Traitement automatique de la parole en réunion par dissémination de capteurs

Théo Mariotte

COLLEGE SCIENCES
DOCTORAL DE L'INGENIERIE
PAYS DE LA LOIRE ET DES SYSTEMES

Le Mans Université

THÈSE DE DOCTORAT DE

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

ÉCOLE DOCTORALE N° 602
Sciences de l'Ingénierie et des Systèmes
Spécialité : *Acoustique*

Par
Théo Mariotte

Traitement automatique de la parole en réunion par dissémination de capteurs

Thèse présentée et soutenue à Le Mans, le 11 Janvier 2024
Unité de recherche : Laboratoire d'Acoustique de l'Université du Mans (LAUM), Laboratoire d'Informatique de l'Université du Mans (LIUM)

Thèse N° : 2024LEMA1001

Rapporteurs avant soutenance :

Jan "Honza" ČERNOKÝ — Professeur, Brno University of Technology, République Tchèque
Emmanuel VINCENT — Directeur de Recherche, Inria Nancy - Grand Est, France

Composition du Jury :

Président :	Gaël RICHARD	Professeur, Télécom Paris, France
Examinatrice :	Julie MAUCLAIR	Maître de Conférences, IRIT, Toulouse, France
Dir. de thèse :	Jean-Hugh THOMAS	Professeur, LAUM, le Mans, France
Encadrants de thèse :	Anthony LARCHER	Professeur, LIUM, Le Mans, France
	Silvio MONTRÉSOR	Maître de Conférences, LAUM, Le Mans, France

Remerciements

Durant ces trois années et quelques, j'ai eu la chance d'être entouré · e de personnes géniales Leur contribution à ces travaux n'est pas négligeable même si difficilement quantifiable. Cette page, c'est pour elleux !

Commençons par les labos, et plus précisément par le LAUM. Cette thèse n'aurait tout simplement pas eu lieu sans la présence de Jean-Hugh, mon directeur de thèse. Bien qu'originaire d'un univers un peu éloigné des directions de recherche prises au cours de ces travaux, tu as toujours été présent et curieux de comprendre les méthodes proposées. Merci aussi pour ton suivi et ton œil aguerri, notamment sur les publications. Merci aussi à Silvio, mon autre encadrant LAUM, pour ta présence discrète mais pertinente.

Au LAUM, en particulier à l'ENSIM, j'ai eu la chance de partager mon bureau avec des personnes très chouettes, avec qui j'ai vécu de beaux moments. Merci à vous : Erwan pour le tennis et les cafés allongés ; Guillaume pour les discussions culinaires à la F. R. Gaudry ; Patrick pour les discussions de geek sur les synthétiseurs et tous les autres sujets qui t'animent ; Chiara pour ton court séjour riche en rigolade ; Ammar pour le tennis dans le bureau et ton rire salvateur ; et tous les autres...

Après le LAUM, il y a le LIUM ! Et au LIUM, il y a Anthony qui a encadré ces travaux de près. Merci pour tes idées, tes relectures diverses et aussi ta réactivité dans les moments de doute. Je n'oublierai pas les jams tardives dans ton salon ! Au LIUM, il y a aussi Marie qui a suivi une partie de ces travaux et avec qui j'ai travaillé au cours du workshop JSALT. Tu es un peu l'encadrante officieuse de ces travaux, et travailler avec toi est toujours un plaisir ! Et comment parler du LIUM sans penser aux doctorant · es avec qui j'ai partagé une bonne partie de mon temps ? Il y a Thibault, qui nous permet d'être à jour sur l'actualité, surtout s'il s'agit du Royaume-Uni. En face, il y a Valentin que j'ai beaucoup (trop ?) sollicité pour régler les petits problèmes de code. Quand on va dans la pièce d'à côté, on trouve Thibault G. C'est peut-être un dinosaure réincarné mais qui ne mord pas. Imbattable en géographie, en dates, et en pas mal d'autres trucs d'ailleurs ! Il fait face à Albane, dont le débit de parole égale la générosité et la gentillesse. De l'autre côté, c'est Simon, le linguiste-informaticien, adepte de synthés modulaires. Et juste en face de lui, Martin, percussionniste de qualité, et acolyte de conférences et d'articles rejetés (mais on va y arriver à publier ensemble, j'y crois) ! Et bien sûr, merci à toute l'équipe ! Notamment Etienne et Gregor qui réparent les bêtises sur le serveur ! Une petite ligne aussi pour Nico et son impitoyable coup droit :)

Mais que serait un labo sans l'administration qui tourne autour ? Merci au pôle doctoral pour leur réactivité et leurs relances pour les formations ! Et merci à Clémence pour ton travail, mais surtout les moments partagés avec nous au bar et au resto !

La thèse ce n'est pas que les labos, c'est aussi un entourage et des moments extra-muros. Et je commencerai par l'orchestre ! Merci à Cécile, Lucas, Camille, Baptiste, Suzanne et toute la team Sympho'Campus pour les beaux moments partagés lors des répétitions du mardi et des concerts.

Les bons moments, c'était aussi au 17 rue Saint-Jacques dans cette chouette collocation à colocataires variables. D'abord, il y a eu Audrey, coloc de cœur et amie de longue date maintenant ! Très hâte de refaire des repas et des soirées avec toi ! Puis Louise, spécialiste de l'acoustique picoseconde et grimpeuse hors pair. Et enfin Amélie, acousticienne et musicienne de la clarinette. Merci à toutes pour ces jolis instants partagés dans ce (très) grand appartement.

Viennent aussi les copain · es de thèse, rencontré · es pour la plupart avec la super association qu'est l'ADOUM. Il y a Valentine, depuis la première année, meilleure comparse pour aller au Septante-Deux ou taper du pied ! Il y a la clique des géologues aussi et notamment Sofyane et Jimmy, qui apportent toujours le sourire, quelle que soit la situation ! Et puis il y a Thomas, Soizic, Simon, Sam, Julie, Amel, Gustave et tout · tes celleux que j'oublie. Merci pour ces moments ! Et que seraient ces moments sans la Brasserie Septante Deux ou le Barouf ? Une thèse sans ces lieux n'aboutierait pas !

Quelques lignes pour les camarades du worksop JSALT 2023 ! Antonio, Pablo G., Victoria, Jazmine, Pablo R., Sergio, Patricia, Alexis, Alfonso, Lara... On a passé deux très beaux mois tous ensemble !

Le Mans c'est petit, et c'est quand même bien (si si), mais tout le monde n'y vit pas. Merci à tou · tes mes ami · es d'un peu partout qui permettent de rendre la vie plus douce. Xavier, Téo, Hiley, Julien, Killian, Hugo, Morgane... Cœur sur vous !

A 4h de train du Mans, se trouve la jolie ville de Strasbourg. J'y ai passé un bon nombre de jours au cours de ces trois années, notamment à partager les 25 m^2 de confinement avec toi et à télétravailler entre quelques piles de partitions et de livres. Merci Camille d'être là, d'estomper mes doutes et de nourrir mes joies.

Un peu plus au sud, il y a la Bourgogne et ses collines dorées. Et dans une petite vallée perdue, il y a mes parents qui m'ont toujours soutenu, et même quand ça ne captait pas très bien ! Merci d'avoir toujours rendu possibles mes ambitions. Et encore plus au sud, à Lyon, il y a ma petite sœur. Un rayon de soleil qui ne laissera jamais l'ombre s'attarder ! Merci d'être là !

Merci à tou · tes ! ♥

TABLE DES MATIÈRES

TABLE DES FIGURES

Liste des tableaux

Acronymes

A :

ACU Antenne Circulaire Uniforme

ALU Antenne Linéaire Uniforme

AP Average Precision, précision moyenne

B :

BFSACC Beamforming Self Attention Channel Combinator

BLSTM Bidirectionnal Long Short-Term Memory

BRNN Bi-directionnal Recurrent Neural Network, Réseau de neurone récurrent bi-directionnel

C :

CNN Convolutional Neural Network, réseau de neurone convolutif

cSACC Complex Self-Attention Channel Combinator, Combination auto-attentive des canaux dans le domaine complexe

CSIPD Cosine and Sine of Interaural Phase Difference, Cosinus et Sinus de la Différence de phase interaurale

D :

DER Diarization Error Rate

DOA Direction of Arrival, direction d'arrivée

E :

EcSACC Explicit SACC, cSACC explicite

EEND End-to-End Diarization, diarisation bout-en-bout

EER Equal Error Rate

G :

GCC Generalized Cross Correlation, Corrélation Croisée Généralisée

GEV Generalized Eigenvalue, Valeurs Propres Généralisées

GMM Gaussian Mixture Model, modèle à mélange de Gaussiennes

GRU Gated Recurrent Unit

H :

HMM Hidden Markov Model, Modèle de Markov

I :

IcSACC Implicit SACC, cSACC implicite

ILD Interaural Level Difference, Différence de niveau interaurale

IPD Interaural Phase Difference, Différence de phase interaurale

J :

JER Jaccard Error Rate

L :

LcSACC Linear complex Self Attention Channel Combinator

LSTM Long Short-Term Memory

M :

MDU Microphone Distant Unique

MFCC Mel Frequency Cepstral Coefficient, coefficients cepstraux à échelle Mel

MLP Multi-Layer Perceptron, perceptron multi-couche

MVDR Minimum Variance Distortion-less

N :

NMV Normalisation des moyenne et variance

O :

OSD Overlapped Speech Detection, détection de parole superposée

P :

PHAT PHAse Transform, Transformation de phase

R :

RIS Réponse Impulsionnelle de Salle

RNA Réseau de Neurones Artificiel

RNN Recurrent Neural Network, Réseau de neurone récurrent

S :

SACC Self-Attention Channel Combinator, Combination auto-attentive des canaux

SCD Speaker Change Detection, détection de changement de locuteur

SER Segmentation Error Rate, taux d'erreur de segmentation

SRP Steered Response Power, Puissance Dirigée

T :

TCD Transformée en Cosinus Discrète

TCN Time Convolutional Network, réseau convolutif temporel

TDOA Time Difference of Arrival, Temps d'arrivée

TF Transformée de Fourier

TFCT Transformée de Fourier à Court Terme

V :

VAD Voice Activity Detection, détection de parole

W :

WER Word Error Rate

Chapitre 1

Introduction

La communication parlée est un élément essentiel pour la transmission d'informations entre individus. La parole transmet des éléments clefs pour les interactions orales telles que le contenu linguistique, c'est-à-dire le message que le locuteur veut transmettre, l'identité de ce dernier ainsi que d'autres informations paralinguistiques (ex : état émotionnel).

La parole est donc vecteur d'information, à la fois sur le locuteur et le message qu'il désire transmettre. Lors d'interactions entre individus, il peut être intéressant d'extraire automatiquement une partie de ces informations. Par exemple, le contenu linguistique peut être transcrit sous forme d'un texte écrit pour une personne malentendante ou l'identité du locuteur actif peut être déterminée afin d'augmenter une vidéoconférence. L'extraction automatique d'information à partir d'un signal de parole fait appel à de nombreuses technologies, rassemblées sous le terme générique *traitement automatique de la parole*.

Le traitement automatique de la parole prend une place de plus en plus importante dans notre quotidien, notamment à travers le développement des assistants vocaux, des environnements virtuels et des moyens de communication. L'intelligence artificielle (IA) constitue le cœur des systèmes de traitement de la parole. Ces technologies, visant à « imiter les fonctions de résolution de problèmes et de prise de décision du cerveau humain » [1] permettent aujourd'hui des performances remarquables. Notamment, l'apprentissage profond supervisé (*deep learning*) offre, depuis une vingtaine d'années, des capacités de modélisation fine de grands jeux de données. Ce type d'algorithme constitue, à la date d'écriture de ce manuscrit, l'état de l'art pour la majorité des systèmes de traitement automatique de la parole.

Parmi les approches de traitement de la parole, on trouve la transcription automatique de conversations (réunions, conseils municipaux, verbatim de tribunaux...). Dans ce contexte, il est utile d'enrichir la transcription en indiquant l'identité des locuteurs intervenants au sein de chaque contenu. C'est ce cas d'usage que nous adressons dans ce document. Plus spécifiquement, les travaux menés considèrent le cas des réunions. Le développement de méthodes de transcriptions enrichies permet de simplifier la prise de notes ou encore de synthétiser les informations importantes (résumé automatique).

Le traitement automatique de la parole passe d'abord par l'acquisition du signal à l'aide d'un microphone. Ce dernier transforme une onde de pression acoustique en un signal électrique. Au cours d'une réunion, l'acquisition de la parole de chaque participant peut être réalisée au moyen

1. https://www.ibm.com/fr-fr/topics/artificial-intelligence consulté le 10 juillet 2023

de microphones individuels (ex : micro-cravate). Dans ce cas, chaque locuteur est enregistré sur un canal séparé. Cette configuration implique d'équiper les locuteurs d'un système individuel, rendant la gestion du nombre de participants délicate. L'utilisation d'un microphone unique, placé au centre de la table et enregistrant l'ensemble de la réunion, semble plus adéquat. Ce type de dispositif lève cependant des problématiques techniques :

- les participants sont loin du dispositif, les signaux acquis tendent à être altérés par le bruit ambiant et la réverbération,
- un seul canal est utilisé pour tous les locuteurs. Le microphone peut alors capter des contributions simultanées de ces derniers.

Ces contraintes sont sources de dégradations des performances des systèmes de traitement automatique de la parole. Pour les limiter, il est possible d'ajouter de l'information supplémentaire en multipliant le nombre de microphones. Les dispositifs multi-microphones sont appelés *antennes de microphones* et permettent d'acquérir un signal *multicanal*. Ce dernier contient des informations sur la répartition des locuteurs dans l'espace pouvant renforcer les systèmes de traitement de la parole.

Les données disponibles pour la mise en œuvre et l'évaluation des systèmes de traitement de la parole multicanale sont rares. La plupart d'entre elles contiennent des signaux de parole acquis au cours de réunions. Les travaux menés au cours de cette thèse sont donc focalisés sur le traitement de la parole en réunion, bien que les approches puissent être étendues à d'autres contextes.

Dans ce chapitre, la section 1.1 introduit le traitement automatique de la parole. Les problématiques rencontrées ainsi que les tâches résolues dans le contexte de la parole distante multicanale sont détaillées en section 1.2. La structure du manuscrit et les contributions apportées sont décrites en section 1.3 avant de lister les publications liées à ces travaux.

1.1 De la production de la parole à son traitement automatique

Le traitement automatique de la parole consiste à analyser et à extraire des informations du signal à l'aide d'un ordinateur. Le signal traité est avant tout issu de l'acquisition d'une onde acoustique convertie à l'aide d'un microphone. Cette section présente brièvement les mécanismes de production de la parole avant d'introduire les méthodes d'acquisition du signal et l'historique du traitement automatique de la parole.

1.1.1 Production de la parole et acquisition

Un signal de parole est une onde acoustique générée par l'appareil vocal humain. Ce dernier est composé d'un mécanisme excitateur (les cordes vocales) et d'un résonateur (les cavités buccale et nasale). Les cordes vocales entrent en vibration suite au passage d'un flux d'air généré par les poumons (expiration). Les cavités buccale et nasale permettent de filtrer l'onde acoustique émise

par les cordes vocales par des effets de résonance. Les caractéristiques sonores de la phonation dépendent du débit d'air, de la tension des cordes vocales et de la forme des cavités.

La physiologie de chaque locuteur est unique, et chacun présente un appareil de production vocale propre. Les caractéristiques acoustiques sont susceptibles de varier d'un locuteur à un autre. Par exemple, la biométrie et la reconnaissance du locuteur font l'hypothèse que ces caractéristiques acoustiques sont propres à chacun. Bien qu'elles dépendent des états physique et émotionnel du locuteur, la parole nous permet de distinguer différents locuteurs au cours d'une conversation. Le contenu acoustique permet alors de caractériser l'identité de la personne ayant émis un signal de parole.

Afin de traiter la parole automatiquement, le signal acoustique doit être converti en un signal électrique à l'aide d'un microphone. Ce dispositif est couramment composé d'une membrane oscillant sous la pression acoustique. Celle-ci est connectée à un transducteur permettant de convertir l'énergie mécanique de la membrane en un signal électrique. Les approches actuelles de traitement du signal sont réalisées dans le domaine numérique. Le signal électrique analogique est donc échantillonné et quantifié afin d'être converti dans le domaine numérique.

Les signaux de parole acquis peuvent ensuite être traités à l'aide d'un ordinateur afin d'en extraire des informations (ex : contenu lexical, identité du locuteur...). Ces opérations peuvent être réalisées à l'aide de méthodes automatiques basées sur le traitement du signal et l'apprentissage machine, regroupées dans le domaine du traitement automatique de la parole.

1.1.2 Traitement automatique de la parole

Le traitement automatique de la parole englobe de nombreuses tâches permettant d'extraire des informations d'un signal émis par un ou plusieurs locuteurs. Les tâches les plus communes sont les suivantes :

- reconnaissance de la parole,
- reconnaissance du locuteur,
- segmentation et regroupement en locuteurs,
- rehaussement et séparation de sources,
- etc.

Les efforts de recherche dans le domaine du traitement automatique de la parole apparaissent avec le développement du téléphone dans les années 1920. Les premières approches permettent de reconnaître certains phonèmes dans le signal de parole émis par un locuteur unique. Elles se sont ensuite développées, principalement pour la transcription automatique, au cours des années 1940 (Juang et al. 2006). L'intérêt pour ces méthodes a grandi au cours des années 1960, notamment avec l'arrivée du codage prédicatif linéaire (CPL) introduit par le *Nippon Telegraph and Telephone* (NTT) (Gray 2010). Des algorithmes de synthèse de la parole embarqués ont commencé à voir le jour dans les années 1970, toujours basés sur le CPL (Gray 2010).

La recherche dans le domaine s'oriente ensuite vers des techniques d'apprentissage machine, basées sur des modèles statistiques tels que les mélanges de Gaussiennes (Gaussian Mixture Models, GMM) et les modèles de Markov (Hidden Markov Models, HMM) (Anguera et al. 2012). Ces approches sont aujourd'hui remplacées par les réseaux de neurones artificiels (RNA), permettant d'obtenir des performances remarquables pour la majorité des tâches (Feng et al. 2023).

1.2 Définition du problème

Les travaux menés portent sur le traitement automatique de la parole en réunions. Les problématiques techniques levées par ces conditions d'acquisition sont décrites dans cette section avant de formaliser les tâches visées.

1.2.1 Acquisition distante d'un signal

Dans ces travaux, nous considérons dans la majorité des cas une antenne de microphones placée au centre du groupe de locuteurs. L'acquisition du signal de parole est alors *distante*, les capteurs étant placés à quelques décimètres, voire plusieurs mètres des participants. Les conditions en champ lointain (distantes) s'opposent au champ proche, où le microphone est à proximité de la bouche des locuteurs. Le traitement de la parole distante introduit des problématiques techniques, décrites dans les sous-sections suivantes.

Définition et problématique

Dans la majorité des jeux de données disponibles pour le traitement automatique de la parole, les données sont acquises en champ proche. Le microphone d'acquisition est donc proche de la source émettant le signal de parole $p(t)$. Le modèle de signal dans cette configuration s'exprime :

$$x(t) = p(t) + n(t), \tag{1.1}$$

où $n(t)$ représente un bruit additif lié à différents facteurs (ex : capteur).

En conditions de captation distante, le microphone va capturer le signal source $p(t)$ mais également les réflexions sur les parois de la salle et les obstacles environnants. Ce phénomène est schématisé en figure 1.1. D'autres phénomènes physiques différencient les signaux en champ proche et en champ lointain comme les ondes évanescentes. La simplification par des chemins de propagation permet cependant de comprendre les phénomènes de réverbération en jeu dans le cas de la parole distante en milieu fermé. Plusieurs chemins de propagation interviennent :

- le chemin direct, entre la source et le microphone (en rouge),
- les premières réflexions, correspondant à une version retardée et atténuée du signal (en bleu),

- les réflexions tardives formant ce qu'on appelle la réverbération (en orange).

Cette même figure présente le profil type d'une réponse impulsionnelle de salle (RIS), c'est-à-dire la réponse de la salle à une impulsion de Dirac. Le signal capté par un microphone distant est filtré par la réponse de la salle, venant altérer le signal. Un signal distant s'écrit :

$$x(t) = (p * h)(t) + n(t), \tag{1.2}$$

avec $h(t)$ la RIS et $*$ le produit de convolution. Un signal capté en conditions distantes est donc sujet à des détériorations par la réverbération. Le niveau sonore du signal source au point d'acquisition est également plus faible. Le rapport signal-à-bruit, c'est-à-dire le rapport des puissances du signal utile (parole) sur ceux du bruit, tend à être plus faible.

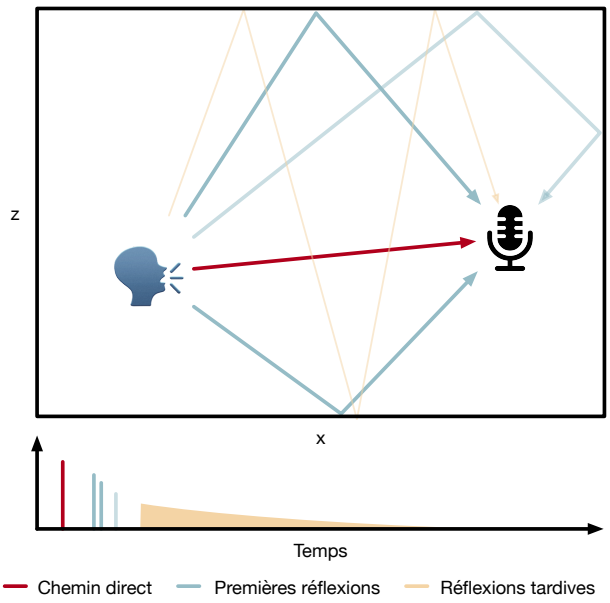


FIGURE 1.1 – Illustration du phénomène de réverbération en cas de propagation acoustique en milieu fermé.

Les signaux de parole distante sont dégradés par rapport aux signaux captés en champ proche. Cette dégradation impacte négativement les systèmes de traitement automatique de la parole. MACIEJEWSKI et al. (2018) évaluent un système de segmentation et regroupement en locuteurs sur des données acquises en réunions à l'aide d'un microphone distant unique. Ils montrent que les performances segmentation et regroupement en locuteurs sont dégradées d'environ 10% par rapoprt à une acquisition des signaux en champ proche. Ce contexte motive l'utilisation de plusieurs microphones afin d'acquérir des informations supplémentaires.

Dispositifs d'acquisition composés de plusieurs microphones

L'utilisation de plusieurs microphones permet d'acquérir des informations supplémentaires sur le champ acoustique. La figure 1.2 schématise quatre chemins de propagation intervenants

lorsque deux microphones sont présents. Pour une même source (locuteur dans notre cas), la géométrie des chemins est différente pour chaque microphone. Le signal issu de chaque chemin sera filtré différemment par la salle, modifiant ainsi le contenu du signal issu de chaque microphone. Finalement, le signal multicanal acquis contient des informations sur la répartition spatiale du champ acoustique. Cette information spatiale peut être utilisée pour localiser les sources actives (Grumiaux et al. 2021) ou réaliser un filtrage spatial (Benesty et al. 2008).

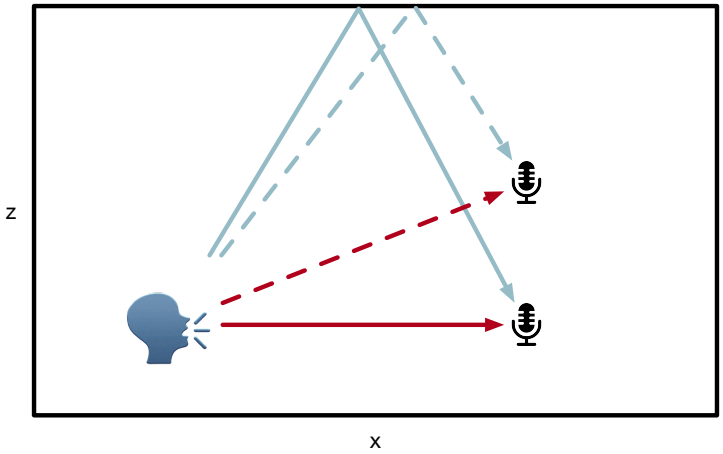


Figure 1.2 – Exemple de plusieurs chemins de propagation lorsque deux microphones sont utilisés pour capturer la scène acoustique.

L'information spatiale contenue dans un signal acquis par plusieurs microphones est corrélée à l'activité des sources. Cette information peut alors être utile dans le contexte du traitement automatique de la parole (Y. Liu et al. 2014 ; Yoshioka et al. 2018). En particulier, elle peut permettre de faciliter la segmentation du signal de parole (Cornell et al. 2022a). La segmentation de la parole est une tâche clef pour la segmentation et le regroupement en locuteurs (Anguera et al. 2012 ; T. J. Park et al. 2022). Les travaux menés portent principalement sur l'utilisation d'information spatiale pour la segmentation de la parole dans le contexte de la diarisation en locuteurs. Les tâches visées sont introduites et formulées dans la sous-section suivante.

1.2.2 Segmentation du signal de parole

Contexte

Les travaux menés portent sur la tâche de segmentation et de regroupement en locuteurs. Cette tâche sera également appelée *diarisation* ou *diarisation en locuteurs* par la suite. La diarisation en locuteurs consiste à partitionner le signal audio en segments homogènes en fonction de l'identité du locuteur. Cela revient à répondre à la question *Qui a parlé et quand ?*, dans un signal de parole.

La majorité des approches de diarisation exploite des modèles en cascade, composés de

plusieurs blocs distincts résolvant une sous-tâche. La figure 1.3 présente la structure actuelle des algorithmes de diarisation en locuteurs en cascade (Bredin et al. 2020). La première étape de segmentation consiste à détecter la parole dans le signal. Il s'agit de la *détection d'activité vocale* (VAD). Pour permettre la détection de segments homogènes en locuteur, les frontières entre les locuteurs doivent être connues. La *détection de changement de locuteur* (SCD) permet de détecter les frontières des segments entre les locuteurs. Une troisième tâche peut être ajoutée pendant la segmentation : *la détection de parole superposée* (OSD). Elle permet de détecter les segments au sein desquels plusieurs locuteurs sont actifs simultanément. Après la segmentation du signal, une représentation est extraite de chaque segment afin de les regrouper par locuteur. Ces représentations, appelées *embeddings*, sont optimisées pour séparer les locuteurs dans leur espace de représentation (Snyder et al. 2018). Les segments avec des caractéristiques de locuteurs similaires seront regroupés. La re-segmentation consiste à affiner les frontières des segments et permet souvent d'assigner les locuteurs aux segments de parole superposée (Otterson et al. 2007).

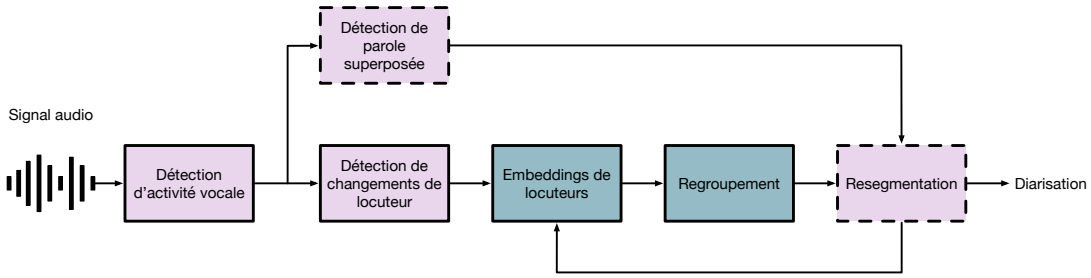


Figure 1.3 – Diagramme de diarisation du locuteur en cascade.

Les travaux menés durant cette thèse se focalisent sur les tâches de segmentation de la parole. Les tâches de VAD, OSD et SCD sont considérées dans le contexte de la parole distante, acquise à l'aide de plusieurs microphones. La figure 1.4 schématise les sorties attendues pour chacune des tâches considérées. Chaque tâche est considérée comme une classification binaire : absence ou présence de l'évènement détecté. Cette classification est réalisée à partir de *caractéristiques* extraites du signal sur de courtes fenêtres appelées *trames*. La sous-section suivante présente les notations et le formalisme utilisé pour résoudre ces tâches.

