> , <i>medium</i> , <i>large</i> et <i>extra large</i> (en %)	107

LISTE DES FIGURES

V.2	Répartition par genre (Hommes/Femmes/Non renseigné) dans les différents ensembles de données LeBenchmark (en heures) . . .	108
VI.1	Proportion en nombre de mots de chacun des corpus présentés dans le CEFC (graphique tiré de Bérard (2020))	122
VI.2	Le “model of affect” de Russell (image tirée de Gonzalez <i>et al.</i> (2020))	129
VI.3	Situation de communication	133
VI.4	Degré d'intimité entre les locuteurs	134
VI.5	Type de communication	134
VI.6	Canal de communication	135
VI.7	Méthode en quatre dimensions pour l'étiquetage de la variation de la parole spontanée. Exemple de l'étiquetage du corpus CID en superposition.	136
VI.8	Exemple d'une entrée dans un fichier <i>.json</i>	138
VI.9	Positionnement des cas d'étude dans l'espace en quatre dimensions de variation de la parole spontanée	141
VI.10	Détail des ensembles d'apprentissage, de développement et d'évaluation <i>All_Spont</i> et Sous-types de spontané (durées = heures effectives de parole)	143
VII.1	Contenu du corpus ETAPE (tiré de (Gravier <i>et al.</i> , 2012))	155
VII.2	Représentation de la technique de validation croisée à trois lots (adaptation d'une image de (Duran-Lopez <i>et al.</i> , 2020))	157
VII.3	WER/enregistrement (gauche) et WER/locuteur (droite) pour chaque lot du cas <i>Usual_close</i>	160
VII.4	WER/enregistrement (gauche) et WER/locuteur (droite) pour chaque lot du cas <i>Usual_close</i>	162
VII.5	WER/enregistrement (gauche) et WER/locuteur (droite) pour chaque lot du cas <i>Usual_distant</i>	164
VII.6	Positionnement des cas d'étude dans l'espace en quatre dimensions de variation de la parole spontanée tel que défini dans ce travail de thèse	173
A.1	Correspondance API-SAMPA pour les consonnes du français . .	207
A.2	Correspondance API-SAMPA pour les voyelles du français . . .	207

Liste des tableaux

I.1	Comparaison de techniques d'extraction de paramètres acoustiques pour la reconnaissance automatique de la parole (Labied et Belangour, 2021)	5
I.2	Modèles pré-appris multilingues incluant le français (liste non-exhaustive)	26
I.3	Exemples d'insertion, substitution et suppression	31
I.4	Corpus d'évaluation monolingues de référence pour la RAP en français	34
I.5	Corpus d'évaluation multilingues de référence pour la reconnaissance automatique de la parole en français	36
I.6	Récapitulatif des performances état de l'art trouvées dans la littérature de systèmes de RAP sur différents corpus pour le français	36
I.7	Résultats tirés de Baevski <i>et al.</i> (2020) rapportant les performances d'un modèle pré-appris de type Wav2Vec 2.0 sur LibriSpeech (anglais), en fonction de la taille du corpus d'adaptation et l'utilisation d'un modèle de langue	38
I.8	Word Error Rate en fonction de niveaux de spontanéité	41
II.1	Caractéristiques acoustiques les plus fréquentes en parole spontanée	52
II.2	Caractéristiques linguistiques les plus retrouvées en parole spontanée	55
II.3	Résultats du système de reconnaissance automatique de Dufour <i>et al.</i> (2014) selon la catégorisation manuelle de segments comme “préparé”, “peu spontané” ou “très spontané”	58
II.4	Évaluation de la classification automatique de segments selon trois types de parole en terme de précision, rappel et F-mesure Dufour <i>et al.</i> (2014)	59
IV.1	Différentes types d'architectures Wav2Vec 2.0 utilisées pour l'entraînement des modèles LeBenchmark	85
IV.2	Hyperparamètres du système état de l'art à base de CTC proposé par Speechbrain	88

LISTE DES TABLEAUX

IV.3	Rappel des différents systèmes utilisés pour les expérimentations, ainsi que des ensembles de données d'apprentissage et de développement.	95
V.1	Statistiques des corpus oraux utilisés dans le projet LeBenchmark en fonction du genre (homme/femme/non renseigné) . . .	106
V.2	Récapitulatif des modèles pré-appris Wav2vec 2.0 appris au cours du projet LeBenchmark	108
V.3	Résultats de Evain <i>et al.</i> (2021a) pour une tâche de RAP <i>End-to-end</i> (%WER) sur les corpus en français Common Voice et ETAPE	110
V.4	Résultats de Evain <i>et al.</i> (2021b) pour une tâche de RAP <i>end-to-end</i> (WER%) sur les corpus Common Voice et ETAPE, obtenus avec des modèles pré-appris de type Wav2Vec 2.0 adaptés avec les données correspondantes	111
V.5	Résultats de Parcollet <i>et al.</i> (2024) pour une tâche de RAP <i>end-to-end</i> (WER%) sur les corpus Common Voice et ETAPE, obtenus avec des modèles pré-appris de type Wav2Vec 2.0 adaptés avec les données correspondantes	113
VI.1	Récapitulatif de l'ensemble des corpus rassemblés étudiés par la suite	124
VI.2	Récapitulatif des facteurs de variation de la parole spontanée et des éléments qui les constituent repérés dans les <i>data papers</i> des différents corpus récoltés	132
VI.3	Liste finale des corpus qui seront utilisés pour les expérimentations	137
VII.1	Rappel des différents systèmes utilisés pour les expérimentations, ainsi que des ensembles de données d'apprentissage et de développement.	149
VII.2	Performances obtenues avec deux systèmes de RAP état de l'art (WER%)	150
VII.3	Performances obtenues suite à l'adaptation du modèle LeBenchmark 7K avec chacun des cas étudié (WER%)	153
VII.4	Détail du contenu de chacune des émissions incluses dans le corpus ETAPE (recherches effectuées par nos soins)	155
VII.5	Détail des lots par cas d'étude constitués pour la validation croisée	158
VII.6	Validation croisée de type <i>k-fold</i> sur les données <i>Usual_close</i> . .	159
VII.7	Validation croisée de type <i>k-fold</i> sur les données <i>Unusual_close</i> .	161
VII.8	Validation croisée de type <i>k-fold</i> sur les données <i>Unusual_distant</i>	163

LISTE DES TABLEAUX

VII.9	Récapitulatif des WER sur chacun des cas d'étude obtenus avec et sans validation croisée	165
VII.10	Performances obtenues avec un modèle adapté au domaine spontané (%WER)	166
VII.11	Performances obtenues suite à l'adaptation spécifique du système adapté au domaine (%WER)	167
VII.12	Performances obtenues sur les différents cas étudiés au fil des expérimentations menées	170
B.1	WER par lot, par enregistrement et par locuteur pour le cas <i>usual_close</i>	209
B.2	WER par lot, par enregistrement et par locuteur pour le cas <i>unusual_close</i>	213
B.3	WER par lot, par enregistrement et par locuteur pour le cas <i>unusual_distant</i>	216

Introduction

Contexte

Apparue au début des années 1950, la reconnaissance automatique de la parole est un procédé permettant d'extraire le texte correspondant au message linguistique contenu dans un signal de parole. Jusqu'au début des années 2010, celle-ci consistait en l'apprentissage de différents composants, à savoir : le modèle acoustique, le modèle de langue et le lexique phonétisé. Le développement et les progrès du *deep learning* et des approches neuronales de bout en bout (ou *end-to-end*) de ces dernières années aura permis aux linguistes-informaticiens de considérer des approches autres que modulaires. En effet, les systèmes entièrement neuronaux sont aujourd'hui en capacité de convertir directement les entrées audio en texte, au sein d'une seule et même architecture. L'entraînement de ce type de système est néanmoins limité par la nécessité d'accéder à une grande quantité de données. Cette limitation est aujourd'hui atténuée par les techniques d'apprentissage auto-supervisé, qui permettent d'apprendre des modèles à partir de données non transcrrites, lesquelles sont disponibles en plus grande quantité. Ces modèles qualifiés de "pré-appris" présentent l'avantage de pouvoir à la fois être utilisés comme simples extracteurs de paramètres acoustiques, utilisés par la suite comme entrées des systèmes de reconnaissance automatique de la parole, mais aussi d'être adaptés, ce qui permet d'améliorer leurs performances pour une tâche particulière (comme la reconnaissance automatique de parole médiatique).

Si les systèmes actuels peuvent afficher des performances impressionnantes, avec une capacité à reconnaître en moyenne neuf mots sur dix, il convient de noter que ces performances varient en fonction du type de parole. Les meilleurs résultats sont souvent obtenus dans des situations de parole préparée ou professionnelle, caractérisées par des phrases grammaticalement correctes, une fluidité de discours et une articulation précise. En revanche, la parole spontanée se produit dans des situations moins contrôlées, inclut des hésitations, des répétitions et des erreurs de prononciation, et représente une source de difficulté pour la transcription automatique. La performance des systèmes sur la parole spontanée est donc toujours

un challenge.

Au delà d'un simple défi technique, l'importance de l'amélioration de la reconnaissance de la parole spontanée trouve sa source dans le besoin des linguistes, sociologues et journalistes de transcrire plus rapidement les enregistrements qui sont au cœur de leur travail. Il faut en effet rappeler que le temps nécessaire à la transcription manuelle de la parole spontanée est nettement plus long que celui requis pour la transcription d'autres types de parole, principalement en raison des répétitions, des hésitations, et d'autres caractéristiques propres à ce type de discours. De plus, outre son utilisation dans le secteur professionnel (transcription automatique de réunions, redirection automatique d'un client à un agent...) ou la sphère personnelle (assistants virtuels...), la reconnaissance automatique de la parole constitue également le premier maillon de diverses chaînes de traitement. Elle peut ainsi représenter un élément de base de systèmes de compréhension automatique de la parole, permettant aux machines de saisir le sens et l'intention derrière les mots prononcés. De même, elle joue un rôle crucial dans la traduction automatique de la parole, facilitant la communication entre personnes de langues différentes. Elle bénéficie également à l'inclusion dans la société des personnes en situation de handicap, par exemple *via* le sous-titrage automatique, rendant les contenus audiovisuels accessibles aux personnes malentendantes ou sourdes (Evain *et al.*, 2020a), ou encore *via* la transcription automatique en pictogrammes (Ormaechea-Grijalba *et al.*, 2023), plus faciles à interpréter pour certaines personnes.

Problèmes et limites

Quelques auteurs, dont Dufour *et al.* (2010) et Deléglise et Lailler (2020), ont pu constater la sensibilité des systèmes de reconnaissance automatique de la parole à la variation des niveaux de spontanéité : à mesure que la spontanéité augmente, le WER augmente. Cependant, la variabilité de situations amenant à la production de différents types de parole spontanée ne se retrouve pas dans les corpus de référence en français actuellement utilisés en reconnaissance automatique de la parole. En effet, ces corpus sont soit constitués de parole lue, soit de parole médiatique, ce dernier étant un contexte spécifique en raison de ses sujets préparés, de son large auditoire et de ses locuteurs professionnels ayant pour mission de capter l'attention et d'être compris de l'ensemble de leurs auditeurs.

Au delà du manque de données variées, nous pouvons également noter l'absence d'une méthode permettant de mettre en parallèle les performances obtenues sur différents sous-types de parole spontanée. Cause ou conséquence, la parole spontanée est ainsi souvent réduite à toute parole qui n'est pas préparée. Ceci représente une limite pour l'analyse fine des performances des systèmes, la parole spontanée produite en contexte d'interview et la parole spontanée apparaissant dans une

discussion entre amis étant, par exemple, assez différentes. Comment, alors, évaluer efficacement les systèmes et nous assurer de leur bonne performance dans des contextes autres que celui constraint par les corpus de référence ? Cette interrogation aura été à la base de ce travail de thèse.

Questions de recherche

Si les corpus de parole spontanée créés pour la reconnaissance automatique de la parole en français sont limités dans leur diversité, d'autres corpus de parole spontanée issus d'études linguistiques sont librement partagés avec la communauté. Ces derniers sont bien plus riches en terme de situations d'enregistrements. Il est probable qu'ils n'aient été, jusqu'à présent, que très peu utilisés en reconnaissance automatique de la parole à la fois à cause de leur petite taille mais aussi à cause de la diversité de situations qu'ils représentent, ce qui ne permet pas de créer des ensembles de données uniformes. Ainsi, nous basons la première partie de notre recherche sur la détermination de dimensions de la variation de la parole spontanée, qui nous permettent :

- de constituer des ensembles de données inter-corpus homogènes, et donc de maximiser la quantité de données correspondant à une même situation,
- de positionner des ensembles de données les uns par rapport aux autres, et ainsi étudier les performances des systèmes en fonction des niveaux de spontanéité capturés dans les données.

Afin d'aller vers une meilleure prise en compte de ces différents types de parole spontanée par les systèmes de reconnaissance automatique de la parole, nous rassemblons les données de parole spontanée pour construire, dans chaque cas, des ensembles d'apprentissage, de développement et de test. Si la spécificité de nos dimensions nous amène à la constitution d'ensembles de données de taille assez réduite, l'utilisation de modèles pré-appris, présentant l'avantage de pouvoir être adaptés avec une petite quantité de données transcrits, nous permet malgré tout de poursuivre notre étude. Une fois les dimensions déterminées et les données étiquetées en fonction de ces dimensions, nous étudions ainsi :

- l'impact de l'adaptation spécifique d'un modèle pré-appris avec une faible quantité de données spécifiques sur les performances de systèmes de reconnaissance automatique de la parole spontanée,
- l'impact de l'adaptation au domaine d'un modèle pré-appris avec une grande quantité de données -cette fois moins contrôlées- sur les performances de systèmes de reconnaissance automatique de la parole spontanée.

Nous espérons, avec ce travail, contribuer à une meilleure analyse des performances des systèmes de reconnaissance automatique de la parole sur la parole spontanée,

mais aussi à l'analyse des capacités de spécialisation des systèmes à différents sous-types de parole spontanée lors de l'étape d'adaptation de modèles pré-apris.

Organisation du manuscrit

Ce manuscrit est structuré en trois parties : l'état de l'art, la méthodologie et les contributions.

Dans la première partie, nous faisons, en **chapitre 1**, une revue des architectures utilisées pour la reconnaissance automatique de la parole, ainsi que des performances relevées sur différents types de parole, dont la parole spontanée. Nous montrons l'apport des modèles pré-apris, et notamment les bonnes performances atteignables en reconnaissance automatique de la parole grâce à eux, ainsi que la possibilité d'adaptation des modèles avec peu de données. Nous montrons également que les systèmes de reconnaissance automatique de la parole sont sensibles à la variation de la spontanéité, bien que l'étude de cet aspect soit limité par le manque de variabilité des corpus de référence utilisés en reconnaissance automatique de la parole spontanée en français. En ce sens, le **chapitre 2** est consacré à l'élaboration d'une revue de ce qu'implique la notion de parole spontanée et de ses différents facteurs de variation relevés dans la littérature, grâce à la confrontation de travaux en reconnaissance automatique de la parole, en traitement automatique des langues, en sociolinguistique, en linguistique et en stylistique. Ce chapitre nous permet de mettre en lumière la nécessité d'étudier la parole spontanée dans toute sa variation, étant donné qu'elle peut tout à la fois varier en fonction du rapport de rôle, du nombre de locuteurs, de lien entre les locuteurs, de leurs présence physique ou encore de la proximité de ceux-ci avec le sujet abordé au cours d'une interaction. Suite à ces deux chapitres sur les modèles, les performances de systèmes de reconnaissance automatique de la parole et la notion de parole spontanée, nous rappelons et approfondissons en **chapitre 3**, nos questions de recherche, à savoir :

- comment déterminer un nombre réduit de facteurs de variation de la parole spontanée afin de constituer des groupes de données homogènes et les positionner les uns aux autres en fonction de leur spontanéité ?
- l'adaptation spécifique d'un modèle pré-apris avec une faible quantité de données représentatives de différents niveaux de spontanéité (et donc contrôlées) peut-elle améliorer les performances de systèmes de reconnaissance automatique de la parole spontanée ?
- quel est l'impact de l'adaptation d'un modèle pré-apris avec une grande quantité de données moins contrôlées sur les performances de systèmes de reconnaissance automatique de la parole spontanée ?

La deuxième partie de ce manuscrit présente, en **chapitre 4**, l'approche méthodologique mise en place dans cette thèse. Nous souhaitions étudier la reconnaissance

automatique de la parole spontanée en nous confrontant à des corpus issus d'études en linguistiques et représentatifs d'une grande variété de situations langagières (dîners entre amis, interactions dans le commerce, enquêtes...). Nous présentons ainsi, tout d'abord, le processus de sélection de corpus de parole spontanée que nous avons mis en place. Ensuite, nous introduisons la façon dont nous exploitons les éléments de conception de ces différents corpus afin de déterminer un nombre réduit de dimensions de variation qui soient utilisables dans un contexte de reconnaissance automatique de la parole. Un nombre restreint de dimensions de variation de la parole spontanée ainsi déterminé, nous présentons leur réutilisation pour la construction d'ensembles de données supposés homogènes, que nous étudions au travers de quelques cas d'étude, représentatifs de différents niveaux de spontanéité. Enfin, nous présentons, dans ce chapitre, la genèse du projet LeBenchmark, dont le but était la mise à disposition de modèles pré-appris pour le français, ainsi que les différents types d'adaptation d'un modèle pré-appris mis en place afin de répondre aux questions de recherche techniques soulevées précédemment.

La troisième et dernière partie de ce manuscrit présente les contributions. Ainsi, le **chapitre 5** présente notre contribution concernant la collecte et la préparation de données liées au projet LeBenchmark, et notamment la constitution d'un corpus de parole lue représentant plus de 6 500 heures de parole. Nous y présentons également les modèles mis à disposition de la communauté, ainsi que leurs performances, pour une tâche de reconnaissance automatique de la parole, sur les deux corpus de référence Common Voice (lecture) et ETAPE (parole préparée/spontanée). Les différentes études effectuées lors de ce projet montrant un réel apport des modèles monolingues LeBenchmark, notamment sur ETAPE, par rapport à des modèles multilingues ou des paramètres acoustiques classiques, nous avons utilisé ces modèles pour la suite de nos travaux. Le **chapitre 6** est tout d'abord consacré à la présentation des corpus de linguistique collectés pour mener nos travaux de recherche, ainsi qu'à la mise en parallèle des facteurs de variation de la parole spontanée issus de la littérature avec les métadonnées de ces corpus. Ensuite, nous y présentons les différents cas d'étude nous permettant d'étudier une partie de la variation de la parole spontanée, notamment l'influence du rôle, du lieu et de le degré d'intimité entre les locuteurs. Le **chapitre 7** présente l'impact de cette variation sur les performances de différents systèmes de reconnaissance automatique de la parole, au travers de différents types d'adaptations d'un modèle pré-appris LeBenchmark. Ainsi, nous étudions tout d'abord l'impact de différentes adaptations spécifiques d'un modèle pré-appris avec de petits ensembles de données correspondant aux différents cas d'étude, avant d'analyser l'impact d'une adaptation avec une grande quantité de données, correspondant à une adaptation sur des données plus diverses, que nous avons regroupées sous la terminologie englobante de "domaine spontané".

Première partie

Etat de l'art

Chapitre I

Les systèmes de reconnaissance automatique de la parole

La Reconnaissance Automatique de la Parole est un procédé consistant à analyser la voix afin d'extraire le texte correspondant au message linguistique contenu dans un signal vocal. Celle-ci peut aussi bien être utilisée à des fins d'interfaces (via l'utilisation d'assistants vocaux tels qu'Alexa d'Amazon, ou la Google Home de Google (Ammari *et al.*, 2019) (CSA-Hadopi, 2019)), servir dans un contexte professionnel pour la transcription automatique de rapports médicaux ou de réunions (Payne *et al.*, 2018; Yu *et al.*, 2022), ou encore s'avérer être utile pour l'inclusion des personnes en situation de handicap (Kafle et Huenerfauth, 2017; Evain *et al.*, 2020b). Nous présentons dans ce chapitre les techniques les plus utilisées pour la résolution de cette tâche. Nous faisons ensuite un état de l'art des performances qu'il est possible d'obtenir sur différents types de parole, et notamment sur la parole spontanée.

Déroulement du chapitre

1	L'approche statistique	2
1.1	L'extraction de paramètres acoustiques	4
1.2	Le modèle acoustique	5
1.3	Le dictionnaire de prononciation	7
1.4	Le modèle de langue	8
2	Les approches neuronales	8
2.1	Les types de réseaux neuronaux utilisés en reconnaissance automatique de la parole	9

2.1.1	Le fonctionnement général d'un réseau de neurones et réseau de neurones à propagation avant (FNN)	10
2.1.2	Les réseaux de neurones récurrents (RNN)	12
2.1.3	Réseaux à mémoire à court et long terme (LSTM)	13
2.1.4	Réseaux de neurones convolutionnels (CNN)	14
2.2	Les systèmes hybrides HMM-DNN	15
2.3	Les systèmes à base de classification temporelle connectionniste (CTC)	17
2.4	Les systèmes à base de transducteur (RNN-T)	19
2.5	Les systèmes à base d'attention	20
2.5.1	L'architecture encodeur-décodeur	20
2.5.2	L'attention	21
2.6	Les modèles pré-appris	24
2.6.1	Le fonctionnement général des modèles pré-appris	24
2.6.2	Les modèles pré-appris pour le français	26
2.6.3	Les modèles multilingues Whisper	27
2.6.4	L'architecture Wav2Vec 2.0	29
3	L'évaluation des systèmes de reconnaissance automatique de la parole	30
3.1	Le WER et ses limites	30
3.2	Les principaux corpus pour la reconnaissance automatique de la parole en français	32
3.2.1	Les corpus monolingues	32
3.2.2	Les corpus multilingues	34
3.3	Les performances de reconnaissance automatique	35
3.3.1	Quelques résultats généraux sur différents types de parole	35
3.3.2	Résultats sur la parole spontanée	38
4	Synthèse	41

1 L'approche statistique

Une première approche de développement d'un système de reconnaissance automatique de la parole est l'approche statistique, basée sur des modèles de Markov cachés (HMM pour *Hidden Markov Model*) (Bourlard et Morgan, 1994). Les systèmes ont pour objectif, à partir d'une séquence de paramètres acoustiques, de trouver la

séquence de mots prononcée la plus probable parmi l'ensemble des séquences possibles (Jelinek, 1998; Athanaselis *et al.*, 2005). Cela se traduit mathématiquement par l'équation suivante :

$$\tilde{W} = \arg \max_w P(W|X) \quad (\text{I.1})$$

où la séquence de mots la plus probable est notée \tilde{W} et la séquence de paramètres du signal acoustique est notée X .

L'application de la formule de Bayes permet de réécrire la formule de la façon suivante :

$$\tilde{W} = \arg \max_w \frac{P(X|W)P(W)}{P(X)} \quad (\text{I.2})$$

où $P(W)$ représente la probabilité *a priori* de la suite de mots W (le modèle de langue), $P(X|W)$ représente la probabilité d'observer le signal de parole X étant donné la suite de mot prononcée W (modèle acoustique), et $P(X)$ est la probabilité d'observer le signal de parole X .

Cette probabilité $P(X)$ ne dépend pas de W et n'est pas utile pour déterminer la meilleure hypothèse. Elle peut donc être ignorée (Barrault, 2008). La formule peut ainsi être simplifiée de cette façon :

$$\tilde{W} = \arg \max_w P(X|W)P(W) \quad (\text{I.3})$$

Dans le cas des systèmes à grand vocabulaire, le modèle acoustique calcule la probabilité d'observer le signal de parole X étant donné une suite de phones¹ U . Un lexique phonétisé² est alors ajouté afin d'estimer la probabilité des prochains phonèmes³ et d'associer un mot à une suite de phonèmes. Cela se représente mathématiquement ainsi :

$$\tilde{W} = \arg \max_w P(X|U)P(U|W)P(W) \quad (\text{I.4})$$

où $P(U|W)$ représente le lexique phonétisé, $P(X|U)$ le modèle acoustique et $P(W)$ le modèle de langue.

1. Réalisations concrètes d'un phonème, celles-ci pouvant varier en fonction du contexte et des locuteurs.

2. ou dictionnaire de prononciation

3. Le plus petit segment phonique (dépourvu de sens) permettant de distinguer les mots. Un phonème peut correspondre à plusieurs phones, selon son contexte de réalisation ou le locuteur et le modèle acoustique prend en compte cette variation.

L’association d’un modèle acoustique, d’un lexique phonétisé et d’un modèle de langue forme donc un système de reconnaissance automatique de la parole, comme le montre la figure I.1. La transcription finale obtenue en sortie du système fait suite à une succession d’allers et retours entre ces différents éléments, afin de proposer en sortie la séquence de mots avec la plus forte probabilité. Dans les sections qui suivent, nous détaillons le principe d’extraction de paramètres acoustiques ainsi que le fonctionnement de chaque module.

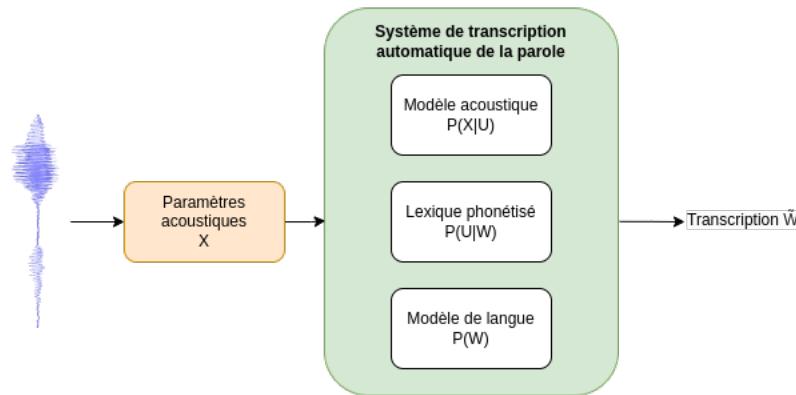


FIGURE I.1 – Principe général d’un système de reconnaissance automatique de la parole

1.1 L’extraction de paramètres acoustiques

Ce paragraphe se base sur les travaux de (Lecouteux, 2008; Pols, 2011; Labied et al., 2022).

L’extraction de paramètres acoustiques sert à capturer, dans un signal vocal, les informations pertinentes pour la reconnaissance de la parole. Ainsi, la quantité d’informations fournie en entrée d’un système est réduite et le traitement est moins complexe et donc plus rapide. On garde, par exemple, la forme de l’enveloppe spectrale qui reflète la forme du conduit vocal et constitue une indication du placement des articulateurs (mâchoire, langue, velum, lèvres).

Avant l’extraction de paramètres acoustiques, le signal est découpé en fenêtres de 25 à 40 ms, afin de travailler sur des portions de signal les plus stables possible. Les vecteurs de paramètres acoustiques sont extraits de ces fenêtres. Afin de permettre de capturer les dynamiques entre ces fenêtres et d’éviter un passage trop brutal de l’un à l’autre, elles se chevauchent de 10 ms. Les effets de bord et discontinuités du signal sont limités grâce à une fenêtre de Hamming. Les étapes successives d’ana-

1 L'approche statistique

Feature extraction	Pros	Cons
MFCC	- High recognition accuracy [27] - Good discrimination and low coefficients correlation [27]	- Inaccurate recognition in noisy speech [27] - high dimensional features vectors [31]
PCA	- Robustness to noises [9] - Reduce the feature vector's size while retaining important information [15]	- Expensive in terms of computing for high-dimensional data [27]
LPC	- Computation speed - Robust for extracting features from speech signals with a low bit rate [33]	- Highly correlated feature coefficients [27] - Unable to distinguish words with similar phonemes
LPCC	- Decorrelate feature coefficients by the cepstral analysis - Robust than LPC [27]	- Unable to analyze local events accurately
PLP	- Low-dimensional feature vector [27] - Reduce the gap between voiced and unvoiced speech	- Altered Spectral balance [27]
DWT	- Denoising speech signal [34] - Compressing speech signal without significant loss of its quality [27]	- Inflexible [27]
RASTA-PLP	- Robustness - Excludes variations between cepstral components and speech signal [27]	- Low performance for noiseless speech [21]

TABLE I.1 – Comparaison de techniques d'extraction de paramètres acoustiques pour la reconnaissance automatique de la parole (Labied et Belangour, 2021)

lyse spectrale propres à chaque technique d'extraction des paramètres acoustiques sont ensuite effectuées.

Il existe plusieurs types de paramètres acoustiques (Markel et Gray JR, 1976; Hermansky et Cox Jr., 1991; Siwat Suksri et Yingthawornsuk, 2012) : les *Mel Filter Banks* (MFB), les *Linear Prediction Cepstral Coefficients* (LPCC), les *Mel Frequency Cepstral Coefficients* (MFCC), les *Perceptual Linear Prediction* (PLP)... Labied et Belangour (2021) ont notamment comparé certains d'entre eux pour leur utilisation en reconnaissance automatique de la parole (voir tableau I.1) Les paramètres acoustiques les plus utilisés sont ceux basés sur le fonctionnement de l'audition humaine, notamment les MFCC (Stevens *et al.*, 1937).

1.2 Le modèle acoustique

C'est au niveau du modèle acoustique qu'interviennent les chaînes de Markov cachée (HMM). Un HMM est un modèle statistique composé d'états, et de connexions entre ces états auxquelles on attribue une probabilité de transition comme le montre la figure I.2. Avec les probabilités de transition entre états, ces modèles permettent de déterminer *a priori* les états suivants les plus probables. Elles prédisent le futur en se basant sur le présent. Pour chaque état, on peut également calculer la probabilité d'émission de chaque événement observable. Les chaînes de Markov cachées peuvent ainsi être utilisées lorsque l'on veut calculer la probabilité d'événements directement observables, que sont les vecteurs de paramètres acoustiques dans le cadre de la reconnaissance automatique de la parole.

On utilise les chaînes de Markov cachées afin de déterminer la probabilité qu'une suite de vecteurs de paramètres acoustiques corresponde à une suite de phones. À

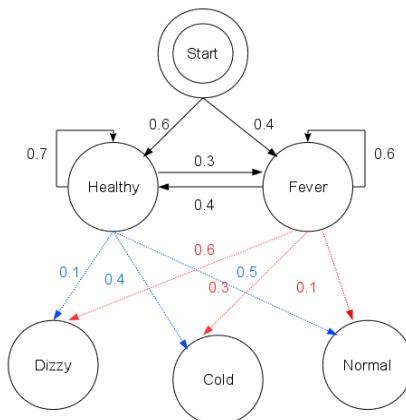


Image tirée de Wikimedia, licence : CC-BY-SA-4.0

FIGURE I.2 – Chaîne de Markov cachée (HMM)

un phone⁴, et donc à un plus ou moins grand nombre de fenêtres de signal selon la durée du phone, correspond un à trois états⁵ (Bromberg *et al.*, 2007).

Couplés à des modèles de mixtures gaussiennes (GMM pour *Gaussian Mixture Model*) qui permettent de calculer la probabilité de distribution des paramètres acoustiques extraits d'un signal, les HMM permettent ainsi de déterminer la probabilité, pour un vecteur de paramètres acoustiques, de correspondre à un état caché déterminé. Cette probabilité se calcule à partir d'un état initial en prenant en compte la probabilité de transition de cet état initial vers un nouvel état possible (probabilité de transition) et la probabilité d'émission de ce vecteur de paramètres acoustiques pour ce nouvel état (GMM de cet état), comme nous le montrons en figure I.3.

Les probabilités d'émission sont donc générées pour chaque fenêtre de signal, et sont utilisées ensuite pour choisir la séquence d'états cachés qui explique le mieux les observations acoustiques obtenues jusqu'au vecteur de paramètres acoustiques en cours de traitement⁶.

Il existe autant de HMM-GMM que de phones lorsque les HMM-GMM correspondent à des phonèmes en contexte comme sur la figure I.3. Les paramètres de

4. Il est d'usage de travailler au niveau du phone pour les systèmes de reconnaissance automatique de la parole grands vocabulaire (Dines et Magimai Doss, 2008). En effet, le nombre de phonèmes étant constant et peu élevé (37 phonèmes pour le français), cela s'avère être une solution intéressante pour modéliser ces unités discrètes intermédiaires ou états.

5. On parle alors de triphone.

6. Voir le cours “*Markov Chains and Hidden Markov Model for ASR*” de K.R. Chowdhary <https://krchowdhary.com/nlsp-lect/nlplect-4.pdf>

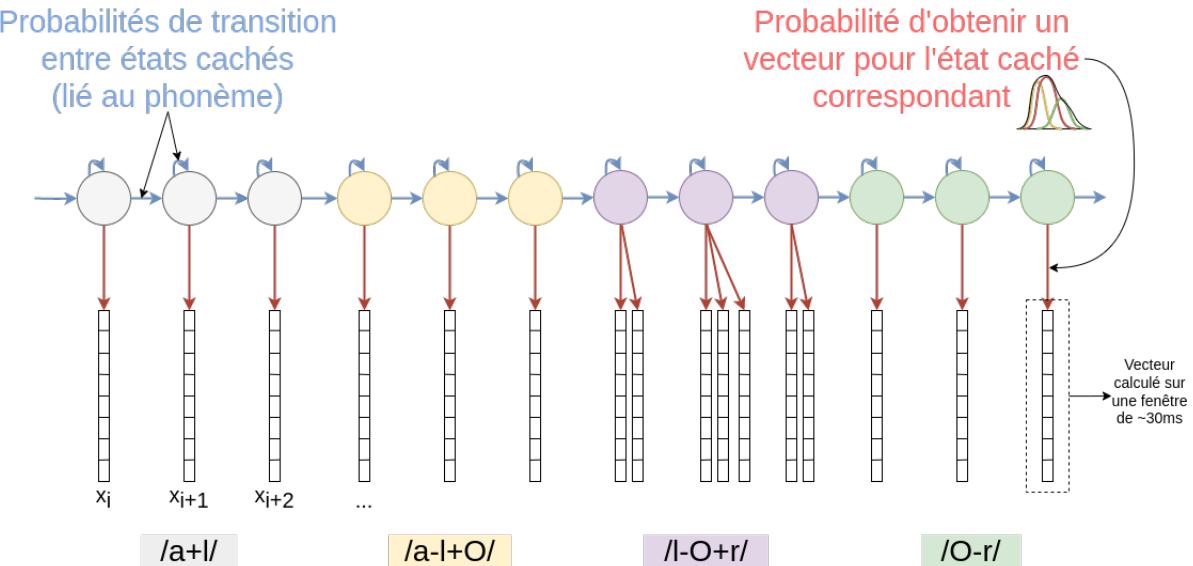


FIGURE I.3 – Exemple de décomposition HMM-GMM (triphones) pour le segment [alOr] où le [O] serait prolongé

ces modèles sont estimés depuis les données d'apprentissage grâce à l'algorithme d'*expectation-maximisation* (Dempster *et al.*, 1977). Les HMM-GMM sont alors appris simultanément grâce à différentes étapes d'alignement forcé entre les signaux acoustiques des données d'apprentissage et leur transcription, ayant pour but de faire correspondre le texte au moment où il est prononcé. L'algorithme de Baum-Welch (algorithme *forward-backward*) (Jiang, 2010) est un algorithme itératif utilisé pour chaque entrée afin de converger vers les probabilités de transition et d'émission de chaque état du modèle. Il maximise la fonction de vraisemblance (*log likelihood*) du modèle sur les données d'apprentissage.

1.3 Le dictionnaire de prononciation

Le dictionnaire de prononciation sert à faire le lien entre le modèle acoustique et le modèle de langue. Il indique pour chaque mot la suite de phonèmes attendue et permet d'estimer la suite d'états HMM-GMM correspondant à ce mot. Contrairement à nos dictionnaires usuels, le dictionnaire de prononciation regroupe toutes les formes possibles d'un mot. Ainsi, nous retrouvons toutes les formes conjuguées des verbes et les formes féminines et/ou au pluriel des noms et adjectifs. Afin de le rendre plus robuste, des variantes de prononciation sont intégrées (Adda-Decker, 2007), ainsi que des phénomènes de liaison (De Calmès *et al.*, 2005).

L'élaboration de ce dictionnaire doit être minutieuse car il est celui qui limite les

sorties du système de reconnaissance automatique de la parole. Sa génération peut être faite manuellement (ce qui est très coûteux en temps (Adda-Decker et Lamel, 2000)), ou automatiquement grâce à des outils comme LIAPhon (Béchet, 2001) ou des bases de données lexicales comme BDLex (Perennou, 1986), mais cela demande tout de même une étape de correction manuelle.

1.4 Le modèle de langue

Le modèle de langue permet d'estimer la probabilité pour un mot de suivre un autre mot ou une suite de mots. Il est là pour forcer le système à générer la suite de mots la plus probable.

L'apprentissage du modèle de langue se fait sur une grande quantité de données textuelles, desquelles on extrait des “n-grams” (Chen et Goodman, 1999), n représentant un nombre de mots. Il est d'usage d'utiliser des modèles de langue d'ordre n de 1 à 5. On obtient alors en sortie un modèle de langue composé d'unigrammes (des mots seuls, hors contexte), de bigrammes, des trigrammes... l'augmentation de n venant complexifier le modèle.

Une des limites connues de cette méthode est que certaines suites de mots pourraient ne pas exister dans le corpus d'apprentissage (et donc avoir une probabilité égale à zéro), alors même qu'elles sont susceptibles d'être prononcées par l'utilisateur du système de reconnaissance automatique de la parole. Une probabilité égale à zéro empêche le système de générer la séquence. Une solution consiste à attribuer une probabilité non nulle à ces suites de mots : c'est ce que l'on appelle le *smoothing* (Chen et Goodman, 1999). Pour cela, les probabilités des suites de mots les plus hautes sont baissées et attribuées aux suites de mots jamais vues.

2 Les approches neuronales

Le développement des moyens informatiques et des techniques de *deep learning* (apprentissage profond) aura permis au début des années 2010 de développer à la fois les techniques d'apprentissage hybrides et les techniques d'apprentissage *end-to-end* (E2E, de bout-en-bout) (Dahl *et al.*, 2012; Graves, 2012). Le qualificatif *end-to-end* fait référence au fait que le système de reconnaissance de la parole est entraîné d'un seul bloc, contrairement aux systèmes stochastiques pour lesquels le modèle acoustique, le lexique et le modèle de langue étaient modélisés séparément les uns des autres. Plutôt que de se baser sur des connaissances linguistiques telles que les phonèmes et la prononciation des mots, les systèmes *end-to-end* apprennent à associer des vecteurs de paramètres acoustiques à des phonèmes, caractères, parties de mots ou mots directement à partir des données. Petit à petit, ces systèmes

deviendront les plus utilisés dans le domaine de la reconnaissance automatique de la parole grâce à leur simplicité de déploiement, leur capacité d'apprentissage et leurs bonnes performances.

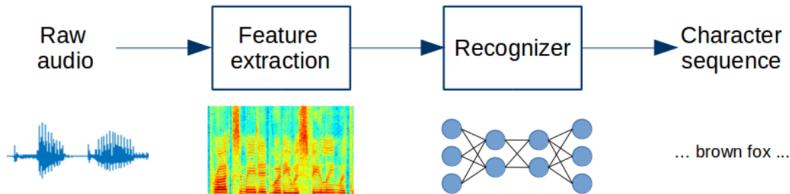


FIGURE I.4 – Architecture générale d'un système de reconnaissance automatique de la parole (*recognizer* sur la figure) *end-to-end* (Vazhenina et Markov, 2020)

Dans cette section, nous commençons par présenter les quatre grands types de réseaux neuronaux utilisés en reconnaissance automatique de la parole : les FNN (*Feedforward Neural Networks*, réseaux de neurones à propagation avant), les RNN (*Recurrent Neural Networks*, réseau de neurones récurrents), les LSTMN⁷ (*Long-Short Term Memory Networks*, réseaux à mémoire à court et long terme) et les CNN (*Convolutional Neural Networks*, réseaux de neurones convolutionnels). Ensuite, nous présentons les quatre grands types d'architectures basées sur les réseaux neuronaux pour la reconnaissance automatique de la parole : les systèmes hybrides, les systèmes à base de CTC (*Connectionist Temporal Classification*, classification temporelle connectioniste), les systèmes à base de transducer et les modèles à base d'attention. Enfin, nous présentons le principe de modèles dits “pré-appris” (*pre-trained models*).

2.1 Les types de réseaux neuronaux utilisés en reconnaissance automatique de la parole

Cette section, bien que dédiée à la présentation des différents types de réseaux neuronaux utilisés en reconnaissance automatique de la parole, n'ira pas trop en profondeur dans le détail des concepts mathématiques qui les régissent. Celle-ci a plutôt vocation à faire comprendre le fonctionnement général des différents types de réseaux et d'en saisir les forces et les faiblesses.

7. La plupart du temps notés simplement LSTM. C'est l'écriture que nous adopterons dans ce manuscrit par la suite.

2.1.1 Le fonctionnement général d'un réseau de neurones et réseau de neurones à propagation avant (FNN)

Les réseaux de neurones artificiels ont été développés avec l'idée de copier le fonctionnement du cerveau humain (Malekian et Chitsaz, 2021). Le premier type de neurone artificiel apprenable a avoir été développé est le perceptron de Rosenblatt (1958), présenté en figure I.5.

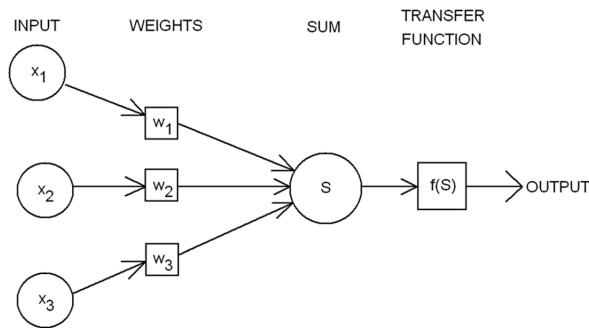


FIGURE I.5 – Le Perceptron de Rosenblatt (1958) (image tirée de Alves (2006))

L'ensemble des entrées x_1 , x_2 et x_3 à gauche de l'image forme ce qu'on appelle la couche d'entrée, le terme de couche se référant à un ensemble de neurones de même niveau. Ces neurones sont multipliés par des poids (*weights*), additionnés et passés au travers d'une fonction d'activation (ou *transfer function*) qui retourne un nombre et permet de déterminer la sortie (*output*) du réseau de neurones. Les poids servent à indiquer au réseau l'importance des informations en entrée du neurone.

Dans l'optique de pouvoir développer des modèles neuronaux plus complexes, le perceptron a ensuite évolué en modèle multicouches, tel que présenté en figure I.6. On retrouve dans le perceptron multicouches les couches d'entrées et de sortie, entrecoupées d'une ou plusieurs couche(s) cachée(s) (*hidden layer*). Le réseau de neurones est dit à propagation avant (*Feedforward*) car les informations transittent d'une couche à l'autre vers l'avant, sans retour en arrière. Les FNN sont souvent utilisés en reconnaissance automatique de la parole. Ils sont parfois caractérisés comme entièrement connectés⁸ lorsque chacun des neurones d'une couche est connecté à chacun de neurones de la couche précédente.

Un biais est ajouté à la somme pondérée afin de déplacer le résultat de la fonction d'activation vers le positif ou le négatif (il sert de seuil). Le calcul effectué à

8. on parle alors de couche entièrement connectée, dense ou linéaire

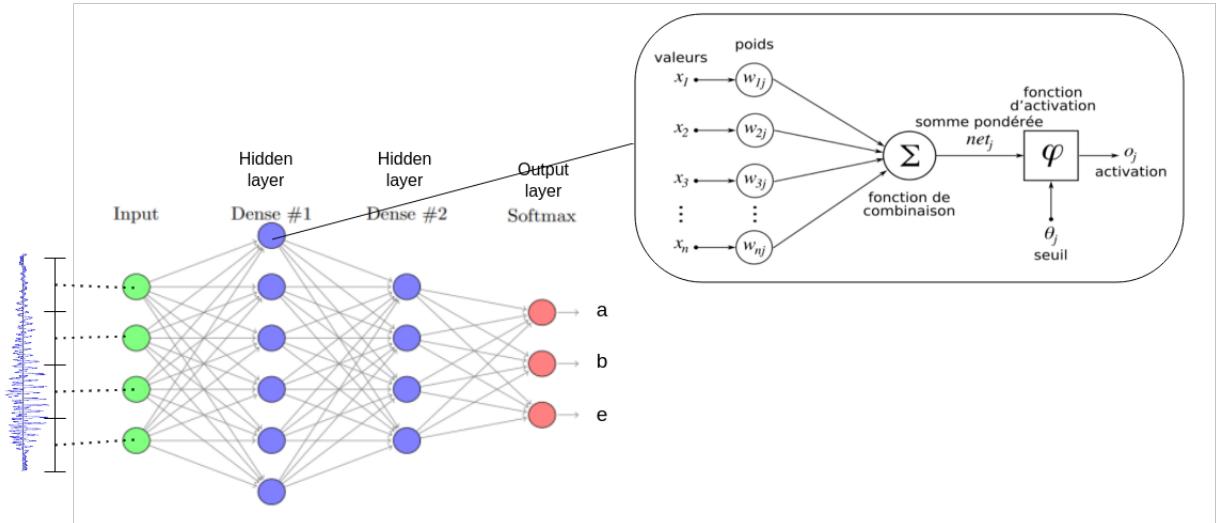


FIGURE I.6 – Réseau de neurones entièrement connectés à propagation avant (sur la base d'une image de (Pelletier *et al.*, 2019))

l'intérieur d'un neurone pour une telle architecture est donc le suivant :

$$g(w_1x_1 + w_2x_2 + \dots + w_nx_n + b) \quad (\text{I.5})$$

où g représente la fonction d'activation, w les poids, x_i les entrées et b le biais. La fonction Softmax est souvent utilisée pour la dernière couche du réseau (Boyer *et al.*, 2022) pour faire de la classification. Elle permet d'attribuer à chaque sortie une probabilité d'appartenance à une classe (caractère, phonème, mot).

L'apprentissage du système se fait grâce à une fonction de coût (aussi appelée *loss* ou fonction d'erreur) permettant de calculer la différence entre la sortie automatique et ce que le système devrait produire, et ainsi de déterminer l'implication de chaque paramètre (poids et biais) dans cette erreur. Les poids et biais du réseau seront corrigés afin de minimiser cette différence. C'est ce que l'on appelle la *backpropagation* (rétropropagation du gradient). Le calcul des poids, le calcul de l'erreur et la rétropropagation/correction des poids forment, lorsque l'ensemble des données d'apprentissage a été vu, une époque. Plusieurs époques sont nécessaires pour optimiser le réseau de neurones.

Les réseaux entièrement connectés à propagation avant présentent une bonne capacité de généralisation. Il est également possible de les utiliser directement pour l'extraction de paramètres acoustiques, sans passer par une représentation intermédiaire tels que les MFCC. Néanmoins, ils demandent un coût de calcul important et ne sont pas bien armés pour traiter les données séquentielles telles que la parole et modéliser les dépendances temporelles entre les entrées (Shewalkar *et al.*, 2019).

Ils nécessitent également un découpage ou une normalisation des données en entrée afin qu'elles aient toutes la même taille. Ces deux derniers points ont pu être traités avec les réseaux de neurones récurrents que nous présentons dans la section suivante.

2.1.2 Les réseaux de neurones récurrents (RNN)

Les réseaux de neurones récurrents sont un type spécifique de réseaux de neurones et ont été développés spécifiquement pour le traitement de données séquentielles pouvant prendre des tailles variables (Jordan, 1989; Elman, 1990). Il en existe plusieurs types (voir la figure I.7), le format *many to many* étant utilisé en reconnaissance automatique de la parole (Robinson, 1994).

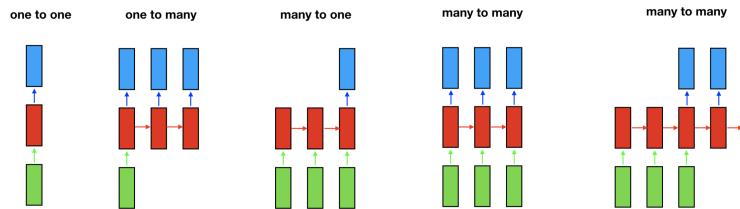


FIGURE I.7 – Différents types de RNN (Li, 2023)

Le principe des RNN est d'ajouter une boucle de rétroaction au niveau des couches cachées. Grâce à cela, le système de reconnaissance automatique de la parole est en capacité de retenir des informations passées (ce qui rappelle l'état caché des HMM), et de les utiliser conjointement aux entrées en cours pour prédire une sortie (voir figure I.8). L'état caché h_0 est la plupart du temps initialisé de façon aléatoire, tout comme les poids des réseaux de neurones traditionnels. Pour chaque entrée, une erreur est calculée. Une fois l'ensemble des entrées traitées, la moyenne est faite et un algorithme de descente de gradient particulier est utilisé afin de mettre à jour les poids et biais : la *Backpropagation Through Time* (descente de gradient à travers le temps) (Werbos, 1990; Ahmad *et al.*, 2004).

Les RNN sont donc une alternative intéressante aux FNN vus précédemment. Toutefois, ce type d'architecture peine à traiter les dépendances à long terme : un contexte trop grand peut entraîner une explosion du gradient, et donc une divergence lors de l'apprentissage (Bengio *et al.*, 1994; Shewalkar *et al.*, 2019). De plus, ils sont encore une fois assez coûteux en termes de calculs.

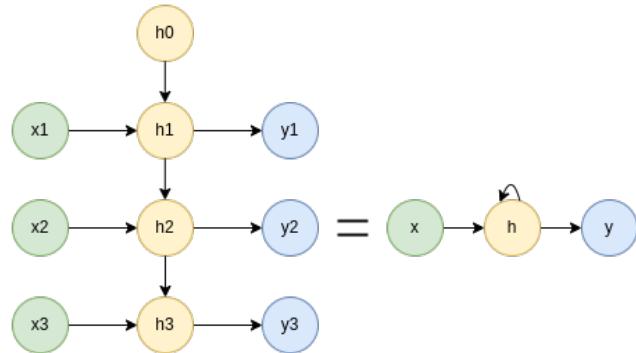


FIGURE I.8 – Architecture d'un réseau de neurones récurrents (RNN)

2.1.3 Réseaux à mémoire à court et long terme (LSTM)

Les réseaux de neurones LSTM sont eux aussi utilisés pour le traitement de données séquentielles (Hochreiter et Schmidhuber, 1997; Graves *et al.*, 2004, 2013). Ils présentent une meilleure résistance au problème d'explosion de gradient que les RNN (Li *et al.*, 2016) et permettent d'utiliser deux types de mémoires complémentaires : la mémoire à long terme ou *cell state* (la mémoire de ce qu'il s'est passé depuis le début de l'époque), et la mémoire à court terme ou *hidden state* (la mémoire immédiate, ce qui s'est passé au temps $n-1$) (Hochreiter et Schmidhuber, 1997).

Chacun des neurones de ce type de réseau est constitué de trois éléments comme le montre la figure I.9 : une porte d'entrée (*input gate*), une porte de sortie (*output gate*) et une porte d'oubli (*forget gate*) (Graves *et al.*, 2004).

La porte d'entrée est celle qui décide si le réseau doit modifier le contenu de la mémoire à long terme ou non. La porte d'oubli permet de remettre la mémoire à zéro. La porte de sortie permet de décider si le contenu de la mémoire doit influer sur la sortie du neurone ou non. Un LSTM prend 3 vecteurs en entrée : la mémoire à long terme, l'état caché ou la mémoire à court terme et le vecteur d'entrée.

Ce type de réseau est, tout comme les FNN et RNN, très gourmand en ressources computationnelles (Masuko, 2017). De plus, des problèmes d’explosion ou disparition du gradient restent possibles et ces réseaux sont sensibles aux situations d’*overfitting* (sur-apprentissage) (Graves *et al.*, 2013), ce qui arrive quand le réseau n’a pas capturé les lois générales qui régissent les données d’apprentissage, mais a appris les données d’apprentissage par cœur.

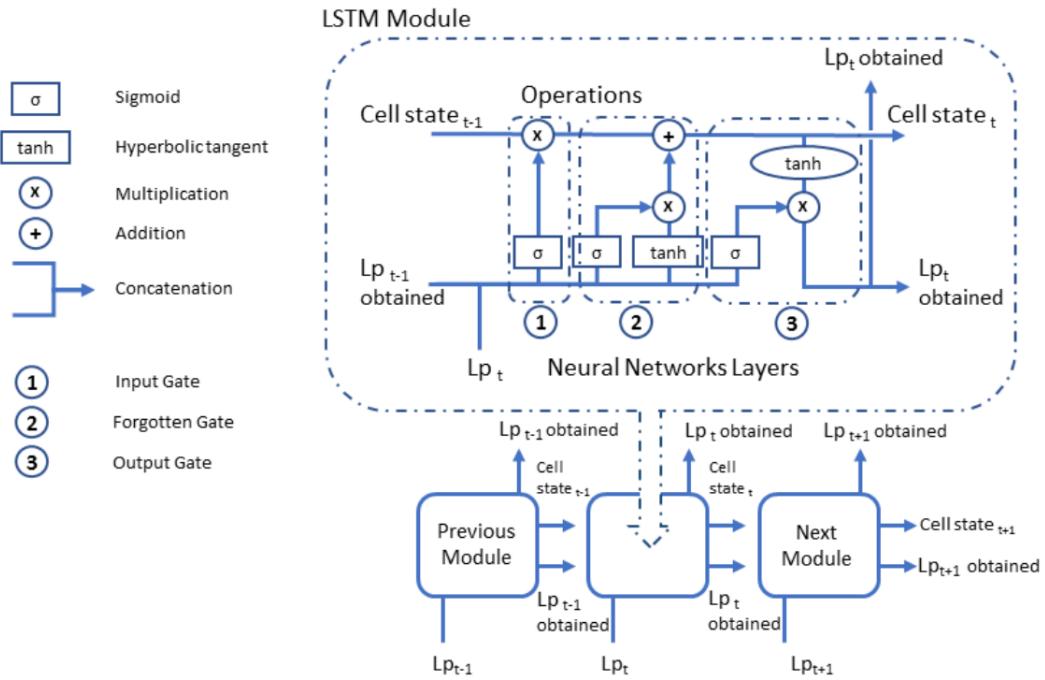


FIGURE I.9 – Réseau de neurones LSTM (*Long-Short Term Memory*) (Navarro et al., 2020)

2.1.4 Réseaux de neurones convolutionnels (CNN)

Le réseau de neurones convolutionnels (Lecun et al., 1998) est un type de réseau de neurones profond qui a d'abord été utilisé en classification d'images et dont le but est d'apprendre à reconnaître des formes en se basant sur des couches de filtres. Palaz et al. (2013, 2015) les ont utilisés pour créer un système de reconnaissance automatique de la parole *end-to-end*, mais ils peuvent aussi être utilisés en remplacement des réseaux de neurones profonds traditionnels utilisés dans une architecture HMM-DNN (Abdel-Hamid et al., 2014).

Ce type de réseau comprend deux étapes : l'extraction de paramètres acoustiques et la classification (voir figure I.10).

Il est possible de donner directement des signaux bruts en entrée. Se suivent ensuite des étapes de convolution et de *pooling*, où chaque couche apprend des représentations de plus en plus abstraites. Une convolution est un filtre qui se déplace sur l'image d'entrée afin de reconnaître des motifs. Ces filtres sont appris par le réseau lors de l'étape d'apprentissage et sont généralement plus petits que la taille de l'image d'entrée⁹. L'étape de *pooling* sert, elle, à sous-échantillonner les entrées,

9. Dans le cadre de la reconnaissance automatique de la parole, l'image d'entrée est un spectrogramme.

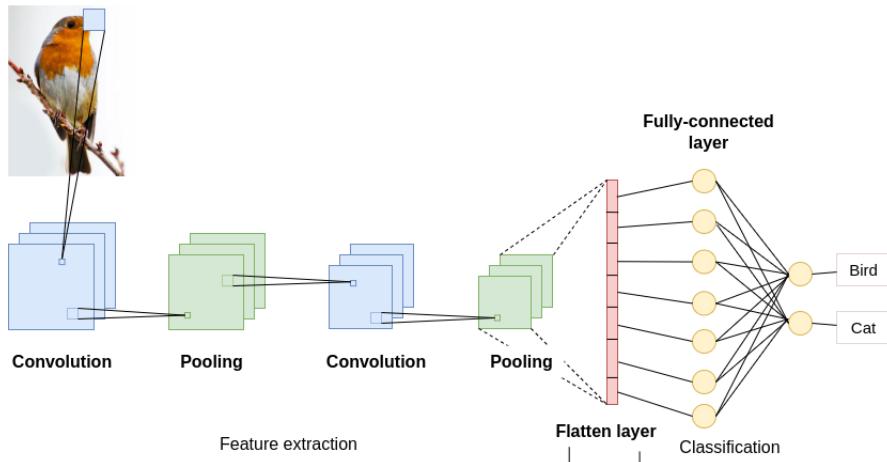


FIGURE I.10 – Réseau de neurones convolutionnels. Figure inspirée de (Phung et Rhee, 2019).

et donc à réduire la complexité des calculs par la suite (Meng *et al.*, 2023). L’ensemble des filtres de ces couches cachées sont concaténés puis suivis par une couche entièrement connectée juste avant la couche de sortie, afin de faire en sorte que la classification puisse se faire sur l’entièreté de l’image et non juste localement.

Le CNN est un réseau de neurones très efficace pour la reconnaissance d’images. Il présente en parole l’avantage de pouvoir composer un système de reconnaissance de la parole entièrement neuronal puisque ce type de réseaux est également en capacité d’extraire des paramètres acoustiques

Les réseaux présentés ici (FNN, RNN, LSTM, CNN) présentent chacun des avantages et inconvénients. Ils peuvent être utilisés comme tels (assemblés ou non), ou bien comme élément de base d’une architecture de plus haut niveau comme nous allons le voir dans la suite de cette section, où nous présenterons les architectures les plus utilisées aujourd’hui.

2.2 Les systèmes hybrides HMM-DNN

L’utilisation de réseaux de neurones en remplacement des GMM a été explorée dès 1994 par (Bourlard et Morgan, 1994). Le manque de ressources computationnelles et de données a néanmoins mis en pause cette piste d’amélioration des systèmes de reconnaissance automatique de la parole (Liu *et al.*, 2015).

C’est au début des années 2010 que des chercheurs de l’université de Toronto, programme (Musaev *et al.*, 2019)

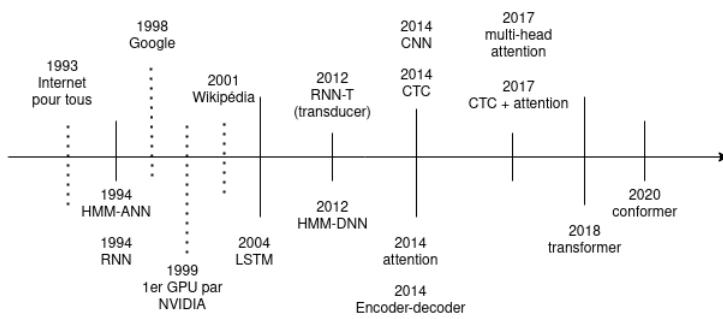
de Microsoft Research, de Google Research et IBM Research ont relancé ces recherches et proposé de remplacer les GMM par des réseaux de neurones profonds (DNN pour *Deep Neural Network*) pour assurer la classification des vecteurs acoustiques en phonèmes dans les systèmes statistiques (Hinton *et al.*, 2012). Ces années marqueront une relance des recherches sur l'utilisation de réseaux de neurones en reconnaissance automatique de la parole comme le montre la figure I.11. Ce nouveau souffle va de paire avec le développement des premiers GPU au début des années 2000¹⁰ et le développement croissant d'Internet depuis avril 1993¹¹, source précieuse de données (Abdel-Hamid *et al.*, 2014). Ces évolutions ont mené aux systèmes de reconnaissance automatique de la parole entièrement neuronaux puis aux modèles pré-appris présentés dans la section 2.6.

10. <https://www.computer.org/publications/tech-news/chasing-pixels/nvidias-geforce-256>

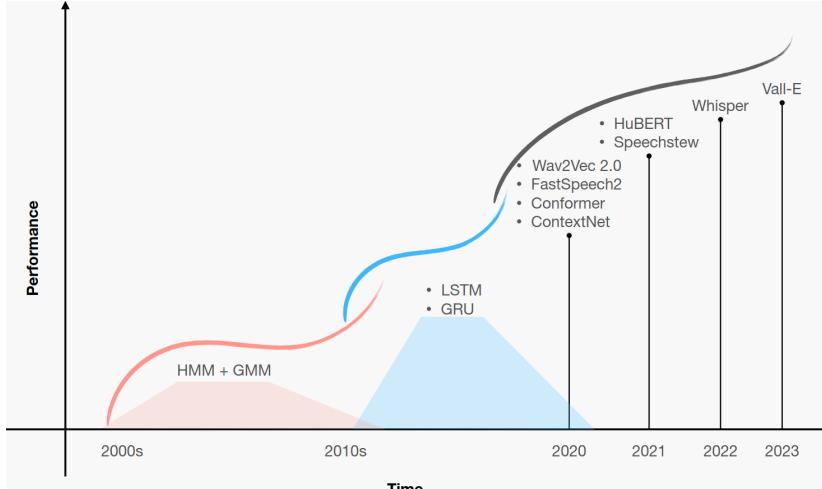
11. <https://www.humanite.fr/culture-et-savoirs/il-etait-une-fois/tout-ce-qui-change-ce-30-avril-1993-quand-le-world-wide-web-est-devenu-public-790504>

2 Les approches neuronales

A.



B.



(Graphique B tiré de (Mehrish *et al.*, 2023))

FIGURE I.11 – Dates-clés des avancées en reconnaissance automatique de la parole neuronale

2.3 Les systèmes à base de classification temporelle connexioniste (CTC)

La classification temporelle connexioniste (Graves *et al.*, 2006) n'est pas à proprement parler une architecture neuronale. C'est une technique qui s'utilise avec des réseaux traitant les données séquentielles et qui permet à la fois d'assurer un alignement monotone entre un signal et sa transcription (Shao et Feng, 2022), et de calculer la *loss* d'un modèle (Talnikar *et al.*, 2021).

Elle présente l'avantage de pouvoir utiliser des données non alignées pour l'entraînement de modèles (Graves et Jaitly, 2014). Ce type de système est composé d'un encodeur qui prend en entrée un signal audio découpé en x fenêtres et sert à encoder une séquence de taille variable en un vecteur de paramètres h de taille fixe. Ces paramètres sont ensuite envoyés à une couche Softmax permettant de prédire

la probabilité d'avoir un caractère y appartenant à l'ensemble des symboles de sortie, selon l'entrée x , comme le montre la figure I.12.

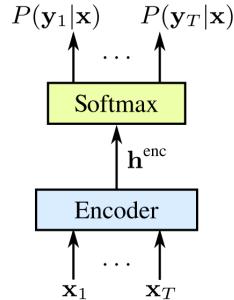


FIGURE I.12 – Classification temporelle connectionniste (CTC) (image tirée de (Prabhavalkar *et al.*, 2017))

La CTC possède une règle de compression, qui lui permet de gérer les différences de taille entre l'entrée et la sortie du réseau en prenant en compte le fait que certains labels (caractères) pourront être répétés en sortie, simplement dû au fait qu'ils auront été prononcés sur plusieurs fenêtres de temps. Néanmoins, toute suite d'un même caractère n'est pas toujours à fusionner. Prenons le mot “balle” par exemple : il faudra pouvoir indiquer au système que deux “l” à la suite sont attendus. Cette spécificité est rendue possible grâce à un caractère blanc (noté ε dans la figure I.13) qui est ajouté aux symboles de sortie et placé entre deux caractères similaires (Anoop et Ramakrishnan, 2021).

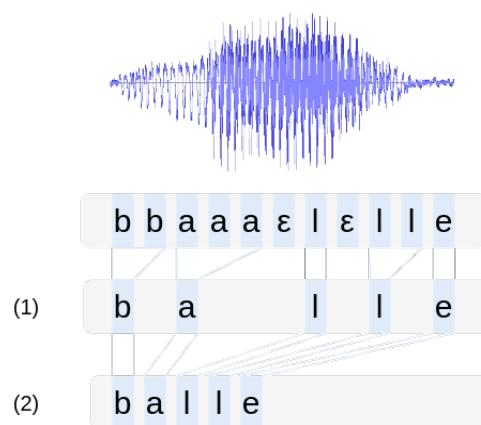


FIGURE I.13 – Exemple de traitement de la sortie d'un réseau de neurones par la CTC : (1) Les caractères qui sont répétés et qui se suivent sont fusionnés. (2) Le caractère blanc est supprimé.

Ce caractère blanc est également utilisé afin de tester tous les alignements possibles entre la séquence $x_1 \dots x_T$ en entrée et la séquence de caractères $y_1 \dots y_T$ cible. Le système apprend ainsi à insérer des espaces au bons endroits grâce au calcul de la *loss*.

Un autre des avantages de ce type de système est sa faible demande computationnelle lorsqu'un algorithme de programmation dynamique est appliqué (comme l'algorithme *forward*). Néanmoins, les sorties du systèmes sont indépendantes les unes des autres, ce qui ne permet pas de prendre en compte le contexte (Lu et Chen, 2023) et le fait perdre en efficacité.

2.4 Les systèmes à base de transducteur (RNN-T)

Le RNN-T a été développé afin de compléter la CTC (Graves, 2012) : le but étant de capturer les dépendances à long terme entre les labels en sortie. Ce type de système est composé d'un encodeur, d'un réseau de prédiction (*prediction network*) et d'un réseau joint comme le montre la figure I.14.

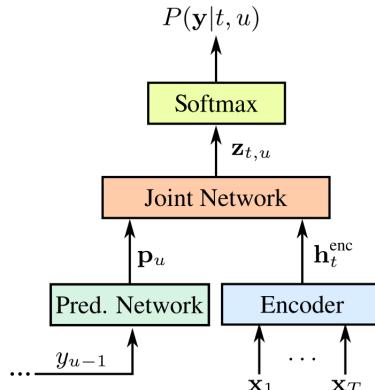


FIGURE I.14 – Schéma d'un RNN-T (image tirée de (Prabhavalkar *et al.*, 2017))

En parallèle de l'encodeur qui extrait les paramètres h au temps t , le réseau de prédiction reçoit en entrée le label précédemment reconnu y_{u-1} et construit un vecteur p qui dépend de l'entièreté de la séquence de labels $[y_0 \dots y_{u-1}]$ ¹² (Albesano *et al.*, 2022). Ce réseau est parfois comparé à un modèle de langue, bien que cela ne fasse pas l'unanimité (Albesano *et al.*, 2022). Les sorties de l'encodeur et du réseau de prédiction sont toutes deux utilisées par le réseau joint afin de prédire la probabilité de distribution du label courant grâce à la fonction Softmax.

12. Le tout premier label passé en entrée du réseau de prédiction étant *SOS* pour *Start of Sentence*.

Ce type d'architecture est très gourmande en ressources computationnelles au moment de l'apprentissage (Kuang *et al.*, 2022) mais beaucoup moins en inférence, étape qui consister à utiliser le modèle sur de nouvelles données une fois qu'il a été appris. Elle est notamment utilisée pour les tâches avec un traitement en *streaming* (He *et al.*, 2019), c'est-à-dire en temps réel.

2.5 Les systèmes à base d'attention

2.5.1 L'architecture encodeur-décodeur

Les systèmes utilisant l'attention se basent sur des architectures encodeur-décodeur. Le type d'architecture encodeur-décodeur a été développée en 2014 (Cho *et al.*, 2014) pour la traduction automatique. Elle est composée de deux blocs (voir figure I.15) : l'encodeur, et le décodeur, qui permet de décoder ce vecteur de taille fixe en une séquence de sortie de taille variable.

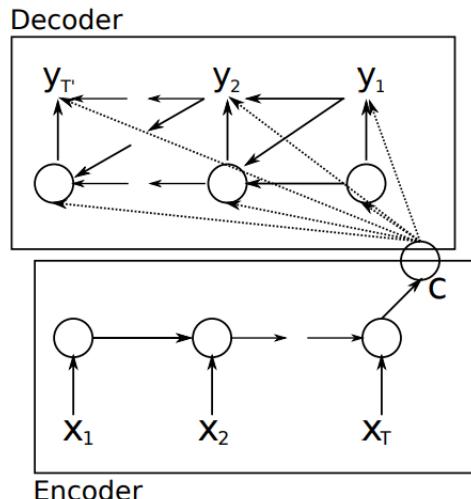


FIGURE I.15 – L'architecture encodeur-décodeur de Cho *et al.* (2014)

Les entrées sont traitées l'une après l'autre, et modifient tour à tour l'état caché de la couche RNN de l'encodeur. Le vecteur C correspond à l'état caché de la dernière entrée, et représente donc l'entièreté de la séquence d'entrée. Des informations sur la sémantique et la structure syntaxique de la phrase y sont capturées. Le décodeur génère la séquence de sortie en s'appuyant sur le vecteur C ainsi que sur les états cachés et les sorties qui précèdent celle en cours.

La limite majeure de ce type d'architecture est la capacité de traitement de longues séquences au sein d'un vecteur C de taille fixe (Bahdanau *et al.*, 2014; Pouget-Abadie *et al.*, 2014). L'étude de (Bahdanau *et al.*, 2014) a permis de pallier ce

problème en utilisant un mécanisme d'attention.

2.5.2 L'attention

2.5.2.1 Mécanismes d'attention globales et locales

Les recherches de (Bahdanau *et al.*, 2014) avaient pour but d'apporter une solution au problème de traduction automatique de longues séquences, cas problématique pour une architecture encodeur-décodeur classique. Ils y ont ajouté un mécanisme d'attention permettant de calculer la probabilité pour chacune des entrées d'être pertinente pour le décodage de l'état en cours. La figure I.16 présente son fonctionnement.

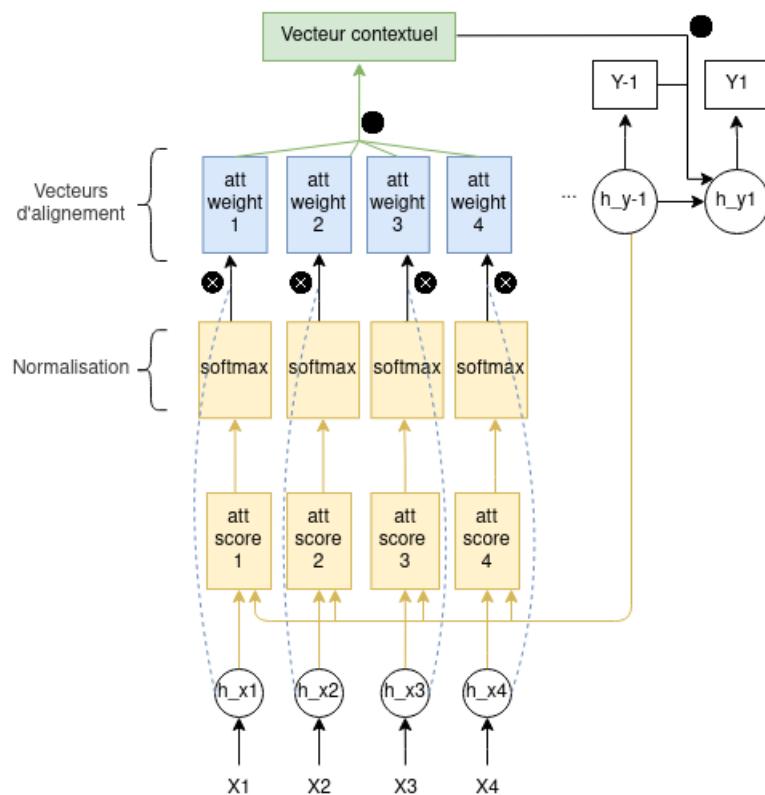


FIGURE I.16 – Schéma explicatif du mécanisme d'attention (adapté de Bahdanau *et al.* (2014))

Tout d'abord, les états cachés de chacune des entrées sont générés. Ensuite, des scores d'alignement sont calculés, permettant de faire ressortir l'importance de chaque entrée pour le décodage en cours. Chacun de ces scores est passé au travers d'une fonction Softmax, ayant pour but de normaliser les scores entre 0 et 1.

Les scores sont ensuite multipliés par les états cachés en entrée afin de calculer les poids d'attention. Ces poids seront utilisés afin de faire ressortir l'importance relative de chacune des entrées entre elles, pour le décodage en cours. Ces poids sont additionnés et forment alors le vecteur contextuel. Celui-ci est concaténé à la sortie précédente (Y_{-1}) et donné en entrée à l'état caché en cours, en plus de l'état caché précédent.

Les travaux de (Luong *et al.*, 2015) dérivent de l'attention de Bahdanau et en proposent une architecture simplifiée (schéma de gauche dans la figure I.17). Ce type d'attention permet de traiter des séquences d'une longueur plus conséquente. Néanmoins, un système d'attention global présente des inconvénients comme le haut coût computationnel et les problèmes d'alignements qui peuvent survenir lorsque les entrées sont trop longues (Tjandra *et al.*, 2017).

En plus de proposer une autre version de l'attention de Bahdanau, Luong propose également dans (Luong *et al.*, 2015) un mécanisme d'attention local. A la différence du système d'attention global, celui-ci ne s'appuie que sur un sous-ensemble des entrées (schéma de droite dans la figure I.17).

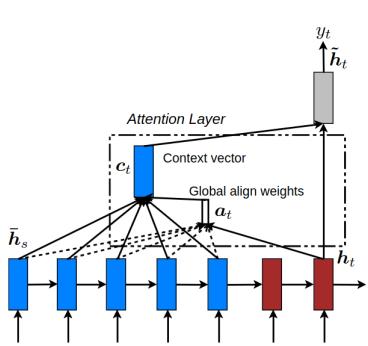


Figure 2: **Global attentional model** – at each time step t , the model infers a *variable-length* alignment weight vector a_t based on the current target state h_t and all source states \bar{h}_s . A global context vector c_t is then computed as the weighted average, according to a_t , over all the source states.

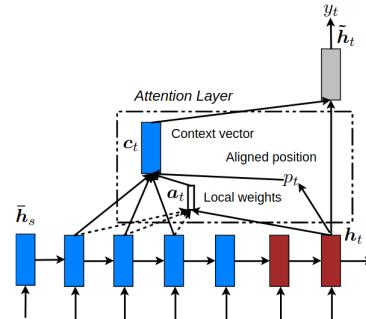


Figure 3: **Local attention model** – the model first predicts a single aligned position p_t for the current target word. A window centered around the source position p_t is then used to compute a context vector c_t , a weighted average of the source hidden states in the window. The weights a_t are inferred from the current target state h_t and those source states \bar{h}_s in the window.

FIGURE I.17 – Attentions globales et locales par (Luong *et al.*, 2015)

2.5.2.2 Auto-attention et transformers

En 2017, des chercheurs de Google ont présenté un moyen pour l'attention de se suffire à elle-même dans l'encodeur, en enlevant les réseaux de neurones récurrents (Vaswani *et al.*, 2017). Ce travail s'inspire des travaux de (Cheng *et al.*, 2016), dont l'idée était de pouvoir capturer grâce au LSTM les relations entre les dif-

férents mots¹³. C'est maintenant le mécanisme d'auto-attention en lui-même qui permet de capturer les dépendances. Par rapport aux mécanismes d'attention vus précédemment, c'est la granularité qui change : l'attention ne fournira plus d'informations concernant l'importance d'un mot en position X par rapport à la séquence entière, mais plutôt l'importance d'un mot Y en position X par rapport à un mot M en position Z dans le texte en entrée. La figure I.18 en présente l'architecture.

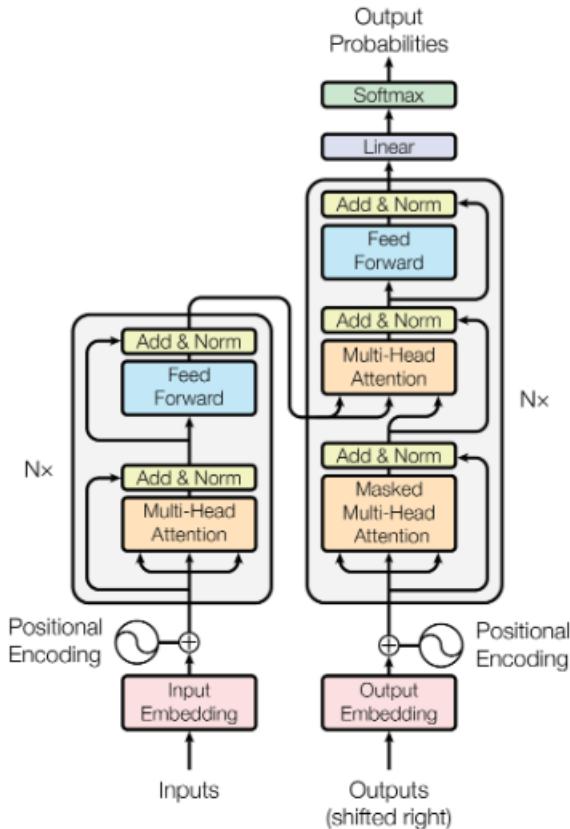


FIGURE I.18 – L'architecture Transformer de Vaswani *et al.* (2017)

Dans l'auto-attention, trois matrices de poids sont utilisées : la *query* (requête), la *key* (clé) et la *value* (valeur). Les mots en entrée, sous forme d'*embedding* (plongement de mot), sont multipliés par chacune de ces matrices générées pendant la phase d'apprentissage. En parallèle est généré un score d'attention grâce au *dot product* (produit scalaire) entre la *query* de l'input courant et la *key* de chacune des autres entrées. On obtient donc autant de scores d'attention qu'il y a

13. Nous parlons ici de mots car cette technique a été développée pour la lecture via une machine. Cette technique aura ensuite été appliquée sur le signal avec comme équivalent aux mots, une séquence sonore.

d'entrées. Après les avoir normalisés, ils sont multipliés par les valeurs correspondantes et additionnées. Ceci donne des scores d'attention relatifs à l'importance de la dépendance entre chaque mot. Les sorties sont ensuite concaténées en un seul vecteur. Dans l'architecture Transformer de (Vaswani *et al.*, 2017), l'auto-attention est utilisée plusieurs fois en même temps avec des matrices de poids différentes. Un processus d'auto-attention est appelée une tête et l'architecture Transformer classique en comprend huit en parallèle, autrement appelée la *multi-head attention*.

2.6 Les modèles pré-appris

Le pré-apprentissage est une technique visant à apprendre un modèle sur une tâche appelée “tâche prétexte”, souvent très générique¹⁴, pour ensuite utiliser les paramètres appris comme base d'un modèle appris pour une autre tâche appelée *downstream task* (tâche en aval) (Mohamed *et al.*, 2022). Cette technique est coûteuse en ressources (Mohamed *et al.*, 2022), mais les possibilités d'adaptation ou de réutilisation telle quelle du modèle pré-appris permettent de partager les modèles à la communauté et donc de limiter l'entraînement de modèles qui ne seraient utilisés que pour une étude. On parle alors de modèles de fondation ou *foundation models*.

2.6.1 Le fonctionnement général des modèles pré-appris

L'apprentissage peut se faire de façon supervisée (Radford *et al.*, 2022), non-supervisée (Chorowski *et al.*, 2019) ou auto-supervisée (Baevski *et al.*, 2020). L'apprentissage supervisé utilise des données étiquetées afin d'en apprendre des représentations. L'apprentissage non supervisé quant à lui utilise des données non-étiquetées. Enfin, l'apprentissage auto-supervisé est à la croisée des deux : le modèle génère ses propres étiquettes à partir des données d'entraînement et les utilise ensuite pour l'apprentissage de représentations intermédiaires grâce à un entraînement supervisé. Cette technique présente l'avantage de faire gagner du temps lors de la préparation des données (Ericsson *et al.*, 2022) (l'annotation manuelle est très coûteuse en temps, notamment la transcription de données audio), mais aussi d'exploiter de grands ensembles de données non transcrrites.

Une première tâche servant à entraîner le modèle pré-appris est appelée la tâche prétexte. Il en existe plusieurs : *predictive coding* (Song *et al.*, 2019; Baevski *et al.*, 2020; Liu *et al.*, 2020), *pretext-task label learning* (Pascual *et al.*, 2019; Ravarrelli *et al.*, 2020), *auto-encoding* (Renshaw *et al.*, 2015; Algayres *et al.*, 2020)¹⁵. Mohamed *et al.* (2022) a classé les différentes tâches prétextes existantes selon

14. comme apprendre à trouver le mot manquant

15. voir (Zaiem *et al.*, 2022) pour une liste plus complète

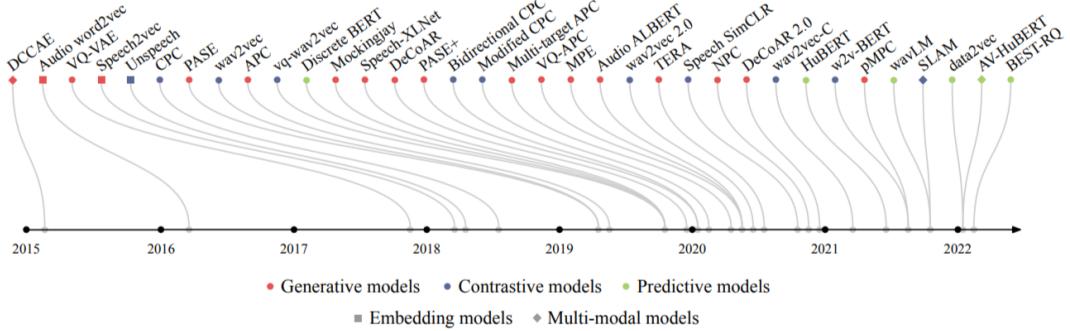


FIGURE I.19 – Classification de tâches prétextes pour l'apprentissage de modèles pré-appris pour l'audio selon (Mohamed *et al.*, 2022)

trois catégories, que nous reprenons ici : les approches génératives, les approches contrastives et les approches prédictives. Celles-ci se distinguent par leur objectif, la fonction de coût utilisée ou encore leur méthode d'apprentissage, comme décrit ci-après.

Approches génératives : l'objectif du modèle est de générer ou reconstruire les données d'entrées. Cela inclut par exemple de prédire le futur grâce au passé, de prédire les éléments masqués grâce aux éléments non masqués ou encore de prédire la version originale d'un élément altéré¹⁶.

Approches contrastives : le modèle apprend à générer des représentations distinctes pour les paires de données similaires et différentes grâce à une fonction de coût contrastive qui attribue un haut score aux paires similaires et un score bas aux paires différentes. Des éléments “négatifs” sont ainsi introduits et comparés avec des exemples “positifs” afin que le modèle puisse apprendre à les distinguer.

Approches prédictives : le modèle apprend à prédire des éléments manquants en se basant sur les données d'entrée. Les fonctions de coût utilisées sont notamment la cross-entropie et l'erreur quadratique (*squared error*). Ce type d'approche utilise la plupart du temps une configuration d'apprentissage dite “professeur-étudiant” (*teacher-student*).

La figure I.19 présente la classification de plusieurs modèles pré-appris en fonction de ces différentes approches.

16. (Mohamed *et al.*, 2022) précisent : “‘Generative’ as used in this paper hence refers to models that target the original input in their pretext task. Note that this differs from generative models, which learn distributions that allow to sample new data.”

Les représentations intermédiaires apprises grâce à la tâche prétexte¹⁷ sont ensuite utilisées pour une tâche en aval. Ainsi, les connaissances plus génériques que le modèle a pu apprendre grâce à la première tâche seront utilisées pour résoudre plus facilement les tâches en aval, nécessitant des connaissances plus spécifiques. La reconnaissance automatique de la parole est un exemple de tâche en aval. L'avantage de l'utilisation de modèles pré-appris est le besoin réduit en données étiquetées ou transcrrites pour la deuxième tâche.

Deux utilisations du modèle pré-appris sont possibles : son adaptation, lui permettant de continuer à apprendre le modèle en le faisant coller le plus possible à la tâche à venir, ou l'utilisation du modèle tel quel, comme simple extracteur de paramètres. Pour la reconnaissance automatique de la parole, le modèle pré-appris pourra être couplé à des couches de FNN, à des couches de bi-LTSM... Ce sont ces couches additionnelles, couplées à une couche de classification ajoutée avant la couche de sortie, qui donneront au système de reconnaissance automatique de la parole la capacité de générer les classes voulues (des caractères, phonèmes, des *byte pair encoding* ou encore des mots en reconnaissance automatique de la parole).

2.6.2 Les modèles pré-appris pour le français

Quelques modèles pré-appris pour le français sont aujourd’hui disponibles et sont principalement multilingues : XLS-R (Arun Babu *et al.*, 2022), XLSR-53 (Conneau *et al.*, 2021), MMS (Pratap *et al.*, 2023), Maestro (Chen *et al.*, 2022), mSLAM (Bapna *et al.*, 2022) et Whisper (Radford *et al.*, 2022), dont le tableau I.2 présente les architectures, le nombre de langues et les données d’apprentissage.

Modèle	Architecture	#Langues	Données d’entraînement
XLSR-53	wav2vec 2.0	53	-
XLS-R	wav2vec 2.0	128	Voxlingua, Babel, CV, Voxpopuli, MLS
Whisper	enc-dec transformer	96	680 000h de données du Web
Maestro	shared text and speech encoders + RNN-T decoder	101	Voxpopuli, CV, MLS, Babel
mSLAM	wav2vec-BERT	51	Babel, CV, Voxpopuli, MLS
MMS	wav2vec 2.0	1 406	extraits de la Bible, MLS, CV, Voxlingua, Babel, Voxpopuli

TABLE I.2 – Modèles pré-appris multilingues incluant le français (liste non-exhaustive)

Les performances des modèles monolingues étant souvent meilleures¹⁸, un groupe-

17. à considérer comme une tâche en amont

18. du moins lorsque la langue en question dispose de beaucoup de ressources

ment de chercheurs a souhaité fournir des modèles monolingues pour le français. C'est ainsi que nous avons pris part au projet LeBenchmark. Ce travail d'équipe a été publié dans (Evain *et al.*, 2021b) et est décrit dans les parties 2 et 3 de ce manuscrit.

Nous présentons dans les sections suivantes les modèles Whisper, ainsi que l'architecture Wav2Vec2.0 (Baevski *et al.*, 2020), largement utilisée dans les systèmes pré-appris et utilisés pour les modèles LeBenchmark.

2.6.3 Les modèles multilingues Whisper

Les modèles Whisper ont été développés dans le but de proposer un système performant sans adaptation spécifique à un ensemble de données, robuste, et unique, permettant divers traitements de la parole. En effet, si les modèles sont multilingues, ils sont également multitâches. Ils permettent de faire à la fois de la reconnaissance automatique de la parole, de la traduction, de l'identification de langues et de la détection de voix. Ils ont été entraînés avec 680 000 heures de d'enregistrements, parmi lesquelles 9 752 heures de français et 4 481 heures de français accompagnées de leur traduction en anglais. La liste des corpus utilisés pour l'apprentissage n'est pas partagée à la communauté.

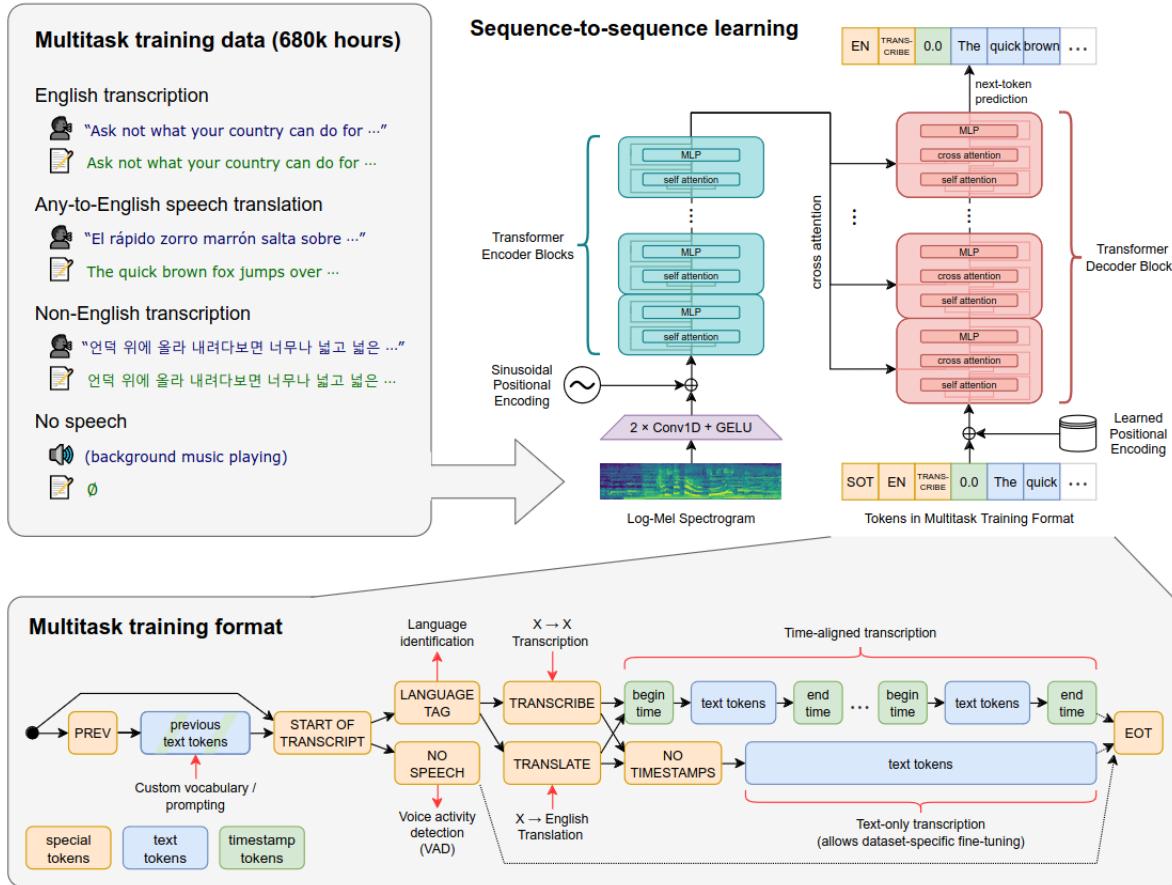
La technique d'apprentissage utilisée est qualifiée de “*large-scale weak supervision*” (faiblement supervisé à grande échelle). La faible supervision signifie que les modèles utilisent à la fois des données étiquetée manuellement¹⁹ et des données non annotées ou dont l'annotation est générée automatiquement et considérée comme bruitée. Les modèles sont appris grâce à l'architecture Transformer encodeur-décodeur de Vaswani *et al.* (2017) déjà présentée en section 2.5.2.2 de ce chapitre.

Plusieurs types de modèles sont mis à disposition de la communauté et dépendent du nombre de paramètres utilisés lors de l'apprentissage :

- *tiny* (minuscule) : 39M de paramètres
- *base* (base) : 74M de paramètres
- *small* (petit) : 244M de paramètres
- *medium* (moyen) : 769M de paramètres
- *large* (grand) : 1 550M de paramètres

La figure I.20 présente les différents types de données, l'architecture du modèle ainsi que le format des données utilisées pour apprendre à résoudre les différentes tâches avec un seul et même modèle.

19. dont la quantité est donc limitée


 FIGURE I.20 – L'approche Whisper de Radford *et al.* (2022)

Le système utilise la transcription du segment précédent afin de fournir du contexte, puis utilise des marqueurs spécifiques afin d'indiquer :

- le début de la prédiction (`<|startoftranscript|>`)
- si l'enregistrement ne contient pas de parole (`<|nospeech|>`)
- la langue identifiée
- le début de la transcription (`<|transcribe|>`)
- le début de la traduction (`<|translate|>`)
- s'il faut prédire les temps de début et de fin (horodatage) des tokens ou non (`<|notimestamps|>` est utilisé dans ce 2e cas)
- la fin de la transcription (`<|endoftranscript|>`)

Il est important de noter que le modèle est entraîné avec des enregistrements de 30 secondes et ne peut décoder des enregistrements plus longs. Lorsque les

enregistrements en entrée sont d'une durée supérieure, une fenêtre de 30 secondes est utilisée et décalée en fonction de l'horodatage prédit par le modèle afin de parcourir l'ensemble de l'enregistrement. Comme les auteurs le précisent, la qualité de la transcription de longs enregistrements repose donc notamment sur la qualité de prédiction de l'horodatage du modèle.

2.6.4 L'architecture Wav2Vec 2.0

La technique d'apprentissage de modèles Wav2Vec 2.0 de Baevski *et al.* (2020) est dite auto-supervisée. La figure I.21 présente le fonctionnement de l'architecture.

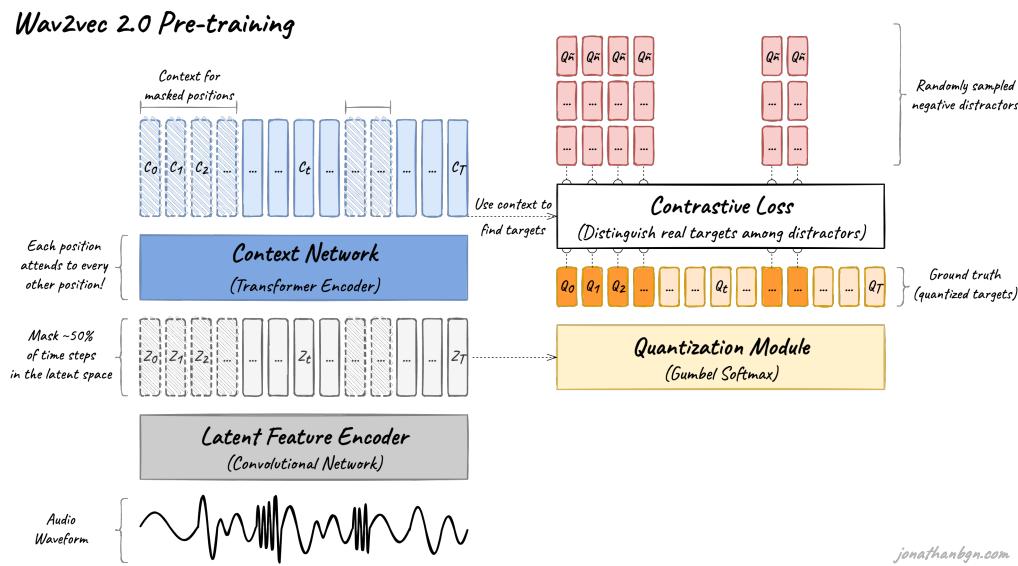


FIGURE I.21 – Architecture autosupervisée Wav2Vec 2.0 (avec l'aimable autorisation de Jonathan Boigne)

Les premières couches (7 couches de CNN, comprises dans le *latent feature encoder*) servent à apprendre des représentations vectorielles du signal de parole en entrée. Suit ensuite une étape de masquage où une partie des données est cachée : c'est la tâche prétexte. Le modèle doit s'appuyer sur le reste des données afin de déterminer ce qui a été caché. Pour cela, il utilise un Transformer, afin de pouvoir apprendre des représentations contextuelles en extrayant des *relative positional embeddings*, c'est-à-dire des *embeddings* permettant de capturer la position relative de deux vecteurs de paramètres acoustiques entre eux.

En parallèle (conjointement), des représentations quantifiées de dimension réduite des représentations latentes des CNN sont apprises. Deux *codebooks* (ou dictionnaires) de 325 unités contiennent l'ensemble des unités discrètes de parole. Ces

représentations discrètes quantifiées servent de cible pour le calcul de la fonction de coût contrastive (*contrastive loss*) dont le but est d'attribuer un score de similarité fort entre un échantillon (appelé l'ancre) et sa représentation latente (appelé échantillon positif), et de pénaliser les hauts scores de similarité entre l'ancre et des distracteurs (échantillons négatifs). Les auteurs notent une forte corrélation entre certaines entrées des deux dictionnaires et les phonèmes.

Il est possible d'apprendre des modèles *base* ou *large*, la différence entre les deux étant le nombre de paramètres constituant le modèle (95 millions *versus* 317 millions). Les modèles partagés par les auteurs sont uniquement monolingues anglais. C'est cette architecture qui a été utilisée pour l'apprentissage des modèles pour le français LeBenchmark Evain *et al.* (2021b) développés par des chercheurs des laboratoires LIG, LIA et Lamsade, ainsi que du laboratoire de recherche industrielle Naver Labs Europe.

3 L'évaluation des systèmes de reconnaissance automatique de la parole

3.1 Le WER et ses limites

Les métriques les plus utilisées pour l'évaluation de la performance de reconnaissance automatique des systèmes sont le *word error rate* (WER, taux d'erreurs de mots), le *phoneme error rate* (PER, taux d'erreurs de phonèmes), le *caractere error rate* (CER, taux d'erreurs de caractères) et le *sentence error rate* (SER, taux d'erreurs de phrases). Ceux-ci se calculent selon le même principe : le ratio du nombre d'erreurs sur le nombre de mots/phrases/caractères/phonèmes total dans la transcription de référence. Par exemple, pour le WER, la formule mathématique est la suivante :

$$\frac{ins + supp + sub}{N} \quad (I.6)$$

avec *ins* représentant le nombre d'insertions de mots, *supp* le nombre de suppression de mots et *sub* le nombre de mots substitués par un autre, le tout divisé par *N*, le nombre total de mots dans la référence.

Une insertion consiste en l'insertion d'un mot par le système de reconnaissance automatique de la parole, là où il ne devrait pas y en avoir. La substitution concerne le remplacement d'un mot par un autre. Les substitutions ne se font pas toujours de façon aléatoire : les mots de remplacement sont la plupart du temps phonétiquement proches des mots de référence. Enfin, la suppression (ou délétion) concerne la suppression d'un mot qui se trouvait dans la transcription de référence. Le tableau I.3 présente un exemple pour chaque type d'erreur.

Référence	le chat mange la souris
Insertion	le chat mange la la souris
Substitution	le chat change la souris
Suppression	le chat mange souris

TABLE I.3 – Exemples d'insertion, substitution et suppression

Bien que largement utilisées, ces métriques sont quelque peu controversées. En effet, les contextes d'utilisation ne sont pas pris en compte dans le calcul de l'erreur (Morris *et al.*, 2004; Munteanu *et al.*, 2006; Apone et O'Connell, 2010; Levit *et al.*, 2012; Helmke *et al.*, 2021). De plus, il peut arriver que les transcriptions de référence elles-mêmes contiennent des erreurs qui impactent le WER, et ce, même si elles sont réalisées de façon manuelle (Levit *et al.*, 2012).

Szymanski *et al.* (2020) indique également que le choix de corpus de test ne couvrant pas le “*full spectrum of inputs encountered in everyday operations*”²⁰ ne permet pas d'avoir une idée précise des performances des systèmes sur des données de la vie réelle. De plus, il précise que certains corpus de référence, notamment pour l'évaluation de la parole spontanée, comprennent de la parole scriptée, très différente de la parole conversationnelle (Shriberg *et al.*, 2001; Levit *et al.*, 2012) et impactent donc l'évaluation.

Différents auteurs ont pu proposer des améliorations du WER ou des alternatives/-compléments à celui-ci, bien qu'il n'y ait pas encore de consensus (Radford *et al.*, 2022). Tout d'abord, il est question de pondérer les erreurs lors du calcul du WER (Garofolo *et al.*, 2000; Nanjo et Kawahara, 2005; Apone et O'Connell, 2010; Levit *et al.*, 2012; Roux *et al.*, 2022) ou de normaliser les transcriptions avant de calculer celui-ci (comme cela a été fait lors de la campagne ETAPE (Gravier *et al.*, 2012), et est utilisé par OpenAI pour évaluer Whisper (Radford *et al.*, 2022)). D'autres métriques ont également été développées dont :

- le MER (*Match Error Rate*) (Morris *et al.*, 2004) qui mesure le pourcentage de mots mal reconnus ou insérés tel que

$$MER = \frac{S + D + I}{N + I} = \frac{S + D + I}{S + D + C + I}$$

où S est le nombre de substitutions, D le nombre de suppressions, I le nombre d'insertions, C le nombre de mots corrects et N le nombre de mots dans la référence soit $N = S + D + C$.

20. spectre complet des entrées rencontrées dans les opérations quotidiennes (trad : chatGPT

- le WIL (*Word Information Lost*) (Morris *et al.*, 2004) qui mesure le pourcentage de mots mal reconnus entre une référence et une hypothèse tel que

$$WIL = 1 - \frac{C}{N} * \frac{C}{H}$$

où C est le nombre de mots corrects, N est le nombre de mots dans la référence et H est le nombre de mots dans l'hypothèse.

- ou encore la mesure de la validité écologique des résultats par des humains (c'est-à-dire la mesure de l'utilité de la sortie d'un système pour un humain dans une situation réelle) (Favre *et al.*, 2013).

Mis à part la normalisation des transcriptions, aucune de ces métriques n'est utilisée de façon régulière en reconnaissance automatique de la parole. Néanmoins, il est important de noter que les métriques permettant de mesurer l'importance et l'impact des erreurs pour les utilisateurs finaux sont très utiles lorsque les systèmes sont voués à être utilisés dans des contextes particuliers comme la transcription automatique pour les sourds et malentendants.

3.2 Les principaux corpus pour la reconnaissance automatique de la parole en français

3.2.1 Les corpus monolingues

Il existe plusieurs corpus du français créés spécifiquement pour la reconnaissance automatique de la parole. Certains de ces corpus ont fait l'objet de campagnes d'évaluation, tels que ESTER 1 et ESTER 2, EPAC, ETAPE et REPERE. Ces derniers sont des corpus de données radiophoniques ou télévisuelles, rassemblant des émissions de radios telles que France-Inter, France Info, RFI, RTM, France Culture et Radio Classique. Ces émissions sont plus ou moins propices à la production de parole spontanée²¹. Les plus anciens corpus Braf100, Bref80 et Bref120 sont tous trois des corpus de lecture. L'ensemble des corpus mentionnés sont brièvement présentés ci-dessous et un récapitulatif est présenté en tableau I.4.

Braf100 (Vaufreydaz *et al.*, 2000) : c'est un corpus comportant 28 heures ou 10 470 phrases de parole lue. Il rassemble la parole de 100 lecteurs (50% hommes, 50% femmes) âgés de 15 à 63 ans ayant eu pour tâche de lire des phrases tirées de pages Web et un extrait de “La science et l'hypothèse” d'Henri Poincaré. Ce corpus ne semble plus être disponible au téléchargement aujourd’hui.

21. Notamment l'émission “Culture et Vous” que Deléglise et Lailler (2020) qualifie de “très spontanée”.

Bref80/Bref120 (Lamel et al., 1991) : Bref80 est un sous-corpus de Bref120 (le corpus de 1991 ayant été augmenté). Il contient 5 330 phrases tirées d'article du quotidien Le Monde, lues par 80 locuteurs. Bref120 a augmenté le nombre de locuteurs à 120, pour réunir en tout 100 heures de parole produites par 65 femmes et 55 hommes. Les enregistrements ont été faits en pièce insonorisée.

ESTER 1/ESTER 2/EPAC (Gravier et al., 2004; Galliano et al., 2005; Estève et al., 2010) : Les corpus ESTER sont des corpus ayant été créés pour deux campagnes d'évaluation des performances des systèmes de transcription d'émissions radiophoniques. La première campagne (ESTER 1) s'est déroulée de 2003 à 2005. Le corpus rassemblait alors des émissions des radios France Inter, France Info, Radio France International, Radio Classique et Radio Télévision Marocaine. Cela représente au total 100 heures de parole transcris orthographiquement et 1 700 heures non transcris (dont des enregistrements de France Culture).

La deuxième campagne (ESTER 2) a eu lieu de 2007 à 2009. Le corpus correspondant reprenait les données de la campagne ESTER 1, auxquelles ont été ajoutées 106 heures d'émissions des radios France Inter, Radio France International, France Culture, Radio Classique, Africa 1, Radio Congo, et TVMA (Radio télévision du Maroc). Ces données sont transcris. L'ajout de nouvelles émissions avait pour but d'insérer plus de variété de types de parole et d'accents dans les données. Ont également été ajoutées 45 heures de données du projet EPAC.

EPAC est un projet qui avait pour but l'étude de la parole conversationnelle. Il contient 100 heures de transcriptions orthographiques manuelles réalisées à partir des 1 700 heures de parole non transcris du corpus ESTER 1, ainsi que la transcription automatique de l'ensemble de ces 1 700 heures.

ETAPE (Gravier et al., 2012) : ETAPE est un corpus de données télévisuelles (29h) et radiophoniques (13.5h) créé en 2012. Sa création a été axée sur la parole spontanée et les phénomènes de parole superposée. Le corpus rassemble plusieurs émissions telles que "BFM Story" et "Top Questions" (LCP) pour les news, "Pile et Face" (LCP), "Ça Vous Regarde" (LCP), "Entre les Lignes" (LCP) pour les débats, "La Place du Village" (TV8) pour le divertissement et "Un Temps de Pauchon", "Service Public", "Le Masque et la Plume", "Comme on nous Parle" et "Le Fou du Roi" pour les émissions de radio.

REPERE (Giraudel et al., 2012) : REPERE n'a pas été créé pour la reconnaissance automatique de la parole, mais est tout de même utilisé dans le domaine. Il a été créé pour la reconnaissance multimodale de personnes dans des vidéos. Pour

les systèmes évalués, l'objectif est de répondre à des questions du type : qui parle ? quelles sont les personnes dont le nom apparaît à l'image ? etc. Ce corpus a été créé entre 2011 et 2014. Les auteurs ont souhaité y réunir des situations diverses, et mettre l'accent sur les différences entre parole préparée et parole spontanée. Il comprend en tout 60 heures de parole venant d'émissions telles que "BFM Story", "Planète Showbiz" (BFM), "Ça Vous Regarde" (LCP), "Entre les Lignes" (LCP), "Pile et Face" (LCP), "LCP Info" ou encore "Top Questions" (LCP).

Corpus	Durée	Transcrit	Licence	Type de parole
BRAF100 (ELRA-S0197)	30 h	oui	ELRA, €	lecture
BREF80 (ELRA-S0006)	≈80 h	oui	ELRA, €	lecture (dictation)
BREF120 (ELRA-S0067)	100 h	oui	ELRA, €	lecture (dictation)
ESTER1 (ELRA-S0241)	1 700 h	≈100 h	ELRA, €	préparée
EPAC (ELRA-S0305)	ESTER1	1 700 h autom., parmi lesquelles 100 h manuel.	ELRA, €	préparée
ESTER2 (ELRA-S0338)	ESTER1 transcrit + EPAC transcrit + 150 h	100 h trans. riche, 50 h trans. rapide	ELRA, €	préparée + spontanée
ETAPE (ELRA-E0046)	30 h	oui	ELRA, €**	spontanée + préparée
REPÈRE (ELRA-E0044)	30 h	oui	ELRA, €**	préparée + spontanée

€ : corpus payant ; **gratuit pour recherche académique sans usage commercial

TABLE I.4 – Corpus d'évaluation monolingues de référence pour la RAP en français

3.2.2 Les corpus multilingues

Certains corpus multilingues sont également devenus ces dernières années (ou sont en train de devenir, pour les plus récents) des corpus de référence pour l'évaluation de système de reconnaissance automatique de la parole en français.

Common Voice (Ardila *et al.*, 2020) : Common Voice est un corpus créé grâce à une plateforme collaborative en ligne. Celle-ci propose aux personnes volontaires de lire des phrases tirées d'articles Wikipédia ou proposées (et validées) par des utilisateurs. Ces enregistrements sont ensuite validés ou non grâce à un système de vote, par d'autres utilisateurs (3 maximum. Deux votes down = audio non valide). De nouvelles versions du corpus sortent régulièrement. Dans l'article présentant Common Voice, il était question de 173 heures validées pour le français et de 38 langues disponibles. Aujourd'hui, pour sa version 13, ce sont 941 heures validées pour le français, et 112 langues qui sont disponibles. La version 13 ne peut réellement être vue comme une extension des versions précédentes puisque les données dans les ensembles d'entraînement, de développement et de test sont redistribuées

à chaque fois.

Fleurs (Conneau et al., 2023) : Ce corpus réunit des enregistrements en 102 langues. Le développement du corpus s'est fait sur la base d'un autre corpus tex-tuel, FloRes-101, contenant 3 001 phrases issues de Wikipédia et traduites dans 101 langues. Pour chaque phrase, trois enregistrements par trois locuteurs natifs différents sont effectués. Les enregistrements ont ensuite été validés, notamment pour vérifier que le signal correspond bien au texte. Le nombre d'heures de parole ou de lecteurs en français n'est pas renseigné.

Multilingual librispeech (Pratap et al., 2020) : Ce corpus est constitué de livres audio téléchargés depuis LibriVox. Il contient des données pour 8 langues différentes. Pour le français, les enregistrements représentent 1 097,12 heures de parole, pour 178 locuteurs (80 hommes, 98 femmes).

Multilingual TEDx (Salesky et al., 2021) : Ce corpus repose sur des enregistrements de conférences TED. Ce type d'enregistrement est plutôt court (10 minutes en moyenne) et préparé à l'avance par les intervenants. Les transcriptions et traductions de ces vidéos sont faites par des volontaires, selon les recommandations TED²². Pour le français, 189 heures de parole sont disponibles, ce qui représente 119 000 segments.

Voxpopuli (Wang et al., 2021) : Voxpopuli est constitué d'enregistrements du parlement européen (sessions plénières, réunions du comité...). Deux corpus sont fournis : un corpus non transcrit, et un transcrit. Le corpus non transcrit rassemble 23 langues pour environ 400 000 heures de parole, dont 22 800 de français. Le corpus transcrit rassemble 16 langues, pour 1791 heures en tout, dont 211 heures pour le français (534 locuteurs, 38.6% de femmes).

3.3 Les performances de reconnaissance automatique

3.3.1 Quelques résultats généraux sur différents types de parole

Le tableau I.6 fait état de différents systèmes de reconnaissance automatique de la parole sur quelques-uns des corpus présentés précédemment, triés par type de parole (lecture ou parole préparée/spontanée), et des performances obtenues.

22. <https://www.ted.com/participate/translate/guidelines>

CHAPITRE I : Les systèmes de reconnaissance automatique de la parole

Corpus	Durée (Fr)	Transcrit	Licence	Type de parole
COMMONVOICE	1 014 h*	oui	CC-0	lecture
FLEURS	N/A	oui	CC-BY	lecture
MULTILINGUAL LIBRISPEECH	1 333,35 h	oui	CC-BY 4.0	lecture
MULTILINGUAL TEDX	189 h	oui	CC BY-NC-ND 4.0	préparé
VOXPOLI	211 h	oui	CC0	professionnel

Durée totale actuelle pour la version 17.0 de Common Voice en français

TABLE I.5 – Corpus d'évaluation multilingues de référence pour la reconnaissance automatique de la parole en français

Corpus	WER	Architecture du système	MPA	ML	Ref.
Lecture					
Braf100	7.14%	HMM-DNN	non	oui	(Leang <i>et al.</i> , 2022)
CV 6.1*	9.75%	W2V2 + lm head + CTC	oui	non	(Evain <i>et al.</i> , 2021b)
CV 9*	13.9%	Enc-dec transformer (Whisper large v2)	oui	non	(Radford <i>et al.</i> , 2022)
FLEURS*	7,7%	Enc-dec transformer (Whisper large)	oui	non	(Radford <i>et al.</i> , 2022)
MLS*	6.58%	enc transformer + linear layer + CTC	oui	non	(Pratap <i>et al.</i> , 2020)
Parole préparée/spontanée					
MTedX*	19,4%	HMM + {CNN-TDNN}	non	oui	(Salesky <i>et al.</i> , 2021)
ESTER 2	13.7%	HMM-TDNN + ivecateurs	non	oui	(Boyer et Rouas, 2019)
ETAPE	25,22%	W2V2 + HMM-TDNN	oui	oui	(Evain <i>et al.</i> , 2021b)
REPÈRE	13,5%	2 systèmes couplés : {HMM-GMM + DNN} & {DNN + ivecateurs}	non	oui	(Rousseau <i>et al.</i> , 2014)
Voxpopuli*	11%	Enc-dec transformer (Whisper large)	oui	non	(Radford <i>et al.</i> , 2022)

Les corpus marqués par * sont des corpus multilingues. // MPA : Modèle Pré-Appris // ML : Modèle de langue

TABLE I.6 – Récapitulatif des performances état de l'art trouvées dans la littérature de systèmes de RAP sur différents corpus pour le français

Dans cette revue de l'état de l'art des performances des systèmes de reconnaissance automatique de la parole, nous avons recherché les meilleures performances obtenues. Les références sont plus ou moins récentes (la majorité étant des études faites entre 2019 et 2022) et les systèmes utilisent tous des réseaux de neurones, qu'ils soient hybrides avec des HMM, ou entièrement neuronaux. Le tableau récapitulatif montre plus d'utilisation de systèmes entièrement neuronaux pour la reconnaissance automatique de la lecture que pour la reconnaissance automatique de la parole préparée/spontanée qui utilise plutôt des systèmes hybrides incluant un modèle de langue.

L'analyse par type de parole rapporte quant à elle une grande disparité des performances obtenues sur des types de paroles *a priori* similaires.

Pour la lecture, les WER se trouvent dans une fourchette allant de 6,58% à 13,9%. Le WER le plus élevé obtenu sur la version 9 de Common Voice pourrait être dû aux accents de certains locuteurs ou encore aux conditions acoustiques variées que peuvent contenir les enregistrements de ce corpus, issus d'une plateforme collaborative en ligne. En effet, toute personne volontaire est invitée à enregistrer sa voix depuis chez elle. Si les performances sur la version 6.1 de ce corpus sont meilleures (9,75%), cela peut s'expliquer par le fait que le modèle pré-appris est ici adapté sur des données de Common Voice. A l'inverse, le système de Radford *et al.* (2022) ne comprend pas d'étape d'adaptation. De plus, le système entier pourrait n'avoir jamais vu des données issues de Common Voice et n'avoir donc pas appris certains accents ou certaines conditions acoustiques. Cela reste cependant une supposition étant donné que les données utilisées pour l'apprentissage du système ne sont pas connues.

L'utilisation de modèles pré-appris est très présente pour la reconnaissance automatique de la lecture. Ceci s'explique sans doute par l'utilisation massive de données de lecture pour l'apprentissage de ce type de modèles, celles-ci étant disponibles en grande quantité. Les modèles sont donc entraînés sur des données proches de celles sur lesquelles les systèmes sont testés ici. Les résultats obtenus sur CV 6.1 notamment ne sont pas sans rappeler les performances mises en avant dans Baevski *et al.* (2020) et reportées dans le tableau I.7 : l'utilisation d'un modèle pré-appris adapté sur une centaine d'heures de parole permet d'obtenir de très bons résultats. L'ensemble d'apprentissage du corpus CV 6.1 comprend en effet 428 heures de parole.

Pour la parole préparée/spontanée, la fourchette de WER s'étend de 11% à 25,22% et est donc deux fois plus grande que la précédente. L'utilisation de Whisper large en *zero shot* permet d'obtenir de bonnes performances sur le corpus Voxpopuli. Il est cependant important de rappeler d'une part la proximité de ce corpus avec de la parole lue : il réunit des enregistrements de parole plutôt très préparée se rapprochant de l'écrit (interventions du parlement européen). D'autre part, ce type de données très préparée représente également une source importante de données utilisées pour l'apprentissage de modèles pré-appris. Ainsi, si les données d'apprentissage des modèles Whisper ne sont pas connues, les chances d'y retrouver des données de parole lue ou préparées sont très hautes, ce qui participe à expliquer les bonnes performances relevées dans ce cas.

Le corpus ETAPE, dont la création a été axée sur la parole spontanée, est celui obtenant les moins bons résultats. Le WER obtenu est de 25,22% pour un système hybride. Les auteurs rapportent également dans leur publication un WER de 26,14% sur ce corpus pour un système *end-to-end* sans modèle de langue. Ces WER élevés pourraient trouver leur origine dans le manque de représentation de

la parole spontanée pour l'apprentissage des modèles Wav2Vec 2.0 : l'adaptation sur une petite quantité de données serait ainsi rendue moins efficace. Toutefois, la parole spontanée n'étant pas un type de parole uniforme mais dépendant notamment de la situation dans laquelle elle est employée et des locuteurs, l'adaptation sur peu de données de parole spontanée pourrait être altérée par une trop grande variabilité.

L'utilisation d'un modèle de langue pourrait également contribuer à améliorer ces performances comme le montre Baevski *et al.* (2020) dont les résultats sont reportés dans le tableau I.7.

Les performances sur ETAPE avec un système utilisant un modèle pré-appris de type Wav2Vec 2.0 nous amène donc à nous interroger sur :

- l'impact de la quantité de parole spontanée dans les données servant au pré-apprentissage de modèles
- l'impact de la variabilité dans la parole spontanée pour l'adaptation de ces modèles
- l'impact de l'utilisation d'un modèle de langue pour le décodage de la parole spontanée avec un système *end-to-end* utilisant un modèle pré-appris.

Système	Apprentissage	Adaptation	ML	WER clean	WER other
HMM-DNN Lüscher <i>et al.</i> (2019)	x	x	4-gram	3.8	8.8
W2V2 Large + randomly initialized linear projection + CTC Baevski <i>et al.</i> (2020)	LV-60k	10 min	transformer	4.8	8.2
	LV-60k	1h	transformer	2.9	5.8
	LV-60k	10h	transformer	2.6	4.9
	LV-60k	100h	transformer	2.0	4.0
	LV-60k	10 min	x	40.2	38.7
	LV-60k	1h	x	17.2	20.3
	LV-60k	10h	x	6.3	10.0
	LV-60k	100h	x	3.1	6.3

LibriSpeech est un corpus de livres audio. // ML : Modèle de langue // clean/other : ensembles de tests du corpus LibriSpeech. "other" est l'ensemble considéré comme le plus compliqué.

TABLE I.7 – Résultats tirés de Baevski *et al.* (2020) rapportant les performances d'un modèle pré-appris de type Wav2Vec 2.0 sur LibriSpeech (anglais), en fonction de la taille du corpus d'adaptation et l'utilisation d'un modèle de langue

3.3.2 Résultats sur la parole spontanée

N'ayant pu trouver de modèle Wav2Vec 2.0 appris sur un ensemble de données en français comprenant plus de parole spontanée que ceux élaborés dans le cadre du projet LeBenchmark, la première question ne peut être étudiée via une analyse de

3 L'évaluation des systèmes de reconnaissance automatique de la parole

la littérature. Nous faisons ensuite le choix de nous concentrer sur l'impact de la variabilité, l'intégration d'un modèle de langue externe venant, selon nous, apporter une couche d'inconnues à l'étude des performances des systèmes de reconnaissance automatique de la parole sur la parole spontanée. Ainsi, nous présentons dans la section suivante quelques résultats sur la reconnaissance automatique de la parole spontanée nous permettant de relever la différence de traitement par les systèmes de différents types de parole spontanée et la source de cette variabilité.

Dans leur rapport de recherche, Tancoigne *et al.* (2020) ont comparé huit outils commerciaux de transcription automatique²³ sur quatre types d'ensembles de test : un texte lu (“Physionomie”, lecture), un cours magistral enregistré en situation (“Comptines”), un entretien avec deux interlocuteurs (“Camille”), une réunion associative avec de nombreux locuteurs (“Harmonie”, spontané)²⁴. La figure I.22 montre les résultats obtenus pour chaque système de reconnaissance automatique de la parole.

	Manuelle	Médiane	Happy Scribe	Go Transcribe	Sonix	Vocapia	Video Indexer	YouTube	Headliner	Vocalmatic
Physionomie (texte lu)	0,6	19,5	14,8	14,8	16,6	18,1	20,8	28,4	28,2	28,5
Comptines (monologue)	1,6	14,8	11,9	12,4	11,1	14,2	18,2	15,4	31,1	26,6
Camille (entretien face à face)	18,9	50,0	51,8	54,2	48,3	28,1	34,0	36,4	76,1	75,5
Harmonie (discussion spontanée)	12,7	88,4	86,2	86,2	90,5	57,6	79,8	107,0	222,6	244,6

Les couleurs indiquent l'écart à la médiane.

FIGURE I.22 – Résultats de Tancoigne *et al.* (2020) sur quatre ensembles de test, obtenus via 8 systèmes commerciaux de reconnaissance automatique de la parole (WER%).

Chacun de ces extraits présentait des difficultés liées au type de parole en lui-

23. Go Transcribe, Happy Scribe, Headliner, Sonix, Video Indexer, Vocalmatic, Vocapia, YouTube

24. L'étiquetage “préparé/spontané” a été effectué par nos soins pour les corpus “Comptines” et “Camille”.

même²⁵, mais aussi au vocabulaire (“Physionomie”), au handicap physique d’un locuteur ou d’une locutrice (“Camille”), ou encore au bruit ambiant (“Harmonie”).

On observe que pour une tâche de reconnaissance de la lecture, pour laquelle la littérature reporte des WER inférieurs à 10%, le WER médian est assez élevé : vers 19,9% alors que les meilleurs systèmes sont à 14,8%. Selon Adda Decker (2006), “la parole lue est certes médialement de l’oral, mais au fond elle reflète parfaitement la langue écrite”. Ils auraient donc dû retrouver ici de bonnes performances.

L’ensemble de test “Comptines” est celui qui est le plus proche d’une parole préparée : la trame du discours est connue et préparée à l’avance. Les auteurs précisent que “le registre de langue employé par l’intervenante peut être qualifié de soutenu. La diction est bonne et peu d’onomatopées ou d’hésitations ponctuent le discours”. Il est difficile de mesurer *a posteriori* si cet ensemble contient également de la parole spontanée. Le WER médian (14,8%) dépasse ici celui obtenu sur la lecture, avec des performances allant de 11,1% à 26,6%. Le corpus “Camille”, ensuite, est un format d’entretien dont le WER médian est à 50% avec des performances toutes différentes, de 28,1% à 76,1%. Le dernier corpus est catégorisé comme spontané par les auteurs : “Cet extrait est représentatif d’une situation de conversation spontanée de groupe dans un contexte associatif ou professionnel.” (p.8). Le WER médian est de 88,4%, et les résultats vont de 57,6% à 244,6%.

Les résultats sur ces extraits montrent que les performances des systèmes semblent varier en fonction du contexte, avec notamment une grande difficulté pour les systèmes lorsqu’il s’agit d’une conversation spontanée dans un contexte associatif ou professionnel.

D’autres études montrent une variabilité des performances des systèmes en fonction de niveaux de spontanéité. Les études de Jousse *et al.* (2008) et Dufour *et al.* (2010) portant sur la détection et la classification automatique de segments de parole spontanée dans de grands ensembles de données montrent effectivement des WER plus élevés au fur et à mesure de l’augmentation de la spontanéité dans des journaux d’information de radios. Nous reportons leurs résultats dans le tableau I.8. Si l’étude de (Jousse *et al.*, 2008) rapporte des recoupements de WER entre les différents niveaux de spontanéité, l’étude de (Dufour *et al.*, 2010) met plus clairement en évidence que l’augmentation du WER suit l’augmentation de la spontanéité. L’étude de Deléglise et Lailler (2020) rapporte également un WER oscillant entre 21,16% et 41,14% (selon le système utilisé), pour l’émission “Culture et Vous” tirée du corpus REPÈRE, performance plutôt moyenne voire mauvaise qu’ils justifient par le côté “très spontané” de l’émission, sans pour autant donner

25. notamment les disfluences présentes en parole spontanée ou la parole superposée lorsque plusieurs locuteurs sont présents

4 Synthèse

de détails sur cette caractérisation. Il est important de noter pour ces études que les WER obtenus sur des émissions ou extraits considérés comme très spontanés ont été obtenus dans des conditions d'enregistrement professionnel (télévision et radio), avec parfois un public présent sur place mais aussi à distance, ce qui oblige à une bonne articulation afin d'être compris de tous. Si l'on est en présence également de parole spontanée dans l'étude de Tancoigne *et al.* (2020), le contexte spécifique de la radio/télévision représente donc un premier point de différenciation. De la même façon, il est important de noter que l'extrait "Camille" de l'étude de Tancoigne *et al.* (2020) présente la particularité de contenir la parole d'un locuteur avec un handicap physique, et que l'extrait "Harmonie" est un corpus bruité, et qui rassemble de nombreux locuteurs. L'ensemble de ces résultats est donc difficilement comparable directement.

Auteur, date	ML	Syst.	Niveaux de spont.	WER
Jousse <i>et al.</i> (2008)	oui	HMM-GMM	1-3 : parole prep., 4-6 : spont. faible, 7-10 : spont. fort	23 à 26% 24 à 30% 28 à 44%
Dufour <i>et al.</i> (2010)	oui	HMM-GMM	1 : parole prep., 2-4 : spont. faible, 5-8 : spont. fort	10,1% 18,4% 28,5%

ML : Modèle de langue

TABLE I.8 – Word Error Rate en fonction de niveaux de spontanéité

4 Synthèse

Ce premier chapitre a d'abord présenté les avancées en termes d'architectures neuronales pour la reconnaissance automatique de la parole. Leur bonne capacité de généralisation aura permis d'égaler voire de dépasser assez rapidement les résultats obtenus avec des systèmes stochastiques. Ces architectures présentent néanmoins toutes l'inconvénient de nécessiter une grande quantité de données pour leur entraînement. La mise à disposition à la communauté de modèles pré-appris et les possibilités de réutilisation et d'adaptation avec très peu de données annotées tout en conservant de bonnes performances permet de contourner cette limite. Si ces modèles nécessitent toujours une grande quantité de données pour être performants, les possibilités d'apprentissage avec des données non annotées changent la donne. Elles sont plus faciles à trouver permet ainsi une création beaucoup plus rapide de nouveaux corpus : c'est en effet l'étape de transcription des corpus qui est chronophage et fastidieuse.

Ensuite, nous avons introduit les corpus de référence en français pour la reconnaissance automatique de la parole. Cette présentation est accompagnée de plusieurs constats. Tout d'abord, les corpus monolingues de parole préparée/spontanée sont tous issus d'un même contexte : la télévision ou la radio. Aucun nouveau corpus monolingue n'a été créé depuis 2012 pour une tâche de reconnaissance automatique de la parole, alors même que l'utilisation de systèmes de reconnaissance automatique de la parole peuvent maintenant faire partie de notre quotidien (transcription automatique de réunions, aide pour les personnes sourdes ou malentendantes) et nécessitent de bonnes performances sur tous types de parole.

Du côté des corpus multilingues, pour la plupart beaucoup plus récents puisque créés entre 2020 et 2023, on observe un manque de représentation de la parole spontanée. Les corpus ne contiennent principalement que de la parole lue ou de la parole très préparée (TEDX).

Enfin, nous avons constaté dans ce chapitre que plusieurs études relèvent de moins bonnes performances des systèmes sur la parole préparée/spontanée que sur la parole lue, mais aussi que la parole spontanée recouvre différents niveaux de spontanéité et que plus la spontanéité augmente, plus le WER augmente. La classification des ensembles comme plus ou moins spontanés semble liée aux différentes situations de parole représentées : un cours magistral, un entretien, une réunion, ou encore en fonction des émissions télévisuelles. En ce sens, le prochain chapitre sera consacré à mieux circonscrire ce que recouvre la dénomination de parole spontanée et étudier les différentes situations dans lesquelles elle apparaît, et avec quel niveau de spontanéité.

Chapitre II

La parole spontanée : un objet d'étude complexe

La parole spontanée est souvent catégorisée dans la littérature comme un style ou un registre sans jamais en proposer une catégorisation systématique. Le CNRTL (Centre National de Ressources Textuelles et Lexicales)¹ définit ces termes comme suit :

Style :

- Ensemble des moyens d'expression (vocabulaire, images, tours de phrase, rythme) qui traduisent de façon originale les pensées, les sentiments, toute la personnalité d'un auteur.
- Mode d'expression verbale qui est spécifique de tel genre ou sujet littéraire, qui correspond ou non à certaines normes formelles.
- Mode d'expression verbale qui correspond idéalement à certaines normes formelles.
- Mode d'expression verbale propre à une école, à une nation, à une époque.
- Mode d'expression verbale propre à une activité, à un groupe professionnel.