Formalisme

La segmentation de la parole est réalisée à partir d'un signal $\boldsymbol{x} \in \mathbb{R}^{M \times L}$ acquis à l'aide de M microphones. L représente le nombre d'échantillons temporels composant le signal. Le signal

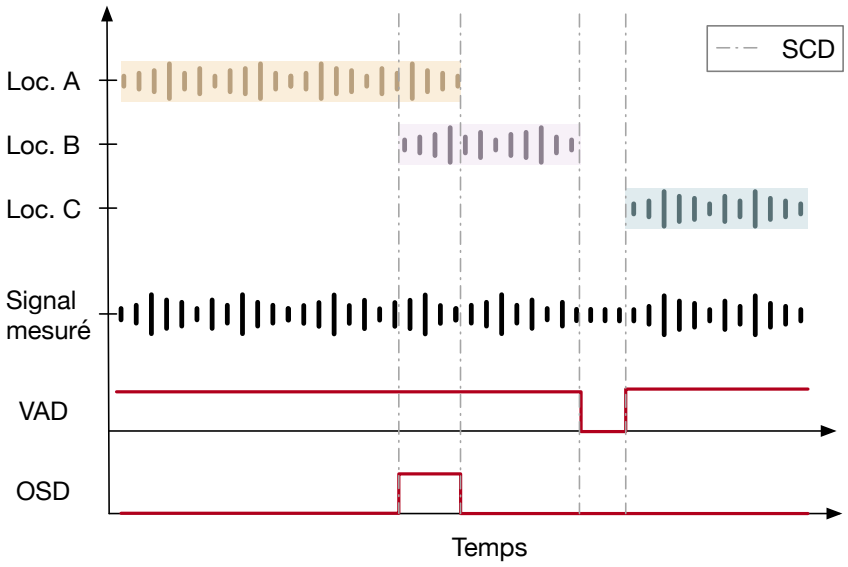


FIGURE 1.4 – Exemple des sorties attendues pour chaque tâche de segmentation de la parole VAD et OSD dans le cas où trois locuteurs (Loc. A, B et C) sont actifs. Les traits mixtes verticaux indiquent les frontières des segments, détectés à l'aide de la tâche de SCD.

acquis par l'ensemble de capteurs s'écrit :

$$\boldsymbol{x}(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_m(t) \\ \vdots \\ x_M(t) \end{bmatrix}, \tag{1.3}$$

avec $m = \{1, \cdots, M\}$ l'indice du microphone courant, t l'indice de l'échantillon temporel et $x_m(t)$ le signal acquis par le microphone m. Pour segmenter la parole, des caractéristiques $\boldsymbol{X} \in \mathbb{R}^{F \times T}$ sont généralement extraites à partir du signal multicanal, avec F le nombre de caractéristiques et T le nombre de trames. Celles-ci sont extraites à l'aide d'une fonction $g : \mathbb{R}^{M \times L} \rightarrow \mathbb{R}^{F \times T}$. La segmentation consiste à prédire la probabilité de chaque trame de $\boldsymbol{X}$ d'appartenir à chaque classe $c \in \{0, 1\}$. Par exemple, dans le cas de la VAD, la classe $c = 1$ correspond à la présence de parole. Les prédictions sont obtenues à l'aide d'une fonction $f : \mathbb{R}^{F \times T} \rightarrow \mathbb{R}^{C \times T}$, avec C le nombre de classes.

Les travaux menés s'intéressent principalement au développement de fonctions g pour extraire des caractéristiques à partir d'un signal acquis à l'aide de M microphones. Ces caractéristiques exploitent l'information spatiale contenue dans le signal multicanal. Leur impact est évalué sur les tâches de segmentation préalablement définies : VAD, OSD et SCD. L'influence de la segmentation en conditions distantes est évaluée sur la diarisation en locuteurs.

1.3 Structure du manuscrit et contributions

Le manuscrit est organisé en deux parties. La première présente l'état de l'art en deux chapitres. Le premier s'intéresse aux méthodes de modélisation et à la segmentation de la parole pour la diarisation en locuteurs. Le second détaille les approches de traitement du signal multicanal et ses applications au traitement de la parole. La seconde partie présente les contributions en quatre chapitres. La figure 1.5 présente la structure du manuscrit, en partant des données (signaux de parole multicanale) vers les contributions obtenues.

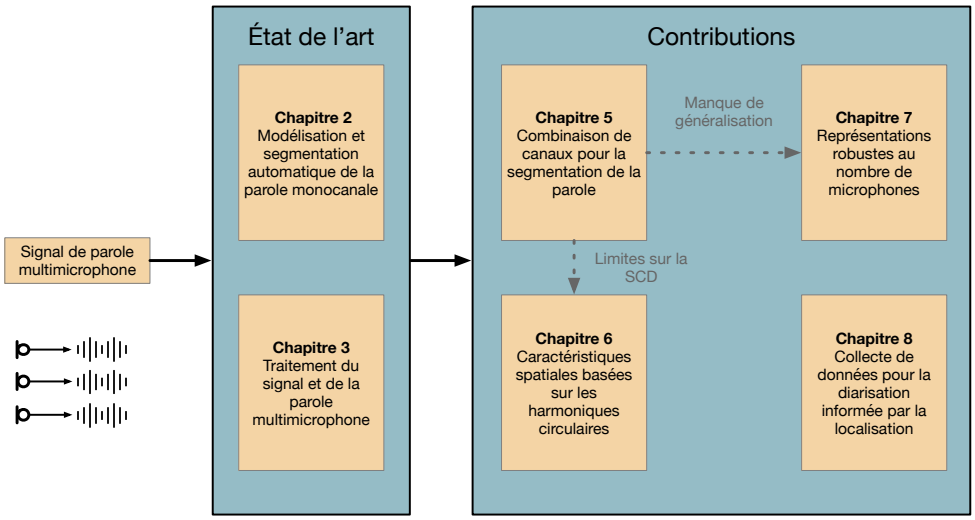


FIGURE 1.5 – Structure du manuscrit sur les travaux menés en segmentation de la parole à l'aide de plusieurs microphones.

Chapitre 2 : état de l'art sur la segmentation de la parole monocanale

Ce chapitre présente l'état de l'art sur la segmentation de la parole dans le contexte de la diarisation du locuteur. Les approches récentes exploitent principalement les réseaux de neurones artificiels pour modéliser le signal de parole ou un ensemble de caractéristiques acoustiques extraites de ce dernier. Une revue de ces technologies est donc proposée. Les différents types de caractéristiques acoustiques, permettant de représenter le signal de parole pour son traitement, sont ensuite introduites. L'état de l'art sur les tâches de segmentation VAD, OSD et SCD est finalement présenté. Enfin, les jeux de données disponibles pour le traitement de la parole en réunion sont détaillés.

Chapitre 3 : état de l'art sur le traitement automatique de la parole multicanal

Ce chapitre présente les approches de la littérature pour le traitement de la parole à l'aide de plusieurs microphones. Les dispositifs d'acquisition sont d'abord présentés en introduisant deux géométries d'antennes de microphones couramment utilisées. La deuxième section présente

les méthodes de formation de voies, permettant de réaliser un filtrage spatial. Les algorithmes classiques et les approches récentes sont présentés. La troisième section présente des méthodes de représentation du signal multicanal. Cela passe par l'extraction de caractéristiques spatiales. La dernière section présente les méthodes de diarisation du locuteur et de segmentation de la parole utilisant plusieurs microphones.

Chapitre 4 : synthèse, axes de recherche et protocole d'évaluation

Ce chapitre résume les deux chapitres d'état de l'art précédents. Les axes de recherches levés par cette étude sont introduits. Le protocole d'évaluation employé au cours de ces travaux est également décrit, celui-ci étant similaire pour chacun des chapitres suivants.

Chapitre 5 : combinaison auto-attentive pour la segmentation de la parole

Ce chapitre étudie l'utilisation de mécanismes d'auto-attention pour la pondération et la combinaison de canaux dans le domaine temps-fréquence. Une approche de la littérature est d'abord appliquée à la VAD et à l'OSD. Trois axes d'extension de cette approche sont ensuite étudiés :

- la représentation temps-fréquence est remplacée par des filtres temporels optimisées,
- le modèle apprend des poids complexes pour permettre une meilleure interprétation des poids appris par le modèle,
- l'auto-attention est utilisée pour sélectionner les signaux de sorties d'un banc de filtres spatiaux.

L'impact de chaque méthode proposée est ensuite évalué sur la tâche de diarisation en locuteurs.

Chapitre 6 : caractéristiques spatiales robustes pour la segmentation de la parole

Ce chapitre propose d'utiliser des caractéristiques spatiales extraites dans le domaine des harmoniques circulaires. Cette méthode permet d'améliorer la robustesse du système de segmentation en cas de changement dans la géométrie de l'antenne, et principalement en cas de désactivation d'un ou plusieurs microphones. Ces caractéristiques sont utilisées pour la VAD, l'OSD et la SCD et montrent leur intérêt principalement sur la dernière tâche. La segmentation est également évaluée dans le contexte de la diarisation en locuteurs.

Chapitre 7 : apprentissage d'une représentation robuste au nombre de canaux

Ce chapitre présente deux approches permettant d'obtenir une représentation robuste au nombre de capteurs disponibles. La première contraint le système de segmentation à produire une séquence de caractéristiques similaire quel que soit le nombre de capteurs disponibles. Deux fonctions de perte sont explorées au sein de trois architectures de détection de parole superposée (OSD). Les résultats montrent que cette méthode d'apprentissage simple améliore

significativement la robustesse du système. La deuxième approche exploite un modèle pré-entraîné pour représenter le signal issu d'un unique microphone de l'antenne. Cette section illustre les capacités de modélisation offertes par ce type de modèles.

Chapitre 8 : vers l'acquisition de données en réunion pour la diarisation informée par la localisation

Ce chapitre introduit le protocole d'acquisition de données développé pour la construction d'un jeu de données de parole en réunion. L'objectif est d'enregistrer des sessions à l'aide d'une antenne de microphones et de fournir les annotations des segments de locuteurs ainsi que de leur localisation. Le protocole d'acquisition et la procédure de pré-annotation sont présentés dans ce chapitre. La collecte de données à grande échelle n'a cependant pas été réalisée.

Chapitre 9 : conclusions et perspectives

Ce chapitre présente les conclusions tirées des différents travaux menés au cours de cette thèse. Les perspectives envisagées pour le traitement automatique de la parole à l'aide d'antennes de microphones sont également présentées.

1.4 Publications

Ces trois années de travaux ont principalement porté sur la segmentation de la parole distante à l'aide de plusieurs microphones. Deux autres publications s'écartent du domaine de la thèse et sont issues des travaux menés au cours du workshop JSALT 2023. Parmi les expériences menées, certaines d'entre elles ont conduit aux publications et communications suivantes :

Conférences nationales

- T. Mariotte, A. Larcher, S. Montrésor, J-H. Thomas, Traitement Multi-Microphone pour la Segmentation Automatique de la Parole en Réunion, *Congrès Français d'Acoustique (CFA)*, Marseille, France, 2022
- T. Mariotte, A. Larcher, S. Montrésor, J-H. Thomas, Détection de parole superposée distante à l'aide d'une antenne de microphones, *Journées d'Étude sur la Parole (JEP)*, Noirmoutier-en-l'île, France, 2022

Conférences internationales avec comité de relecture

- T. Mariotte, A. Larcher, S. Montrésor, J-H. Thomas, Microphone Array Channel Combination Algorithms for Overlapped Speech Detection, *Interspeech 2022*, Incheon, Corée du Sud, 2022

- T. Mariotte, A. Larcher, S. Montrésor, J-H. Thomas, Multi-microphone Automatic Speech Segmentation in Meetings Based on Circular Harmonics Features, *Interspeech 2023*, Dublin, Irlande, 2023
- T. Mariotte, A. Almudevar, M. Tahon, A. Ortega, An Explainable Proxy Model For Multilabel Audio Segmentation, *ICASSP*, Séoul, Corée du Sud, 2024
- A. Almudevar, T.Mariotte, A. Ortega, M. Tahon, Unsupervised Multiple Domain Translation through Controlled Disentaglement in Variational Autoencoder, *ICASSP*, Séoul, Corée du Sud, 2024

Revue avec comité de relecture

- T. Mariotte, A. Larcher, S. Montrésor, J-H. Thomas, Channel-Combination Algorithms for Robust Distant Voice Activity and Overlapped Speech Detection, **Soumission en cours** *Transactions on Audio, Speech and Language Processing (TASLP)*, 2023

Première partie

État de l'art

Chapitre 2

Modélisation et représentation de la parole monocanale pour la segmentation automatique

La segmentation et regroupement en locuteur, ou diarisation du locuteur, est une tâche clef pour le traitement automatique de la parole. Elle permet de déterminer l'activité de chaque locuteur actif au cours d'une conversation. Les approches actuelles reposent notamment sur la modélisation des signaux de parole afin d'en extraire des segments homogènes en locuteur. Elles reposent notamment sur deux tâches :

- la segmentation, consistant à trouver les frontières temporelles dans le signal,
- la modélisation du locuteur, visant à développer des représentations discriminant les locuteurs.

La majorité de ces méthodes utilisent aujourd'hui l'apprentissage automatique et les réseaux de neurones profonds, permettant une représentation et une modélisation fine des données. Ce chapitre présente d'abord une vue d'ensemble des architectures neuronales pouvant être utilisées pour la diarisation du locuteur.

La modélisation des signaux de parole requiert d'en extraire une représentation adaptée pour la tâche visée. Bien que des approches exploitant les signaux bruts voient le jour, il est souvent requis d'en extraire des caractéristiques. Celles-ci permettent d'obtenir une représentation plus explicite de l'information contenue dans le signal. Les caractéristiques usuelles pour la segmentation et le regroupement en locuteur sont donc présentées.

Les travaux menés au cours de cette thèse sont focalisés sur la segmentation du signal de parole. La troisième section de ce chapitre dresse un état de l'art des méthodes de segmentation automatique de la parole. Les choix d'évaluation pour chaque méthode sont justifiés avant de présenter l'impact de la segmentation sur les performances de la tâche finale (diarisation du locuteur).

Enfin, les réseaux de neurones nécessitent de grandes quantités de données annotées pour l'apprentissage. Dans le contexte du traitement automatique de la parole distante à l'aide de plusieurs microphones, peu de données sont actuellement disponibles. Les jeux de données AMI, AISHELL-4 et CHIME-6 sont introduits pour le traitement automatique de la parole distante à l'aide de plusieurs microphones.

2.1 Réseaux de neurones artificiels

Les réseaux de neurones artificiels (RNA) ont d'abord été développés pour modéliser le fonctionnement de neurones biologiques (McCulloch et al. 1943). Ils sont destinés à résoudre une tâche définie (ex : reconnaissance d'images, traduction automatique...) à l'aide d'un algorithme d'apprentissage. L'apprentissage est réalisé sur jeu de données adapté à la tâche envisagée. Dans le contexte de l'apprentissage supervisé, ces données sont accompagnées d'étiquettes servant de cibles d'apprentissage et permettant la mise à jour des paramètres du modèle à via un algorithme d'optimisation. Depuis les premiers travaux de McCulloch et al. (1943), de nombreuses architectures neuronales ont été développées. Bien que toute fonction mathématique puisse être modélisée par un RNA à trois couches (Hornik et al. 1989), des architectures plus profondes et plus complexes ont été proposés dans la littérature en fonction des tâches visées.

Chaque architecture a été conçue pour répondre à une problématique précise. Par exemple, les réseaux de neurones convolutifs (Convolutional Neural Network, CNN) ont été développés dans le cadre de la reconnaissance d'images (LeCun et al. 1989). Les réseaux de neurones récurrents (Recurrent Neural Network, RNN) permettent de modéliser des séquences temporelles en prenant en compte le contexte. Ils s'appliquent ainsi à la transcription automatique ou la traduction de la parole.

Cette section propose une vue d'ensemble des architectures neuronales couramment utilisées pour l'apprentissage automatique. La sous-section 2.1.1 présente le perceptron multicouche ainsi que l'algorithme de rétro-propagation pour optimiser le modèle. La sous-section 2.1.2 présente les CNN suivis des réseaux récurrents en sous-section 2.1.3. Les mécanismes d'attention et les architectures dérivées sont décrits en sous-section 2.1.4.

2.1.1 Perceptron multi-couche

Le perceptron multicouche (Multi-Layer Percreptron, MLP) est une architecture basée sur des neurones artificiels tels que proposés par McCulloch et al. (1943). Comme précisé plus haut, son fonctionnement est inspiré des neurones biologiques.

Réseau de neurones artificiels

Les RNA sont composés de neurones artificiels. Il s'agit d'une opération mathématique appliquant une transformation à des données. Un neurone prend un vecteur $\mathbf{x} = [x_1, \cdots, x_i, \cdots, x_I]$ de I éléments en entrée et lui applique une transformation non-linéaire pour obtenir une sortie y. Cette transformation consiste en une projection linéaire suivie d'une fonction d'activation non-linéaire σ :

$$y = \sigma\Big(\sum_{i=1}^{I} w_i x_i + b\Big), \tag{2.1}$$

avec $\mathbf{w} = [w_1, \cdots, w_i, \cdots, w_I]$ les *poids* du modèle et b un *biais*. Un neurone seul comme décrit dans l'équation (2.1) offre des capacités de modélisation limitées. Les neurones sont donc organisés en réseau (RNA) pour permettre la modélisation de fonctions plus complexes. Un RNA est composé d'une couche d'entrée, d'un ensemble de couches cachées et d'une couche de sortie. En notant h l'indice de la couche courante et x_i^{h-1} la sortie de la couche $h-1$, la sortie x_j^h du neurone j s'écrit :

$$x_j^h = \sigma\Big(\sum_{i=1}^{I_{h-1}} w_{i,j}^h x_i^{h-1} + b_j^h \Big), \tag{2.2}$$

avec $w_{i,j}$ le poids appliqué à la liaison entre les neurones i et j. L'équation (2.2) peut être reformulée sous forme matricielle :

$$\mathbf{x}^h = \sigma(\mathbf{W}^h \mathbf{x}^{h-1} + \mathbf{b}^h). \tag{2.3}$$

Das l'équation (2.3), la matrice $\mathbf{W}$ représente les poids associés à chaque neurone de chaque couche et $\mathbf{b}$ l'ensemble des biais. Les poids et les biais sont les *paramètres* du modèle définis tels que $\boldsymbol{\theta} = \{\mathbf{W}, \mathbf{b}\}$. Ces paramètres sont ensuite optimisés afin de minimiser une fonction de perte. Pour cela, les poids du modèle sont mis à jour à l'aide de l'algorithme de rétro-propagation présenté dans la section suivante.

Algorithme de rétro-propagation

L'algorithme de rétro-propagation permet d'optimiser les poids d'un RNA selon un schéma itératif. Une fonction de perte (ou fonction de coût) $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$ évalue l'erreur entre la prédiction $\hat{\mathbf{y}}$ fournie par le RNA et une vérité de terrain $\mathbf{y}$. L'algorithme de rétro-propagation consiste à minimiser la fonction $\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})$ en optimisant les paramètres $\boldsymbol{\theta}$ du modèle. La méthode de descente du gradient est utilisée pour mettre à jour les paramètres tels que :

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \frac{\partial \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y})}{\partial \boldsymbol{\theta}^t}, \tag{2.4}$$

où η représente le taux d'apprentissage (*learning rate*) et t l'indice de l'itération de l'algorithme d'optimisation. Il existe de nombreuses fonctions de perte en fonction de la tâche visée. Les plus courantes sont présentées dans le tableau 2.1.

Nom	Tâche	Expression
Erreur quadratique moyenne	Régression	$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{N} \sum_{n=1}^{N} (y_n - \hat{y}_n)^2$
Entropie croisée	Classification	$\mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) = -\sum_{n=1}^{N} y_n \log(\hat{y}_n)$
Entropie croisée binaire [1]	Classification	$\mathcal{L}(\hat{y}_n, y_n) = -y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n)$

TABLE 2.1 – Fonctions de perte couramment utilisées pour l'optimisation des réseaux de neurones.

Bien que les RNA soient des estimateurs universels à partir de trois couches (Hornik et al. 1989), leur optimisation peut s'avérer délicate, notamment à cause du sur-apprentissage (Goodfellow et al. 2016). Pour pallier ce problème, de nombreuses architectures ont vu le jour dans la littérature. Celles-ci sont détaillées dans les sous-sections suivantes.

2.1.2 Réseaux convolutifs

Les réseaux de neurones convolutifs (Convolutional Neural Network, CNN) utilisent l'opération de convolution pour transformer les données d'entrée pour une tâche cible. Ces réseaux ont été proposés par LeCun et al. (1989) pour la détection de nombres écrits à la main. Ils sont conçus pour modéliser des données sous forme de grilles telles que des images. Cette section introduit le concept de convolution et présente les avantages des architectures convolutives face aux RNA.

Principe

Les CNN utilisent la convolution pour produire une sortie à partir des données d'entrées. L'opération de convolution est réalisée entre les données $x(n)$ et un noyau $k(n)$. Les données considérées étant discrètes, la convolution à une dimension (1-D) est définie par la relation suivante :

$$y(n) = (x \star k)(n) = \sum_{i=0}^{L-1} x(n-i)k(i) \tag{2.5}$$

avec n l'indice temporel et L la dimension du noyau.

La convolution s'exprime également en 2-D (ex : images) par la relation suivante :

$$y(m,n) = (x \star k)(m,n) = \sum_{i=0}^{L-1}\sum_{j=0}^{H-1} x(n-i,m-j)k(i,j) \tag{2.6}$$

avec n et m les indices associés à chaque dimension des représentations x et y, et L et H les dimensions du noyau dans chaque dimension. Le principe de la convolution 2-D est présenté en figure 2.1.

La taille du noyau k conditionne les dimensions de la représentation de sortie. La taille du noyau étant indépendante de l'entrée, les CNN peuvent traiter des données de dimensions variables. Le *zero padding* peut être utilisé afin de garantir une taille de sortie identique à l'entrée. Il consiste à ajouter des zéros autour de l'image. La convolution peut également permettre de réduire les dimensions des données à l'aide du *stride*. Cette opération consiste à augmenter le pas s du noyau entre deux opérations de convolution successives. En choisissant $s > 1$, la longueur de la sortie y est divisée par s.

Les CNN consistent à optimiser les valeurs du noyau k à l'aide de l'algorithme de rétro-propagation. Ainsi, la représentation y est optimisée selon la fonction de perte choisie. Les

1. Les valeurs de $\mathcal{L}(y_n, \hat{y}_n)$ associées aux éléments y_n et $\hat{y}_n$ sont ensuite sommées pour tout $n \in [1, N]$.

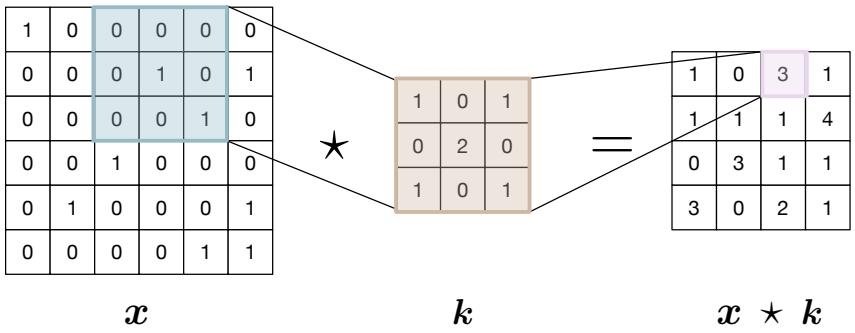


FIGURE 2.1 – Principe de la convolution à deux dimensions.

couches des CNN sont souvent composées de plusieurs noyaux, appelés *canaux* comme illustré en figure 2.2. Chaque noyau est optimisé indépendamment et permet d'obtenir une représentation spécifique des données d'entrée x. En extrayant plusieurs représentations de l'entrée, les CNN sont capables d'extraire différentes caractéristiques.

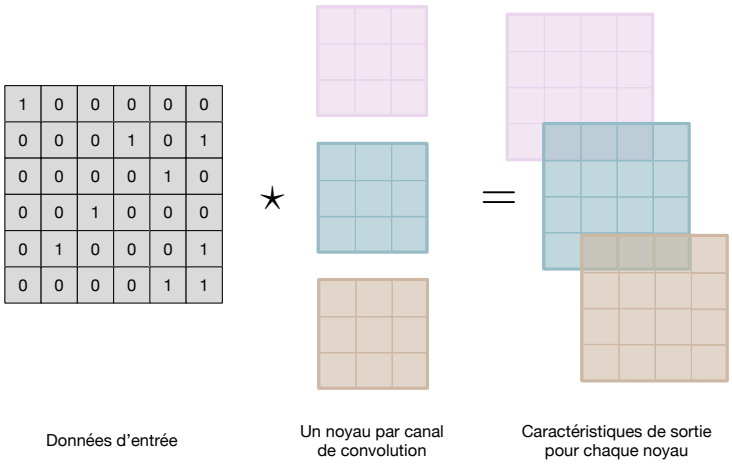


FIGURE 2.2 – Principe de la convolution à deux dimensions avec plusieurs canaux. Chaque noyau est convolué aux données d'entrée et produit une sortie spécifique. Les noyaux sont optimisés individuellement.

Propriétés

Les CNN exploitent trois propriétés de la convolution motivant leur utilisation et permettant d'améliorer les performances des modèles d'apprentissage automatique : la parcimonie, le partage des paramètres et les représentations équivariantes.

Les RNA traditionnels mettent en relation chaque élément des données d'entrée avec tous les éléments de la représentation de sortie par un produit matriciel. Chaque élément de sortie interagit donc avec toutes les données d'entrée. Les CNN permettent une relation parcimonieuse entre les éléments de sortie et les données d'entrée. Cela est possible par l'utilisation d'un noyau de dimension inférieure à celle de l'entrée, réduisant ainsi le nombre de connexions entre les

entrées et les sorties. La parcimonie permet de réduire le nombre de paramètres optimisés, ce qui facilite la convergence du modèle et limite le sur-apprentissage (Goodfellow et al. 2016).

Le partage des paramètres consiste à apprendre un jeu de paramètres unique pour un noyau. Cela permet de réduire l'espace mémoire occupé par les paramètres du modèle. De plus, le fait d'apprendre un jeu de paramètres unique mène à la troisième propriété des CNN : l'équivariance en translation.

La convolution est une opération équivariante en translation (Goodfellow et al. 2016). Cela implique qu'un décalage de l'entrée x entraîne un décalage similaire de la sortie y. La combinaison de la convolution et d'une opération de *pooling* rend ces systèmes invariants en translation. Cette propriété explique les performances remarquables des CNN en reconnaissance d'images (Shuying Liu et al. 2015). Quelle que soit la position du sujet à détecter dans l'image, la représentation obtenue en sortie sera similaire, à une translation près.

Dilatation

La *dilatation* a été proposée dans le contexte des CNN par (Fisher Yu et al. 2015). Cette opération consiste à dilater le noyau de convolution d'un facteur $d \in \mathbb{N}$. La convolution est ainsi réalisée sur une version décimée des données d'entrée. La convolution dilatée 1-D s'exprime par la relation suivante :

$$(x \star k)(n) = \sum_{i=0}^{L-1} x(n - di)k(i). \tag{2.7}$$

Pour $d = 1$ dans l'équation (2.7), l'opération est équivalente à la convolution classique (Eq. (2.5)). Pour $d > 1$, la convolution s'applique à une version décimée d'un facteur d de données d'entrée x. Cette considération permet d'augmenter le champ de réception du modèle sans augmenter le nombre de paramètres optimisés. Le champ de réception d'un modèle peut augmenter de façon exponentielle en doublant le facteur d entre deux couches successives sans perte de résolution.

Connexions résiduelles

Les architectures à connexions résiduelles sont introduites par K. He et al. (2016) pour pallier deux problèmes inhérents aux CNN profonds. La première est l'apparition du phénomène de *disparition du gradient*. Les valeurs du gradient (Eq. (2.4)) tendent vers zéro, rendant difficile l'optimisation des paramètres du modèle. La seconde est la perte de précision des modèles de classification lorsque les architectures sont très profondes.

K. He et al. (2016) font l'hypothèse que ces deux problèmes sont liés à un défaut de circulation de l'information au sein du réseau. Ils proposent de conserver la représentation d'entrée d'un groupe de couches convolutives en l'ajoutant à sa sortie. En reprenant les notations de l'équation (2.5), une connexion résiduelle s'exprime :

$$y = f(x) + x, \tag{2.8}$$

avec $f(\cdot)$ la fonction réalisée par un groupe de couches convolutives. Le principe de la connexion résiduelle est également illustré en figure 2.3. Une couche de convolution supplémentaire peut être ajoutée pour garantir des dimensions similaires entre l'entrée $x(n)$ et la sortie $y(n)$.

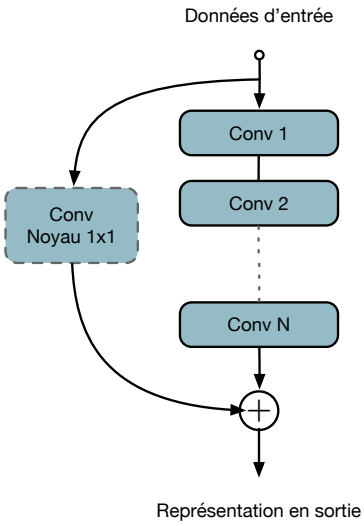


FIGURE 2.3 – Vue schématique d'une connexion résiduelle. Les données d'entrée sont ajoutées à la représentation de sortie. Une couche intermédiaire peut être ajoutée afin de conserver les mêmes dimensions lors de l'addition.

2.1.3 Réseaux récurrents

La section précédente présente les réseaux de neurones convolutifs. Ces architectures permettent d'obtenir des représentations locales précises des données d'entrée, rendant les CNN performants sur certaines tâches (ex : reconnaissance d'images). Cependant, ces architectures ne sont pas conçues pour traiter des séries temporelles. En effet, les CNN n'ont pas d'effet mémoire et ne permettent pas de conserver une information d'un instant à un autre. C'est dans ce contexte que les réseaux de neurones récurrents (Recurrent Neural Network, RNN) ont été développés.