Registre :

- Registres de langue, de discours. “Usages divers qui sont faits de cette langue (de ce discours) selon les milieux où elle est employée ou selon les situations psychosociologiques dans lesquelles se trouve l'émetteur” (Éduc. 1979). *On dira que chaque locuteur dispose de plusieurs registres habituels ou préférentiels dans l'usage qu'il fait d'une langue donnée (D. D. L. 1976).*

1. <https://www.cnrtl.fr/definition/>

Ainsi, Veiga *et al.* (2012); Lancien (2020) et Boula de Mareüil (2014) parlent respectivement de style, de macro-style et de phonostyle² spontané, et Torreira *et al.* (2010) de registre spontané. De nombreux auteurs ont souligné que la parole spontanée peut apparaître aussi bien lors d'une discussion entre amis (Torreira *et al.*, 2010; Lancien et Côté, 2018; Bodur *et al.*, 2022), que lors d'une interview/émission radiophonique (Dufour *et al.*, 2010; Garnerin, 2022). Cela amène effectivement à considérer la parole spontanée comme un style ou un registre, car ce sont là des aspects qui se rapportent aux situations dans lesquelles peut s'inscrire le locuteur (registre), aux normes qui les régissent et au mode d'expression qui peut être employé par un groupe ou lors d'une activité spécifique (style). Néanmoins, il faut noter qu'une discussion entre amis et une interview sont des situations bien différentes.

La notion de “genre” apparaît également parfois pour qualifier la parole spontanée (Caelen-Haumont et Bel, 2000; Lai, 2019). Cependant, si cet usage n'est pas expliqué dans Lai (2019), cette désignation qui selon le CNRTL se rapporte en linguistique à une “idée générale ou classe d’êtres ou d’objets qui possèdent un ou plusieurs caractères communs” rassemble pour Caelen-Haumont et Bel (2000) différentes formes d’oralité dont le dialogue de parole et la poésie improvisée. Ainsi, la parole spontanée peut être considérée comme un genre si on l'observe à travers ses caractéristiques intrinsèques ou saillantes, parmi lesquelles les disfluences (Dufour *et al.*, 2010; Clavel *et al.*, 2013; Koch et Oesterreicher, 2001; Adda-Decker *et al.*, 2003), les réductions³ (Llisterri, 1992; Wu et Adda-Decker, 2020) ou encore le débit de parole (Labov, 1973; Goldman *et al.*, 2009).

Nous décidons, pour ce travail de thèse, de rester ouverts à chacune de ces différentes manières de voir la parole spontanée. Nous rapportons aussi bien des éléments se rattachant au genre, au style ou au registre spontané, et nos recherches reposent à la fois sur des travaux issus du domaine du traitement de la parole, de la phonétique, de la sociolinguistique et de la phonostylistique. Afin de ne pas confusionner le lecteur, nous utiliserons dès à présent la locution “type de parole spontanée” pour désigner l'ensemble des différentes formes (interview, discussion entre amis, dialogue...) de l'objet d'étude qu'est la parole spontanée.

Nous commençons ce chapitre par une définition du type de parole spontanée en reconnaissance automatique de la parole, puis rapportons ses caractéristiques

2. “Styles de parole perçus comme identifiant une situation de communication, via un genre, une image acoustique typifiée.” (Goldman *et al.*, 2009)

3. Réduction : zones de parole où les prononciations des mots seraient seulement partiellement réalisées et contiendraient potentiellement moins de segments que le nombre prévu par une prononciation canonique. Exemple tiré de Wu et Adda-Decker (2021) : le mot “quatre” prononcé [katR] admet comme variante réduite [kat]

acoustiques et linguistiques et les difficultés de traitement de ces dernières par les systèmes de reconnaissance automatique de la parole. Ces caractéristiques sont également mises en relation avec les niveaux de spontanéité observés dans le chapitre précédent. Nous étudions ensuite la parole spontanée au travers de travaux de recherche non rattachés au domaine de la reconnaissance automatique de la parole. Ainsi, nous présentons ses facteurs de variation ainsi que les modèles à une ou plusieurs dimensions qui caractérisent ses variations. Ce chapitre a pour but de comprendre et circonscrire le type de parole spontané : Comment se caractérise-t-il ? Où et quand apparaît-il ? Quelles en sont ses formes ?

Déroulement du chapitre

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole	46
1.1 Les caractéristiques de la parole spontanée	47
1.1.1 Les caractéristiques acoustiques	47
1.1.2 Les caractéristiques linguistiques	53
1.1.3 Les difficultés induites pour la reconnaissance automatique de la parole	54
1.1.4 Déterminer des niveaux de spontanéité en fonction de ces caractéristiques	57
2 La parole spontanée : caractériser ses variations	60
2.1 Définir la parole spontanée	60
2.2 Les facteurs de variation de la spontanéité	62
2.2.1 Les facteurs interpersonnels	62
2.2.2 Les facteurs environnementaux	64
2.2.3 Les facteurs plus personnels	65
2.3 Revue des modèles à une ou plusieurs dimensions relatifs à la variation stylistique	65
2.3.1 Le niveau de formalité	66
2.3.2 L'attention portée au discours	66
2.3.3 Le niveau de contrôle	67
2.3.4 Modèles prenant en compte d'autres dimensions	68
3 Synthèse	70

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole

Le chapitre précédent et notamment l'étude des performances de systèmes de reconnaissance automatique de la parole sur différents types de parole nous a montré que la reconnaissance de la parole spontanée est plus difficile que la reconnaissance de la lecture ou de la parole préparée. Ceci est d'autant plus vrai lorsque le niveau de spontanéité est élevé. (Szaszák *et al.*, 2016, p. 1) indiquent que “*the treatment of spontaneous speech constitutes a big challenge in spoken language technology, because it violates standards and assumptions valid for formal speaking style or written language and hence constitutes a much more complex challenge in terms of modelling and processing algorithms*”⁴. Les auteurs précisent un peu plus loin que la parole spontanée se différencie de l'écrit de par les disfluences, pauses remplies, répétitions, réparations et amorces de mots et de la syntaxe particulière qui la composent. Gabler *et al.* (2023) ajoutent que tout ce qui représente un signe de la production linéaire de la parole spontanée, dont les réductions, est ce qui la rend plus difficile à transcrire automatiquement. Ces deux études représentent la parole spontanée à la fois sous sa forme différentielle, c'est-à-dire basée sur une comparaison avec la lecture/l'écrit oralisé ou la parole préparée, mais aussi sur la base de ses caractéristiques linguistiques et acoustiques. La détermination du type de parole comme spontané dépend donc :

- de la distance entre le type de parole et l'écrit
- et/ou de la présence forte de certaines caractéristiques linguistiques ou acoustiques.

Ces deux conditions, bien que différentes dans leur forme, portent finalement le même sens. En effet, la parole spontanée n'est pas préparée à l'avance et de fait, s'éloigne des structures usuelles de la “langue écrite”. Et cette distance est rendue visible par la présence de certaines caractéristiques linguistiques et phonétiques (telles que les répétitions, les faux départs, les troncations, les réductions, les hésitations etc. (Adda-Decker *et al.*, 2003; Bazillon *et al.*, 2008b; Gabler *et al.*, 2023)). Au-delà d'être définie comme un objet linguistique, la parole spontanée est définie en reconnaissance automatique de la parole comme un objet statistique menant

4. le traitement de la parole spontanée constitue un défi majeur dans la technologie de la langue parlée, car elle viole les normes et les hypothèses valables pour le style de discours formel ou la langue écrite, ce qui en fait un défi beaucoup plus complexe en termes de modélisation et d'algorithmes de traitement. (trad : chatGPT 3.5)

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole

toujours au même type de conclusion : plus certaines de ses caractéristiques sont présentes, plus les performances des systèmes baissent. S'il est possible de définir le type de parole spontanée en reconnaissance automatique de la parole de ces deux façons, certaines études privilégient l'une ou l'autre de ces définitions. Ainsi, Nakamura *et al.* (2005); Horii *et al.* (2022) optent pour une description différentielle, tandis que Ulasik *et al.* (2020); Bang *et al.* (2020); Candido Junior *et al.* (2023) la décrivent uniquement via ses caractéristiques ou les situations dans lesquelles elle apparaît (interviews, dialogues informels, conversations, contextes naturels, environnements bruités).

1.1 Les caractéristiques de la parole spontanée

Nous présentons dans cette section les caractéristiques acoustiques et linguistiques liées à la parole spontanée les plus fréquemment relevées dans la littérature. Nous avons choisi de les séparer en fonction de leur implication aux niveaux i) acoustique et ii) linguistique.

1.1.1 Les caractéristiques acoustiques

Une première caractéristique acoustique relevée en parole spontanée est la présence de **prononciations déviante**s, encore appelées prononciations non-standard ou variantes de prononciation. Cette notion réfère à la prononciation d'un mot (ou parfois d'un ensemble de mots) ne correspondant pas à sa prononciation de référence que l'on peut retrouver dans des dictionnaires de prononciation tel que Lexique380 (New *et al.*, 2007). Il est important de préciser que les qualificatifs “déviants” et “non-standard” utilisés par certains auteurs ne font pas l'unanimité. En effet, comme le précise Claire Blanche-Benviste⁵ : «certaines prononciations courantes, différentes de la prononciation académique, sont d'un usage si répandu qu'on ne peut pas les considérer comme des phénomènes marginaux.». Ainsi, les écarts remarqués entre la langue écrite ou académique et la langue orale ne doivent pas être considérés comme des fautes ou un parler uniquement populaire. L'utilisation de “variantes de prononciation” est donc préférable. Adda Decker (2006) montre qu'elles sont plus présentes en parole journalistique qu'en lecture. Bigi et Meunier (2018) précisent que ce phénomène peut parfois mener à des prononciations inattendues telles que le mot “exemple” prononcé [Ekszap]⁶ au lieu de [Egza~pl].

Les variantes de prononciation peuvent avoir plusieurs origines comme un accent,

5. propos rapportés dans (Thomas, 2002, p. 14)

6. L'écriture SAMPA est utilisée pour plus de facilité. La correspondance API-SAMPA se trouve en annexe A.

une **hypoarticulation** (réduction de l'effort articulatoire) ou encore l'influence de ce qui est dit avant ou après. L'hypoarticulation est plus forte en parole spontanée qu'en lecture selon Bodur *et al.* (2022) et entraîne parfois ce que Fohr *et al.* (2015) qualifient de “suites de syllabes incompréhensibles”. Cela est dû à la réduction et à la centralisation de l'espace acoustique des voyelles les rendant moins distinguables les unes des autres, plus fortement présentes en parole spontanée qu'en parole de laboratoire, lecture ou parole préparée (Lancien et Côté, 2018; Lancien, 2020).

Les phénomènes d'**assimilation** sont une autre caractéristique acoustique observable, tel que le relève Bazillon *et al.* (2008b). Ils concernent souvent le voisement comme l'indique (Bazillon, 2011, p. 90), qui parle d’“une variation phonétique entraînant la modification de la prononciation d'une consonne sourde au contact d'une consonne voisine sonore (ou l'inverse)”. L'auteur donne l'exemple de “je crois” prononcé [SkRwa] ou encore “cheval” prononcé [Sfall] suite à la disparition du schwa. Ces phénomènes sont eux aussi à l'origine de l'apparition de variantes de prononciation. Duez (2001) rapporte un degré plus élevé d'assimilation en parole spontanée qu'en parole lue.

La disparition du schwa mentionnée précédemment est un type d'**élision**, à savoir un “effacement, dans le compte des syllabes, dans la prononciation, de la voyelle finale d'un mot qui précède un autre mot commençant par une voyelle ou un « h » muet.”⁷. Certaines élisions sont courantes et existent à l'écrit (telles que “l'école” pour “la école”) tandis que d'autres sont considérées comme des variantes “non standard” qui s'écartent d'une prononciation de référence (comme “petit” prononcé [pti] ou encore “t'as” pour “tu as”). Le qualificatif “non standard” se rapporte ici au fait que ces élisions n'apparaissent pas dans la grammaire acceptée de l'écrit⁸, et non à leur fréquence d'apparition.

Ensuite, il est fréquent de trouver des **réductions** en parole spontanée. (Wu et Adda-Decker, 2020, p. 628) les définissent comme des “zones de parole où les prononciations des mots seraient seulement partiellement réalisées et contiendraient potentiellement moins de segments que le nombre prévu par une prononciation canonique”. Elles peuvent aussi bien concerner un mot qu'un ensemble de mots comme le montrent les autrices en figure II.1 avec “par exemple” prononcé [paRa~p].

Les réductions sont plus présentes en parole spontanée que dans d'autres types de parole (Duez, 2001). Cela va de paire avec un **débit de parole** élevé, environnement propice à l'apparition de tels phénomènes selon Llisterri (1992). En effet,

7. Définition tirée du dictionnaire de l'Académie Française : <https://www.dictionnaire-academie.fr/article/A9E0789>

8. hors cas spéciaux comme tweets, SMS...

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole

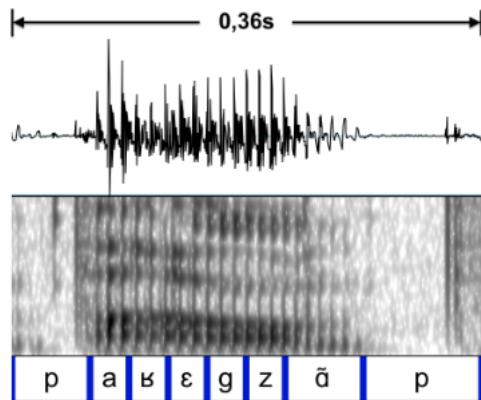


FIGURE II.1 – Exemple de réduction tiré de Wu et Adda-Decker (2020). “par exemple” est prononcé [paRa~p]

le débit de parole que Adda Decker *et al.* (2012) définissent comme “le nombre moyen de mots par seconde, de syllabes par seconde, de segments phonétiques par seconde, incluant ou non les silences inter-mots” est souvent qualifié de rapide en parole spontanée. Ainsi, Llisterri (1992) rapporte un débit de parole plus élevé en parole spontanée. Au contraire, Goldman *et al.* (2009) rapporte un débit de parole calculé sur la syllabe plus élevé en lecture qu’en parole spontanée. Si la mesure de (Goldman *et al.*, 2009) inclut les pauses, la précision n’est pas donnée par (Llisterri, 1992). Plutôt qu’un début plus rapide ou plus lent, il apparaît que la variabilité du débit est un phénomène plus visible en parole spontanée que pour d’autres types de parole (Christodoulides, 2020), constat que (Llisterri, 1992) complète en précisant que la plage de variabilité est plus grande en parole spontanée qu’en lecture. Les mesures de débit d’articulation ne prennent pas en compte les pauses rapportent quand à elles un niveau plus élevé en lecture qu’en parole spontanée selon Goldman *et al.* (2009) pour un calcul se basant sur le nombre de syllabes, et plus élevé en parole spontanée (26 phonèmes par seconde) dite “familière” qu’en parole journalistique (15 phonèmes par seconde) selon Adda Decker *et al.* (2012)⁹. Bazillon *et al.* (2008b) font également état d’un débit phonémique qui varie peu dans un journal d’informations en comparaison à une situation d’entretien, mais note toutefois qu’une certaine variabilité dans les mesures est présente du fait du sujet de la discussion ou encore de l’état émotionnel du locuteur.

Autre caractéristique observable, les **pauses** (pleines ou non) sont également très présentes en parole spontanée. Les pauses pleines se caractérisent par des **allons-**

9. Les autrices parlent ici de débit de parole mais celui-ci étant calculé au niveau du mot sans prise en compte des pauses, il s’agit en fait d’un débit d’articulation

gements, “procédé expressif consistant à allonger la durée de la voyelle tonique pour marquer que l’on s’attarde dans la contemplation d’une idée”¹⁰. Le locuteur peut marquer une pause aussi bien pour reprendre son souffle ou déglutir, que pour marquer **l’hésitation**, le temps de réflexion qu’il s’accorde avant de continuer sa prise de parole (Bazillon *et al.*, 2008b).

Les **disfluences** sont une autre caractéristique importante de la parole spontanée. (Clavel *et al.*, 2013, p. 1107) reprend la définition de (Blanche-Benveniste, 1990) et considère comme disfluence “*words and sentences that are cut off mid-utterance, phrases that are restarted or repeated, repeated syllables, grunts or unrecognizable utterances occurring as fillers, and repaired utterances*”¹¹. Elles ont un impact sur le rythme. (Clavel *et al.*, 2013, p. 1113) précisent que ces disfluences sont d’autant plus présentes en parole spontanée lorsque le locuteur est traversé par certaines émotions. L’état émotionnel du locuteur avait déjà été mentionné par Bazillon *et al.* (2008b) comme un facteur pouvant avoir un impact sur le débit phonémique.

D’un point de vue intonatif, Bazillon *et al.* (2008b) relèvent un “**manque de marque intonative**” en parole spontanée. Ils y observent des descentes prosodiques indiquant la fin d’un énoncé, puis des remontées prosodiques subites indiquant la continuation de la prise de parole. Ceci rend l’introduction de la ponctuation dans les transcriptions particulièrement compliquée. Goldman *et al.* (2009) ajoute que le registre de la fréquence fondamentale (f_0)¹² est plus réduit en lecture qu’en parole spontanée, tandis que Llisterri (1992) relève des valeurs de f_0 plus hautes en lecture, une F_0 moyenne assez haute et une plage plus large de variation de la f_0 .

Enfin, la parole spontanée se caractérise également, notamment lorsqu’elle est conversationnelle, par des passages de **parole superposée**¹³, des **backchannels** et des **bruits**. Le terme “backchannel” regroupe “l’ensemble des signaux verbaux, vocaux et gestuels, émis par l’interlocuteur d’un dialogue pour montrer son écoute, sa compréhension, son accord, etc. au discours produit” (Bertrand *et al.*, 2009, p 3). Ces signes comprennent notamment les bâillements, rires, interjections… Ainsi, les rires et les bruits de bouche ne se retrouvent quasiment pas en lecture (Bigi et Meunier, 2018).

10. Définition du dictionnaire en ligne CNRTL

11. mots et phrases coupés en plein énoncé, les segments repris ou répétées, les syllabes répétées, les grognements ou énoncés non reconnaissables se produisant comme des remplissages, et les énoncés réparés (trad : chatGPT)

12. “Registre” : désigne ici l’étendue de l’échelle vocale, de la note la plus grave à la note la plus aiguë. “ f_0 ” : hauteur de voix correspondant à la fréquence la plus basse du signal produit, telle que produite par les cordes vocales.

13. lorsque plusieurs locuteurs parlent en même temps

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole

Les caractéristiques acoustiques mentionnées sont récapitulées dans le tableau II.1. Dans l'ensemble, les études relèvent des conséquences acoustiques sur le signal de parole cohérentes hormis pour le débit de parole et la variation prosodique où certaines études relèvent des résultats contradictoires.

Prononciations déviante s/ non-standard	(Fohr <i>et al.</i> , 2015; Bigi et Meunier, 2018)
Hypoarticulation	(Lancien et Côté, 2018; Wu et Adda-Decker, 2020)
Elisions	(Fohr <i>et al.</i> , 2015) non-standard (Bigi et Meunier, 2018)
Disparition des schwas et assimilations	(Bazillon <i>et al.</i> , 2008b)
Réductions temporelles	(Adda Decker, 2006; Bigi et Meunier, 2018; Wu et Adda-Decker, 2020) voyelles réduites sur mots grammaticaux (Llisterri, 1992)
Débit de parole	rapide (Adda-Decker et Lamel, 2000; Bigi et Meunier, 2018) irrégulier (Bigi et Meunier, 2018)
Pauses	(Koch et Oesterreicher, 2001; Fohr <i>et al.</i> , 2015) pleines (Jousse <i>et al.</i> , 2008; Torreira <i>et al.</i> , 2010; Dufour <i>et al.</i> , 2010) allongements phoniques (Koch et Oesterreicher, 2001)
Disfluences	(Bortfeld <i>et al.</i> , 2001; Clavel <i>et al.</i> , 2013) hésitations (Luzzati, 2007; Fohr <i>et al.</i> , 2015; Bigi et Meunier, 2018) morphèmes spécifiques (ehu, ben...) marquant l'hésitation (Koch et Oesterreicher, 2001; Bazillon <i>et al.</i> , 2008b; Torreira <i>et al.</i> , 2010) répétitions (Koch et Oesterreicher, 2001; Jousse <i>et al.</i> , 2008; Dufour <i>et al.</i> , 2010; Bazillon <i>et al.</i> , 2008b; Bigi et Meunier, 2018) répétitions de mots fonction (Adda-Decker <i>et al.</i> , 2003) réparations (Jousse <i>et al.</i> , 2008; Dufour <i>et al.</i> , 2010) faux départs (Jousse <i>et al.</i> , 2008; Bazillon <i>et al.</i> , 2008b; Dufour <i>et al.</i> , 2010) troncations (Bazillon <i>et al.</i> , 2008b; Bigi et Meunier, 2018)
Emotions	Clavel <i>et al.</i> (2013)
Manque de marque intonative	(Bazillon <i>et al.</i> , 2008b)
Parole superposée	Clavel <i>et al.</i> (2013)
Backchannels	Bigi et Meunier (2018)
Nombreux bruits	respirations Fohr <i>et al.</i> (2015) rires Fohr <i>et al.</i> (2015) exclamations Fohr <i>et al.</i> (2015)

TABLE II.1 – Caractéristiques acoustiques les plus fréquentes en parole spontanée

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole

1.1.2 Les caractéristiques linguistiques

Certaines caractéristiques de la parole spontanée donnent lieu à des processus qui peuvent être lexicalisés. On retrouve ainsi aux côtés des pauses et des allongements, les éléments lexicalisés “euh” ou encore “ben” qui peuvent également marquer l'**hésitation**, ainsi que “hum”, “ouais” et diverses **interjections** utilisées comme éléments de **Backchannelling**. Le morphème “euh” se retrouve peu en lecture d'après Bigi et Meunier (2018). Selon Koch et Oesterreicher (2001), les interjections (“ah !”, “oh !”, “zut !”) “se rattachent directement à l'émotionnalité forte du parlé”.

D'autres caractéristiques se rapportent aux **registres de langue**¹⁴. Il est courant d'associer à ces registres de langue des niveaux de formalité. Le registre familier serait alors plutôt informel et le registre soutenu très formel, donc très porté sur la forme. Blanche-Benveniste et Bilger (1999) indiquent que le locuteur peut être amené à utiliser un registre très soutenu lorsqu'il parle en tant que représentant d'une profession, et à changer de registre dans une même prise de parole. Selon Torreira *et al.* (2010), les **mots informels** comme “mec” ou “gars” (pour désigner un garçon) et “truc” (pour désigner une chose), en **verlan** (“ouf” pour “fou”, “meuf” pour “femme”...) ou encore les **gross-mots** peuvent être vus comme des indicateurs d'une certaine spontanéité. Nous pouvons en conclure que l'ensemble des registres de langue est susceptible d'être utilisé en parole spontanée. Du point de vue lexical, Fohr *et al.* (2015) mentionne également la présence de **néologismes**.

Aux niveaux syntaxique et grammatical, Bazillon *et al.* (2008b) et Luzzati (2004) rapportent des **fenêtres syntaxiques** (ou ruptures de construction) courtes¹⁵, pas nécessairement conjointes et superposables en parole spontanée. Ce constat est à mettre en opposition à celui fait sur la parole préparée où les fenêtres sont plutôt longues, conjointes et sans intersection. En effet, à ce type de parole sont associés différents types d'**interruptions**, telles que les **parenthèses ou commentaires**, et les **corrections**. Les auteurs donnent en exemple le fenêtrage d'une séquence tirée du corpus CaFE (Gournay *et al.*, 2018) en figure II.2. Ils y présentent les différents types de fenêtres syntaxiques avant de proposer le fenêtrage syntaxique de la séquence.

Enfin, certains auteurs rapportent une certaine **agrammaticalité**¹⁶. Sous ce terme controversé sont réunis notamment les accords non réalisés et l'absence de doubles négations. (Luzzati, 2007, p. 25) précise que “l'oral spontané est destiné à provoquer une réaction et non une transcription”, et risque de paraître “fragmentaire”

14. Il en existe trois : le registre familial, le registre courant et le registre soutenu.

15. empan moyen inférieur à huit « mots »

16. L'éloignement qui peut se produire avec la grammaire de l'écrit ne veut pas forcément dire que la grammaire de l'oral est fautive.

(a)	fenêtre normale	[-----]
(b)	fenêtre interrompue	[----<
(c)	fenêtre non initiée	>----]
(d)	fenêtres de bafouillage	[---[---[---
(e)	fenêtres de recherche lexicale	---]----]]
(f)	fenêtres avec mise en commun ou apo kōinou :	
(f1)	mise en commun du segment central	[1a...[1b...1a]...1b]
(f2)	mise en commun du segment gauche	[1---[---1a]---1b]
(f3)	mise en commun du segment droit	[1a...[1b ---]---1]
<i>I¹le défaut qu'ils ont¹ / I²ils ont une chambre [pour eux^{2a}] / pour payer moins cher^{2b}] et // I³ils prennent un copain ou deux³] et alors voilà / mais I⁴les bains qui c'est qui les paye⁴] I⁵ils payent pour un bain⁵] I⁶ ils payent pas pour trois⁶] / ah</i>		

Séquence : “le défaut qu’ils ont ils ont une chambre pour eux pour payer moins cher et ils prennent un copain ou deux et alors voilà mais les bains qui c’est qui les paye ils payent pour un bain ils payent pas pour trois ah”

FIGURE II.2 – Exemple de fenêtrage syntaxique de la séquence du corpus CaFE (Bazillon *et al.*, 2008b)

si l’on cherche à l’analyser au travers d’une syntaxe basée sur le modèle de la proposition, contrairement à la parole préparée, *a fortiori* écrite, où le locuteur vise “une formulation à la fois ‘bien formée’, élaborée, explicite et compacte” (Koch et Oesterreicher, 2001). Les linguistes aujourd’hui prennent en compte les acceptations de l’écrit et de l’oral. On peut ainsi citer la “grande grammaire du français” (Abeillé et Godard, 2021).

L’ensemble des caractéristiques linguistiques présentées est récapitulé dans le tableau II.2.

1.1.3 Les difficultés induites pour la reconnaissance automatique de la parole

Les caractéristiques acoustiques et linguistiques décrites pour la parole spontanée par rapport à la lecture ou la parole préparée entraînent notamment une forte baisse des performances lorsque des systèmes appris sur de la lecture ou une parole très préparée sont utilisés pour transcrire automatiquement de la parole spontanée (Furui, 2002). Des études ont étudié certaines de ces caractéristiques et les difficultés qui en résultent pour les systèmes de reconnaissance automatique de la parole.

Le traitement des **disfluences** représente une première difficulté pour les systèmes à cause de l’interruption de la structure syntaxique du message d’après de nombreuses études (Bortfeld *et al.*, 2001; Adda-Decker *et al.*, 2003; Stouten *et al.*, 2006; Bazillon *et al.*, 2008a; Goldwater *et al.*, 2008; Clavel *et al.*, 2013). D’après Stouten

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole

Hésitation	Bigi et Meunier (2018)
Backchannels	Bigi et Meunier (2018)
Interjections	(Koch et Oesterreicher, 2001)
Registre de langue	très soutenu lorsque le locuteur parle en tant que représentant de sa profession (Blanche-Benveniste et Bilger, 1999) peut changer dans une même prise de parole (Blanche-Benveniste et Bilger, 1999)
Mots informels	(Torreira et al., 2010) verlan (Torreira et al., 2010) gros-mots/insultes (Torreira et al., 2010; Clavel et al., 2013)
Néologismes	(Fohr et al., 2015)
Fenêtres syntaxiques	particulières (Luzzati, 2004)
Interruptions	parenthèses ou commentaires (Blanche-Benveniste et Bilger, 1999) corrections (Koch et Oesterreicher, 2001)
Agrammaticalité	(Dufour, 2008) accords non réalisés (Koch et Oesterreicher, 2001) doubles négations absentes (Torreira et al., 2010)

TABLE II.2 – Caractéristiques linguistiques les plus retrouvées en parole spontanée

et al. (2006), ces interruptions impactent notamment les capacités de prédiction des modèles de langue et peuvent amener à une erreur de décodage qui impactera probablement à son tour le décodage des mots suivants. Goldwater *et al.* (2008) précisent que toutes les disfluences n'ont cependant pas le même impact. Dans leur étude sur les effets des répétitions, fragments de mots et pauses pleines, ils démontrent que les répétitions de mots en position non finale et les mots situés autour de fragments de mots sont les plus sujets aux erreurs.

Si les disfluences sont effectivement problématiques, Adda-Decker *et al.* (2003) a pu montrer dans leur étude qu'elles n'étaient la cause que de 12,5% des erreurs, contre 25,1% pour les **réductions**. Les difficultés amenées par les réductions ont également été notées par Dufour (2008) et Gabler *et al.* (2023). Selon Dufour (2008), l'ajout de variantes de prononciations, dont celles liées aux réductions, dans le lexique de prononciation d'un système de reconnaissance de la parole peut être bénéfique et résoudre certains problèmes de suppression de mots ou de substitutions. Malheureusement, cela peut aussi finir par perdre le système et créer des problèmes d'insertions lorsque le nombre de variants est trop élevé. Adda-Decker *et al.* (2003) font également le lien entre réductions et insertions. En effet, les réductions de type "y a" ou "c'est pas" seraient régulièrement transcris par les systèmes dans leur forme complète ("il y a" ou "ce n'est pas" dans ce cas). Les ré-

ductions peuvent également être à l'origine de substitutions, comme “ça” transcrit comme “cela”. Cette hypercorrection est introduite en raison d'un apprentissage des systèmes de reconnaissance automatique de la parole sur une parole proche de l'écrit. Ce type de réductions est très fréquent et leur impact sur le WER n'est donc pas négligeable comme le soulignent les auteurs.

Une autre difficulté pour les systèmes est la **parole superposée** (Adda-Decker *et al.*, 2003; Bazillon *et al.*, 2008b; Kneubühler, 2022; Gabler *et al.*, 2023). Il est d'usage de retirer les passages de parole superposée des ensembles d'évaluation des systèmes (Adda-Decker *et al.*, 2003; Elloumi *et al.*, 2018). Adda-Decker *et al.* (2003) précisent que la parole superposée produite par un locuteur comme *backchannel* à ce qu'un second locuteur est en train de dire ajoute du bruit et est à l'origine de nombreuses erreurs.

Enfin, la dernière difficulté que nous abordons est celle due à la difficulté de production d'une **transcription** humaine parfaite lorsqu'il s'agit de parole spontanée. (Gabler *et al.*, 2023, p. 2) précise même que “*The perfect reconstruction of speech is fundamentally impossible]. This is what] distinguishes spontaneous from read speech data. (...) [There is] no real ground truth for annotations in the case of spontaneous speech data.*”¹⁷. Cette absence d'une vérité de terrain fait que les systèmes apprennent des transcriptions approximatives et que les performances sont comparées à une transcription possible mais pas certaine.

Selon une étude de Zayats *et al.* (2019) dont les résultats sont rapportés par Gabler *et al.* (2023), les variations dans la transcription humaine pourraient par exemple concerner 5 à 10% des mots du corpus Switchboard¹⁸. Cette étude montre que les transcripteurs sont notamment susceptibles de mal comprendre les mots peu fréquents, moins porteurs d'informations, bien que cette tendance ne se vérifie pas pour les morphèmes utilisés comme interjections, *backchannels* ou pauses pleines en parole spontanée, ce qui est également reporté par Stolcke et Droppo (2017). Les disfluences et répétitions sont également sources d'erreurs, ce qui est également rapporté par Lickley et Bard (1996) et Adda-Decker *et al.* (2003). Pour Gabler *et al.* (2023), les erreurs de transcriptions sont notamment du fait de la structure grammaticale particulière de la parole spontanée et de ses réductions. Bazillon *et al.* (2008b) relèvent également comme sources d'erreurs les problèmes d'homonymies (ces/ses, été/était...), les problèmes d'accords de participes passés de verbes pronominaux, ou encore les pluriels incertains (pas de problème/ problèmes), les noms propres, mais aussi la superposition de la parole puisqu'il est parfois difficile de

17. La reconstruction parfaite de la parole est fondamentalement impossible. C'est ce qui distingue les données de parole spontanée des données de parole lue. (...) Il n'y a pas de vérité absolue pour les annotations dans le cas des données de parole spontanée. (trad : ChatGPT)

18. Switchboard est un corpus américain de conversations téléphoniques.

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole

bien discerner ce qui est prononcé, et par qui. Enfin, le texte n'existant pas en amont de la prise de parole, les transcripteurs peuvent se voir confrontés à des erreurs de perception les menant à transcrire quelque chose qui n'a pas été dit. C'est ainsi que certains chercheurs désignent également les transcriptions comme des interprétations (Bilger et Gedo, 1997; Bazillon *et al.*, 2008b). Si ces variations dans la transcription ne représentent pas vraiment une caractéristique de la parole spontanée en elle-même, ce sont bien les caractéristiques de la parole spontanée qui rendent la tâche plus complexe.

1.1.4 Déterminer des niveaux de spontanéité en fonction de ces caractéristiques

Dufour *et al.* (2014) ont travaillé sur la catégorisation automatique de segments en fonction de niveaux de spontanéité. Afin de pouvoir évaluer leur système et s'assurer de sélectionner des caractéristiques pertinentes pour la classification, deux personnes ont annoté un corpus de 11 heures d'enregistrements de journaux d'informations en français selon trois classes : préparé, peu spontané, très spontané. La méthode d'annotation n'est pas précisée, mais un haut coefficient de concordance Kappa a été obtenu sur une heure d'enregistrements annotée par deux annotateurs (0,852). La classe de ces segments est ensuite mise en relation avec les caractéristiques de la parole spontanée sélectionnées. Ainsi, les auteurs démontrent lorsque la spontanéité augmente :

- une hausse de la durée moyenne et de la déviation standard de chacun des paramètres acoustiques étudiés : durée des voyelles, durée d'une syllabe en fin de mot (appelée *mélisme*), débit phonémique (estimé d'abord avec les pauses et *fillers*¹⁹ puis sans), et la f0 (*pitch*)
- une hausse du pourcentage de *fillers* et répétitions
- une baisse du pourcentage de noms propres
- une hausse de la taille moyenne des groupes syntaxiques
- une gradation ascendante des mesures de confiance d'un système de reconnaissance automatique de la parole : les systèmes ont en effet plus de difficulté à bien transcrire les segments de parole spontanée que les segments de parole préparée.

Le *Word Error Rate* obtenu sur les données²⁰ est reporté dans le tableau II.3. Il est intéressant de noter que le WER augmente lorsque la spontanéité augmente et

19. hum, heu...

20. Les auteurs ont utilisé la technique du *leave-one-out* consistant à réserver ici 1 enregistrement pour le test et les 10 autres pour l'apprentissage, puis de faire tourner ce processus jusqu'à ce que toutes les données aient été vues.

que le WER global n'est pas représentatif du WER obtenu sur la parole qualifiée de très spontanée.

Niveau de spontanéité	Durée	#Segments	#Mots	WER
préparé	3h40	1 228	39 984	10,1%
peu spontané	3h50	1 339	44 245	18,4%
très spontané	3h30	1 247	42 515	28,5%
total	11h	3 814	126 744	15%

TABLE II.3 – Résultats du système de reconnaissance automatique de Dufour *et al.* (2014) selon la catégorisation manuelle de segments comme “préparé”, “peu spontané” ou “très spontané”

Du point de vue de la classification automatique de segments, l'étude montre que l'utilisation conjointe des informations acoustiques, linguistiques et de la mesure de confiance du système sur les transcriptions automatiques permet d'obtenir les meilleurs résultats sur la classification de chaque segment individuellement (que les auteurs qualifient de “décision locale”). L'approche de “décision globale” permettant de prendre en compte la classification des segments voisins pour la classification d'un segment améliore encore les résultats. Le tableau II.4 rapporte les résultats obtenus en terme de précision, rappel et F-mesure²¹.

Il est intéressant de retenir de cette étude que la classification de segments comme préparés et très spontanés fonctionne plutôt bien. Cependant, la classification de segments comme peu spontanés fonctionne plutôt mal ce qui, comme le précisent les auteurs “is not surprising as these segments can be easily misclassified as prepared speech on one side or high spontaneous on the other side”²².

Les auteurs eux-mêmes ont qualifié cette tâche d'annotation manuelle des segments comme plus ou moins spontanés de “compliquée” (Dufour *et al.*, 2010)²³. La catégorisation ayant été effectuée au niveau du segment, les annotateurs se sont appuyés sur les caractéristiques “observables” de la parole telles que le début de parole, les disfluences (reprises, répétitions, hésitations, *etc.*)... La production de parole spontanée est sensible à la variabilité individuelle comme le soulignent les

21. La F-mesure ou le F-score est une mesure de la performance d'un modèle de classification. Elle combine les mesures de précision et rappel, elles-mêmes basées sur les taux de vrais positifs, faux positifs et faux négatifs. L'idée de la F-mesure est de s'assurer qu'un classificateur fait de bonnes prédictions de la classe pertinente (bonne précision) en suffisamment grand nombre (bon rappel) sur un jeu de données cible. (Définition de Wikipédia)

22. N'est pas surprenant car ces segments peuvent être facilement classés à tort comme un discours préparé d'un côté ou comme hautement spontané de l'autre côté (traduction : chatGPT)

23. La publication de 2014 est une version étendue et complétée de leurs travaux de 2010

1 Les caractéristiques de la parole spontanée et les difficultés en reconnaissance automatique de la parole

Mesure	décision locale %	décision globale %
Parole préparée		
Précision	59,8	66,8
Rappel	65,1	69,6
F-mesure	62,3	68,2
Parole peu spontanée		
Précision	47,0	54,3
Rappel	43,5	51,8
F-mesure	45,2	53,0
Parole très spontanée		
Précision	66,7	73,0
Rappel	66,3	73,5
F-mesure	66,5	73,2

L'ensemble des valeurs a été multiplié par 100 par les auteurs

TABLE II.4 – Évaluation de la classification automatique de segments selon trois types de parole en terme de précision, rappel et F-mesure Dufour *et al.* (2014)

auteurs dans l'introduction de leur article de 2014 : “In addition to disfluencies, spontaneous speech is also characterized by ungrammaticality and a language register different from the one that can be found in written texts. Depending on the speaker, the emotional state and the context, the language used can be very different”²⁴. D'autres auteurs ont pu montrer que les caractéristiques de la parole présentées en sections 1.1.1 et 1.1.2 sont sensibles à cette variabilité intra-locuteur. Ainsi, Lancien (2020) montre que la variation de durée vocalique (durée d'une voyelle) varie de façon différente chez chaque locuteur. Llisterri (1992) rappelle que certains locuteurs peuvent donner l'impression d'une grande précision et donc d'une grande clarté, même dans un discours rapide. Selon Bodur *et al.* (2022), les réductions peuvent, elles aussi, être liées au locuteur.

Si l'on regarde les disfluences, Bortfeld *et al.* (2001) précise qu'elles peuvent être les témoins de l'état émotionnel d'un locuteur (des difficultés de planification de sa parole, ou des problèmes de confiance), ce qui implique qu'elles peuvent être présentes en grand nombre dans une parole pourtant peu spontanée. Adda-Decker *et al.* (2003) renforce ce constat en mesurant des différences significatives entre les locuteurs selon leur rôle. Les auteurs précisent que les répétitions sont dominantes chez les journalistes, notamment lors de leurs tentatives d'interruption des

24. En plus des disfluences, le discours spontané se caractérise également par l'agrammaticalité et un registre de langue différent de celui que l'on peut trouver dans les textes écrits. Selon le locuteur, son état émotionnel et le contexte, le langage utilisé peut être très différent.

interviewés, tandis que les reprises sont dominantes chez les interviewés. De même, Bazillon *et al.* (2008a) précise qu'une parole préparée contenant de nombreux faux départs ou répétitions peut sonner comme une parole spontanée, et que certains discours spontanés sont très similaires à des discours préparés. En effet, les répétitions peuvent également être utilisées pour des questions de rythme, d'esthétique, ou pour garder la parole (Andersen, 2012). De même, Christodoulides (2020) précise que les pauses peuvent être utilisées pour ajouter un effet rhétorique.

La catégorisation du niveau de spontanéité de chaque segment est chronophage (puisque elle demande d'analyser manuellement de nombreux segments) et complexe : une des caractéristiques de la parole spontanée réside dans son aspect irrégulier et parfois contradictoire, possiblement en raison de fortes différences entre les locuteurs/langues. Ainsi, les catégorisations portent souvent sur un plus haut niveau comme le corpus (Gabler *et al.*, 2023; Szaszák *et al.*, 2016) ou encore l'émission (donc l'enregistrement) dans un corpus d'émissions de télévision (Deléglise et Lailler, 2020). Cependant, la catégorisation n'en devient, à première vue, pas plus facile étant donné que l'on peut retrouver de la parole spontanée dans une prise de parole préparée et inversement (Bazillon *et al.*, 2008a; Dufour *et al.*, 2010). Nous proposons de revenir sur la notion de “parole spontanée” pour mieux cerner les contours de cette parole multiple.

2 La parole spontanée : caractériser ses variations

2.1 Définir la parole spontanée

Luzzati (2007, p 27) offre une définition complète de la parole spontanée. Il caractérise d'abord le discours oral spontané comme “insaisissable et fluctuant”, et précise plus loin dans son article :

“lorsqu'il s'agit de conversation, on bascule sauf exception dans l'oral spontané, dans la 'parole' ou plutôt pour être plus précis, dans la 'parole conversationnelle', que nous définissons comme l'ensemble des énoncés oraux conçus et perçus dans le fil de leur énonciation. (...) On quitte de fait l'univers de la phrase, celui de la langue préparée, celui de l'erreur qui s'efface et se rature, pour basculer du côté des énoncés non prémédités, dont l'émetteur est le premier auditeur, dans lesquels l'erreur se traduit par un allongement du message”

L'aspect spontané n'est plus vu ici dans son aspect “résultat”, mais plutôt au travers de sa production. Ce qui rend la parole spontanée, c'est le fait de l'improviser et de la partager comme elle vient. Le locuteur s'offre la possibilité de prendre

2 La parole spontanée : caractériser ses variations

un temps pour réfléchir à la suite du message (pouvant entraîner des hésitations), d'amorcer un message et d'y renoncer (pouvant entraîner des faux départs ou des répétitions), ou encore de le modifier (pouvant entraîner des reprises et réparations). La définition caractérise le processus de production/perception des énoncés et non les énoncés en eux même, ainsi rien n'empêche un locuteur de produire une parole spontanée ne comprenant pas ou peu de manifestations de ces processus.

(Fujisaki, 1997, p 38) quant à lui propose la définition suivante²⁵ :

“The word spontaneous is understood here to describe the attribute of something that occurs by its own internal force, motivation etc., rather than by external ones.”

La parole spontanée serait donc une parole dont le locuteur est le seul initiateur, en réponse à une impulsion interne qui lui est propre. Cette définition reprend le côté "non prémédité" des énoncés dont parlait Luzzati (2007). La définition de Andersen (2012) reprend, elle, la notion de processus en cours :

"l'oral est perçu comme un processus. Les énoncés sont créés spontanément et linéairement, ce qui fait que les corrections obligent de revenir en arrière et invitent ainsi à des reprises, des reformulations, des faux-départs (...) : l'oral est un brouillon (...) l'oral est la vie de tous les jours"

De ces définitions, nous pouvons retenir :

- l'aspect linéaire de ce type de parole : "conçu dans le fil de leur énonciation", "processus", "énoncés créés linéairement", "on quitte l'univers de la langue préparée"
- le fait que les énoncés sont non préparés et créés en réponse à une impulsion, une motivation interne et dont l'émetteur est le premier auditeur : "non prémédités", "énoncés créés spontanément", "se produit par sa propre force interne"
- les énoncés ainsi conçus invitent à des reprises, des reformulations et des faux départs et engendrent des 'erreurs' ou corrections visibles : "l'erreur se traduit par un allongement du message", "les corrections obligent de revenir en arrière", "brouillon".

La parole spontanée est donc un processus de production/perception d'énoncés "au fin de leur énonciation" et cette contrainte génère les caractéristiques de la

25. Le mot 'spontané' est ici compris pour décrire l'attribut de quelque chose qui se produit par sa propre force interne, motivation, etc., plutôt que par des influences externes." (trad : ChatGPT)

parole spontanée mentionnées par les notions d'erreurs, de corrections, de reprises, reformulations et faux départs qui se rapportent principalement aux disfluences.

Certains auteurs mentionnent l'existence d'un continuum des types de parole dont l'un des extrêmes pourrait être la lecture (Fujisaki, 1997; Eshkol-Taravella *et al.*, 2011), la parole préparée (Gabler *et al.*, 2023) ou encore la parole spontanée formelle (Labov, 1973; Torreira *et al.*, 2010), et l'autre extrême étant la parole spontanée non planifiée (Gabler *et al.*, 2023), informelle (Labov, 1973; Fujisaki, 1997; Torreira *et al.*, 2010), naturelle (Eshkol-Taravella *et al.*, 2011) ou libre (Fujisaki, 1997). Ce qui se trouve au milieu est parfois qualifié de “parole contrainte” (Adda-Decker *et al.*, 2003), parole semi-préparée (Gabler *et al.*, 2023), parole contrôlée ou semi-spontanée (Bertrand *et al.*, 2022). Les termes utilisés pour caractériser la parole le long d'un continuum depuis la lecture jusqu'à une parole non planifiée ou spontanée sont très variés et peuvent recouvrir des réalités différentes. Plutôt que de chercher à caractériser le type de parole spontané, nous avons cherché les facteurs de variation qui soutiennent sa diversité.

2.2 Les facteurs de variation de la spontanéité

2.2.1 Les facteurs interpersonnels

L'un des facteurs fréquemment relevé dans la littérature est la notion de rôle. Il se rapporte aux interactions dans lesquelles les locuteurs ne sont pas sur un pied d'égalité. Il peut y avoir une différence de rôle social (ou statut social), comme un conseiller municipal s'adressant à des représentants de comité de quartier (Blanche-Benveniste et Bilger, 1999), un chef d'état prenant la parole (Goldman *et al.*, 2009) ou le représentant d'une profession, qui aura tendance à utiliser un discours plutôt soutenu (Blanche-Benveniste et Bilger, 1999; Bilger et Cappéau, 2004) et lent, pour marquer l'aspect solennel dans le cadre d'une prise de parole politique (Goldman *et al.*, 2009). Il existe également des différences de rôles “dans le discours” parmi lesquels *commentateur*, *correspondant spécial*, *présentateur*, *invité*, *auditeur*, *interviewé* et *interviewer* pour lesquels les niveaux de préparation (et donc les niveaux de spontanéité) ne vont pas être les mêmes : beaucoup de parole préparée chez le commentateur, le correspondant spécial et le présentateur, une parole plus spontanée chez l'invité, l'auditeur et l'interviewé et un mélange des deux pour l'expert et l'interviewer (Dufour *et al.*, 2014).

Le nombre d'interlocuteurs est également très souvent considéré comme un facteur important. Il est fait une différence entre une parole adressée à une seule personne, où l'enjeu premier sera de ne pas rompre la relation qui existe (Agha, 2006; Romera Ciria, 2019), et une parole publique, liée à la notion de performance

et de pouvoir²⁶ où les locuteurs travaillent leur éloquence et leur rhétorique²⁷ pour convaincre leur auditoire. La spontanéité est moindre en parole publique puisqu'une attention très forte est portée au discours afin d'éviter une erreur ou inexactitude (dont la correction impliquerait une nouvelle intervention publique) et pour maintenir l'attention de l'auditoire. Labov (1973) précise que lorsque l'attention qu'un locuteur porte à sa parole se réduit, cela l'amène à l'emploi d'un style de parole relâché et informel. La spontanéité dans la parole peut ainsi varier en fonction du nombre de personnes à qui un locuteur s'adresse.

Le degré d'intimité entre les locuteurs est également un facteur important. Il est, d'ailleurs, souvent celui qui revient en premier lorsque l'on veut expliquer à quelqu'un ce qu'est la parole spontanée : c'est la parole que l'on utilise entre amis (Torreira *et al.*, 2010; Bodur *et al.*, 2022). Ce type de conversation permet de positionner les locuteurs dans une situation aussi naturelle que possible (Llisterri, 1992), et ainsi de faire émerger une parole plus spontanée. Dans une étude sur les phénomènes de réductions qui se base sur une analyse du corpus CID (Corpus of Interactional Data), (Bodur *et al.*, 2022, p 3) précise que les personnes enregistrées sont des "collègues entretenant des relations assez familières". Les phénomènes de réduction étant plus présents en parole spontanée que dans les autres types de parole, les auteurs considèrent que le degré d'intimité entre les locuteurs joue sur la spontanéité et donc sur l'apparition de ces phénomènes. Lancien (2020) le note également. Dans son étude sur le rôle de la distance sociale entre locuteurs, elle met en évidence une réduction de l'espace acoustique vocalique, avec des voyelles plus proches les unes des autres dans les "styles" où les locuteurs sont les plus intimes (comparaison de discussions entre conjoints et entre inconnus). Elle montre ainsi une hypoarticulation plus forte lorsque les locuteurs ont des relations familiales. Pour autant, il est important de noter que le degré d'intimité ne semble pas influer sur la production de toutes les caractéristiques de la parole spontanée. Bortfeld *et al.* (2001) a par exemple pu démontrer qu'une conversation entre un couple marié et des étrangers ne montre pas de différence de taux de reprises, répétitions ou autres tics de langage (*fillers*).

Les créateurs du PFC (Laks *et al.*, 2009) et du NCCFr (Torreira *et al.*, 2010) ont également mentionné que les conversations recueillies sont plus ou moins formelles selon le niveau d'intimité entre les locuteurs. Cela rejoint les observations de Labov (1973) lorsque, en contexte d'interview, l'interviewé a quelque fois produit une parole spontanée informelle lorsqu'il s'adressait à une personne du foyer, sa conjointe ou ses enfants. Les études de Traverso (1996); Andersen (2012) ont également noté

26. La parole est régulièrement comparée à une arme, voir notamment les livres de Périer (2017) et Viktorovitch (2021)

27. Il est intéressant de noter qu'en date du 19/10/2022, la page Wikipédia 'public speaking' en anglais amenait sur une page intitulée 'rhétorique' lorsque l'on affichait sa version française.

la prédominance de l'informel et du “léger” dans la conversation familiale, avec une volonté des locuteurs d'insister sur leur complicité, leurs savoirs et expériences partagées.

2.2.2 Les facteurs environnementaux

Les facteurs que nous qualifions d'environnementaux concernent le lieu de l'interaction, la présence (physique ou distante) des interlocuteurs, ainsi que le canal de communication.

Le lieu, tout d'abord, semble avoir une influence sur le type de parole. En effet, Equipe Delic *et al.* (2004) parlent d'ailleurs de “parole institutionnelle” pour désigner la parole produite sur le lieu de travail. Llisterri (1992) distingue trois types de parole spontanée en fonction du lieu dans lequel elle est produite : les laboratoires (parole contrainte), la TV/radio et l'environnement naturel des locuteurs.

Le lieu est également un facteur à prendre en compte pour les créateurs des corpus ESLO, CLAPI et PFC. Ils ont en effet fait le choix de varier les lieux d'enregistrement (centre médico-psychopédagogique, espaces publics, maison, université²⁸, institutions, services publics, chez le médecin...). Le choix de lieux publics ou familiers (la rue, les bars, le métro, la plage... ou toute visite à un ami) pour faire émerger une parole plus relâchée et informelle était déjà relevé par Labov (1973). Nous sommes là au croisement entre les facteurs interpersonnels et les facteurs environnementaux.

La présence physique ou non des locuteurs semble également être un facteur de variation. La proximité et l'échange direct amènent à une économie des gestes articulatoires (hypoarticulation), notamment dans les situations informelles (Duez, 2001). Par ailleurs, (Bortfeld *et al.*, 2001) a pu mesurer plus de disfluences (corrections, faux-départs, fillers, répétitions) pour des conversations ne comprenant que le canal audio, en comparaison à des conversations où les locuteurs se trouvent dans le même lieu, avec un contact visuel.

Enfin, la présence d'un auditoire, c'est-à-dire de personnes qui rentrent pas en interaction mais à qui le message est tout de même destiné a également un effet sur la parole. Dans le cadre d'une émission radiophonique, d'une conférence ou d'un sermon par exemple, l'auditoire distant amène les locuteurs à adopter une articulation claire afin d'être compris de tous (Duez, 2001; Adda Decker, 2006; Lancien, 2020).

28. L'université étant désignée comme un lieu neutre dans le PFC.

2.2.3 Les facteurs plus personnels

Nous avons relevé plusieurs autres facteurs de variation de la spontanéité de la parole, telles que les émotions générées par la situation d'enregistrement ou la proximité du locuteur avec le thème de la discussion.

En effet, l'état émotionnel est très souvent nommé comme un facteur de variation de la parole (Llisterri, 1992; Dufour *et al.*, 2010; Laukka *et al.*, 2011; Alghowinem *et al.*, 2012). Llisterri (1992) précise que certains auteurs font parfois le choix d'interviewer des locuteurs professionnels pour cette raison : le niveau de stress créé par certains lieux d'enregistrement (studio) et la présence du matériel d'enregistrement serait moins élevé chez eux que chez une personne lambda.

Concernant le thème de la discussion, il existe quelques corpus pour lesquels un thème est imposé (soit pour l'ensemble des enregistrements soit pour une partie d'entre eux) : Nijmegen Corpus of Casual French (NCCFr) (Torreira *et al.*, 2010), Phonologie du Français Contemporain (PFC) (Laks *et al.*, 2009), Traitement de Corpus Oraux en Français (TCOF) (André et Canut, 2010)... Dans le PFC et TCOF, les locuteurs sont invités à parler de leurs activités, de leur enfance, leurs expériences ou d'un savoir-faire, et ce afin de faire émerger une parole naturelle. Les locuteurs sont invités à parler d'eux ou de ce qu'ils connaissent, ce qui les invite à réduire l'attention portée au discours (Wagner *et al.*, 2015). Les facteurs de variation que sont les émotions, le stress ou plus globalement l'état du locuteur ont donc leur importance, mais une grande liberté est laissée aux locuteurs de raconter ou non les difficultés rencontrées dans les différentes périodes de leur vie.

2.3 Revue des modèles à une ou plusieurs dimensions relatifs à la variation stylistique

Nous pouvons noter que certaines situations favorisent une parole plus spontanée, à savoir :

- l'absence de rapport hiérarchique entre les locuteurs
- la présence d'un nombre réduit d'interlocuteurs
- le lien amical ou familial
- l'environnement naturel des locuteurs (lieux publics/familiers)
- la présence physique des locuteurs
- la proximité du locuteur avec le thème de la discussion

Certains de ces facteurs sont étroitement liés comme les rôles des locuteurs et le lieu de l'enregistrement et des études ont proposé de définir des modèles à une ou plusieurs dimensions de description de la variabilité de la parole spontanée. Nous avons réuni ces études suivant les dimensions envisagées.

2.3.1 Le niveau de formalité

Plusieurs auteurs considèrent que les styles de parole varient le long d'un axe représentant la formalité (Zwicky, 1972; Tarone, 1988; Eskenazi, 1993; Dewaele, 1996; Joos, 2012; Gabler *et al.*, 2023). Ainsi, dans les études de (Zwicky, 1972; Joos, 2012), ce sont jusqu'à cinq styles qui sont identifiés : *intime*, *informel*, *consultatif* (professionnels), *formel* et *figé*²⁹ (très formel). Pour Dewaele (1996), la formalité dépend de deux facteurs : le contexte (ou situation) dans lequel est produit le discours et la personnalité du locuteur, à savoir son degré d'extraversion. Dans une situation supposée très formelle, plus le locuteur serait extraverti, moins l'entretien respecterait cette formalité (voir figure II.3). A l'inverse, l'auteur précise qu'un locuteur introverti produira un discours "moins ancré dans le contexte spatio-temporel" ce qui lui confère une "impression de discours écrit".

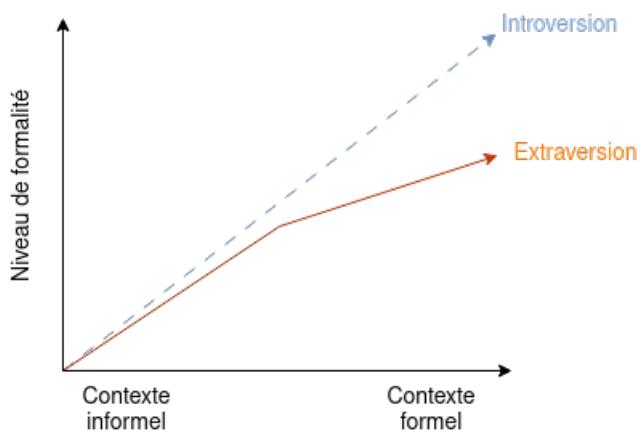


FIGURE II.3 – Différence de niveau de formalité en fonction du contexte plus ou moins formel entre locuteurs introvertis et extravertis.

2.3.2 L'attention portée au discours

Pour Labov (1973), l'évolution des styles de parole sur un continuum se fait en fonction du niveau d'attention porté au discours. Ainsi, il place les paires minimales³⁰ d'un côté de son continuum, et la parole informelle de l'autre. Il précise dans son chapitre que le style informel (vernaculaire) est utilisé essentiellement entre ceux qui partagent le plus de connaissances et où le minimum d'attention est porté à la parole ("the more casual or vernacular style is used primarily with those

29. Original : *frozen*

30. Paire de quasi-homonymes dont le rapprochement permet d'étudier les traits distinctifs d'un phonème. La paire minimale tard/dard permet de distinguer /t/ et /d/ comme voisé-non-voisé (Mounin1974). (définition du CNRTL, consulté le 24/01/2024)

who share the most knowledge together, where the minimum amount of attention is paid to speech"). Au-delà du niveau d'attention porté au discours, le style de parole évoluerait également en fonction du niveau d'intimité entre les locuteurs, ces deux axes variant finalement en même temps, mais dans des sens positifs et négatifs opposés (voir Figure II.4).

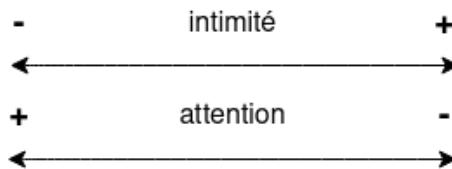


FIGURE II.4 – Dimensions de variation des styles de parole selon Labov (1973)

2.3.3 Le niveau de contrôle

Wagner *et al.* (2015) considèrent le style comme un terme générique englobant des situations de parole conduisant à une variation phonétique, où le niveau de contrôle est le principal facteur descriptif³¹. Ils placent notamment sur leur continuum les types de situations de parole tels que “lu”, “scripté”, “non scripté”, “informel” et “conversationnel” et précise que “lu” et “conversationnel” en forment les deux extrêmes dont il est possible de retrouver des éléments dans chacune des situations préalablement citées. Il précise un peu plus loin dans son article que l'on peut considérer le niveau de contrôle sur la parole et le niveau de formalité comme directement liés, ainsi que le montre la figure II.5.

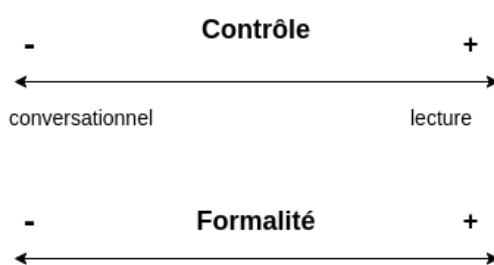


FIGURE II.5 – Variation des styles de parole selon le contrôle (Wagner *et al.*, 2015)

31. “A broad cover term for speaking situations leading to phonetic variation where the level of control is the main descriptive factor” (p.2)

2.3.4 Modèles prenant en compte d'autres dimensions

Selon Goldman *et al.* (2009), les genres ou styles de parole se distingueraient sur un même continuum en fonction des trois axes suivants : le degré de préparation du discours, le type d'audience ("micro", "face-à-face", "beaucoup", tel que rapporté dans l'article), et la médiatisation ou non (voir figure II.6).

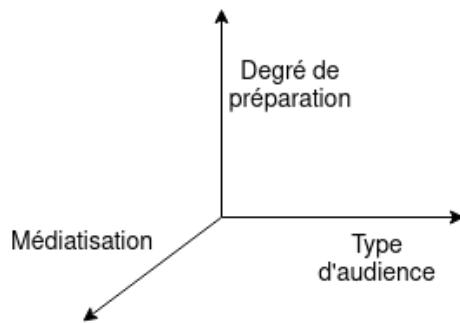


FIGURE II.6 – Catégorisation des genres/styles de parole de (Goldman *et al.*, 2009)

Eskenazi (1993) propose le même nombre de dimensions mais des axes différents, à savoir la familiarité entre les interlocuteurs, la classe sociale (capturant le degré de sophistication depuis le "*colloquial and/or 'lower class' tone*" - le ton familier ou de la classe sociale basse- au "*'highly cultivated' and/or 'upper class' tone*" - le ton très cultivé ou de la classe sociale haute) et l'intelligibilité³². Il est intéressant de noter que sur la figure II.7 issue de Eskenazi (1993), la dimension de la classe sociale correspond à un niveau de formalité.

Le modèle proposé par Koch et Oesterreicher (2001) définit un "continuum communicatif", ayant pour modalités la parole et l'écrit, l'axe horizontal de leur modèle représentant une distance communicative vs une immédiateté communicative, et la dimension verticale représentant le rapprochement ou l'éloignement du passage d'un code graphique (texte) à un code phonique (parole), ou inversement. Ainsi, le cas *a* représentatif d'une conversation spontanée se situe, sur l'axe horizontal, au plus proche de l'immédiat communicatif et sur l'axe vertical, au plus éloigné du code graphique. *A contrario*, la lecture à haute voix d'un texte de loi représentée par *i'* est du côté d'une grande distance communicative sur l'axe horizontal, et proche du code graphique sur l'axe vertical (voir la figure II.8).

Les auteurs introduisent également des valeurs paramétriques du langage parlé et du langage écrit (voir figure II.9). Celles-ci "caractérisent le comportement

32. Il précise également que l'on pourrait prendre en compte un quatrième axe de variation : l'interaction entre le locuteur et son environnement, mais ne l'intègre pas dans la suite de son

2 La parole spontanée : caractériser ses variations

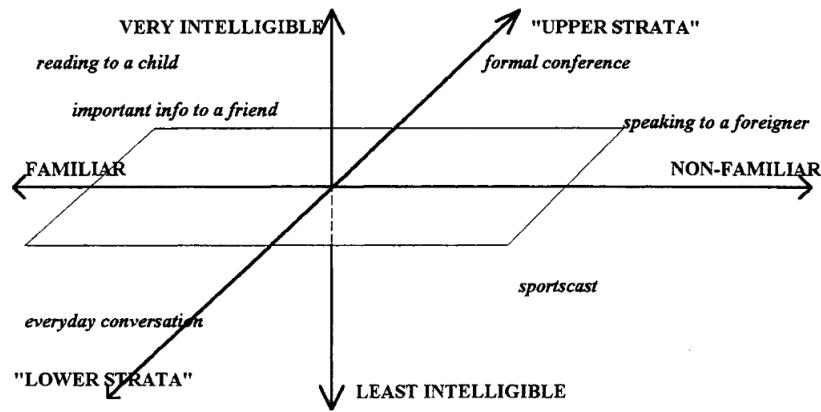


FIGURE II.7 – Les dimensions le long desquelles les styles de parole peuvent être situés selon Eskenazi (1993)

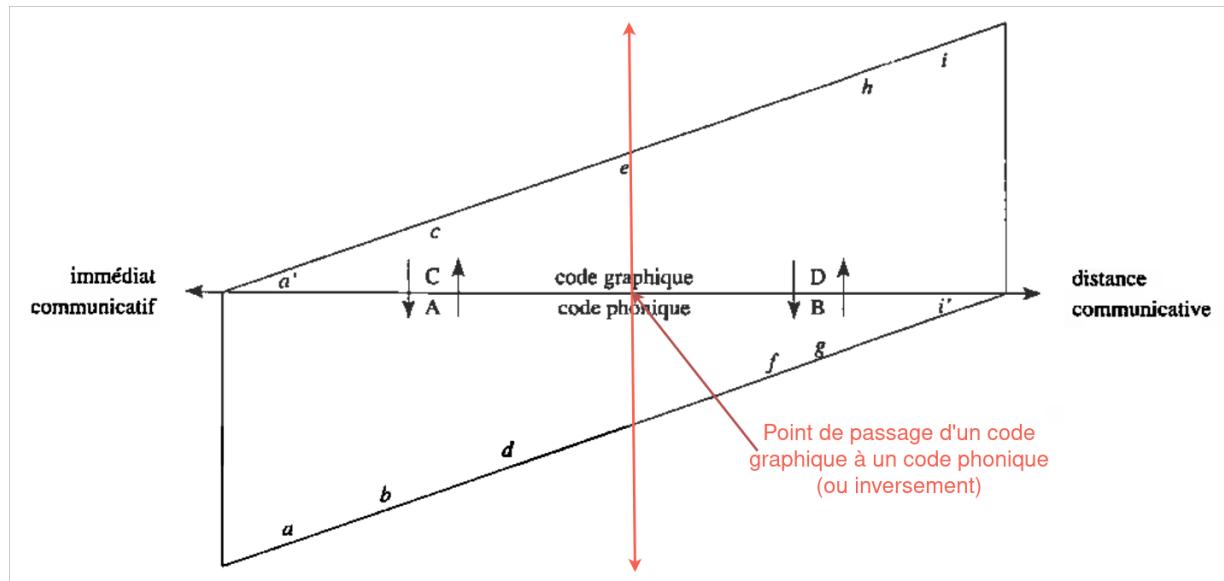


FIGURE II.8 – Le continuum communicatif de Koch et Oesterreicher (2001)

[HAUT] a' : transcription de conversation spontanée, c : lettre personnelle entre amis, e : interview dans la presse, h : article de loi, i : texte de loi [BAS] a : conversation spontanée, b : coup de téléphone, d : entretien professionnel, f : sermon, g : conférence scientifique, i' : lecture à voix haute d'un texte de loi // A,B,C et D représentent des secteurs // Les annotations en orange ont été ajoutées par nos soins pour faciliter la compréhension du schéma.

communicatif des interlocuteurs par rapport aux déterminants situationnels et contextuels". La validation de l'ensemble des valeurs de gauche représente l'immédiat communicatif et l'ensemble des valeurs de droite représente au contraire une étude.

① communication privée	communication publique ①
② interlocuteur intime	interlocuteur inconnu ②
③ émotionnalité forte	émotionnalité faible ③
④ ancrage actionnel et situationnel	détachement actionnel et situationnel ④
⑤ ancrage référentiel dans la situation	détachement référentiel de la situation ⑤
⑥ coprésence spatio-temporelle	séparation spatio-temporelle ⑥
⑦ coopération communicative intense	coopération communicative minime ⑦
⑧ dialogue	monologue ⑧
⑨ communication spontanée	communication préparée ⑨
⑩ liberté thématique	fixation thématique ⑩
etc.	etc.

FIGURE II.9 – Valeurs paramétriques du langage parlé (à gauche) et du langage écrit (à droite) selon Koch et Oesterreicher (2001)

grande distance communicative. Les styles se situant entre ces deux extrêmes sur le continuum communicatif (comme l'entretien professionnel) emprunteraient à la fois des caractéristiques de l'immédiat communicatif et de la distance communicative, résultant en un “relief conceptionnel” tel que celui donné en exemple en figure II.10.

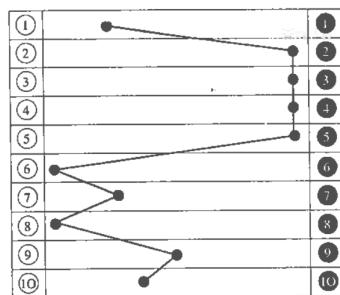


FIGURE II.10 – Relief conceptionnel de l'entretien professionnel par Koch et Oesterreicher (2001)

Ces reliefs conceptionnels empruntent des éléments relevant du nombre d'interlocuteurs (communication privée/publique), du degré d'intimité, de l'état émotionnel, du rôle du locuteur, de la situation, de la présence physique, et du thème, nombre d'éléments que nous avons déjà vus.

3 Synthèse

Ce chapitre de l'état de l'art sur la parole spontanée fait ressortir que la parole spontanée ne peut pas s'opposer *stricto sensu* aux autres types de paroles qu'ils soit préparé ou lu. En effet, la locution “parole spontanée” regroupe un ensemble de

3 Synthèse

sous-types de parole : quand on parle entre amis, quand on demande notre chemin dans la rue, quand on répond à une interview... La parole spontanée n'est donc pas uniforme mais représentative de la richesse des situations interactionnelles. Elle est un processus linéaire, dont les énoncés sont créés sur le fil de leur énonciation, sur simple impulsion du locuteur lui-même et contient ainsi des "erreurs" et corrections visibles, et dont l'émetteur est le premier auditeur.

De nombreux facteurs sont à l'origine de différents sous-types de parole spontanée. Ainsi, le rôle, le nombre d'interlocuteurs, le degré d'intimité entre les locuteurs, le lieu de l'interaction, la présence physique des locuteurs, le canal de communication, la présence d'un auditoire, les émotions et la proximité des locuteurs avec le thème de la discussion sont tout autant de paramètres qu'il faudrait prendre en compte afin de pouvoir déterminer des niveaux de spontanéité. En effet, certaines conditions semblent favoriser une parole plus spontanée, à savoir :

- l'absence de rapport de rôle entre les locuteurs
- la présence de peu d'interlocuteurs
- le lien amical ou familial
- la présence physique des locuteurs
- la proximité du locuteur avec le thème de la discussion.

Ces facteurs ne se retrouvent jamais tous dans les modèles relatifs à la variation stylistique de la parole spontanée que nous avons présentés. Certains y sont regroupés, tels que la relation entre les locuteurs et la situation (interview, conversation) qui participent à la dimension de la formalité ou encore l'intimité, l'émotionnalité et la coprésence spacio-temporelle notamment rassemblés dans le relief conceptionnel. D'autres font leur apparition comme le degré d'extraversion, l'attention ou encore le contrôle. Nous finissons donc cet état de l'art sur la parole spontanée avec une idée plus précise de ce qu'est la parole spontanée et de ce qui la fait varier. Néanmoins, un travail de clarification des dimensions à intégrer à un modèle de représentation de la variation stylistique reste à effectuer pour faciliter et favoriser l'analyse de l'impact de la parole spontanée en reconnaissance automatique de la parole.

Chapitre III

Questions de recherche

Les systèmes de reconnaissance automatique de la parole donnent aujourd’hui des performances satisfaisantes sur la lecture. Nous avons effectivement observé dans l’état de l’art des WER régulièrement inférieurs à 10% sur ce type de parole. L’évaluation des performances sur les corpus comprenant à la fois de la parole préparée et de la parole spontanée montre au contraire des WER d’une plus grande disparité, avec des scores allant de 11% à 25%. De plus, nous avons pu voir que lorsqu’un ensemble de test est représentatif d’une parole très spontanée, les systèmes sont souvent mis en grande difficulté et peuvent présenter des WER allant jusqu’à 80%, ce qui ne permet pas l’exploitation de la transcription automatique.

Quelques études font effectivement un lien entre l’augmentation du WER et l’augmentation de la spontanéité. Les raisons de ces difficultés sont recherchées dans les caractéristiques de cette parole telles que les prononciations déviantes, les réductions, le manque de marque intonative, les hésitations et autres disfluences qui seraient plus présentes lorsque la spontanéité est élevée. Mais ces caractéristiques sont nombreuses et montrent souvent une très grande variabilité.

La parole spontanée est définie par (Luzzati, 2007, p. 27) comme “l’ensemble des énoncés oraux conçus et perçus dans le fil de leur énonciation.”. Si cette définition présente la parole spontanée comme un ensemble, la littérature aura pu nous montrer que de très nombreux facteurs jouent sur la spontanéité, qu’ils soient interpersonnels (rôle, nombre d’interlocuteurs, degré d’intimité), environnementaux (lieu, présence physique) ou individuels (émotions, proximité du locuteur avec le thème de la discussion). De ces facteurs dépendent donc de nombreux sous-types de parole spontanée, comme la conversation entre amis, la conversation entre collègues sur et en-dehors du lieu de travail, la conversation téléphonique *etc.* Ainsi, la “parole spontanée” ne désignerait pas un type de parole uniforme, mais bel et bien

une locution générique pour désigner le processus de production de cette parole en lui-même.

Des études ont proposé de placer ces différents sous-types de parole spontanée sur un continuum de styles de parole grâce à des modèles multidimensionnels. Les modèles présentés incluent certains des facteurs relevés précédemment, parfois en les rassemblant sur une dimension telle que le niveau de formalité, parfois en les dissociant. De nouvelles dimensions peuvent également avoir été ajoutées comme le côté plus ou moins extraverti des locuteurs ou encore le niveau d'attention ou de contrôle du locuteur sur sa parole. Le modèle rassemblant le plus grand nombre de facteurs précédemment cités permet de déterminer ce qu'ils caractérisent comme le "relief conceptionnel" d'une prise de parole, c'est-à-dire de déterminer en quels points la prise de parole se rapproche plutôt du langage parlé ou du langage écrit. Si ce modèle place notamment sur le même plan l'évaluation de la spontanéité, l'intimité entre les locuteurs ou encore l'aspect privé ou public de la prise de parole, la littérature montre que l'évaluation de la spontanéité est, de fait, capturée par l'évaluation de l'intimité, de l'aspect privé ou public et autres facteurs relevés précédemment. Ainsi, l'absence de rapport de rôle entre les locuteurs, le faible nombre d'interlocuteurs, le fort niveau d'intimité partagé entre les interlocuteurs, leur présence physique, et la proximité du locuteur avec le thème de la discussion semblent représenter les conditions les plus propices à la production d'une parole très spontanée.

Notre objectif dans cette thèse est de pouvoir catégoriser de la parole comme plus ou moins spontanée et d'étudier le comportement de systèmes de reconnaissance automatique de la parole au travers de ces données. Les modèles issus d'études en linguistique relevés dans l'état de l'art présentent des limites pour une utilisation directe dans nos travaux de recherche, de part :

- la non prise en compte de certains facteurs influant sur la spontanéité et pourtant souvent renseignés dans les métadonnées tel que le nombre d'interlocuteurs
- le mélange de plusieurs facteurs sur une même dimension ne permettant pas de dissocier certains sous-types de parole
- la prise en compte de facteurs pour lesquels aucune métadonnée n'est disponible dans les corpus (tel que le degré d'extraversion)
- la dissociation du lien entre spontanéité et certains facteurs comme l'intimité, la présence ou encore l'aspect privé ou public de la prise de la parole.

Ainsi, notre première question de recherche concerne la **détermination d'un nombre réduit de facteurs, nous permettant à la fois :**

- **de constituer des sous-groupes de données langagières supposées similaires ou du moins homogènes,**

-
- de positionner des sous-groupes de données langagières les uns par rapport aux autres, en fonction de leur niveau de spontanéité et notamment des facteurs impactant cette spontanéité.

La littérature dans le domaine de la reconnaissance automatique de la parole montre que la spontanéité est l'un des facteurs pouvant faire varier les performances des systèmes. Nous étudierons ainsi les performances de systèmes de reconnaissance automatique de la parole sur différents sous-groupes de données langagières que nous supposons comme plus ou moins spontanés.

Notre deuxième question de recherche repose sur la bonne capture de niveaux de spontanéité au travers des facteurs pris en compte. Les données de parole spontanée étant en quantité limitée par rapport aux corpus de lecture ou de parole préparée et les modèles pré-appris actuels permettant de tirer bénéfice de la faible taille de corpus, nous nous demandons **si l'adaptation spécifique d'un modèle pré-appris avec une faible quantité de données représentatives de différents niveaux de spontanéité pourrait améliorer les performances des systèmes**. En effet, au même titre qu'un système appris sur de la parole préparée donnera de moins bonnes performances sur de la parole spontanée, peut-être que l'apprentissage d'un système avec des données représentatives d'un niveau de spontanéité spécifique permettra d'améliorer les performances sur ce même sous-type de parole spontané.

Enfin, étant donné que l'étude de Baevski *et al.* (2020) rapporte de meilleurs résultats lorsque l'adaptation est effectuée sur une grande quantité de données, nous étudierons au travers de notre troisième question de recherche **l'impact d'une grande quantité de données moins contrôlées sur l'adaptation d'un modèle pré-appris dans le cadre d'une tâche de reconnaissance automatique de la parole**. Ceci se fera en deux étapes : une adaptation au domaine spontané, suivie d'une adaptation spécifique similaire à celle présentée précédemment.

L'étude de ces questions implique donc premièrement de définir un modèle permettant de capturer des niveaux de spontanéité, qui nous servira ensuite à étiqueter une partie des données de parole spontanée que nous aurons pu rassembler. La définition de ce modèle se fera grâce à une mise en parallèle des facteurs relevés dans l'état de l'art et de l'analyse des corpus que nous utiliserons dans cette étude. Cette thèse débutant simultanément avec le projet LeBenchmark dont l'objectif est de fournir des modèles pré-appris pour le français à la communauté, nous aurons à notre disposition des modèles monolingues pour nous confronter à ce défi de la transcription automatique de la parole spontanée. Nous développons la méthodologie adoptée pour répondre à ces questions de recherche dans la partie II de ce manuscrit.

Deuxième partie

Méthodologie

Chapitre IV

Cadre méthodologique

Ce chapitre présente les méthodes que nous avons adoptées afin de répondre à nos questions de recherche. Nous présentons tout d'abord le processus de sélection de données de parole spontanée, ainsi que la façon dont nous nous servons de ces données pour, à la fois nous aider à déterminer des dimensions de variation de la parole spontanée avec peu de facteurs, mais aussi étiqueter des ensembles de données pour l'étude de la variation de la spontanéité en reconnaissance automatique de la parole.

Ensuite, nous présentons le projet LeBenchmark dont le but est de fournir à la communauté des modèles pré-appris de type Wav2Vec 2.0 sur le français, ainsi que les systèmes état de l'art que nous utiliserons comme systèmes de référence pour la suite de nos expérimentations.

Enfin, nous présentons les expérimentations utilisant l'un des modèles LeBenchmark, imaginées pour étudier l'amélioration des performances des systèmes de reconnaissance automatique de la parole sur la parole spontanée en fonction de différentes quantités de données.

Déroulement du chapitre

1 Les données de parole spontanée	80
1.1 Le processus de sélection des corpus de parole spontanée . . .	81
1.2 Les dimensions de variation de la parole spontanée	82

1.3	L'étiquetage des corpus et ensembles de données pour les expérimentations	83
2	Les systèmes de reconnaissance automatique de la parole de référence	84
2.1	Les modèles pré-appris sur le français : le projet LeBenchmark	84
2.2	Deux systèmes état de l'art comme systèmes de référence	86
2.2.1	Le système Speechbrain	86
2.2.2	Le système Whisper en <i>zero-shot</i>	89
2.3	Métriques d'évaluation	89
3	Plans d'expérimentations	89
3.1	Mesure des performances des systèmes état de l'art selon les sous-types de parole spontanée	90
3.2	Les adaptations aux différents sous-types de parole spontanée	91
3.3	L'utilisation des données de parole spontanée non catégorisées	91
4	Synthèse	92

1 Les données de parole spontanée

L'état de l'art sur la parole spontanée nous aura montré que ce type de parole est complexe. Il existe différents niveaux de spontanéité qui dépendent à la fois de facteurs sociaux (rôle, nombre d'interlocuteurs, degré d'intimité), de facteurs environnementaux (lieu, présence physique ou non des locuteurs, canal de communication, présence d'un auditoire) mais aussi d'autres facteurs comme les émotions ou la proximité des locuteurs avec le thème de la discussion. Les corpus de parole spontanée de référence pour la reconnaissance automatique de la parole ne sont pas, à eux seuls, représentatifs de cette diversité. De plus, ces corpus se restreignent majoritairement au contexte radiophonique ou télévisuel. Du côté des corpus multilingues, bien que pour la plupart très récents, la parole spontanée n'est pas représentée. Dans cette section nous présentons dans un premier temps la méthode mise en place afin de réunir davantage de données représentatives de différents types de parole spontanée. Dans un second temps, nous discutons de la méthode mise en place afin de déterminer des dimensions de variation de la parole spontanée permettant de capturer différents niveaux de spontanéité. Enfin, nous présentons la méthode d'étiquetage des données de parole spontanée en fonction des dimensions de la parole spontanée déterminées que nous avons suivie.

1.1 Le processus de sélection des corpus de parole spontanée

La récolte de données représentatives de différents sous-types de parole spontanée a un objectif double. Les corpus serviront tout d'abord d'appui à l'élaboration de dimensions de variations de la parole spontanée. Ils serviront ensuite à créer des ensembles de données pour les expérimentations en reconnaissance automatique de la parole spontanée. De ces objectifs découlent des contraintes fortes sur le processus de sélection.

Les deux premières contraintes sont liées au domaine dans lequel nous effectuons ce travail de recherche : la reconnaissance automatique de la parole. Ainsi, les données récoltées ne devront pas être trop bruitées (environnement bruyant) ou de trop mauvaise qualité (enregistrement dégradé avec le temps, faible fréquence d'échantillonnage...), car s'ajoutent alors des difficultés de transcription supplémentaires. Il sera également nécessaire de veiller à récolter de préférence des enregistrements comprenant un nombre restreint locuteurs, la parole superposée étant encore aujourd'hui mal traitée par les systèmes. Des signaux trop bruités ou comprenant une trop grande proportion de parole superposée viendraient de fait dégrader les performances de reconnaissance automatique de la parole (cette dégradation ne serait ainsi pas du seul fait de la spontanéité).

La troisième contrainte est liée à la licence attribuée au corpus. En effet, nous adhérons au principe FAIR¹ en faveur d'une recherche reproductible. La première étape de la reproductibilité des travaux se faisant notamment grâce au libre accès aux données utilisées au cours des travaux de recherche, le choix de l'utilisation de données avec une licence ouverte est important. Il existe un grand nombre de licences plus ou moins ouvertes, la plus connue étant la licence Creative Commons à laquelle peut être associé un ou plusieurs des attributs suivants² :

- BY : l'auteur doit être crédité
- SA : les adaptations doivent être partagées dans les mêmes conditions
- NC : seul un usage non commercial est autorisé
- ND : aucun dérivé et aucune adaptation ne sont permis

Certaines données peuvent également être estampillées d'une licence CC-0 lorsqu'elles rentrent dans le domaine public (ce qui peut être un choix délibéré, ou se faire automatiquement 70 ans après la mort d'un auteur).

La quatrième contrainte est induite par ce qui est au cœur de ce projet de recherche : étudier la parole spontanée dans sa variabilité. Ainsi, les données de

1. pour *Findable, Accessible, Interoperable and Reusable* (trouvable, accessible, interopérable, réutilisable) Wilkinson *et al.* (2016)

2. Pour plus d'explications : <https://creativecommons.org/share-your-work/cclICENSES/>

télévision et de radio étant déjà largement représentées par les corpus ETAPE et REPERE, nous choisissons de nous concentrer sur la récolte de corpus représentant une diversité de situations propices à l'apparition de différentes formes de parole spontanée. De plus, ces données devront être accompagnées de *data papers* (publications relatives aux données) et/ou de métadonnées complètes afin de pouvoir qualifier nos données selon plusieurs facteurs dont nous savons qu'ils influencent les niveaux de spontanéité de la parole.

1.2 Les dimensions de variation de la parole spontanée

Les performances des systèmes de reconnaissance automatique de la parole sont sensibles aux différents niveaux de spontanéité : plus la spontanéité augmente, plus le WER augmente. Néanmoins, les facteurs de variation de la spontanéité sont nombreux et concernent aussi bien l'état émotionnel du locuteur, son état physique, l'environnement dans lequel la production de parole a lieu, le lien entre les personnes qui échangent... Le nombre de facteurs et les différents niveaux d'analyse qu'ils représentent complexifient la tâche d'analyse des performances sur le type de parole qu'est la parole spontanée. L'objectif ici est de répondre à la première question de recherche : quelles dimensions de variation de la parole spontanée peut-on prendre en compte afin de pouvoir affiner l'analyse des performances des systèmes sur ce type de parole ?

Afin de répondre à cette question, nous mettrons en parallèle la liste de facteurs et de dimensions liés à la parole spontanée rapportés dans le chapitre 2, avec les facteurs de variation relevés dans les publications et fichiers de métadonnées relatives aux corpus de parole spontanée que nous aurons sélectionnés. Il est important de croiser les deux car certaines informations, comme les émotions traversées par les locuteurs pendant leur prise de parole, ne sont que très peu renseignées par les créateurs des corpus. Or, nous souhaitons que la future catégorisation de données selon des dimensions de variation de la parole spontanée puisse se faire sur un maximum de données. Certains facteurs ou certaines dimensions devront donc soit être laissés de côté, soit retravaillés.

Nous souhaitons également que cette catégorisation puisse se faire aussi facilement et rapidement que possible. Cela implique donc de limiter le nombre de facteurs étudiés et donc de dimensions, mais aussi de considérer un haut niveau d'annotation. La catégorisation doit pouvoir se faire simplement à partir des *data papers* lorsque ceux-ci sont bien renseignés. Les fichiers de métadonnées ne devraient servir qu'à lever le doute sur le lien entre deux personnes par exemple. De plus, nous souhaitons que la catégorisation puisse se faire à l'échelle d'un corpus ou des enregistrements qui le composent. En effet, la catégorisation de segments de parole est une tâche chronophage.

Enfin, il nous semble important de réussir à capturer la variabilité individuelle via nos dimensions, mais de ne pas nous reposer dessus pour déterminer ces dimensions. Ainsi, aucune de nos dimensions ne devra reposer sur des éléments comme le nombre de disfluences, le débit de parole ou autre caractéristique pouvant varier d'un locuteur à l'autre. De plus, nous avons vu dans le chapitre 2 que pour une même situation, le type de parole spontanée produit pourra sembler différent en fonction du rôle social (selon si on interviewe un animateur de radio ou une personne *lambda* par exemple). Nos dimensions devront donc refléter différents types de productions induits par une source autre que l'individu en lui-même, c'est-à-dire plutôt par des normes sociales, l'intimité entre les locuteurs, la communication publique ou privée. En effet, les facteurs individuels tels que l'état émotionnel du locuteur, ses traits de personnalité plus ou moins extraverti, sa proximité avec le ou les thèmes abordés au cours de l'enregistrement sont le plus souvent très difficiles à mesurer et non renseignés dans les métadonnées, excepté lorsque le corpus a été construit volontairement sur ce principe.

1.3 L'étiquetage des corpus et ensembles de données pour les expérimentations

Les dimensions de variation de la parole spontanée déterminées permettront d'étiqueter les corpus rassemblés comme représentatifs de sous-types de parole avec des niveaux de spontanéité supposés différents. Ainsi, il deviendra possible, grâce à un référentiel commun, de regrouper les données similaires et de comparer ces ensembles entre eux. Nous formerons des ensembles de données d'apprentissage, de validation et de test destinés à la fois à l'évaluation des systèmes, mais aussi à l'amélioration de leurs performances sur ce type de parole. Certains des ensembles pouvant être identifiés formeront nos cas d'étude, pour l'étude de la parole en fonction de différents niveaux de spontanéité. Les différents cas ne seront plus forcément uniquement constitués des enregistrements d'un seul corpus mais pourront regrouper les enregistrements de plusieurs corpus étant donné que les critères de regroupement seront les dimensions précédemment déterminées. De plus, cela pourra permettre d'éviter d'évaluer ou d'entraîner les systèmes sur un seul et même type de données et ainsi de rendre les systèmes trop liés à un corpus en particulier.

L'étude de Baevski *et al.* (2020) confirme que plus l'ensemble de données utilisé pour l'adaptation d'un modèle est grand, meilleures sont les performances du système de reconnaissance automatique de la parole. Les auteurs montrent également que l'utilisation de moins de 10 heures de parole pour l'adaptation d'un modèle pré-appris utilisé pour une tâche de reconnaissance automatique de la parole lorsqu'aucun modèle de langue n'est utilisé dégrade fortement les résultats : +36,4 points de WER sur LibriSpeech *clean*, par rapport aux performances obtenues par

Lüscher *et al.* (2019)³, avec 10 minutes parole et +13 points avec 1 heure, contre +2,5 points de WER seulement, avec 10 heures de parole. Nous basant sur cette étude, nous nous sommes fixés comme objectif de réunir pour chaque cas d'étude au moins dix heures de parole dédiées à l'apprentissage.

Nous veillerons à ne pas inclure de données de locuteurs non natifs afin de ne pas fausser l'évaluation de l'apport de nos dimensions sur la reconnaissance automatique de la parole spontanée. De plus, nous retirerons les segments contenant de la parole superposée lorsque cela sera possible, c'est-à-dire lorsque l'information sera disponible dans les fichiers de transcription. Enfin, nous choisissons de n'intégrer à ces ensembles de sous-types de parole spontanée que des enregistrements avec deux locuteurs maximum afin d'éviter les situations de schismes interactionnels, c'est-à-dire les multiples conversations en parallèles qui apparaissent parfois lorsqu'il y a plusieurs locuteurs et qui pourraient ajouter une couche de difficulté pour les systèmes de reconnaissance automatique de la parole.

2 Les systèmes de reconnaissance automatique de la parole de référence

Nous présentons dans cette section le projet LeBenchmark, ainsi que les deux systèmes de reconnaissance automatique de la parole donnant des résultats état de l'art qui nous serviront de systèmes de référence tout au long de ce travail de thèse.

2.1 Les modèles pré-appris sur le français : le projet Le-Benchmark

La reconnaissance automatique de la parole a largement pu profiter de l'apprentissage et la distribution de *foundations models* ces dernières années. C'est ainsi que le projet LeBenchmark a vu le jour au début de ce travail de doctorat (Evain *et al.*, 2021b; Parcollet *et al.*, 2024) et que nous avons eu l'occasion de participer à cette entreprise collective. Le but du projet était de fournir à la communauté des modèles pré-appris de façon auto-supervisée sur des données en français. Ceci n'avait jusqu'alors été fait que dans le cadre de l'apprentissage de modèles multilingues.

L'entraînement, l'évaluation et la mise à disposition de tels modèles représentent une grande charge de travail et nécessitent des compétences diverses. Ce projet a donc demandé l'alliance de nombreux chercheurs et chercheuses issus de différents laboratoires de recherche publics et privés : le Laboratoire d'Informatique

3. Performances reportées en tableau I.7

de Grenoble (LIG), le Laboratoire Informatique d’Avignon (LIA), le Laboratoire d’Analyse et de Modélisation de Systèmes pour l’Aide à la Décision (LAMSADE), Naver Labs Europe et Samsung.

D’un point de vue plus pratique, l’entraînement de ce type de modèle demande de nombreuses ressources computationnelles, indisponibles à l’échelle individuelle des équipes de recherche travaillant sur la parole. L’utilisation du supercalculateur Jean-zay⁴ s’est donc avérée indispensable.

Le comité de pilotage du projet a fait le choix de l’architecture Wav2Vec 2.0 pour ces modèles, telle que présentée en section 2.6.4 du chapitre 1. Différents types de modèles ont ainsi été appris : *light*, *base*, *large*, et *xlarge*. Les architectures *base* et *large* sont celles présentées dans Baevski *et al.* (2020). Les deux autres ont été ajoutées afin de tester les modèles dans deux contextes différents : pour une utilisation au sein d’un dispositif avec ressources limitées (*light*) et dans des environnements non contraignants et centralisés (*xlarge*). Le tableau IV.1 présente le nombre de paramètres de chacun des modèles, ainsi que la taille des vecteurs de sortie.

Modèle	#param.	Dimension sortie
<i>light</i>	26M	512
<i>base</i>	90M	768
<i>large</i>	330M	1024
<i>xlarge</i>	965M	1280

TABLE IV.1 – Différentes types d’architectures Wav2Vec 2.0 utilisées pour l’entraînement des modèles LeBenchmark

L’apprentissage de *foundation models* requiert une grande quantité de données diverses, afin de créer des modèles “génériques”. Le projet a été divisé en trois lots :

- récolte et préparation des données
- entraînement des modèles
- évaluation des modèles.

C’est sur le premier lot que ma contribution a été la plus importante. La méthode de recherche de corpus de parole présente des similitudes avec celle adoptée pour la

4. La supercalculateur Jean-Zay a été acquis par le ministère de l’Enseignement supérieur, de la Recherche et de l’Innovation par l’intermédiaire de la société GENCI (Grand équipement national de calcul intensif). Il a été installé à l’IDRIS, centre national de calcul du CNRS en 2019.

sélection de corpus de parole spontanée, la contrainte du type de parole spontanée en moins. Ainsi, les corpus doivent :

- respecter au mieux l'ensemble des principes FAIR, notamment le principe d'accès facile aux données
- être libres d'accès pour la recherche
- être de bonne qualité audio
- être de grande taille afin de réduire le temps de préparation des données
- représenter au mieux la variété des types de parole (lu, préparée, spontanée, radiophonique/télévisuelle, professionnelle, émotionnelle).

Concernant les métadonnées, l'information sur le genre des locuteurs est un plus, mais non une contrainte forte.

Les modèles appris sont ensuite utilisés pour l'évaluation de six tâches en aval, à savoir : la reconnaissance automatique de la parole, la compréhension de la parole, la traduction automatique de la parole vers le texte, la reconnaissance automatique d'émotions, l'analyse syntaxique et la vérification automatique du locuteur. Chaque tâche a sa propre architecture neuronale. Les modèles LeBenchmark ainsi testés ont été appris soit indépendamment de la tâche d'évaluation visée (utilisation des modèles comme simples extracteurs de paramètres), soit spécifiquement pour cette tâche (ajout d'une phase d'apprentissage sur les données du domaine, avant ou sans les labels correspondants). Ces modèles ont servi au développement de différents systèmes de reconnaissance automatique de la parole spontanée utilisés dans ce travail de thèse.

2.2 Deux systèmes état de l'art comme systèmes de référence

Deux systèmes sont utilisés afin de fournir des résultats de référence qui nous serviront tout au long de cette étude. Nous présentons ainsi dans cette section les systèmes Speechbrain et Whisper utilisés et précisons les corpus de référence sur lesquels ils seront évalués.

2.2.1 Le système Speechbrain

Les modèles LeBenchmark étant les premiers modèles pré-appris monolingues disponibles pour le français et ce type de modèle donnant la plupart du temps de meilleures performances que l'utilisation de modèles multilingues, nous souhaitions qu'un de nos systèmes de référence se base sur l'un de ces modèles fournis à la

communauté. Ainsi, le premier système utilisé est appris grâce à une recette constituée et partagée par l'équipe Speechbrain. Speechbrain⁵ (Parcollet *et al.*, 2022) est une boîte à outils qui a pour but de faciliter le développement, la portabilité et l'utilisation de différentes technologies de traitement de la parole (reconnaissance de la parole, compréhension de la parole, des émotions etc) et qui est de plus en plus utilisée par la communauté. L'un des auteurs du projet LeBenchmark faisant partie de l'équipe de développement de Speechbrain, les modèles LeBenchmark sont intégrés dans le projet Speechbrain via HuggingFace, et notamment dans certaines des “recettes” proposées pour développer des systèmes de reconnaissance automatique de la parole.

Nous nous sommes intéressés aux recettes CommonVoice, les seules disponibles au début de ce travail de thèse pour la transcription automatique du français. Bien que ces recettes soient optimisées pour la reconnaissance de données issues de Common Voice, elles nous permettent de nous assurer de la reproductibilité de nos travaux. La sélection du type d'architecture à utiliser parmi les quatre proposées s'est faite selon (i) l'intégration déjà présente d'un modèle de type Wav2Vec 2.0 dans l'architecture et (ii) la performance du système sur un corpus de référence. Sur les deux recettes disponibles intégrant des modèles Wav2Vec 2.0, le WER sur l'ensemble de test de la version 6.1 de CommonVoice est de 9,96% pour la recette CTC, contre 13,34% pour la recette seq2seq. Ainsi, nous avons sélectionné la recette CTC proposant l'architecture suivante : un modèle pré-appris issu du projet LeBenchmark, suivi de trois couches DNN⁶ et d'une couche de sortie dont la dimension dépend du nombre de caractères dans les données d'apprentissage. La fonction de coût utilisée est la CTC, telle que présentée en section 2.3 du chapitre 1.

Nous présentons en chapitre VII des résultats avec un apprentissage du système sur les données d'apprentissage de Common Voice 6.1 et une optimisation sur l'ensemble de développement Common Voice 6.1. Ce système est évalué sur les ensembles de test des corpus de référence Common Voice et ETAPE. Quelques-uns des hyperparamètres utilisés pour l'apprentissage de ce système sont présentés en tableau IV.2 et expliqués à la suite.

Le filtre haut correspond à un filtrage des données. Les systèmes traitant difficilement les segments de plus de 30 secondes, nous avons fixé cette limite en conséquence. Cet hyperparamètre était initialement fixé à 10 secondes, durée d'enregistrement qui correspond, pour Common Voice, à un oubli de coupure du microphone après la lecture. Le *learning rate* attribué à l'adaptation du modèle pré-appris est très faible afin d'éviter une situation d'oubli catastrophique, c'est-à-dire l'oubli des informations déjà apprises par le modèle lors du pré-apprentissage. L'option

5. <https://speechbrain.github.io/>

6. DNN, tel qu'utilisé dans Speechbrain, se rapporte à *Dense Neural Network*

Filtre haut	30 secondes
Nombre d'époques	30
Learning rate (LR)	1.0
Learning rate Wav2Vec	0.0001
Taille batch	2
Taille batch test	4
Type de token	char
Freeze Wav2Vec	False
Nombre de neurones de sortie	76
LR modèle annealing facteur	0.8
LR modèle improvement threshold	0.0025
LR Wav2Vec annealing facteur	0.9
LR Wav2Vec improvement threshold	0.0025

TABLE IV.2 – Hyperparamètres du système état de l’art à base de CTC proposé par Speechbrain

Freeze Wav2Vec permet choisir si l’on veut figer le modèle pré-appris, et ainsi l’utiliser comme simple extracteur de paramètres. Sa configuration comme *False* nous permet donc de signifier que l’on souhaite une adaptation du modèle pré-appris. Le nombre de neurones en sortie correspond au nombre de caractères que le système pourra générer. Celui-ci est calculé sur l’ensemble de données d’apprentissage. Certains caractères surprenant tels que Ñ ou encore Ø, qui ne sont pas utilisés en français, sont présents ici et issus de la transcription de noms propres dans Common Voice⁷. Enfin, la méthode d’*annealing* change la valeur du *learning rate* selon la condition suivante :

Si

$$(past_validLoss - current_validLoss) / past_validLoss < improvement_threshold$$

alors

$$lr = lr * annealing_factor$$

Cela aide l’algorithme à converger et ainsi assure ainsi sa stabilité. Cette technique est appelée le *learning rate decay* (Nakamura *et al.*, 2021).

7. exemple : “UN SILENCE DE QUELQUES SECONDES SUIVIT LE RÉCIT DE LA SEÑORA MENDEZ”

3 Plans d’expérimentations

2.2.2 Le système Whisper en *zero-shot*

Le deuxième système état de l’art utilisé est le système Whisper, présenté en section 2.6.3 du chapitre 1. Nous testons ses performances en *zero-shot*, c’est-à-dire sans adaptation du modèle à des données spécifiques, tel que recommandé par les auteurs (Radford *et al.*, 2022). Nous avons créé une pipeline afin de pouvoir utiliser les mêmes fichiers d’entrées que ceux utilisés pour Speechbrain. En effet, étant donné que nous souhaitons retirer le plus de parole superposée possible des ensembles de test, il nous était impossible de donner simplement les enregistrements entiers à décoder au système. De même, nous avons mis en place un script qui permet de faire le lien entre les sorties Whisper et le calcul du WER proposé par Speechbrain. Nous assurons ainsi une certaine comparabilité entre nos systèmes état de l’art, bien qu’il soit important de rappeler que la quantité de données d’apprentissage est largement plus conséquente dans le cas du système Whisper et que le détail de ces données n’est pas connu.

Le modèle Whisper utilisé est le large-v2, celui-ci étant le modèle donnant la plupart du temps les meilleurs résultats sur la transcription automatique du français (voir Radford *et al.* (2022)). Ce système est évalué sur les corpus de test de Common Voice 6.1 et d’ETAPE, tout comme le système Speechbrain.

2.3 Métriques d’évaluation

La mesure de performance de l’ensemble de nos systèmes est le *Word Error Rate* (WER). Nous n’avons pas souhaité pondérer les résultats en fonction de catégories syntaxiques ou type d’erreurs car nos systèmes sont uniquement voués à l’analyse des performances sur de la parole spontanée en fonction de divers sous-types de parole spontanée. Le WER est très largement utilisé dans le domaine, et si cette mesure a des limites comme nous l’avons montré en section 3.1 du chapitre 1, elle a le mérite de rendre nos résultats comparables à ceux de travaux précédents. Nous calculons le WER par ensemble de données, mais aussi le WER par locuteur et le WER par enregistrement afin d’étudier la variabilité dans nos données de test. Nous souhaitons ainsi vérifier si les systèmes sont effectivement sensibles aux sous-types de parole spontanée relevés ou s’ils capturent la façon de parler d’un locuteur en particulier ou les conditions acoustiques particulières d’un enregistrement.

3 Plans d’expérimentations

L’apprentissage et la mise à disposition des modèles LeBenchmark nous aura permis de considérer des approches entièrement neuronales pour notre étude sur la parole spontanée. Ainsi, si l’usage de systèmes état de l’art nous permettra d’étu-

dier les performances des systèmes sur les sous-types de parole spontanée que nous aurons identifiés, l'apprentissage de nouveaux systèmes adaptés de différentes façons à la parole spontanée nous permettra de mesurer l'impact à l'apprentissage de données représentatives des ces sous-types de parole spontanée sur les performances. Les systèmes présentés sont tous utilisés sans modèle de langue externe.

3.1 Mesure des performances des systèmes état de l'art selon les sous-types de parole spontanée

Tout d'abord, nous souhaitons tester la sensibilité des performances de systèmes de reconnaissance automatique de la parole état de l'art à la variation des dimensions de la parole spontanée. Pour cela, nous utilisons les deux systèmes état de l'art décrits précédemment : le système Speechbrain, et le système Whisper, en *zero-shot*. Les architectures des systèmes état de l'art utilisés étant différentes, elles nous permettent de vérifier que les différences de performance sont observées indépendamment de l'architecture neuronale utilisée. Logiquement, nous nous attendons à trouver des performances dégradées lorsque les sous-types sélectionnés correspondent à des dimensions favorisant une plus grande spontanéité.

En plus d'être testés sur Common Voice et ETAPE, deux corpus de référence, les performances des systèmes sont également calculées sur des données pour lesquelles la catégorisation précise en sous-types de parole spontanée n'a pas été possible. Cet ensemble *All_spont* est séparé en 3 parties, *train*, *dev* et *test*. Enfin, nous testons les systèmes sur des ensembles représentatifs de différents sous-types de parole spontané, capturés dans nos différents cas d'études.

Le système Speechbrain, bien que disponible sur HuggingFace⁸, est réentraîné. En effet, la recette étant réutilisée pour les expérimentations suivantes, nous souhaitons nous assurer que l'entraînement en lui-même ne présente aucun problème sur le serveur utilisé, à savoir Jean-Zay. Nous testons à la fois une optimisation sur les données Common Voice et sur les données représentatives de sous-types de parole spontanée afin de voir si cela a une incidence sur les performances, et assurer la comparabilité avec les systèmes suivants.

Les systèmes et les données utilisés sont récapitulés en figure IV.1.

8. <https://huggingface.co/speechbrain/asr-wav2vec2-commonvoice-fr>

3 Plans d'expérimentations

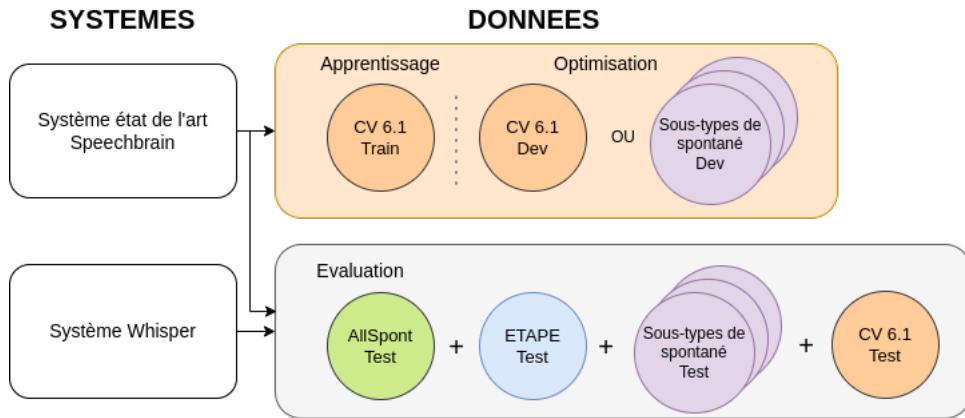


FIGURE IV.1 – Détail des données d'apprentissage, d'optimisation et de test utilisées pour l'apprentissage et/ou l'évaluation des systèmes état de l'art Speechbrain et Whisper.

3.2 Les adaptations aux différents sous-types de parole spontanée

Nous nous intéressons ensuite à l'impact que pourraient avoir des données représentatives de différents niveaux de spontanéité sur l'apprentissage et l'adaptation d'un système. Nous cherchons ainsi à améliorer les performances sur les différents sous-types de parole spontanée que nous étudions. Pour répondre à cette question, nous testons différentes adaptations spécifiques du modèle 7k-large de LeBenchmark. Le même modèle pré-appris et la même architecture que le système état de l'art Speechbrain sont utilisés et les adaptations spécifiques se font avec de petites quantités de données, correspondant aux différents sous-types de parole spontanée capturés dans les cas d'étude. Ces différents systèmes sont évalués sur les mêmes corpus que précédemment, à savoir Common Voice, ETAPE, *All_spont* et les ensembles de test de nos cas d'étude. Les systèmes et données sont récapitulés en figure IV.2.

3.3 L'utilisation des données de parole spontanée non catégorisées

Après l'apprentissage et l'adaptation avec de faibles quantités de parole spontanée, nous souhaitons étudier l'impact de l'utilisation d'une plus grande quantité de données. C'est pourquoi nous utilisons cette fois-ci comme ensemble d'apprentissage, les données de parole spontanée "tout-venant" rassemblées dans *All_spont train*. Cette adaptation plus large nous permettra d'obtenir un système adapté au

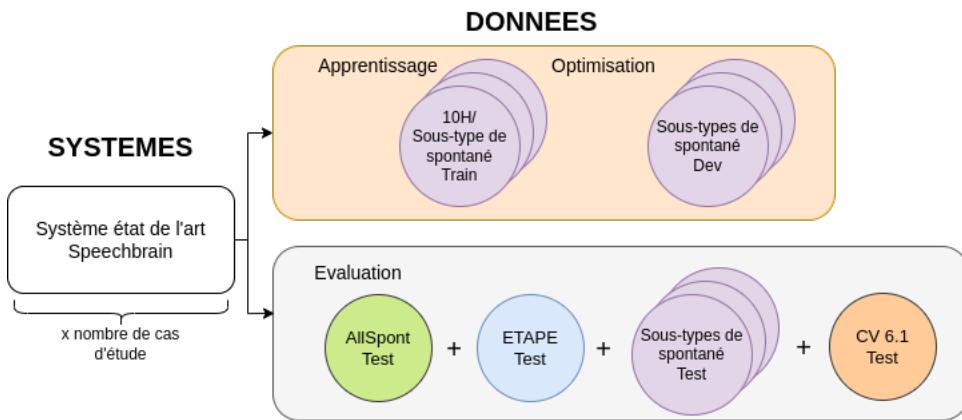


FIGURE IV.2 – Détail des données d'apprentissage, d'optimisation et de test utilisées pour l'entraînement et l'évaluation des systèmes adaptés aux sous-types de parole spontanée capturés dans les cas d'étude

domaine spontané. Nous espérons pouvoir rassembler plusieurs centaines d'heures de parole pour ce faire.

Nous souhaitons ensuite étudier l'impact d'une réadaptation spécifique de ce système adapté au domaine spontané, en procédant de la même façon que précédemment. Ainsi, nous apprendrons et adapterons différents systèmes selon nos cas d'étude.

L'ensemble de ces systèmes est optimisé sur les données représentatives de sous-types de parole spontanée. Ils sont testés sur les ensembles de référence Common Voice et ETAPE, ainsi que sur *All_spont test* et sur les données représentatives de nos différents cas. La figure IV.3 présente un récapitulatif des systèmes et données utilisés ici.

4 Synthèse

Ce chapitre a présenté les méthodes mises en place pour pouvoir étudier les questions de recherche suivantes :

- Quels facteurs pour constituer des groupes de données langagières supposées comme similaires et positionner les uns par rapport aux autres des ensembles représentant différents niveaux de spontanéité ?
- L'adaptation spécifique d'un modèle pré-appris avec une faible quantité de données peut-elle améliorer les performances sur la parole spontanée ?

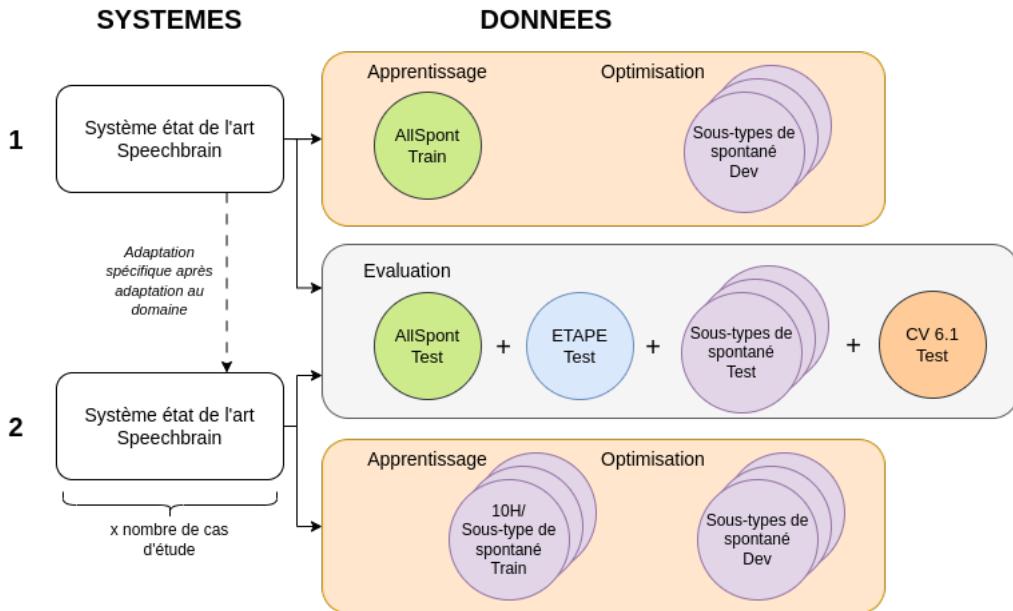


FIGURE IV.3 – Détail des données d'apprentissage, d'optimisation et de test utilisées pour l'entraînement et l'évaluation du système adapté au domaine puis adapté aux sous-types de parole spontanée

- Quel impact de l'adaptation au domaine d'un modèle pré-appris avec une grande quantité de données moins contrôlées ?

Nous commencerons par mettre en parallèle les facteurs de variation de la parole spontanée relevés dans la littérature avec une analyse des métadonnées rattachées à des corpus de linguistiques existants et que nous pourrons utiliser dans nos travaux grâce à leur licence ouverte. Ceci nous permettra de relever un nombre réduit de facteurs et ainsi de former des ensembles de données homogènes, représentatifs de sous-types de parole spontanée. La figure IV.4 présente le récapitulatif des ensembles de données qui seront créés pour mener à bien les expérimentations.

Ensuite, nous sélectionnerons différents ensembles de données que nous souhaitons représentatifs de différents niveaux de spontanéité. Cette spontanéité plus ou moins forte dans nos cas d'étude sera vérifiée grâce aux performances de systèmes de reconnaissance automatique de la parole sur ces cas. Ainsi, nous devrions retrouver le WER le plus élevé sur le cas que nous supposons le plus spontané et le WER le moins élevé sur le cas que nous supposons le moins spontané. Cette tendance pourra notamment être observée grâce à l'utilisation de deux systèmes état de l'art : le système Speechbrain et le système Whisper. Cette validation nous permettra de poursuivre notre étude sur l'amélioration des performances des systèmes sur la

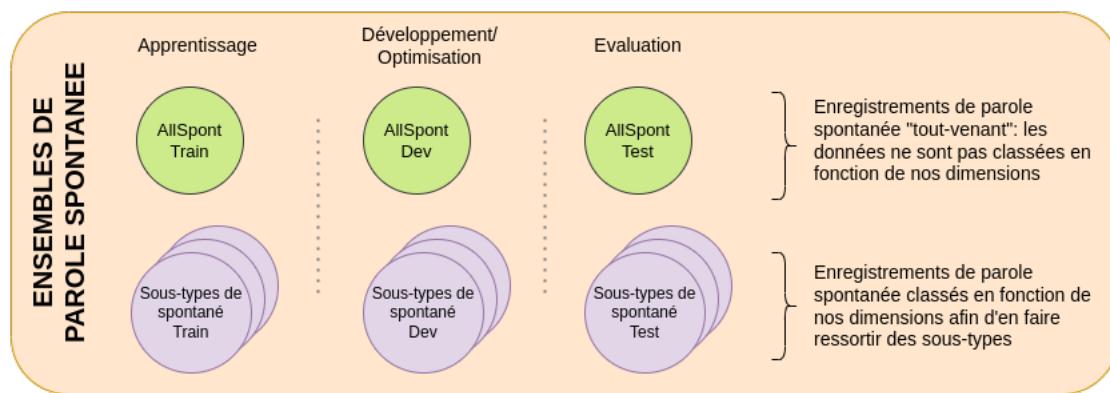


FIGURE IV.4 – Ensembles de données d'apprentissage, de développement et de test qui seront utilisés pour l'apprentissage et l'évaluation de nos systèmes

parole spontanée.

Enfin, nous effectuerons des adaptations spécifiques d'un modèle pré-appris Le-Benchmark dans deux conditions différentes. Tout d'abord, nous adapterons directement le modèle auto-supervisé grâce à des ensembles d'apprentissage représentatifs de nos cas d'études, c'est-à-dire supposés représentatifs de différents niveaux de spontanéité. Les données seront ici en quantité limitée (10 heures). Nous reproduirons ensuite cette adaptation spécifique sur un modèle préalablement adapté au domaine spontané grâce à une grande quantité de parole spontanée “tout-venant”, c'est-à-dire non étiquetées comme appartenant à un sous-type ou à un autre sous-type de parole spontanée. Nous récapitulons l'ensemble des expérimentations qui seront menées dans cette thèse dans le tableau VII.1.

Système	Modèle pré-appris	Adapt.	Données d'apprentissage	Données de validation
Systèmes état de l'art				
Whisper	Whisper-large	Non	-	-
SB-CV	LB-7K-large	Oui	CV	CV
SB-CV2	LB-7K-large	Oui	CV	All_cases
Systèmes adaptés aux sous-types de parole spontanée				
SB- <i>Usual_close</i>	LB-7K-large	Oui	<i>Usual_close</i>	All_cases
SB- <i>Unusual_close</i>	LB-7K-large	Oui	<i>Unusual_close</i>	All_cases
SB- <i>Unusual_distant</i>	LB-7K-large	Oui	<i>Unusual_distant</i>	All_cases
Système adapté au domaine spontané				
*SB- <i>All_spont</i>	LB-7K-large	Oui	All_spont	All_cases
Système adapté au domaine + adapté au sous-types de parole spontanée				
*SB- <i>All_spont-Usual_close</i>	LB-7K-large	Oui	All_spont, <i>Usual_close</i>	All_cases
*SB- <i>All_spont-Unusual_close</i>	LB-7K-large	Oui	All_spont, <i>Unusual_close</i>	All_cases
*SB- <i>All_spont-Unusual_distant</i>	LB-7K-large	Oui	All_spont, <i>Unusual_distant</i>	All_cases

*correspond à un système adapté au domaine spontané // Les systèmes en gras sont les systèmes état de l'art.

TABLE IV.3 – Rappel des différents systèmes utilisés pour les expérimentations, ainsi que des ensembles de données d'apprentissage et de développement.

Troisième partie

Contributions

Chapitre V

Modèles pré-appris pour le français : le projet LeBenchmark

En parallèle de notre travail sur la parole spontanée, nous avons pris part au projet LeBenchmark au début de ce doctorat. Le projet avait pour but de fournir à la communauté de chercheurs en parole les premiers modèles pré-appris pour la parole en français. Des chercheurs de différents laboratoires de recherche publics et privés (LIG, LIA, LAMSADe, Naver Labs Europe et Samsung) ont ainsi unis leurs forces et ont permis, entre 2021 et 2023, la mise à disposition de 14 modèles de type Wav2Vec 2.0 pour le français. Au sein de ce travail, mon rôle aura été d'aider à la recherche, la collecte et la préparation de données nécessaires à l'apprentissage des différents modèles.

Nous présentons dans ce chapitre les données collectées, les modèles appris ainsi que les performances de ceux-ci pour une tâche de reconnaissance automatique de la parole.

Déroulement du chapitre

1 Collecte et pré-traitement des données	100
1.1 Méthode de sélection des données	100
1.2 Présentation des corpus	101
1.2.1 Augmenter la quantité de données disponible : le corpus audiocite.net	102
1.3 Préparation des données	103

1.4	Création d'ensembles d'apprentissage	104
2	Modèles disponibles	107
3	Performances des modèles pour une tâche de reconnaissance automatique de la parole	109
3.1	Modèles pré-appris utilisés comme extracteurs de paramètres : Interspeech 2021	109
3.2	Modèles pré-appris adaptés à la tâche : NeurIPS 2021	110
3.3	LeBenchmark 2.0 : <i>Computer Speech & Language</i>	112
4	Synthèse	114

1 Collecte et pré-traitement des données

1.1 Méthode de sélection des données

L'apprentissage de modèles pré-appris nécessite une grande quantité de données. Dans la littérature, on retrouve des modèles état de l'art appris sur plusieurs milliers d'heures de parole. Les plus gros modèles Wave2Vec 2.0 sur l'anglais rapportées dans Baevski *et al.* (2020) sont ainsi appris sur 53 000 heures de données issues de LibriVox¹, une plateforme en ligne qui rassemble une grande quantité de livres audio. De même, les modèles multilingues XLSR-53 de Conneau *et al.* (2021) sont appris sur 56 000 heures de parole, issues de Common Voice, Babel² et Multilingual LibriSpeech (MLS). Nous cherchons donc pour LeBenchmark à réunir un maximum de données, celles-ci ne nécessitant pas d'être transcrites.

Ainsi, nous avons cherché en priorité des corpus de grande taille afin de réduire au maximum le temps nécessaire à la préparation des données. Deux corpus multilingues venaient d'être partagés à la communauté lorsque nous avons démarré le projet : MLS et Voxpopuli. Le premier rassemble 1 096 heures de lectures en français de livres rentrés dans le domaine public et hébergés sur le site Librivox³. Le second rassemble, pour sa partie française, 4 743 heures⁴ d'enregistrements is-

1. Les auteurs proposent également des modèles appris sur 1 000 heures de données de LibriSpeech.

2. Un corpus multilingue de conversations téléphoniques pour les langues d'Asie et d'Afrique, dont le bengali, le cantonais, le géorgien, l'haïtien, le Kurmanji, le Pashto, le Tamil, le Turc, le Tokpisin et le Vietnamien, langues ayant servi pour l'apprentissage.

3. <https://librivox.org/>

4. Au moment du téléchargement, seule la première version du corpus en français était disponible. Peu après, celle-ci a été étendue à 22 000 heures.

sus de différents types de rencontres du parlement européen entre 2009 et 2020. Ces corpus représentent à eux seuls une quantité assez conséquente de données mais sont assez peu représentatifs de différents types de parole (lecture et parole très préparée seulement). Nous avons souhaité augmenter la quantité de données et diversifier les données disponibles pour le projet LeBenchmark avec des corpus pour la plupart plus petits, mais représentatifs d'une plus grande diversité. Notre recherche a consisté à la fois à l'exploration de plateformes en ligne telles qu'Ortolang, OpenSLR et Cocoon pour la collecte de corpus issus de la recherche en linguistique et en traitement automatique des langues.

Sur 86 corpus considérés, nous en avons conservé 15 qui répondaient à nos exigences. Pour rappel, nous souhaitions les corpus :

- libres de réutilisation pour la recherche,
- de bonne qualité audio,
- de grande taille afin de réduire le temps de préparation des données,
- représentatifs de la variété des types de parole existants (lu, préparée, spontanée, radiophonique/télévisuelle, professionnelle, émotionnelle...),
- dont les auteurs respectent au mieux l'ensemble des principes FAIR, notamment le principe d'accès facile aux données.

Nos recherches nous ont ainsi permis d'obtenir un ensemble de données comprenant :

- de la parole accentuée : MLS, African Accented French, CaFE et Niger-Mali Audio Collection,
- de la parole émotionnelle actée : GEMEP, CaFE et Att-Hack,
- de la parole téléphonique : PORTMEDIA,
- de la lecture : MLS, African Accented French et Mass,
- de la parole spontanée : CFPP, ESLO2, MPF, TCOF, NCCF,
- de la parole radiophonique : EPAC,
- de la parole professionnelle : Voxpopuli.

1.2 Présentation des corpus

Le corpus MLS⁵ rassemble des livres audio et les corpus CFPP, ESLO2, MPF, TCOF et NCCF⁶ recouvrent différentes situations de parole spontanée. Nous présentons ci-après les autres corpus utilisés :

African Accented French (Unk, 2003) : Ce corpus est divisé en trois parties :

5. section 3.2.2 du chapitre 1
6. section 1 du chapitre 6

- Le sous-ensemble *Yaounde* collecté par une équipe du CTELL⁷ de l'académie militaire des États-Unis en 2003 à Yaoundé au Cameroun.
- Le sous-ensemble *CA16* collecté par une équipe scientifique RDECOM à Libreville au Gabon en 2016.
- Le sous-ensemble *Niger* collecté en 2015 à Niamey au Niger.

Att-Hack (Le Moine et Obin, 2020) : C'est un corpus de parole actée comprenant différentes attitudes sociales (amical, séducteur, dominant, distant). Il rassemble en tout 100 énoncés produit selon les 4 attitudes sociales mentionnées.

Canadian French Emotional speech dataset - CaFE (Gournay et al., 2018) : Ce corpus a été créé en demandant à 12 acteurs de prononcer 6 phrases différentes avec 6 émotions basiques (colère, dégoût, joie, peur, surprise, tristesse) ainsi que de façon neutre. Chacune des émotions est jouée selon deux niveaux d'intensité.

GEneva Multimodal Emotion Portrayals - GEMEP (Bänziger et al., 2012) : Le GEMEP est une collection d'enregistrements audio et vidéo de 10 acteurs mettant en scène 18 états affectifs, avec différents contenus verbaux et différents modes d'expression. Les émotions incluses sont l'allégresse, l'amusement, la fierté, la peur, le désespoir, le plaisir, le soulagement, l'intérêt, la colère, la joie, l'irritation, l'anxiété et la tristesse.

MaSS (Boito et al., 2020) : Le MaSS est un corpus multilingue aligné d'enregistrements de la Bible issus de *Bible.is*. Les enregistrements ont tous été effectués par des locuteurs natifs.

Niger-Mali Audio Collection (Zanon Boito et al., 2022b,c) : Ce corpus comprend les enregistrements distribués lors du challenge “Low-resource Speech Translation” de la conférence IWSLT 2022. Les données proviennent des sites *Studio Kalangou* et *Studio Tamani*, avec l'autorisation de la fondation Hirondelle.

PORTRMEDIA (Lefèvre et al., 2012) : Ce corpus réunit des interactions entre humain et machine (technique du magicien d'Oz). Il existe une partie en français réunissant des dialogues à propos de la réservation de tickets pour un festival et une partie en italien réunissant des dialogues à propos du tourisme.

1.2.1 Augmenter la quantité de données disponible : le corpus audio-cite.net

Afin de pouvoir agrandir considérablement la quantité de données disponible pour l'apprentissage des modèles, nous avons également créé un corpus sur la base de

7. Center for Technology Enhances Language Learning

livres audio téléchargés sur le site audiocite.net⁸. La sélection des livres audio s'est faite sur la base de la licence Creative Commons associée aux enregistrements. Les lecteurs proposant des enregistrements ont le choix entre lire des livres rentrés dans le domaine public ou dont les auteurs ont donné l'autorisation d'enregistrer et mettre à disposition la version audio sur *audiocite.net*.

La version initiale du corpus utilisée pour le projet LeBenchmark rassemble 6 698 heures d'enregistrements pour 130 locuteurs. Les enregistrements ne sont pas accompagnés de leur transcription mais ceci ne présentait pas de problème dans le cadre de l'apprentissage de modèles auto-supervisés.

Nous avons décidé de mettre cette ressource à disposition de la communauté en la partageant sur OpenSLR⁹. Cette mise en ligne a été précédée d'un second nettoyage des données : retrait de quelques enregistrements dupliqués, vérification du genre des locuteurs. Nous proposons également dorénavant un découpage des données en ensembles d'apprentissage, de développement et de test. Les enregistrements sont distribués dans leur version originale (*mp3* pour la plupart). Nous les partageons accompagnés de métadonnées renseignées sur le site par les lecteurs, telles que la durée de l'enregistrement, l'identifiant du lecteur, le titre du livre lu, l'auteur du livre, la catégorie associée au livre et la licence. Le corpus complet est décrit dans (Felice *et al.*, 2024).

1.3 Préparation des données

L'apprentissage d'un modèle de type Wav2Vec 2.0 nécessite des enregistrements de parole au format mono, 16 bits, 16 kHz, de 30 secondes maximum. Les fichiers audio récoltés sont bien plus longs et nous avons donc effectué une première étape de segmentation en nous appuyant, lorsque les données étaient transcrrites, sur le découpage effectué par les auteurs dans les fichiers de transcriptions.

Certains formats de fichiers d'annotation (trs notamment) nous auront permis également de retirer les sections de parole superposée. Les corpus ne comprenant pas de fichiers d'annotations ont été découpés en segments de 30 secondes. De plus, nous avons retiré l'ensemble des segments inférieurs à 1 seconde.

Quand l'information était disponible dans les métadonnées, nous avons récupéré le genre des locuteurs ainsi que leur identifiant afin de pouvoir calculer le nombre de locuteurs par corpus. Le nombre de locuteurs dans le corpus EPAC n'est toutefois pas exact. En effet, ce corpus étant un corpus d'émissions de radio, certains locuteurs se retrouvent dans plusieurs émissions. Une grande partie des données

8. Nous remercions l'administrateur du site pour son autorisation.

9. <https://www.openslr.org/139/>

ayant été annotées automatiquement, deux identifiants différents ont alors pu leur être attribués.

L'ensemble des scripts liés aux données est disponible dans le répertoire Github du projet¹⁰.

1.4 Création d'ensembles d'apprentissage

Les corpus rassemblés ont été répartis en cinq ensembles : *small dataset*, *medium-clean dataset*, *medium dataset*, *large dataset* et *extra large dataset*¹¹ que nous présentons et dont nous commentons le contenu ci-après. Ces différents ensembles nous auront permis de tester à la fois l'impact de la quantité de données et l'apport d'une certaine diversité dans les données de pré-apprentissage sur la performance des modèles.

Small dataset ($\approx 1\,000$ heures) : cet ensemble comprend uniquement les données en français issues du corpus MLS. Il a permis d'apprendre un modèle uniquement sur de la lecture et ainsi créer un modèle similaire à celui entraîné dans Baevski *et al.* (2020) sur l'anglais. Cela représente 1 096 heures d'enregistrement.

Medium-clean dataset ($\approx 2\,700$ heures) : Cet ensemble comprend les données de MLS (1 096 heures de lecture) et d'EPAC (1 626 heures de données radiophoniques plutôt spontanées). Ces corpus sont représentatifs de la plupart des corpus utilisés pour la reconnaissance automatique de la parole en français.

Medium dataset ($\approx 3\,000$ heures) : cet ensemble comprend l'ensemble des données du *medium-clean dataset* auxquelles nous avons ajouté un corpus de lecture, quatre corpus de parole actée et quatre corpus de parole spontanée. Ceci représente 2 933 heures de parole, dont 67 heures de parole actée, 123 h de parole spontanée, 1 115 h de lecture et 1 626 h de radio (parole préparée/spontanée en grande partie professionnelle). Cet ensemble se veut plus diversifié que le *medium-clean* qui a été constitué ainsi afin de pouvoir étudier l'impact de données de parole spontanée sur l'apprentissage des modèles Wav2Vec 2.0..