Principe

Les RNN consistent à représenter une série temporelle $\mathbf{x} \in \mathbb{R}^{F \times T}$, avec T le nombre de trames et F le nombre d'éléments par trame, par une séquence $\mathbf{y} \in \mathbb{R}^{E \times T}$ où E représente le nombre d'éléments dans la séquence de sortie. La trame de sortie $\mathbf{y}^t$ à un instant t contient une partie de l'information de l'entrée $\mathbf{x}^t$ et une partie de celle des instants passés $\mathbf{x}^{t-i}$ avec $i \in [1, t]$. La conservation d'information passée est permise par un état caché, noté $\mathbf{h}^t$. L'état caché joue le rôle de mémoire en conservant une partie de l'information passée pour la transmettre aux instants suivants. Le principe des RNN est schématisé en figure 2.4.

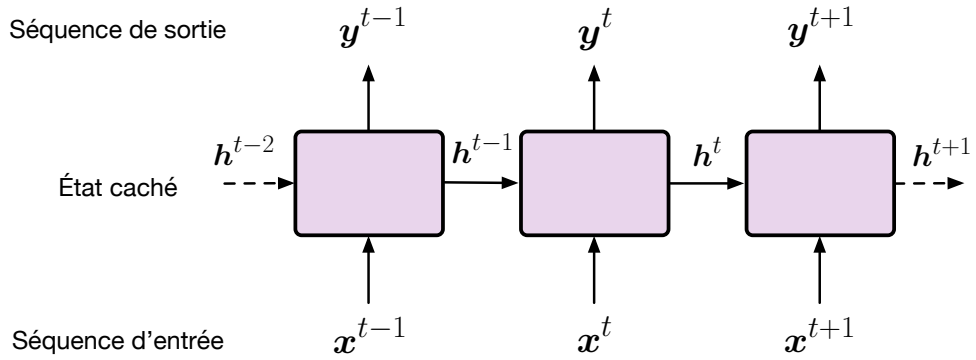


FIGURE 2.4 – Principe de fonctionnement d'un réseau récurrent.

Les réseaux récurrents classiques présentent cependant plusieurs problèmes (GOODFELLOW et al. 2016). Premièrement, l'état caché conserve difficilement l'information à long terme. Deuxièmement, ces architectures présentent des instabilités lors de l'apprentissage telles que l'explosion du gradient ou, à l'inverse, la disparition du gradient. Pour répondre à ces problématiques, plusieurs architectures récurrentes ont été proposées dans la littérature. Elles sont présentées ci-après.

Long Short-Term Memory

Les réseaux récurrents de type *Long Short-Term Memory* (LSTM) sont introduits par HOCHREITER et al. (1997) pour contourner les problèmes de gradient des RNN. Les LSTM utilisent un mécanisme de portes (*gates*) permettant de sélectionner automatiquement l'information retenue et l'information oubliée d'une trame $t-1$ à une trame t. Le principe d'une cellule LSTM est schématisé en figure 2.5.

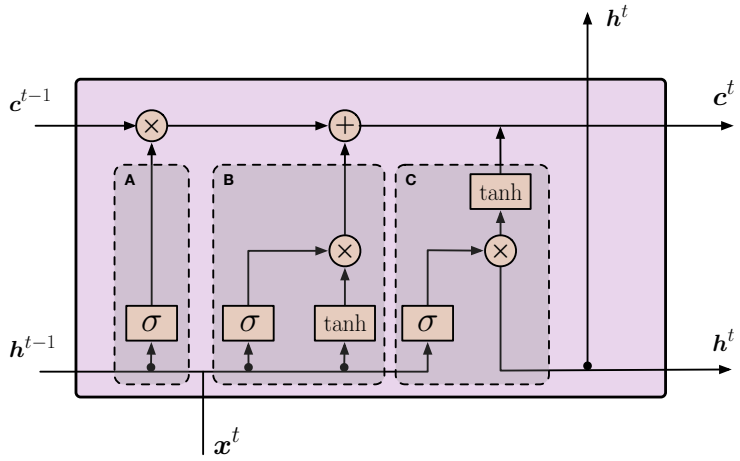


FIGURE 2.5 – Schéma d'une cellule récurrente de type LSTM avec **A** la porte d'oubli, **B** la porte d'entrée et **C** la porte de sortie. Les activations σ et tanh représentent respectivement la sigmoïde et la tangente hyperbolique.

Une cellule LSTM calcule l'état caché $\mathbf{h}^t$ et l'état de la cellule $\mathbf{c}^t$ à la trame t courante. Pour cela, un ensemble d'opérations est appliqué aux états de la trame passée $t-1$, notés $\mathbf{h}^{t-1}$ et $\mathbf{c}^{t-1}$.

La première étape consiste à calculer les sorties des portes d'oubli $\mathbf{f}^t$, d'entrée $\mathbf{i}^t$ et de sortie $\mathbf{o}^t$. Chaque porte possède deux jeux de poids $\mathbf{W}_\alpha$ appliqué à l'entrée $\mathbf{x}^t$ et $\mathbf{U}_\alpha$ appliqué à l'état caché passé $\mathbf{h}^{t-1}$. Un biais $\mathbf{b}_\alpha$ est également ajouté avec $\alpha = \{f, i, o\}$ en fonction de la porte considérée. La sortie de chaque porte s'exprime :

$$\mathbf{f}^t = \sigma(\mathbf{W}_f \mathbf{x}_t + \mathbf{U}_f \mathbf{h}_{t-1} + \mathbf{b}_f), \tag{2.9}$$

$$\mathbf{i}^t = \sigma(\mathbf{W}_i \mathbf{x}^t + \mathbf{U}_i \mathbf{h}^{t-1} + \mathbf{b}_i), \tag{2.10}$$

$$\mathbf{o}^t = \sigma(\mathbf{W}_o \mathbf{x}^t + \mathbf{U}_o \mathbf{h}^{t-1} + \mathbf{b}_o). \tag{2.11}$$

Un état intermédiaire $\tilde{\mathbf{c}}^t$ est également calculé avant d'appliquer les différentes portes :

$$\tilde{\mathbf{c}}^t = \sigma(\mathbf{W}_c \mathbf{x}^t + \mathbf{U}_c \mathbf{h}^{t-1} + \mathbf{b}_c). \tag{2.12}$$

Cet état est calculé à l'aide d'un autre jeu de paramètres $\mathbf{W}_c$, $\mathbf{U}_c$ et $\mathbf{b}_c$. Les états du LSTM à la trame t sont calculés par les relations suivantes :

$$\mathbf{c}^t = \mathbf{f}^t \odot \mathbf{c}^{t-1} + \mathbf{i}^t \odot \tilde{\mathbf{c}}^t, \tag{2.13}$$

$$\mathbf{h}^t = \mathbf{o}^t \odot \sigma(\mathbf{c}^t). \tag{2.14}$$

Les LSTM permettent de limiter les problèmes liés à l'explosion ou l'évanescence du gradient. Cependant, ils restent limités pour l'apprentissage d'information sur le long-terme (Bengio et al. 1994).

Gated Recurrent Units

Les LSTM présentent une architecture complexe permettant de sélectionner l'information conservée et oubliée d'une trame à une autre. Un autre type de cellule récurrente est introduit par Cho et al. (2014) afin de simplifier l'architecture LSTM. Cette architecture, nommée *Gated Recurrent Unit* (GRU), consiste à réduire le nombre de portes au sein de la cellule. Bien que plus récentes, les cellules GRU n'apportent par nécessairement un gain de performance (Goodfellow et al. 2016). Celles-ci n'étant pas utilisées dans le cadre de cette thèse, les détails de fonctionnement des GRU ne sont pas présentés ici.

Réseaux récurrents bi-directionnels

Les RNN modélisent un échantillon $\mathbf{y}^t$ de la séquence temporelle de sortie à partir de l'échantillon courant $\mathbf{x}^t$ des échantillons passés $\mathbf{x}^{t-i}$. Les réseaux récurrents sont donc des systèmes causaux étant donné qu'une sortie $\mathbf{y}^t$ ne dépend que des éléments passés et courant Dans certaines conditions, la causalité n'est pas requise. Dans ce cas, les réseaux récurrents bi-directionnels (BRNN) peuvent être utilisés (Schuster et al. 1997). Le principe des BRNN est présenté en figure 2.6. Ils consistent à utiliser deux RNN en parallèle. Le premier modélise la série temporelle entrante dans le sens des t croissants (système causal). Le second modélise la série temporelle entrante dans le sens des t décroissants. Deux séquences d'états cachés sont donc obtenues. En pratique, ces deux représentations sont concaténées. La dimension de l'état caché $\mathbf{h}^t$ est donc doublée par rapport au RNN causal.

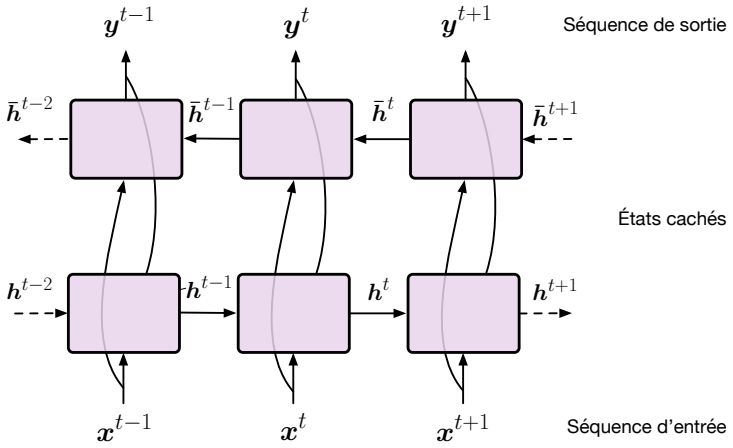


Figure 2.6 – Principe d'un réseau récurrent bi-directionnel. La séquence d'entrée $\mathbf{x}$ est modélisée par deux RNN observant la séquence selon les temps croissant et décroissant respectivement. Deux jeux d'états cachés, $\mathbf{h}$ et $\bar{\mathbf{h}}$, sont donc obtenus.

Réseau convolutif temporel

Les réseaux convolutifs temporels (Temporal Convolutional Network, TCN) ont été proposés par (Bai et al. 2018). Cette architecture consiste à modéliser une séquence temporelle à l'aide de couches convolutives 1-D. Le TCN exploite la dilatation de la convolution pour prendre en compte un large contexte temporel. En pratique, une trame t de la séquence de sortie $\mathbf{y}$ contient de l'information sur toutes les trames passées de la séquence d'entrée $\mathbf{x}$. Cette architecture a l'avantage d'être causale, la trame de sortie t dépend uniquement des trames passées de la séquence d'entrée.

2.1.4 Mécanismes d'attention

Cette section décrit brièvement le principe d'attention introduit dans les réseaux de neurones. L'attention est introduite par Bahdanau et al. (2014) puis étendue par Luong et al. (2015).

L'attention permet à un réseau de neurones de sélectionner l'information utile dans la séquence d'entrée pour générer une sortie. Elle est d'abord développée dans le contexte de la traduction automatique pour permettre la prise en compte de longues séquences. Le concept d'attention est également un élément clef du Transformer (Vaswani et al. 2017), composant essentiel des modèles pré-entraînés tels que BERT (Devlin et al. 2018) et GPT (Bronstein et al. 2021) pour le traitement du langage naturel ou WavLM (S. Chen et al. 2022) pour le traitement de la parole.

Architecture encodeur-décodeur et attention

Les architectures encodeur-décodeur permettent de modéliser une série temporelle en passant par une représentation intermédiaire de dimension fixe appelée *vecteur de contexte*. Un premier réseau de neurones (encodeur) projette la séquence d'entrée $\mathbf{x} \in \mathbb{R}^{F \times T_e}$ vers un vecteur de contexte de taille C fixe $\mathbf{c} \in \mathbb{R}^{C}$. Un second réseau (décodeur) génère la séquence de sortie $\mathbf{x} \in \mathbb{R}^{F \times T_d}$ de façon récursive à partir du vecteur de contexte. La longueur de la séquence d'entrée T_e peut être différente de celle de la sortie T_d.

Les architectures encodeur-décodeur font l'hypothèse que la séquence d'entrée peut être modélisée par un vecteur de taille fixe $\mathbf{c}$. Cependant, la compression de l'information devient difficile pour les longues séquences (Bahdanau et al. 2014). Les mécanismes d'attention tentent de pallier ce problème en sélectionnant l'information utile au sein de la séquence encodée $\mathbf{x}$ pour générer chaque élément de sortie $\mathbf{y}_i$. La trame de sortie $\mathbf{y}_i$ est calculée à partir d'un vecteur de contexte $\mathbf{c}_i$ mis à jour à chaque trame :

$$\mathbf{c}_i = \sum_{j=0}^{N-1} \alpha_{ij} h_j, \tag{2.15}$$

avec h_j la sortie de l'encodeur associée à la trame d'entrée $\mathbf{x}_j$. Les poids α_{ij} sont obtenus à l'aide d'un modèle d'alignement (réseau de neurones). Celui-ci prend en entrée l'état caché passé du décodeur $\mathbf{s}_{i-1}$ et la séquence encodée h_j. Un score est ensuite calculé sur la sortie du modèle d'alignement pour obtenir des poids normalisés (Bahdanau et al. 2014). La figure 2.7 schématise le principe d'attention au sein d'un modèle encodeur-décodeur.

D'autres approches de calcul des poids d'attention α_{ij} ont ensuite été proposées. Par exemple, Luong et al. (2015) proposent le concept d'attention locale et globale en entraînant deux réseaux de neurones distincts pour générer les poids d'attention.

Transformer et mécanisme d'auto-attention

Les encodeur-décodeur attentifs utilisent des RNN pour l'encodage et le décodage des séquences d'entrée et de sortie. L'aspect récursif de ces réseaux rend leur parallélisation difficile. L'apprentissage de modèles encodeur-décodeur récurrents requiert alors un temps conséquent.

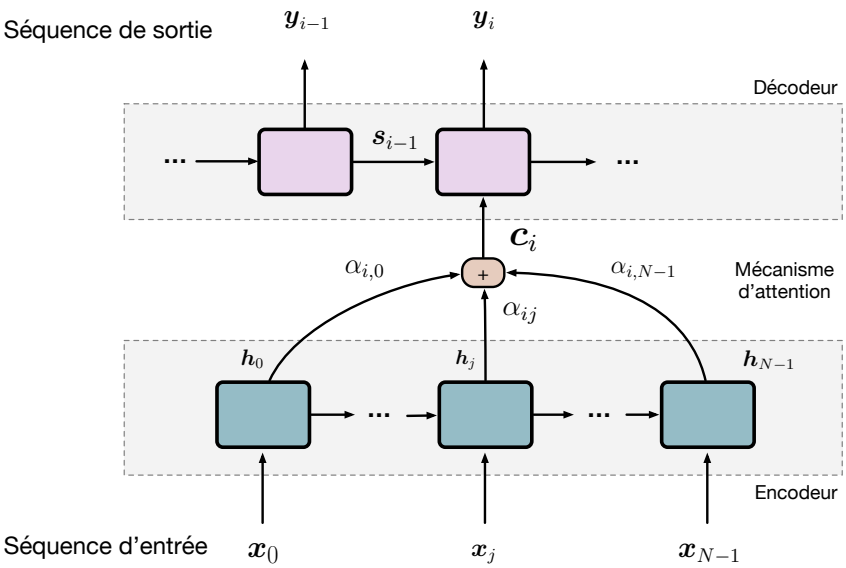


FIGURE 2.7 – Architecture encodeur-décodeur avec mécanisme d'attention. L'élément i de la séquence de sortie $\mathbf{y}$ est généré à partir de l'état caché de sortie passé $\mathbf{s}_{i-1}$ et du vecteur de contexte $\mathbf{c}_i$. Ce dernier est calculé par la somme pondérée de tous les éléments de la séquence encodée $\mathbf{h}$.

Le *Transformer* est une architecture neuronale basée uniquement sur les mécanismes d'attention (VASWANI et al. 2017). En comparaison aux architectures encodeur-décodeur précédemment présentées, il permet une meilleure parallélisation des opérations tout en améliorant les performances et réduisant le temps d'apprentissage.

Le mécanisme d'auto-attention est une opération clef du transformer. L'attention, telle que formulée par BAHDANAU et al. (2014), calcule les poids à partir de la séquence d'entrée et de l'information sur l'élément précédent de la séquence de sortie. À l'inverse, l'auto-attention calcule les poids uniquement à partir de la séquence d'entrée. Pour cela, trois représentations d'une trame i de la séquence d'entrée $\mathbf{x} = [\mathbf{x}_1, \cdots, \mathbf{x_i}, \cdots, \mathbf{x_T}]$ sont calculées : la *clef* $\mathbf{K}_i$, la *requête* $\mathbf{Q}_i$ et la *valeur* $\mathbf{V}_i$:

$$\begin{aligned} \mathbf{Q}_i &= \mathbf{x}_i^T \mathbf{W}_Q, \\ \mathbf{K}_i &= \mathbf{x}_i^T \mathbf{W}_K, \\ \mathbf{V}_i &= \mathbf{x}_i^T \mathbf{W}_V. \end{aligned} \tag{2.16}$$

Dans l'équation (2.16), les matrices $\mathbf{W}_Q$, $\mathbf{W}_K$ et $\mathbf{W}_V$ représentent les poids de trois MLP permettant de projeter la séquence vers chaque représentation. Les poids d'attention sont obtenus à l'aide d'un score de similarité calculé entre la clef et la requête :

$$\mathbf{s} = \text{softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_j^T}{\sqrt{D}}\right), \tag{2.17}$$

avec D la dimension de l'espace dans lequel sont projetées la clef et la requête.

La séquence de sortie du module d'auto-attention $\mathbf{z}_i$ est obtenue par le produit matriciel des poids d'attention et de la valeur $\mathbf{V}_j$:

$$\mathbf{z}_i = \sum_{j=1}^{T} s_{ij} \mathbf{V}_j. \tag{2.18}$$

Pour chaque élément d'entrée $\mathbf{x}_i$, un nouveau vecteur $\mathbf{z}_i$ est obtenu. Ce dernier prend en compte l'information courante ainsi que celle des éléments passés et futurs de la séquence $\mathbf{x}$.

Attention multi-tête

Les mécanismes d'auto-attention permettent de sélectionner l'information utile au sein d'une séquence d'entrée. Vaswani et al. (2017) ont étendu ce concept avec l'auto-attention multi-tête. Cette approche consiste à utiliser plusieurs clefs, requêtes et valeurs afin de calculer plusieurs représentations de la séquence d'entrée en parallèle. Chaque tête possède un trio de poids $\mathbf{W}_Q$, $\mathbf{W}_K$ et $\mathbf{W}_V$. En notant H le nombre de têtes, une représentation $\bar{\mathbf{z}}_{ih}$ de l'élément d'entrée $\mathbf{x}_i$ est obtenue pour chaque tête avec $h \in [1, H]$. Les représentations sont ensuite concaténées et projetées vers un nouvel espace à l'aide d'une autre matrice de poids $\mathbf{W}_O$ telle que :

$$\mathbf{z}_i = \mathrm{cat}(\bar{\mathbf{z}}_{i,1}, \ldots, \bar{\mathbf{z}}_{i,H})\mathbf{W}_O. \tag{2.19}$$

L'attention multi-tête permet plus de flexibilité dans la représentation de la séquence d'entrée sans augmenter le nombre de paramètres.

2.2 Représentation du signal de parole

Les réseaux de neurones artificiels sont de puissants outils pour la modélisation des données. Ils permettent de résoudre des tâches complexes avec une grande précision. Dans le contexte du traitement automatique de la parole, les données considérées sont des signaux audio. Cependant, il peut être difficile pour une architecture neuronale d'utiliser les données brutes en entrée. Pour faciliter la convergence des systèmes de traitement automatique de la parole, des caractéristiques sont extraites à partir du signal. Ils permettent d'obtenir une représentation du signal contenant des informations utiles pour la tâche visée.

Il existe de nombreuses approches d'extraction de caractéristiques pour le traitement automatique de la parole. Deux catégories peuvent être discriminées : les caractéristiques acoustiques et les caractéristiques neuronales. Les premières sont obtenues à l'aide de méthodes de traitement du signal classiques. Les secondes sont extraites par des réseaux de neurones. Ces modèles neuronaux peuvent être entraînés simultanément avec le système effectuant la tâche de traitement de la parole. Il s'agit alors de systèmes bout-en-bout (*end-to-end*). Plus récemment, les caractéristiques pré-entraînées ont permis de produire des représentations des signaux de parole (Baevski et al. 2020 ; S. Chen et al. 2022 ; Schneider et al. 2019) complexes et optimisées.

2.2.1 Caractéristiques acoustiques

Les caractéristiques *acoustiques* sont extraites du signal à l'aide de méthode de traitement des signaux. Elles sont ensuite fournies au réseau de neurones résolvant la tâche en aval.

Transformée de Fourier à court terme

La transformée de Fourier (TF) permet de représenter un signal temporel sur une base de fonctions harmoniques. Elle donne ainsi accès au contenu fréquentiel du signal. Soit $x(n)$ un signal temporel discret de N échantillons où $n = \{0, \ldots, N-1\}$ représente l'indice temporel d'un échantillon. La transformée de Fourier discrète du signal s'exprime :

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-2\pi jkn/N}, \tag{2.20}$$

avec $k = \{0, \ldots, K\}$ l'indice fréquentiel, K le nombre de fréquences représentées et $j = \sqrt{-1}$.

Cependant, la transformée de Fourier ne permet pas de conserver l'information temporelle du signal. Les signaux de parole n'étant pas stationnaires, il est nécessaire de connaître l'évolution du contenu fréquentiel en fonction du temps. La transformée de Fourier à court terme (TFCT) est donc privilégiée. Elle consiste à calculer la TF sur une fenêtre glissante $w(n)$. Ainsi, une représentation fréquentielle est obtenue pour chaque position m de la fenêtre d'analyse. La TFCT s'exprime :

$$X(m,k) = \sum_{n=0}^{N-1} x(n) w(n-m) e^{-2\pi jkn/N}. \tag{2.21}$$

Le *spectrogramme* est couramment utilisé pour représenter les signaux de parole à partir de la TFCT :

$$S(m,k) = |X(m,k)|^2, \tag{2.22}$$

avec $|.|$ le module.

La TFCT est une représentation d'un signal temporel dans le domaine temps-fréquence. Cette représentation équivaut à représenter le signal dans des bandes de fréquences linéairement espacées. La parole étant principalement localisée dans les bandes de fréquence basses (600-6000 Hz), d'autres représentations ont été proposées pour permettre une meilleure représentation de ce type de signal.

Spectrogramme à échelle Mel

Le Mel est une unité représentant la hauteur d'un son. Sa définition est basée sur la perception de la fréquence par les humains (Stevens et al. 1937). Elle a été déterminée à la suite d'expériences

psychoacoustiques pour être linéaire d'un point de vue perceptif. La conversion des fréquences f en Hertz vers les $mels$ en Mel est donnée par la relation suivante :

$$mels = 2595 \log_{10}\Big(1 + \frac{f}{700}\Big). \tag{2.23}$$

Le spectrogramme à échelle Mel est une représentation temps-fréquence du signal obtenue à partir d'un spectrogramme. La conversion d'un spectrogramme en Hertz est réalisée à l'aide d'un banc de filtres triangulaires dont les fréquences centrales suivent l'échelle Mel.

Coefficients cepstraux à échelle Mel

Les coefficients cepstraux à échelle Mel (Mel Frequency Cepstral Coefficient, MFCC) ont été proposés comme une alternative au spectrogramme à échelle Mel. Les MFCC sont obtenus en appliquant une transformée en cosinus discrète (TCD) au spectrogramme à échelle Mel. Il est également possible d'extraire des paramètres dynamiques des MFCC. Ces derniers sont obtenus en estimant la dérivée (Δ) et la dérivée seconde ($\Delta\Delta$) des coefficients.

2.2.2 Caractéristiques adaptées

Les caractéristiques acoustiques telles que les MFCC sont extraites du signal à l'aide d'une chaîne de traitement fixe. L'extraction de caractéristiques ne peut donc pas s'adapter au modèle de traitement de la parole placé en aval. Afin d'optimiser l'extraction de caractéristiques pour la tâche visée, celle-ci peut être adaptée au cours de l'apprentissage.

Banc de filtres optimisés

Une première approche consiste à entraîner un réseau convolutif 1-D (Eq. (2.5)) à extraire des caractéristiques pour la tâche aval. Cette approche a notamment été utilisée dans le contexte de la séparation de la parole (Luo et al. 2019) et la reconnaissance de la parole (Sainath et al. 2017). Cependant, l'apprentissage de filtres initialisés aléatoirement et optimisés à partir du signal temporel s'avère délicat.

Pour régulariser l'apprentissage, Pariente et al. (2020) proposent de combiner l'apprentissage de filtres convolutifs avec le formalisme des filtres analytiques (Flanagan 1980). En notant $\mathbf{h}(n)$ un filtre à réponse impulsionnelle finie (RIF), le filtre analytique $\mathbf{h}_a(n)$ associé est obtenu via la transformée de Hilbert $\mathcal{H}$:

$$\mathbf{h}_a(n) = \mathbf{h}(n) + j\mathcal{H}\Big(\mathbf{h}(n)\Big). \tag{2.24}$$

Banc de filtres paramétrés

Les bancs de filtres paramétrés sont une alternative aux filtres adaptés. Ils consistent à apprendre un jeu de paramètres restreint suffisant pour définir un filtre. Ravanelli et al. (2018)

proposent SincNet, un banc de filtres RIF paramétrés par la largeur de bande passante et la fréquence centrale des filtres. Seuls ces paramètres sont optimisés lors de la rétro-propagation. Le nombre de paramètres à optimiser est ainsi réduit. Les filtres SincNet ont été largement utilisés dans la littérature (Bullock et al. 2020) sur diverses tâches de traitement automatique de la parole.

E. Bavu et al. (2019) introduisent TimeScaleNet. Il s'agit d'un banc de filtres à réponse impulsionnelle infinie biquadratiques dont les coefficients sont optimisés avec la tâche aval. Les coefficients du filtre sont directement optimisés, contrairement au modèle SincNet. De plus, le nombre de paramètres est réduit, les filtres biquadratiques nécessitant seulement quatre coefficients pour être définis. Cela permet de réduire le nombre de paramètres à optimiser tout en favorisant l'interprétabilité du modèle. Ce type de filtres présente cependant des contraintes de stabilité.

2.2.3 Caractéristiques pré-entraînées

Les modèles pré-entraînés consistent à apprendre une représentation optimale d'une modalité d'entrée. Ce concept a d'abord été utilisé avec des modèles de langage tels que BERT (Devlin et al. 2018) ou GPT (Brown et al. 2020). Ces modèles apprennent une représentation du langage à l'aide de l'apprentissage auto-supervisé. En pratique, il s'agit d'architectures Transformer (Vaswani et al. 2017) prenant en entrée du texte dont certains mots ont été masqués. Le modèle est entraîné à reconstruire les mots masqués. Cela permet d'obtenir une représentation optimisée des mots dans une phrase à partir du contexte.

Les modèles pré-entraînés ont également été développés pour le traitement automatique de la parole. Baevski et al. (2020) introduisent wav2vec 2.0. Il s'agit de la première approche appliquant l'apprentissage auto-supervisé à des données audio. La phase d'apprentissage consiste à reconstruire les segments d'un signal de parole préalablement masqués. La modèle apprend ainsi une représentation optimisée du signal en prenant en compte un large contexte temporel. La reconstruction optimisée à l'aide d'une fonction de coût contrastive et 53 000 h de signaux de parole issus de divers jeux de données de la littérature sont utilisées pour l'apprentissage.

Suite aux performances de wav2vec 2.0, S. Chen et al. (2022) proposent Wavlm. L'architecture est proche du modèle précédent et l'apprentissage est réalisé sur 96 000 h de parole. Les auteurs ajoutent une phase d'augmentation de données en créant de la parole superposée artificielle. Cela permet au modèle de représenter ce type de d'évènement. La méthode d'apprentissage est également différente. Le signal de parole masqué n'est pas reconstruit comme le propose wav2vec 2.0. L'optimisation du modèle est formulée comme une classification de phonèmes à l'aide d'une fonction de perte similaire à BERT (Devlin et al. 2018).

Les représentations pré-entraînées encodent un large contexte temporel et une représentation optimisée du signal. Elles peuvent être utilisées comme caractéristiques d'entrée des systèmes de traitement automatique de la parole. Il est cependant nécessaire de rappeler que ces systèmes ont une grande complexité algorithmique et que les représentétions ainsi obtenues sont difficiles

à interpréter.