Large dataset ($\approx 7\,700$ heures) : cet ensemble comprend l'ensemble des données du *medium dataset* ainsi que les données d'un corpus de lecture (MaSS), d'un corpus de parole spontanée (NCCFr) et d'un corpus de parole professionnelle (Vox-populi *unlabeled + transcribed*). Ceci représente 7 739 heures de parole, dont 67 heures de parole actée, 165 heures de parole spontanée, 1 135 heures de parole lue, 1 626 heures de radio (parole préparée/spontanée en grande partie professionnelle)

10. <https://github.com/LeBenchmark>

11. La publication originale présentant le projet ayant été faite à Interspeech, nous avons conservé les noms en anglais pour une meilleure compréhension.

et 4 744 heures de parole professionnelle (parole préparée). Cet ensemble a été constitué afin de pouvoir étudier les performances sur les différentes tâches lorsque le modèle est entraîné avec plus de données.

Extra large dataset ($\approx 14\,000$ heures) : cet ensemble comprend l'ensemble des données du *large dataset* ainsi que les données de deux corpus : Niger-Mali Audio Collection et Audiocite.net. L'ensemble de données comprend donc 67 heures de parole actée, 165 heures de parole spontanée, 1 737 heures de radio , 4 744 heures de parole professionnelle et 7 834 heures de lecture. De même que le précédent, cet ensemble a été créé afin d'évaluer les performances sur les différentes tâches d'un modèle avec plus de données. De plus, il permet également d'évaluer l'impact de l'ajout d'un grand ensemble de données lues.

Le tableau V.1 présente les corpus regroupés selon ces cinq ensembles et récapitule les caractéristiques globales de chacun d'entre eux.

La segmentation effectuée en fonction des fichiers de transcription apporte à l'ensemble de données une représentation de tours de parole correspondant mieux à ceux de la vie réelle que dans les corpus MLS, Voxpopuli et Audiocite.net, les 29 secondes de durée moyenne des segments pour ces deux derniers corpus correspondant à un découpage strict à 30 secondes lorsque les transcriptions n'étaient pas disponibles. Il ressort de ces durées moyennes que les segments de parole sont plus courts en parole spontanée (entre 1,9 et 6 secondes) qu'en parole professionnelle (entre 9 et 10 secondes). Le nombre de locuteurs, quant à lui, n'est pas forcément proportionnel à la taille du corpus. Ainsi, les 6 698 heures rassemblées dans Audiocite.net ne sont produites que par 130 locuteurs alors que les 18 heures d'African Accented French sont produites par 232 locuteurs.

Nous présentons, en figure V.1, la répartition de chaque type de parole dans chacun des ensembles de données présentés sous forme graphique. Nous avons inclus le type radiophonique dans le type professionnel, plus global. Les ensembles *small* et *extra large* sont ceux comprenant la plus grande proportion de lecture. L'ensemble *medium* est celui comprenant la plus grosse proportion de parole spontanée, toutefois celle-ci ne dépasse pas 4,2% de l'ensemble. Dans l'ensemble *large*, la parole professionnelle est de loin la plus représentée (82%). Les types de parole spontanée, émotionnelle et téléphonique étant largement sous-représentés dans l'ensemble des modèles, il est probable que l'évaluation sur de tels types de parole donne des performances plus réduites qu'avec des modèles qui auraient été appris sur données par types de parole mieux réparties. Cependant, nous espérons que leur inclusion permet tout de même aux différents modèles entraînés d'avoir pu apprendre sur ces types de parole. Il est en effet important de noter qu'en comparaison avec les modèles état de l'art Wav2Vec 2.0 et XLSR-53, les données de pré-apprentissage uti-

CHAPITRE V : Modèles pré-appris pour le français : le projet LeBenchmark

Corpus	# Segments	Durée [hh :mm]	# Locuteurs	Durée moy. segment	Type de parole
Small dataset – 1K					
MLS French CC BY 4.0	263,055 124,590 / 138,465 / -	1,096 :43 520 :13 / 576 :29 / -	178 80 / 98 / -	15 s 15 s / 15 s / -	Lecture
Medium-clean dataset – 2.7K					
EPAC** ELRA NC	623,250 465,859 / 157,391 / -	1,626 :02 1,240 :10 / 385 :52 / -	Unk - / - / -	9 s - / - / -	Émissions de radio
Medium-clean dataset total	886,305 590,449 / 295,856 / -	2,722 :45 1,760 :23 / 962 :21 / -	-	-	-
Medium dataset – 3K					
African Accented French Apache 2.0	16,402 373 / 102 / 15,927	18 :56 - / - / 18 :56	232 48 / 36 / 148	4 s - / - / -	Lecture
Att-Hack CC BY NC ND	36,339 16,564 / 19,775 / -	27 :02 12 :07 / 14 :54 / -	20 9 / 11 / -	2.7 s 2.6 s / 2.7 s / -	Émotionnelle actée
CaFE CC NC	936 468 / 468 / -	1 :09 0 :32 / 0 :36 / -	12 6 / 6 / -	4.4 s 4.2 s / 4.7 s / -	Émotionnelle actée
CFPP2000* CC BY NC SA	9853 166 / 1,184 / 8,503	16 :26 0 :14 / 1 :56 / 14 :16	49 2 / 4 / 43	6 s 5 s / 5 s / 6 s	Spontanée
ESLO2 CC BY-NC-SA	62,918 30,440 / 32,147 / 331	34 :12 17 :06 / 16 :57 / 0 :09	190 68 / 120 / 2	1.9 s 2 s / 1.9 s / 1.7 s	Spontanée
GEMEP User agreement	1,236 616 / 620 / -	0 :50 0 :24 / 0 :26 / -	10 5 / 5 / -	2.5 s 2.4 s / 2.5 s / -	Émotionnelle actée
MIPF CC BY-NC-SA 4.0	19,527 5,326 / 4,649 / 9,552	19 :06 5 :26 / 4 :36 / 9 :03	114 36 / 29 / 49	3.5 s 3.7 s / 3.6 s / 3.4 s	Spontanée
PORTMEDIA (FR) ELRA NC	19,627 9,294 / 10,333 / -	38 :59 19 :08 / 19 :50 / -	193 84 / 109 / -	7.1 s 7.4 s / 6.9 s / -	Dialogue téléphonique acté
TCOF (Adults) CC BY-NC-SA	58,722 10,377 / 14,763 / 33,582	53 :59 9 :33 / 12 :39 / 31 :46	749 119 / 162 / 468	3.3 s 3.3 s / 3.1 s / 3.4 s	Spontanée
Medium dataset total	1,111,865 664,073 / 379,897 / 67,895	2,933 :24 1,824 :53 / 1,034 :15 / 74 :10	-	-	-
Large dataset – 7K					
MaSS MIT	8,219 8,219 / - / -	19 :40 19 :40 / - / -	Unk - / - / -	8.6 s 8.6 s / - / -	Lecture
NCCFr User agreement	29,421 14,570 / 13,922 / 929	26 :35 12 :44 / 12 :59 / 00 :50	46 24 / 21 / 1	3 s 3 s / 3 s / 3 s	Spontanée
Voxpopuli Unlabeled CC0	568,338 - / - / -	4,532 :17 - / - / 4,532 :17	Unk - / - / -	29 s - / - / -	Parole professionnelle
Voxpopuli transcribed CC0	76,281 - / - / -	211 :57 - / - / 211 :57	327 - / - / -	10 s - / - / -	Parole professionnelle
Large dataset total	1,814,242 682,322 / 388,217 / 99,084	7,739 :22 1,853 :02 / 1,041 :07 / 4,845 :07	-	-	-
Extra Large dataset – 14K					
Audiocite.net CC BY + ND/NC/SA	817 295 425 033 / 159 691 / 232 571	6698 :35 3477 :24 / 1309 :49 / 1911 :21	130 35 / 32 / 63	29 s 29 s / 29 s / 29 s	Lecture
Niger-Mali Audio Collection CC BY NC ND	38 332 18 546 / 19 786 / -	111 :01 52 :15 / 58 :46 / -	357 192 / 165 / -	10 s 10 s / 10 s / -	Émissions de radio
Extra Large dataset total	2 669 869 1 125 901 / 567 694 / 331 655	14 548 :58 5 382 :41 / 2 409 :42 / 6 756 :28	-	-	-

*Composé d'enregistrements non inclus dans le corpus CEFC v2.1, 02/2021 ; **les locuteurs n'ont pas un identifiant unique

TABLE V.1 – Statistiques des corpus oraux utilisés dans le projet LeBenchmark en fonction du genre (homme/femme/non renseigné)

lisées pour l'apprentissage des modèles LeBenchmark contiennent une plus grande diversité de types de parole.

La répartition en genre dans chacun des ensembles de données est représentée en figure V.2. Celle-ci nous montre que plus la taille des ensembles de données grossit, plus la présence de locuteurs dont le genre n'est pas renseigné augmente. Il nous

2 Modèles disponibles

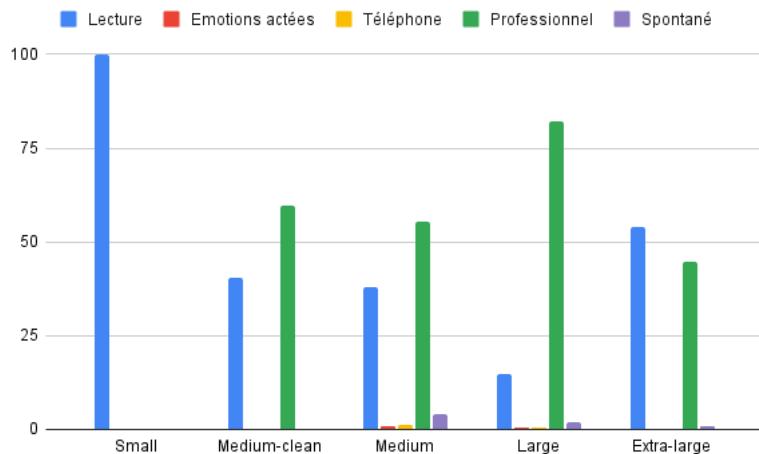


FIGURE V.1 – Représentation des différents types de parole dans les ensembles de données *small*, *medium*, *large* et *extra large* (en %)

semble donc important de rappeler l'effort qui a été fait pour fournir cette information au sein du corpus Audiocite.net : sur 6 698 heures au total, seules 1 911 heures d'enregistrement ne se sont pas vues attribuer un genre. Le graphique nous montre également que la proportion de locutrices est toujours plus basse que la proportion de locuteurs, excepté pour l'ensemble *small* constitué du corpus MLS uniquement où le ratio est équilibré, ainsi que le voulaient les créateurs du corpus. Dans tous les autres ensembles de données rassemblés pour LeBenchmark, la proportion de femmes équivaut à chaque fois à peu près à la moitié de la proportion d'hommes.

2 Modèles disponibles

Les différents ensembles de données rassemblés ont servi à l'apprentissage de 10 modèles de type Wav2Vec 2.0 pour le français. Quatre modèles supplémentaires ont été partagés par la suite par certains membres du projet pour une étude spécifique sur le genre (Zanon Boito *et al.*, 2022a). L'ensemble de ces modèles est à retrouver sur le répertoire HuggingFace du projet¹². Cette plateforme permet une intégration rapide aux scripts de différentes boîtes à outils telle que Speechbrain¹³.

Le tableau V.2 présente, pour chacun des 10 modèles appris au sein du projet, la quantité de données d'apprentissage, le nombre de paramètres, la taille des vecteurs en sortie, le nombre de mises à jour, le nombre de GPU utilisés et la durée d'heures

12. <https://huggingface.co/LeBenchmark>

13. <https://speechbrain.github.io/>

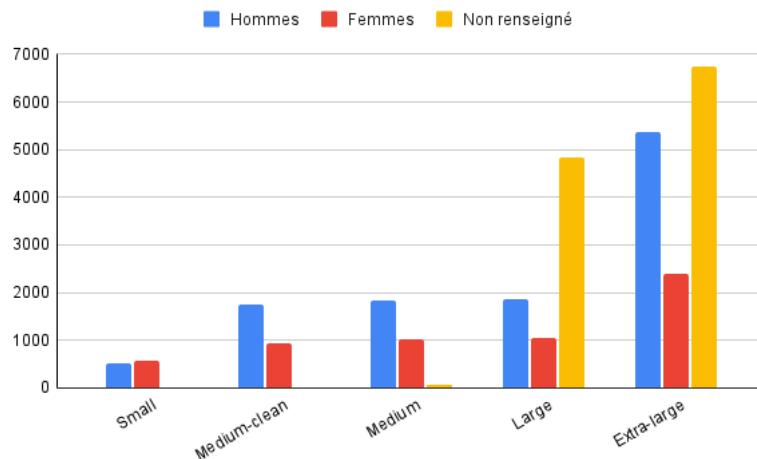


FIGURE V.2 – Répartition par genre (Hommes/Femmes/Non renseigné) dans les différents ensembles de données LeBenchmark (en heures)

GPU nécessaire à l'entraînement.

Modèle	Données de pré-apprentissage	#Paramètres	Dimension de sortie	Mises à jour	#GPU	Heures GPU
1K-base	1,096 h	90M	768	200K	4	1,000
1K-large	1,096 h	330M	1024	200K	32	3,700
2.7K-base	2,773 h	90M	768	500K	32	4,100
3K-base	2,933 h	90M	768	500K	32	4,100
3K-large	2,933 h	330M	1024	500K	32	10,900
7K-base	7,739 h	90M	768	500K	64	7,900
7K-large	7,739 h	330M	1,024	500K	64	13,500
14K-light	14,000 h	26M	512	500K	32	5,000
14K-large	14,000 h	330M	1,024	1M	64	28,800
14K-xlarge	14,000 h	965M	1,280	1M	104	54,600

"Heures GPU" réfère au temps total d'apprentissage cumulé sur le "#GPU" pour atteindre le nombre de "Mises à jour"

TABLE V.2 – Récapitulatif des modèles pré-appris Wav2vec 2.0 appris au cours du projet LeBenchmark

Les cinq ensembles d'apprentissage correspondent au début des noms des modèles. Le nombre de paramètres dépend des dimensions de sortie, et les heures et le nombre de GPU sont très variables, depuis 4 GPU pour 1 000 heures d'apprentissage pour le plus petit modèle 1K-base, jusqu'à 104 GPU pour 54 600 heures d'apprentissage pour le modèle 14K-xlarge.

3 Performances des modèles pour une tâche de reconnaissance automatique de la parole

Nous présentons dans cette section les performances obtenues avec des systèmes *end-to-end*. Deux corpus de référence pour l'évaluation de systèmes de reconnaissance de la parole ont été utilisés : Common Voice et ETAPE. La version 6.1 de Common Voice contient 428 heures dédiées à l'apprentissage, 24 heures pour le développement et 25 heures pour le test. Les données d'ETAPE sont réparties respectivement en ensembles de 22 heures, 7 heures et 7 heures pour les sets d'apprentissage, de développement et de test respectivement. Ces deux corpus de taille différente ont permis de tester les modèles à la fois dans un scénario où beaucoup de données sont disponibles, mais aussi dans un scénario avec peu de ressources. Les expérimentations ont été menées avec la boîte à outils Speechbrain.

3.1 Modèles pré-appris utilisés comme extracteurs de paramètres : Interspeech 2021

Nous reportons ici une partie des résultats de la tâche de reconnaissance automatique de la parole présentée à Interspeech 2021 (Evain et al., 2021a).

L'utilisation d'un modèle pré-appris comme extracteur de paramètres revient à utiliser un modèle tel qu'il est, sans l'entraîner davantage et sans l'adapter à la tâche visée. On dit alors qu'il est figé. Pour cette première expérimentation, deux systèmes sont utilisés. Tout d'abord, lorsque l'extraction de paramètres est faite à partir de *Mel Filter Banks* (MFB), le système est de type “encodeur-décodeur”, l'encodeur étant constitué d'un CRDNN (*Convolutionnal Recurrent Deep Neural Network* - Réseau neuronal profond récurrent convolutionnel), et le décodeur étant un réseau à long et court terme (LSTM) avec CTC/attention conjointe. Ensuite, lorsque les paramètres sont de type Wav2Vec 2.0, la couche de convolution du CRDNN est retirée. Aucun modèle de langue n'est utilisé. La couche de sortie correspond à un vocabulaire de 500 BPE (pour *Byte Pair Encoding*¹⁴).

Le premier tableau de résultats présenté V.3 permet de comparer l'utilisation de paramètres de type MFB (*Mel Filter Banks*) et de paramètres issus de trois modèles pré-appris :

- notre modèle 3K-*large* appris sur l'ensemble *medium*
- le modèle Wav2Vec 2.0 *large* appris sur 960 heures en anglais du corpus MLS (Baevski et al., 2020)

14. Les BPE consistent à remplacer les paires d'octets qui se répètent régulièrement par un octet unique.

- et le modèle multilingue XLSR-53 *large* appris sur 56 000 heures de données et 53 langues (Conneau *et al.*, 2021).

Corpus	Common Voice		ETAPE		
	Paramètres	Dev	Test	Dev	Test
MFB	20,19	23,40	54,55	56,17	
3K- <i>large</i>	20,23	24,06	55,56	57,04	
En- <i>large</i>	34,07	37,29	98,79	99,10	
XLSR-53- <i>large</i>	30,07	32,72	(*)	(*)	

(*) indique que l'algorithme d'apprentissage n'a pas convergé à un WER inférieur à 100%.

TABLE V.3 – Résultats de Evain *et al.* (2021a) pour une tâche de RAP *End-to-end* (%WER) sur les corpus en français Common Voice et ETAPE

Les résultats montrent que le système utilisant les MFB donne de meilleures performances, et ce, peu importe le corpus de test. Cependant, le système utilisant le modèle LeBenchmark 3K-*large* pour l'extraction de paramètres permet d'obtenir des performances très proches de ce qui est obtenu avec les MFB. *A contrario*, les performances obtenues avec les modèles pré-appris sur de l'anglais ou plusieurs langues font largement perdre en performance sur Common Voice et sont même inefficaces sur ETAPE. Pourtant, le modèle multilingue XLSR-53 inclut du français. Les résultats montrent donc ici de meilleures performances pour notre modèle monolingue, appris sur la langue de test cible, le français.

3.2 Modèles pré-appris adaptés à la tâche : NeurIPS 2021

*Nous reportons ici une partie des résultats de la tâche de reconnaissance automatique de la parole présentée à NeurIPS 2021 (Evain *et al.*, 2021b).*

Un des apports les plus remarquables des modèles pré-appris est le fait que l'on peut les adapter avec peu de données. Ainsi, le modèle n'est plus utilisé comme simple extracteur de paramètres acoustiques : on continue de l'entraîner en même temps que la tâche (ici la reconnaissance automatique de la parole) afin qu'il apprenne à ressortir les caractéristiques les plus intéressantes dans les données utilisées pour la tâche. Notre publication à NeurIPS en 2021 présente les résultats sur les mêmes corpus qu'auparavant, avec cette adaptation *end-to-end* en plus. Nous reportons dans le tableau V.4 les résultats obtenus avec des MFB et les modèles pré-appris adaptés suivants :

- notre modèle 2.7K-base appris sur l'ensemble *medium-clean*
- nos modèles 3K-base et 3K-*large* appris sur l'ensemble *medium*

3 Performances des modèles pour une tâche de reconnaissance automatique de la parole

- nos modèles 7K-base et 7K-*large* appris sur l’ensemble *large*
- le modèle Wav2Vec 2.0 *large* appris sur 960 heures en anglais du corpus MLS (Baevski *et al.*, 2020)
- et le modèle multilingue XLSR-53 *large* appris sur 56 000 heures de données et 53 langues (Conneau *et al.*, 2021).

L’architecture des systèmes est cette fois simplifiée : un modèle pré-appris + une couche cachée + une couche de sortie dont la taille dépend du nombre de caractères dans l’ensemble d’apprentissage. Encore une fois, aucun modèle de langue n’est utilisé.

Corpus	CommonVoice		ETAPE		
	Paramètres	Dev	Test	Dev	Test
MFB	17,67±0,37	20,59±0,41	54,03±1,33	54,36±1,32	
2,7K- <i>base</i>	11,04±0,27	13,09±0,24	26,23±0,78	29,08±0,80	
3K- <i>base</i>	11,25±0,23	13,22±0,24	26,14±0,70	28,86±0,79	
3K- <i>large</i>	8,34 ±0,18	9,75 ±0,20	23,51 ±0,68	26,14 ±0,77	
7K- <i>base</i>	10,84±0,21	12,88±0,24	25,13±0,68	28,16±0,79	
7K- <i>large</i>	8,55±0,18	9,94±0,21	24,14±0,70	27,25±0,78	
En- <i>large</i>	12,05±0,23	14,17±0,52	42,14±0,72	44,82±0,74	
XLSR-53- <i>large</i>	16,41±0,27	19,40±0,29	58,55±0,65	61,03±0,70	

TABLE V.4 – Résultats de Evain *et al.* (2021b) pour une tâche de RAP *end-to-end* (WER%) sur les corpus Common Voice et ETAPE, obtenus avec des modèles pré-appris de type Wav2Vec 2.0 adaptés avec les données correspondantes

Les résultats de cette nouvelle étude reportés dans le tableau V.4 montrent que les meilleures performances ne sont plus obtenus avec les MFB, mais avec le modèle 3K-*large*, et ce pour CommonVoice comme pour ETAPE. L’adaptation des modèles pré-appris est donc réellement une méthode efficace pour améliorer les résultats, y compris avec 22 heures de données d’apprentissage pour ETAPE. L’adaptation du modèle 3K-*large* et l’architecture utilisée ici nous permettent en effet d’obtenir un WER de 9,75% sur Common Voice et 26,14% sur ETAPE, contre 20,59% et 54,36% lors de l’utilisation de MFB, ce qui représente une amélioration remarquable.

Sur le corpus ETAPE contenant des données de radio et télévision plutôt spontanées, nous remarquons que les résultats sont meilleurs avec le modèle 3K-*large* qu’avec le modèle 2.7K-*base*. Pourtant, la parole professionnelle utilisée pour l’apprentissage du modèle 2.7K-*base* est de même type que celle contenue dans ETAPE (radio), et représente une plus grande partie des données d’apprentissage que celles

ayant servi à l'apprentissage du modèle 3K-*large*. Se posent alors les questions de l'impact la taille des modèles pré-appris, l'importance de l'ajout de données et de l'impact du type de parole représenté dans les données ajoutées.

La comparaison entre les modèles 2.7K-*base*, 3K-*base* et 3K-*large* nous montre que la taille de l'architecture Wav2Vec 2.0 utilisée a effectivement son importance. Le WER est meilleur avec l'architecture *large*. Ensuite, les résultats obtenus avec le modèle 7K-*large* nous montrent que l'ajout d'environ 4 000 heures de données n'améliore pas les résultats obtenus avec le modèle 3K-*large*. Le doublement des heures de données d'apprentissage n'a donc finalement que peu d'impact pour cette tâche de reconnaissance automatique de la parole, sur Common Voice, comme sur ETAPE. De plus, si nous considérons la parole radiophonique et télévisuelle comme une parole professionnelle, alors l'ajout des 4 700 heures de données de Voxpopuli aurait dû permettre l'obtention de meilleurs résultats avec le 7K-*large* qu'avec le 3K-*large* sur ETAPE, qui contient également de la parole professionnelle, mais énoncée dans un contexte radiophonique.

Les résultats sur Common Voice montrent eux-aussi l'importance de la taille du modèle pré-appris. Cependant, l'impact de l'ajout de données de parole lue ne peut être ici réellement évalué, la lecture représentant seulement 0,3% des données ajoutées entre les modèles 3K et les modèles 7K.

Enfin les résultats obtenus avec les modèles En-*large* et XLSR-53-*large* montrent un apport des modèles appris sur l'anglais ou sur de nombreuses langues, en comparaison aux MFB, et ce sur Common Voice comme sur ETAPE. Toutefois, cet apport n'est que peu visible avec XLSR-53-*large* sur Commonvoice et inexistant sur ETAPE. Nos modèles pré-appris monolingues sont tous plus performants que ces deux modèles état de l'art. Ces résultats soulignent la puissance de ces modèles pré-appris lorsqu'ils sont adaptés, même avec une quantité de données de taille réduite.

3.3 LeBenchmark 2.0 : Computer Speech & Language

Nous reportons ici une partie des résultats de la tâche de reconnaissance automatique de la parole inclus dans notre article publié dans la revue Computer Speech & Language (Parcollet et al., 2024).

Nous avons souhaité vérifier l'impact de la quantité et du type de données utilisées pour l'apprentissage de modèles Wav2Vec 2.0. Ainsi, nous présentons de nouveaux résultats obtenus avec le modèle 1K *large* appris sur 1 000 heures de données issues de MLS, et les modèles 14K *light/large/xlarge* appris sur 14 000 heures de données. Les WER obtenus avec les modèles 3K *large* et 7K *large* ne sont pas exactement les mêmes que dans notre publication à NeurIPS 2021 du fait de quelques changements

3 Performances des modèles pour une tâche de reconnaissance automatique de la parole

de méthode entre les deux études dont le changement d'un optimiseur, l'utilisation de la version HuggingFace des modèles LeBenchmark au lieu de la version Fairseq et une technique d'augmentation de données utilisée lors de l'adaptation. Les résultats sont à retrouver dans le tableau V.5.

Corpus	CommonVoice		ETAPE		
	Paramètres	Dev	Test	Dev	Test
1K-large	9.49±0.20	11.21±0.23	28.57±0.79	30.58±0.88	
3K-large	8.00±0.19	9.27±0.20	22.26±0.76	24.21±0.85	
7K-large	8.02±0.18	9.39±0.21	21.34±0.74	23.46±0.83	
14K-light	19.86±0.28	22.81±0.34	58.30±0.66	59.82±0.7	
14K-large	8.39±0.19	9.83±0.21	23.67±0.81	26.03±0.89	
14K-xlarge	8.26±0.19	9.83±0.21	22.38±0.95	24.67±0.83	

TABLE V.5 – Résultats de Parcollet *et al.* (2024) pour une tâche de RAP *end-to-end* (WER%) sur les corpus Common Voice et ETAPE, obtenus avec des modèles pré-appris de type Wav2Vec 2.0 adaptés avec les données correspondantes

Cette étude nous permet tout d'abord de constater que pour Common Voice, le modèle permettant d'obtenir le meilleur WER est le 3K-*large* (9,27%). Si l'on observe un apport de la quantité de données entre le modèle 1K-*large* et 3K-*large*, cela ne se vérifie plus ensuite. Les performances ont en effet tendance à stagner voir à se dégrader légèrement avec les modèles de 7K et 14K de type *large* et *xlarge*.

Pour ETAPE, le meilleur modèle est le 7K-*large*. Tout comme pour Common Voice, on observe que l'ajout de données de pré-apprentissage supplémentaire n'est pas toujours bénéfique. Ainsi, le modèle 14K-*large* offre de moins bonnes performances.

L'analyse de l'ajout de données représentant un type de parole similaire à celui compris dans les ensembles de test, nous montre ensuite que pour Common Voice, cet ajout spécifique ne semble pas avoir d'impact. En effet, le WER obtenu avec le modèle 14K-*large* est plus élevé que ce qui est obtenu avec le modèle 3K-*large* et le 7K-*large*. Pourtant, les données ajoutées pour constituer l'ensemble d'apprentissage *extra-large* utilisé pour l'apprentissage des modèles 14K sont principalement de la lecture (98% des 6 809 heures ajoutées). Cette absence d'amélioration pourrait être due à la différence entre les données Common Voice et les données audiocite.net. En effet, si l'on est chaque fois en présence de lecture, les données issues de Common Voice incluent différents accents, différentes conditions d'enregistrement et représentent une tâche bien différente de la lecture d'un livre entier. Les locuteurs sont en effet invités à enregistrer uniquement des phrases. Si les en-

registrement se trouvant dans audiocite.net sont également issus d'une plateforme collaborative, aucune information n'est donnée sur la présence de parole accentuée.

Sur ETAPE et contrairement aux résultats présentés dans (Evain *et al.*, 2021b), le modèle 7K-*large* permet d'obtenir les meilleures performances. Les changements effectués lors de l'apprentissage et l'adaptation des modèles (changement d'optimiseur, HuggingFace vs Fairseq, augmentation des données) semblent donc avoir affecté cette tendance.

Enfin, du point de vue de la taille de l'architecture Wav2Vec 2.0, on remarque que le modèle 14K-*light* (26M de paramètres) est celui avec les WER les plus élevés sur chacun des ensembles de test, ce qui n'est pas étonnant du fait que nous avions déjà observé une différence entre les modèles *base* (90M de paramètres) et *large* (330M de paramètres) dans (Evain *et al.*, 2021a,b). L'architecture étant ici encore plus petite, cela confirme l'importance d'un certain nombre de paramètres pour la performance des modèles. Le modèle 14K-*xlarge* quant à lui n'apporte pas d'amélioration par rapport au modèle 14K-*large* sur Common Voice, mais apporte une amélioration de 1,36 points de WER sur ETAPE. Cependant, malgré cette amélioration, les meilleurs résultats restent obtenus avec le modèle 7K-*large* sur ce corpus et avec le modèle 3K-*large* sur Common Voice.

4 Synthèse

Dans ce chapitre, nous avons présenté le projet LeBenchmark auquel nous avons pu prendre part au début de cette thèse avec des collègues de différents laboratoires privés et publics (LIG, LIA, Lamsade, Naver Labs Europe et Samsung). Mon rôle dans ce projet aura été de participer à la collection et à la création de corpus pour l'entraînement de modèles pré-appris de type LeBenchmark. Nous sommes parvenu à créer cinq ensembles de données, regroupant de 1 000 heures à 14 500 heures de parole, notamment grâce à la création du corpus de livres audio audiocite.net qui représente à lui seul 6 698 heures.

Dix modèles pré-appris pour le français ont pu être appris au sein du projet (+4 issus de travaux additionnels). Ceux-ci sont tous mis à la disposition de la communauté via la plateforme HuggingFace. Dans une démarche de recherche reproduitible, nous partageons l'ensemble de nos scripts ainsi que le détail des données utilisées¹⁵.

Les expérimentations sur une tâche de reconnaissance automatique de la parole montrent plusieurs choses. Tout d'abord, on observe un réel apport de nos modèles

15. <https://huggingface.co/LeBenchmark>

monolingues par rapport aux modèles état de l'art Wav2Vec 2.0 *large* (Baevski *et al.*, 2020) et XLSR-53 *large* (Conneau *et al.*, 2021), respectivement monolingues anglais et multilingues. Ce travail collectif nous permet donc, dans le cadre de cette thèse, de pouvoir travailler par la suite avec des systèmes état de l'art sur la reconnaissance automatique de la parole en français.

Nous observons ensuite un impact limité de la quantité de données sur les résultats. En effet, les modèles 14K appris sur 14 000 heures de données ne sont pas les plus performants.

L'adaptation du modèle 7K-*large* avec 22 heures de données permet d'obtenir un WER de 23% sur ETAPE. Ceci ouvre des perspectives intéressantes pour l'amélioration des systèmes de reconnaissance automatique de la parole sur de la parole spontanée avec une faible quantité de données.

Enfin, nous nous interrogeons sur l'impact de la diversité et/ou de la spécificité des données utilisées pour l'apprentissage de ces modèles. En effet, si les modèles 14K ne donnent pas les meilleures performances sur Common Voice malgré l'ajout d'une grande quantité de parole lue, l'ajout de données similaires aux données d'ETAPE (spontané + professionnel) dans le modèle 7K-*large* pourrait expliquer au moins en partie la meilleure performance de ce modèle sur ce corpus. Les comparaisons et conclusions sont toutefois rendues difficiles par le fait qu'il n'existe pas qu'un seul type de parole spontanée, tout comme il n'existe pas qu'un seul type de parole professionnelle, de parole préparée et de lecture. Si cette difficulté n'a pas été mise en avant dans le projet LeBenchmark, les travaux de recherche de cette thèse prennent néanmoins tout leur sens ici. Nous espérons en effet pouvoir aider à une meilleure compréhension de la parole spontanée et de ses effets sur les systèmes de reconnaissance automatique de la parole.

Chapitre VI

Dimensions de variation de la parole spontanée et étiquetage de données

Afin de pouvoir étudier la parole spontanée dans sa variété, nous avons fait le choix de nous tourner vers de nouveaux corpus, en écartant dans ce travail les corpus de référence utilisés en reconnaissance automatique de la parole. Ainsi, nous présentons dans ce chapitre les corpus librement accessibles issus de la linguistique, représentatifs d'un grand nombre de situations de parole spontanée différentes, que nous avons pu trouver. Ces corpus nous servent tout d'abord à mettre en parallèle les facteurs de variation de la parole spontanée relevés dans la littérature avec les métadonnées associées aux différents corpus étudiés. Ceci nous permet de déterminer un nombre restreint de dimensions de variation de la spontanéité et d'étiqueter, dans un deuxième temps, les données rassemblées afin de construire des cas d'étude représentatifs de différents sous-types de parole spontanée. La constitution de cas d'études est contrainte par l'étiquetage selon les dimensions de la parole spontanée et donc par la richesse des métadonnées qui les accompagnent, mais aussi par la qualité des données (non bruitées, nombre restreint de locuteurs). Pour rappel, nous cherchons à réunir *a minima* 10 heures de parole pour les ensembles d'apprentissage de chacun des cas que nous étudierons. Les données non étiquetées serviront à l'élaboration d'un grand ensemble de parole spontanée, et sera utilisé pour l'adaptation d'un système de reconnaissance automatique de la parole au domaine spontané.

Déroulement du chapitre

CHAPITRE VI : *Dimensions de variation de la parole spontanée et étiquetage de données*

1.1	Les données issues du Corpus d'Étude pour le Français Contemporain (CEFC)	119
1.2	Autres corpus rassemblés	121
2	Dimensions de variation de la parole spontanée	123
2.1	Analyse des <i>data papers</i>	124
2.1.1	Le rôle des locuteurs	125
2.1.2	Le degré d'intimité entre les locuteurs	126
2.1.3	Le thème de la discussion	127
2.1.4	Les émotions	128
2.1.5	Le lieu	129
2.1.6	Le nombre d'interlocuteurs	130
2.1.7	Le canal de communication et la présence physique des locuteurs	130
2.1.8	La présence d'un auditoire	130
2.1.9	Discussion	131
2.2	Détermination de niveaux de parole spontanée pour la reconnaissance automatique de la parole : une méthode en quatre dimensions	132
3	Préparation des données pour la reconnaissance automatique de la parole	135
3.1	Tri gros grain des données finales	135
3.2	Normalisation des transcriptions et fichiers audio	138
3.3	Étiquetage des données pour la création d'ensembles représentatifs de sous-types de parole spontanée	139
4	Synthèse	143

1 La collection de corpus de parole spontanée

Afin de pouvoir rassembler des corpus et créer des ensembles de données de parole spontanée “tout-venant” et des ensembles représentatifs de sous-types de parole spontanée, nous avons commencé par rechercher et télécharger des corpus de parole spontanée. Nous présentons chacun d’entre eux ci-après, en commençant par les données CEFC qui lui-même rassemble de nombreux corpus.

1.1 Les données issues du Corpus d'Étude pour le Français Contemporain (CEFC)

Le Corpus d'Étude pour le Français Contemporain est issu du projet ORFEO (Benzitoun *et al.*, 2016) qui a pour but de mener des études comparatives sur des données de genres variés (notamment écrit vs oral) sur un corpus unifié et outillé. Les auteurs ont réuni plusieurs corpus existants libres de droits ou mis à disposition par les ayants droits afin de constituer une base de données diversifiée. Un travail d'harmonisation des transcriptions et d'alignement texte/son a été effectué. Le CEFC réunit tout ou une partie des corpus de parole spontanée suivants dans sa partie orale :

CFPP - Corpus de Français Parlé Parisien (Branca-Rosoff *et al.*, 2012) : Ce corpus a été développé à des fins d'étude du français de Paris et de sa banlieue limitrophe, et plus spécifiquement du français “de communication” qui apparaît lors d'interviews conversationnelles. Les habitants ont été interrogés chacun pendant environ une heure sur leur vie et les rapports qu'ils entretiennent à leur quartier. Les auteurs qualifient le type de parole de “commun” ou encore “d'oralité ordinaire” comprenant des “façons de dire familières” et le placent entre le vernaculaire¹ et le français standard “décrit dans les grammaires” [p.18]. Ils précisent également que la parole est variée en ceci qu'elle inclut “des variations, en fonction des appartенноances sociales des habitants interrogés, des activités discursives des interlocuteurs et des moments de la conversation” [p.3].

CFPB - Corpus du Français Parlé à Bruxelles (Dister et Labeau, 2017) : Le CFPB reprend le principe du corpus CFPP, mais se concentre sur Bruxelles. Ainsi, les participants sont interrogés sur leur rapport à leur quartier, leur commune et leur ville. On y retrouve des entretiens laboviens semi-dirigés, des dialogues “assez formels étant donné le contexte de l'interview” et des multilogues, comprenant “des échanges spontanés entre informateurs qui se connaissent et sont susceptibles d'entrer dans des échanges naturels” [p.11].

CLAPI - Corpus de Langue PArlée en Interaction (Balthasar et Bert, 2005; Baldauf-Quilliatre *et al.*, 2016) : À la base de CLAPI se trouve la volonté de la part des auteurs d'archiver, préserver et mettre à disposition les corpus régulièrement développés au sein du laboratoire ICAR. C'est ainsi que la base de données a été lancée dès la fin des années 1990. Les données recueillies se concentrent sur la parole en interaction. Les différents corpus disponibles sont ainsi représentatifs d'une grande variété de situations sociales (réunions, interactions en site commercial, repas en famille et entre amis, consultations médicales *etc.*),

1. Langue communément parlée dans les limites d'une communauté (CNRTL, consulté le 28/02/2024)

et majoritairement écologiques, c'est-à-dire que les situations et données qui en résultent ne sont pas créées par les chercheurs, mais que les situations sont -de base- existantes et capturées pour en étudier les interactions.

C-ORAL-ROM (Cresti et al., 2004) : Le corpus C-ORAL-ROM est une collection de corpus de parole spontanée en français, italien, portugais et espagnol. Le but des créateurs était de rassembler une certaine variété d'actes de parole présents dans la parole de tous les jours et d'en étudier les structures syntaxiques et prosodiques. On y retrouve à la fois de la parole formelle et informelle, dans des contextes privés et publics ainsi que de la parole téléphonique, médiatique et produite en contexte dit "naturel" [p.576].

CRFP - Corpus de Référence du Français Parlé (Equipe Delic et al., 2004) : Les auteurs ont souhaité, via la création du CRFP, mettre à disposition de la communauté des linguistiques, chercheurs et enseignants, un témoignage de la langue française alors parlée dans les principales villes de France. Le type de parole visé était un français "d'usage général et courant" ce qui a obligé les auteurs à faire "des choix touchant aussi bien aux caractéristiques des locuteurs qu'aux situations de parole" (privée, professionnelle, publique) [p. 12].

FLEURON (André, 2016; André, 2017) : FLEURON est un site internet dédié à l'apprentissage du français et destiné à des étudiants qui souhaitent faire un séjour universitaire en France. Il leur permet de se préparer aux différentes interactions liées à la vie universitaire qu'ils pourront rencontrer. Ainsi, c'est un véritable corpus multimédia qui est mis à leur disposition, composé d'interactions authentiques se déroulant à la scolarité, à la bibliothèque universitaire, au restaurant universitaire mais aussi à la banque ou encore au guichet de vente de billets de transport en commun.

OFROM - Corpus Oral de Français de Suisse Romande (Avanzi et al., 2016) : Le corpus regroupe des enregistrements de français parlé en Suisse romande. Ces enregistrements sont majoritairement des entretiens guidés. On y retrouve également des interactions de deux personnes parlant "à bâtons rompus". Plusieurs thèmes de discussion sont représentés comme les métiers, les voyages, les passe-temps des locuteurs, leurs relations de voisinage...

Réunions de travail (Husianycia, 2013) : Ce corpus regroupe des interactions sur le lieu de travail : réunions, séances de travail dans le milieu associatif et conversations entre collègues avant les réunions. Il a été développé par Magali Husianycia (ATILF) dans le cadre d'un travail de thèse.

TCOF - Traitement des Corpus Oraux en Français (André et Canut, 2010) : Le projet TCOF a vu le jour suite à la volonté d'un groupe de linguistes de

partager à la communauté les différents corpus de données orales développés depuis des années dans deux domaines de recherches : l'analyse syntaxique des productions orales d'adultes en français parlé et l'analyse des interactions entre adultes et enfants âgés de 2 à 7 ans en linguistique de l'acquisition. Il vise notamment à la description de la langue orale, l'étude des processus interactionnels qui régissent les pratiques langagières mais aussi les processus interactionnels d'apprentissage au cours de la période d'acquisition du langage. Dans la partie "Adultes" du corpus, on retrouve des interactions sollicitées (entretiens sur des récits de vie, d'expériences ou explication d'un savoir-faire professionnel, et conversations à bâtons rompus ou portant sur des thématiques spécifiques) et non sollicitées (situations publiques ou professionnelles : réunions, activités professionnelles).

TUFS - Tokyo University of Foreign Studies : Le corpus TUFS, dont la création a été coordonnée par Y. Kawaguchi (Tokyo University of Foreign Studies), rassemble des enregistrements d'étudiants des universités Aix-Marseille et Paris XIII. Les enregistrements sont plutôt longs ce qui favorise la production de parole de plus en plus spontanée (Bérard, 2020).

Valibel (Dister *et al.*, 2009) : Valibel est un regroupement de corpus créés depuis 1987. La base de données est ouverte (de nouvelles données ou annotations peuvent être ajoutées) et regroupe aujourd'hui plus de 300 heures d'enregistrements (seules quelques-unes sont incluses dans le CEFC). Bolly *et al.* (2016) rapporte 24 corpus exploitables. Parmi les corpus disponibles, il est possible de retrouver des enquêtes sur l'accent, sur les représentations linguistiques, sur la liaison *etc.*

French Oral Narrative Corpus² : Ce corpus réunit un ensemble de contes présentés à un public d'adultes ou d'enfants (87 histoires, 18 conteurs).

La figure VI.1 présente la proportion de chacun de ces corpus dans le CEFC.

Nous avons cherché à compléter les corpus présents dans le CEFC en téléchargeant de nouveaux enregistrements. Ainsi, nous avons pu compléter les données des corpus CFPP, CFPB, TCOF, CLAPI et Valibel. Les enregistrements supplémentaires ont pu être trouvés sur Ortolang ou encore directement sur les sites internet des différentes projets.

1.2 Autres corpus rassemblés

Afin de compléter les données des corpus présents dans le CEFC, nous avons cherché d'autres corpus de parole spontanée. Cette recherche nous a permis d'ajouter

2. Nous avons fait le choix de l'écartier, la représentation de contes étant trop éloignée de la parole spontanée par ses aspects préparé et acté.

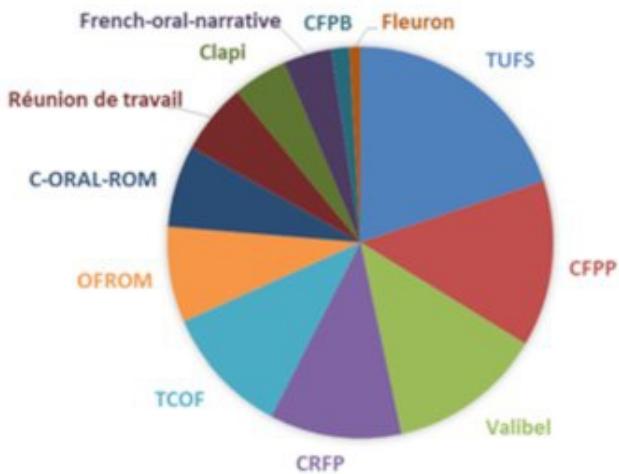


FIGURE VI.1 – Proportion en nombre de mots de chacun des corpus présentés dans le CEFC (graphique tiré de Bérard (2020))

les corpus suivants :

CID - Corpus of Interactional Data (Bertrand *et al.*, 2008) : Le CID est un corpus audio-vidéo d’interactions. Il est annoté aux niveaux phonétique, prosodique, syntaxique, discursif et mimo-gestuel. Il contient deux types d’enregistrements : l’évocation de conflits professionnels puis de situations insolites. Les participants étaient invités à s’écarter des consignes s’ils le souhaitaient.

ESLO - Enquêtes SocioLinguistiques à Orléans (Eshkol-Taravella *et al.*, 2011; Baude et Dugua, 2016) : Le corpus ESLO a été composé en deux temps : ESLO1 (1968-1971) et ESLO2 (2008-2012). La première partie a été dirigée par des professeurs de français de l’université d’Essex au Royaume-Uni. Le but était de construire le portrait sonore de la ville d’Orléans. Quarante ans plus tard, le Laboratoire Ligérien de Linguistique de l’université d’Orléans se fixe comme objectifs de constituer un nouveau corpus du français parlé à Orléans et de diffuser largement ESLO2. On retrouve dans ces corpus des entretiens face-à-face, des enregistrements libres dans des situations privées ou professionnelles, des interviews de personnalités, des conférences-débats... Il permet des analyses sur la variation socio-linguistique en français.

MPF - Multicultural Paris French (Gadet et Guerin, 2016) : Le corpus MPF visait à recueillir des enregistrements de populations “jeunes” connaissant des contacts multiculturels réguliers afin d’étudier les effets sur le français du contact avec les langues de l’immigration. Les locuteurs de chacun des enregistrements se connaissaient en amont du recueil afin de respecter le souhait de proximité com-

municative des auteurs. Le corpus comprend à la fois des entretiens traditionnels, entretiens de proximité et événements écologiques (sans enquêteur).

PFC - Phonologie du Français Contemporain (Durand *et al.*, 2002) : Le projet PFC avait pour but de permettre la conservation du patrimoine linguistique des espaces francophones du monde, mais aussi de proposer des analyses phonologiques de certains phénomènes linguistiques. Les témoins ont participé à quatre types d'enregistrements : lecture d'une liste de mots, lecture d'un texte, conversation semi-dirigée en face-à-face et conversation spontanée ou libre entre pairs. Ceux-ci ont été effectués soit chez les témoins, soit dans un lieu “neutre, comme à l'université” [p.98]. Pour chaque enquête, il était recommandé aux enquêteurs de former un duo au sein duquel l'un des enquêteurs est familier du témoin, et mène alors la conversation spontanée pour accéder au vernaculaire de ce dernier, tandis que l'autre était en charge de la conversation semi-dirigée.

Nous avons donc à disposition les données de 15 corpus se différenciant par le choix des participants, leurs rôles respectifs, la méthode d'enregistrement, la diversité des situations de parole capturées ou encore le lieu d'enregistrement. Le tableau VI.1 récapitule les corpus récoltés et y associe leur licence d'utilisation, ainsi que le nombre d'heures d'enregistrement approximatif qu'ils représentent. Les différents corpus collectés rassemblent ainsi près de 1 000 heures d'enregistrements de parole spontanée, représentant des situations très variées qu'il nous semble important de mieux caractériser.

2 Dimensions de variation de la parole spontanée

Nous avons présenté dans l'état de l'art sur la parole spontanée différents facteurs à l'origine de la production de différentes formes de parole spontanée, dont certains se retrouvent capturés dans les dimensions relatives à la variation stylistique également recensées. Nous proposons d'analyser les *data papers* des corpus présentés ci-avant afin d'étudier ce qui est relaté par les auteurs comme ayant une incidence sur la spontanéité de la parole, et de croiser ces facteurs avec ceux que nous avons pu trouver lors de l'élaboration de l'état de l'art sur la parole spontanée. Nous en faisons ensuite une synthèse qui nous permet de proposer une méthode de catégorisation d'un enregistrement comme plus ou moins spontané, applicable dans le domaine de la reconnaissance automatique de la parole.

Corpus	Licence	Heures d'enregistrement*
Issus du corpus CEFC v1.5		
CFPP	CC-BY-NC-SA 4.0	38 h
CFPB	CC-BY-NC-SA 4.0	5 h
CLAPI	CC-BY-NC-SA 4.0	17 h
C-ORAL-ROM	CC-BY-NC-SA 4.0	22 h
CRFP	CC-BY-NC-SA 4.0	34 h
FLEURON	CC-BY-NC-SA 4.0	3 h
FROM	CC-BY-NC-SA 4.0	25 h
Réunions	CC-BY-NC-SA 4.0	18 h
TCOF	CC-BY-NC-SA 4.0	6 h
TUFS	CC-BY-NC-SA 4.0	53 h
Valibel	CC-BY-NC-SA 4.0	4 h
Total		225 h
Corpus indépendants et complétés		
CFPP	CC-BY-NC-SA 4.0	20 h
CFPB	CC-BY-NC-SA 3.0	9 h
CID	CC-BY-NC-SA 4.0	16 h
CLAPI*	CC-BY-NC-SA 4.0	5 h
ESLO2	CC-BY-NC-SA	117 h
MPF	CC-BY-NC-SA	47 h
NCCF	Convention	35 h
PFC	CC-BY-NC	113h
TCOF Adultes	CC BY-NC-SA	62 h
Valibel	Convention	318 h
Total		742 h
Total global		967 h

*durée approximative

TABLE VI.1 – Récapitulatif de l'ensemble des corpus rassemblés étudiés par la suite

2.1 Analyse des *data papers*

L'état de l'art nous a permis de comprendre que les différents niveaux de spontanéité dépendent à la fois de facteurs sociaux (rôle, nombre d'interlocuteurs, degré d'intimité), environnementaux (lieu, présence physique ou non des locuteurs, canal de communication, présence d'un auditoire) mais aussi des émotions et de la proximité des locuteurs avec le thème de la discussion. Nous présentons ainsi les

différents éléments des *data papers* se rapportant à chacun de ces facteurs.

2.1.1 Le rôle des locuteurs

Parmi les rôles, nous avons retrouvé dans les corpus CFPP, OFROM, MPF, CFPB et PFC ceux d'intervieweur et interviewé parfois également appelés enquêteur et enquêté/informateur. Le rôle correspond ici à un statut à un moment donné. Ainsi, un même locuteur peut très bien avoir un rôle précis dans un enregistrement et avoir un autre rôle dans une autre situation, comme cela est le cas dans le PFC ainsi que l'indiquent les auteurs : “Ce qui différencie la discussion libre et la discussion guidée c'est que dans le premier cas, il n'y a aucune asymétrie de rôle. L'enquêteur, s'il est présent, est un simple membre du groupe. Dans le deuxième cas, c'est un enquêteur.” (Durand *et al.*, 2008, p.10). Les auteurs du corpus CID caractérisent également ce premier cas comme une “symétrie de statut et de place”.

Dans le NCCF, c'est le rôle de complice qui est présenté, impliquant qu'une mission spécifique a été donnée à une personne parmi celles enregistrées (dans ce cas, celle de relancer la conversation sur un sujet familier lorsque celle-ci s'essouffle). Les auteurs du CFPP précisent qu'un effet mimétique automatique est régulièrement observé dans les enquêtes : les enquêtés tendent alors à se rapprocher de la variété linguistique employée par l'enquêteur. Ainsi, si l'enquêteur fait usage d'un style détendu, le registre de parole peut passer de formel à un style plus “vernaculaire”.

Les auteurs du CFPP mentionnent également l'existence de deux types de rôles chez les enquêteurs : celui qu'ils s'attribuent et celui qu'on leur attribue dans l'interaction, leur accordant ainsi le statut de *figure normative*³. Dans FLEURON, ce sont des interactions avec des membres du personnel administratif de l'université qui sont enregistrées. Ce statut implique que l'échange entre les locuteurs est contraint par des normes sociales telles que le respect de la politesse, le vouvoiement ou encore l'utilisation d'un vocabulaire plutôt soutenu. Ces normes sociales sont définies par (Becker et Chapoulie, 1985, p. 25) de la façon suivante : “Tous les groupes sociaux instituent des normes et s'efforcent de les faire appliquer, au moins à certains moments et dans certaines circonstances. Les normes sociales définissent des situations et les modes de comportement appropriés à celles-ci : certaines actions sont prescrites (ce qui est «bien»), d'autres sont interdites (ce qui est «mal»). Quand un individu est supposé avoir transgressé une norme en vigueur, il peut se faire qu'il soit perçu comme un type particulier d'individu, auquel on ne peut faire confiance pour vivre selon les normes sur lesquelles s'accorde le groupe. Cet individu est considéré comme étranger au groupe”. La notion de style de parole plus ou moins approprié est relevée par les créateurs de FLEURON

3. dans ce cas représentée par les enquêteurs universitaires

qui précisent que “pour être compétent dans une langue, il fait savoir produire les bonnes pratiques langagières de façon appropriée” (André, 2017, p. 299). Les auteurs du CFPP précisent que le registre oral adopté lors de l’enquête est celui qui est considéré comme *approprié* dans le contexte spécifique de la conversation entre personnes qui n’appartiennent pas au même espace intime, amical ou professionnel.

2.1.2 Le degré d’intimité entre les locuteurs

Le rapport qu’ont les locuteurs entre eux et l’ensemble des éléments s’y rapportant sont réunis sous la locution “degré d’intimité”. Les locuteurs peuvent aussi bien être des inconnus les uns pour les autres (ESLO), des amis d’amis comme dans le CFPP, des collègues comme dans le CID, des voisins (CLAPI), de bons amis (NCCF), de la famille (ESLO, CLAPI) ou encore des personnes se connaissant de façon intime⁴ (PFC). Ces différentes relations, que nous avons présentées de la moins intime à la plus intime, ont une influence notamment sur le niveau de formalité de la parole produite, tel que noté dans le PFC et le CFPB. Ainsi, les auteurs du PFC précisent que lorsqu’enquêteurs et enquêtés ne se connaissent pas en amont de l’enregistrement, alors la production d’un registre véritablement informel demanderait de poursuivre l’enregistrement plusieurs heures. Si aucune explication complémentaire n’est donnée, nous pouvons en déduire qu’un allongement de la session d’enregistrement permettrait aux personnes en interaction de mieux en mieux se connaître. Les auteurs du CFPB rapportent une certaine spontanéité lorsque les locuteurs se connaissent, et qualifient alors l’échange de “naturel”. Les auteurs du PFC, quant à eux, suivent les recommandations données dans Bourdieu (1993) et considèrent que la proximité entre les locuteurs permet de réduire le paradoxe de l’observateur et d’enregistrer une parole naturelle. Ce paradoxe est défini par Labov (1973) comme le fait de devoir observer la façon dont les locuteurs parlent quand ils ne sont pas observés pour obtenir des données fondamentales pour la théorie linguistique, alors que ces mêmes locuteurs changent leurs pratiques langagières lorsqu’ils sont observés.

Si nous reprenons ce qui a été relevé dans le CFPP dans la section précédente, les locuteurs tendent à respecter des normes sociales et ainsi un comportement approprié selon des contextes spécifiques. Ainsi, lorsqu’ils n’appartiennent pas à des espaces intimes, amicaux ou quotidiens partagés et qu’ils communiquent dans le cadre d’une enquête/interview, l’utilisation du vouvoiement, le respect de la politesse et l’utilisation d’une parole peu familière est de rigueur. Nos recherches complémentaires sur les changements qu’impliquent la relation entre les locuteurs sur le discours nous montre que plus les locuteurs sont dans une relation intime,

4. Dans le sens “Qui est très proche, profondément lié par la parenté, l’affection.” (CNRTL, consulté le 8/03/2024)

2 Dimensions de variation de la parole spontanée

plus la parole sera informelle. Ainsi, plus le degré d'intimité entre les locuteurs est fort, moins ceux-ci respecteront les normes sociales liées au contexte de l'entretien. Romera Ciria (2019) parle d'*habitus* interactionnel, concept que Pierre Bourdieu utilise “pour rendre compte de l'ajustement qui s'opère le plus souvent « spontanément », c'est-à-dire sans calcul ni intention expresse, entre les contraintes qui s'imposent objectivement aux agents, et leurs espérances ou aspirations subjectives” (Wagner, 2012, p. 1). Dans FLEURON par exemple, il est précisé que la relation entre les locuteurs peut conditionner leur tutoiement.

Néanmoins, il est important de préciser que la relation entre des locuteurs n'est pas, de fait, quelque chose d'inaliénable. Ainsi, tel que le précisent les auteurs de FLEURON, “la situation peut changer suite à des paroles injurieuses des locuteurs qui peuvent décider de rompre leur relation qui peut donc passer d'amicale à inamicale” (André, 2017, p. 299). Ceci est particulièrement valable dans le cadre des interactions conversationnelles où l'objectif est qualifié d'interne car centré sur la relation, au contraire des interactions produites en réponse à une consigne alors centrées sur cette consigne et dont l'objectif est qualifié d'externe tel que précisé dans le CID. Romera Ciria (2019) précise que les liens amicaux, notamment, sont les garants d’“expériences partagées par des partenaires qui créent entre eux un “code culturel”. Pour la constitution du CID, les auteurs ont ainsi choisi les locuteurs “en fonction de leur degré de familiarité et de leur habitude à converser ensemble”. Ils précisent que cela leur assure des échanges plus spontanés. C'est ainsi également que dans le MPF, les auteurs distinguent les entretiens traditionnels des entretiens de proximité dans lesquels l'enquêteur a réussi à établir une “interaction connivente”, c'est-à-dire qu'une complicité entre lui et le(s) locuteur(s) est apparue. Cette connivence peut apparaître lorsqu'une passion commune est abordée, et représente un facteur de l'évaluation de la proximité de communication entre les interactants précisée dans les métadonnées. Plus les locuteurs partagent une connivence forte, plus l'immédiat communicatif est authentique Koch et Oesterreicher (2001). À l'inverse, le fait d'aborder un sujet non connu par l'un des locuteurs augmente la distance communicationnelle. Les auteurs du MPF précisent que cette distance peut apparaître au sein même d'un échange entre des personnes entretenant un lien très familier. Ainsi, la proximité communicationnelle dépend à la fois “de la connivence et du partage d'un ensemble dense de savoirs et d'expériences à l'origine de nombreux implicites” [p.5].

2.1.3 Le thème de la discussion

L'analyse des différentes publications présentant les corpus de parole spontanée récoltés nous montre, de façon très logique, que les échanges peuvent se rapporter à un vaste ensemble de thèmes :

- pour le PFC : viticulture, traditions locales, conditions de travail, gastronomie, maladies, décès, scènes de vie, voyages, activités, travail, projets, enfance, questions d'actus...
- pour le CFPP et le CFPB : relation des habitants à leur quartier, leur commune et leur ville
- pour le TCOF : récits de vie, récits d'expériences ou d'événements, explications sur savoir-faire professionnel *etc.*
- pour OFROM : métiers, voyages, passe-temps, relations de voisinage, projets ou situations incongrues, système politique, situation linguistique de la Suisse...
- pour le NCCF : examens, grève, fêtes, voyages...

Il est important de préciser que si le thème de la discussion peut assez aisément être rapporté lorsque les entretiens sont sollicités et guidés, la tâche se complexifie largement dans le cadre d'une conversation à bâtons rompus. Si le partage de la connaissance d'un des thèmes abordés au cours de l'enregistrement entre les locuteurs semble important comme nous l'avons relevé auparavant, nous n'avons pas pu trouver de corpus autre que le MPF qui relève cette information.

2.1.4 Les émotions

La mention de l'existence d'échange naturel des opinions entre intervieweur et interviewé reporté dans le CFPP est associée à la précision suivante : “l'enquêteur a eu soin de ne pas heurter trop ses partenaires” [p.4]. Cette consigne n'est pas sans rappeler ce que nous avons rapporté de FLEURON précédemment : “la situation peut changer suite à des paroles injurieuses des locuteurs qui peuvent décider de rompre leur relation qui peut donc passer d'amicale à inamicale” (André, 2017, p. 299). Ainsi, si une relation très intime se fonde plutôt sur une relation solide et de longue date difficile à briser, le lien nouveau qui se crée entre un intervieweur et un interviewé qui ne se connaissaient pas en amont est plus fragile et plus susceptible de se rompre face à l'immédiateté des émotions, notamment négatives, traversées par les locuteurs. Les auteurs du CFPP mentionnent également l'importance de la construction d'une relation de confiance entre les enquêteurs et enquêtés. La notion de confiance est catégorisée comme une émotion positive dans le modèle de Russell (1980), régulièrement utilisé dans le domaine de la reconnaissance automatique des émotions et présenté en figure VI.2.

Si les émotions traversées par les locuteurs ne sont jamais renseignées dans les corpus étudiés ici⁵, leur importance est tout de même bien relevée. Le “contrô-

5. cette catégorisation est très difficile à mettre en place et nécessite de faire appel aux locuteurs eux-mêmes afin qu'ils précisent les émotions par lesquelles ils ont été traversés en un temps

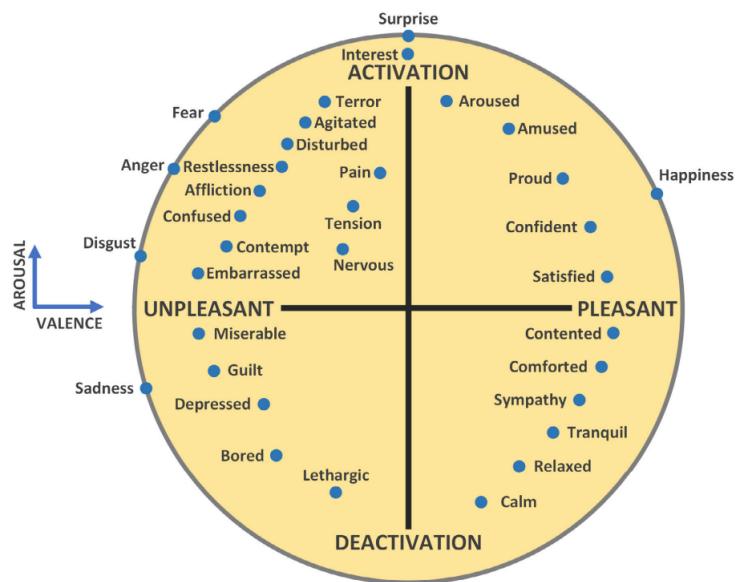


FIGURE VI.2 – Le “model of affect” de Russell (image tirée de Gonzalez *et al.* (2020))

le” des émotions repose alors sur des facteurs extérieurs au locuteur interviewé. C'est en effet l'intervieweur qui s'adapte, ou tout le protocole qui est revu afin de mettre les locuteurs dans de bonnes conditions. Ainsi, pour le corpus CID, cela passe par la sélection de locuteurs travaillant sur le lieu de l'enregistrement afin que l'environnement leur paraisse familier et éviter un stress et un embarras trop importants.

2.1.5 Le lieu

Le lieu physique⁶ influence le type de parole produit. Dans le PFC, il est indiqué que l'enquêteur a échangé avec les témoins à la fois sur leur lieu d'habitation et dans des lieux neutres⁷ comme l'université afin d'obtenir des enregistrements plus formels.

L'analyse des articles nous a permis de relever différents lieux tels que les lieux publics (magasins, marchés, ateliers...) dans ESLO, différents types d'institutions, de services publics ou d'entreprises privées (poste, mairies, études notariales, commerces, cabinet, hôpital) dans CLAPI et le lieu de travail pour le CRFP, le CID

donné

6. à distinguer du lieu géographique qui serait précisé par le nom d'un pays ou d'une ville par exemple

7. qualifié comme tel par les auteurs

et CLAPI. Les auteurs du CRFP qualifient la parole utilisée par les locuteurs sur leur lieu de travail comme plus institutionnelle que lors des interactions sur un lieu privé (le même sujet de discussion étant utilisé dans les deux cas). Enfin, dans le corpus FLEURON, les lieux amènent à des interactions différentes, produites en contexte par des locuteurs réels, compétence que devront développer les étudiants étrangers. Ainsi, on retrouve dans ce corpus des enregistrements réalisés à la scolarité, à la bibliothèque universitaire, au restaurant universitaire, mais aussi à la banque ou au guichet de vente de billet de transport en commun, à la préfecture, ou dans des lieux de culture...

2.1.6 Le nombre d'interlocuteurs

Le nombre d'interlocuteurs est variable selon les enregistrements. Ainsi, il est possible de retrouver des échanges impliquant deux locuteurs, trois locuteurs ou un groupe restreint. Néanmoins, comme stipulé par les auteurs du PFC, les échanges à deux ou trois locuteurs sont souvent privilégiés afin d'éviter les phénomènes de schismes interactionnels⁸ qu'ils qualifient de “très brouillés au plan phonétique” et sont donc difficiles à transcrire (Laks *et al.*, 2009, p. 9). Les échanges avec un groupe se retrouvent plutôt dans des contextes professionnels tels que les réunions de travail ou associatives dans le CRFP.