2.3 Segmentation automatique de la parole monocanale pour la diarisation en locuteurs

La segmentation automatique de la parole permet d'extraire des segments homogènes pour la diarisation (Garcia-Perera et al. 2020 ; T. J. Park et al. 2022). Elle se décompose couramment en trois sous-tâches : la détection d'activité vocale (VAD), la détection de parole superposée (OSD) et la détection de changements de locuteur (SCD). Ces tâches permettent respectivement d'identifier des segments du signal contenant de la parole, de la parole superposée et les frontières entre les tours de parole. L'intérêt pour ces tâches est présent depuis plusieurs décennies dans le contexte du traitement automatique de la parole. Aujourd'hui, les approches pour les résoudre convergent vers des méthodes similaires (Bredin et al. 2020). Cette section propose un état de l'art des méthodes de segmentation de la parole pour la diarisation. La section 2.3.1 présente les méthodes de détection d'activité vocale. La section 2.3.2 introduit les approches de détection de parole. Les méthodes de détection de changements de locuteur sont présentées en section 2.3.3 avant que ne soit abordé l'impact de la segmentation sur la tâche de diarisation en section 2.3.4.

2.3.1 Détection d'activité vocale

La détection d'activité vocale (Voice Activity Detection, VAD), consiste à détecter les segments contenant de la parole dans un signal audio. Il s'agit couramment de la première étape de segmentation dans les systèmes de diarisation (T. J. Park et al. 2022). Les premières approches proposées sont basées sur des méthodes de traitement du signal et des modèles statistiques. Les algorithmes plus récents exploitent les capacités de modélisation des réseaux de neurones et l'apprentissage supervisé.

Traitement du signal et modèles statistiques

Les premières approches de VAD utilisent l'énergie du signal pour détecter les échantillons de parole. Chengalvarayan (1999) proposent une méthode de normalisation du signal à l'aide de l'énergie. Un algorithme à base de règles permet de discriminer les échantillons de parole des silences. Cette méthode prend notamment en compte la moyenne des MFCC dans le calcul pour faciliter cette discrimination. Woo et al. (2000) introduisent une méthode de modélisation du bruit pour faciliter la détection. D'autres approches utilisent les fonctions d'auto-corrélation afin de déterminer les échantillons contenant de la parole. Par exemple, Kristjansson et al. (2005) introduisent des caractéristiques basées sur l'auto-corrélation des segments de parole. D'autres caractéristiques du signal telles que le taux de passage par zéro sont également utilisées (Ghaemmaghami et al. 2010). Le codage prédicatif linéaire (Nemer et al. 2001) a aussi été exploré pour cette tâche.

Un autre pan de la littérature propose l'utilisation de modèles statistiques pour la détection de segments de parole. Sohn et al. (1999) proposent un modèle de décision dans le domaine temps-fréquence permettant de détecter la présence de parole au sein de chaque trame du signal. Une chaîne de Markov est utilisée pour prendre en compte la temporalité de la séquence. La plupart des approches statistiques utilisent ensuite les modèles de Markov (HMM) et les mélanges de Gaussiennes (GMM) (Ng et al. 2012 ; Pfau et al. 2001). La généralisation de l'utilisation des réseaux de neurones a permis de développement de nouvelles méthodes de VAD bien plus performantes.

Approches neuronales

Les méthodes de VAD basées sur le signal requièrent de poser des hypothèses sur le signal afin de définir des règles de décision. Les réseaux de neurones permettent d'optimiser un modèle pour une tâche cible uniquement à partir des données. L'utilisation de ce type de système s'est généralisée pour la VAD. Ryant et al. (2013) proposent l'utilisation d'un perceptron multicouches pour la détection de segments de parole. Le modèle prend en entrée des MFCC extraits sur des trames de 25 ms. Un MLP permet ensuite de prédire la présence de parole à chaque trame. Les auteurs montrent que leur système surpasse un modèle HMM-GMM de l'état de l'art. Hughes et al. (2013) introduisent l'utilisation de réseaux récurrents pour la VAD. L'architecture est composée de deux couches récurrentes prenant en entrée des caractéristiques acoustiques (PLP [2]) et détectant la présence de parole à la trame. Les auteurs montrent également que le réseau récurrent permet d'améliorer les performances de détection par rapport à un modèle GMM. Thomas et al. (2014) ont ensuite exploré l'utilisation des CNN pour détecter la parole. Le système proposé prend un spectrogramme à échelle Mel en entrée et prédit la présence de parole à la trame. Les auteurs montrent un gain de performance par rapport aux modèles MLP, notamment en conditions acoustiques difficiles. Les premières méthodes de détection de parole à l'aide des réseaux de neurones se basent sur un formalisme similaire. Ces systèmes prennent une séquence de caractéristiques acoustiques puis le modèle est optimisé pour classer les trames. Cette approche a été conservée dans les méthodes proposant les performances à l'état de l'art.

Les RNN sont majoritairement utilisés dans les approches courantes de VAD. Ils permettent de prédire la présence de parole en prenant en compte le contexte temporel (mémoire des trames passées). Gelly et al. (2017) proposent l'utilisation de LSTM pour détecter les segments de parole à partir de caractéristiques acoustiques. Les auteurs comparent différentes fonctions de pertes et différents contextes d'optimisation. Ils montrent que l'utilisation de LSTM permet une meilleure détection que les approches MLP. L'utilisation des LSTM pour la VAD est étendue par Lavechin et al. (2020). Les auteurs introduisent un banc de filtres paramétré SincNet pour l'extraction de caractéristiques. Cette considération permet un gain de performance significatif par rapport aux caractéristiques acoustiques (MFCC). Les auteurs proposent également une méthode d'apprentissage permettant de réduire la sensibilité du système au domaine considéré.

2. Les PLP sont similaires aux MFCC.

Évaluation

Les sous-sections précédentes présentent les approches de la littérature pour résoudre la tâche de détection de parole. Cette section introduit les métriques utilisées pour évaluer la tâche de VAD. Les premières approches utilisent le taux d'erreur sur les mots (Word Error Rate, WER) (Lippmann 1997) pour évaluer les modèles. Cette métrique est couramment utilisée pour la reconnaissance de la parole. Elle correspond au rapport entre le nombre d'erreurs sur les mots et le nombre de mots dans la référence. L'*Equal Error Rate* (EER) (Doddington et al. 2000) est également utilisé. Il correspond au point où le taux de détection manquée ($Miss$) et le taux de fausse alarme (F_A) sont identiques sur la courbe définissant la région de convergence (courbe ROC). Pour ces deux métriques, plus la valeur est basse, meilleur est le score.

Le tableau 2.2 synthétise les méthodes et les performances des systèmes présentés dans les deux sections précédentes. La comparaison entre les approches est délicate. Les contributions utilisent différentes données et les métriques ne sont pas homogènes. Cependant, Lavechin et al. (2020) proposent l'utilisation de taux d'erreurs pour évaluer la segmentation. Ces taux se rapprochent des méthodes d'évaluation de la diarisation. Par souci d'homogénéisation, ces métriques sont choisies pour évaluer les systèmes de VAD développés au cours de cette thèse.

	Auteurs	Modèle	Données	Métrique	Score
Signal	(Chengalvarayan 1999)	Énergie	n.d.	$\text{WER}_{\%\downarrow}$	4,18
	(Sohn et al. 1999)	HMM	n.d.	$P_{d\%\uparrow}$	97,3
	(Nemer et al. 2001)	LPC	TIA	$P_{d\%\uparrow}$	79,9
	(Pfau et al. 2001)	HMM	n.d.	$\text{WER}_{\%\downarrow}$	41,4
	(Kristjansson et al. 2005)	HMM	n.d.		-12,93
	(Ng et al. 2012)	GMM	RATS	$\text{EER}_{\%\downarrow}$	1,42
Neuronal	(Ryant et al. 2013)	LSTM	HAVIC	$\text{EER}_{\%\downarrow}$	19,6
	(Hughes et al. 2013)	RNN	n.d	$F_{A\%\downarrow}$	10,5
	(Thomas et al. 2014)	CNN	RATS	$\text{EER}_{\%\downarrow}$	2,2
	(Gelly et al. 2017)	LSTM	AMI	$\text{FER}_{\%\downarrow}$	5,9
	(Lavechin et al. 2020)	LSTM	DIHARD 3	$F_A + Miss_{\%\downarrow}$	9,9

Table 2.2 – Performance des modèles de détection d'activité vocale dans la littérature. Le sigle n.d. indique que les données ont été obtenues par les auteurs et ne font pas partie de jeux de données publiés. Les métriques indiquées ↓ sont meilleures quand le score est faible.

La VAD est une tâche de détection binaire. Deux types d'erreurs interviennent :

- Fausse alarme : une trame de silence est détectée comme contenant de la parole,
- Détection manquée : une trame de parole n'est pas détectée.

Le taux de fausses alarmes $F_{A\downarrow}$ correspond au rapport de la durée cumulée T_{F_A} des segments associés à la classe *parole* par erreur sur la durée totale des signaux T_{Tot} :

$$F_{A\downarrow} = \frac{T_{F_A}}{T_{Tot}}. \tag{2.25}$$

La flèche ↓ indique que, plus la valeur d'une métrique est faible, meilleure est la performance. Le taux de détections manquées $Miss$ est défini de façon similaire à partir de la durée cumulée T_{Miss} des segments associés à la classe *silence* par erreur :

$$Miss_{\downarrow} = \frac{T_{Miss}}{T_{Tot}}. \tag{2.26}$$

Ces deux métriques sont complémentaires et peuvent être combinées pour obtenir un taux d'erreur de segmentation (SER). Un système de VAD performant offre des valeurs de fausse alarme et de détection manquée faibles.

2.3.2 Détection de parole superposée

La parole superposée apparaît lorsque plusieurs locuteurs sont simultanément actifs. Elle est source de dégradation des performances dans les systèmes de diarisation (Bullock et al. 2020 ; Garcia-Perera et al. 2020). Il est donc nécessaire de détecter ces évènements afin d'en prévoir un traitement spécifique. Les premières études sur la détection de parole superposée (OSD) utilisent des méthodes statistiques. Comme pour la VAD, ces méthodes sont remplacées par les réseaux de neurones dont les capacités de modélisation des signaux sont accrues.

Approches statistiques

Les premières études sur la détection de parole superposée utilisent des modèles HMM-GMM. Boakye et al. (2008) font partie des premiers auteurs à s'intéresser à cette tâche dans le contexte de la diarisation. Ils proposent un modèle de Markov à trois classes utilisant diverses caractéristiques acoustiques. Chaque trame peut soit contenir de la parole superposée, soit un seul locuteur actif, soit aucun locuteur. Ils montrent également que la prise en compte de la parole superposée est bénéfique pour la diarisation. Les travaux sur l'OSD se sont ensuite développés et sont principalement basés sur les modèles HMM/GMM (Boakye et al. 2011 ; Charlet et al. 2013 ; Lee et al. 2016 ; Yella et al. 2014). En particulier, les travaux de Charlet et al. (2013) et Yella et al. (2014) proposent de réassigner les segments de parole superposée à partir d'une modélisation de la structure du discours. Cette étape améliore significativement la diarisation.

D'autre part, Vipperla et al. (2012) proposent une approche basée sur l'analyse du signal. Les auteurs appliquent une méthode de factorisation en matrices non-négatives sur les spectrogrammes extraits des segments de parole. Cela permet de déterminer l'activité de chaque locuteur au cours du temps et d'identifier les segments de parole superposée.

Suite au développement des réseaux de neurones et de l'apprentissage pour le traitement automatique de la parole, les modèles statistiques ont ensuite été remplacés par ce type d'approches. La sous-section suivante décrit les approches neuronales de l'état de l'art pour la détection de parole superposée.

Approches neuronales

La première approche neuronale pour la détection de parole superposée est proposée par Geiger et al. (2013). Les auteurs utilisent un réseau récurrent LSTM pour détecter la parole superposée suite au succès de cette architecture pour la VAD. Les expériences menées montrent que le LSTM surpasse les performances d'un HMM. Andrei et al. (2017) ont ensuite proposé l'utilisation de réseaux convolutifs pour l'OSD. Cependant, les expériences sont menées sur des données acquises par les auteurs et offrant une représentativité très restreinte (uniquement des locuteurs masculins). Il est difficile d'évaluer l'efficacité de cette approche par rapport à l'état de l'art. Kunesová et al. (2019) proposent également une architecture CNN pour la détection de parole superposée. Ils montrent que l'architecture développée permet de bonnes performances de détection. Les auteurs évaluent également l'impact de la parole superposée sur un système de diarisation.

Sajjan et al. (2018) comparent plusieurs architectures neuronales pour la détection de parole superposée. Les modèles sont évalués et comparés sur deux jeux de données. Les auteurs montrent que les architectures LSTM offrent les meilleures performances de détection. Ils montrent également que la détection de segments de parole superposée permet un gain significatif sur les performances de diarisation. Bullock et al. (2020) proposent également une architecture LSTM pour l'OSD. L'architecture est similaire à celle proposée par Lavechin et al. (2020) pour la VAD. Il s'agit de deux couches LSTM prenant en entrée des caractéristiques extraites par SincNet. Les auteurs introduisent également une méthode d'augmentation de données consistant à combiner des segments de parole au cours de l'apprentissage afin de simuler de la parole superposée. Cela permet de corriger le déséquilibre entre les classes et améliorer les performances de détection. Un algorithme Bayésien d'assignation des segments de parole superposée est également proposé. Lebourdais et al. (2022) proposent l'utilisation de caractéristiques pré-entraînées extraites à l'aide de WavLM. Un système TCN est utilisé pour la modélisation de séquence. Cette architecture permet d'atteindre des résultats à l'état de l'art pour l'OSD.

Approches à trois classes

Les tâches de détection de parole et de parole superposée peuvent être considérées comme complémentaires. En effet, la VAD consiste à détecter les segments contenant de la parole. L'OSD détecte ceux où au moins deux locuteurs sont actifs simultanément. Un modèle peut alors être entraîné à prédire la présence d'un locuteur ou de plusieurs locuteurs actifs à chaque trame. Jung et al. (2021) proposent un modèle neuronal composé de couches convolutives et récurrentes (CRNN). Le système prend en entrée un spectrogramme en échelle Mel et prédit simultanément la présence d'un (VAD) ou plusieurs locuteurs (OSD). L'apprentissage joint des deux tâches permet un gain significatif sur la tâche d'OSD.

Cornell et al. (2020) introduisent également un modèle multi-classes. Le système proposé est entraîné à compter le nombre de locuteurs actifs à l'échelle de la trame. Ce modèle permet

également de détecter la présence de parole (N_{spk}=1) et de parole superposée ($N_{spk} > 1$). Cette étude est étendue par Cornell et al. (2022a) où les auteurs développent une architecture Transformer pour résoudre la même tâche. Diverses caractéristiques acoustiques sont également expérimentées. Cette étude est menée sur des données de parole distante et détaillée dans la section 3.4.

Bredin et al. (2021) introduisent un modèle bout-en-bout pour la segmentation de la parole permettant prédire l'activité simultanée de plusieurs locuteurs. Le modèle prédit l'activité des locuteurs à partir du signal audio. En considérant qu'un nombre maximal N_{spk}^{max} peut être actif dans un segment audio, le modèle prédit une séquence binaire pour chaque locuteur. Cette approche permet donc de détecter la présence de parole et de parole superposée. Elle permet également d'affiner la segmentation après une première étape de diarisation (re-segmentation). Le modèle est composé d'un ensemble de couches BLSTM prenant en entrée des caractéristiques SincNet. Cette approche permet une meilleure détection de la parole superposée que Bullock et al. (2020) et Jung et al. (2021).

Évaluation

Les sous-sections précédentes présentent les approches de détection de parole superposée de la littérature. Les approches de l'état de l'art présentées ainsi que leurs performances sont synthétisées dans le tableau 2.3. Contrairement à la VAD, les différentes contributions sont évaluées avec des métriques similaires. Les données utilisées ne sont cependant pas homogènes entre les contributions, rendant là encore la comparaison des approches difficiles.

En se basant sur la littérature récente, quatre métriques d'évaluation sont retenues pour ces travaux. La parole superposée étant couramment formulée comme une tâche de classification binaire, les performances des systèmes peuvent être modélisées par la matrice de confusion $\begin{bmatrix} t_n & f_n \\ f_p & t_p \end{bmatrix}$ avec :

- Vrai positif (t_p) : nombre de vrais positifs, soit les échantillons correctement assignés à la classe positive,
- Vrai négatif (t_n) : nombre de vrais négatifs, soit les échantillons correctement assignés à la classe négative,
- Faux positif (f_p) : nombre de faux positifs, soit les échantillons négatifs dans la référence assignés par erreur à la classe positive,
- Faux négatif (f_n) : nombre de faux négatifs, soit les échantillons positifs dans la référence assignés par erreur à la classe négative.

Deux métriques sont couramment utilisées pour résumer la matrice de confusion. La *précision* évalue le nombre d'échantillons correctement assignés à la classe positive par rapport à tous les éléments classés positifs. Elle est définie par la relation suivante :

1. Les auteurs évaluent leur système avec une *detection accuracy* et non un F1-score.
2. Les auteurs évaluent leur système avec la précision moyenne et non un F1-score.

	Auteurs	Caractéristiques	Modèle	Données	F1-score$_{\%\uparrow}$
Statistique	(Boakye et al. 2008)	Multiple	HMM	ISCI	47,0
	(Vipperla et al. 2012)	TFCT	CNSC	NIST RT	22,8
	(Charlet et al. 2013)	MFCC+PLP	HMM	ETAPE	59,8
	(Yella et al. 2014)	Multiple	HMM	AMI	51,0
	(Lee et al. 2016)	Multiple	HMM	GRID	75,4
2 classes	(Geiger et al. 2013)	MFCC	LSTM	AMI*	34,9
	(Andrei et al. 2017)	Multiple	CNN	Acquises†	72,0
	(Sajjan et al. 2018)	Mel spec.	LSTM	AMI*	71,0[1]
	(Kunesová et al. 2019)	TFCT	CNN	AMI	56,2
	(Bullock et al. 2020)	SincNet	LSTM	AMI	74,9
	(Lebourdais et al. 2022)	WavLM	TCN	DIHARD3	63,4
3 classes	(Jung et al. 2021)	Mel spec.	CRNN	DIHARD3	60,9
	(Bredin et al. 2021)	SincNet	LSTM	AMI	75,3
	(Bredin et al. 2021)	SincNet	LSTM	DIHARD3	59,9
	(Cornell et al. 2022a)	Mel spec.+Spat.	Transformer	AMI*	60,4[2]
	(Cornell et al. 2022a)	Mel spec.+Spat.	Transformer	CHIME6	52,2[2]

Table 2.3 – Performance des modèles de détection de parole superposée dans la littérature. L'exposant $\star$ indique que les résultats sont obtenus sur les données du corpus AMI en conditions distantes.

$$P_\uparrow = \frac{t_p}{t_p + f_p}. \tag{2.27}$$

Le *rappel* évalue la proportion d'éléments positifs détectés par rapport au nombre total d'éléments positifs dans la référence. Il informe sur la robustesse du modèle et est défini par la relation suivante :

$$R_\uparrow = \frac{t_p}{t_p + f_n}. \tag{2.28}$$

Les systèmes de détection sont couramment évalués à l'aide du F1-score, qui correspond à la moyenne harmonique de la précision et du rappel. La flèche $\uparrow$ indique que, plus la valeur d'une métrique est élevée, meilleure est la performance. Cette métrique est définie par la relation suivante :

$$F1_\uparrow = 2\frac{P.R}{P + R}. \tag{2.29}$$

Les valeurs de précision et de rappel dépendent des seuils de décision utilisés en sortie des modèles d'OSD. Le F1-score dépend donc également de ces seuils. Pour évaluer ces modèles sans dépendre de ces seuils, la précision moyenne est également utilisée (Cornell et al. 2020, 2022a). Cette métrique calcule la moyenne de la précision, pondérée par le rappel, pour un ensemble de

seuils de décision th :

$$AP_{\uparrow} = \sum_{th}(P_{th} - P_{th-1})R_{th}. \tag{2.30}$$

Cette métrique apporte des valeurs légèrement plus élevées que le F1-score.

2.3.3 Détection de changements de locuteur

Les deux tâches de segmentation préalablement décrites (VAD et OSD) consistent respectivement à détecter la présence de parole et de parole superposée au sein d'un signal audio. La détection de changements de locuteur (Speaker Change Detection, SCD) vise à détecter la fin d'un tour de parole dans le signal. Cette tâche permet d'obtenir des segments ne contenant qu'un locuteur actif afin d'extraire des caractéristiques du locuteur (embedding) de meilleure qualité pour la diarisation (*cf.* section 2.3.4). Cette section présente une revue des travaux de la littérature menés sur la SCD.

Approches statistiques

Les premières approches de SCD ont été développées dans le contexte de la transcription automatique (S. S. Chen et al. 1998 ; Siegler et al. 1997). Elles consistent à comparer les caractéristiques du signal entre deux fenêtres glissantes successives. Pour cela, les systèmes de la littérature utilisent des modèles statistiques. D. Liu et al. (1999) introduisent un modèle statistique à base de règles pour détecter les changements de locuteur à partir de caractéristiques acoustiques. Sachant que les tours de parole interviennent 80% du temps après un silence, la méthode proposée modélise également les zones de silence. Ils montrent que leur approche permet une amélioration de la transcription par rapport à l'état de l'art.

Les contributions sur la détection de changements de locuteur par des méthodes statistiques sont cependant limitées. Les approches récentes proposées pour résoudre cette tâche sont principalement basées sur les réseaux de neurones et l'apprentissage automatique. Les architectures sont similaires à celles proposées pour la VAD et l'OSD.

Approches neuronales

La majorité des approches neuronales utilisent les réseaux de neurones récurrents pour modéliser une séquence de caractéristiques acoustiques. La présence de changement de locuteur est ensuite prédite à partir de la séquence modélisée. Bredin (2017) propose l'utilisation d'une fonction coût par triplet afin d'apprendre deux représentations de la séquence. Une première représentation est optimisée pour représenter les exemples négatifs, soit l'absence de tour de parole. La seconde modélise la présence de tours de parole. Le modèle apprend ainsi à discriminer les trames correspondant à un changement de locuteur des autres. L'auteur compare cette approche avec deux méthodes statistiques de l'état de l'art et montre un gain significatif.

Yin et al. (2017) proposent l'utilisation d'un modèle BLSTM pour modéliser une séquence de caractéristiques acoustiques puis prédire la présence de tours de parole à l'échelle de la trame. Le système de SCD proposé est similaire à celui développé par Bullock et al. (2020) pour l'OSD. La SCD est formulée comme une tâche de classification binaire. Cependant, les changements de locuteur sont des évènements rares. Une méthode d'augmentation de données est proposée et consiste à définir les tours de parole par une fenêtre de plusieurs trames.

D'autre part, Hrúz et al. (2017) introduisent un système de SCD utilisant des couches convolutives. Dans la référence, les tours de parole sont modélisés par des fonctions triangulaires centrées sur les changements. La SCD est formulée comme une régression de ces fonctions. Les changements de locuteur sont ensuite détectés en appliquant un seuil de détection à la séquence prédite et en identifiant la position des maxima. Les auteurs montrent que cette approche améliore la détection des frontières entre les segments, ainsi que les performances de diarisation.

Évaluation

Les sections précédentes présentent les méthodes de détection de changement de locuteur. Les performances ainsi que les différentes approches sont synthétisées dans le tableau 2.4. La majorité des méthodes de SCD sont évaluées à l'aide de la *pureté* et de la *couverture*. Ces métriques permettent également d'évaluer la segmentation et non plus la classification.

	Auteurs	Modèle	Données	Métrique	Score
Stat.	(Siegler et al. 1997)	Gaussien	ETAPE	$P(\mathcal{R},\mathcal{H})_{\%\uparrow}$	91.0[3]
	(S. S. Chen et al. 1998)	BIC	ETAPE	$P(\mathcal{R},\mathcal{H})_{\%\uparrow}$	90.5[3]
Neuronal	(Bredin 2017)	LSTM	ETAPE	$P(\mathcal{R},\mathcal{H})_{\%\uparrow}$	93.0[3]
	(Yin et al. 2017)	LSTM	ETAPE	$P(\mathcal{R},\mathcal{H})_{\%\uparrow}$	94.7[3]
	(Hrúz et al. 2017)	CNN	CallHome	$\text{EER}_{\%\downarrow}$	17.5
	(Kalda et al. 2022)	LSTM	HUB4	$\text{F1-score}_{\%\uparrow}$	73.0

Table 2.4 – Performance des modèles de détection de changement de locuteur dans la littérature.

La *couverture* évalue la durée de l'intersection entre tous les segments r de la référence $\mathcal{R}$ et chaque segment r de l'hypothèse $\mathcal{H}$. La métrique est normalisée par la durée totale de la référence :

$$C(\mathcal{R},\mathcal{H}) = \frac{\sum_{r\in\mathcal{R}} \max_{h\in\mathcal{H}} |r \bigcap h|}{\sum_{r\in\mathcal{R}} |r|}, \tag{2.31}$$

avec $|.|$ la durée. Cette métrique est maximale lorsque l'intersection entre la référence et l'hypothèse est identique à la référence.

3. Les scores présentés sont issus de (Yin et al. 2017) où les auteurs comparent les méthodes en suivant le même protocole.

L'expression de la *pureté* est similaire, en inter-changeant la référence et l'hypothèse :

$$P(\mathcal{R},\mathcal{H}) = \frac{\sum_{h\in\mathcal{H}} \max_{r\in\mathcal{R}} |h \bigcap r|}{\sum_{h\in\mathcal{H}} |h|}. \tag{2.32}$$

2.3.4 Impact de la segmentation sur la diarisation

Cette section présente un bref état de l'art des méthodes de diarisation. Les études sur l'influence de la segmentation sur les performances sont également reportées.

Approches pour la diarisation

La diarisation (ou segmentation et regroupement de locuteurs) consiste à déterminer *Qui a parlé et quand ?* dans un signal audio (Anguera et al. 2012). Depuis l'utilisation des RNA pour la diarisation, deux types d'approches sont considérés pour résoudre cette tâche. La première concerne les architectures *bout-en-bout*. Il s'agit de modèles neuronaux prédisant l'activité de chaque locuteur en sortie à partir du signal audio brut comme illustré en figure 2.8. La diarisation *bout-en-bout* (End-to-End Neural Diarization, EEND) a d'abord été proposée par Fujita et al. (2019). De nombreux travaux ont ensuite été proposés à partir de ce formalisme (Kinoshita et al. 2021). Bien qu'offrant des performances encourageantes, les modèles EEND nécessitent l'utilisation de données simulées pour l'apprentissage afin d'obtenir de bonnes performances. Les conversations modélisées par le modèle ne sont donc pas réalistes, entraînant une dégradation des performances sur des données réelles. La modélisation de conversations (Landini et al. 2022a) peut être une alternative à cette limitation des modèles EEND.

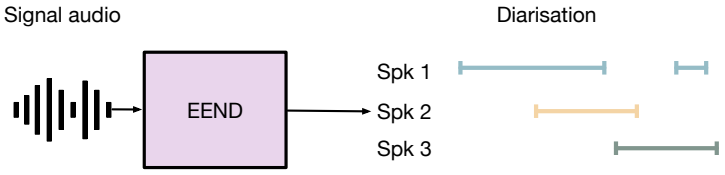


Figure 2.8 – Schéma de principe du modèle EEND (Fujita et al. 2019). Le modèle prédit la diarisation à partir du signal audio brut.

La seconde approche concerne les modèles en cascade. Ils sont couramment composés d'une étape de segmentation, pouvant inclure la VAD, l'OSD et la SCD. Une étape d'extraction d'embeddings de locuteur permet ensuite de modéliser les segments dans un espace de dimension fixe. Le regroupement est appliqué sur les embeddings afin de les regrouper par locuteur (T. J. Park et al. 2022). Les différents blocs de ces architectures sont optimisés séparément, incluant les tâches de segmentation décrites précédemment. La figure 2.9 schématise ce type d'architecture.

Les embeddings de locuteurs représentent les locuteurs par des vecteurs de taille fixe. Ils sont définis de sorte à discriminer les locuteurs dans leur espace de définition. Initiés avec les i-vecteurs (Dehak et al. 2010), ces représentations sont désormais extraites à l'aide de réseaux de

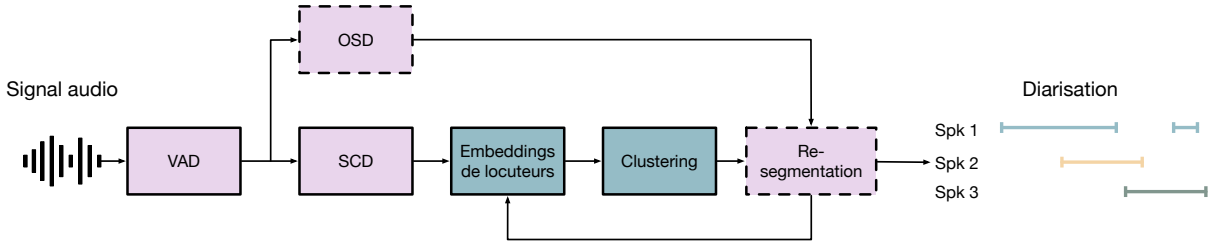


FIGURE 2.9 – Schéma de principe de la diarisation en cascade. La diarisation est obtenue à partir d'embeddings de locuteurs qui sont regroupés à l'aide d'un algorithme de *clustering*. Les étapes de segmentation permettent l'obtention de segments ne contenant qu'un locuteur actif pour une meilleure extraction des embeddings. Les blocs pointillés indiquent des étapes facultatives.

neurones profonds. Les embeddings neuronaux tels que les x-vecteurs (SNYDER et al. 2018) ou les d-vecteurs (Q. WANG et al. 2018 ; A. ZHANG et al. 2019) ont permis des avancées remarquables en diarisation (BREDIN et al. 2020 ; DAWALATABAD et al. 2021 ; LARCHER et al. 2021).

Le regroupement (ou *clustering*) consiste à rassembler des embeddings selon certains critères (ex : distance). Pour la diarisation, ils permettent de regrouper les segments de parole ayant des caractéristiques de locuteur similaires. Des algorithmes tels que le regroupement hiérarchique (Agglomerative Hierarchical Clustering, AHC) (T. J. PARK et al. 2022) ou le regroupement spectral (T. J. PARK et al. 2022) sont couramment utilisés dans ce contexte. Des systèmes en diarisation en cascade comme Pyannote sont disponibles dans la littérature (BREDIN et al. 2020). L'assignation des segments aux locuteurs à partir d'embeddings peut également être réalisée à l'aide d'un modèle statistique. En particulier, LANDINI et al. (2022b) proposent le modèle VBx (Variational Bayes and x-vector) pour la diarisation. Le regroupement est initialisé à l'aide du AHC. Un modèle HMM permet ensuite de modéliser les changements de locuteurs et d'affiner l'assignation.

Évaluation

La tâche de diarisation du locuteur est couramment évaluée à l'aide du taux d'erreur de diarisation (Diarization Error Rate, DER) et du taux d'erreur de Jacquard (Jacquard Error Rate, JER) (T. J. PARK et al. 2022). Le DER combine trois types d'erreurs : le taux de fausse alarme (FA), le taux de détection manquée ($Miss$) et la confusion entre les locuteurs ($Conf$). Cette combinaison d'erreurs est normalisée par la durée des signaux évalués T_{tot} et s'exprime en pourcent :

$$DER = \frac{FA + Miss + Conf}{T_{tot}}. \tag{2.33}$$

En pratique, une tolérance de 0,25 s peut être ajoutée sur les frontières des segments détectés lors de l'évaluation (FISCUS et al. 2006). La somme des différentes erreurs pouvant être supérieure à la durée totale du sous-ensemble d'évaluation, le DER peut atteindre des valeurs supérieures à 100%.

Le JER vise à évaluer la segmentation de chaque locuteur avec le même poids. Le taux d'erreur est calculé pour chacun des locuteurs de la référence puis moyenné sur le nombre de segments N :

$$JER = \frac{1}{N} \sum_{i=1}^{N_{ref}} \frac{FA_i + Miss_i}{T_i} \tag{2.34}$$

avec N_{ref} le nombre de locuteurs dans la référence. T_i représente l'union entre le temps de parole du locuteur i dans l'hypothèse et dans la référence. A l'inverse du DER, le JER ne peut être supérieur à 100%.

Impact de la segmentation sur les performances

La sous-section précédente présente les approches courante pour la diarisation. Dans les systèmes en cascade, la segmentation du signal de parole permet de sélectionner les parties du signal contenant un seul locuteur actif. Elle permet également d'isoler les segments de parole superposée (OSD). Il est cependant nécessaire de questionner l'impact de la qualité de la segmentation sur la diarisation finale. Des études ont été menées sur des systèmes de segmentation statistiques (Huijbregts et al. 2011). Ils montrent que la qualité de la segmentation impacte directement la qualité de la diarisation. La parole superposée est également la première source d'erreurs dans ces systèmes. Cependant, ces études n'ont pas été reproduites sur les systèmes actuellement à l'état de l'art. Quelques études évaluent l'influence de la segmentation sur la diarisation.

Garcia-Perera et al. (2020) étudient l'impact de différents traitements pour la diarisation. Les auteurs montrent notamment qu'un algorithme de VAD performant permet jusqu'à 15% de gain sur le DER. Ils montrent également que la détection de parole superposée puis l'assignation de ces segments aux locuteurs apporte un gain significatif. Ces résultats sont confirmés par l'étude de Bullock et al. (2020) puis par les travaux de Landini et al. (2021).

2.4 Jeux de données

La section 2.3 montre que la plupart des modèles de segmentation et de diarisation utilisent les réseaux de neurones profonds. Ces approches requièrent une grande quantité de données pour être optimisées. Il existe de nombreux jeux de données développés pour la diarisation des signaux monocanal pour divers domaines (sources audiovisuelles, livres audio, téléphone...). La majorité de ces signaux sont acquis en champ proche. Dans le contexte de la parole distante, peu de jeux de données sont disponibles. Il est également préférable d'utiliser plusieurs microphones afin d'acquérir des informations spatiales. Deux solutions se présentent alors :

- l'utilisation d'antennes à géométrie fixe et connue (ex : antenne circulaire uniforme (ACU), linéaire uniforme (ALU)...),

- l'utilisation d'un jeu de microphones dont les positions sont inconnues, parfois appelé antenne *ad-hoc*.

Cette section présente trois jeux de données représentatifs de ces deux cas de figure.

2.4.1 AMI

Le corpus AMI (Carletta et al. 2006) contient environ 100 h de données multimodales enregistrées au cours de réunions. Les données contiennent notamment des enregistrements audio réalisés à l'aide de divers dispositifs (microphone-cravate, antenne...). Les réunions enregistrées sont en partie scénarisées. Les données contiennent donc de la parole actée et de la parole spontanée en anglais. Chaque réunion fait intervenir entre quatre et cinq participants Lors de l'acquisition des données, les participants ont été assignés à divers rôles dans un projet virtuel. Chaque locuteur a donc un rôle précis.

Les signaux audio du corpus AMI ont été enregistrés à l'aide de divers dispositifs tels que des antennes de microphones (deux modèles) et des microphones placés sur les participants. Des signaux acquis en champ proche et en conditions distantes sont donc disponibles.

2.4.2 AISHELL-4

Le corpus AISHELL-4 (Fu et al. 2021) propose 120 h d'enregistrement audio de réunions en mandarin. À l'exception du thème de la réunion qui est fixé au début des réunions, aucun scénario n'est imposé. La parole est donc majoritairement spontanée. Les réunions font intervenir jusqu'à huit participants dans différents environnements acoustiques.

Les données disponibles ont été acquises à l'aide d'une antenne de microphone uniquement. Seuls les signaux distants sont donc disponibles.

2.4.3 CHIME-6

Les données de CHIME-6 (Watanabe et al. 2020) sont issues du challenge éponyme axé sur la transcription de signaux en conditions acoustiques difficiles (bruit de fond élevé et parole distante). Le corpus contient 60 h d'enregistrements de parole en anglais réalisés au cours de soirées entre amis. Les signaux audio sont enregistrés à l'aide de six antennes de quatre microphones placées à différentes positions dans les appartements. Il s'agit donc de conditions *ad-hoc*, la position des antennes n'étant pas connue.

2.5 Conclusions

Ce chapitre introduit les architectures neuronales couramment utilisés pour l'apprentissage automatique dans le cadre du traitement de la parole. Les méthodes de représentation de ce type de signaux sont également présentées avant de détailler l'état de l'art sur la segmentation monocanale. Le chapitre dresse une synthèse des approches pour la détection de parole (VAD), de parole

superposée (OSD) et de changement de locuteur (SCD). L'étude bibliographique montre que la segmentation est un élément clef pour la diarisation en locuteur. Les architectures récurrentes sont privilégiées afin de modéliser les dépendances temporelles entre les trames de caractéristiques extraites du signal audio. Elles permettent aujourd'hui les meilleures performances sur les trois tâches présentées. Les jeux de données AMI, AISHELL-4 and CHIME-6 sont également présentés. Ils permettent l'apprentissage de systèmes dans le contexte de la parole distante. Les travaux sont menés sur les données AMI, qui permettent de comparer les performances de systèmes en conditions proches et distantes. L'utilisation d'antennes à géométrie fixe permet également la mise en œuvre de méthodes spécifiques (*cf.* chapitre 6). Les antennes *ad-hoc* ne sont donc pas considérées dans ces travaux. Bien que l'utilisation de plusieurs microphones puisse être bénéfique pour la segmentation (Cornell et al. 2022a), peu de travaux sont menés sur la diarisation dans ce contexte. Les signaux multicanaux ont cependant été utilisés dans divers domaines du traitement de la parole et requièrent des traitements spécifiques. Le chapitre suivant introduit les méthodes de traitement des signaux multicanaux pour le traitement automatique de la parole.

Chapitre 3

Acquisition et traitement multi-microphones pour la diarisation en locuteurs

La majorité des algorithmes de traitement automatique de la parole sont conçus pour l'analyse de signaux enregistrés dans des conditions acoustiques favorables (faible bruit de fond, peu de réverbération). Dans le contexte des réunions, l'obtention de signaux de haute qualité nécessite l'utilisation de microphones individuels pour chaque participant. Cette configuration présente cependant des limites logistiques, notamment si une personne arrive en cours de réunion. Il est alors préférable d'installer un dispositif au centre de la table afin d'enregistrer la scène. Les signaux acquis dans ces conditions sont cependant sujets à des dégradations causées par le bruit ambiant (ex : vidéo-projecteur) et la réverbération. L'impact du bruit ambiant et de la réverbération est lié à la distance accrue entre les sources et les récepteurs. Le terme *parole distante* est employé pour décrire ces conditions d'acquisition.

En conditions distantes, il est courant d'utiliser plusieurs microphones afin d'acquérir des informations supplémentaires sur les sources. Ce chapitre introduit la notion d'antenne acoustique et les différents traitements pouvant être réalisés à partir des signaux obtenus avec ces dispositifs. Les signaux ainsi enregistrés sont composés de plusieurs canaux et permettent l'utilisation d'algorithmes spécifiques afin de réduire l'influence du bruit et de la réverbération. En particulier, les canaux peuvent être combinés afin de réaliser un filtrage spatial en sélectionnant une direction spécifique. Ces filtres, obtenus par formation de voies, sont couramment utilisés pour le traitement de la parole multicanale. Les signaux multicanaux permettent également l'extraction d'information sur la répartition spatiale des sources dans l'espace. Les caractéristiques spatiales ainsi extraites apportent des informations supplémentaires aux modèles de traitement de la parole distante.

Les travaux menés dans cette thèse portent sur l'exploitation de signaux issus d'antennes de microphones pour la segmentation de la parole, dans le contexte de la diarisation. Un état de l'art sur les méthodes multicanales de diarisation du locuteur est donc présenté.

La section 3.1 présente les antennes acoustiques. Les sections 3.2 et 3.3 introduisent respectivement la formation de voies et les caractéristiques spatiales couramment utilisées. La section 3.4 dresse un état de l'art des méthodes de diarisation multicanale avant de conclure en section 3.5.

3.1 Antennes de microphones

Les antennes de microphones sont des dispositifs d'acquisition d'un signal acoustique composés de plusieurs microphones. Le principe de ces dispositifs ainsi que les géométries couramment utilisées pour le traitement de la parole sont présentés dans cette section.

3.1.1 Principe général

Les antennes de microphones sont des réseaux de capteurs arrangés selon une certaine géométrie. La position relative de chaque microphone est donc connue. Ces dispositifs permettent d'acquérir un signal acoustique en plusieurs points de l'espace. Le signal ainsi obtenu est composé de plusieurs canaux, chacun étant associé à un microphone. L'acquisition spatiale du champ acoustique permet au signal multicanal d'encoder cette information. Elle peut ensuite être utilisée pour localiser les locuteurs, séparer les sources, réduire le bruit au sein du signal (rehaussement), etc.

Les antennes de microphones ont vu le jour dans les années 1960 (Benesty et al. 2015). Le terme *antenne* est issu des similarités entre ces dispositifs et les radars et sonars électromagnétiques. Les antennes acoustiques font cependant face à divers problèmes liés aux conditions d'acquisition des signaux (principalement distantes) et aux propriétés du signal (Benesty et al. 2015) :

(i) les signaux acoustiques, et notamment la parole, ont une large bande passante, nécessitant de travailler dans plusieurs bandes de fréquence,

(ii) ces signaux sont peu stationnaires rendant les estimations statistiques délicates,

(iii) les acquisitions étant principalement réalisées en milieu fermé, la réverbération dégrade le contenu utile du signal,

(iv) le nombre de microphones est souvent réduit, ce qui limite la quantité d'information spatiale acquise.

De nombreuses méthodes de traitement des signaux multicanaux ont cependant été développées. Le principe de fonctionnement d'une antenne de microphones est détaillé dans la sous-section suivante à partir de deux géométries couramment utilisées pour le traitement automatique de la parole.

3.1.2 Géométries courantes

Pour permettre la mise en œuvre d'algorithmes de traitement des signaux issus d'une antenne de microphones, plusieurs géométries ont vu le jour. Cette section présente deux géométries courantes : linéaire et circulaire uniformes.

Antenne linéaire uniforme

L'antenne linéaire uniforme (ALU) est la géométrie la plus simple. Elle consiste à placer un jeu de M microphones séparés d'une distance δ sur un axe, comme illustré en figure 3.1.

En considérant une source unique émettant un signal $s(t)$ et distante des microphones, l'onde acquise peut être supposée plane. Chaque microphone m mesure un signal :

$$x_m(t) = s(t - \tau_m) + v_m(t), \tag{3.1}$$

où $v_m(t)$ représente la somme des échos précoces, de la réverbération et du bruit additif (stationnaire ou non). Cette grandeur est propre à chaque microphone et porte donc l'indice m. Le retard τ_m correspond au temps de vol entre la source et le microphone m. En pratique, il est plus aisé d'estimer le retard relatif entre une paire de microphones. Dans le cas d'une antenne linéaire uniforme (ALU), et avec $m = 1$ pour référence, ce retard s'exprime :

$$\tau_m - \tau_1 = (m-1)\frac{\delta \cos\theta_s}{c}, \quad m = 1, 2, \ldots, M, \tag{3.2}$$

avec c la vitesse du son et θ_s l'angle d'incidence de l'onde plane. Le retard défini en équation (3.2) varie en fonction de l'angle d'incidence θ_s. Le signal multicanal $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_M(t)]$ obtenu sera donc différent en fonction de la position de la source et encode ainsi l'information spatiale.

L'antenne linéaire présente une limitation majeure : elle encode la position de la source pour $\theta_s \in [0^\circ, 180^\circ]$. En effet, deux sources symétriques par rapport à l'axe de l'antenne induisent le même retard τ_m.

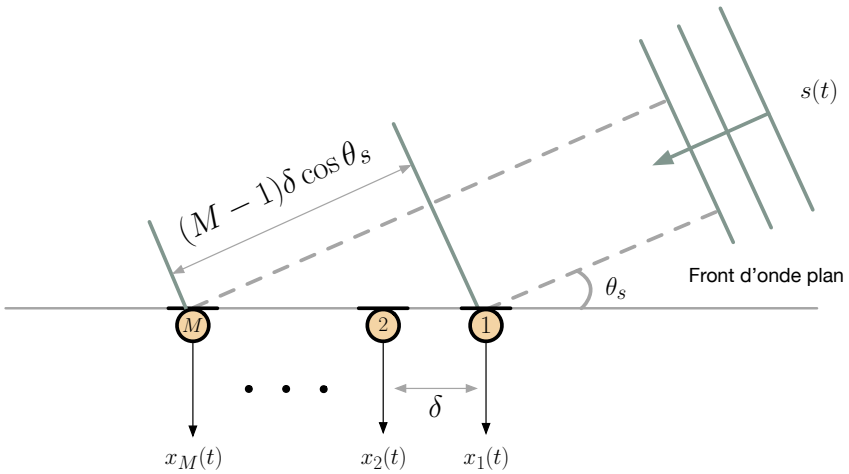


FIGURE 3.1 – Illustration d'une antenne linéaire composée de M microphones.

Antenne circulaire uniforme

L'antenne circulaire uniforme (ACU) est une alternative à l'antenne linéaire. Les microphones sont répartis à équidistance sur un cercle de rayon r comme illustré en figure 3.2. Contrairement à l'ALU, elle permet une modélisation de la position de la source pour tous les angles $\theta_s \in [0^\circ, 360^\circ]$.

Dans le cas d'une ACU, le retard entre un microphone m et le centre de l'antenne est défini

par Benesty et al. (2015) :

$$\tau_m = \frac{r}{c}\cos(\theta_s - \psi_m), \tag{3.3}$$

où $\psi_m = 2\pi(m-1)/M$ représente la position angulaire du microphone m. La distance entre deux microphones est définie par :

$$\delta = 2r\sin\left(\frac{\pi}{M}\right). \tag{3.4}$$

Tout comme pour l'échantillonnage temporel, le phénomène de repliement intervient dans le domaine spatial. Pour éviter son apparition, la distance entre deux microphones doit être inférieure à la moitié de la longueur d'onde acoustique $\delta < c/2f$. Cela permet de déterminer la fréquence limite d'utilisation d'une ACU :

$$f < \frac{cM}{4\pi r}. \tag{3.5}$$

D'autres géométries d'antennes existent dans la littérature, mais ne sont pas mentionnées ici. Il existe également des géométries *ad-hoc* (ex : CHIME-6 (Watanabe et al. 2020)) pour lesquelles la position des microphones dans l'espace est inconnue. Les antennes de microphones permettent l'utilisation d'algorithmes de traitement du signal multicanal (section 3.2).

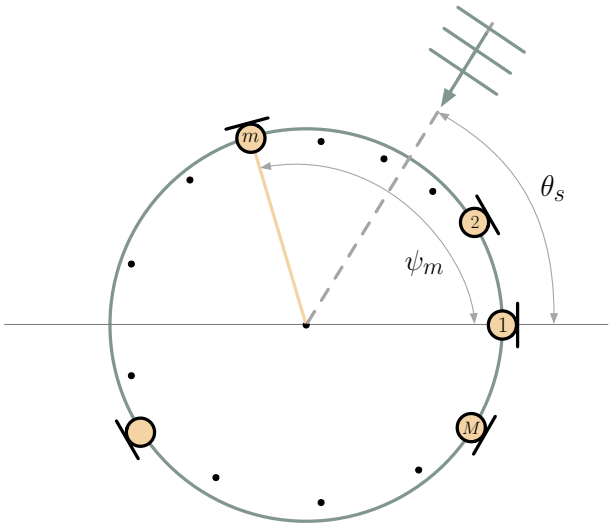


Figure 3.2 – Illustration d'une antenne circulaire composée de M microphones.

3.2 Formation de voies

La *formation de voies* est une approche permettant d'exploiter l'information spatiale contenue dans un signal multicanal. Elle consiste à pondérer et combiner les canaux afin d'appliquer un filtrage d'un signal dans l'espace. Le signal audio résultant est filtré dans une certaine direction de focalisation. En principe, le filtre (ou faisceau) est dirigé dans la direction de la source active afin

d'obtenir une reconstruction du signal source $\hat{s}(t)$ en réduisant le bruit environnant. Aligner le fitre à la direction de la source assure que la reconstruction ne soit pas distordue. Ces approches permettent d'obtenir un signal de sortie monocanal dont le rapport signal-à-bruit est rehaussé. Dans le contexte de la diarisation, cela permet d'obtenir des signaux captés en conditions distantes dont la qualité s'approche de ceux obtenus en champ proche. Plusieurs algorithmes permettent d'obtenir des coefficients de filtres spatiaux.

Une première sous-section présente trois algorithmes usuels pour la formation de voies. Ces algorithmes requièrent l'estimation de paramètres statistiques des signaux et peuvent bénéficier des capacités de modélisation offertes par les réseaux de neurones artificiels (RNA). Une seconde sous-section présente deux types d'approches basées sur les RNA : les algorithmes augmentés par cette technologie et les méthodes bout-à-bout.

3.2.1 Algorithmes usuels

La formation de voies consiste à appliquer un filtrage spatial. Ce filtrage est obtenu en pondérant les canaux par des coefficients et en les combinant. En notant $X_m(f)$ la transformée de Fourier du signal $x_m(t)$ mesuré par le microphone m, la formation de voies s'exprime :

$$\begin{aligned} Y(f) &= \sum_{m=1}^{M} W_m^*(f) X_m(f) \\ &= \mathbf{w}^H(f)\mathbf{X}(f), \end{aligned} \tag{3.6}$$

où $\mathbf{w}(f) = [W_1(f), W_2(f), \dots W_M(f)]^T$ représentent les coefficients du filtre spatial, * représente le conjugué et ^{H} le transposé-conjugué d'un nombre complexe. Le signal $Y(f)$ est la version filtrée du signal multicanal $\mathbf{X}(f)$. Plusieurs algorithmes permettent de calculer les coefficients $\mathbf{w}(f)$ et sont présentés dans les sous-sections suivantes.

Retard et somme

L'algorithme retard et somme consiste à compenser le retard entre les microphones de l'antenne afin d'aligner les signaux dans le temps, puis à en calculer la moyenne. Le coefficient $\mathbf{W}_m(f)$ appliqué au microphone m s'exprime :

$$W_m(f) = \frac{1}{M} e^{-j2\pi f \tau_m}, \quad m = 1, \dots, M. \tag{3.7}$$

Les coefficients de l'équation (3.7) permettent de compenser le retard en chaque microphone dû à la propagation de l'onde plane. Pour estimer τ_m, il est nécessaire de choisir une direction de focalisation θ_s. La reconstruction du signal source $s(t)$ est obtenue lorsque la direction de focalisation correspond à la direction d'arrivée de l'onde.

Minimum Variance Distortionless

L'algorithme Minimum Variance Distortionless (MVDR) (Benesty et al. 2008 ; Souden et al. 2009) repose sur le même principe d'alignement des signaux. Les coefficients $\mathbf{w}$ sont optimisés pour minimiser la puissance du signal filtré (*minimum variance*) sans distordre le signal source (*distortionless*). Le calcul des coefficients est obtenu en résolvant le problème d'optimisation sous contrainte suivant :

$$\min_{\mathbf{w}} \mathbf{w}^H \mathbf{\Phi}_{xx} \mathbf{w} \quad \text{contraint par} \quad \mathbf{w}^H \mathbf{d} = 1, \tag{3.8}$$

où $\mathbf{\Phi}_{xx}(f) = \mathrm{E}[\mathbf{X}(f)\mathbf{X}(f)^T]$ représente la matrice interspectrale du signal $\mathbf{X}(f)$ et $\mathbf{d}$ représente le vecteur de focalisation. Les coefficients du filtre spatial $\mathbf{w}(f)$, solution de (3.8), sont définis par :

$$\mathbf{w}(f) = \frac{\mathbf{\Phi}_{vv}^{-1}(f)\mathbf{d}(f)}{\mathbf{d}^H(f)\mathbf{\Phi}_{vv}^{-1}(f)\mathbf{d}(f)}, \tag{3.9}$$

avec $\mathbf{\Phi}_{vv}(f) = \mathrm{E}[\mathbf{V}(f)\mathbf{V}(f)^T]$ la matrice interspectrale du bruit $\mathbf{V}(f)$. L'expression (3.9) dépend du vecteur de focalisation. Il est donc nécessaire de connaître la direction de focalisation θ_s.

Pour pallier ce problème, Souden et al. (2009) proposent une approche basée sur la sélection d'un microphone de référence. Avec cette formulation, les poids s'expriment :

$$\mathbf{w}(f) = \frac{\mathbf{\Phi}_{vv}^{-1}(f)\mathbf{\Phi}_{xx}(f)}{\mathrm{tr}(\mathbf{\Phi}_{vv}^{-1}(f)\mathbf{\Phi}_{xx}(f))}\mathbf{u}, \tag{3.10}$$

où $\mathrm{tr}()$ représente la trace d'une matrice. $\mathbf{u}$ est un vecteur *one-hot* permettant de sélectionner le microphone de référence. Seul l'élément u_{m_r} associé au microphone de référence m_r vaut 1, les autres sont à zéro.

Valeurs propres généralisées

L'algorithme de formation de voies aux valeurs propres généralisées (Generalized Eigenvalue, GEV) est obtenu en maximisant le rapport signal-à-bruit en sortie du filtre spatial (Deng et al. 2020) :

$$\max_{\mathbf{w}} \frac{\mathbf{w}^H(f)\mathbf{\Phi}_{xx}(f)\mathbf{w}(f)}{\mathbf{w}^H(f)\mathbf{\Phi}_{vv}(f)\mathbf{w}(f)}. \tag{3.11}$$

La résolution de l'équation (3.11) mène au problème des valeurs propres généralisé suivant :

$$\mathbf{\Phi}_{xx}(f)\mathbf{W} = \lambda \mathbf{\Phi}_{vv}(f)\mathbf{W}, \tag{3.12}$$

où $\mathbf{W}$ et λ représentent respectivement les vecteurs propres et les valeurs propres de la matrice $\mathbf{\Phi}_{vv}(f)^{-1}\mathbf{\Phi}_{xx}(f)$. Les coefficients du filtre spatial $\mathbf{w}(f)$ correspondent au vecteur propre associé à la plus grande valeur propre.

L'algorithme GEV peut cependant créer des distorsions dans le signal. Un filtrage peut être

requis sur la sortie du filtre pour limiter ces effets (Deng et al. 2020).

Le calcul des coefficients MVDR et GEV nécessite la connaissance des densités spectrales de puissance du bruit dans l'équation (3.9) et également du signal source en équation (3.10) et (3.11). En conditions réelles, le bruit et le signal utile ne sont pas séparés (Eq. (3.1)). Il est donc nécessaire d'estimer les matrices $\mathbf{\Phi}_{xx}(f)$ et $\mathbf{\Phi}_{vv}(f)$. Certaines approches se basent uniquement sur des approches de traitement du signal pour déterminer ces deux matrices (Schwarz et al. 2015 ; Vincent et al. 2018). Cependant, les approches courantes de la littérature utilisent les RNA pour estimer ces paramètres. D'autres méthodes combinent les canaux directement à partir du signal sans calcul explicite des coefficients. Ces approches sont présentées dans la sous-section suivante.

3.2.2 Méthodes augmentées par les réseaux de neurones

Les réseaux de neurones sont utilisés pour combiner les canaux d'un signal issus d'une antenne de microphones. Ils permettent d'estimer un masque dans le domaine temps-fréquence afin de séparer les éléments contenant du bruit de ceux contenant le signal utile. Les RNA sont également utilisés pour combiner implicitement les canaux, sans le calcul des coefficients de filtre. De nombreux travaux portent sur la formation de voies dans le contexte du traitement automatique de la parole. Les travaux présentés ici se limitent donc à ce contexte.

Masquage temps-fréquence

Une partie des méthodes développées dans la littérature estime des masques dans le domaine temps-fréquence. Ces derniers permettent de séparer les éléments dominés par le bruit de ceux dominés par le signal utile. Les densités spectrales de puissance du signal $\mathbf{\Phi}_{xx}(f)$ et du bruit $\mathbf{\Phi}_{vv}(f)$ sont ensuite estimées à partir des signaux masqués.

Heymann et al. (2016) proposent l'utilisation de BLSTM pour estimer les masques temps-fréquence. Ils sont utilisés pour calculer les coefficients d'un filtre spatial GEV. L'estimation des masques est réalisée de bout-en-bout avec la tâche de reconnaissance de la parole. Les auteurs montrent que cette approche permet de meilleures performances de transcription.

Ochiai et al. (2017) introduisent l'utilisation d'un encodeur-décodeur avec des mécanismes d'attention pour l'estimation des masques. Les densités spectrales de puissance sont utilisées pour calculer les coefficients MVDR. Le modèle de masquage est optimisé simultanément avec la tâche de reconnaissance de la parole. Les auteurs montrent que l'utilisation de mécanismes d'attention permet une meilleure séparation du bruit et du signal, améliorant ainsi la reconnaissance de la parole.

Cornell et al. (2022b) proposent l'utilisation d'un banc de filtres adaptés pour remplacer la TFCT. Un réseau convolutif permet ensuite le calcul de masques à partir des représentations extraites. Les coefficients du filtre spatial ne sont pas fixés dans le temps comme Heymann et al. (2016) et Ochiai et al. (2017). Une estimation de la densité spectrale de puissance est réalisée à

chaque trame. Les coefficients sont donc adaptatifs. L'utilisation de filtres analytiques adaptés pour l'estimation de masques permet de meilleures performances de rehaussement de la parole.

Z.-Q. Wang et al. (2020) développent un modèle neuronal permettant d'estimer directement la TFCT du signal source, en réduisant le bruit de fond. Le système est entraîné à reconstruire les parties réelle et imaginaire de la TFCT à partir du signal brut. Estimer le signal source permet une estimation robuste des matrices de densité spectrale de puissance du signal et du bruit. Les auteurs montrent que cette approche améliore le rehaussement et la reconnaissance de la parole en environnement bruyant.

Les travaux présentés ici utilisent les réseaux de neurones pour améliorer l'estimation des densités spectrales de puissance. Les modèles neuronaux permettent également de combiner implicitement les canaux, sans nécessiter l'estimation de ces paramètres. Les approches de formation de voies neuronales sont présentées en section suivante.