2.1.7 Le canal de communication et la présence physique des locuteurs

La présence physique ou non des locuteurs étant déterminée par le canal de communication, nous avons réuni ces deux catégories. Le canal de communication est la plupart du temps précisé dans les publications. Celui-ci peut référer à une interaction en face-à-face comme dans le PFC, à des communications téléphoniques (dans ESLO et CLAPI) ou encore à des conversations en ligne (dans CLAPI). Les interactions en face-à-face et les interactions téléphoniques et donc distantes sont les plus courantes. Les communications téléphoniques et conversations en ligne sont toutes deux des communications distantes mais diffèrent sur un point : l'utilisation de la vidéo qui permet aux interlocuteurs de capturer les gestes qui appuient la parole.

2.1.8 La présence d'un auditoire

L'auditoire est défini dans le CNRTL comme une “assemblée de ceux qui écoutent quelqu'un ou écoutent quelque chose présenté ou exécuté par quelqu'un : un orateur (un discours, un sermon), un professeur (un cours), une émission radiophonique,

8. Cette locution qualifie les multiples conversations en parallèle qui peuvent survenir notamment lorsqu'il y a plusieurs locuteurs.

une plaidoirie, etc.”. Celui-ci ne suppose donc pas de moment d’interaction entre les locuteurs et les auditeurs. Un exemple de ce type est les émissions de radio incluses dans ESLO. Cette absence d’interaction amène le locuteur à bien articuler sa parole afin d’être compris de tous les auditeurs.

2.1.9 Discussion

Nous pouvons ressortir de l’analyse de ces *data papers* la dépendance de certains facteurs les uns avec les autres. Effectivement, il paraît difficile de réussir à étudier les différents types de la parole spontanée si l’on ne prend pas en compte la relation intime partagée entre les locuteurs. Cette dernière conditionne effectivement les pratiques langagières qui régissent les échanges.

La force de la relation qui se construit entre deux locuteurs qui ne se connaissaient pas en amont d’un enregistrement semble largement dépendre du thème de la discussion ainsi que du type d’émotions ressenties, permettant un certain niveau de connivence entre les locuteurs. Une relation peu solide peut être rapidement rompue suite à des paroles déplacées. Les émotions négatives sont alors une des conséquences directes de propos inappropriés et peuvent entraîner une rupture ou du moins un changement dans la relation partagée par les locuteurs. Ces situations sont néanmoins précautionneusement évitées lors de la constitution des corpus. Les aspects émotionnels peuvent ainsi modifier momentanément la force de la relation établie par le degré d’intimité entre les locuteurs. Toutefois, il nous semble important de noter que la variété d’émotions pouvant être traversées par un locuteur au cours d’un enregistrement ne peut que difficilement être observée. Le plus souvent, des précautions sont ainsi prises au niveau du protocole d’enregistrement. Nous écarterons donc cet aspect émotionnel de la suite de nos travaux.

Le thème de la discussion est également un facteur compliqué à considérer pour l’analyse de la spontanéité. Celui-ci peut être contrôlé en amont de l’échange, notamment grâce à un guide d’entretien fourni aux enquêteurs comme dans le CFPP et le CFPB ayant pour but de recueillir des témoignages d’habitants sur leur relation à leur quartier, commune ou ville. Néanmoins, lors d’une conversation à bâtons rompus, les locuteurs peuvent aborder une variété de thèmes impressionnante : métiers, voyages, maladies, récits de vie, examens etc. Le détail des thèmes n’est, dans de tels cas, pas toujours annoté. Il est donc difficile d’établir un lien entre le thème et la parole spontanée. Pour cette raison, nous écarterons également cet aspect de la suite de nos travaux.

Enfin, nous pouvons également noter le lien fort qui existe entre certains rôles et le lieu de l’interaction. C’est notamment le cas pour les échanges enregistrés à l’université avec le personnel administratif. Ce n’est pas le lieu en lui-même qui

contraint la communication⁹, mais les normes sociales de l’interaction avec des personnes représentant une institution. Néanmoins, c’est parce que ces personnes se trouvent sur leur lieu de travail et donc au sein de l’institution au moment de l’interaction qu’elles ont ce statut particulier, et c’est ce qui fait que le lieu et le rôle peuvent partager un lien fort.

2.2 Détermination de niveaux de parole spontanée pour la reconnaissance automatique de la parole : une méthode en quatre dimensions

La section précédente nous aura donné un aperçu des différents facteurs de variation de la parole spontanée ainsi que des éléments que ces facteurs comprennent. Nous les récapitulons dans le tableau VI.2.

Facteur	Eléments
Rôle	symétrie de statut et de place, enquêteur/enquêté, complice, figure normative, personnel administratif
Lieu	lieu d’habitation, lieu neutre (univ), lieux publics, institutions, services publics, entreprises privées, lieu de travail, scolarité, bibliothèque universitaire, restaurant universitaire, banque, guichet transport, préfecture, lieux de culture
Degré d’intimité entre les locuteurs	inconnus, amis d’amis, collègues, voisins, bons amis, famille, lien intime
Nombre d’interlocuteurs	deux locuteurs, trois locuteurs, un groupe restreint
Présence d’un auditoire	groupe large et distant
Canal de communication/présence des locuteurs	face-à-face, téléphone, en ligne

TABLE VI.2 – Récapitulatif des facteurs de variation de la parole spontanée et des éléments qui les constituent repérés dans les *data papers* des différents corpus récoltés

Nous avions déjà noté à la fin du chapitre de l’état de l’art sur la parole spontanée que certaines conditions semblent être plus propices à une parole très spontanée :

- la symétrie de statut et de place entre les locuteurs
- la présence de peu d’interlocuteurs
- le lien amical ou familial

9. Notez que l’université est considérée comme un lieu neutre par les créateurs du PFC.

2 Dimensions de variation de la parole spontanée

- la présence physique des locuteurs

Nous partons donc de ce constat afin de construire, à partir des éléments relevés dans les *data papers*, des dimensions avec une graduation allant du plus propice à une parole très spontanée au moins propice. Nous espérons ainsi pouvoir capturer des niveaux de spontanéité grâce à l'étiquetage de nos données.

Nous rassemblons tout d'abord les facteurs “rôle” et “lieu” en une dimension que nous appelons *situation de communication*. En effet, nous avons pu noter que ces deux facteurs pouvaient être étroitement liés et il nous semble important de les rassembler. La graduation des rôles proposée est la suivante : (i) symétrie de statut et de place entre les locuteurs, (ii) dissymétrie du rôle dans le discours ou du rôle/statut social. Par cette dernière, nous souhaitons capturer les situations imposant un certain niveau de formalité, amenant à un respect fort de la politesse et d'un discours non familier, notamment dans le cadre d'échanges entre locuteurs avec un déséquilibre hiérarchique. La graduation liée aux lieux proposée est la suivante : (i) lieux quotidiens/familiers, (ii) lieux professionnels/institutionnels. Le premier se rapporte, par exemple, à l'habitation et aux lieux publics. Le deuxième se rapporte aux institutions, services publics, entreprises, lieux de travail, banque... Ainsi, nous proposons la graduation globale suivante :

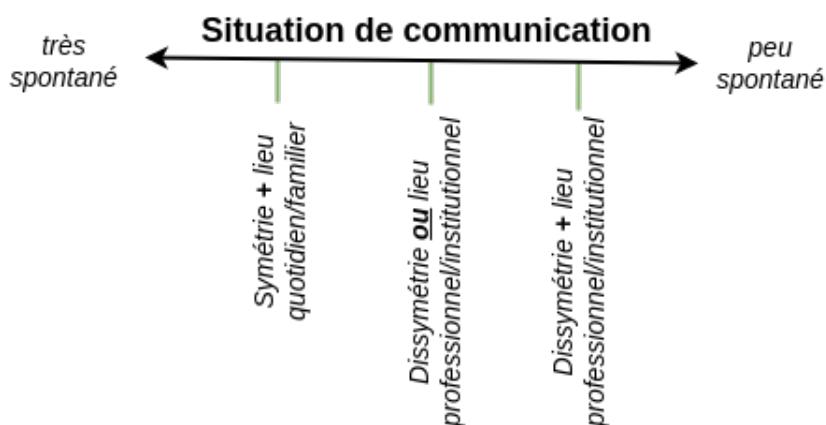


FIGURE VI.3 – Situation de communication

Le facteur *degré d'intimité entre les locuteurs* représente une dimension à part entière. Nous en proposons la graduation suivante : (i) famille/amis proches, (ii) relations amicales liées à un contexte, (iii) lien faible, (iv) inconnus ainsi qu'ilustré en figure VI.4.

La première graduation se rapporte donc aux liens solides et plutôt d'ordre intime, familiaux ou amicaux. La deuxième graduation représente les relations de type

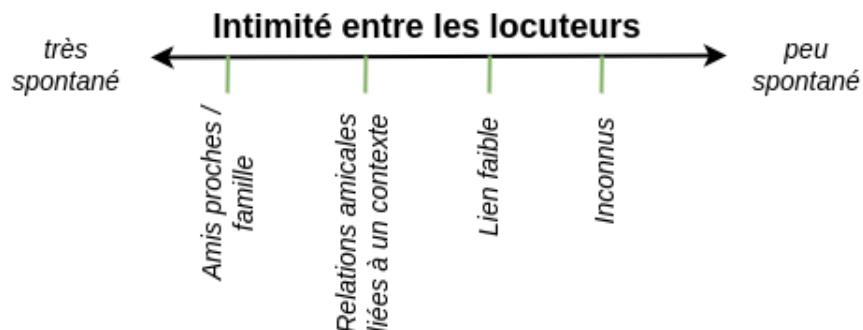


FIGURE VI.4 – Degré d'intimité entre les locuteurs

amicales liées au partage entre les locuteurs de lieux ou d'activités communes tel qu'un quartier ou le lieu de travail. La graduation "lien faible" se rapporte au lien entre deux personnes qui se connaissent très peu ou qui se rencontrent grâce à un ami en commun. Enfin, la dernière graduation représente l'absence de tout lien antérieur à l'enregistrement entre les locuteurs.

Nous proposons ensuite de rassembler les facteurs “nombre d’interlocuteurs” et “présence d’un auditoire” sous la dimension *type de communication*. En effet, nous retrouvons dans les deux la notion d’adresse et proposons donc la graduation suivante : (i) interpersonnel, (ii) de groupe, (iii) de masse. La graduation interperson-

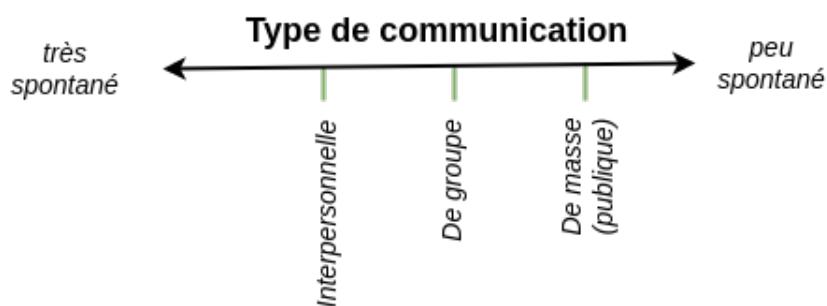


FIGURE VI.5 – Type de communication

nelle implique donc deux, voire trois personnes. La graduation de groupe implique que les interlocuteurs, parmi lesquels certains peuvent n'être qu'auditeurs, sont nombreux mais "identifiables". Par cela, nous entendons que le groupe est rassemblé dans un objectif commun et que le message est ciblé (réunion de travail, assemblée générale d'association...). Enfin, le type de communication de masse réduit voire rend inexistant le nombre d'interlocuteurs directs, mais augmente le nombre de personnes dans l'auditoire. C'est le cas des discours politiques dans un

contexte de campagne électorale par exemple, ou encore des émissions diffusées à la télévision ou la radio que l'on retrouve dans les corpus ESTER et ETAPE.

Enfin, nous regroupons sous l'appellation générale *canal de communication* les trois graduations suivantes : (i) face-à-face, (ii) distant avec visio, (iii) distant sans visio. La notion de distance qui était rattachée au facteur “présence d'un auditoire” se retrouve capturée ici.



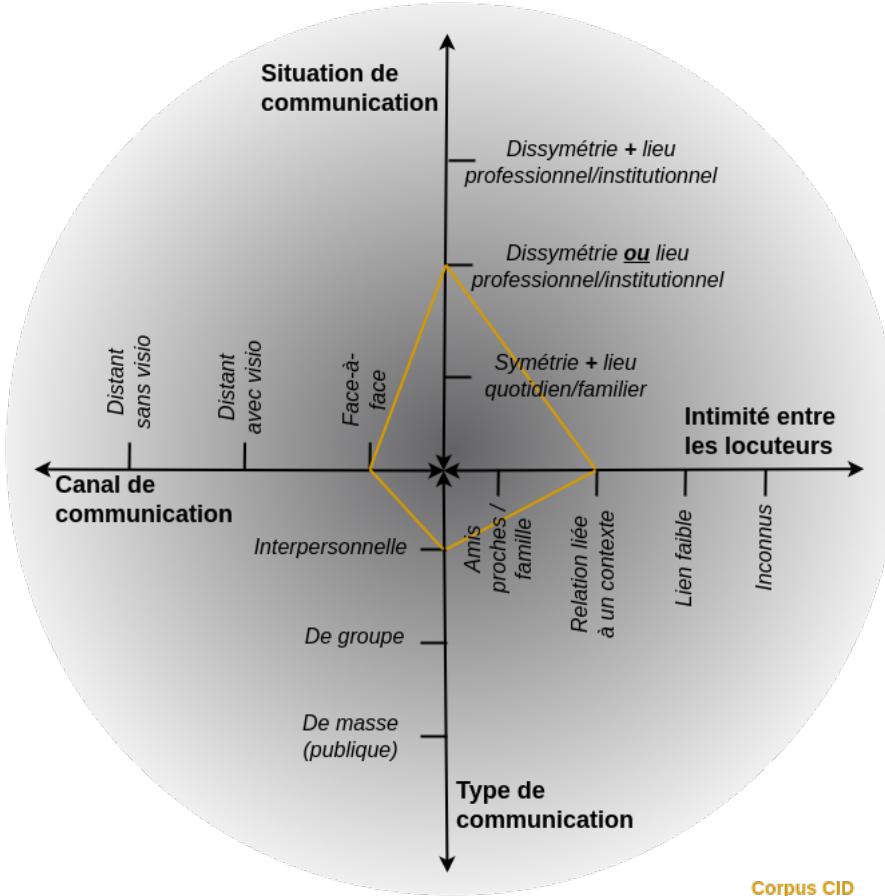
FIGURE VI.6 – Canal de communication

La figure VI.7 présente notre méthode d'étiquetage de la variation de la parole spontanée dans son ensemble. Nous y faisons figurer chacun de nos axes, le milieu de la figure représentant une situation très propice à la production d'une parole très spontanée. Ainsi, plus l'étiquetage d'un corpus sur un des axes se rapproche de l'extérieur du cercle, et donc plus l'aire du quadrilatère est grande, moins ce corpus est spontané. Un exemple d'étiquetage du corpus CID, dans lequel tous les enregistrements correspondent à une seule et même situation, est donné.

3 Préparation des données pour la reconnaissance automatique de la parole

3.1 Tri gros grain des données finales

Nous avons trié nos données et séparé les enregistrements en quatre catégories (CEFC, site internet, Ortolang, LeBenchmark) afin de faciliter la construction d'ensembles d'apprentissage, de développement et de test. En effet, certains corpus sont parfois disponibles au téléchargement à de multiples endroits : sur Ortolang au sein du CEFC, sur Ortolang en tant que corpus indépendant, et via le site internet du projet. Cette duplication des données étant rarement parfaite, une comparaison des fichiers audio inclus dans chaque catégorie nous aura parfois permis de récupérer des enregistrements supplémentaires. Cela concerne les corpus CLAPI (concerné par les trois cas), TCOF (disponible à la fois dans le CEFC et sur Ortolang seul) et Valibel (disponible au sein du CEFC et sur demande). Ensuite, nous



Plus la couleur tend vers le gris foncé (le centre), plus la situation est propice à une parole très spontanée. Plus elle tend vers le gris clair (l'extérieur du cercle), moins elle est propice à une parole spontanée.

FIGURE VI.7 – Méthode en quatre dimensions pour l'étiquetage de la variation de la parole spontanée. Exemple de l'étiquetage du corpus CID en superposition.

avons étiqueté et mis de côté les données déjà utilisées pour l'apprentissage des modèles LeBenchmark afin d'être sûrs de ne pas les retrouver dans les ensembles de développement et de test qui seront créés. Ce tri a été fait manuellement pour les corpus TCOF, Valibel, CFPP et CLAPI¹⁰.

Les données des corpus CLAPI (hors CEFC) et Valibel ne seront finalement pas utilisées pour ce travail de thèse. Les fichiers de transcription pour les données de parole spontanée que nous avons récupéré sur le site et qui ont fait l'objet d'une convention étant de type .pdf ou .doc, difficiles à parser automatiquement.

10. Les tableurs utilisés pour ces étiquetages sont à retrouver ici : <https://drive.proton.me/urls/V819T2A39W#TnnxNhQIdTkR>

3 Préparation des données pour la reconnaissance automatique de la parole

Des fichiers TEI étaient disponibles pour certains enregistrements, mais le temps demandé pour l'élaboration d'un script d'extraction des transcriptions avec leur temps de début et de fin aurait été très important pour très peu de données. Nous nous sommes donc résolus à les laisser de côté devant le temps de traitement que cela nécessitait. D'autre part, l'analyse des fichiers de métadonnées accompagnant certains des enregistrements du corpus Valibel nous aura appris qu'une très large partie comprend :

- des fichiers audio issus de la numérisation d'un enregistrement sur cassette ou VHS
- des données de télévision
- des enregistrements ne comprenant pas que du français
- des doublons de ce que l'on retrouve dans le PFC
- des doublons de ce que l'on retrouve dans le CEFC

Pour toutes ces raisons, nous avons également écarté ce corpus des données disponibles pour nos expérimentations. La liste de corpus finalement utilisables pour notre étude est récapitulée en tableau VI.3. L'ensemble de ces données représente environ 365 heures de parole effective.

Corpus
CFPP
CFPB
CID
CLAPI (CEFC)
C-ORAL-ROM
CRFP
ESLO2
FLEURON
FROM
MPF
NCCF
PFC
Réunions
TCOF
TUFS

TABLE VI.3 – Liste finale des corpus qui seront utilisés pour les expérimentations

```
“CFPB-1190-1_69.765-71.620”: {  
    “spk_id”: “Charles”,  
    “start_seg”: 69.765,  
    “end_seg”: 71.62,  
    “duration”: 1.855000000000004,  
    “spk_gender”: “U”,  
    “wav”: “/.../Dev/All_spont/CFPB-1190-1.wav”,  
    “wrд”: “J AVAIS DIX HUIT”,  
    “chars”: “J \_ A V A I S \_ D I X \_ H U I T”,  
},
```

FIGURE VI.8 – Exemple d'une entrée dans un fichier *.json*

3.2 Normalisation des transcriptions et fichiers audio

Après l'étape de téléchargement et de tri de nos données, nous avons procédé à la normalisations des transcriptions et des enregistrements.

La boîte à outil Speechbrain prend en entrée des fichiers *.json* ou *.csv*. Nous avons choisi le format *.json*, où chaque entrée correspond à un segment¹¹. Pour chaque entrée, est renseigné l'identifiant du locuteur, les temps de début et de fin ainsi que la durée du segment en secondes, le genre du locuteur, le chemin vers le fichier *.wav*, la transcription en mots normalisée et la transcription en caractères normalisée. Un exemple est donné en figure VI.8. La normalisation a consisté notamment en la suppression des tirets marquant les amorces de mots, les parenthèses, crochets ou accolades pouvant marquer des commentaires, deux options de transcription en mots et caractères, des bruits ou prononciations spécifiques, les apostrophes et la capitalisation de l'ensemble des caractères, tel qu'initialement proposé dans la recette Speechbrain.

La récupération de ces informations est faite à partir des fichiers de transcriptions, disponibles en différents formats : *.textgrid*, *.csv*, *.orfeo*, *.trs*. Les conventions de transcriptions utilisées étant différentes pour chaque corpus, nous avons normalisé les transcriptions de chacun des segments. Pour plus de facilité, nous avons utilisé le format intermédiaire *.tsv* pour le traitement de chaque corpus. Ainsi, nous avons des scripts *csv/textgrid/trs* vers *.tsv* pour assurer la capture des informations à intégrer au fichier *.json*, et un script *.tsv* vers *.json* pour assurer la normalisation des transcriptions, calculer la durée des segments et normaliser le chemin vers le fichier *.wav*.

11. Il est également possible de découper les enregistrements. Dans ce cas, l'entrée correspond à un enregistrement et les temps de début et de fin de chaque entrée n'est pas renseigné.

Les fichiers audio ont tous été converti au format *.wav*, mono, 16 bits, 16kHz grâce à la librairie *sox*. L'ensemble des scripts est disponible sur le répertoire Gitlab de cette thèse¹².

3.3 Étiquetage des données pour la création d'ensembles représentatifs de sous-types de parole spontanée

Nous faisons le choix d'étudier dans cette thèse l'influence de deux dimensions sur les quatre présentées en section 2 : la situation de communication et le degré d'intimité entre les locuteurs. L'ensemble des données traitées ici considèrent les caractéristiques les plus favorables à la parole spontanée sur les dimensions *type de communication* et *canal de communication*, à savoir des communications de type interpersonnel et en face-à-face. Ce faisant, nous écartons ainsi de notre étude les extrêmes que représentent la communication de masse et la communication à distance, situations déjà largement représentées et étudiées au travers des corpus médiatiques de référence en reconnaissance automatique de la parole en français. Ainsi, si les enregistrements sont bel et bien étiquetés en fonction de chacune des dimensions, chacun d'entre eux partage le même étiquetage des dimensions “canal de communication” (face-à-face) et “type de communication” (interpersonnelle).

Nous nous sommes donc focalisés sur l'étude des deux dimensions “situation de communication” et “degré d'intimité entre les locuteurs”, en commençant par leurs graduations extrêmes :

- symétrie + lieu quotidien/familier *vs* dissymétrie + lieu professionnel/institutionnel pour la situation de communication
- et amis proches/famille *vs* inconnus pour le degré d'intimité entre les locuteurs.

Ce choix nous a amené à rechercher dans les corpus recueillis, des enregistrements à rassembler pour former des cas d'études homogènes correspondant aux situations suivantes :

Usual_close¹³ : Amis proches ou membres d'une même famille et situation de communication symétrique dans un lieu quotidien/familier.

exemple : *Des amis qui discutent chez l'un d'entre eux.*

Unusual_close : Amis proches ou membres d'une même famille qui échangent sur un lieu de travail ou au sein d'une institution. Les locuteurs ne sont pas dans

12. <https://gitlab.com/solene-evain/recops++>

13. Pour plus de facilité, nous gardons les dénominations en anglais adoptées dans notre publication à LREC/COLING 2024. Nous proposons toutefois la traduction suivante : *Usual_close* → *Usuel_proche*, *Unusual_close* → *inhabituel_proche*, *Usual_distant* → *usuel_distant*, *Unusual_distant* → *inhabituel_distant*

une situation symétrique.

exemple : *Quelqu'un qui interview un ami au travail pour récolter des données.*

Usual_distant : Une situation de communication symétrique dans un lieu quotidien/familier, et des locuteurs qui ne se connaissent pas.

exemple : *Deux personnes qui ne se connaissent pas qui ont une interaction dans la rue pour indiquer un chemin.*

Unusual_distant : Deux personnes qui ne se connaissent pas et qui échangent au travail ou au sein d'une institution. Les locuteurs sont dans une situation dissymétrique.

exemple : *Une interview sur un lieu de travail, entre des personnes qui ne se connaissent pas.*

Nous représentons en figure VI.9 chacun de ces cas sur les axes de variation de la parole spontanée préalablement déterminés. Le cas *Usual_close* est représenté avec le quadrilatère ayant l'aire la plus petite, ce qui inclut donc les enregistrements supposés comme les plus spontanés. A l'inverse, le cas *Unusual_distant* a l'aire la plus grande, et inclut donc, en ce sens, les enregistrements supposés comme les moins spontanés.

Afin de procéder à l'étiquetage des données, nous avons analysé chaque corpus un par un. Étant donné qu'il nous aurait été impossible sur le temps imparti d'étudier chaque enregistrement un par un afin de pouvoir les étiqueter, nous avons adopté la méthode suivante :

- si un corpus a été créé en suivant un seul et même protocole, alors l'étiquetage est effectué à partir des informations présentes dans le *data paper*. Les fichiers de métadonnées peuvent ensuite être consultés si une information manque. Cela a été le cas pour les corpus CFPP, CFPB, CID.
- si un corpus est composé de multiples sous-ensembles et donc de multiples protocoles, alors un maximum d'information est tiré du *data paper* et le reste est recherché sur le site internet du corpus ou dans les métadonnées. Les fichiers de métadonnées sont donc d'une très grande importance dans ce cas. Cette méthode a été adoptée pour les corpus ESL02, TCOF, PFC.

La figure VI.10 donne le détail du nombre d'heures et de la provenance des données des ensembles d'apprentissage, de développement et de test réunies pour ce travail de thèse.

Le temps total de recueil, d'analyse, d'étiquetage et de normalisation des données a été effectué en parallèle de notre implication dans le projet LeBenchmark et de la détermination de dimensions de variation de la parole spontanée, et s'est donc étalé sur 2 ans. Le travail d'étiquetage des données pour les corpus représentatifs

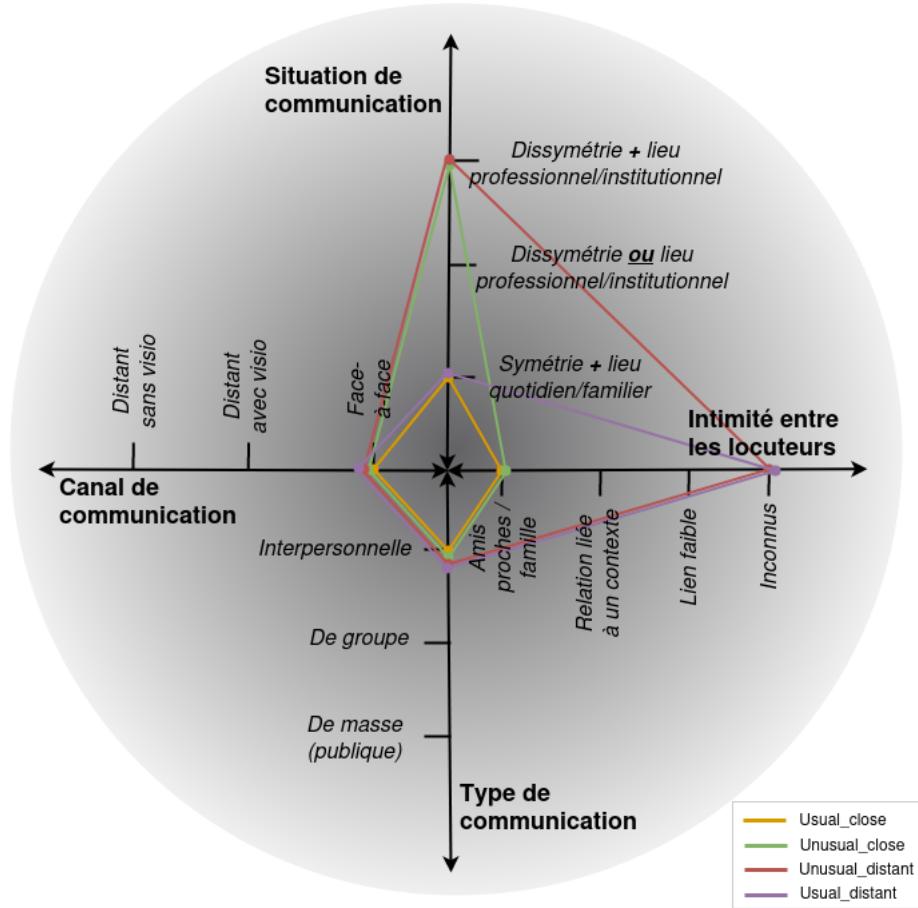


FIGURE VI.9 – Positionnement des cas d'étude dans l'espace en quatre dimensions de variation de la parole spontanée

d'une multitude de situations, comme TCOF ou ESLO2, aura été particulièrement chronophage, nous obligeant à aller chercher les informations nécessaires à l'étiquetage parfois jusque dans les fichiers de métadonnées de chacun des enregistrements. Certains doutes liés à l'étiquetage auront quant à eux parfois pu être levés grâce à l'étude des informations renseignées sur les sites Web des projets ou encore grâce à une prise de contact directe avec les auteurs des corpus. Cette dernière démarche est toutefois restée exceptionnelle. Faute d'avoir pu trouver une quantité de données suffisante lorsque le nombre de locuteurs était limité à deux, nous avons été contraints d'inclure des enregistrements avec trois locuteurs dans certains cas.

Grâce au travail d'étiquetage effectué en fonction du rôle, du lieu, du type de communication, du canal de communication et du degré d'intimité entre les locuteurs,

CHAPITRE VI : Dimensions de variation de la parole spontanée et étiquetage de données

nous obtenons un total de 9h35 de parole pour le cas *Usual_close*, 16h24 pour le cas *Unusual_close* et 12h51 pour le cas *Unusual_distant*. Le sous-type *Usual_distant* n'est pas présenté car nous n'avons pas pu rassembler assez de données pour ce cas d'étude. Nous n'avons donc pas pu l'étudier.

Afin de pouvoir utiliser ces données pour une tâche de reconnaissance automatique de la parole, nous avons défini des ensembles d'apprentissage, de développement et de test. Les ensembles d'apprentissage représentent approximativement 80% des heures de parole, tandis que les ensembles de développement et de test sont constitués d'environ 10% de ces heures, chacun. Ces proportions sont approximatives car aucun enregistrement n'a été découpé afin d'être sûr qu'aucune des situations de développement et ou de test n'ait déjà été vue lors de l'apprentissage.

Le nombre d'heures total des ensembles d'apprentissage, de développement et de test des différents cas représente respectivement 29h34, 4h43 et 4h35 de parole. Le cas *Usual_close* est celui pour lequel nous avons pu rassembler le moins d'enregistrements (6h50 de parole effective pour l'ensemble d'apprentissage) n'atteignant ainsi pas notre objectif de 10 heures de parole. Nous avons toutefois souhaité le conserver, celui-ci représentant *a priori* le plus haut niveau de spontanéité parmi les sous-types étudiés. Les ensembles de développement et de test représentent respectivement 1h17 et 1h28 de parole. L'ensemble d'apprentissage du cas *Unusual_distant* représente également un peu moins de 10 heures de parole (9h36), contrairement au cas *Unusual_close* pour lequel nous avons pu rassembler 13 heures. Cette différence pourrait s'expliquer par la plus grande facilité que les linguistes peuvent avoir à trouver des volontaires parmi leurs amis ou les membres de leur famille pour participer à des enquêtes linguistiques, que parmi des inconnus (cas pour lequel nous avons tout de même pu rassembler 9h36 de données grâce à la combinaison d'enregistrements des deux corpus TCOF et ESL02). Les ensembles de développement et de test de ces deux derniers cas contiennent 1h35 et 1h40 pour le cas *Unusual_distant* et 1h51 et 1h25 pour le cas *Unusual_close*.

Les données non étiquetées ont finalement été rassemblées en un ensemble appelé *All_spont*, également découpé en ensembles d'apprentissage, de développement et de test représentant respectivement 268h, 34h et 34h de parole effective. Ces ensembles sont à considérer comme de la parole spontanée "tout-venant", c'est-à-dire que le détail de la représentation de nos différentes dimensions dans ces données n'est pas connue.

4 Synthèse

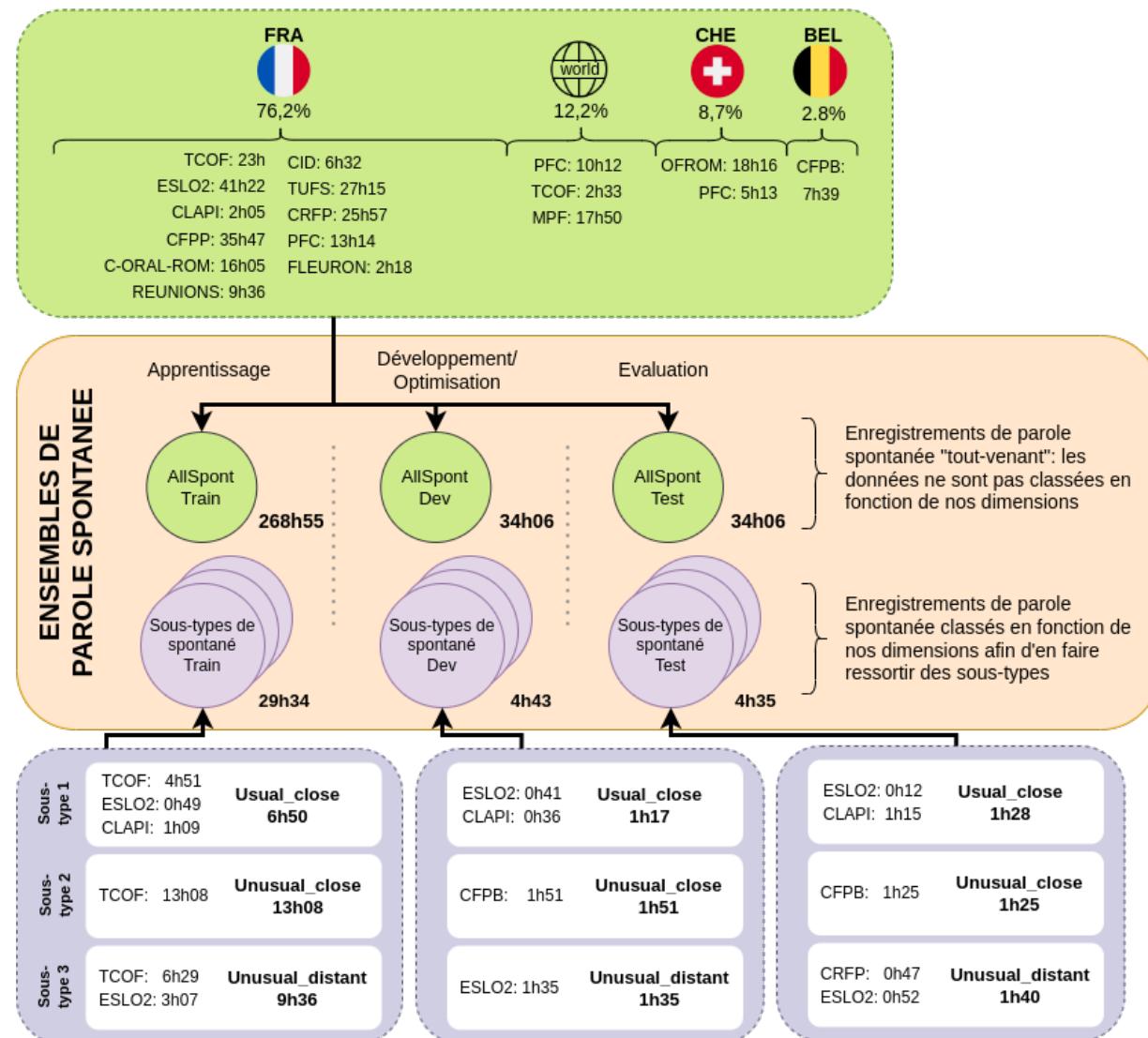


FIGURE VI.10 – Détail des ensembles d'apprentissage, de développement et d'évaluation *All_Spont* et Sous-types de spontané (durées = heures effectives de parole)

4 Synthèse

Nous avons présenté, dans ce chapitre, les corpus issus de la linguistique sur lesquels nous nous sommes basés, tout d'abord, afin de déterminer un nombre restreint de dimensions de variation de la parole spontanée. Plusieurs d'entre eux étant déjà inclus partiellement dans le Corpus d'Etude pour le Français Contemporain (CEFC), nous avons procédé à leur complémentation. D'autres ont nécessité un triage enregistrement par enregistrement lorsqu'ils étaient disponibles en plusieurs endroits. En

effet, certains enregistrements étaient parfois disponibles sur le site web relatif au corpus, sur la plateforme Ortolang ou au sein du CEFC et pas ailleurs. Or, nous souhaitions maximiser la quantité de données disponibles. Nous avons collecté en parallèle de ces corpus les *data papers* et fichiers de métadonnées les accompagnant. Grâce à ces informations et en nous appuyant sur l'état de l'art (chapitre II) nous avons sélectionné des facteurs de variations de la parole spontanée que nous présentons selon quatre dimensions :

- la situation de communication
- le degré d'intimité entre les locuteurs
- le canal de communication
- et le type de communication.

Nous proposons pour chacune de ces dimensions des graduations permettant de déterminer *a priori* si une situation est plus ou moins propice à l'apparition d'une parole très spontanée. Nous espérons ainsi pouvoir capturer différents niveaux de spontanéité grâce à l'étiquetage de nos données.

Les données collectées, triées et étiquetées nous ont permis, dans un deuxième temps, de créer des ensembles représentatifs de sous-types de parole spontanée. En effet, nous concentrant sur l'étude des deux dimensions “situation de communication” et “degré d'intimité entre les locuteurs” dans leurs aspects extrêmes, nous avons établis les cas d'étude suivants :

- ***Usual_close*** : Amis proches ou membres d'une même famille et situation de communication symétrique dans un lieu quotidien/familier.
- ***Unusual_close*** : Amis proches ou membres d'une même famille qui échangent sur un lieu de travail ou au sein d'une institution. Les locuteurs ne sont pas dans une situation symétrique.
- ***Usual_distant*** : Une situation de communication symétrique dans un lieu quotidien/familier, et des locuteurs qui ne se connaissent pas.
- ***Unusual_distant*** : Deux personnes qui ne se connaissent pas et qui échangent au travail ou au sein d'une institution. Les locuteurs sont dans une situation asymétrique.

Pour chacun de ces cas, nous sommes parvenus à étiqueter 9h35 de parole pour *Usual_close*, 16h24 pour *Unusual_close* et 12h51 pour *Unusual_distant*. Le cas *Usual_distant* est écarté de notre étude, faute d'une quantité suffisante de données. Les données de parole spontanée non étiquetées sont rassemblées pour former l'ensemble *All_spont*, rassemblant 337 heures de parole. En termes de quantité de données, le nombre d'heures moyen de données étiquetées pour chacun de nos cas représente 3,56% de *All_spont*, ce qui est très faible. Cette faible proportion vient, dans une grande partie, de l'absence de métadonnées sur lesquelles repose notre

étiquetage en fonction de nos dimensions. Mais elle vient également du fait que nous avons choisi de nous intéresser aux cas extrêmes des deux dimensions *situation de communication* et *degré d'intimité entre les locuteurs*, délaissant ainsi les gradations intermédiaires. Ce choix s'inscrit dans une volonté d'explorer les différentes dimensions de la parole spontanée à partir d'ensembles d'enregistrements homogènes au sein d'un même cas, lesquels devaient présenter un fort contraste les uns envers les autres.

Chapitre VII

Résultats d'expérimentations

L'état de l'art nous aura montré que les systèmes de reconnaissance automatique de la parole sont sensibles à différents niveaux de spontanéité. En effet, plus la spontanéité augmente, plus la performance baisse. Néanmoins, si nous savons que différents facteurs sont à l'origine de la production de différents sous-type de parole spontanée, il reste encore à déterminer lesquels de ces facteurs sont les plus problématiques pour les systèmes de reconnaissance automatique de la parole. Pour ce faire, nous utiliserons les modèles état de l'art LeBenchmark et l'étiquetage de différents enregistrements en fonction de notre modèle en quatre dimensions effectué auparavant.

Ainsi, la première expérimentation présentée dans ce chapitre nous permet de vérifier l'homogénéité et l'existence de différents niveaux de spontanéité dans les trois cas étudiés. Pour cela, nous utilisons des systèmes de reconnaissance automatique de la parole état de l'art et vérifions le lien entre nos différents cas et une hausse du WER. Ces résultats nous permettent par la suite d'étudier l'amélioration des performances de reconnaissance du moins spontané au plus spontané. Nous commençons en effet par explorer l'impact de différentes adaptations spécifiques d'un modèle pré-appris LeBenchmark avec de faibles quantités de parole. Ces adaptations spécifiques correspondent à l'adaptation d'un modèle avec des données représentatives de chacun de nos cas d'étude. Ensuite, une dernière expérimentation nous permet d'étudier l'impact d'une adaptation au domaine avec une grande quantité de parole, avant une adaptation spécifique.

Déroulement du chapitre

1 Performances de reconnaissance avec des systèmes état de l'art	149
--	-----

1.1	Analyse des résultats	151
2	Adaptation spécifique d'un modèle pré-appris	152
2.1	Analyse des résultats	153
2.2	Validation croisée sur chacun des cas d'étude	156
2.2.1	Validation croisée pour le cas <i>Usual_close</i>	158
2.2.2	Validation croisée pour le cas <i>Unusual_close</i>	160
2.2.3	Validation croisée pour le cas <i>Unusual_distant</i>	162
2.3	Analyse générale de la validation croisée	163
3	Apports d'une adaptation au domaine	165
3.1	Un système adapté au domaine spontané	165
3.2	Adaptation spécifique d'un système adapté au domaine	167
3.2.1	Analyse des résultats	168
4	Synthèse	168

Nous rappelons dans le tableau VII.1 l'ensemble des systèmes qui seront utilisés dans ce chapitre.

L'ensemble des scripts de préparation des données, d'entraînement et d'évaluation des modèles est partagé sur le répertoire Gitlab de cette thèse¹. Les systèmes suivants sont à retrouver sur HuggingFace² :

- l'ensemble des systèmes directement adaptés à nos cas spécifiques (SB-*Usual_close*, SB-*Unusual_close* et SB-*Unusual_distant*)
- le système adapté au domaine spontané (SB-*All_spont*)
- l'ensemble des systèmes issus de l'adaptation spécifique du système adapté au domaine spontané (SB-*All_spont-Usual_close*, SB-*All_spont-Unusual_close* et SB-*All_spont-Unusual_distant*).

Enfin, les fichiers *.json* de données ainsi que les résultats de transcriptions automatiques accompagnées du calcul du WER sont en ligne³⁴.

1. <https://gitlab.com/solene-evain/recops>
 2. <https://huggingface.co/Sevain>
 3. Fichiers json : <https://drive.proton.me/urls/AKVFBP03QG6TGLRS15AjLF>
 4. Transcriptions automatiques : <https://drive.proton.me/urls/MV6TRAM83GIYa1MDAKovjQ>

Système	Modèle pré-apris	Adapt.	Données d'apprentissage	Données de validation
Whisper	Whisper-large	Non	-	-
SB-CV	LB-7K-large	Oui	CV	CV
SB-CV2	LB-7K-large	Oui	CV	All_cases
SB- <i>Usual_close</i>	LB-7K-large	Oui	<i>Usual_close</i>	All_cases
SB- <i>Unusual_close</i>	LB-7K-large	Oui	<i>Unusual_close</i>	All_cases
SB- <i>Unusual_distant</i>	LB-7K-large	Oui	<i>Unusual_distant</i>	All_cases
*SB- <i>All_spont</i>	LB-7K-large	Oui	All_spont	All_cases
*SB- <i>All_spont_Usual_close</i>	LB-7K-large	Oui	All_spont, <i>Usual_close</i>	All_cases
*SB- <i>All_spont_Unusual_close</i>	LB-7K-large	Oui	All_spont, <i>Unusual_close</i>	All_cases
*SB- <i>All_spont_Unusual_distant</i>	LB-7K-large	Oui	All_spont, <i>Unusual_distant</i>	All_cases

*correspond à un système adapté au domaine spontané // Les systèmes en gras sont les systèmes état de l'art.

TABLE VII.1 – Rappel des différents systèmes utilisés pour les expérimentations, ainsi que des ensembles de données d'apprentissage et de développement.

1 Performances de reconnaissance avec des systèmes état de l'art

Nous avons, dans cette première section, pour objectif d'observer le comportement de différents systèmes état de l'art sur les ensembles de test de nos différents cas d'étude. La littérature nous ayant déjà montré que l'absence de rapport hiérarchique entre les locuteurs et que les liens amicaux ou familiaux favorisent une parole plutôt très spontanée, notre cas *Usual_close* devrait ainsi représenter une parole plus spontanée que *Unusual_close*, lui-même plus spontané que *Unusual_distant*. Ainsi, nous nous attendons à retrouver, pour cette première expérimentation, un WER plus haut sur notre cas le plus spontané et une baisse du WER en fonction de la baisse de la spontanéité capturée dans les deux autres cas.

Le premier système état de l'art que nous utilisons est le système à base de CTC de Speechbrain, présenté en section 2.2.1 du chapitre IV. La recette étant réutilisée

pour la suite des expérimentations, nous avons fait le choix d'entraîner de nouveau le système afin de nous assurer du bon fonctionnement de l'apprentissage de modèles à partir de celle-ci. Le système est adapté sur les données de développement du corpus Common Voice. Nous avons obtenu, après apprentissage, un *Word Error Rate* de 9,97% sur l'ensemble de test de Common Voice, ce qui rejoint les résultats annoncés par l'équipe Speechbrain (9,96%⁵).

Le système état de l'art suivant est Whisper. Une différence majeure avec le système à base de CTC utilisé précédemment est son utilisation en *zero-shot*, c'est-à-dire sans adaptation à un ensemble de données spécifiques. Nous avons utilisé, pour l'évaluation de ce système, le script de calcul du WER fourni par l'équipe Speechbrain déjà utilisé pour le calcul du WER du premier système présenté. Cela a nécessité un travail de normalisation sur les sorties du système Whisper.

Un décodage avec ces deux systèmes état de l'art est effectué :

- sur les ensembles de test de nos cas représentatifs de différents niveaux de spontanéité (*Usual_close*, *Unusual_close* et *Unusual_distant*),
- sur l'ensemble des ensembles de test de nos cas (*All_cases*),
- sur l'ensemble de test de données de parole spontanée “tout-venant” que nous avons rassemblées (*All_Spont*),
- sur l'ensemble de test du corpus ETAPE,
- sur l'ensemble de test du corpus Common Voice.

L'ensemble des résultats est à retrouver dans le tableau VII.2.

Système	Modèle	<i>Usual_close</i>	<i>Unusual_close</i>	<i>Unusual_distant</i>	<i>All_cases</i>	<i>All_spont</i>	<i>ETAPE</i>	<i>CV</i>
SB-CV	LB-7K-large	80,85	52,66	32,16	55,14	51,2	36,55	9,97
Whisper	large-v2	65,29	38,68	27,57	43,67	40,45	28,33	12,93

Les valeurs en gras représentent le meilleur WER obtenu par ensemble de test.

TABLE VII.2 – Performances obtenues avec deux systèmes de RAP état de l'art (WER%)

5. <https://huggingface.co/speechbrain/asr-wav2vec2-commonvoice-fr>

1.1 Analyse des résultats

Les résultats nous montrent que d'un point de vue global, Whisper est le système donnant les meilleures performances sur chacun des ensembles de test, excepté sur Common Voice (2,97 point de différence avec la référence Speechbrain). Le système Speechbrain ayant été appris et adapté sur ce corpus -ce qui n'est pas le cas de Whisper, du moins concernant l'adaptation- il n'est pas étonnant que le WER soit meilleur (9,97% contre 12,93%).

Sur le corpus ETAPE, référence pour la reconnaissance automatique de la parole spontanée en français, les performances de Whisper (28,33%) dépassent de 8,22 points les performances du système Speechbrain (36,55%). À nouveau, il est important de rappeler que le système Speechbrain a été appris et adapté sur les données du corpus Common Voice. Ce corpus comprenant uniquement de la parole lue, il n'est pas étonnant que les résultats sur les corpus comprenant de la parole spontanée soient moins bons. La même tendance est observée sur le corpus *All_spont* (40,45% pour Whisper et 51,2% pour Speechbrain, soit 10,75 points d'écart) regroupant divers corpus de parole spontanée.

Nous avions émis l'hypothèse que nos cas étant représentatifs de différents niveaux de spontanéité, nous devrions retrouver un WER plus élevé pour notre cas le plus spontané et moins élevé pour notre cas le moins spontané. Cette tendance est effectivement observée, avec une gradation ascendante des WER obtenus sur *Unusual_distant*, *Unusual_close* et *Usual_close*. Quel que soit le système, le WER le plus bas est ainsi obtenu sur le cas incluant des situations telles que des entretiens entre locuteurs qui ne se connaissent pas, tandis que le plus haut est obtenu pour des situations telles que des repas entre amis. L'analyse des résultats montre un écart moins important entre *Unusual_close* et *Unusual_distant* qu'entre *Usual_close* et *Unusual_close* (11,11 points vs 26,61 points pour le système Whisper). Ceci laisse envisager un impact plus fort de la dimension *situation de communication* que de la dimension *degré d'intimité entre les locuteurs* sur le WER.

Les WER sur *All_spont* quant à eux sont chaque fois assez proches des WER obtenus sur *All_cases* et sur *Unusual_close*. Pour rappel, *All_spont* regroupe un échantillon des enregistrements de parole spontanée "tout-venant", sélectionnés de façon aléatoire. Ainsi, l'ensemble *All_spont* pourrait tout à la fois comprendre uniquement des enregistrements représentatifs de la situation *Unusual_close* comme une situation d'interview entre amis, mais aussi comprendre des enregistrements représentatifs de chacun des cas présentés. Dans ce dernier cas, le WER pourrait être considéré comme non représentatif des performances pouvant être obtenues avec ce système. En effet, comme les différents WER sur *All_cases* nous le

montrent, lorsqu’un ensemble de test est constitué de données correspondant à la fois à *Usual_close*, *Unusual_close* et *Unusual_distant*, le WER obtenu est “lisé”. Celui-ci ne permet pas d’identifier les potentielles difficultés d’un système sur des enregistrements très spontanés, ni ne permet d’identifier les cas dans lesquels les performances sont les meilleures. C’est ainsi que pour le système Whisper par exemple, le WER sur *All_cases* est de 43,67% alors même que notre étiquetage nous permet d’identifier à la fois des performances autour de 65% pour le cas très spontané *Usual_close*, autour de 38% pour le cas moins spontané *Unusual_close* et autour de 27% pour le cas le moins spontané *Unusual_distant*. Or, si l’exploitation d’une transcription avec 7 mots sur 10 bien reconnus est envisageable, une transcription comprenant plus de 65% d’erreurs ne l’est pas vraiment. La combinaison, dans un ensemble de test, de données de parole spontanées issues de situations trop différentes ne permet donc pas d’évaluer précisément les performances d’un système sur ce type de parole.

2 Adaptation spécifique d’un modèle pré-appris

Nous étudions dans cette section l’adaptation du modèle LeBenchmark avec différents ensembles de données de parole spontanée représentatifs de chacun de nos cas d’étude. Le système SB-CV présenté précédemment étant adapté sur de la lecture, une adaptation du modèle LeBenchmark avec des données de parole spontanée, quelles qu’elles soient, devrait améliorer les résultats sur les ensembles représentatifs de ce type de parole. Ensuite, nous pensons que l’adaptation du modèle à un cas spécifique devrait toujours permettre d’obtenir le meilleur WER sur ce cas (le système SB-*Usual_close* devrait par exemple permettre d’obtenir le meilleur WER sur l’ensemble de test *Usual_close*). Si tel est le cas, cela voudrait dire que les systèmes sont en capacité d’apprendre des représentations différentes pour chacun des sous-types de parole spontanée étudiés ici.

Ainsi, trois systèmes de reconnaissance automatique de la parole sont appris sur *Usual_close* (6h50 de parole effective), *Unusual_close* (13h08 de parole) et *Unusual_distant* (9h36 de parole). Nous testons ainsi dans des conditions quasiment similaires à celles présentées dans (Baevski *et al.*, 2020), montrant qu’une dizaine d’heures de parole est suffisante pour obtenir des résultats sur LibriSpeech proches de ceux de l’état de l’art lorsqu’aucun modèle de langue n’est utilisé⁶.

L’ensemble de développement utilisé est constitué de l’ensemble des données de développement de chacun des cas (*All_cases* dev). À des fins de comparaisons, nous incluons dans le tableau les résultats obtenus avec le système Speechbrain

6. 4,2/8,6% sur les ensembles *clean* et *other* avec l’approche de (Radford *et al.*, 2018), contre 8,0/12,1% pour Wav2Vec 2.0.

2 Adaptation spécifique d'un modèle pré-appris

état de l'art cette fois adapté sur *All_cases* dev. L'apprentissage de ce système et des trois systèmes spécifiques a été arrêté lorsque l'amélioration du WER ne dépassait pas 0,5 points sur trois époques consécutives. Nous présentons l'ensemble des résultats obtenus dans le tableau VII.3.

Système	Modèle	Épq	<i>Usual_close</i>	<i>Unusual_close</i>	<i>Unusual_distant</i>	<i>All_cases</i>	<i>All_spont</i>	ETAPE	CV
Whisper	large-v2	-	65,29	38,68	27,57	43,67	40,45	28,33	12,93
SB-CV2	LB-7K-large	23	80,87	52,87	32,15	55,20	51,05	36,22	9,94
SB- <i>Usual close</i>	LB-7K-large	32	63,74 ±1,3	39,83 ±0,9	25,54 ±0,69	43,37 ±0,66	41,75	38,57	36,85
SB- <i>Unusual close</i>	LB-7K-large	32	63,51 ±1,25	33,88 ±0,95	19,12 ±0,65	39,12 ±0,72	36,92	32,02	28,48
SB- <i>Unusual distant</i>	LB-7K-large	36	70,84 ±1,13	34,52 ±0,98	18,14 ±0,59	40,78 ±0,72	36,99	32,00	28,60

Épq : époque // Les cases grisées correspondent à un apprentissage et un décodage sur des données appartenant à un même cas. // Les valeurs en gras représentent le meilleur WER obtenu sur chacun des ensembles de test. // Les valeurs en gris représentent l'intervalle de confiance du WER moyen.

TABLE VII.3 – Performances obtenues suite à l'adaptation du modèle LeBenchmark 7K avec chacun des cas étudié (WER%)

2.1 Analyse des résultats

Quel que soit le système de reconnaissance automatique de la parole spécifique utilisé, les WER obtenus sur les données de test très spontanées *Usual_close* sont chaque fois plus élevés (WER >63%) que ceux obtenus sur les données moyennement spontanées *Unusual_close* (>33% et <40%), eux-mêmes chaque fois plus élevés que ceux obtenus sur les données peu spontanées *Unusual_distant* (<26%). Les différents corpus de test ont donc des difficultés intrinsèques que l'existence de données similaires dans les données d'apprentissage ne suffit pas à réduire. Tout comme nous l'avions noté avec les systèmes état de l'art, l'écart de WER entre *Usual_close* et *Unusual_close* est toujours plus grand que celui entre les WER de *Unusual_close* et *Unusual_distant*. Ceci semble encore une fois indiquer que la dimension *situation de communication*, et notamment la situation “usuelle”

représentée dans *Usual_close*, est celle apportant le plus de difficulté pour les systèmes de reconnaissance automatique de la parole. Les deux ensembles *All_cases* et *All_spont* obtiennent, quant à eux, des performances intermédiaires, soit comprises entre 40,78% et 43,37% pour le premier, et entre 36,92% et 41,75% pour le deuxième, ce qui, encore une fois, n'est absolument pas informatif quant aux mauvaises performances obtenues sur *Usual_close* et les bonnes voire très bonnes performances obtenues sur *Unusual_distant*. Tout comme avec les systèmes état de l'art, on observe pour les trois nouveaux systèmes que les WER sur *All_spont* sont chaque fois assez proches des WER obtenus sur *Unusual_close* et *All_cases*. Encore une fois, cela pourrait tout aussi bien indiquer qu'*All_spont* contient uniquement des données du même type que celles contenues dans *Unusual_close*, ou encore qu'une certaine représentativité de chacun de nos cas d'étude s'est trouvée ici capturée aléatoirement.

Une comparaison avec les systèmes état de l'art nous montre que les systèmes appris sur chacun de nos cas d'étude sont chaque fois plus performants sur nos corpus de parole spontanée. Les corpus Common Voice et ETAPE, en revanche, ne bénéficient pas de performances améliorées par rapport à ce qui a déjà été obtenu avec Whisper. En ce qui concerne Common Voice, l'écart important entre le WER obtenu grâce au système état de l'art (9,94%) et nos systèmes adaptés à des cas spécifiques de parole spontanée (de 28% à 36%) rappelle ce que nous avons déjà rapporté dans l'état de l'art : les systèmes adaptés sur de la lecture ne parviennent pas à donner de bons résultats sur de la parole spontanée et inversement. Pour ETAPE, une analyse de son contenu nous permet de voir que celui-ci ne comporte que très peu voire pas de situations se rapprochant de celles incluses dans nos cas d'étude (voir figure VII.1 et tableau VII.4). En effet, si le degré d'intimité entre les locuteurs n'est pas précisé, il apparaît que le contexte de l'émission de radio ou de télévision fait que les échanges entre les locuteurs se déroulent dans des situations plutôt cadrées (débat, interview...). Or, nos différents cas de parole spontanée rassemblent des communications en face-à-face, en petits comités (2 à 3 personnes), et non des débats et émissions médiatisés. Ceci pourrait expliquer pourquoi nos systèmes spécifiques ne permettent pas d'améliorer -voire dégradent de façon significative- les performances de Whisper (28% de WER contre 32 à 38%) sur cet ensemble de test. Le système Whisper reste donc le plus performant sur le corpus de test ETAPE. Nous pouvons supposer que les données d'apprentissage de ce système contiennent des données médiatiques, mais sans information précise sur ces données, nous sommes dans l'incapacité de vérifier cette hypothèse.

Ensuite, si un début de tendance montrant que le meilleur WER est obtenu avec les systèmes appris sur les ensembles d'apprentissage correspondants est observé sur *Unusual_close* et *Unusual_distant*, cela n'est pas tout à fait le cas pour

2 Adaptation spécifique d'un modèle pré-appris

genre	train	dev	test	sources
TV news	7h30	1h35	1h35	BFM Story, Top Questions (LCP)
TV debates	10h30	2h40	2h40	Pile et Face, à vous regarde, Entre les lignes (LCP)
TV amusements	–	1h05	1h05	La place du village (TV8)
Radio shows	7h50	3h00	3h00	Un temps de Pauchon, Service Public, Le masque et la plume, Comme on nous parle, Le fou du roi
Total	25h30	8h20	8h20	42h10

FIGURE VII.1 – Contenu du corpus ETAPE (tiré de (Gravier *et al.*, 2012))

Émission	Description
BFM Story	Cette émission fait place à l'analyse et aux débats autour des événements marquants de la journée. BFM Story reçoit des invités de tous horizons.
Top questions	Retransmission de trois questions au Gouvernement posées lors des séances à l'Assemblée Nationale.
Pile et Face	Rendez-vous de trente minutes, où le présentateur anime un débat entre deux politiques, interrogés sur un sujet au cœur d'actualité.
ça vous regarde	En présence de députés, d'experts, de personnalités mais aussi de citoyens, ce magazine consacré à la vie parlementaire et politique développe chaque soir un thème au cœur de l'actualité
Entre les lignes (LCP)	Émission en deux parties. Le Kiosque (une revue de presse de l'actualité) : C. Ruaults orchestre le débat entre quatre rédacteurs en chef de news magazines sur les choix éditoriaux réalisés par les rédactions respectives. Les idées : grand entretien avec une personnalité.
La place du village	Le concept consiste en une visite des Pays de Savoie (petits villages comme grands quartiers), à la rencontre des habitants qui apportent un témoignage authentique, et de superbes images montrant toute la beauté de la région.
Un temps de Pauchon	H. Pauchon va à la rencontre d'anonymes, dans des espaces publics, lors de manifestations etc. et leur tend son micro. Un reportage qui s'intéresse à tout et ne s'interdit rien.
Service public	Chaque jour un (ou plusieurs thèmes) d'actualité sont abordés. Régulièrement des experts, entreprises, associations de consommateurs sont invités. Une chronique est souvent diffusée sur un sujet qui peut être différent du sujet principal, mais généralement en rapport avec la consommation.
Le masque et la plume	Le Masque et la Plume est une émission de radio dédiée à la critique de livres, de films de cinéma et de pièces de théâtre.
Comme on nous parle	Interviews d'artistes et d'acteurs de l'actualité.
Le fou du roi	Chaque jour, flanqué d'une équipe de chroniqueurs et d'humoristes, S. Bern reçoit en direct des invités du monde du spectacle, des médias ou de la politique.

TABLE VII.4 – Détail du contenu de chacune des émissions incluses dans le corpus ETAPE (recherches effectuées par nos soins)

Usual_close. Pour celui-ci, on peut en effet considérer que la meilleure performance est à la fois obtenue avec le système SB-*Usual_close* et le système SB-*Unusual_close* étant donné la faible différence de WER entre les deux (63,74% vs 63,51% respectivement) et les écarts-types correspondants. L'apport moins flagrant d'un apprentissage spécifique sur ce cas pourrait s'expliquer par le fait que la quantité de données d'apprentissage destinées à une adaptation spécifique était assez faible : 6h50 de parole, contre 13h08 pour *Unusual_close* et 9h36 pour *Unusual_distant*. La taille des ensembles d'apprentissage de chaque cas n'étant pas exactement la même pour chacun des cas, nous avons souhaité vérifier la stabilité de nos résultats et poursuivons ainsi nos analyses grâce à une technique de validation croisée que nous présentons dans la section suivante.

Enfin, nous pouvons remarquer que le système SB-*Usual_close* donne quasiment le même WER sur *Usual_close* que le système SB-*Unusual_close*, mais que contrairement à ce dernier, il ne performe pas aussi bien sur les deux autres cas (39,83% vs 33,88% sur *Unusual_close* et 25,54% vs 19,12% sur *Unusual_distant*). Ensuite, le système SB-*Unusual_close* permet d'obtenir de bonnes performances sur tous les cas, ainsi que sur *All_spont*. Ceci est sans doute dû au fait que ce système a été appris à la fois à partir de données représentatives d'une grande intimité entre les locuteurs -ce qui est également capturé dans *Usual_close*- mais aussi à partir de données représentatives de situations inhabituelles -ce qui est également capturé dans *Unusual_distant*. Il pourrait donc avoir appris à modéliser un peu de chacun des cas, ce qui lui permet alors de donner de bonnes performances partout. Le système SB-*Unusual_distant* montre, quant à lui, des performances quasiment similaires à SB-*Unusual_close* sur les deux cas les moins spontanés (34,52% vs 33,88% sur *Unusual_close*, et 18,14% vs 19,12% sur *Unusual_distant* respectivement), mais donne le WER le plus élevé sur *Usual_close* (70,84%). Cette analyse générale semble ainsi montrer qu'il existe une différence notable entre les cas *Usual_close* et *Unusual_distant*, qui fait qu'un système appris sur l'un de ces cas ne parvient pas à donner de bons résultats sur l'autre cas.

2.2 Validation croisée sur chacun des cas d'étude

La validation croisée consiste à vérifier si une décision ou un score obtenu à partir d'un ensemble de données se confirme avec un autre ensemble de données, indépendant du premier mais acquis avec le même instrument. Il en existe deux méthodes : (1) la validation croisée d'un contre tous (*leave-one-out cross-validation*) et (2) la validation croisée à k -lots (*k-fold cross-validation*). La première méthode consiste à retirer, dans un ensemble de données, une observation et de la réserver pour la validation d'un modèle. Ainsi, l'apprentissage se fait sur l'ensemble des données restantes (hors ensemble de test). Ce procédé est répété autant de fois que néces-

saire, jusqu'à ce que chacune des observations ait pu servir à la validation d'un modèle. L'observation retirée à un temps t est réinjectée dans l'ensemble d'apprentissage une fois la validation effectuée. La méthode du k -fold suit le même principe, à la différence que ce n'est pas une observation unique qui est retirée des données, mais un ensemble de taille k , comme l'illustre la figure VII.2 où trois lots sont utilisés.