Formation de voies bout-en-bout

B. Li et al. (2016) proposent un modèle de pré-traitement des signaux issus de plusieurs microphones dans le contexte de la reconnaissance de la parole. Il prend en entrée un signal multicanal. Un premier module prédit les coefficients d'un banc de filtres fréquentiels. Un banc différent est appliqué à chaque canal. Les signaux sont ensuite filtrés et sommés sur la dimension des canaux avant d'alimenter un modèle acoustique pour la reconnaissance de la parole. Ces travaux sont étendus par Sainath et al. (2017) où deux architectures supplémentaires sont introduites. Elles consistent à estimer des bancs de filtres fréquentiels indépendants entre les canaux. Les signaux filtrés par bande sont ensuite combinés comme pour la formation de voies. L'estimation de filtres fréquentiels par bandes et par canal permet de visualiser les directions spatiales dans lesquelles le modèle extrait des informations utiles.

Minhua et al. (2019) proposent trois systèmes de traitement du signal multicanal en entrée d'un système de reconnaissance de la parole bout-en-bout. L'un des modèles proposés est notamment initialisé avec les coefficients de plusieurs filtres spatiaux de type retard et somme. Les directions de focalisation d'initialisation peuvent également être mises à jour au cours de l'apprentissage. Lors de la phase d'inférence, le système sélectionne les directions utiles à la reconnaissance de la parole. T. Park et al. (2020) améliorent la stabilité du système précédent en ajoutant un traitement dépendant de la fréquence.

Gong et al. (2021) exploitent les mécanismes d'auto-attention pour combiner les canaux issus de plusieurs microphones. Le système proposé prend en entrée le module de la TFCT du signal issu de chaque microphone. Un mécanisme d'attention estime un jeu de poids associé à chaque canal. Ces poids varient dans le temps et sont appliqués à la TFCT de chaque canal avant de combiner les canaux. Cette approche simple permet une amélioration significative des performances de reconnaissance de la parole en conditions distantes.

Z.-Q. Wang et al. (2020) montrent également que la reconstruction de la TFCT permet directement d'obtenir un signal filtré dans l'espace. Un réseau de neurones est entraîné à recons-

truire le signal source à partir d'un signal multicanal. Ces travaux sont étendus par Tan et al. (2022). Les auteurs analysent l'impact du nombre de microphones et de la position de la source sur les performances de la méthode. Ils montrent qu'elle permet des performances similaires, voire meilleures que les approches basées sur l'utilisation de masques temps-fréquence.

3.2.3 Conclusions

La formation de voies est une technique de traitement des signaux issus d'une antenne de microphones. Elle consiste à pondérer les canaux par des coefficients puis à les combiner pour obtenir un signal focalisé dans une direction de l'espace. De nombreux algorithmes permettent de calculer ces coefficients. Leur calcul requiert cependant l'estimation de paramètres statistiques sur le signal et le bruit. Des méthodes basées sur les réseaux de neurones permettent de masquer le bruit afin d'améliorer cette estimation. Certaines approches neuronales combinent les canaux au sein du réseau, rendant l'estimation des poids implicite. Les approches bout-en-bout restent cependant moins interprétables que les approches avec estimation de masques.

3.3 Caractéristiques spatiales

Les signaux audio acquis par une antenne de microphones contiennent des informations sur la répartition spatiale du champ acoustique. La section précédente montre que cette information permet de construire des filtres spatiaux via la formation de voies. L'information spatiale peut également être extraite sous forme de caractéristiques afin d'enrichir les caractéristiques acoustiques (Section 2.2). Cette section présente les caractéristiques spatiales couramment utilisées pour le traitement automatique de la parole.

3.3.1 Caractéristiques inter-canal

Les caractéristiques spatiales peuvent être extraites à l'aide d'une paire de microphones. Leur utilisation est en partie issue du traitement du signal pour les appareils auditifs et sont donc désignées par les *caractéristiques inter-canal*.

Différences interaurales

Une paire de microphones permet d'acquérir deux signaux différents en fonction de la position des capteurs. Pour une position de source donnée, le niveau du signal acoustique capté par chaque microphone sera différent. Cette variation de niveau peut être représentée par la différence de niveau interaurale (Interaural Level Difference, ILD) (Vincent et al. 2018). L'ILD est calculée à partir des TFCT $X_i(t,f)$ et $X_j(t,f)$ des signaux issus de chaque microphone de la paire $\{i,j\}$:

$$ILD_{ij}(t,f) = 20\log_{10}\left|\frac{X_i(t,f)}{X_j(t,f)}\right|. \tag{3.13}$$

De plus, chaque microphone de la paire capte une version retardée du signal source (*cf.* figure 3.1). Ce retard induit une différence de phase entre les deux signaux mesurés. Celle-ci peut être représentée par la différence de phase interaurale (Interaural Phase Difference, IPD) (Vincent et al. 2018) :

$$IPD_{ij}(t,f) = \angle \frac{X_i(t,f)}{X_j(t,f)}, \tag{3.14}$$

où $\angle$ représente la phase d'une grandeur complexe.

La phase est une fonction discontinue par définition. Elle est représentée sur un cercle telle que $\angle X_i(t,f) \in [0, 2\pi)$. La discontinuité de cette grandeur rend sa modélisation délicate par les réseaux de neurones. Pour contourner cette contrainte, les caractéristiques IPD peuvent être encodées à l'aide de fonctions sinusoïdales (CSIPD) telles que :

$$CSIPD_{ij}(t,f) = \begin{bmatrix} \cos\big(IPD_{ij}(t,f)\big) \\ \sin\big(IPD_{ij}(t,f)\big) \end{bmatrix}. \tag{3.15}$$

Les CSIPD définis en équation (3.15) peuvent être concaténées afin d'obtenir une représentation à une dimension. Les IPD sont utilisées pour la séparation de la parole (Gu et al. 2020b ; Q. Wang et al. 2018), pour la segmentation de la parole (Cornell et al. 2022a) ou encore la localisation de locuteur (Sivasankaran 2020).

Corrélation croisée généralisée

La corrélation croisée généralisée (Generalized Cross Correlation, GCC) a été développée pour estimer le retard entre deux signaux (Knapp et al. 1976). Elle est couramment utilisée pour déterminer la position d'une source acoustique dans l'espace. La différence de temps d'arrivée (Time-Difference of Arrival, TDOA) $\hat{\tau_s}$ entre les signaux mesurés par deux microphones est obtenue à partir d'une fonction d'inter-corrélation. L'argument associé au maximum de cette fonction correspond au retard estimé entre les deux signaux. La fonction GCC est définie par la relation suivante :

$$GCC_{ij}(\tau) = \int_{-\infty}^{\infty} \xi_{ij}(f) X_i(f) X_j^{\star}(f) e^{j2\pi f\tau} df \tag{3.16}$$

avec τ un décalage temporel entre le signal mesuré par le microphone i et celui mesuré par le microphone j.

En équation (3.16), $\xi_{ij}(f)$ représente un facteur de normalisation. La transformation de phase (PHAse Transform, PHAT) est couramment utilisée :

$$\xi_{ij}(f) = \frac{1}{|X_i(f) X_j^{\star}(f)|}. \tag{3.17}$$

La fonction GCC obtenue entre deux microphones d'une paire peut être utilisée comme

caractéristique pour le traitement de la parole (Z.-Q. Wang et al. 2018). La TDOA $\hat{\tau}_s$ obtenue par la relation

$$\hat{\tau}_s = \arg\max_{\tau} GCC_{ij}(\tau), \tag{3.18}$$

peut également être employée (Evans et al. 2009 ; Hu et al. 2015 ; Vijayasenan et al. 2011).

Les caractéristiques binaurales, obtenues à partir d'une paire de microphones, peuvent également être estimées à l'aide d'antennes composées de plusieurs microphones. Elles permettent notamment l'estimation de l'énergie acoustique comme présenté dans la sous-section suivante.

3.3.2 Énergie acoustique

L'énergie acoustique permet de visualiser la répartition spatiale du champ acoustique dans l'espace. Cette grandeur est calculée sur un ensemble de points de l'espace. Cet ensemble est une grille $\boldsymbol{\mathcal{G}} = \{\mathbf{r}_1, \ldots, \mathbf{r}_G\}$, où $\mathbf{r}_g = [x_g, y_g, z_g]$ représentent les coordonnées cartésiennes de l'élément g. L'énergie acoustique peut être évaluée en tout point de la grille en dirigeant un filtre spatial en chaque point $\boldsymbol{\mathcal{G}}$ puis en calculant l'énergie du signal résultant. L'algorithme de la puissance dirigée (Steered Response Power, SRP) (Cobos et al. 2010) permet le calcul de l'énergie en tout point de la grille. Le calcul de la SRP peut également s'exprimer en fonction de la GCC. Pour cela, chaque paire de microphones disponible dans une antenne de M capteurs est considérée :

$$P(\mathbf{r}_g) = \sum_{i=1}^{M} \sum_{j=i+1}^{M} GCC_{ij}(\tau(\mathbf{r}_g)), \tag{3.19}$$

où $P(\mathbf{r}_g)$ représente l'énergie acoustique à la position de l'élément g. L'équation (3.19) nécessite le calcul du retard $\tau(\mathbf{r})$ entre une paire de microphones en fonction de l'élément $\mathbf{r}_g$ courant. Il est défini tel que :

$$\tau(\mathbf{r}_g) = \frac{\|\mathbf{r}_g - \mathbf{r}_i\| - \|\mathbf{r}_g - \mathbf{r}_j\|}{c}, \tag{3.20}$$

avec $\mathbf{r}_i$ et $\mathbf{r}_j$ les coordonnées cartésiennes des microphones i et j.

Si la GCC est calculée avec une pondération PHAT, l'algorithme est désigné par SRP-PHAT. Les caractéristiques SRP sont principalement utilisées pour la localisation du locuteur (Cobos et al. 2010 ; Diaz-Guerra et al. 2020 ; Marti et al. 2011). L'utilisation de ces caractéristiques pour le traitement automatique de la parole reste rare (C. Wu et al. 2021 ; Zheng et al. 2021).

3.3.3 Caractéristiques neuronales

Les réseaux de neurones artificiels permettent également d'extraire des caractéristiques spatiales à partir d'un signal multicanal. Cela permet notamment d'adapter l'extraction de caractéristiques à la tâche visée dans les systèmes bout-en-bout.

Exemple de caractéristiques

Inspirés par les bénéfices des caractéristiques IPD pour la séparation de sources, Gu et al. (2020a) proposent la différence convolutive interaurale. Il s'agit d'un modèle convolutif 1-D appliquant un filtrage à chaque canal d'une paire de microphones et permettant la combinaison de ces dernières :

$$ICD_{ij}^{(f_n)} = \mathbf{w}^a \left(x_i \star k^{(f_n)}(t) \right) + \mathbf{w}^b \left(x_j \star k^{(f_n)}(t) \right), \tag{3.21}$$

où $k^{(f_n)}$ représente le noyau de convolution du filtre f_n et $\mathbf{w}^a$ et $\mathbf{w}^b$ représentent deux fenêtres de pondération de L échantillons. D'après les auteurs, en choisissant des filtres k dont les fréquences sont linéairement espacées (TFCT) et les fenêtres $w_i^a = 1 \ \forall i \in [1, L]$ et $w_j^b = -1 \ \forall j \in [1, L]$, l'équation (3.21) est similaire à la IPD (eq. (3.14)). Cette analogie est cependant discutable. La sortie des filtres k ne correspond pas à une transformée de Fourier à court terme, mais à plusieurs versions du signal temporel, filtré par bandes de fréquences. La différence réalisée dans ce cas de figure ne correspond donc pas à une différence de phase entre les signaux d'une paire de microphones. Les auteurs proposent d'apprendre les filtres k simultanément avec le système de séparation de la parole, rendant ainsi l'extraction de caractéristiques spatiales plus flexible. Ils montrent que cette approche est bénéfique pour la tâche visée.

Cornell et al. (2022a) utilisent un modèle de localisation du locuteur pour obtenir des caractéristiques spatiales. Un réseau de neurones est entraîné à prédire la position des locuteurs dans l'espace à partir du signal multicanal.Cette prédiction est réalisée à partir du signal audio. Il est supposé que le modèle de localisation utilise l'information spatiale contenue dans les données d'entrée pour prédire la position des locuteurs. La représentation des données obtenue avant la couche de prédiction est donc supposée contenir des caractéristiques spatiales permettant de discriminer les locuteurs dans l'espace. Les auteurs utilisent donc cette représentation (*embedding*) comme caractéristique spatiale. Ils montrent que l'utilisation d'embeddings spatiaux peut améliorer les performances d'un système de comptage de locuteurs.

Robustesse à la géométrie de l'antenne

Les caractéristiques spatiales extraites ou modélisées par des réseaux de neurones dépendent fortement de la géométrie de l'antenne utilisée dans les données d'apprentissage. L'utilisation d'une géométrie différente lors de la phase d'inférence peut conduire à de fortes dégradations des performances (Luo et al. 2020 ; Taherian et al. 2022). Pour palier ce problème, Luo et al. (2020) introduisent le modèle *Transform-Average-Concatenate* (TAC). Il s'agit d'un RNA qui projette les données d'entrées dans un nouvel espace. Les données projetées sont ensuite moyennées sur la dimension des canaux. La représentation du signal multicanal ainsi obtenue est moins sensible au nombre de canaux disponibles en entrée. Les auteurs montrent un gain significatif des performances de séparation de la parole avec cette approche, et notamment en cas de non-conformité de l'antenne par rapport à l'apprentissage.

Taherian et al. (2022) proposent une approche similaire. Elle consiste à calculer la moyenne de la TFCT du signal multicanal sur la dimension des canaux :

$$\bar{X}(f,t) = \frac{1}{M}\sum_{m=1}^{M} X_m(f,t), \tag{3.22}$$

puis d'extraire les caractéristiques IPD entre chaque microphone disponible et le signal moyen $I\bar{P}D_i = \angle(X_i(f,t)/\bar{X}(f,t))$. La représentation $I\bar{P}D_i$ est concaténée avec la TFCT de chaque canal disponible. Les auteurs montrent que cette approche permet d'améliorer les performances de reconnaissance de la parole et de rendre le modèle robuste à des antennes non conformes. Notons cependant que la moyenne réalisée en équation (3.22) n'a pas de sens physique. Le calcul des IPD à partir de cette représentation, $I\bar{P}D$, n'en a donc pas plus.

3.3.4 Conclusions

Les caractéristiques spatiales permettent d'extraire des informations sur la répartition du champ acoustique dans l'espace. Ces approches restent cependant marginales par rapport à l'utilisation de la formation de voies pour le traitement automatique de la parole (ex : reconnaissance de la parole, séparation de sources...). La répartition spatiale du champ acoustique est cependant corrélée à l'activité des locuteurs. Cette information peut donc potentiellement faciliter la segmentation et le regroupement en locuteur. La section suivante propose un état de l'art sur l'utilisation de données acquises par plusieurs microphones pour la diarisation.

3.4 Diarisation en locuteur multicanale

Les performances de diarisation tendent à se dégrader en conditions de parole distante (Maciejewski et al. 2018). L'utilisation d'antennes de microphones permet d'acquérir des informations supplémentaires dans ce contexte. De plus, la position des locuteurs dans l'espace influence la répartition spatiale du champ acoustique. L'information spatiale contenue dans un signal acquis par une antenne de microphones est donc corrélée à l'activité des locuteurs. Elle peut permettre d'améliorer les performances de diarisation.

L'utilisation d'information spatiale pour la diarisation a été étudiée au sein des modèles statistiques. Peu de travaux considèrent ce type de données dans les modèles neuronaux actuels.

3.4.1 Diarisation à l'aide de plusieurs microphones

Cette sous-section introduit les méthodes de diarisation utilisant l'information spatiale contenue dans un signal multicanal. Les travaux sont séparés entre les premières approches basées sur des modèles statistiques et les approches neuronales désormais utilisées. Une partie est

spécifiquement dédiée à la segmentation de la parole distante et multicanale, les travaux menés au cours de cette thèse s'intéressant principalement à cette tâche.

Modèles statistiques

Les premiers travaux sur la diarisation à partir de signaux multicanaux exploitent principalement des caractéristiques spatiales TDOA. Celles-ci sont utilisées pour déterminer les paramètres d'un algorithme de formation de voies (Anguera et al. 2007) ou directement comme caractéristiques (Anguera et al. 2007 ; Evans et al. 2009 ; Vijayasenan et al. 2012).

Anguera et al. (2007) proposent un algorithme de formation de voies de type retard et somme pour la diarisation en réunions. Cet algorithme permet de traiter des signaux issus de divers types d'antennes. Il est composé d'une étape de débruitage, suivi d'une estimation des TDOA de chaque source à l'aide de l'algorithme GCC. Les TDOA les plus adaptées à la diarisation sont sélectionnées à l'aide d'un décodage de Viterbi. Elles sont ensuite utilisées pour le filtrage spatial, rehaussant ainsi le signal de parole. Les auteurs montrent que la combinaison de canaux est bénéfique et peut apporter jusqu'à 10% de gain absolu sur le DER.

L'algorithme de combinaison des canaux proposé en (Anguera et al. 2007) permet d'obtenir un signal de meilleure qualité ainsi que les TDOA. Cependant, ces caractéristiques ne permettent pas de discriminer explicitement des locuteurs. Evans et al. (2009) introduisent l'utilisation d'un algorithme non supervisé pour regrouper les TDOA pouvant être associées à un même locuteur, et éloigner celles appartenant à un autre locuteur. Cette approche permet d'améliorer la séparation des locuteurs dans l'espace des TDOA. Les résultats obtenus montrent que la projection des TDOA améliore significativement le DER.

Les caractéristiques spatiales TDOA sont aussi combinées aux caractéristiques acoustiques (MFCC) par Vijayasenan et al. (2011, 2012). Les auteurs montrent que les caractéristiques spatiales améliorent la diarisation lorsqu'elles sont combinées aux MFCC. D'autres caractéristiques acoustiques sont étudiées et leur combinaison est évaluée au sein de deux types de modèles de diarisation.

L'impact de l'intégration de données spatiales pour la diarisation a été largement étudiée au sein des modèles statistiques. Les études ont montré que ce type d'information permet d'améliorer les performances des systèmes sur cette tâche. Les modèles statistiques de diarisation ont ensuite été remplacés par les réseaux de neurones artificiels. La section suivante présente les méthodes de diarisation neuronale utilisant des données acquises à l'aide de plusieurs microphones.

Approches neuronales

Les réseaux de neurones ont permis une amélioration significative des systèmes de diarisation à la fois pour la segmentation des signaux (Bredin et al. 2021 ; Lavechin et al. 2020), l'extraction d'embeddings de locuteurs (Larcher et al. 2021 ; Snyder et al. 2018 ; Q. Wang et al. 2018)

ou la diarisation bout-en-bout (Fujita et al. 2019 ; Kinoshita et al. 2021). Cependant, peu de contributions explorent l'utilisation de signaux issus d'antennes de microphones. Un regain d'intérêt pour ce type de données est cependant à noter suite à l'organisation des campagnes d'évaluation CHIME-5/6 (Watanabe et al. 2020) et M2MeT (Fan Yu et al. 2022).

Kang et al. (2020) proposent de combiner des d-vecteur (Q. Wang et al. 2018) avec une carte d'énergie SRP-PHAT. Les d-vecteurs sont extraits sur un canal de l'antenne à l'aide d'une fenêtre glissante. Une carte d'énergie acoustique radiale est également extraite à partir du signal multicanal sur les mêmes fenêtres. Ces deux caractéristiques sont fusionnées selon deux schémas différents. Les auteurs montrent que l'ajout d'information spatiale est bénéfique à la diarisation.

Zheng et al. (2021) utilisent la formation de voies et l'information de localisation des locuteurs pour la diarisation en réunion. La formation de voies réalise un filtrage spatial dans plusieurs directions angulaires autour de l'antenne de microphones. Les locuteurs sont ensuite localisés à l'aide de la SRP-PHAT. L'information de position des locuteurs permet de sélectionner la direction de formation de voies la plus proche du locuteur actif et sert de caractéristique d'entrée au module de segmentation.

Medennikov et al. (2020) introduisent la VAD ciblée sur les locuteurs (TS-VAD, target-speaker VAD). Ce modèle prend en entrée des caractéristiques acoustiques ainsi que des i-vecteur (Dehak et al. 2010). Le système produit une prédiction de l'activité de chaque locuteur à partir de leurs caractéristiques et de la représentation du signal. Ce modèle est étendu aux données multicanales et a atteint les meilleures performances de diarisation lors de la campagne d'évaluation CHiME-6 (Watanabe et al. 2020).

Peu de travaux s'intéressent à l'utilisation des modèles bout-en-bout EEND sur des données multicanales. Horiguchi et al. (2022) proposent une méthode de distillation permettant d'adapter un modèle EEND à des données mono et multicanales.

A l'exception des modèles bout-en-bout, la segmentation est une étape clef de la diarisation. Les performances des systèmes de diarisation en cascade dépendent fortement de la qualité de la segmentation (Garcia-Perera et al. 2020). La sous-section suivante présente quelques travaux réalisés sur la segmentation de la parole à l'aide de plusieurs microphones.

3.4.2 Travaux dédiés à la segmentation à l'aide de plusieurs microphones

Comme décrit précédemment, les erreurs de segmentation impactent l'extraction d'embeddings de locuteur, car les segments de parole manquent d'homogénéité. Peu de publications portent cependant sur l'utilisation d'antennes de microphones pour la segmentation de la parole.

Hu et al. (2015) proposent l'utilisation du rapport signal-à-réverbération pour la détection de changement de locuteurs dans le contexte de la diarisation. Les caractéristiques extraites correspondent au rapport entre l'énergie de la première partie de la réponse impulsionnelle de la salle et l'énergie de la queue de cette dernière. Les auteurs comparent cette approche avec des caractéristiques GCC et montre un gain significatif sur la segmentation. L'extraction de ces caractéristiques nécessite cependant de connaître la réponse impulsionnelle de la salle. Les

expériences sont donc menées sur des données simulées rendant difficile la comparaison avec d'autres approches.

Cornell et al. (2022a) combinent des caractéristiques acoustiques (MFCC, spectrogramme à échelle Mel...) avec des caractéristiques spatiales IPD et neuronales. Ils montrent que l'information spatiale permet d'améliorer la détection de parole, de parole superposée et le comptage de locuteurs actifs. Ils montrent également que le TCN est une architecture adaptée pour la segmentation de la parole et permet de meilleures performances que les architectures LSTM jusqu'alors utilisées.

3.5 Conclusions

Ce chapitre présente les dispositifs d'acquisition couramment utilisés pour le traitement automatique de la parole. Les antennes de microphones linéaire et circulaire sont présentées afin d'introduire les notions de traitement du signal multicanal. Quelques algorithmes de traitement de ce type de signaux sont ensuite présentés. Deux catégories se dessinent : la formation de voies et l'extraction de caractéristiques spatiales. La première approche consiste à pondérer et à combiner les canaux dans le domaine temporel ou fréquentiel. Cette opération permet d'améliorer le rapport signal-sur-bruit du signal en alignant sur la source cible. Trois algorithmes sont couramment utilisés dans la littérature. L'algorithme MVDR est majoritairement utilisé dans le contexte du traitement de la parole. Il consiste à minimiser le bruit dans le signal et à limiter les distorsions du signal de sortie. Les méthodes de formation de voies requièrent cependant l'estimation de paramètres statistiques du signal (matrices interpectrales). Les réseaux de neurones permettent d'estimer des masques dans le domaine temps-fréquence afin de faciliter l'estimation de ces paramètres. D'autres approches neuronales permettent de combiner directement les canaux à partir du signal brut.

La seconde approche consiste à extraire des caractéristiques représentatives de l'information spatiale contenue dans le signal. Cette information peut être extraite à partir d'une paire de microphones (caractéristiques binaurales) comme les IPD ou le GCC. L'énergie acoustique peut également être calculée sur une grille préalablement définie à l'aide de l'algorithme SRP. Quelques travaux utilisent également les réseaux de neurones pour extraire des représentations spatiales optimisées pour les tâches visées.

Ce chapitre dresse également un état de l'art sur les méthodes de diarisation utilisant plusieurs microphones. Les premières études explorent principalement l'intégration des TDOA comme caractéristiques supplémentaires au sein de modèles statistiques. Les méthodes neuronales exploitant les signaux multicanaux restent marginales. Quelques travaux intègrent des caractéristiques SRP afin de faciliter la discrimination des locuteurs. D'autres approches proposent un traitement implicite des canaux. Peu de travaux portent sur l'utilisation de signaux issus de plusieurs microphones pour les tâches de segmentation, bien qu'elles représentent une étape clefs pour la diarisation. Les seules contributions identifiées explorent l'utilisation de caractéristiques spatiales (ex : IPD) pour segmenter le signal audio. Celles-ci sont couramment fusionnées aux caractéristiques acoustiques.

Chapitre 4

Synthèse, axes de recherche et protocole d'évaluation

Ce chapitre présente une synthèse de l'état de l'art présenté dans les deux chapitres précédents. Les axes de recherche levés par l'étude de la bibliographie sont ensuite introduits. Enfin, le protocole d'évaluation utilisé pour les expériences menées au cours de cette thèse est décrit.

4.1 Synthèse bibliographique

Le chapitre 2 montre que les modèles de diarisation en cascade reposent sur une étape de segmentation puis une phase de regroupement en locuteurs des segments extraits. La segmentation est réalisée à l'aide de trois tâches de détection :

- Détection d'activité vocale (VAD),
- Détection de parole superposée (OSD),
- Détection de changement de locuteur (SCD).

Les systèmes résolvant ces trois tâches suivent le même formalisme. Ils consistent à prédire la présence d'un évènement à partir d'une séquence de caractéristiques. La probabilité de présence d'un évènement est obtenue pour chaque trame de la séquence d'entrée. Les représentations temps-fréquence sont majoritairement utilisées pour représenter le signal. Elles peuvent être extraites à l'aide de méthodes de traitement du signal (Spectrogramme, MFCC) ou de réseaux de neurones (SincNet, filtres optimisés). Des réseaux de neurones prenant en compte les dépendances temporelles (BLSTM, TCN) sont utilisés pour modéliser la séquence de caractéristiques et prédire la présence d'un évènement. Les tâches de segmentation, notamment la VAD et l'OSD influencent fortement les performances des systèmes de diarisation en locuteurs, rendant ces tâches indispensables dans ce contexte.

Les performances des systèmes de segmentation de l'état de l'art tendent à se dégrader en présence de parole distante, couramment rencontrée dans le contexte des réunions. Les bases de données disponibles dans la littérature (AMI, AISHELL-4) fournissent des données enregistrées à l'aide d'antennes de microphones. Ce type de dispositif permet d'acquérir des informations sur la répartition spatiale du champ acoustique et sur la position des sources. Il requiert cependant l'utilisation d'algorithmes spécifiques dont une partie est décrite dans le chapitre 3.

Le chapitre 3 présente les méthodes de la littérature pour le traitement de signaux audio issus d'antennes de microphones. Deux types d'approches existent. La première catégorie de méthodes consiste à pondérer puis combiner les canaux dans le domaine fréquentiel. Ces algorithmes, regroupés sous le terme de formation de voies, permettent de réaliser un filtrage spatial afin d'éliminer les sources de bruit. Ils exploitent désormais les capacités de modélisation des réseaux de neurones soit pour améliorer l'estimation de paramètres statistiques, soit pour combiner directement les canaux. En particulier, Gong et al. (2021) proposent un modèle basé sur l'auto-attention pour combiner les canaux à partir d'une représentation temps-fréquence.

La seconde catégorie consiste à extraire des caractéristiques spatiales à partir du signal. Ces caractéristiques peuvent être combinées avec des caractéristiques acoustiques (ex : MFCC, spectrogramme...). L'utilisation de caractéristiques spatiales reste marginale pour le traitement de la parole, notamment pour la segmentation. Cornell et al. (2022a) proposent cependant de les utiliser pour le comptage de locuteurs à la trame et démontrent l'efficacité de ce type d'approche en conditions distantes. Enfin, les méthodes neuronales de traitement multicanal dépendent fortement de la configuration de l'antenne utilisée lors de l'apprentissage. Peu de méthodes permettent de rendre les systèmes robustes à cette géométrie et au nombre de microphones disponibles (Luo et al. 2020 ; Taherian et al. 2022).

4.2 Axes de recherche

En considérant la littérature étudiée, les travaux menés dans cette thèse se focalisent sur la segmentation de la parole acquise à l'aide d'une antenne de microphones, dans le contexte des réunions. Les tâches de détection d'activité vocale (VAD), de parole superposée (OSD) et de changements de locuteurs (SCD) sont considérées. Certaines expériences sont également menées sur l'impact de la segmentation sur la tâche de diarisation en locuteurs. Les axes de recherche investigués suite à l'étude de l'état de l'art sont listés ci-dessous :

- La combinaison auto-attentive des canaux (Gong et al. 2021) est explorée comme extracteur de caractéristiques pour la segmentation de la parole (VAD et OSD).
- L'approche de Gong et al. (2021) est étendue en utilisant des filtres optimisés en suivant l'approche proposée par Pariente et al. (2020) afin de prendre en compte l'information de phase.
- La combinaison auto-attentive est également étendue dans le domaine complexe. Deux formulations sont proposées dont l'une s'approche d'un filtre spatial linéaire. Cette considération permet de visualiser la réponse spatiale du modèle (Benesty et al. 2008) et fournit une meilleure interprétation du modèle.
- Les travaux de Gong et al. (2021) montrent que l'auto-attention sélectionne efficacement les canaux. Cette approche est également explorée pour sélectionner les sorties d'un banc de filtres spatiaux. Cette approche vise à améliorer l'interprétabilité du système en connaissant les directions angulaires sélectionnées.

- Les travaux de Cornell et al. (2022a) montrent que l'utilisation de caractéristiques IPD améliore la VAD et l'OSD. Ces travaux proposent d'étudier ces caractéristiques pour la détection de changement de locuteurs,
- Les caractéristiques spatiales ne sont pas contraintes à être robustes au nombre de microphones actifs. Le formalisme des harmoniques circulaires (SongGong et al. 2021) est exploré afin de construire des caractéristiques spatiales robustes au nombre de microphones actifs.
- Peu d'approches de traitement de la parole multicanal considèrent la conception de modèles robustes au nombre de capteurs disponibles (Gu et al. 2020a ; Taherian et al. 2022). Une méthode d'apprentissage est développée afin de garantir la robustesse des systèmes de segmentation au nombre de capteurs disponibles.
- Les jeux de données pour le traitement de la parole en réunion sont rares (Carletta et al. 2006 ; Fu et al. 2021), et la position des locuteurs n'est pas annotée dans ces derniers. Un protocole d'acquisition de données multicanales en réunion a été conçu dans l'objectif d'enrichir l'existant.

Les travaux présentés dans ce manuscrit considèrent les trois tâches de segmentation (VAD, OSD et SCD) indépendamment. Elles sont donc résolues séparément à l'aide de différents modèles. Les tâches de VAD et OSD sont cependant complémentaires, et peuvent être résolues simultanément à l'aide d'un modèle unique à trois classes. Une étude a été réalisée durant cette thèse afin d'inclure un traitement multicanal au sein d'un système de segmentation à trois classes. La méthode, le protocole et les résultats ne sont pas présentés dans ce document, mais sont disponibles dans un rapport technique [1].

4.3 Protocole d'évaluation

Cette section présente le jeu de données AMI (Carletta et al. 2006) utilisé pour ces travaux, puis introduit le protocole d'évaluation.

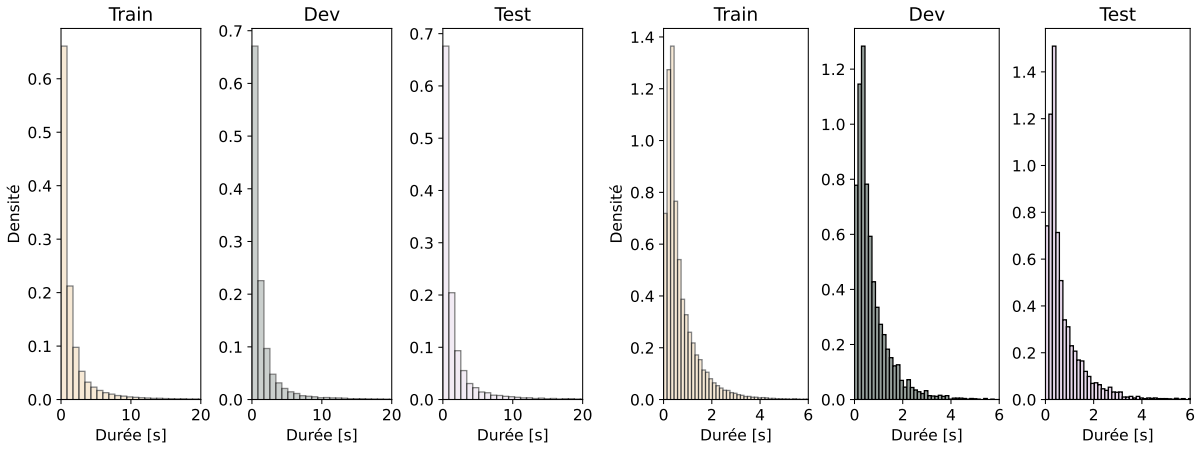
Corpus AMI

Sauf mention contraire, le protocole d'évaluation est identique pour chaque expérience présentée dans les contributions. La section 2.4 a présenté un ensemble de jeux de données pour le traitement de la parole distante multicanal. L'apprentissage et l'évaluation des modèles sont réalisés sur les données du corpus AMI (Carletta et al. 2006) qui propose des enregistrements de réunions à la fois en conditions d'acquisition proche et distante. Les données du corpus AMI ont été enregistrées au cours de réunions incluant quatre à cinq participants. Ces derniers devaient

1. Lebourdais, M., Mariotte, T., Tahon, M., Larcher, A., Laurent, A., Montresor, S., et al. (2023). Joint speech and overlap detection : a benchmark over multiple audio setup and speech domains, https://hal.science/hal-04133268/document consulté le 20/07/2023

mener un projet en passant par toutes les étapes de développement. Les réunions sont parfois libres, avec de la parole spontanée, ou scénarisées.

Lors de la publication initiale des données, plusieurs protocoles ont été proposés par les auteurs (Carletta et al. 2006). Le protocole utilisé pour évaluer des systèmes de diarisation du locuteur était donc susceptible de varier d'une étude à une autre. Récemment, Landini et al. (2022b) ont introduit un protocole unique afin d'unifier la distribution des fichiers entre les sous-ensembles d'apprentissage (*Train*), de développement (*Dev*) et d'évaluation (*Test*). Ce protocole est utilisé dans ces travaux. Les données d'apprentissage contiennent environ 80 h de signal audio avec 152 locuteurs différents. 10% des données d'apprentissage sont réservées à la validation et ne sont pas observées par le système lors de l'apprentissage. Elles permettent de calculer la métrique cible (ex : F1-score) afin de sélectionner le meilleur modèle. Les sous-ensembles de développement et d'évaluation contiennent chacun 10 h de réunions avec 21 et 16 locuteurs respectivement, différents de ceux du jeu d'apprentissage. Les données contiennent 24,7% de parole superposée, faisant des données AMI un bon candidat pour l'OSD. Les annotations pour les tâches de segmentation visées (VAD, OSD et SCD) sont extraites des annotations manuelles des segments, disponible avec les données. Les figures 4.1a et 4.1b présentent la densité des durées des segments de parole et de parole superposée pour chaque sous-ensemble des données AMI. La première figure montre que les segments de parole sont majoritairement localisés dans l'intervalle]0,10] secondes. Les segments de parole superposée sont majoritairement courts, leur durée dépassant rarement les 4 secondes.



(a) Segments de parole (b) Segments de parole superposée

Figure 4.1 – Densité de la durée des segments de parole et de parole superposée pour chaque partition du corpus AMI.

Les signaux de parole du corpus AMI sont acquis à l'aide de divers dispositifs. Les signaux en champ proche sont obtenus à l'aide de microphones placés sur les casques des participants. Un mixage des canaux issus de chaque locuteur a ensuite été réalisé dans la version disponible de la base de données. Ces données sont appelées *headset-mix* par la suite. Les signaux de parole

distante sont enregistrés à l'aide d'une ACU de $M = 8$ microphones omnidirectionnels de rayon $r = 10$ cm. L'appellation *Antenne 1* est ensuite utilisée pour faire référence aux signaux issus de ce dispositif. Une seconde antenne de $M = 4$ microphones est parfois utilisée et placée sur une extrémité de la table au cours des réunions. La géométrie de l'antenne est variable en fonction des séances d'enregistrement (circulaire ou linéaire). Les signaux issus de ce dispositif sont référencés par l'appellation *Antenne 2*.

Protocole d'évaluation

Les modèles de segmentation développés sont entraînés sur des segments de 2 s extraits aléatoirement des données d'apprentissage. Les signaux du corpus AMI sont échantillonnés à 16 kHz. Les étiquettes associées à l'évènement détecté (ex : présence de parole) sont extraits toutes les 10 ms, à l'échelle de la trame. Plusieurs types de modèles sont considérés par la suite en fonction des données d'apprentissage utilisées. Quelle que soit l'expérience considérée, le tirage aléatoire des données est contrôlé par une graine aléatoire (*seed*) afin de permettre la comparaison des performances. Les paramètres des modèles sont mis à jour à l'aide de l'algorithme d'optimisation ADAM (Kingma et al. 2014) après passage des données de chaque lot. Sauf mention contraire, le taux d'apprentissage est fixé à $l_r = 10^{-3}$. Le meilleur modèle est choisi comme étant celui obtenant le meilleur score sur les données de validation. Les hyperparamètres (nombre de couches, dimensions, taille des caractéristiques, etc) choisis sont ceux qui atteignent les meilleurs scores sur les données de développement.

Pour certaines expériences, un modèle de référence entraîné sur les données en champ proche est utilisé. Ce modèle est référencé par l'acronyme CT. Il est entraîné sur les données issues du *headset-mix* en suivant le protocole précédemment décrit.

Un modèle de référence est également entraîné sur les signaux issus d'un microphone distant unique (MDU). Ce modèle est entraîné sur le premier canal de l'Antenne 1 en suivant le protocole précédent. Il s'agit donc d'un modèle monocanal, référencé par l'acronyme MDU. Les modèles exploitant plusieurs microphones sont entraînés sur les données issues de l'Antenne 1 en utilisant tous les microphones disponibles.

L'évaluation des modèles est réalisée sur des segments de 2 s extraits à l'aide d'une fenêtre glissante avec un pas de 25 ms. Les prédictions sont donc obtenues pour chaque fichier séparément. Ces dernières sont moyennées sur les parties recouvertes des fenêtres glissantes (Bredin et al. 2020). Comme présenté en sections 2.3.1 et 2.3.2, les modèles de VAD sont évalués à l'aide des taux de fausse alarme (FA) et de détections manquées (Miss). Le taux d'erreur de segmentation (SER) est également calculé. Les modèles d'OSD sont évalués à l'aide de la précision, du rappel et du F1-score. Pour certaines études, la précision moyenne (AP) est également présentée et permet de résumer les performances du modèle sans tenir compte des seuils de détection. Ces seuils de détection permettent d'obtenir une classification binaire à partir d'une prédiction continue. Le seuil σ_+ indique la valeur de passage de la classe négative à la classe positive. Le seuil σ_- permet

l'opération inverse. Les seuils sont déterminés sur les données de développement puis leur valeur est fixée sur les données d'évaluation.

Sur la tâche de SCD, les modèles sont évalués à l'aide de la pureté et de la couverture. La moyenne harmonique de ces deux valeurs, appelée S-score, est également calculée. Sur cette dernière tâche, un seuil unique de détection est utilisé tel que $\sigma_+ = \sigma_-$.

Sauf mention contraire, la différence statistique entre deux systèmes est évaluée à l'aide du test des rangs signés de Wilcoxon (Demšar 2006). Le test est réalisé entre les performances obtenues pour chaque fichier du jeu de données d'évaluation (respectivement de développement). Un score est considéré statistiquement différent si le test vérifie $p < 0.01$. Les scores indiqués en gras sont ceux issus des systèmes qui vérifient ce test.

Deuxième partie

Contributions

Chapitre 5

Combinaison auto-attentive de canaux pour la segmentation de la parole

L'analyse de la littérature montre que peu de travaux exploitent les signaux multicanaux pour la segmentation automatique de la parole. Peu de travaux exploitent le filtrage spatial ou la combinaison de canaux sûr les tâches de VAD et d'OSD pour la parole distante. Ce chapitre étudie l'intégration de méthodes de combinaison de canaux pour la segmentation automatique de la parole. Les algorithmes développés sont basés sur la même architecture, illustrée en figure 5.1. Le signal multicanal est d'abord représenté dans le domaine temps-fréquence. Cette représentation permet de calculer un jeu de *poids de combinaison*. Les poids sont associés à chaque canal et varient en fonction du temps. Ils permettent de pondérer les canaux dans le domaine temps-fréquence avant de les combiner afin d'obtenir une séquence de caractéristiques composée d'un canal unique.

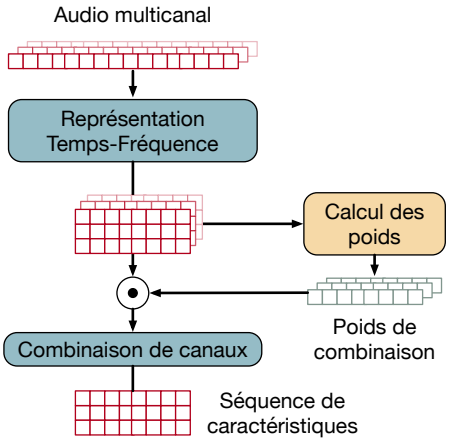


Figure 5.1 – Principe de l'extraction de caractéristiques par combinaison de canaux. La représentation temps-fréquence et la méthode de calcul des poids de combinaison diffèrent en fonction de l'algorithme considéré.

Toutes les approches développées dans ce chapitre s'inspirent de la méthode de Gong et al. (2021) exploitant les mécanismes d'auto-attention pour la pondération des canaux. La section

5.1 présente les systèmes utilisés, le protocole expérimental ainsi que les données utilisées. La section 5.2 présente les résultats obtenus avec l'approche proposée par Gong et al. (2021). Nous proposons trois extensions de cette approche. La première utilise des filtres optimisés comme représentation temps-fréquence (Sect. 5.3). La seconde étend l'approche de (Gong et al. 2021) dans le domaine complexe. Deux formalisations sont proposées en sections 5.4 et 5.5 respectivement. La troisième extension utilise l'auto-attention pour sélectionner les sorties d'un banc de filtres spatiaux (Sect. 5.6). L'impact des modèles de VAD et d'OSD est évalué sur la tâche de diarisation du locuteur en section 5.7.

Questions de recherche Les questions de recherche étudiées dans ce chapitre sont les suivantes :

Q1 L'utilisation de la combinaison auto-attentive des canaux permet-elle d'améliorer la détection d'activité vocale et de parole superposée ? (Section 5.2)

Q2 Est-il bénéfique de permettre au modèle d'apprendre la représentation temps-fréquence avec des filtres optimisés ? (Section 5.3)

Q3 La prise en compte de la phase de la transformée de Fourier dans le traitement améliore-t-elle les performances ? (Sections 5.4 et 5.5)

Q4 Quelles directions spatiales sont utilisées par les modèles complexes afin de détecter la parole superposée ? (Section 5.5)

Q5 La combinaison auto-attentive peut-elle être appliquée à la sélection de filtre spatial ? Quelles possibilités d'interprétation cette approche offre-t-elle ? (Section 5.6)

Q6 Quel est l'impact de ces prétraitements sur la tâche de diarisation en locuteur ? (Section 5.7)

5.1 Méthodologie

Cette section présente le formalisme des tâches de détection d'activité vocale et de parole superposée. Les modèles de référence ainsi que les protocoles d'apprentissage et d'évaluation sont introduits.

5.1.1 Tâches visées

Dans cette section, deux tâches de segmentation automatique de la parole distante sont considérées : la VAD et l'OSD. Ces deux tâches sont basées sur le même formalisme consistant à assigner les trames de caractéristiques à une classe. Pour ces travaux, seules deux classes sont considérées.

Soit $\boldsymbol{X} \in \mathbb{R}^{F \times T}$ une séquence de caractéristiques extraites du signal audio multicanal. La VAD (respectivement l'OSD) consiste à prédire une séquence $\tilde{\boldsymbol{y}} = \{\tilde{\boldsymbol{y}}_1, \ldots, \tilde{\boldsymbol{y}}_N\} \in \mathbb{R}^{C \times T}$ avec $\tilde{y}_n = \{p(c = 0|\boldsymbol{X}), p(c = 1|\boldsymbol{X})\}$. Ici, p représente la pseudo-probabilité de la trame n d'appartenir

à une classe c. En pratique, la séquence $\tilde{\boldsymbol{y}}$ est obtenue à l'aide d'un modèle f_θ de paramètres θ tel que :

$$\tilde{\boldsymbol{y}} = f_\theta(\boldsymbol{X}). \tag{5.1}$$

Les paramètres θ sont déterminés par apprentissage supervisé à l'aide de l'algorithme de rétro-propagation. L'entropie croisée binaire est utilisée comme fonction de perte.

Dans le cas de la combinaison de canaux, les caractéristiques $\boldsymbol{X}$ sont extraites du signal audio multicanal $\boldsymbol{x} \in \mathbb{R}^{M \times L}$ à l'aide d'une fonction g_Θ. Les paramètres Θ sont optimisés simultanément avec les paramètres θ du modèle de classification. La fonction d'extraction de caractéristiques peut également être fixée (ex : formation de voies).

5.1.2 Protocole d'évaluation

Le protocole d'évaluation utilisé est celui présenté en section 4.3. Les modèles sont entraînés sur des lots de 64 segments. Les tâches résolues consistent à classer les trames parmi différentes classes. L'entropie croisée est donc utilisée comme fonction de perte, en moyennant les valeurs à l'échelle des lots. L'algorithme d'optimisation ADAM (Kingma et al. 2014) est utilisé avec un taux d'apprentissage fixé à $l_r = 10^{-3}$.

Les modèles sont ensuite évalués sur les données de développement et d'évaluation du corpus AMI en suivant la procédure décrite en section 4.3. La tâche de VAD est évaluée à l'aide du taux de fausse alarme (FA), du taux de détection manquée (Miss) et du taux d'erreur de segmentation (SER). La tâche d'OSD est évaluée à l'aide de la précision, du rappel et du F1-score.

Dans ce chapitre, les résultats affichés en gras dans les tableaux présentent des performances statistiquement supérieures. Pour cela, le test des rangs signés de Wilcoxon est appliqué entre un modèle et toues les concurrents. Les valeurs des scores entre deux systèmes sont considérées significativement différentes si $p < 0.01$ (Cornell et al. 2022a ; Demšar 2006).

5.1.3 Modèles de référence

Trois types de modèles de référence sont considérés afin d'évaluer les performances des modèles proposés :

- Référence en champ proche (référence *haute*) : le modèle est entraîné sur des signaux acquis en champ proche,
- Référence distante Microphone Distant Unique (MDU, référence *basse*) : le modèle est entraîné sur des signaux issus d'un microphone distant unique (monocanal),
- Référence distante Formation de voies (référence *mixte*) : le modèle est entraîné sur des signaux distants issus de plusieurs microphones et pré-traités à l'aide d'un algorithme de formation de voies.

Les modèles de référence décrits ci-dessus sont présentés en détail dans les sous-sections suivantes.

Modèle en champ proche

Le premier modèle de référence considéré est entraîné sur des signaux acquis en champ proche. Ces derniers sont acquis par les microphones placés sur les casques des participants. Il est présenté par le sigle CT MFCC. Ce modèle étant entraîné sur des données de meilleure qualité, il doit donner des performances cibles pour les modèles distants. Il sert de référence *haute*.

L'architecture se compose d'une couche d'extraction de caractéristiques MFCC, suivie d'un système de modélisation de séquence. Deux types de modèles sont alors considérés pour cette seconde partie. Le premier, inspiré par Bullock et al. (2020), utilise des réseaux récurrents BLSTM suivis d'un RNA. Le second est inspiré par les travaux de Cornell et al. (2022a) et utilise un réseau TCN pour la modélisation de séquence. Les deux types d'architecture sont illustrés en figures 5.2a et 5.2b.

En pratique, 20 coefficients MFCC sont extraits sur une fenêtre glissante de 25 ms avec un pas de 10 ms. Le premier coefficient (énergie) est supprimé pour réduire les instabilités. La vitesse (Δ) et l'accélération ($\Delta\Delta$) des coefficients sont également calculées. Finalement, un vecteur de caractéristiques composé de $F = 59$ coefficients est obtenu.

Le premier système de modélisation de séquence, schématisé en figure 5.2a, est composé de deux couches BLSTM de 256 cellules chacune. Un RNA composé de trois couches linéaires de dimensions $L_1 = 128$, $L_2 = 128$ et $L_3 = 2$ permet de projeter la séquence de sortie du réseau récurrent vers l'espace de décision. Une fonction d'activation Softmax est appliquée afin d'obtenir les pseudo-probabilités $\tilde{y}_n \in [0, 1]$ dont la somme vaut un.

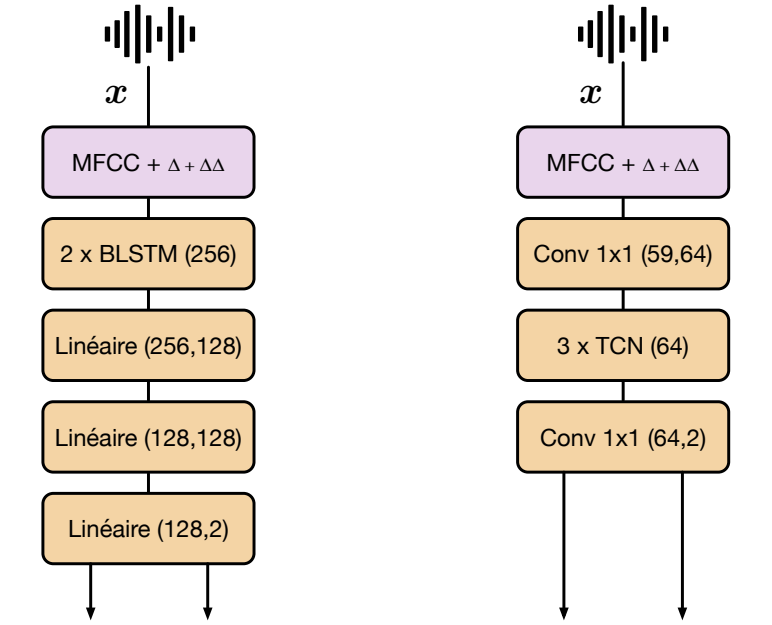
Le second modèle, présenté en figure 5.2b, est composé d'une première couche convolutive, permettant de projeter les caractéristiques vers un espace à 64 dimensions. Cette couche est suivie de trois blocs TCN. Ces blocs contiennent cinq couches convolutives composées d'un noyau de 3 poids et 128 canaux. Le facteur de dilatation est multiplié par deux entre deux couches successives. Une couche convolutive de projection est également utilisée en sortie du dernier bloc TCN pour projeter la séquence modélisée vers l'espace de décision (2 dimensions).

Microphone distant unique

Le modèle de référence distant est entraîné sur les données issues d'un microphone unique. Les deux architectures précédemment introduites sont utilisées. Seules les données d'apprentissage diffèrent. En pratique, il est attendu que les performances de ce modèle soient inférieures à celles du modèle précédent, la qualité des signaux étant dégradée en conditions distantes. Ce modèle est désigné par l'acronyme MDU MFCC et sert de référence *basse* pour l'évaluation.

Formation de voies MVDR

La formation de voies MVDR est utilisée comme méthode de référence pour la combinaison des canaux. Cet algorithme est appliqué dans le domaine temps-fréquence. La TFCT des signaux issus de chaque canal est donc calculée sur des fenêtres de 25 ms avec un pas de 10 ms. Pour



(a) BLSTM (b) TCN

FIGURE 5.2 – Architectures des modèles de référence pour la segmentation avec modélisation de séquence.

chaque fenêtre, 512 fréquences sont extraites. L'algorithme MVDR proposé par SOUDEN et al. (2009) permet de déterminer les poids de combinaison des canaux (*cf.* équation 3.10).

Cette formulation requiert le calcul des matrices interspectrales du signal et du bruit dont l'estimation est réalisée avec la méthode introduite par SCHWARZ et al. (2015). Les auteurs proposent plusieurs méthodes d'estimation du rapport cohérent-diffus (coherent-to-diffuse ratio, CDR). Cette quantité mesure le rapport entre le contenu direct, et donc utile, et diffus du signal. Le calcul de cette quantité repose sur les matrices interspectrales du signal et du bruit. Plusieurs méthodes d'estimation des matrices sont proposées. Dans notre cas, nous choisissons l'approche où la position de la source est inconnue. Pour cela, le bruit est modélisé par un champ isotrope (voir SCHWARZ et al. (2015), eq. (25)).

Après pondération et combinaison des canaux, la TFCT obtenue est monocanale. Le spectrogramme de sortie est converti en échelle Mel à l'aide de $F = 64$ filtres triangulaires. Les caractéristiques ainsi obtenues sont utilisées en entrée des modèles de segmentation présentés en figure 5.2.

Ce modèle traite donc des signaux de parole distante composés de plusieurs canaux. Les modèles de référence utilisant la formation de voies servent de référence *mixte* et sont dénotés par l'acronyme MVDR.

5.2 Combinaison auto-attentive des canaux dans le domaine de Fourier

Dans cette section, nous proposons d'utiliser le modèle Self Attention Channel Combinator SACC (Gong et al. 2021) comme approche d'extraction de caractéristiques pour la VAD et l'OSD. Il exploite le mécanisme d'auto-attention afin d'estimer les poids appliqués aux canaux pour chaque trame de la TFCT. Le modèle est décrit en sous-section 5.2.1 avant la présentation des résultats en section 5.2.2.

5.2.1 Présentation de l'algorithme

Le modèle SACC consiste à estimer des poids $\mathbf{w}$ à appliquer à chaque canal issu de l'antenne. Ces poids sont déterminés à partir du module de la TFCT $|\mathbf{X}_{TFCT}| \in \mathbb{R}^{M\times K\times T}$, où M représente le nombre de canaux, K de fréquences et T de trames. Un mécanisme d'auto-attention permet d'estimer un jeu de poids de combinaison à chaque trame. Trois représentations sont calculées : la *requête* $\mathbf{Q} = q(|\mathbf{X}_{TFCT}|)$, la *clef* $\mathbf{K} = k(|\mathbf{X}_{TFCT}|)$ et la *valeur* $\mathbf{V} = v(|\mathbf{X}_{TFCT}|)$. Les fonctions q, k et v sont implémentées par des couches neuronales linéaires. Les représentations résultantes ont les dimensions suivantes : $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{T\times M\times D}$ avec D la dimension de la couche linéaire et $\mathbf{V} \in \mathbb{R}^{T\times M\times 1}$. Les poids de combinaison sont déterminés par la relation suivante :

$$\mathbf{w} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}}\right)\mathbf{V}. \tag{5.2}$$

Le produit matriciel de l'équation (5.2) est appliqué pour chaque trame $t \in [1, T]$ entre les matrices $\mathbf{Q}_t \in \mathbb{R}^{M\times D}$ et $\mathbf{K}_t^T \in \mathbb{R}^{D\times M}$. L'opérateur $\cdot^T$ représente la transposition.

Les poids résultants sont de dimensions $\mathbf{w} \in \mathbb{R}^{T\times M\times 1}$. Le choix est fait dans l'algorithme original (Gong et al. 2021) de ne pas conserver la dimension des fréquences dans les poids. Le module de la TFCT $|\mathbf{X}_{TFCT}| \in \mathbb{R}^{T\times M\times F}$ est ensuite pondérée par $\mathbf{w}$ avant de combiner les canaux par la relation suivante :

$$\mathbf{X}_{att} = \sum_{m=1}^{M} \text{softmax}(\mathbf{w}) \odot |\mathbf{X}_{TFCT}|, \tag{5.3}$$

avec $\odot$ le produit terme-à-terme tel qu'un poids w_t à la trame t est appliqué à toutes les fréquences du spectrogramme à la même trame $|\mathbf{X}_t|$. La fonction softmax est appliquée sur la dimension des canaux et permet de garantir $w_{t,m} \in [0, 1]$. La somme est appliquée sur la dimension des canaux et permet d'obtenir le spectrogramme moyenné $\mathbf{X}_{att} \in \mathbb{R}^{T\times F}$. Une normalisation des moyenne et variance (NMV) est appliquée avant l'estimation des poids. Cela permet de réduire la sensibilité du modèle aux variations d'amplitude dans le signal d'entrée.

Pour les expériences menées, les couches linéaires q et k ont une dimension $d = 256$. La couche v ne contient qu'une sortie afin de déterminer des poids indépendants de la fréquence (Gong

et al. 2021). Le spectrogramme $\mathbf{X}_{att}$ est converti en échelle Mel à l'aide de $F = 64$ filtres afin d'obtenir la séquence de caractéristiques $\mathbf{X}$. Le système SACC est schématisé en figure 5.3. Par la suite, ce modèle est dénoté SACC TFCT.

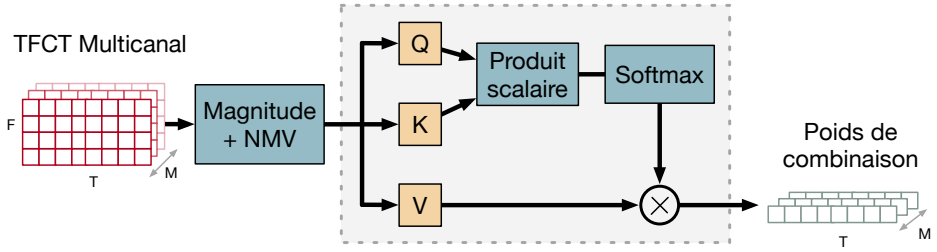


FIGURE 5.3 – Calcul des poids de combinaison à l'aide de l'algorithme SACC proposé en (GONG et al. 2021).