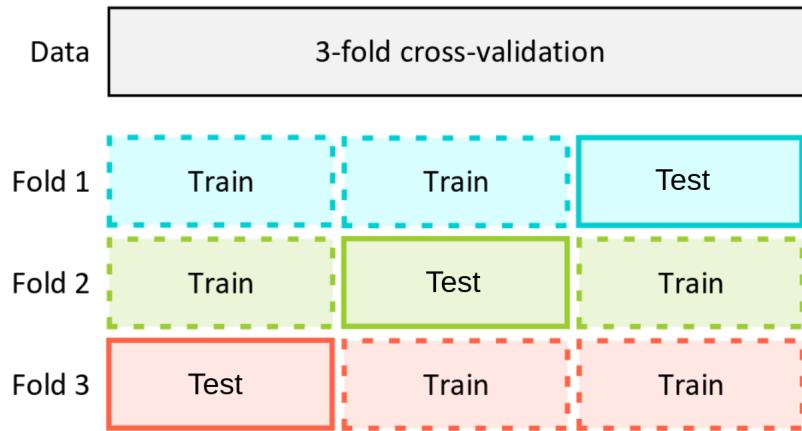


FIGURE VII.2 – Représentation de la technique de validation croisée à trois lots (adaptation d'une image de (Duran-Lopez *et al.*, 2020))

L'utilisation de cette technique amène à entraîner tout autant de modèles qu'il y a d'observations ou de lots. Des mesures statistiques sont ensuite effectuées afin d'avoir une idée plus précise de la tendance globale. Cela est particulièrement intéressant lorsque la quantité de données est plutôt faible puisqu'elle permet à la fois d'apprendre les modèles sur des ensembles de données plus grand, mais aussi d'augmenter la quantité de données destinée à l'évaluation. Ainsi, il devient plus aisément de détecter un cas de sur-apprentissage.

Dans notre cas, afin de limiter le nombre d'expérimentations à effectuer, nous avons adopté la technique du k -fold. Nos lots ont été créés afin de respecter au mieux les contraintes suivantes :

- durée d'un pli : entre 1h20 et 1h30
- un corpus par pli
- rassembler les sous-corpus similaires dans un même pli (ex : diners entre amis)
- rassembler les enregistrements d'un même pays et d'une même région dans un pli

Cela nous donne, pour le cas *Usual_close*, 7 lots d'une durée moyenne d'1h22, pour *Unusual_close*, 11 lots d'une durée moyenne d'1h29, et pour *Unusual_distant*, 10 lots d'une durée moyenne d'1h32. Le tableau VII.5 récapitule ces informations.

Cas	Durée moy.	[Max.-Min.]	Nombre de lots
<i>Usual_close</i>	1h22	[1h05 - 1h30]	7
<i>Unusual_close</i>	1h29	[1h22 - 1h41]	11
<i>Unusual_distant</i>	1h17	[0h59 - 1h32]	10

TABLE VII.5 – Détail des lots par cas d'étude constitués pour la validation croisée

Nous présentons dans les sections suivante la validation croisée pour chacun de ces cas d'étude. Les WER indiqués sont obtenus pour chaque itération sur :

- le lot de test
- l'ensemble de test *All_Spont*
- l'ensemble de test du corpus ETAPE
- et l'ensemble de test de Common Voice 6.1

Les performances obtenues par lot, par enregistrement et par locuteur pour chacun des cas étudiés sont à retrouver en annexe B.

2.2.1 Validation croisée pour le cas *Usual_close*

Le tableau VII.6 présente les résultats de la validation croisée obtenus pour ce cas.

Tout d'abord, nous observons que les résultats sont stables sur les ensembles de test de référence des corpus ETAPE et Common Voice. En effet, les écart-types (ET) sont respectivement de 0,82 et 0,84. La même stabilité est observée sur l'ensemble de test *All_Spont* ($ET=0,44$). Les différents systèmes appris en validation croisée montrent ainsi une relative stabilité.

Sur les lots de test, on observe au contraire une plus grande variabilité des résultats. En effet, le WER moyen est de 55,61% avec un écart-type de 9,83.

Les moins bons résultats sont obtenus sur le premier lot de test. Cependant, ce lot ne contient que 3 enregistrements. La figure VII.3 montre que le lot 1 se détache effectivement des autres lots au niveau du WER par enregistrement et du WER par locuteur. Cette différence de WER n'est pas due à un corpus ou un dialecte particulier étant donné que les données des lots 1 et 2, présentant 10 points de WER d'écart, sont tous deux issus de CLAPI et représentent des enregistrements faits en France, dans la même région (Rhône-Alpes). Il est important de rappeler que

Lot 1	Lot 2	Lot 3	Lot 4	Lot 5	Lot 6	Lot 7	WER/lot	# enregistrements	WER AllSpont	WER ETAPE	WER CV
							70,2*	3	42,05	38,77	37,18
							60,98	4	42,4	40,85	38,62
							65,4	5	42,56	39,7	37,85
							45,46	6	42,46	40,15	38,97
							48,04	9	43,47*	41,26*	39,67*
							47,39	8	42,65	40,23	38,5
							51,8	9	42,8	40,6	39,18
Moyenne		55,61		42,63		40,22		38,57			
Écart-type		9,83		0,44		0,82		0,84			

En gras : meilleur WER pour un ensemble de test // * : WER le moins bon pour un ensemble de test // Case colorée : lot destiné à la validation

TABLE VII.6 – Validation croisée de type *k-fold* sur les données *Usual_close*

CLAPI n'est pas un corpus à proprement parler mais plutôt une base de données regroupant une multitude de corpus enregistrés dans différentes conditions. Une analyse plus approfondie nous montre qu'un des enregistrements est à 85,8% de WER et que le WER de deux locuteurs sur trois au sein de cet enregistrement est supérieur ou égal à 90%. Cet enregistrement faisant partie du corpus CLAPI et étant audio-visuel, le visionnage d'une partie de la vidéo nous aura permis de constater que la caméra utilisée pour l'enregistrement est assez éloignée des locuteurs et que ceux-ci ne semblent pas porter de microphone individuel. Cet enregistrement est celui ayant le WER le plus élevé sur l'ensemble des lots du cas *Usual_close*. Le retrait de ce lot ramène la moyenne à 53,1 et l'écart-type à 8,14.

Le meilleur WER (45,46%) est obtenu sur le lot 4, qui regroupe des enregistrements du corpus TCOF (conversations) et du corpus ESLO2 (repas), enregistrés en France (Pays-de-la-Loire, Centre-Val-de-Loire) et aux Pays-bas⁷. Les lots 5, 6 et 7 montrent quant à eux des performances relativement stables, à défaut d'être les meilleures.

7. Ainsi qu'un enregistrement dont le lieu n'est pas précisé.)

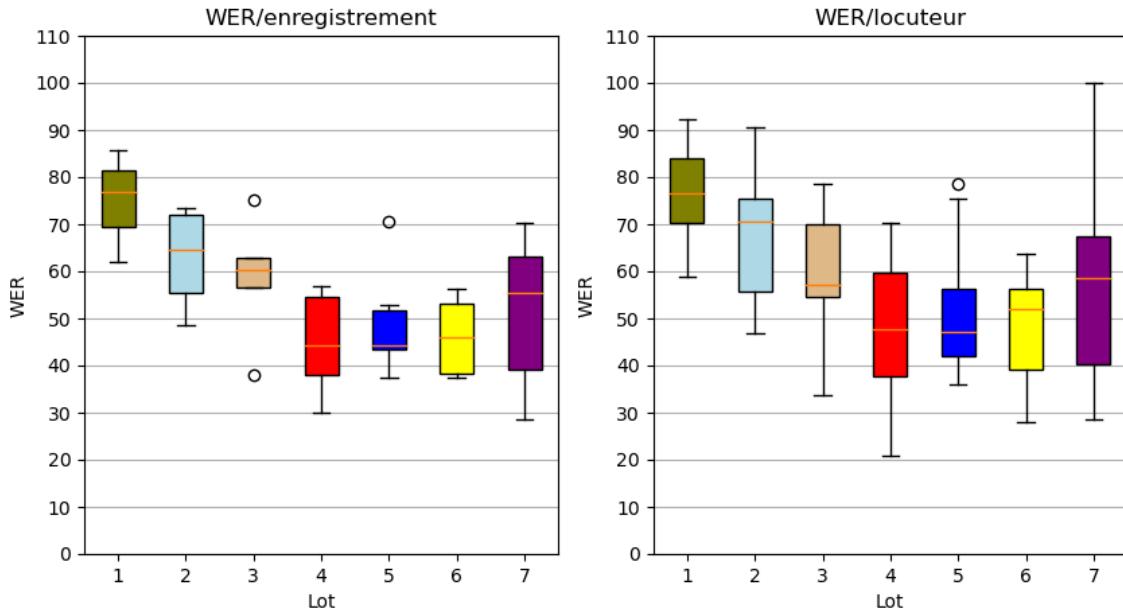


FIGURE VII.3 – WER/enregistrement (gauche) et WER/locuteur (droite) pour chaque lot du cas *Usual_close*

2.2.2 Validation croisée pour le cas *Unusual_close*

Les résultats de la validation croisée sur le cas *Unusual_close* sont reportés dans le tableau VII.7.

Le WER sur les corpus de référence ETAPE et CV est stable ($ET = 0,27$), tout comme sur l'ensemble de test *All_Spont* ($ET : 0,40$). Cependant, tout comme pour le cas *Usual_close*, nous observons une grande variabilité des résultats suivant les lots testés ($ET = 7,13$), avec un WER moyen à 29,96%.

Le lot sur lequel le WER est le moins bon est le lot 9 avec 43,31%, composé d'entretiens du corpus TCOF. Une analyse du WER par enregistrement et du WER par locuteur (voir figure VII.4) nous montre la présence d'une valeur considérée comme aberrante (86,06 % de WER) au niveau du WER par enregistrement. Une analyse du WER par locuteur nous permet de préciser qu'il s'agit ici de deux locuteurs pour lesquels les WER sont de 87% et 75% ce qui est nettement au-dessus du WER par locuteur pour les autres enregistrements de ce lot (entre 22% et 58%). L'écoute du fichier audio (internat_bea.wav) nous permet de conclure à un problème au niveau de l'enregistrement, le signal étant bruité tout du long. Le retrait de ce lot ramène le WER moyen à 28,62 et l'écart-type à 5,90. Le meilleur WER

	Lot 1	Lot 2	Lot 3	Lot 4	Lot 5	Lot 6	Lot 7	Lot 8	Lot 9	Lot 10	Lot 11	WER/lot	# enregistrements	WER AllSpont	WER ETAPE	CV
	31.17	2	37.27*	32.34*	28.01											
	26.84	2	36.49	31.97	28.27											
	28.70	7	36.89	32.17	28.61											
	19.69	5	36.48	31.58	28.16											
	25.97	5	36.11	31.56	28.24											
	22.92	6	37.25	31.96	28.39											
	28.73	8	36.93	31.71	28.76*											
	26.66	4	37.17	31.90	28.50											
	43.31*	5	36.19	31.43	28.03											
	35.25	3	36.55	31.87	27.83											
	40.32	5	36.59	31.61	28.41											
Moyenne			29,96		36,72	31,82	28,29									
Ecart-type			7,13		0,40	0,27	0,27									

En gras : meilleur WER pour un ensemble de test // * : WER le moins bon pour un ensemble de test // Case colorée : lot destiné à la validation

TABLE VII.7 – Validation croisée de type *k-fold* sur les données *Unusual_close*

de 19,69% est quand à lui obtenu sur le lot 4, correspondant à des enregistrements du sous-corpus “conversations” du corpus TCOF.

Les deux WER à 100% observés pour les lots 5 et 6 en figure VII.4 sur le graphique de droite s’expliquent par le fait que les locuteurs ne prononcent qu’un mot et que celui-ci est mal reconnu. Cela n’impacte pas la reconnaissance générale de l’enregistrement et du lot. Dans le cas du lot 5, l’enregistrement en question fait partie du sous-corpus “conversations” de TCOF. Il peut donc sembler assez étonnant d’y retrouver un enregistrement avec un locuteur ne prononçant qu’un mot. L’explication repose probablement sur le sujet abordé : un jeune homme raconte le décès de son grand-père et à quel point cela a été difficile pour lui. L’interviewer prononce le mot “pardon” à un moment où elle fait du bruit et cela interrompt le discours du jeune homme. Ce manque d’échange ne correspond en effet pas au schéma que l’on retrouve dans la plupart des autres enregistrements de ce sous-corpus et dans

ce cas, l'interviewer a laissé le jeune homme parler.

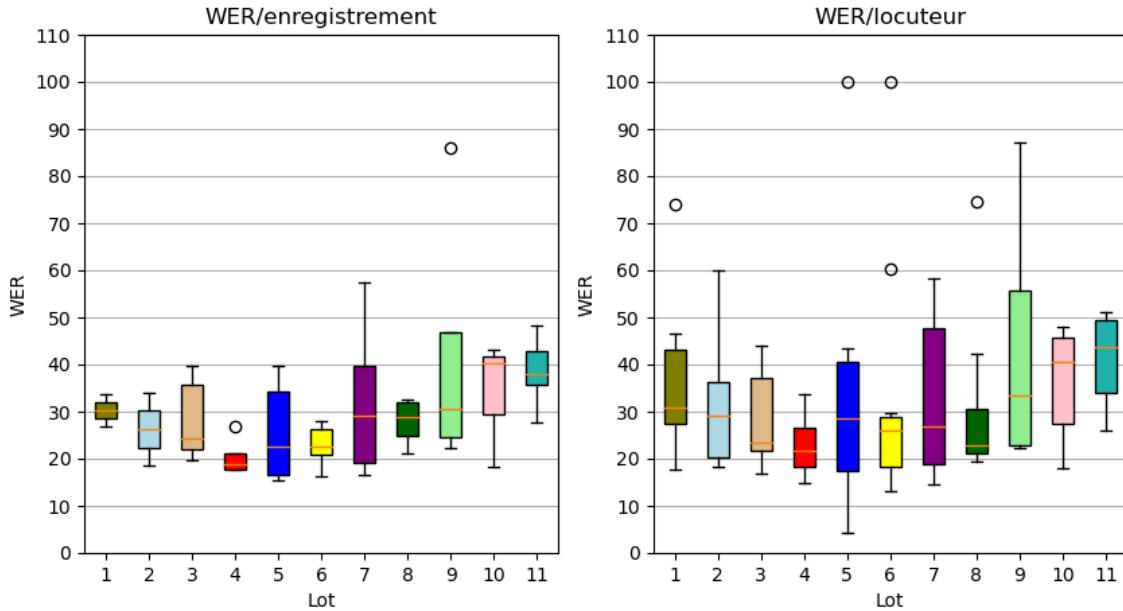


FIGURE VII.4 – WER/enregistrement (gauche) et WER/locuteur (droite) pour chaque lot du cas *Usual_close*

2.2.3 Validation croisée pour le cas *Unusual_distant*

Le tableau VII.8 présente les résultats de la validation croisée sur le cas *Unusual_distant*.

Tout comme pour les deux autres cas, le WER sur *All_Spont*, ETAPE et CV est stable, tandis que le WER sur les différents lots varie de 18,19% à 37,57%.

Les WER les plus hauts sont obtenus sur les lots 7 et 8. Les enregistrements de ces lots sont issus du sous-corpus “entretiens” de TCOF. Le lot 7 comporte une certaine disparité des WER par enregistrements allant de 15,47% à 94,56%⁸. L'enregistrement de ce lot avec le WER le plus haut compte deux locuteurs avec des WER à 94,54% et 100%. Cela trouve son explication dans le fait que l'enregistrement ne contient plus aucun son à partir de la seconde 56 (sur 16 min 50 d'enregistrement). Ce problème n'a pas été détecté au moment de l'étiquetage de nos données. L'écartement de ce lot donne une nouvelle moyenne à 25,62 et un écart-type à 6,58. Pour ce qui est du lot 8, il s'agit d'enregistrements de locuteurs

8. 15,47 ; 22,64 ; 23,7 ; 43,36 ; 94,56

Lot 1	Lot 2	Lot 3	Lot 4	Lot 5	Lot 6	Lot 7	Lot 8	Lot 9	Lot 10	WER/lot	# enregistrements	WER AllSpont	WER ETAPE	WER CV
										19.02	3	36.44	31.91	29.00
										22.41	3	36.57	32.23	29.00
										18.19	2	37.57*	32.52*	29.04
										24.05	2	36.56	32.04	28.25
										22.15	1	37.18	31.92	28.88
										34.96	4	37.05	32.06	29.21*
										37.57*	5	36.73	31.55	28.48
										37.25	3	37.45	32.34	29.05
										26.12	8	35.64	31.89	28.60
										26.46	6	36.80	31.92	28.72
Moyenne				26,81			36,80		32,05		28,82			
Ecart-type				7,26			0,56		0,28		0,30			

En gras : meilleur WER pour un ensemble de test // * : WER le moins bon pour un ensemble de test // Case colorée : lot destiné à la validation

TABLE VII.8 – Validation croisée de type *k-fold* sur les données *Unusual_distant*

avec un accent de la région Provence-Alpes-Côte-d'Azur. L'enregistrement avec le WER le plus élevé dans ce lot, soit 46,11% contient un bruit de fond. Le retrait de ces deux lots donne une moyenne à 24,17 et un écart-type à 5,27.

Le meilleur WER de 18,19% est quant à lui obtenu sur le lot 3 qui rassemble deux entretiens tirés du corpus ESLO2.

2.3 Analyse générale de la validation croisée

Le tableau VII.9 rappelle les WER moyens et leur écart-type obtenus sur les lots de validation croisée pour les différents cas d'étude, ainsi que les WER moyens obtenus pour un apprentissage et un décodage sur un même cas, issus du tableau VII.3.

Le cas *Usual_close*, le plus spontané, est celui avec le WER moyen le plus élevé

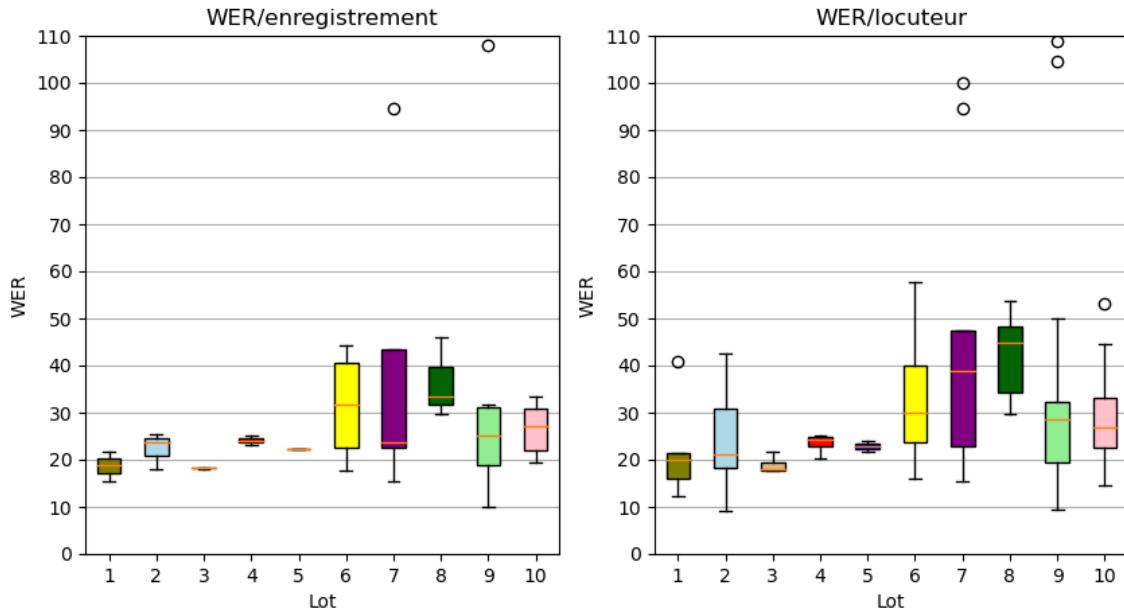


FIGURE VII.5 – WER/enregistrement (gauche) et WER/locuteur (droite) pour chaque lot du cas *Usual_distant*

en validation croisée : 53%. L’écart du résultat moyen sur ce cas avec le cas *Unusual_close* est de plus de 24 points, tandis que l’écart entre *Unusual_close* et *Unusual_distant* est de 4,45 points. La dimension “situation de communication” et notamment la notion de rôle qui la compose semble donc effectivement être la source d’une certaine complexité pour les systèmes de reconnaissance automatique de la parole.

La comparaison des résultats avec et sans la validation croisée nous montre que la tendance d’augmentation du WER lorsque la spontanéité augmente se retrouve dans les deux cas, mais que celle-ci est moins flagrante dans le cas de la validation croisée, notamment en ce qui concerne la différence entre *Unusual_close* et *Unusual_distant* (4,45 dans le premier cas, contre 15,74 dans le second). Cela montre l’existence d’une certaine variabilité des données à l’intérieur même des cas supposés homogènes élaborés à partir de nos dimensions. Nos cas capturent une partie de la variation de la parole spontanée, mais de nombreux autres facteurs pouvant influencer les performances des systèmes ne sont pas pris en compte dans cette approche.

3 Apports d'une adaptation au domaine

Cas	WER Moyen (valid. croisée)*	Ecart-type (va-lid. croisée)*	WER Moyen**
<i>Usual_close</i>	53,1	8,14	63,74
<i>Unusual_close</i>	28,62	5,90	33,88
<i>Unusual_distant</i>	24,17	5,27	18,14

*Tels que recalculés suite à l'écartement d'un lot problématique. // **Issus du tableau VII.3, colonnes grisées

TABLE VII.9 – Récapitulatif des WER sur chacun des cas d'étude obtenus avec et sans validation croisée

3 Apports d'une adaptation au domaine

La création des ensembles de données pour chacun de nos cas d'étude nous aura demandé un long travail d'analyse et d'étiquetage des données de corpus de parole spontanée existants, nous amenant à écarter de nombreux corpus de notre étude sur l'adaptation spécifique d'un modèle pré-apris. Ainsi, après avoir rassemblé et analysé près de 1000 heures d'enregistrements, nous n'avons finalement pas toujours été en mesure d'obtenir les 10 heures de parole effective visées pour les ensembles d'apprentissage de chacun de nos cas d'étude. Si nous pensons qu'il serait bénéfique d'augmenter cette quantité de parole afin de lisser les variations dues à un ou deux enregistrements en particulier, nous rencontrons là une limite de notre étude basée sur les données.

Toutefois, nous avons à notre disposition une large quantité de données de parole spontanée qui pourrait nous permettre de spécialiser un premier système au domaine avant d'entraîner de nouveaux des systèmes spécifiques à chacun de nos cas. Notre hypothèse ici est que l'utilisation d'un système déjà adapté au domaine spontané devrait permettre une spécialisation plus efficace des systèmes par la suite.

3.1 Un système adapté au domaine spontané

Nous utilisons, pour ce système, la même architecture que celle utilisée précédemment. L'apprentissage du système se fait cette fois avec l'ensemble *All_spont*, représentatif de données de parole spontanée "tout-venant". Le système est entraîné jusqu'à ce que le WER ne s'améliore pas de plus de 0,5 points sur trois époques consécutives. Les résultats, ainsi que ceux précédemment obtenus avec le système Speechbrain état de l'art et les systèmes adaptés à chacun des trois cas d'étude, sont reportés dans le tableau VII.10.

Les résultats nous montrent que l'augmentation de la quantité de données est profitable à l'apprentissage. En effet, le système adapté au domaine avec les don-

Système	Modèle	Épq	<i>Usual - close</i>	<i>Unusual - close</i>	<i>Unusual - distant</i>	<i>All - cases</i>	<i>All - spont</i>	ETAPE	CV
Whisper	large-v2	-	65,29	38,68	27,57	43,67	40,45	28,33	12,93
SB-CV2	LB-7K-large	23	80,87	52,87	32,15	55,20	51,05	36,22	9,94
SB- <i>Usual close</i>	LB-7K-large	32	63,74	39,83	25,54	43,37	41,75	38,57	36,85
SB- <i>Unusual close</i>	LB-7K-large	32	63,51	33,88	19,12	39,12	36,92	32,02	28,48
SB- <i>Unusual distant</i>	LB-7K-large	36	70,84	34,52	18,14	40,78	36,99	32,00	28,60
*SB- <i>All_spont</i>	LB-7K-large	18	51,97	23,36	13,44	29,41	26,80	27,81	21,69

Épq : époque // Les cases grisées correspondent à un apprentissage et un décodage sur des données appartenant à un même cas. // Les valeurs en gras représentent le meilleur WER obtenu sur chacun des ensembles de test.
 // *correspond à un système adapté au domaine

TABLE VII.10 – Performances obtenues avec un modèle adapté au domaine spontané (%WER)

nées de *All_spont* est chaque fois meilleur que les autres systèmes, excepté sur Common Voice. Sur ETAPE, l'apport du système adapté au domaine est moindre comparé à l'apport sur tous les autres ensembles de test. Des améliorations importantes s'observent sur *Usual_close* (-11,54 points), sur *Unusual_close* (-10,52 points) et sur *Unusual_distant* (-4,7 points) et donc logiquement sur *All_cases* (-9,71 points). Sur *All_spont* ont retrouvé également une amélioration conséquente (-10,12 points).

Enfin, la tendance de baisse de performance lorsque la spontanéité augmente est encore et toujours observée avec ce nouveau système, tout comme avec les précédents. De même, l'écart de WER entre *Usual_close* et *Unusual_close* (28,61 points) est toujours plus élevé qu'entre *Unusual_close* et *Unusual_distant* (9,92 points).

3.2 Adaptation spécifique d'un système adapté au domaine

Après l'adaptation d'un système au domaine spontané, nous avons effectué une deuxième étape d'adaptation spécifique à chaque cas, afin de voir si le système peut parvenir à se spécialiser sur chacun de nos cas d'études. Nous reprenons le même procédé qu'en section 2 et adaptions le système SB-domaine avec les données d'apprentissage de chacun de nos cas, l'un après l'autre. Nous avons procédé à la deuxième adaptation en continuant l'apprentissage de notre système SB-domaine jusqu'à la 40ième époque tout en changeant les données d'apprentissage. Les résultats sont présentés en tableau VII.11.

Système	Modèle	Épq	Usual - close	Unusual - close	Unusual - distant	All - cases	All - spont	ETAPE	CV
Whisper	large-v2	-	65,29	38,68	27,57	43,67	40,45	28,33	12,93
SB-CV2	LB-7K-large	23	80,87	52,87	32,15	55,20	51,05	36,22	9,94
*SB- All_spont	LB-7K-large	18	51.97	23.36	13.44	29.41	26.80	27.81	21.69
*SB- All_spont- Usual close	LB-7K-large	30	50,85 $\pm 1,25$	23,80 $\pm 0,82$	13,14 $\pm 0,51$	29,17 $\pm 0,62$	26,86	28,51	22,49
*SB- All_spont- Unusual close	LB-7K-large	22	51,65 $\pm 1,28$	23,18 $\pm 0,82$	13,25 $\pm 0,51$	29,25 $\pm 0,62$	26,92	27,21	21,80
*SB- All_spont- Unusual distant	LB-7K-large	34	54,77 $\pm 1,29$	23,15 $\pm 0,78$	13,38 $\pm 0,53$	29,99 $\pm 0,66$	26,81	27,24	22,08

Épq : époque // Les cases grisées correspondent à un apprentissage et un décodage sur des données appartenant à un même cas. // Les valeurs en gras représentent le meilleur WER obtenu sur l'ensemble de test de chacun des cas étudiés. // *correspond à un système adapté au domaine

TABLE VII.11 – Performances obtenues suite à l'adaptation spécifique du système adapté au domaine (%WER)

3.2.1 Analyse des résultats

La comparaison de ces systèmes spécifiques avec le système uniquement adapté au domaine spontané nous montre que la deuxième étape d'adaptation n'améliore pas vraiment les résultats. Les WER obtenus sur un ensemble de test avec différents systèmes sont chaque fois extrêmement proches les uns des autres, ce qui montre que la spécialisation ne se fait plus. Sur nos cas, les intervalles de confiance associés aux WER en gras -censées représenter le WER le plus bas obtenu pour chaque ensemble de test- nous montrent que les meilleurs WER obtenus avec ces systèmes spécifiques ne sont pas significativement meilleurs que ceux obtenus avec le système SB-*All_spont*. L'apport d'une adaptation à un cas spécifique pour la transcription automatique de données représentatives de ce même cas ne s'observe donc pas ici, bien qu'il reste intéressant de noter que les performances sur *Usual_close* présentent une dégradation entre l'adaptation spécifique du système SB-*All_spont-Unusual_distant* et l'adaptation spécifique du système SB-*All_spont-Usual_close* (54,77% vs 50,85%). Cette absence de systématичit  montre que l'apprentissage d'un système sur une large quantité de données vari es est plus b n fique que l'apprentissage sur un cas unique avec peu de donn es. 茅tant donn  les bonnes performances obtenues sur l'ensemble des corpus de test suite 脿 l'adaptation d'un syst me avec *All_spont* uniquement, nous mettons en avant dans cette derni re tude, l'importance de la quantit , mais surtout de la variabilit  des donn es d'apprentissage, plus que de leur sp cificit .

4 Synth se

Les exp rimentations men es dans ce chapitre nous auront permis, de facon g n rale, de montrer que le WER est toujours plus  lev  pour notre cas *Usual_close* repr sentatif de situations comme des repas entre amis, et toujours moins  lev  pour notre cas *Unusual_distant* repr sentatif d'entretiens entre personnes qui ne se connaissent pas. Si la variation de la spontan t e n'est probablement pas le seul facteur faisant varier le WER, il semble tout de m me que nous soyons parvenus 脿 capturer diff rents niveaux de spontan t e 脿 travers de nos. Nous montrons 茅galement que l' cart de WER entre *Usual_close* et *Unusual_close* n'est pas  gal 脿 l' cart de WER entre *Unusual_close* et *Unusual_distant*. En effet, on observe un  cart plus important dans ce premier cas, ce qui semble indiquer que la dimension *situation de communication* est la plus compliqu e 脿 traiter pour les syst mes. De la m me facon, nous retrouvons pour chacun des syst mes utilis s des performances proches sur les ensembles *All_cases*, *Unusual_close* et *All_spont*. Si nous savons que l'ensemble *All_cases* repr sente la r union des ensembles de test *Usual_close*, *Unusual_close* et *Unusual_distant*, la repr sentation de chacun de ces cas n'est

pas connue au sein de *All_spont*. Ces performances similaires pourraient donc à la fois signifier que l'ensemble *All_spont* ne contient que des enregistrements représentatifs du cas *Unusual_close*, mais pourraient également signifier que *All_spont* est constitué d'enregistrements représentatifs de chacun des trois cas étudiés, au même titre que *All_cases*. Si tel est le cas, l'observation des performances obtenues sur chacun des cas inclus dans *All_cases* nous permet de mettre en lumière l'apport de notre approche en quatre dimensions pour l'analyse des performances de systèmes de reconnaissance automatique de la parole sur la parole spontanée. En effet, les différents WER obtenus sur *All_cases* ne sont jamais représentatifs des mauvaises performances obtenues sur le cas le plus spontané *Usual_close*, mais ne sont jamais représentatifs non plus des bonnes performances qui peuvent être obtenus sur le cas le moins spontané *Unusual_distant*. Ainsi, notre meilleur système SB-*All_spont* affiche une performance sur *All_cases* de 29,41% qui ne reflète ni les 51% de WER obtenus sur *Usual_close* ni les 13% obtenus sur *Unusual_distant*.

De façon plus précise, nous ressortons également de notre étude qu'une adaptation spécifique avec une petite quantité de parole ne permet pas de résoudre les difficultés intrinsèques contenues dans les ensembles de test représentatifs de différents niveaux de spontanéité. En effet, la différence de WER entre nos trois cas n'est pas réduite par cette approche. Ces adaptations spécifiques nous permettent cependant d'observer une légère tendance, semblant indiquer qu'un système adapté sur un certain sous-type de parole spontanée permettra ensuite d'atteindre les meilleures performances sur ce même sous-type de parole, à savoir 63% pour *Usual_close*, 33% pour *Unusual_close* et 18% pour *Unusual_distant*. Si pour le cas *Usual_close* l'apport d'une adaptation spécifique n'est pas flagrant, il est important de rappeler que c'est aussi le cas pour lequel nous avons pu réunir le moins de données d'apprentissage (6h50, contre 13h08 pour *Unusual_close* et 9h36 pour *Unusual_distant*). L'utilisation d'une technique de validation croisée nous aura permis de contourner ce manque de données et de tester la stabilité de nos systèmes. Cette première adaptation spécifique nous montre également que l'utilisation d'un système adapté au cas peu spontané dégrade les performances originellement obtenues avec Whisper (70% contre 65%). Ces observations indiquent que la spécialisation d'un système peut à la fois être bénéfique et néfaste, et n'est à encourager que dans des cas applicatifs restreints.

La validation croisée, nous montre quant à elle que pour chaque cas, nous retrouvons une stabilité des différents systèmes appris, confirmée par les performances sur les corpus ETAPE et Common Voice, mais aussi une certaine instabilité des WER sur les différents lots. Cette instabilité témoigne de la variabilité capturée au sein de nos cas considérés comme homogènes. Celle-ci apparaît d'autant plus fortement que les ensembles de test sont de petite taille : les enregistrements ou

les locuteurs qui se démarquent des autres influencent d'autant plus les résultats. Enfin, nous retrouvons, avec cette validation croisée, la différence de WER en fonction des niveaux de spontanéité capturés dans nos cas, avec toutefois cette fois un écart beaucoup plus resserré entre *Unusual_close* et *Unusual_distant* que ce que nous avons pu observer auparavant (28% vs 24%).

Enfin, l'étape d'adaptation au domaine spontané nous aura permis d'obtenir les meilleures performances sur chacun de nos cas (51% sur *Usual_close*, 23% sur *Unusual_close* et 13% sur *Unusual_distant*). L'adaptation spécifique qui aura suivi cet apprentissage ne montre toutefois pas d'apport, ni de réelle spécialisation des différents systèmes à un sous-type de parole spontanée. En effet, nous obtenons 23% de WER sur *Unusual_close* et 13% de WER sur *Unusual_distant* pour chacun des systèmes. Cependant, il faut noter que la spécialisation à *Unusual_distant* dégrade encore une fois les performances sur *Usual_close*. Cette dernière expérimentation nous montre donc la prévalence de l'utilisation d'une grande quantité de données diversifiées sur l'utilisation de systèmes spécifiques, adaptés avec peu de données. Le tableau VII.12 reporte les différents résultats obtenus au cours de ces différentes expérimentations.

Système	%WER		
	<i>Usual_close</i>	<i>Unusual_close</i>	<i>Unusual_distant</i>
Whisper	65	38	27
SB-spécifique*	63	33	18
SB-spécifique* validation croisée	53	28	24
SB- <i>All_spont</i>	51	23	13
SB- <i>All_spont</i> -spécifique*	50	23	13

*Nous rapportons ici les meilleurs WER obtenus sur les cas grâce à des systèmes adaptés aux différents cas

TABLE VII.12 – Performances obtenues sur les différents cas étudiés au fil des expérimentations menées

Conclusion, limites et perspectives

Nous avons présenté, dans ce manuscrit, nos travaux pour l'étude inter-corpus des performances de systèmes de reconnaissance automatique en fonction de différents sous-types de parole spontanée.

Après une étude de l'état de l'art présentant les apports des systèmes entièrement neuronaux et des modèles pré-appris sur les performances de systèmes de reconnaissance automatique de la parole, la première étape de ce travail de thèse s'est inscrite dans le cadre du projet LeBenchmark, dont les modèles de type Wav2vec 2.0, appris sur du français, ont été mis à disposition de la communauté et présentent l'avantage de pouvoir être adaptés avec peu de données. Si notre apport au sein de ce projet aura été le rassemblement et la préparation de corpus existants, ainsi que la constitution d'un corpus d'environ 7 000 heures de parole lue, nos collègues auront pu démontrer l'intérêt de ces modèles pour une tâche de reconnaissance automatique de la parole avec une architecture *end-to-end*, et ce, à la fois sur un corpus de lecture (Common Voice, 9% de WER) et sur un corpus contenant de la parole préparée et spontanée (ETAPE, 23% de WER). Les modèles, une fois adaptés, présentent de façon générale meilleures performances que lors de l'utilisation de modèles multilingues (Common Voice : 19%, ETAPE : 61%) ou appris sur l'anglais (Common Voice : 14%, ETAPE : 44%), ou encore que l'extraction de Mel Filter Banks (Common Voice : 20%, ETAPE : 54%). Ainsi, nous avons fait le choix de placer ces modèles au coeur de notre travail de recherche.

La deuxième étape de ce travail de thèse a ensuite consisté en une étude approfondie du concept de parole spontanée pour la prise en compte de sa variabilité dans les études en reconnaissance automatique de la parole. En effet, bien que certains corpus de référence en français incluent de la parole spontanée, ils ne capturent pas toute la diversité de situations dans lesquelles peut apparaître la parole spontanée. Ceux-ci se limitent principalement à de la parole médiatique, contexte particulier dans lequel les locuteurs préparent leurs interventions et montrent ainsi une parole construite et bien articulée qui leur permet d'être compris de tous les auditeurs. C'est pourquoi nous nous sommes tournés vers les corpus de linguistique, captu-

rant à la fois des interactions usuelles, quotidiennes, ainsi que des entretiens, et offrant donc plus de variété. Si ces corpus sont souvent de taille assez réduite en comparaison à ce qui est utilisé d'ordinaire en reconnaissance automatique de la parole (>100h), la mise à disposition des modèles LeBenchmark pré-appris sur du français et pouvant être adaptés avec une petite quantité de parole aura rendu possible ce travail de recherche. Cependant, si la diversité représentée par ces corpus permet d'étudier la reconnaissance automatique de la parole spontanée de façon plus précise, elle rend nécessaire une catégorisation des différents sous-types de parole spontanée, afin de pouvoir faire ressortir des ensembles de données homogènes. C'est ainsi que s'est formée notre première question de recherche : *quels facteurs de variation de la parole spontanée sélectionner, dans le but de constituer des sous-groupes de données similaires, et de positionner ces sous-groupes les uns par rapport aux autres ?*

Afin de répondre à cette première question, nous avons tout d'abord recueilli des données provenant de corpus de linguistique tels que le CFPP, le CFPB, TCOF, ESLO2... qui se sont faits les témoins de la variabilité des sous-types de parole spontanée (conversations entre amis, entre collègues, entretiens formels et informels, etc.). En croisant les métadonnées des corpus avec les facteurs de variation de la parole spontanée identifiés dans les travaux en linguistique, sociolinguistique et stylistique, nous avons pu faire ressortir quatre dimensions de variation exploitables :

- la situation de communication : rapport symétrique entre les locuteurs ou non, et importance du lieu (quotidien/familier *vs* professionnel/institutionnel)
- le degré d'intimité entre les locuteurs : de très proches à inconnus
- le canal de communication : en face-à-face ou à distance, avec ou sans visio
- le type de communication : interpersonnelle, de groupe ou de masse

À partir des dimensions définies précédemment, nous avons ainsi commencé par identifier et étiqueter des ensembles de données supposés représentatifs de différents niveaux de spontanéité. Trois cas d'étude ont été définis, que nous décrivons ci-après et plaçons graphiquement sur un espace de variation en quatre dimensions (voir figure VII.6) :

- ***Usual_close*** : Amis proches ou membres d'une même famille et situation de communication symétrique dans un lieu quotidien/familier.
- ***Unusual_close*** : Amis proches ou membres d'une même famille qui échangent sur un lieu de travail ou au sein d'une institution. Les locuteurs ne sont pas dans une situation symétrique.

- ***Unusual_distant*** : Deux personnes qui ne se connaissent pas et qui échangent au travail ou au sein d'une institution. Les locuteurs sont dans une situation asymétrique.

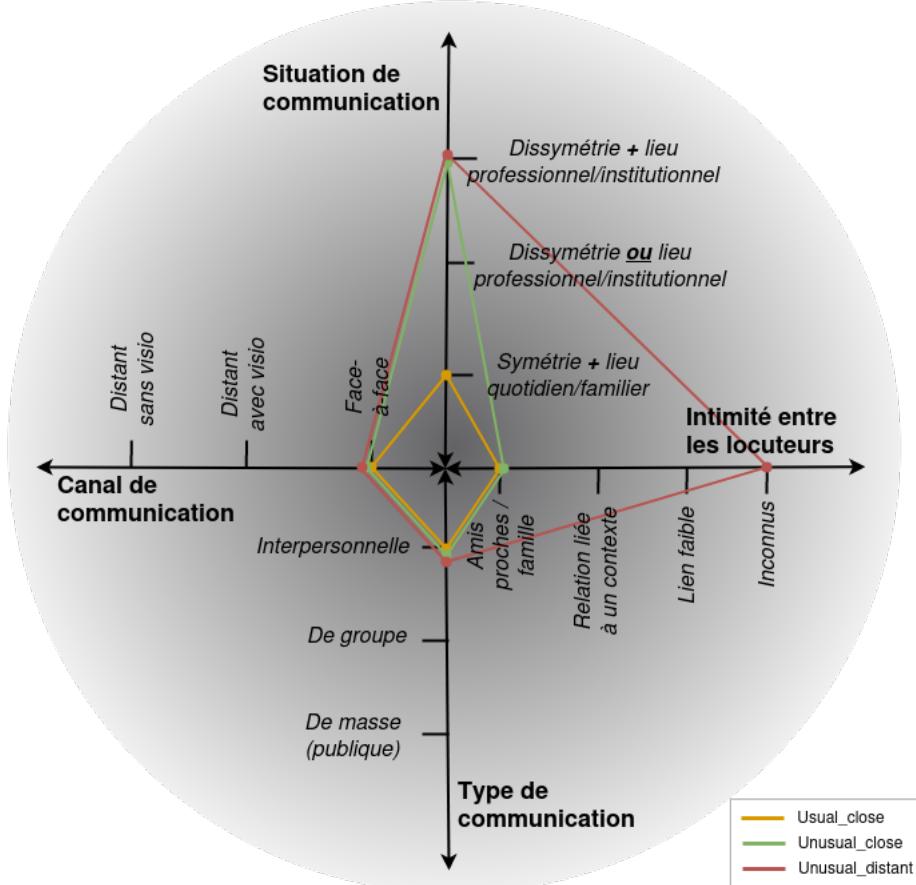


FIGURE VII.6 – Positionnement des cas d'étude dans l'espace en quatre dimensions de variation de la parole spontanée tel que défini dans ce travail de thèse

Chacun de ces cas a été découpé en ensembles d'apprentissage, de développement et de test. Nous avons ainsi pu étiqueter 6h50 de parole pour l'ensemble d'apprentissage du cas *Usual_close*, 13h08 pour *Unusual_close* et 9h36 pour *Unusual_distant*. Il est important de noter que sans le travail de précision des linguistes et leur connaissances approfondie de l'objet de recherche qu'est la parole, l'identification des facteurs de variation de la parole spontanée au sein de ces données aurait été bien plus laborieuse. Les données de parole spontanée non étiquetées ont ensuite été compilées dans un ensemble complet nommé “All_spont” et totalisant 337 heures de parole.

La troisième étape de notre étude aura consisté en l'étude inter-corpus des perfor-

mances de systèmes de reconnaissance automatique de la parole dans différentes conditions : lors de l'adaptation d'un modèle pré-appris avec peu de données étiquetées en fonction des dimensions déterminées *versus* lors de l'adaptation d'un modèle pré-appris avec une grande quantité de données moins contrôlées (c'est-à-dire non étiquetées en fonction des dimensions). Nous avons étudié, tout d'abord, l'adaptation d'un modèle pré-appris avec une petite quantité de données, adaptant chaque fois un modèle LeBenchmark avec l'un des ensembles d'apprentissage de nos cas d'étude. Nous souhaitions, par cette méthode, apporter une réponse à notre deuxième question de recherche : *l'adaptation spécifique d'un modèle pré-appris avec une faible quantité de données représentatives de différents niveaux de spontanéité peut-elle améliorer les performances des systèmes sur différents sous-types de parole spontanée ?* Afin de nous constituer une référence, nous avons tout d'abord utilisé deux systèmes état de l'art : l'un appris à partir d'une recette Speechbrain (réutilisée par la suite) et utilisant un modèle LeBenchmark adapté sur Common Voice, l'autre étant Whisper, un système multilingue appris sur 680 000 heures de parole. Il ressort, de cette première évaluation, que le WER augmente lorsque la spontanéité capturée dans nos cas d'étude augmente. En effet, notre évaluation a montré des WER variables selon les sous-types de parole spontanée étiquetés. Ainsi, on observe les WER les plus hauts sur *Usual_close* (65% pour Whisper, 80% pour le système Speechbrain) et les plus bas sur *Unusual_distant* (27% pour Whisper, 32% pour le système Speechbrain). Le WER observé sur *Unusual_close* est quant à lui de 38% avec Whisper et 52% pour le système Speechbrain. Whisper apparaissait donc comme le système état de l'art le plus performant, ce qui n'est pas étonnant étant donné que le modèle LeBenchmark utilisé pour l'entraînement du système Speechbrain de référence était adapté sur de la parole lue.

Les différentes adaptations du modèle LeBenchmark avec, cette fois, les données de nos différents cas d'étude ont, quant à elles, chaque fois montré une amélioration des résultats sur la parole spontanée par rapport aux deux systèmes de référence. Ainsi, si l'amélioration est faible sur *Usual_close* avec un WER minimal de 63%, les gains sur *Unusual_close* et sur *Unusual_distant* sont plus importants : ce sont 33% et 18% de WER qui sont obtenus, ce qui correspond à des améliorations respectives de 5 et 9 points de WER par rapport au meilleur système état de l'art. Les différentes adaptations ont également montré qu'adapter un système à un sous-type de parole spontanée spécifique peut faire stagner, voire dégrader les performances sur d'autres sous-types. Ainsi, une adaptation avec *Usual_close* donne, sur *Unusual_close*, quasiment la même performance qu'avec Whisper (39% vs 38%). Plus flagrant encore, une adaptation avec *Unusual_distant* dégrade le WER sur *Usual_close* de 5 points, le faisant passer de 65% à 70%. Nous avons également observé, à ce stade, que le système présentant un apport global sur tous les cas d'études était celui adapté sur les données de type *Unusual_close*.

Cependant, ces observations sont à mettre en parallèle avec les différentes quantités de données ayant pu être rassemblées pour ces adaptations spécifiques. Nous avons effectivement pu réunir deux fois plus de données d'apprentissage pour le cas *Unusual_close* que pour *Usual_close*. Afin de vérifier la stabilité des résultats, nous avons ainsi procédé à une étape de validation croisée. Celle-ci aura montré, pour chacun des cas, une stabilité des différents systèmes appris, mais aussi une certaine variabilité des WER obtenus sur les différents lots de test. Ainsi, la validation croisée nous montre que les performances se situent plutôt autour de 53% pour *Usual_close*, 28% pour *Unusual_close* et 24% pour *Unusual_distant*.

Après cette étude sur l'adaptation spécifique d'un modèle, nous avions pour but de répondre à notre troisième et dernière question de recherche : *quel est impact de l'adaptation, avec une bien plus grande quantité de données, moins bien contrôlées qu'auparavant, d'un modèle pré-appris sur les performances de reconnaissance de différents sous-types de parole spontanée ?* Nous avons, cette fois, adapté le modèle LeBenchmark avec les données de *All_spont* (337 heures de parole), avant d'effectuer une seconde fois les différentes adaptations spécifiques en fonction de nos cas. Nous montrons que l'adaptation au domaine spontané avec l'ensemble *All_spont* permet de réduire très significativement le WER, notamment pour les sous-types *Usual_close* et *Unusual_close* pour lesquels on observe un gain important par rapport à ce qui peut être obtenu avec les systèmes spécifiques : 51% vs 63% pour ce premier, et 23% vs 33% pour ce deuxième. Le WER sur le cas *Unusual_distant* montre une baisse plus légère (13% vs 18%), ce qui reste une bonne amélioration si l'on considère qu'il était déjà plutôt bas. Les adaptations spécifiques supplémentaires ont, quant à elles, montré une limite à l'amélioration des résultats. Le WER aura ainsi tendance à stagner, voire à se dégrader sur le cas *Usual_close*, par rapport à ce qui était obtenu avec la seule étape d'adaptation au domaine. Cela peut suggérer à la fois une bonne représentation de nos cas d'étude dans *All_spont*, mais aussi un manque de données représentatives de chacun de nos cas, qui pourrait limiter cette deuxième étape d'amélioration des performances des systèmes.

D'un point de vue global, on observe avec chacun des systèmes utilisés pour ce travail de thèse une différence importante d'écart de WER entre *Usual_close* et *Unusual_close* d'un côté et *Unusual_close* et *Unusual_distant* d'un autre côté, ce qui semble indiquer que les systèmes de reconnaissance automatique de la parole ont plus de difficultés avec le haut niveau de spontanéité induit par la situation de communication, que par le degré d'intimité entre les locuteurs. Cela a notamment été rendu très visible avec les WER moyens obtenus par la validation croisée (4 points d'écart entre *Unusual_distant* et *Unusual_close*, contre 25 points d'écart entre *Unusual_close* et *Usual_close*). Notre travail montre également l'importance

de caractériser la parole spontanée pour l'évaluation des performances des systèmes, plus que pour leur amélioration. En effet, la moyenne des WER de nos différents cas, observée au travers de l'ensemble *All_cases*, nous montre que les performances obtenues sur cet ensemble ne sont jamais représentatives de la réelle variabilité des WER sur différents sous-types de parole spontanée. Ainsi, le WER obtenu sur *All_cases* avec le système adapté au domaine est de 29%, alors même que cet ensemble contient des données de type *Usual_close* avec un WER de 51%, des données de type *Unusual_close* avec un WER de 23% et des données de type *Unusual_distant* avec un WER de 13%.

Limites et perspectives

Notre étude souligne l'actuelle nécessité de la création de grands jeux de données diversifiés pour l'étude de la reconnaissance automatique de la parole spontanée. En effet, malgré les presque 1000 heures de parole rassemblées au début de cette thèse, seule une dizaine d'heures auront pu être étiquetées et utilisées au sein de nos différents cas. Les 10 heures de parole préalablement souhaitées pour l'étude de l'impact des différentes adaptations spécifiques ont notamment été atteintes que partiellement pour les cas *Usual_close* et *Unusual_distant*. Bien que nous ayons fait usage d'une technique de validation croisée pour nous assurer de la stabilité des résultats, des recherches complémentaires, avec plus de données, permettraient d'approfondir la compréhension des adaptations spécifiques selon les sous-types de parole spontanée. Les projets en cours comme le corpus Common Voice pour la parole spontanée⁹ et le programme culturel et scientifique "Ecouter-Parler"¹⁰ initié par la DGLFLF sont des initiatives prometteuses pour l'expansion des données de parole spontanée. Ces données pourraient s'avérer cruciales pour l'analyse des performances de systèmes de reconnaissance automatique de la parole en fonction de différents types de parole spontanée, mais aussi pour l'amélioration des systèmes, ce qui reste encore un défi.

Une autre limite de notre étude réside en l'absence d'utilisation d'un modèle de langue. Nous ne souhaitions en effet pas biaiser le décodage et avoir la possibilité d'étudier le seul impact de différentes adaptations d'un modèle pré-appris. Cependant, nous pensons que l'utilisation d'un modèle de langue pourrait améliorer les performances des systèmes, notamment sur la parole peu spontanée.

Il est important de noter également que nous ne couvrons, au travers de nos cas d'étude, qu'une surface limitée de l'espace de variation de la parole spontanée en quatre dimensions. Pour pallier cette limite, nous avons choisi de nous concentrer sur deux dimensions, et d'en étudier les extrêmes afin de baser nos recherches sur

9. <https://foundation.mozilla.org/fr/blog/common-voice-2024-roadmap/>

10. <https://ecouter-parler.fr/presentation/>

trois cas bien séparés. Les recherches futures sur ce sujet devront explorer les dimensions non étudiées ici (canal de communication et type de communication), ainsi que les variations plus fines de la spontanéité. Une expansion et amélioration des dimensions déterminées ici est également envisageable, notamment via l'analyse de travaux en sociologie afin de mieux comprendre les rapports de rôle, mais aussi la question des différents groupes socio-culturels, dont on sait, grâce aux études en sociolinguistique, que les habitudes langagières peuvent largement différer. De même, la prise en compte d'autres conditions de prises de parole telles que le *streaming*, le direct, le différé ou encore les messages vocaux, aujourd'hui largement utilisés via les services de messagerie instantanée (Glikman et Fauth, 2022). Une extension de notre approche pourrait ainsi mener à l'élaboration d'un benchmark, grâce auquel les systèmes de reconnaissance automatique de la parole ne seraient plus évalués avec une seule valeur de WER, mais plutôt sur une diversité de situations, permettant de rendre compte de façon plus réelle et plus fiable des performances des systèmes.

Enfin, si nous ne mesurons pas, à l'intérieur de nos cas d'étude, la variation de certaines caractéristiques reportées dans la littérature comme liées à la spontanéité (variation du débit de parole, hypoarticulation, disfluences...), il pourrait être intéressant de le faire, notamment sur une plus large quantité de données étiquetées. Ceci permettrait de croiser notre espace de variation avec les études en linguistique sur la parole spontanée. L'ensemble de ces travaux pourrait également être étendu à l'étude de la variation de la parole spontanée pour d'autres langues que le français, le rapport aux normes sociales pouvant notamment varier en fonction des pays. Les usages langagiers sont en effet structurés par des "normes" qui dépendent des groupes sociaux et des cultures. Une part du défi à venir pourrait donc consister en la proposition de systèmes de transcriptions adaptés à ces différents usages, étant donné que leur diffusion large et parfois multilingue implique forcément une utilisation au sein de situations langagières très disparates, reflétant l'usage que les humains font du langage au quotidien.

Références

- Ossama ABDEL-HAMID, Abdel-rahman MOHAMED, Hui JIANG, Li DENG, Gerald PENN et Dong YU : Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech and Language processing*, 22(10): 1533–1545, 2014. 14, 16
- Anne ABEILLÉ et Danielle GODARD : *La Grande Grammaire du français*. 2021. ISBN 978-2-330-14239-1. 54
- M. Adda DECKER : De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. *In Actes des XXVIe Journées d'Etude sur la Parole (JEP2006)*, Dinard, France, 2006. 40, 47, 52, 64
- Martine ADDA-DECKER : Problèmes posés par le schwa en reconnaissance et en alignement automatiques de la parole. *In Actes des 5èmes Journées Linguistiques de Nantes*, 2007. 7
- Martine Adda DECKER, Elisabeth Delais ROUSSARIE, Fougeron CECILE, Cédric GENDROT et Lori LAMEL : La liaison dans la parole spontanée familiale : explorations semi-automatiques de grands corpus. *In Actes de la conférence conjointe JEP-TALN-RECITAL*, Grenoble, France, 2012. 49
- Martine ADDA-DECKER, B HABERT, C BARRAS, G ADDA, Philippe Boula de MAREÜIL et Patrick PAROUBEK : A disfluency study for cleaning spontaneous speech automatic transcripts and improving speech language models. *In proceedings of ISCA Tutorial and Research Workshop on Disfluency in Spontaneous Speech*, janvier 2003. 44, 46, 52, 54, 55, 56, 59, 62
- Martine ADDA-DECKER et Lori LAMEL : *in Lexicon Development for Speech and Language Processing*, chapitre The Use of Lexica in Automatic Speech Recognition, pages 235–266. Text, Speech and Language Technology series. Springer Netherlands, 2000. ISBN 9789401094580. 8, 52
- Asif AGHA : *Language and Social Relations*. Cambridge University Press, 2006. 62

RÉFÉRENCES

- A.M. AHMAD, S. ISMAIL et D.F. SAMAON : Recurrent neural network with back-propagation through time for speech recognition. In *IEEE International Symposium on Communications and Information Technology, 2004. ISCIT 2004.*, volume 1, octobre 2004. 12
- Dario ALBESANO, Jesús ANDRÉS-FERRER, Nicola FERRI et Puming ZHAN : On the prediction network architecture in rnn-t for asr. In *proceedings of INTERSPEECH 2022*, Incheon, Korea, septembre 2022. 19
- Robin ALGAYRES, Mohamed Salah ZAIEM, Benoît SAGOT et Emmanuel DUPOUX : Evaluating the reliability of acoustic speech embeddings. In *proceedings of INTERSPEECH 2020 - Annual Conference of the International Speech Communication Association*, Shanghai / Virtual, China, octobre 2020. 24
- Sharifa ALGHOWINEM, Roland GOECKE, Michael WAGNER, Julien EPPS, Michael BREAKSPEAR, Gordon PARKER *et al.* : From joyous to clinically depressed : Mood detection using spontaneous speech. In *FLAIRS*, 2012. 65
- Eduardo ALVES : Earthquake Forecasting Using Neural Networks : Results and Future Work. *Nonlinear Dynamics*, 44:341–349, juin 2006. xv, 10
- Tawfiq AMMARI, Jofish KAYE, Janice Y. TSAI et Frank BENTLEY : Music, search and iot : How people (really) use voiceassistants. In *proceedings of ACM Transactions on Computer-Human Interaction*, 2019. 1
- Hanne Leth ANDERSEN : *La complexité en langue et son acquisition*, chapitre Complexité des systèmes linguistiques : La complexité à l'oral et à l'écrit, pages 19–36. Numéro 394. Towarzystwo Naukowe KUL, 2012. 60, 61, 63
- Virginie ANDRÉ : FLEURON : Français Langue Étrangère Universitaire – Ressources et Outils Numériques. Origine, démarches et perspectives. *Mélanges CRAPEL*, 37:69–92, 2016. 120
- Virginie ANDRÉ : Un corpus multimédia pour apprendre à interagir en situations universitaires en France. In *Troisième colloque international de l'ATPF “Enseigner le français : s’engager et innover”*, Bangkok, Thailand, octobre 2017. 120, 126, 127, 128
- Virginie ANDRÉ et Emmanuelle CANUT : Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français). *Pratiques [Online]*, (147-148), 2010. 65, 120
- C. S. ANOOP et A. G. RAMAKRISHNAN : CTC-Based End-To-End ASR for the Low Resource Sanskrit Language with Spectrogram Augmentation. In *procee-*

RÉFÉRENCES

- dings of the 2021 National Conference on Communications (NCC)*, juillet 2021. 18
- Marcia APONE, Tom Brooks et Trisha O'CONNELL : Caption viewer survey : Error ranking of real-time captions inlive television news programs. Caption Accuracy Metrics Project, décembre 2010. 31
- Rosana ARDILA, Megan BRANSON, Kelly DAVIS, Michael KOHLER, Josh MEYER, Michael HENRETTY, Reuben MORAIS, Lindsay SAUNDERS, Francis TYERS et Gregor WEBER : Common Voice : A Massively-Multilingual Speech Corpus. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, mai 2020. 34
- Thirunavukkarasu ARUN BABU, Changhan WANG, Andros TJANDRA, Kushal LAKHOTIA, Qiantong XU, Naman GOYAL, Kritika SINGH, Patrick PLATEN, Yatharth SARAF, Juan PINO, Alexei BAEVSKI, Alexis CONNEAU et Michael AULI : XLS-R : Self-supervised Cross-lingual Speech Representation Learning at Scale. *In proceedings of INTERSPEECH 2022*, Incheon, Korea, novembre 2022. 26
- T. ATHANASELIS, S. BAKAMIDIS, I. DOLOGLOU, R. COWIE, E. DOUGLAS-COWIE et C. COX : ASR for emotional speech : Clarifying the issues and enhancing performance. *Neural Networks*, 18(4):437–444, mai 2005. 3
- Mathieu AVANZI, Marie-José BÉGUELIN et Federica DIÉMOZ : De l'archive de parole au corpus de référence : la base de données orales du français de Suisse romande (OFRON). *Corpus*, (15), octobre 2016. 120
- Alexei BAEVSKI, Henry ZHOU, Abdelrahman MOHAMED et Michael AULI : wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. *In proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, Vancouver, Canada, juin 2020. xix, 24, 27, 29, 37, 38, 75, 83, 85, 100, 104, 109, 111, 115, 152
- Dzmitry BAHDANAU, Kyunghyun CHO et Y. BENGIO : Neural Machine Translation by Jointly Learning to Align and Translate. <https://doi.org/10.48550/arXiv.1409.0473>, septembre 2014. xv, 20, 21
- H. BALDAUF-QUILLIATRE, I. Colón de CARVAJAL, C. ETIENNE, E. JOUIN-CHARDON, S. TESTON-BONNARD et V. TRAVERSO : CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. *Corpus*, (15), octobre 2016. 119
- Lukas BALTHASAR et Michel BERT : La plateforme « Corpus de langues parlées en interaction » (CLAPI). *Lidil. Revue de linguistique et de didactique des langues*, (31):13–33, juin 2005. 119

RÉFÉRENCES

- Jeong-Uk BANG, Seung YUN, Seung-Hi KIM, Mu-Yeol CHOI, Min-Kyu LEE, Yeo-Jeong KIM, Dong-Hyun KIM, Jun PARK, Young-Jik LEE et Sang-Hun KIM : KsponSpeech : Korean Spontaneous Speech Corpus for Automatic Speech Recognition. *Applied Sciences*, 10(19):6936, janvier 2020. 47
- Ankur BAPNA, Colin CHERRY, Yu ZHANG, Ye JIA, Melvin JOHNSON, Yong CHENG, Simran KHANUJA, Jason RIESA et Alexis CONNEAU : mslam : Massively multilingual joint pre-training for speech and text. <https://doi.org/10.48550/arXiv.2202.01374>, 2022. 26
- Loïc BARRAULT : *Diagnostic pour la combinaison de systèmes de reconnaissance automatique de la parole*. phdthesis, Université d'Avignon, 2008. NNT : . tel-00424699. 3
- Olivier BAUDE et Céline DUGUA : Les ESLO, du portrait sonore au paysage digital. *Corpus*, (15), octobre 2016. 122
- Thierry BAZILLON : *Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique*. phdthesis, Université du Maine, 2011. 48
- Thierry BAZILLON, Yannick ESTÈVE et Daniel LUZZATI : Manual vs Assisted Transcription of Prepared and Spontaneous Speech. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, mai 2008a. 54, 60
- Thierry BAZILLON, Vincent JOUSSE, Frédéric BÉCHET, Georges LINARÈS et Daniel LUZZATI : La parole spontanée : transcription et traitement. *Traitemet Automatique des Langues*, volume 49, 2008b. xvi, 46, 48, 49, 50, 52, 53, 54, 56, 57
- Howard S. BECKER et Jean-Michel CHAPOULIE : Le double sens de « outsider ». In *Outsiders, Leçons De Choses*, pages 25–42. Éditions Métailié, Paris, 1985. ISBN 9782864249184. 125
- Y. BENGIO, P. SIMARD et P. FRASCONI : Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, mars 1994. 12
- Christophe BENZITOUN, Jeanne-Marie DEBAISIEUX et Henri-José DEULOFEU : Le projet ORFÉO : un corpus d'étude pour le français contemporain. *Corpus*, (15), octobre 2016. 119
- Roxane BERTRAND, Morgane ADER, Philippe BLACHE, Gaëlle FERRÉ, Robert ESPESSER et Stéphane RAUZY : Représentation, édition et exploitation de

RÉFÉRENCES

- données multimodales : le cas des backchannels du corpus CID. *Cahiers de Linguistique*, 33(2):183–212, 2009. 50
- Roxane BERTRAND, Philippe BLACHE, Robert ESPESSER, Gaëlle FERRÉ, Christine MEUNIER, Béatrice PRIEGO-VALVERDE et Stéphane RAUZY : Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle. *Revue TAL : traitement automatique des langues*, 49 (3):105–134, 2008. 122
- Roxane BERTRAND, James Sneed GERMAN, Sophie HERMENT, Daniel J. HIRST, Amandine MICHELAS, Caterina PETRONE, Cristel PORTES, Anne TORTEL et Pauline WELBY : La prosodie au Laboratoire Parole et Langage : histoire, recherches actuelles et perspectives. *Travaux Interdisciplinaires sur la Parole et le Langage*, 38, 2022. 62
- Brigitte BIGI et Christine MEUNIER : Automatic Segmentation of Spontaneous Speech. *Revista De Estudos Da Linguagem*, 26(4):1489–1530, octobre 2018. 47, 50, 52, 53, 55
- Mireille BILGER et Paul CAPPEAU : L'oral ou la multiplication des styles. *Langage et societe*, numéro 109:pages 13–30, 2004. 62
- Mireille BILGER et Collectif Du Groupe d'Etude Sur Les Données Orales GEDO : Transcription de l'oral et interprétation. Illustration de quelques difficultés. *Recherches sur le français parlé*, (14):pp. 57–87, 1997. 57
- Claire BLANCHE-BENVENISTE : *Le français parlé. Etudes grammaticales*. 1990. 50
- Claire BLANCHE-BENVENISTE et Mireille BILGER : "Français parlé - oral spontané". Quelques réflexions. *Revue française de linguistique appliquée*, IV(2):21–30, 1999. 53, 55, 62
- Kübra BODUR, Christine MEUNIER et Corinne FREDOUILLE : Formes réduites en conversation : caractéristiques des séquences et des locuteurs. *In proceedings XXXIVe Journées d'Études sur la Parole JEP 2022*, 2022. 44, 48, 59, 63
- Marcelly Zanon BOITO, William HAVARD, Mahault GARNERIN, Éric LE FERRAND et Laurent BESACIER : MaSS : A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. *In Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France, mai 2020. European Language Resources Association. 102
- Catherine T. BOLLY, George CHRISTODOULIDES et Anne Catherine SIMON : Dis-

RÉFÉRENCES

- fluences et vieillissement langagier. De la base de données VALIBEL aux corpus outillés en français parlé. *Corpus*, (15), octobre 2016. ISSN 1638-9808. 121
- Heather BORTFELD, Silvia D. LEON, Jonathan E. BLOOM, Michael F. SCHOBER et Susan E. BRENNAN : Disfluency rates in conversation : Effects of age, relationship, topic, role, and gender. *Language and Speech*, 44(2):123–147, 2001. 52, 54, 59, 63, 64
- Philippe Boula de MAREÜIL : Qu'est-ce qu'un (phono)style ? *Nouveaux cahiers de linguistique française*, (31):9–19, janvier 2014. 44
- Pierre BOURDIEU : *La misère du monde*. Le Seuil, Paris, 1993. 126
- Herve BOURLARD et Nelson MORGAN : *Connectionist Speech Recognition : A Hybrid Approach*. Kluwer Press, 1994. ISBN 978-1-4613-6409-2. 2, 15
- Florian BOYER et Jean-Luc ROUAS : End-to-End Speech Recognition : A review for the French Language. <https://doi.org/10.48550/arXiv.1910.08502>, octobre 2019. 36
- Florian BOYER, Yusuke SHINOHARA, Takaaki ISHII, Hirofumi INAGUMA et Shinji WATANABE : A Study of Transducer based End-to-End ASR with ESPnet : Architecture, Auxiliary Loss and Decoding Strategies. In *proceedings of 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, janvier 2022. 11
- S. BRANCA-ROSOFF, S. FLEURY, F. LEFEUVRE et M. PIRES : Discours sur la ville. Présentation du Corpus de Français parlé Parisien des années 2000 (CFPP2000). <http://cfpp2000.univ-paris3.fr/CFPP2000.pdf>, 2012. 119
- Ilana BROMBERG, Qian QIAN, Jun HOU, Jinyu LI, Chengyuan MA, Brett MATTHEWS, Antonio MORENO-DANIEL, Jeremy MORRIS, Marco SINISCALCHI, Yu TSAO et Yu WANG : Detection-based ASR in the automatic speech attribute transcription project. In *proceedings of INTERSPEECH 2007*, août 2007. 6
- Tanja BÄNZIGER, Marcello MORTILLARO et Klaus SCHERER : Introducing the Geneva Multimodal Expression Corpus for Experimental Research on Emotion Perception. *Emotion (Washington, D.C.)*, 12(5):1161–79, 2012. 102
- Frédéric BÉCHET : Lia_phon - un système complet de phonétisation de textes. *Traitemen Automatique des Langues*, 42(1):47–67, 2001. ISSN 1965-0906. 8
- Lolita BÉRARD : La partie orale du Corpus d'Étude pour le Français Contemporain (CÉFC). *Langages*, 219(3):25–37, 2020. xvii, 121, 122
- Geneviève CAELEN-HAUMONT et Bernard BEL : Le caractère spontané dans la

RÉFÉRENCES

- parole et le chant improvisés : de la structure intonative au mélisme. *Revue PAROLE*, (15-16):251–302, 2000. 44
- Arnaldo CANDIDO JUNIOR, Edresson CASANOVA, Anderson SOARES, Frede-
rico Santos de OLIVEIRA, Lucas OLIVEIRA, Ricardo Corso Fernandes JUNIOR,
Daniel Peixoto Pinto da SILVA, Fernando Gorgulho FAYET, Bruno Baldissera
CARLOTTO, Lucas Rafael Stefanel GRIS et Sandra Maria ALUÍSIO : CORAA
ASR : a large corpus of spontaneous and prepared speech manually validated for
speech recognition in Brazilian Portuguese. *Language Resources and Evaluation*,
57(3):1139–1171, 2023. 47
- Stanley F. CHEN et Joshua GOODMAN : An empirical study of smoothing tech-
niques for language modeling. *Computer Speech & Language*, 13(4):359–394,
octobre 1999. ISSN 0885-2308. 8
- Zhehuai CHEN, Yu ZHANG, Andrew ROSENBERG, Bhuvana RAMABHADRAN, Pe-
dro MORENO, Ankur BAPNA et Heiga ZEN : MAESTRO : Matched Speech Text
Representations through Modality Matching. In *proceedings of INTERSPEECH*
2022, Incheon, Korea, septembre 2022. 26
- Jianpeng CHENG, Li DONG et Mirella LAPATA : Long Short-Term Memory-
Networks for Machine Reading. In *Proceedings of the 2016 Conference on Em-
pirical Methods in Natural Language Processing*, Austin, Texas, novembre 2016.
22
- Kyunghyun CHO, Bart van MERRIËNBOER, Caglar GULCEHRE, Dzmitry BAHDA-
NAU, Fethi BOUGARES, Holger SCHWENK et Yoshua BENGIO : Learning Phrase
Representations using RNN Encoder–Decoder for Statistical Machine Transla-
tion. In *Proceedings of the 2014 Conference on Empirical Methods in Natural
Language Processing (EMNLP)*, Doha, Qatar, octobre 2014. xv, 20
- Jan CHOROWSKI, Ron J. WEISS, Samy BENGIO et Aaron van den OORD :
Unsupervised Speech Representation Learning Using WaveNet Autoencoders.
IEEE/ACM Transactions on Audio, Speech and Language Processing, 27(12):
2041–2053, décembre 2019. 24
- George CHRISTODOULIDES : Variation prosodique des styles de parole et interface
syntaxe-prosodie : Étude sur corpus à grande échelle. In *actes des 6e confé-
rence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement
Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étu-
diants Chercheurs en Informatique pour le Traitement Automatique des Langues
(RÉCITAL, 22e édition)*, Nancy, France, juin 2020. 49, 60
- Chloé CLAVEL, Gilles ADDA, Frederik CAILLIAU, Martine GARNIER-RIZET,

RÉFÉRENCES

Ariane CAVET, Géraldine CHAPUIS, Sandrine COURCINOUS, Charlotte DANESI, Anne-Laure DAQUO, Myrtille DELEDOSSI, Sylvie GUILLEMIN-LANNE, Marjorie SEIZOU et Philippe SUGNARD : Spontaneous speech and opinion detection : mining call-centre transcripts. *Language Resources and Evaluation*, 47(4):1089–1125, 2013. 44, 50, 52, 54, 55

Alexis CONNEAU, Alexei BAEVSKI, Ronan COLLOBERT, Abdelrahman MOHAMED et Michael AULI : Unsupervised cross-lingual representation learning for speech recognition. In *proceedings of INTERSPEECH 2021*, Brno, Czechia, 2021. 26, 100, 110, 111, 115

Alexis CONNEAU, Min MA, Simran KHANUJA, Yu ZHANG, Vera AXELROD, Siddharth DALMIA, Jason RIESA, Clara RIVERA et Ankur BAPNA : FLEURS : FEW-Shot Learning Evaluation of Universal Representations of Speech. In *proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, Doha, Qatar, janvier 2023. 35

Emanuela CRESTI, Fernanda Bacelar do NASCIMENTO, Antonio Moreno SANDOVAL, Jean VERONIS, Philippe MARTIN et Khalid CHOUKRI : The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, mai 2004. 120

CSA-HADOPI : Assistants vocaux et enceintes connectées - l'impact de la voix sur l'offre et les usages culturels et médias. <https://www.csa.fr/Informer/Collections-du-CSA/Thema-Toutes-les-etudes-realisees-ou-co-realisees-par-le-CSA-sur-des-themes-specifiques/Les-etudes-corealisees-avec-le-CSA/Etude-HADOPI-CSA-2019-Assistants-vocaux-et-enceintes-connectees>, 2019. 1

George DAHL, Dong YU, li DENG et Alex ACERO : Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 20:30 – 42, 2012. 8

Martine DE CALMÈS, Jérôme FARINAS, Isabelle FERRANÉ et Julien PINQUIER : Campagne ESTER : une première version d'un système complet de transcription automatique de la parole grand vocabulaire (Atelier ESTER , Avignon, 30/03/05-31/03/05), mars 2005. 7

Paul DELÉGLISE et Carole LAILLER : Quel type de systèmes utiliser pour la transcription automatique du français ? Les HMM font de la résistance (What system for the automatic transcription of French in audiovisual broadcasts?). In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition)*,

RÉFÉRENCES

- Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, Nancy, France, juin 2020. xxiv, 32, 40, 60
- A.P. DEMPSTER, N.M. LAIRD et D.B. RUBIN : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, 1* (Series B. 39):1–38, 1977. 7
- Jean-Marc DEWAELE : Variation dans la composition lexicale des styles oraux. *International Review of Applied Linguistics in Language Teaching*, 34(4):261–282, janvier 1996. 66
- John DINES et Mathew MAGIMAI DOSS : A Study of Phoneme and Grapheme Based Context-Dependent ASR Systems. *In proceedings of Machine Learning for Multimodal Interaction*, Berlin, Heidelberg, 2008. 6
- Anne DISTER, Michel FRANCARD, Philippe HAMBYE et Anne-Catherine SIMON : Du corpus à la banque de données. Du son, des textes et des métadonnées. L'évolution de la banque de données textuelles orales VALIBEL (1989-2009). *Cahiers de Linguistique*, 33:113–129, 2009. 121
- Anne DISTER et Emmanuelle LABEAU : Le corpus de français parlé à bruxelles ; origines, hypothèses, développements et prédictions. *Cahiers AFLS eJournal*, 21(1):1–22, 2017. 119
- Danielle DUEZ : Manifestation acoustique et phonétique de la réduction et de l'assimilation contextuelle des segments de la parole conversationnelle. *Parole*, pages 89–111, 2001. 48, 64
- Richard DUFOUR : From prepared speech to spontaneous speech recognition system : a comparative study applied to French language. *In Proceedings of the 5th international conference on Soft computing as transdisciplinary science and technology (CSTST'08)*, New York, NY, USA, 2008. 55
- Richard DUFOUR, Yannick ESTÈVE et Paul DELÉGLISE : Automatic indexing of speech segments with spontaneity levels on large audio database. *In Proceedings of the 2010 international workshop on Searching spontaneous conversational speech - SSCS'10*, Firenze, Italy, 2010. xxiv, 40, 41, 44, 52, 58, 60, 65
- Richard DUFOUR, Yannick ESTÈVE et Paul DELÉGLISE : Characterizing and detecting spontaneous speech : Application to speaker role recognition. *Speech Communication*, 56:1–18, janvier 2014. xix, 57, 58, 59, 62
- Lourdes DURAN-LOPEZ, Juan P. DOMINGUEZ-MORALES, Antonio Felix CONDE-

RÉFÉRENCES

MARTIN, Saturnino VICENTE-DIAZ et Alejandro LINARES-BARRANCO : Prometeo : A cnn-based computer-aided diagnosis system for wsi prostate cancer detection. *IEEE Access*, 8:128613–128628, 2020. xvii, 157

Jacques DURAND, Bernard LAKS et Chantal LYCHE : La phonologie du français contemporain : usages, variétés et structure. In C. Pusch W. Raible (ÉDS), éditeur : *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/-Romance Corpus Linguistics – Corpora and Spoken Language*, pages 93–106. Gunter Narr Verlag Tübingen, 2002. 123

Jacques DURAND, Bernard LAKS, Chantal LYCHE, Meqqori Abderrahim DELAIS-ROUSSARIE, Elisabeth et Jean-Michel TARRIER : Bulletin pfc n°1. <https://www.projet-pfc.net/2008/11/24/bulletin-pfc-n1/>, 2008. 125

Zied ELLOUMI, Laurent BESACIER, Olivier GALIBERT, Juliette KAHN et Benjamin LECOUTEUX : Asr performance prediction on unseen broadcast programs using convolutional neural networks. In *proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018. 56

Jeffrey L. ELMAN : Finding structure in time. *Cognitive Science*, 14(2):179–211, avril 1990. ISSN 0364-0213. 12

EQUIPE DELIC, Sandra TESTON-BONNARD et Jean VÉRONIS : Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé*, 18:11–42, 2004. 64, 120

Linus ERICSSON, Henry GOUK, Chen Change LOY et Timothy M. HOSPEDALES : Self-Supervised Representation Learning : Introduction, advances, and challenges. *IEEE Signal Processing Magazine*, 39(3):42–62, mai 2022. 24

Iris ESHKOL-TARAVELLA, Olivier BAUDE, Denis MAUREL, Linda HRIBA, Celine DUGUA et Isabelle TELLIER : Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. *Traitemet Automatique des Langues*, 53(2):17–46, 2011. 62, 122

Maxine ESKENAZI : Trends in speaking styles research. In *Proc. 3rd European Conference on Speech Communication and Technology (Eurospeech 1993)*, 1993. xvi, 66, 68, 69

Yannick ESTÈVE, Thierry BAZILLON, Jean-Yves ANTOINE, Frédéric BÉCHET et Jérôme FARINAS : The EPAC Corpus : Manual and Automatic Annotations of Conversational Speech in French Broadcast News. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, mai 2010. 33

RÉFÉRENCES

Solène EVAIN, Benjamin LECOUTEUX, François PORTET, Isabelle ESTÈVE et Marion FABRE : Towards Automatic Captioning of University Lectures for French Students who are Deaf. *In ACM SIGACCESS Conference on Computers and Accessibility*, Athènes, Greece, 2020a. xxiv

Solène EVAIN, Benjamin LECOUTEUX, François PORTET, Isabelle ESTÈVE et Marion FABRE : Towards Automatic Captioning of University Lectures for French students who are Deaf. *In proceedings of the 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA, 2020b. 1

Solène EVAIN, Ha NGUYEN, Hang LE, Marcely Zanon BOITO, Salima MDHAFFAR, Sina ALISAMIR, Ziyi TONG, Natalia TOMASHENKO, Marco DINARELLI, Titouan PARCOLLET, Alexandre ALLAUZEN, Yannick ESTÈVE, Benjamin LECOUTEUX, François PORTET, Solange ROSSATO, Fabien RINGEVAL, Didier SCHWAB et Laurent BESACIER : LeBenchmark : A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. *In proceedings of INTERSPEECH 2021*, août 2021a. xx, 109, 110, 114

Solène EVAIN, Manh Hà NGUYEN, Hang LE, Marcely ZANON BOITO, Salima MDHAFFAR, Sina ALISAMIR, Ziyi TONG, Natalia TOMASHENKO, Marco DINARELLI, Titouan PARCOLLET, Alexandre ALLAUZEN, Yannick ESTÈVE, Benjamin LECOUTEUX, François PORTET, Solange ROSSATO, Fabien RINGEVAL, Didier SCHWAB et Laurent BESACIER : Task agnostic and task specific self-supervised learning from speech with lebenchmark. *In proceedings of the Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS 2021)*, on-line, United States, décembre 2021b. xx, 27, 30, 36, 84, 110, 111, 114

Benoit FAVRE, Kyla CHEUNG, Siavash KAZEMIAN, Adam LEE, Yang LIU, Cosmin MUNTEANU, Ani NENKOVA, Dennis OCHEI, Gerald PENN, Stephen TRATZ, Clare VOSS et Frauke ZELLER : Automatic Human Utility Evaluation of ASR Systems : Does WER Really Predict Performance ? *In proceedings of Interspeech 2013*, Lyon, France, janvier 2013. 32

Soline FELICE, Solène EVAIN, Solange ROSSATO et François PORTET : Audio-cite.net : A large spoken read dataset in french. *proceedings of LREC-COLING 2024*, 2024. 103

Dominique FOHR, Odile MELLA et Denis JOUVET : De l'importance de l'homogénéisation des conventions de transcription pour l'alignement automatique de corpus oraux de parole spontanée. *In 8es Journées Internationales de Linguistique de Corpus (JLC2015)*,, septembre 2015. 48, 52, 53, 55

RÉFÉRENCES

- Hiroya FUJISAKI : Prosody, Models, and Spontaneous Speech. *In Computing Prosody*. Springer, New York, NY, 1997. 61, 62
- S. FURUI : Recent progress in spontaneous speech recognition and understanding. *In 2002 IEEE Workshop on Multimedia Signal Processing.*, décembre 2002. 54
- Philipp GABLER, Bernhard C. GEIGER, Barbara SCHUPPLER et Roman KERN : Reconsidering Read and Spontaneous Speech : Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition. *Information*, 14 (2):137, 2023. 46, 55, 56, 60, 62, 66
- Françoise GADET et Emmanuelle GUERIN : Construire un corpus pour des façons de parler non standard : « Multicultural Paris French ». *Corpus*, (15), octobre 2016. ISSN 8. 122
- Sebastian GALLIANO, Edouard GEOFFROIS, Djamel MOSTEFA, Khalid CHOUKRI, Jean-François BONASTRE et Guillaume GRAVIER : The ESTER phase II evaluation campaign for the rich transcription of French broadcast news. *In proceedings of INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, septembre 2005. 33
- Mahault GARNERIN : *Des données aux systèmes : étude des liens entre données d'apprentissage et biais de performance genrés dans les systèmes de reconnaissance automatique de la parole*. phdthesis, Université Grenoble-Alpes, Grenoble, France, 2022. 44
- John S. GAROFOLLO, Cedric G. P. AUZANNE et Ellen M. VOORHEES : The TREC spoken document retrieval track : a success story. *In Content-Based Multimedia Information Access - Volume 1*, Paris, FRA, avril 2000. 31
- Aude GIRAUDEL, Matthieu CARRÉ, Valérie MAPELLI, Juliette KAHN, Olivier GALIBERT et Ludovic QUINTARD : The REPERE Corpus : a multimodal corpus for person recognition. *In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, mai 2012. 33
- Julie GLIKMAN et Camille FAUTH : Un nouvel accès à la parole spontanée : les vocaux. *In proceedings of the JEP 2022*, juin 2022. 177
- Jean-Philippe GOLDMAN, Antoine AUCHLIN et Anne-Catherine SIMON : Discrimination de styles de parole par analyse prosodique semi-automatique. *In proceedings of Interface Discours Prosodie (IDP2009)*, 2009. xvi, 44, 49, 50, 62, 68
- Sharon GOLDWATER, Dan JURAFSKY et Christopher D. MANNING : Which words

RÉFÉRENCES

- are hard to recognize ? prosodic, lexical and disfluency factors that increase asr errors. *In in proceedigs of ACL-08 : HLT*, 2008. 54, 55
- Hector GONZALEZ, Shahzad MUZAFFAR, Jerald YOO et Ibrahim (Abe) ELFADEL : BioCNN : A Hardware Inference Engine for EEG-based Emotion Detection. *IEEE Access*, 8, janvier 2020. xvii, 129
- Philippe GOURNAY, Olivier LAHAIE et R. LEFEBVRE : A canadian french emotional speech dataset. *In proceedings of the 9th ACM Multimedia Systems Conference*, 2018. 53, 102
- Alex GRAVES : Sequence Transduction with Recurrent Neural Networks. *In Representation Learning Workshop, ICML 2012*, Edinburgh, Scotland, novembre 2012. 8, 19
- Alex GRAVES, Douglas ECK, Nicole BERINGER et Juergen SCHMIDHUBER : Biologically Plausible Speech Recognition with LSTM Neural Nets. *In Proceedings of the 1st International Worshop on Biologically Inspired Approaches to Advanced Information Technology*, mai 2004. 13
- Alex GRAVES, Santiago FERNÁNDEZ, Faustino GOMEZ et Jürgen SCHMIDHUBER : Connectionist temporal classification : labelling unsegmented sequence data with recurrent neural networks. *In Proceedings of the 23rd international conference on Machine learning*, juin 2006. 17
- Alex GRAVES et Navdeep JAITLEY : Towards End-To-End Speech Recognition with Recurrent Neural Networks. *In Proceedings of the 31st International Conference on Machine Learning*, juin 2014. 17
- Alex GRAVES, Abdel-rahman MOHAMED et Geoffrey HINTON : Speech Recognition with Deep Recurrent Neural Networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, mars 2013. 13
- G. GRAVIER, J-F. BONASTRE, E. GEOFFROIS, S. GALLIANO, K. MCTAIT et K. CHOUKRI : The ESTER Evaluation Campaign for the Rich Transcription of French Broadcast News. *In Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, 2004. 33
- Guillaume GRAVIER, Gilles ADDA, Niklas PAULSON, Matthieu CARRÉ, Aude GI-RAUDEL et Olivier GALIBERT : 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*, Turkey, 2012. xvii, 31, 33, 155

- Yanzhang HE, Tara N. SAINATH, Rohit PRABHAVALKAR, Ian McGRAW, Raziel ALVAREZ, Ding ZHAO, David RYBACH, Anjuli KANNAN, Yonghui WU, Ruoming PANG, Qiao LIANG, Deepti BHATIA, Yuan SHANGGUAN, Bo LI, Golan PUNDAK, Khe Chai SIM, Tom BAGBY, Shuo-yiin CHANG, Kanishka RAO et Alexander GRUENSTEIN : Streaming End-to-end Speech Recognition for Mobile Devices. *In proceedings of ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, mai 2019. 20
- Hartmut HELMKE, Shruthi SHETTY, Matthias KLEINERT, Oliver OHNEISER, Amrutha PRASAD, Petr MOTLICE, Aneta CERNA et Christian WINDISCH : How to Measure Speech Recognition Performance in the Air Traffic Control Domain ? The Word Error Rate is only half of the truth. *In proceedings of Interspeech 2021*, janvier 2021. 31
- H. HERMANSKY et L.A. COX JR. : Perceptual Linear Predictive (PLP) Analysis-Resynthesis Technique. *In proceedings of Final Program and Paper Summaries 1991 IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, octobre 1991. 5
- Geoffrey HINTON, Li DENG, Dong YU, George E. DAHL, Abdel-rahman MO-HAMED, Navdeep JAITLEY, Andrew SENIOR, Vincent VANHOUCKE, Patrick NGUYEN, Tara N. SAINATH et Brian KINGSBURY : Deep Neural Networks for Acoustic Modeling in Speech Recognition : The Shared Views of Four Research Groups. *IEEE Signal Processing Magazine*, 29(6):82–97, novembre 2012. 16
- Sepp HOCHREITER et Jürgen SCHMIDHUBER : Long Short-term Memory. *Neural computation*, 9(8):1735–80, décembre 1997. 13
- Koharu HORII, Meiko FUKUDA, Kengo OHTA, Ryota NISHIMURA, Atsunori OGAWA et Norihide KITAOKA : End-to-End Spontaneous Speech Recognition Using Disfluency Labeling. *In Proc. Interspeech 2022*, 2022. 47
- Magali HUSIANYCIA : « Genre » ou « type » de discours ? *Pratiques. Linguistique, littérature, didactique*, (157-158):133–152, juin 2013. 120
- Frederick JELINEK : *Statistical Methods for Speech Recognition*. Bradford Books, fourth printing édition, 1998. ISBN 978-0262100663. 3
- Hui JIANG : Discriminative training of HMMs for automatic speech recognition : A survey. *Computer Speech & Language*, 24(4):589–608, octobre 2010. 7
- Martin JOOS : THE ISOLATION OF STYLES. *In Readings in the Sociology of Language*, pages 185–191. De Gruyter Mouton, mai 2012. 66
- Michael I. JORDAN : Serial order : A parallel, distributed processing approach. *In*

RÉFÉRENCES

- Jeffrey L. ELMAN et David E. RUMELHART, éditeurs : *Advances in Connectionist Theory : Speech*. Erlbaum, Hillsdale, NJ, 1989. 12
- Vincent JOUSSE, Yannick ESTEVE, Frederic BECHET, Thierry BAZILLON et Georges LINARES : Caractérisation et détection de parole spontanée dans de larges collections de documents audio. *In proceedings of the Journées d'Études de la Parole, JEP*, 2008. 40, 41, 52
- Sushant KAFLE et Matt HUENERFAUTH : Evaluating the Usability of Automatically Generated Captions for People who are Deaf or Hard of Hearing. *In Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, New York, NY, USA, 2017. 1
- Marine KNEUBÜHLER : Qu'advient-il de la sémantique et de l'interaction dans les transcriptions automatiques d'un logiciel ? *Revue d'anthropologie des connaissances*, 16(2), juin 2022. 56
- Peter KOCH et Wulf OESTERREICHER : Langage parlé et langage écrit (gesprochene sprache and geschriebene sprache). *Lexikon der Romanistischen Linguistik*, I(2):585–627, 2001. xvi, 44, 52, 53, 54, 55, 68, 69, 70, 127
- Fangjun KUANG, Liyong GUO, Wei KANG, Long LIN, Mingshuang LUO, Zengwei YAO et Daniel POVEY : Pruned rnn-t for fast, memory-efficient asr training. *In proceedings of INTERSPEECH 2022*, Incheon, Korea, septembre 2022. 20
- Maria LABIED et Abdessamad BELANGOUR : Automatic speech recognition features extraction techniques : A multi-criteria comparison. *International Journal of Advanced Computer Science and Applications*, 12(8), 2021. xix, 5
- Maria LABIED, Abdessamad BELANGOUR, Mouad BANANE et Allae ERRAISSI : An overview of automatic speech recognition preprocessing techniques. *In proceedings of 2022 International Conference on Decision Aid Sciences and Applications (DASA)*, 2022. 4
- William LABOV : *Sociolinguistic patterns*, chapitre The isolation of contextual styles, pages 70–110. Philadelphia, University of Pennsylvania Press, 2nd edition édition, 1973. ISBN 0-8122-1052-2. xvi, 44, 62, 63, 64, 66, 67, 126
- Jean-Pierre LAI : Le recueil de données dialectales : enjeux et difficultés. *Flambeau*, 45:15–29, 2019. 44
- Bernard LAKS *et al.* : Le projet PFC (Phonologie du Français Contemporain) : une source de données primaires structurées. *In Phonologie, variation et accents du français*. Hermès, 2009. 63, 65, 130

RÉFÉRENCES

- Lori F. LAMEL, Jean-Luc GAUVAIN et Maxine ESKENAZI : BREF, a large vocabulary spoken corpus for French. *In proceedings of the 2nd European Conference on Speech Communication and Technology (Eurospeech 1991)*, septembre 1991. 33
- Mélanie LANCIEN : Caractérisation de la variation liée à la situation de communication : apport de l'acoustique à la phonostylistique. *Glottopol*, 33, 2020. 44, 48, 59, 63, 64
- Mélanie LANCIEN et Marie-Hélène CÔTÉ : Phonostyle et réduction vocalique en français laurentien. *In actes du 6e Congrès Mondial de Linguistique Française*, volume 46, 2018. SHSWeb Conf. 44, 48, 52
- Petri LAUKKA, Daniel NEIBERG, Mimmi FORSELL, Inger KARLSSON et Kjell ELENIUS : Expression of affect in spontaneous speech : Acoustic correlates and automatic detection of irritation and resignation. *Computer Speech & Language*, 25(1):84–104, janvier 2011. 65
- Clément LE MOINE et Nicolas OBIN : Att-HACK : An Expressive Speech Database with Social Attitudes. *In Speech Prosody*, 2020. 102
- Sotheara LEANG, Sotheara CASTELLI, Dominique VAUFREYDAZ et Sethserey SAM : Preliminary study on ssrf-derived polar coordinate for asr. *In proceedings of ACET 2022*, hnom Penh, Cambodia, décembre 2022. 36
- Benjamin LECOUTEUX : *Reconnaissance automatique de la parole guidée par des transcriptions a priori.* phdthesis, Université d'Avignon et des Pays de Vaucluse,, 2008. NNT : tel-01381704. 4
- Y. LECUN, L. BOTTOU, Y. BENGIO et P. HAFFNER : Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, novembre 1998. 14
- Fabrice LEFÈVRE, Djamel MOSTEFA, Laurent BESACIER, Yannick ESTÈVE, Matthieu QUIGNARD, Nathalie CAMELIN, Benoit FAVRE, Bassam JABAIAN et Lina ROJAS-BARAHONA : Robustesse et portabilités multilingue et multi-domaines des systèmes de compréhension de la parole : le projet PortMedia. *In Actes de la conférence conjointe JEP-TALN-RECITAL 2012*, Grenoble, France, juin 2012. 102
- Michael LEVIT, Shuangyu CHANG, Bruce BUNTSCUH et Nick KIBRE : End-to-end speech recognition accuracy metric for voice-search tasks. *In proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2012)*, mars 2012. 31

RÉFÉRENCES

- Hui Lin LI, Ming : *Practitioner's Guide to Data Science*. Chapman and Hall/CRC, New York, 1st édition, 2023. ISBN 9781351132916. xv, 12
- Jinyu LI, Abdelrahman MOHAMED, Geoffrey ZWEIG et Yifan GONG : Exploring multidimensional lstms for large vocabulary ASR. In *proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4940–4944, mars 2016. 13
- Robin LICKLEY et Ellen BARD : On not Recognizing Disfluencies in Dialogue. In *in proceedings of the Spoken Language Conference, ICSLP 96*, novembre 1996. 56
- Andy LIU, Shu-wen YANG, Po-Han CHI, Po-chun HSU et Hung-yi LEE : Mockingjay : Unsupervised Speech Representation Learning with Deep Bidirectional Transformer Encoders. In *proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, mai 2020. 24
- Yuzong LIU, Rishabh IYER, Katrin KIRCHHOFF et Jeff BILMES : Svtchboard ii and fisver i : High-quality limited-complexity corpora of conversational english speech. In *proceedings of INTERSPEECH 2015*, 2015. 15
- Joaquim LLISTERRI : Speaking styles in speech research. In *actes de ELSNET/ESCA/SALT Workshop on Integrating Speech and Natural Language*, Dublin, Ireland, 1992. 44, 48, 49, 50, 52, 59, 63, 64, 65
- Ke-Han LU et Kuan-Yu CHEN : A Context-Aware Knowledge Transferring Strategy for CTC-Based ASR. In *proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, janvier 2023. 19
- Thang LUONG, Hieu PHAM et Christopher D. MANNING : Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, septembre 2015. xv, 22
- Daniel LUZZATI : Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané. In *MIDL 2004*, Paris, France, 2004. 53, 55
- Daniel LUZZATI : Le dialogue oral spontané : quels objets pour quels corpora. *Revue d'Interaction Homme-Machine*, 8(2), 2007. 52, 53, 60, 61, 73
- Christoph LÜSCHER, Eugen BECK, Kazuki IRIE, Markus KITZA, Wilfried MICHEL, Albert ZEYER, Ralf SCHLÜTER et Hermann NEY : RWTH ASR Systems for LibriSpeech : Hybrid vs Attention. In *proceedings of INTERSPEECH 2019*, septembre 2019. 38, 84
- Arash MALEKIAN et Nastaran CHITSAZ : *Advances in Streamflow Forecasting -*

From Traditional to Modern Approaches, chapitre Chapter 4 - Concepts, procedures, and applications of artificial neural network models in streamflow forecasting, pages 115–147. Elsevier, janvier 2021. ISBN 9780128206737. 10

John. D. MARKEL et Augustine H. GRAY JR : *Linear prediction of speech*, volume 12 de *Communication and cybernetics*. Springer-Verlag, Berlin, Heidelberg, New York, 1976. 5

Takashi MASUKO : Computational cost reduction of long short-term memory based on simultaneous compression of input and hidden state. *In proceesinds of 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, décembre 2017. 13

Ambuj MEHRISH, Navonil MAJUMDER, Rishabh BHARADWAJ, Rada MIHALCEA et Soujanya PORIA : A review of deep learning techniques for speech processing. *Information Fusion*, 99, novembre 2023. 17

Yen MENG, Hsuan-Jui CHEN, Jiatong SHI, Shinji WATANABE, Paola GARCIA, Hung-yi LEE et Hao TANG : On Compressing Sequences for Self-Supervised Speech Models. *In proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)*, janvier 2023. 15

Abdelrahman MOHAMED, Hung-yi LEE, Lasse BORGHOLT, Jakob D. HAVTORN, Joakim EDIN, Christian IGEL, Katrin KIRCHHOFF, Shang-Wen LI, Karen LIVESCU, Lars MAALOE, Tara N. SAINATH et Shinji WATANABE : Self-Supervised Speech Representation Learning : A Review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210, octobre 2022. xv, 24, 25

Andrew MORRIS, Viktoria MAIER et Phil GREEN : From WER and RIL to MER and WIL : improved evaluation measures for connected speech recognition. *In proceedings of Interspeech 2004*, Jeju Island, Korea, octobre 2004. 31, 32

Cosmin MUNTEANU, Ronald BAECKER, Gerald PENN, Elaine TOMS et David JAMES : The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. *In Proceedings of the 2006 Conference on Human Factors in Computing Systems, CHI 2006*, Montréal, Québec, Canada, avril 2006. 31

Muhammadjon MUSAEV, Ilyos XUJAYOROV et Mannan OCHILOV : Image Approach to Speech Recognition on CNN. *In proceedings of the 2019 3rd International Symposium on Computer Science and Intelligent Control*, septembre 2019. 15

Kensuke NAKAMURA, Bilel DERBEL, Kyoung-Jae WON et Byung-Woo HONG :

RÉFÉRENCES

- Learning-Rate Annealing Methods for Deep Neural Networks. *Electronics*, 10 (16):2029, janvier 2021. 88
- Masanobu NAKAMURA, Iwano KOJI et Sadaoki FURUI : Analysis of spectral space reduction in spontaneous speech and its effects on speech recognition performances. In *proceedings of NTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology*, septembre 2005. 47
- H. NANJO et T. KAWAHARA : A new ASR evaluation measure and minimum Bayes-risk decoding for open-domain speech understanding. In *Proceedings of ICASSP '05. IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, mars 2005. 31
- Juan M. NAVARRO, Raquel MARTÍNEZ-ESPAÑA, Andrés BUENO-CRESPO, Ramón MARTÍNEZ et José M. CECILIA : Sound Levels Forecasting in an Acoustic Sensor Network Using a Deep Neural Network. *Sensors*, 20(3):903, janvier 2020. xv, 14
- Boris NEW, Marc BRYSBERT, Jean VERONIS et Christophe PALLIER : The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, 28 (4):661–677, 2007. 47
- Lucia ORMAECHEA-GRIJALBA, Pierrette BOUILLON, Maximin COAVOUX, Emmanuelle ESPERANÇA-RODIER, Johanna GERLACH, Jérôme GOULIAN, Benjamin LECOUTEUX, Cécile MACAIRE, Jonathan David MUTAL, Magali NORRÉ, Adrien PUPIER, Didier SCHWAB et Hervé SPECHBACH : PROPICTO : Développer des systèmes de traduction de la parole vers des séquences de pictogrammes pour améliorer l'accessibilité de la communication. In *18e Conférence en Recherche d'Information et Applications, 16e Rencontres Jeunes Chercheurs en RI, 30e Conférence sur le Traitement Automatique des Langues Naturelles, 25e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues*, Paris, France, 2023. xxiv
- Dimitri PALAZ, Ronan COLLOBERT et Mathew MAGIMAI.-DOSS : End-to-end phoneme sequence recognition using convolutional neural networks. In *proceedings of NIPS Deep learning Workshop*, 2013. 14
- Dimitri PALAZ, Mathew MAGIMAI.-DOSS et Ronan COLLOBERT : Convolutional Neural Networks-based continuous speech recognition using raw speech signal. In *proceedings of 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, avril 2015. 14
- Titouan PARCOLLET, Ha NGUYEN, Solène EVAIN, Marcely ZANON BOITO, Adrien PUPIER, Salima MDHAFFAR, Hang LE, Sina ALISAMIR, Natalia TOMASHENKO, Marco DINARELLI, Shucong ZHANG, Alexandre ALLAUZEN, Maximin COA-

RÉFÉRENCES

VOUX, Yannick ESTÈVE, Mickael ROUVIER, Jérôme GOULIAN, Benjamin LE-COUTEUX, François PORTET, Solange ROSSATO, Fabien RINGEVAL, Didier SCHWAB et Laurent BESACIER : LeBenchmark 2.0 : A standardized, replicable and enhanced framework for self-supervised representations of French speech. *Computer Speech & Language*, 86:101622, juin 2024. xx, 84, 112, 113

Titouan PARCOLLET, Mirco RAVANELLI, Peter PLANTINGA, Aku ROUHE, Samuele CORNELL, Loren LUGOSCH, Cem SUBAKAN, Nauman DAWALATABAD, Abdelwahab HEBA, Jianyuan ZHONG, Ju-Chieh CHOU, Sung-Lin YEH, Szu-Wei FU, Chien-Feng LIAO, Elena RASTORGUEVA, François GRONDIN, William ARIS, Hwidong NA, Yan GAO, Renato de MORI et Yoshua BENGIO : SpeechBrain : A General-Purpose Speech Toolkit. Preprint, mars 2022. 87

Santiago PASCUAL, Mirco RAVANELLI, Joan SERRÀ, Antonio BONAFONTE et Y. BENGIO : Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. *In proceedings of Interspeech 2019*, septembre 2019. 24

Thomas H. PAYNE, W. David ALONSO, J. Andrew MARKIEL, Kevin LYBARGER et Andrew A. WHITE : Using voice to create hospital progress notes : Description of a mobile application and supporting system integrated with a commercial electronic health record. *Journal of Biomedical Informatics*, 77:91–96, 2018. ISSN 1532-0464. 1

Charlotte PELLETIER, Geoffrey WEBB et François PETITJEAN : Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. *Remote Sensing*, 11(5):523, mars 2019. xv, 11

G. PERENNOU : B.D.L.E.X. : A data and cognition base of spoken French. *In proceedings of ICASSP '86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, avril 1986. 8

Hiep Van PHUNG et Eun Joo RHEE : A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets. *Applied Sciences*, 9(21), 2019. xv, 15

Louis C.W. POLS : Speech dynamics. *In ICPHS XVII*, Hong Kong, 2011. 4

Jean POUGET-ABADIE, Dzmitry BAHDANAU, Bart van MERRIËNBOER, Kyunghyun CHO et Yoshua BENGIO : Overcoming the Curse of Sentence Length for Neural Machine Translation using Automatic Segmentation. *In Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, Doha, Qatar, octobre 2014. 20

Rohit PRABHAVALKAR, Kanishka RAO, Tara SAINATH, Bo LI, Leif JOHNSON et

RÉFÉRENCES

- Navdeep JAITLEY : A Comparison of Sequence-to-Sequence Models for Speech Recognition. *In proceedings of INTERSPEECH 2017*, Stockholm, Sweden, août 2017. xv, 18, 19
- Vineel PRATAP, Andros TJANDRA, Bowen SHI, Paden TOMASELLO, Arun BABU, Sayani KUNDU, Ali ELKAHKY, Zhaoheng NI, Apoorv VYAS, Maryam FAZEL-ZARANDI, Alexei BAEVSKI, Yossi ADI, Xiaohui ZHANG, Wei-Ning HSU, Alexis CONNEAU et Michael AULI : Scaling Speech Technology to 1,000+ Languages. 10.48550/arXiv.2305.13516, mai 2023. 26
- Vineel PRATAP, Qiantong XU, Anuroop SRIRAM, Gabriel SYNNAEVE et Ronan COLLOBERT : Mls : A large-scale multilingual dataset for speech research. *In proceedings of INTERSPEECH 2020*, 2020. Updated ArXiV version : <https://arxiv.org/pdf/2012.03411.pdf>. 35, 36
- Bertrand PÉRIER : *La parole est un sport de combat*. Broché, 2017. 63
- Alec RADFORD, Jong KIM, Tao XU, Greg BROCKMAN, Christine MCLEAVEY et Ilya SUTSKEVER : Robust Speech Recognition via Large-Scale Weak Supervision. <https://doi.org/10.48550/arXiv.2212.04356>, décembre 2022. xvi, 24, 26, 28, 31, 36, 37, 89
- Alec RADFORD, Karthik NARASIMHAN, Tim SALIMANS et Ilya SUTSKEVER : Improving language understanding by generative pre-training. *preprint*, 2018. 152
- Mirco RAVANELLI, Jianyu ZHONG, Santiago PASCUAL, Pawel SWIETOJANSKI, Joao MONTEIRO, Jan TRMAL et Yoshua BENGIO : Multi-Task Self-Supervised Learning for Robust Speech Recognition. *In proceedings of ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, mai 2020. 24
- Daniel RENSHAW, Herman KAMPER, Aren JANSEN et Sharon GOLDWATER : A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. *In Proceedings of Interspeech 2015*, 2015. 24
- A.J. ROBINSON : An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, mars 1994. ISSN 1941-0093. 12
- Magdalena ROMERA CIRIA : Relationships as regulators of discourse interaction in Spanish. *Círculo de Lingüística Aplicada a la Comunicación*, numéro 79:pages 297–322, 2019. 62, 127

RÉFÉRENCES

- F. ROSENBLATT : The perceptron : A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958. xv, 10
- Anthony ROUSSEAU, Gilles BOULIANNE, Paul DELÉGLISE, Yannick ESTÈVE, Vishwa GUPTA et Sylvain MEIGNIER : Lium and crim asr system combination for the repere evaluation campai. *In proceedings of the International Conference on Text, Speech, and Dialogue*, 2014. 36
- Thibault Bañeras ROUX, Mickael ROUVIER, Jane WOTTAWA et Richard DUFOUR : Qualitative Evaluation of Language Model Rescoring in Automatic Speech Recognition. *In proceedings of Interspeech 2022*, septembre 2022. 31
- James RUSSELL : A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39:1161–1178, décembre 1980. 128
- Elizabeth SALESKY, Matthew WIESNER, Jacob BREMERMAN, Roldano CATTONI, Matteo NEGRI, Marco TURCHI, Douglas OARD et Matt POST : The Multilingual TEDx Corpus for Speech Recognition and Translation. *In proceedings of Interspeech 2021*, août 2021. 35, 36
- Chenze SHAO et Yang FENG : Non-monotonic latent alignments for ctc-based non-autoregressive machine translation. *In proceedings of NeurIPS 2022*, 2022. 17
- Apeksha SHEWALKAR, Deepika NYAVANANDI et Simone LUDWIG : Performance Evaluation of Deep neural networks Applied to Speech Recognition : Rnn, LSTM and GRU. *Journal of Artificial Intelligence and Soft Computing Research*, 9:235–245, octobre 2019. 11, 12
- Elizabeth SHRIBERG, Andreas STOLCKE et Don BARON : Observations on overlap : findings and implications for automatic processing of multi-party conversation. *In 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, septembre 2001. 31
- Chadawan Ittichaichareon SIWAT SUKSRI et Thaweesak YINGTHAWORNSUK : Speech recognition using mfcc. *In International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012)*, Pattaya (Thailand), 2012. 5
- Xingcheng SONG, Guangsen WANG, Zhiyong WU, Yiheng HUANG, Dan SU, Dong YU et Helen M. MENG : Speech-XLNet : Unsupervised acoustic model pretraining for self-attention networks. 2019. 24
- S. S. STEVENS, J. VOLKMANN et E. B. NEWMAN : A scale for the measurement of

RÉFÉRENCES

- the psychological magnitude. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937. 5
- Andreas STOLCKE et Jasha DROPO : Comparing Human and Machine Errors in Conversational Speech Transcription. *In proceedings of INTERSPEECH 2017*, août 2017. 56
- Frederik STOUTEN, Jacques DUCHATEAU, Jean-Pierre MARTENS et Patrick WAMBACQ : Coping with disfluencies in spontaneous speech recognition : Acoustic detection and linguistic context manipulation. *Speech Communication*, 48 (11):1590–1606, novembre 2006. 54
- György SZASZÁK, Máté Ákos TÜNDIK et András BEKE : Summarization of Spontaneous Speech using Automatic Speech Recognition and a Speech Prosody based Tokenizer. *In Proceedings of the International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Setubal, PRT, novembre 2016. 46, 60
- Piotr SZYMANSKI, Piotr ZELASKO, Mikolaj MORZY, Adrian SZYMCZAK, Marzena ZYLA-HOPPE, Joanna BANASZCZAK, Lukasz AUGUSTYNIAK, Jan MIZGAJSKI et Yishay CARMIEL : Wer we are and wer we think we are. *In findings of EMNLP 2020*, janvier 2020. 31
- Chaitanya TALNIKAR, Tatiana LIKHOMANENKO, Ronan COLLOBERT et Gabriel SYNNAEVE : Joint Masked CPC And CTC Training For ASR. *In proceedings of ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, juin 2021. 17
- Elise TANCOIGNE, Jean-Philippe CORBELLINI, Gaëlle DELETRAZ, Laure GAY-RAUD, Sandrine OLLINGER et Daniel VALERO : La transcription automatique : un rêve enfin accessible ? Analyse et comparaison d’outils pour les SHS. Nouvelle méthodologie et résultats. Research report, MATE-SHS, août 2020. xvi, 39, 41
- Elaine TARONE : *Variation in interlanguage*. Hodder Arnold, 1988. 66
- Alain THOMAS : La variation phonétique en français langue seconde au niveau universitaire avancé. *Acquisition et interaction en langue étrangère*, (17):101–121, décembre 2002. 47
- Andros TJANDRA, Sakriani SAKTI et Satoshi NAKAMURA : Local Monotonic Attention Mechanism for End-to-End Speech And Language Processing. *In Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1 : Long Papers)*, Taipei, Taiwan, novembre 2017. 22
- Francisco TORREIRA, Martine ADDA-DECKER et Mirjam ERNESTUS : The Nij-

RÉFÉRENCES

- megen Corpus of Casual French. *Speech Communication*, 52(3), 2010. 44, 52, 53, 55, 62, 63, 65
- Véronique TRAVERSO : *La conversation familière : analyse pragmatique des interactions*. 1996. 63
- Malgorzata Anna ULASIK, Manuela HÜRLIMANN, Fabian GERMANN, Esin GEDIK, Fernando BENITES et Mark CIELIEBAK : CEASR : A Corpus for Evaluating Automatic Speech Recognition. *In Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, mai 2020. 47
- UNK : African Accented French, 2003. URL <https://www.openslr.org/57/>. 101
- Ashish VASWANI, Noam M. SHAZER, Niki PARMAR, Jakob USZKOREIT, Llion JONES, Aidan N. GOMEZ, Lukasz KAISER et Illia POLOSUKHIN : Attention is All you Need. *In proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, juin 2017. xv, 22, 23, 24, 27
- D. VAUFREYDAZ, C. BERGAMINI, J.F. SERIGNAT, L. BESACIER et M. AKBAR : A New Methodology for Speech Corpora Definition from Internet Documents. *In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece, mai 2000. 32
- Daria VAZHENINA et Konstantin MARKOV : End-to-End Noisy Speech Recognition Using Fourier and Hilbert Spectrum Features. *Electronics*, 9(7):1157, juillet 2020. xv, 9
- Arlindo VEIGA, Sara CANDEIAS, Dirce CELORICO, Jorge PROENÇA et Fernando PERDIGÃO : Towards Automatic Classification of Speech Styles. *In Proceedings of the 10th international conference on Computational Processing of the Portuguese Language*, Coimbra, Portugal, 2012. 44
- Clément VIKTOROVITCH : *Le pouvoir rhétorique*. 2021. 63
- Anne-Catherine WAGNER : Habitus. *Sociologie*, mars 2012. 127
- Petra WAGNER, Jürgen TROUVAIN et Frank ZIMMERER : In defense of stylistic diversity in speech research. *Journal of Phonetics*, 48:1–12, janvier 2015. ISSN 0095-4470. xvi, 65, 67
- Changhan WANG, Morgane RIVIERE, Ann LEE, Anne WU, Chaitanya TALNIKAR, Daniel HAZIZA, Mary WILLIAMSON, Juan PINO et Emmanuel DUPOUX : VoxPopuli : A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th*

RÉFÉRENCES

International Joint Conference on Natural Language Processing, volume Volume 1 : Long Papers, Online, août 2021. 35

P.J. WERBOS : Backpropagation through time : what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, octobre 1990. ISSN 1558-2256. 12

Mark WILKINSON, Michel DUMONTIER, IJsbrand Jan AALBERSBERG, Gaby AP-PLETON, Myles AXTON, Arie BAAK, Niklas BLOMBERG, Jan-Willem BOITEN, Luiz Olavo Bonino da SILVA SANTOS, Philip BOURNE, Jildau BOUWMAN, Anthony BROOKES, Tim CLARK, Merce CROSAS, Ingrid DILLO, Olivier DUMON, Scott EDMUND, Chris EVELO, Richard FINKERS et Barend MONS : The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 2016. 81

Yaru WU et Martine ADDA-DECKER : Réduction temporelle en français spontané : où se cache-t-elle ? Une étude des segments, des mots et séquences de mots fréquemment réduits. In *Actes de la 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)*, Nancy, France, 2020. xvi, 44, 48, 49, 52

Yaru WU et Martine ADDA-DECKER : Réduction des segments en français spontané : apports des grands corpus et du traitement automatique de la parole. *Corpus*, 22, 2021. [en ligne] consulté le 17 janvier 2024. 44

Fan YU, Shiliang ZHANG, Yihui FU, Lei XIE, Siqi ZHENG, Zhihao DU, Weilong HUANG, Pengcheng GUO, Zhijie YAN, Bin MA, Xin XU et Hui BU : M2met : The icassp 2022 multi-channel multi-party meeting transcription challenge. In *proceedings of ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 1

Salah ZAIEM, Titouan PARCOLLET, Slim ESSID et Abdelwahab HEBA : Pretext Tasks Selection for Multitask Self-Supervised Audio Representation Learning. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1439–1453, octobre 2022. 24

M. ZANON BOITO, L. BESACIER, N. TOMASHENKO et Y. ESTÈVE : A study of gender impact in self-supervised models for speech-to-text systems, 2022a. arXiv :2204.01397. 107

Marcely ZANON BOITO, Fethi BOUGARES, Florentin BARBIER, Souhir GAH-BICHE, Loïc BARRAULT, Mickael ROUVIER et Yannick ESTÈVE : Niger-

RÉFÉRENCES

mali audio collection, 2022b. URL <https://demo-lia.univ-avignon.fr/studios-tamani-kalangou/>. 102

Marcely ZANON BOITO, Fethi BOUGARES, Florentin BARBIER, Souhir GAH-BICHE, Loïc BARRAULT, Mickael ROUVIER et Yannick ESTÈVE : Speech resources in the tamasheq language. *In Proceedings of LREC 2022*, 2022c. URL <https://arxiv.org/pdf/2201.05051.pdf>. 102

Vicky ZAYATS, Trang TRAN, Richard WRIGHT, Courtney MANSFIELD et Mari OSTENDORF : Disfluencies and Human Speech Transcription Errors. *In proceedings of Interspeech 2019*, 2019. 56

A. ZWICKY : On casual speech. *In proceedings of the 8th regional meeting of the Chicago Linguistic Society*, 1972. 66

Quatrième partie

Annexes

Annexe A

Correspondance API-SAMPA

SAMPA	API	Exemples
p	p	<i>pont</i>
b	b	<i>bon</i>
t	t	<i>temps</i>
d	d	<i>dans</i>
k	k	<i>coût, quand, koala</i>
g	g	<i>gant</i>
f	f	<i>femme</i>
v	v	<i>vent</i>
s	s	<i>sans, dessus, cerise</i>
z	z	<i>zone, rose</i>
S	ʃ	<i>champ</i>
Z	ʒ	<i>gens, jambon</i>
j	j	<i>ion [jo~]</i>
m	m	<i>mont</i>
n	n	<i>nom</i>
J	ɲ	<i>oignon</i>
N	ŋ	<i>camping</i>
l	l	<i>long</i>
R	r	<i>rond</i>
w	w	<i>quoi [kwa]</i>
H	ɥ	<i>juin [ZHe~]</i>

FIGURE A.1 – Correspondance API-SAMPA pour les consonnes du français

SAMPA	API	Exemples
i	i	<i>si</i>
e	e	<i>Blé</i>
E	ɛ	<i>seize</i>
a	a	<i>patte</i>
A	ɑ	<i>pâtre</i>
O	ɔ	<i>comme</i>
o	o	<i>gros</i>
u	u	<i>doux</i>
y	y	<i>du</i>
ə	ø	<i>deux</i>
ø	œ	<i>neuf</i>
@	ə	<i>justement</i>
e~	ɛ̃	<i>vin</i>
a~	ã	<i>vent</i>
o~	ɔ̃	<i>bon</i>
ø~	œ̃	<i>brun</i>

FIGURE A.2 – Correspondance API-SAMPA pour les voyelles du français

Annexe B

Performances de validation croisée (WER) sur les différents cas d'étude

TABLE B.1 – WER par lot, par enregistrement et par locuteur pour le cas *usual_close*

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.
1	70,2	aperitif_glasgow	61,98	JUD	58,9
				PAT	64,76
		montage_meuble	76,94	MIC	76,44
				CLA	77,14
		aperitif_rupture	85,8	ALB	90,8
				ARN	92,25
2	60,98	repas_français	48,43	JUD	48,99
				MAR	46,78
		repas_olives	73,42	FLO	73,15
				EMI	68,35
		repas_kiwi	57,88	PIE	79,93
				ELI	54,18
3	65,4	repas_pois	71,38	BEA	60,5
				MAR	76,17
		aperitif_pois	75,04	ROM	90,53
				ANN	72,53
				JUL	66,9
				spk2	78,58
				spk1	72,25
				Suite en page suivante	

CHAPITRE B : *Performances de validation croisée (WER) sur les différents cas d'étude*

Table B.1 – suite de la page précédente

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.
		ESLO2 REPAS_1259	60,32	spk2	55,85
				spk1	66,35
		ESLO2 REPAS_1263	37,93	spk2	33,77
				spk1	41,54
		ESLO2 REPAS_1271	56,64	spk3	58,16
				spk1	54,65
				spk2	56,07
		ESLO2 REPAS_1269	62,97	spk2	53,82
				spk1	73,1
				spk3	69,08
4	45,46	rae_ash_sd	38,1	spk2	29,91
				spk1	48,45
		raei_leh_sd	29,92	spk1	25,18
				spk2	39,04
		fete_lec_07	55,97	spk3	49,4
				spk2	61,88
				spk1	59,72
		experiences_yer_12	37,81	spk2	59,62
				spk1	34,3
				spk3	20,73
5	48,04	ESLO2 REPAS_1256	50,34	spk2	46,61
				spk1	46,18
				spk3	65,4
		ESLO2 REPAS_1267	57	spk1	56,05
				spk3	46,87
				spk2	70,21
		cadeaux_bon_08	51,61	spk2	58,41
				spk1	45,53
		stag_bad_08	43,89	spk2	43,56
				spk1	46,42
		famille_fer_08_2	51,46	spk1	55,69
				spk2	48,38
		sports_pet_07	37,3	spk2	35,87
				spk1	38,73
		teatime_lam_12	70,55	spk1	63,1
				spk2	75,37
		ang_jul_07	53	spk1	52,2
				spk2	55,28
		famille_lem_13	44,24	spk2	41,58
				spk1	47,74

Suite en page suivante

Table B.1 – suite de la page précédente

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.
		stage_son_12	39,67	spk1	38,52
				spk2	41,99
				spk3	78,57
		garcons_tou_15	43,5	spk1	46,37
				spk2	37,66
				spk3	58,91
6	47,39	pri_mam_06	37,5	spk1	28,1
				spk2	59,77
		foot_mar_12	38,44	spk1	37,92
				spk3	34,33
				spk2	51,03
		automobile_gue_08	41,26	spk1	38,79
				spk2	55,56
		etudes_rem_13	50,97	spk3	52,99
				spk2	60,86
				spk1	41,81
		fraise_ber_13	55,75	spk2	56,73
				spk3	53,14
				spk1	56,4
		salledebain_sch_13	56,16	spk2	63,8
				spk1	56,79
				spk3	46,66
		decesprofesseur_sd	52,35	spk1	50,81
				spk3	56,16
				spk2	56,31
		cours_mau_15	38,36	spk2	38,11
				spk1	40,71
				spk3	35,98
7	51,8	mai_web_07	55,46	spk1	64,27
				spk2	40,7
		dav_gan_06	70,16	spk2	66,9
				spk1	72,28
		siderurgie_per_10	39,05	spk1	36,35
				spk4	73,33
				spk2	58,82
		siderurgie_guy_10	28,64	spk1	28,58
				spk2	100
		masc_dom_sd	63,6	spk2	67,98
				spk1	59,71
		grandsparents_bar_13	57,21	spk2	54,19

Suite en page suivante

CHAPITRE B : *Performances de validation croisée (WER) sur les différents cas d'étude*

Table B.1 – suite de la page précédente

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.
				spk3	81,6
				spk1	49,89
		famille_fer_08_1	39,89	spk2	37,31
				spk1	58,2
		conv_cai_6	63,22	spk1	67,19
				spk2	53,48
		politique_ham_15	37,64	spk2	35,86
				spk1	38,72
Ecart- 9,831590241 type	ET		13,59225727	ET	15,61055306

TABLE B.2 – WER par lot, par enregistrement et par locuteur pour le cas *unusual_close*

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.
1	31,17	CFPB-1083-3B	33,71	Léon_Neyssens	33,31
				Eva_di_Dio	27,09
				Arlette	46,6
		CFPB-1070-2A	26,79	Bernadette	28,36
				Elise	17,73
				Léa	74,02
2	26,84	CFPB-1000-1	34,09	Matthieu	37,27
				Grand-Jojo	33,11
				serveur	60
		CFPB-1020-1	18,43	Dimitri	18,35
				Camille	18,72
				Philippe	25
3	28,7	photographie_cou_14	20,45	spk2	21,66
				spk1	16,99
		film_tre_15	39,74	spk1	34,56
				spk2	43,91
		conversation_mat_08	36,46	spk1	33,11
				spk2	40,4
		chee_alb_sd	34,72	spk2	37,21
				spk1	21,84
		politique_car_14	23,39	spk2	23,39
				spk2	19,66
		infirmier_aud_14	19,8	spk1	21,93
				spk1	20,75
		jeu_thi_14	24,23	spk2	38,22
				spk2	22,06
4	19,69	nat_hou_07	26,86	spk1	33,62
				spk1	20,09
		enfant_lem_10	21,23	spk2	27,95
				spk2	14,97
		dos_cou_14	17,74	spk1	22,48
				spk1	16,03
5	25,97	how_pro_14	17,59	spk2	28,82
				spk1	21,37
		etudes_gil_13	18,79	spk2	17,71
				spk2	15,82
		loisirs_nat_06	16,57	spk1	4,16
				spk2	15,71

Suite en page suivante

CHAPITRE B : *Performances de validation croisée (WER) sur les différents cas d'étude*

Table B.2 – suite de la page précédente

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.
6	22,92	spiritualite_cel_14	23,92	spk1	24,44
				spk1	22,5
				spk2	100
				spk1	32,84
				spk2	43,53
				spk2	40,91
				spk1	39,41
				spk1	27,86
				spk2	18,8
				spk1	28,48
7	28,73	voyage_con_15	21,08	spk2	25,73
				spk1	13,15
				spk1	21,15
				spk3	17,15
				spk1	60,16
				spk1	16,32
				spk2	100
				spk1	26,26
				spk2	29,71
				spk2	37,03
8	26,66	cha_hey_07	39,59	spk1	49,34
				spk1	19,14
				spk2	25,54
				spk1	42,99
				spk2	22,62
				spk2	28,07
				spk1	47,01
				spk1	24,94
				spk2	54,54
				spk1	14,86
		accident_cat_14	17,47	spk2	18,02
				spk2	14,43
				spk1	15,04
				spk1	58,33
				spk2	52,94
		msf_blan_06	26,11	spk2	26,72
				spk1	19,44
				spk2	23,55
				spk1	74,64
		memoire_yun_15	31,68	spk1	20,95
				spk1	Suite en page suivante

Table B.2 – suite de la page précédente

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.	
9	43,31	deltaplane_tus	21,22	spk2	42,32	
				spk2	21,2	
				spk1	22,22	
		sncf_dez_11	22,37	spk2	22,23	
			46,79	spk1	22,65	
				spk2	45,96	
			30,57	spk1	58,8	
				spk2	34,03	
			86,06	spk1	23,81	
				spk2	87,08	
		psychologue_dum_08	24,53	spk1	75,75	
			40,39	spk2	22,2	
				spk1	33,04	
10	35,25	christine_pru	18,23	spk1	38,78	
				spk2	48	
		mac_cle_sd	43,16	spk1	18,01	
			43,15	spk2	23,65	
		guerre_coc_sd	42,86	spk2	42,38	
11	40,32			spk1	46,78	
	48,36		spk2	43,15		
			spk1	35,57		
	35,65		spk1	49,48		
			spk2	44,16		
	27,84		spk1	32,1		
			spk2	48,97		
	38,03		spk1	25,94		
			spk2	49,45		
	12,61754847		spk1	33,62		
			spk2	51,14		
ET	7,139186228	ET			17,79848273	

CHAPITRE B : *Performances de validation croisée (WER) sur les différents cas d'étude*

TABLE B.3 – WER par lot, par enregistrement et par locuteur pour le cas *unusual_distant*

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.
1	70,2	aperitif_glasgow	61,98	JUD	58,9
				PAT	64,76
1	19,02	PRO-COR-1	21,85	L2	21,41
				L1	40,9
		PRO-GRE-1	15,34	L2	15,17
				L3	21,55
				L1	12,24
2	22,41	PRO-LIM-1	18,87	L1	18,87
				spk2	21,43
		ESLO2_ENT_1015	25,32	spk1	42,43
				spk2	17,34
3	18,19	ESLO2_ENT_1043	18,01	spk1	20,71
				spk2	33,96
		ESLO2_RUMEUR_1334	23,62	spk1	9,04
				spk2	21,66
4	24,05	ESLO2_ENT_1014	18,08	spk1	17,66
				spk2	18,54
		ESLO2_ENT_1050	18,36	spk1	17,6
				spk2	23,73
5	22,15	ESLO2_ENT_1013	23,19	spk1	20,27
				spk2	25,1
		ESLO2_ENT_1058	25,05	spk1	24,88
				spk2	21,81
6	34,96	joueurmusique_02	44,37	spk1	24,02
				spk2	49,13
		siderurgie_mar_10	24,31	spk1	24,13
				spk2	22,44
7	37,57	escalade_mic	39,25	spk1	31,62
				spk2	36,79
		apiculteur_sd	17,73	spk1	57,85
				spk2	15,96
		explorationssonores	94,56	spk1	28,64
				spk2	94,54
		2crossfit_mao_15	23,7	spk1	100
				spk2	22,8
		39_45_eva_14	15,47	spk1	47,5
				spk2	15,46
		rechercheancer_06	22,64	spk1	21,09
				spk2	

Suite en page suivante

Table B.3 – suite de la page précédente

Lot	WER/ lot	Enregistrement	WER/ audio	Locuteur	WER/ loc.
8	37,25	rwanda_lam_sd	43,36	spk1	29,44
				spk2	38,75
				spk1	44,4
		mili_89	29,84	spk1	29,63
		pomp_prov_sd	46,11	spk2	53,7
				spk1	43,03
9	26,12	incen_prov	33,43	spk2	46,59
				spk1	31,51
		quen_quen_sd	10,05	spk2	48,74
				spk1	9,55
		gestapo_sd	23,23	spk2	50
				spk1	22,95
		apprendrealycee	14,23	spk2	32,65
				spk1	10,53
		informaticien_bio_10	20,58	spk2	17,28
				spk1	20,11
10	26,46	orthophoniste_gom_12	30,88	spk2	28,44
				spk1	27,94
		bmx_min	27,22	spk2	32,09
				spk1	15,59
		orthophoniste_lat_12	31,66	spk2	28,51
				spk1	32,21
		professeur_cez_08	108,14	spk2	29,91
				spk1	108,98
		pompe_bli_sd	27,73	spk1	104,65
				spk2	53,13
ET	7,268450546	etudiantesalariee_sd	33,52	spk2	23,6
				spk1	29,34
		tatouagepolynesie_06	20,73	spk1	44,69
				spk2	21,25
		tromboniste	31,89	spk1	14,64
				spk2	30,62
		orthophoniste_sow_13	26,3	spk1	36,69
				spk2	24,32
		australie_mes_sd	19,33	spk1	31,97
				spk2	18,78
				spk1	22,87
ET	7,268450546	ET	19,39480216	ET	20,43277236