5.2.2 Performances de segmentation

L'extraction de caractéristiques à l'aide de SACC est évaluée sur les tâches de détection d'activité vocale VAD et de parole superposée OSD. Les résultats obtenus sont présentés dans les sous-sections suivantes.

VAD

Les résultats sur la tâche de VAD obtenus à l'aide du modèle SACC sont présentés dans la table 5.1. La première partie de la table présente les performances du système BLSTM. Comme espéré, le modèle de référence en champ proche CT MFCC obtient les meilleurs résultats avec un SER de 5,70% sur les données d'évaluation. D'autre part, le modèle MDU MFCC obtient les performances de détection les plus basses avec un SER à 8,38% sur les données d'évaluation. L'architecture employée étant la même, cette dégradation est uniquement liée au type de signal utilisé lors de l'apprentissage. Pour la VAD, les signaux issus du MDU mènent à une dégradation absolue de -2,68% du SER d'évaluation. L'extraction de caractéristiques à l'aide de la formation de voies permet un gain significatif sur les performances de VAD. Le modèle MVDR atteint un SER de 6,09% sur les données de développement et de 6,73% en évaluation. L'utilisation de la combinaison de canaux auto-attentive SACC atteint des résultats statistiquement équivalents à MVDR. Ce système atteint 5,95% et 6,62% de SER sur les données de développement et d'évaluation respectivement.

La deuxième partie de la table 5.1 présente les performances du système TCN sur la tâche de VAD pour chaque type de caractéristiques. Le modèle en champ proche atteint les meilleures performances avec 5,83% et 5,85% de SER sur les sous-ensembles de développement et d'évaluation. Ces performances sont légèrement dégradées par rapport à celles obtenues avec le BLSTM. Le MDU dégrade les performances par rapport au champ proche avec 6.95% de SER en évaluation. Les pré-traitements MVDR et SACC mènent à des résultats similaires, le premier atteignant

TABLE 5.1 – Performances obtenues à l'aide de chaque type de caractéristique sur la tâche de VAD au sein de deux architectures. Les scores en gras indiquent les meilleurs systèmes en conditions distantes pour chaque modélisation de séquence. Les valeurs en italique représentent les scores en champ proche.

		FA$_{\%\downarrow}$		**Miss**$_{\%\downarrow}$		**SER**$_{\%\downarrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
BLSTM	CT MFCC	*3.32*	*3.12*	*2.36*	*2.61*	*5.68*	*5.70*
	MDU MFCC	4.07	2.08	2.80	6.29	6.88	8.38
	MVDR	3.65	3.07	2.45	3.66	6.09	6.73
	SACC TFCT	3.52	3.07	2.43	3.55	**5.95**	**6.62**
TCN	CT MFCC	*3.14*	*2.41*	*2.69*	*3.4*	*5.83*	*5.85*
	MDU MFCC	4.76	2.95	1.82	3.99	6.57	6.95
	MVDR	3.15	2.65	2.59	3.78	**5.74**	**6.43**
	SACC TFCT	3.51	3.08	2.30	3.39	**5.81**	**6.47**

6,43% de SER et le second 6,47% sur les données d'évaluation. Ces scores sont proches de ceux obtenus sur le sous-ensemble de développement, montrant que le modèle généralise correctement sur ces données.

Le modèle SACC permet d'obtenir des performances proches, voire meilleures (ex : BLSTM) à la formation de voies MVDR sur la tâche de VAD. Cette méthode ne requiert cependant aucun *a priori* sur la statistique du signal et ne nécessite pas l'application de méthodes de masquage dans le domaine temps-fréquence pour estimer les matrices inter-spectrales du signal et du bruit.

OSD

Les systèmes préalablement présentés sont évalués sur la tâche d'OSD. Les résultats sont présentés dans la table 5.2. La première partie de cette table présente les performances du système BLSTM. Sur cette tâche, le modèle en champ proche atteint les meilleures performances de détection avec 70,3% et 67,9% de F1-score respectivement sur les sous-ensembles Dev et Eval. Le MDU obtient un F1-score de 63,7% et 59,5%. Ce modèle affiche une dégradation absolue de -8,4% du F1-score d'évaluation par rapport au CT MFCC. Il présente également un écart relatif de -6,6% entre le développement et l'évaluation. Ici encore, la qualité du signal impacte significativement les performances de détection. L'utilisation de techniques de combinaison de canaux permet un gain significatif des performances en champ lointain. Le modèle MVDR atteint un F1-score de 68,8% et 65,3% sur chaque sous-ensemble. Les caractéristiques SACC permettent d'obtenir des F1-scores de 69,2% et 65,3%. Cela représente un gain absolu de +5,8% sur les données d'évaluation par rapport au MDU. L'écart relatif entre Dev et Eval est réduit à -5,6%.

La deuxième partie de la table 5.2 présente les performances de l'architecture TCN sur la tâche d'OSD. Le modèle en champ proche atteint un F1-score de 73,9% sur les données de développement et 70,2% sur le sous-ensemble d'évaluation. Il s'agit du meilleur modèle obtenu

TABLE 5.2 – Performances obtenues à l'aide de chaque type de caractéristique sur la tâche d'OSD au sein de deux architectures. Les scores en gras indiquent les meilleures performances avec chaque modélisation de séquence. Les valeurs en italique représentent les scores en champ proche.

		Précision$_{\%\uparrow}$		**Rappel**$_{\%\uparrow}$		**F1-score**$_{\%\uparrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
BLSTM	CT MFCC	*68.9*	*76.9*	*71.8*	*60.9*	*70.3*	*67.9*
	MDU MFCC	64.4	65.7	63.0	54.3	63.7	59.5
	MVDR	68.0	71.9	69.6	59.9	**68.8**	**65.3**
	SACC TFCT	68.5	69.9	69.9	61.4	**69.2**	**65.3**
TCN	CT MFCC	*73.8*	*81.4*	*74.1*	*61.8*	*73.9*	*70.2*
	MDU MFCC	69.5	72.8	68.1	59.5	68.8	65.5
	MVDR	74.3	73.1	70.0	66.2	**72.1**	**69.5**
	SACC TFCT	72.5	72.9	72.3	65.2	**72.4**	**68.8**

avec les caractéristiques MFCC. Les performances subissent une dégradation absolue de -4,7% avec le MDU sur le F1-score d'évaluation avec 65,5%. Ici encore, les algorithmes MVDR et SACC obtiennent des performances similaires. Le premier atteint 72,1% et 69,5% en F1-score respectivement sur les sous-ensembles *Dev* et *Eval*. Le second obtient respectivement 72,4% et 68,8% sur ces mêmes ensembles.

Conclusions et discussions

Le modèle SACC a été appliqué à deux tâches de segmentation de la parole distante multicanale. Les performances de VAD (5,81% SER Dev) et d'OSD (72,4% F1-score Dev) obtenues au sein de deux architectures neuronales surpassent le MDU MFCC et égalent ou surpassent la formation de voies MVDR. Les expériences montrent que l'utilisation d'algorithmes utilisant plusieurs microphones améliore significativement la détection, comme le montrent les résultats SACC et MVDR. La combinaison auto-attentive des canaux ne permet pas un gain significatif par rapport à la formation de voies. L'approche SACC ne requiert cependant pas l'estimation des matrices interspectrales du signal et du bruit. Les poids de combinaison sont appris simultanément avec la tâche de classification. L'estimation des matrices interspectrales, requise pour calculer les coefficients MVDR, nécessite des étapes de calcul supplémentaires ajoutant un coût non négligeable.

L'algorithme SACC est un candidat intéressant pour la segmentation de la parole multicanale en conditions distantes. Cependant, la formulation proposée par GONG et al. (2021) n'exploite pas la phase de la TFCT. Cette information renseigne pourtant sur le retard des signaux reçus à chaque microphone et peut-être bénéfique pour la segmentation. Les trois prochaines sections présentent deux formalismes pour intégrer l'information de la phase au sein du modèle SACC. Par la suite, les extensions proposées sont comparées au modèle SACC TFCT.

5.3 Banc de filtres optimisés

La formulation initiale du module SACC estime les poids à partir du module de la TFCT. L'information de la phase n'est donc pas prise en compte. Dans cette section, nous proposons l'utilisation de filtres fréquentiels appliqués dans le domaine temporel et optimisés simultanément avec la tâche. Le traitement est appliqué dans le domaine temporel, ce qui permet de conserver implicitement l'information de la phase dans les caractéristiques extraites.

5.3.1 SACC basé sur des filtres analytiques optimisés

La TFCT est remplacée par des filtres optimisés simultanément avec la tâche de segmentation. Cette approche permet au modèle de sélectionner le contenu du signal utile à la tâche visée. Dans le cas du module SACC, ces filtres permettent d'extraire une représentation temps-fréquence du signal pour chaque canal. Celle-ci est également utilisée pour estimer les poids de combinaison. L'architecture est donc similaire à celles présentées en figures 5.1 et 5.3 en remplaçant la représentation temps fréquence par des filtres optimisés.

Pour extraire les caractéristiques à partir du signal, une couche convolutive 1-d de noyau $K = 400$ est utilisée. Par la suite, le nombre de filtres optimisés sera noté N_f. L'optimisation de filtres temporels d'ordres élevés peut s'avérer délicat (Pariente et al. 2020). Pour régulariser l'apprentissage, le formalisme des filtres analytiques est utilisé (Pariente et al. 2020). Cette approche consiste à apprendre N_f filtres puis à les orthogonaliser à l'aide de la transformée de Hilbert (*cf.* section 2.2). Les caractéristiques obtenues contiennent alors $F = 2N_f$ dimensions et vérifient $\mathbf{X} \in \mathbb{R}^{2N_f \times T}$.

5.3.2 Choix du nombre de filtres

Cette section évalue l'impact du nombre de filtres choisi sur les performances d'OSD. Les filtres sont également visualisés dans le domaine fréquentiel.

Étude de dimensionnement

La table 5.3 présente les résultats obtenus sur la tâche d'OSD sur les données de développement du corpus AMI en fonction du nombre de filtres analytiques choisi. Afin d'évaluer les systèmes indépendamment des seuils de détection, la précision moyenne (AP) est présentée en plus du F1-score. D'après la table, il semble préférable d'utiliser un nombre de filtres réduit. Ici, $N_f = 32$ permet d'obtenir un F1-score de 69,3% et une AP de 73,6%. Les performances varient ensuite fortement en fonction du nombre de filtres utilisé. Par exemple, $N_f = 64$ dégrade les performances en descendant à 62,9% de F1-score. Le cas $N_f = 128$ offre également des performances acceptables. Par la suite, 32 filtres seront considérés. Comme décrit précédemment, la transformation analytique des filtres double le nombre de caractéristiques extraites. Ce système extrait donc $F = 64$ caractéristiques.

N_f	**F1-score**$_{\%\uparrow}$	**AP**$_{\%\uparrow}$
32	**69,3**	**73,6**
64	62,9	65,6
128	66,7	70,6
256	65,1	69,4
512	66,8	70,2

TABLE 5.3 – Influence du nombre de filtres sur les performances de détection de parole superposée. Les résultats sont obtenus sur les données de développement du corpus AMI.

Pour essayer de comprendre la cause de la dispersion des performances entre les systèmes, les filtres appris sont analysés dans la sous-section suivante.

Analyse des filtres

Le modèle SACC basé sur des filtres analytiques optimisés présente des performances variables en fonction du nombre de filtres. Ces derniers sont analysés dans le domaine temps-fréquence en visualisant les poids après apprentissage. Soit $h_i(n)$ le filtre d'indice i. Après convergence du modèle de détection de parole superposée, celui-ci peut être représenté dans le domaine fréquentiel par sa réponse en fréquence :

$$H_i(k) = TF\left[h_i\right](n) = |H_i(k)|e^{j\angle H_i(k)}, \tag{5.4}$$

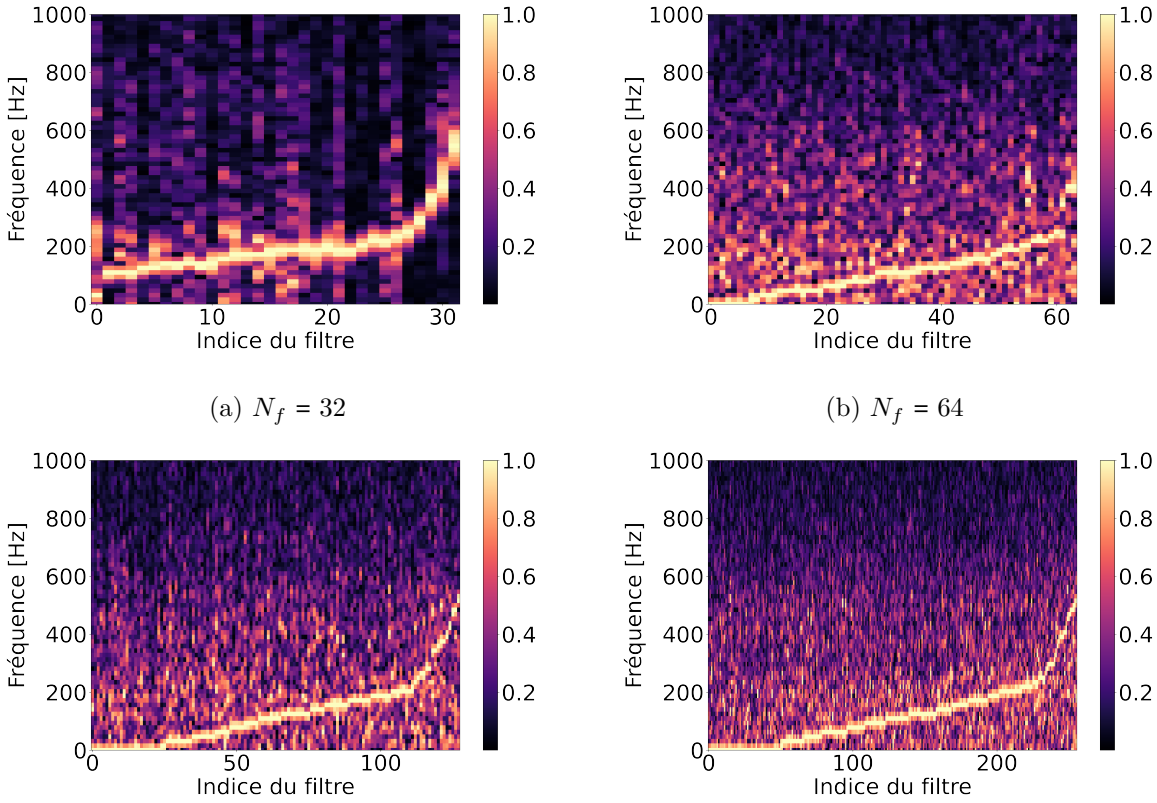
où TF représente la transformée de Fourier, k l'indice fréquentiel et $||.||$ et $\angle$ respectivement le module et la phase d'une grandeur complexe. La figure 5.4 présente le module de la réponse en fréquence de chaque filtre. Pour faciliter la visualisation, les filtres sont triés de la fréquence centrale la plus faible à la plus élevée.

La figure 5.4 montre que les filtres sont principalement regroupés dans la bande [0,600] Hz. Dans le cas où le modèle offre des performances moindres ($N_f = 64$), les filtres sont restreints dans la bande [0,400] Hz. De plus, beaucoup d'entre eux sont redondants en étant centrés sur des fréquences proches. Dans les autres cas, les filtres sont répartis linéairement en fréquence puis présentent une plus forte variabilité sur les derniers indices (ex : filtres 25 à 32 pour $N_f = 32$).

Le fait de réduire le nombre de filtres limite l'apprentissage de filtres redondants. Cela peut expliquer les meilleures performances obtenues par ce système. Enfin, dans les cas $N_f > 32$, de nombreux filtres présentent une fréquence centrale proche de 0 Hz. Ces filtres présentent donc peu d'intérêt pour l'analyse des signaux de parole.

5.3.3 Évaluation par rapport aux références

Cette sous-section présente les performances du modèle SACC basé sur des filtres analytiques sur les tâches de VAD et OSD. Les performances sont comparées à celles des modèles de référence



(a) N_f = 32

(b) N_f = 64

(c) N_f = 128

(d) N_f = 256

FIGURE 5.4 – Module de la réponse en fréquence des filtres analytiques optimisés au sein du modèle SACC pour la tâche d'OSD. Chaque figure correspond à un nombre de filtres N_f différent. Les filtres sont triés par fréquence centrale croissante.

et du système SACC original. Les modèles SACC dotés de filtres analytiques sont référencés comme SACC $\mathcal{A}_{N_f}$. Dans cette section, le modèle N_f = 32 est considéré.

VAD

Les résultats obtenus sur la tâche de VAD avec les filtres analytiques adaptés sont présentés dans la table 5.4. La première partie de la table présente les résultats au sein d'un système BLSTM. L'utilisation de filtres analytiques (SER Dev : 6,61% ; Eval : 7,02%) ne permet pas d'améliorer les performances par rapport au modèle SACC. Dans le cas du système TCN, les observations sont similaires. Les filtres analytiques obtiennent un SER de 6,61% et 7,09% respectivement sur les données de développement et d'évaluation. L'utilisation de filtres analytiques adaptés au sein du module SACC ne permet pas un gain de performance sur la tâche de VAD.

TABLE 5.4 – Résultats sur la tâche de VAD à l'aide du modèle SACC basé sur des filtres adaptés analytiques par rapport aux modèles de référence et à l'algorithme SACC original. Les scores en gras indiquent les meilleures performances avec chaque modélisation de séquence. Les valeurs en italiques correspondent à la référence en champ proche.

		FA$_{\%\downarrow}$		**Miss**$_{\%\downarrow}$		**SER**$_{\%\downarrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
BLSTM	CT MFCC	*3.32*	*3.12*	*2.36*	*2.61*	*5.68*	*5.70*
	MDU MFCC	4.07	2.08	2.80	6.29	6.88	8.38
	MVDR	3.65	3.07	2.45	3.66	6.09	6.73
	SACC TFCT	3.52	3.07	2.43	3.55	**5.95**	**6.62**
	SACC $\mathcal{A}_{32}$	4,62	3,95	2.20	3,07	6,61	7,02
TCN	CT MFCC	*3.14*	*2.41*	*2.69*	*3.4*	*5.83*	*5.85*
	MDU MFCC	4.76	2.95	1.82	3.99	6.57	6.95
	MVDR	3.15	2.65	2.59	3.78	**5.74**	**6.43**
	SACC TFCT	3.51	3.08	2.30	3.39	**5.81**	**6.47**
	SACC $\mathcal{A}_{32}$	3,88	3,31	2.73	3,78	6,61	7,09

OSD

Les résultats obtenus sur la tâche d'OSD sont présentés dans la table 5.5. Ils montrent que le modèle SACC doté de filtres analytiques optimisés n'est pas aussi performant que le modèle original (SACC TFCT). Au sein de l'architecture BLSTM, les filtres analytiques atteignent un F1-score de 65,0% et de 58,7% sur les partitions Dev et Eval respectivement. Ces performances sont similaires à celles du système MDU MFCC. Cependant, une dégradation de -6,6% est observée sur les données d'évaluation par rapport au modèle SACC TFCT. Les conclusions sont similaires au sein de l'architecture TCN. L'utilisation de filtres analytiques comme représentation temps-fréquence du module SACC offre des performances mitigées (F1-score Dev : 69,3% ; Eval : 65,4%). Les résultats obtenus n'égalent pas ceux atteints par la formation de voies (Eval : 69,5%) ou le modèle SACC original (Eval : 68,8%).

L'apprentissage de filtres optimisés à partir du signal ne permet d'améliorer ni la VAD, ni l'OSD sur la parole distante. La figure 5.4 montre que les filtres optimisés sont uniquement localisés dans la bande [0,600] Hz. La dégradation des performances peut être liée au manque de représentation des bandes supérieures. Dans la sous-section suivante, nous proposons d'initialiser les filtres optimisés pour permettre au modèle de prendre en compte un plus large domaine fréquentiel.

5.3.4 Initialisation des filtres

Poids initiaux des filtres

L'initialisation des filtres au début de la phase d'apprentissage peut faciliter l'apprentissage de filtres couvrant la bande passante utile pour la parole. Cette sous-section propose une méthode

TABLE 5.5 – Résultats sur la tâche d'OSD à l'aide du modèle SACC basé sur des filtres analytiques par rapport aux modèles de référence et à l'algorithme SACC original. Les scores en gras indiquent les meilleurs systèmes distants pour chaque modélisation de séquence. Les valeurs en italiques correspondent à la référence en champ proche.

		Précision$_{\%\uparrow}$		**Rappel**$_{\%\uparrow}$		**F1-score**$_{\%\uparrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
BLSTM	CT MFCC	*68.9*	*76.9*	*71.8*	*60.9*	*70.3*	*67.9*
	MDU MFCC	64.4	65.7	63.0	54.3	63.7	59.5
	MVDR	68.0	71.9	69.6	59.9	**68.8**	**65.3**
	SACC TFCT	68.5	69.9	69.9	61.4	**69.2**	**65.3**
	SACC $\mathcal{A}_{32}$	64,9	66,4	65,0	52,5	65,0	58,7
TCN	CT MFCC	*73.8*	*81.4*	*74.1*	*61.8*	*73.9*	*70.2*
	MDU MFCC	69.5	72.8	68.1	59.5	68.8	65.5
	MVDR	74.3	73.1	70.0	66.2	**72.1**	**69.5**
	SACC TFCT	72.5	72.9	72.3	65.2	**72.4**	**68.8**
	SACC $\mathcal{A}_{32}$	71,4	73,9	67,3	58.6	69.3	65.4

d'initialisation inspirée du modèle SincNet (RAVANELLI et al. 2018). Les filtres sont initialisés comme des filtres à réponse impulsionnelle finie (RIF) de réponse en fréquence rectangulaire :

$$H_i(f) = \text{rect}\left(\frac{f}{2f^i_{sup}}\right) - \text{rect}\left(\frac{f}{2f^i_{inf}}\right), \tag{5.5}$$

où f^i_{sup} représente la fréquence de coupure supérieure du filtre, f^i_{inf} la fréquence de coupure inférieure et rect(.) la fonction rectangle. La fréquence centrale des filtres est définie par $f^i_c = \frac{f^i_{inf}+f^i_{sup}}{2}$. La réponse impulsionnelle de ces filtres est obtenue dans le domaine temporel à l'aide de la transformée de Fourier inverse :

$$h_i(n) = 2f^i_{sup}\text{sinc}\left(2\pi n f^i_{sup}\right) - 2f^i_{inf}\text{sinc}\left(2\pi n f^i_{inf}\right), \tag{5.6}$$

avec $\text{sinc}(x) = sin(x)/x$. Les poids de la couche convolutive effectuant le filtrage sont initialisés à l'aide de l'équation (5.6). Les fréquences centrales des filtres sont espacées linéairement et les fréquences inférieures et supérieures sont déterminées en choisissant une largeur de bande Δf fixe. La sous-section suivante présente les résultats obtenus sur la tâche d'OSD lorsque les filtres sont initialisés.

Résultats sur la tâche d'OSD

La table 5.6 présente les résultats obtenus sur la tâche d'OSD en initialisant les filtres analytiques par des filtres rectangulaires de largeur de bande $\Delta f = 100$ Hz. Les performances sont comparées à celles obtenues en sous-section 5.3.2.

Les résultats montrent que l'initialisation des filtres analytiques ne permet pas de gain

significatif sur les performances de détection de parole superposée dans le cas $N_f = 32$. Les meilleurs résultats sont d'ailleurs obtenus avec cette configuration avec un F1-score de 69,4% et une AP de 73,7%. Augmenter le nombre de filtres tend à dégrader la qualité de la détection. L'initialisation des filtres permet une légère amélioration de l'OSD. Par exemple, le cas $N_f = 64$, les filtres initialisés atteignent une AP de 70,9% contre 65,6% sans.

TABLE 5.6 – Performances d'OSD sur les données de développement du corpus AMI en fonction du nombre de filtres choisi, avec et sans initialisation.

	Initialisé		Libre	
N_f	**F1-score**$_{\%\uparrow}$	**AP**$_{\%\uparrow}$	**F1-score**$_{\%\uparrow}$	**AP**$_{\%\uparrow}$
32	**69,4**	**73,7**	**69,3**	**73,6**
64	67,6	70,9	62,9	65,6
128	66,9	70,2	66,7	70,6
256	66,2	69,3	65,1	69,4
512	67,7	70,9	66,8	70,2

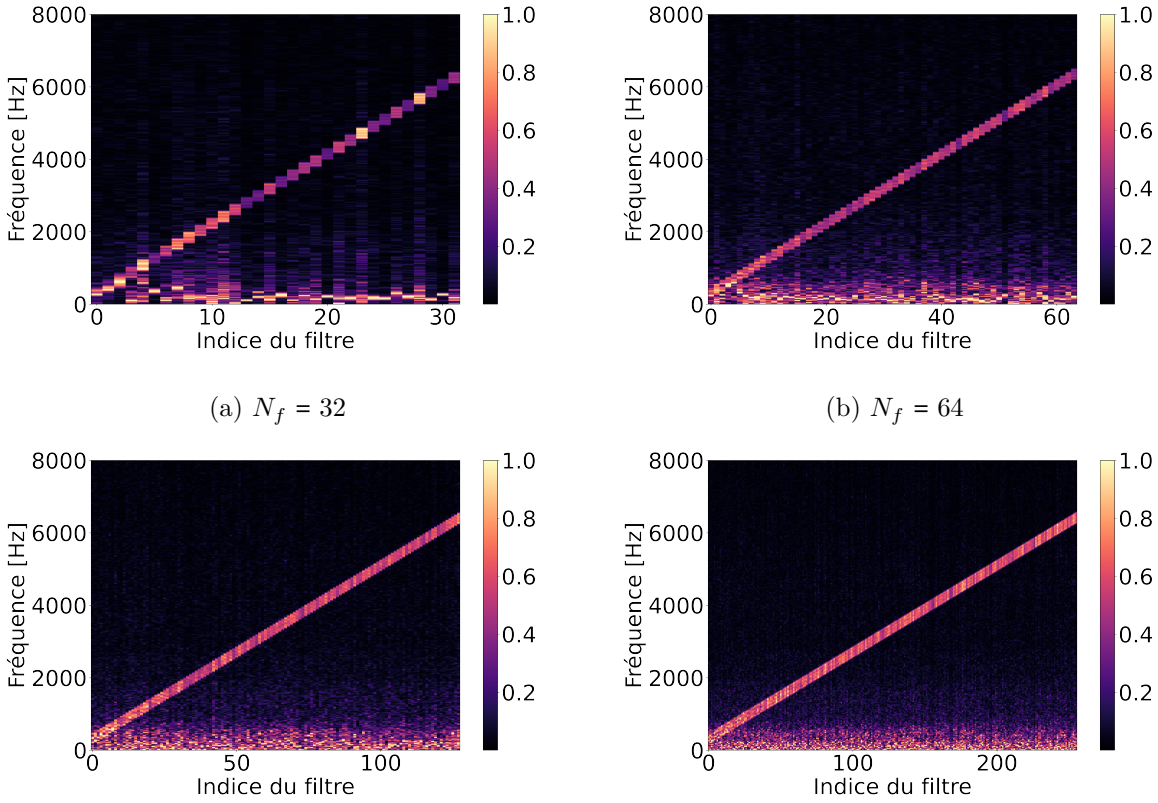
L'initialisation des filtres SACC ne permet cependant pas d'améliorer les performances de détection de parole superposée. Par exemple, le modèle $N_f = 32$ conserve des performances similaires entre les cas initialisé (AP : 73,7%) et libre (AP : 73,6%). La section suivante présente l'analyse des filtres obtenus après la phase d'apprentissage.

Analyse des filtres

Les filtres initialisés sont analysés en suivant l'approche de la section 5.3.2. Le module de la réponse en fréquence des filtres est présentée en figure 5.5. Chaque figure présente une diagonale. Elle correspond à la réponse des filtres lors de l'initialisation. Cependant, l'amplitude de la réponse en fréquence dans cette diagonale n'est pas constante. Les filtres sont modifiés au cours de l'apprentissage. Sur les quatre exemples présentés, une partie des filtres migre vers des fréquences centrales plus basses. Cela est particulièrement visible sur la figure 5.5a. L'initialisation des filtres semble également limiter leur redondance.

Conclusions

Cette section explore l'utilisation de filtres fréquentiels optimisés avec la tâche de segmentation afin de remplacer la TFCT pour la combinaison de canaux avec le module SACC. Le formalisme des filtres analytiques est utilisé, ces derniers ayant montré leurs avantages dans le contexte de la séparation de sources et la formation de voies (CORNELL et al. 2022b ; PARIENTE et al. 2020). La meilleure configuration, avec $N_f = 32$ filtres appris, permet d'atteindre un SER de 7,02% pour la VAD et un F1-score de 65,4% sur la tâche d'OSD sur les données d'évaluation. Les résultats montrent que l'utilisation des filtres appris ne permet pas d'égaler les performances du modèle original, basé sur la TFCT. L'analyse fréquentielle du banc de filtres après apprentissage montre



(a) N_f = 32

(b) N_f = 64

(c) N_f = 128

(d) N_f = 256

FIGURE 5.5 – Module de la réponse en fréquence des filtres analytiques appris. Ces derniers sont initialisés par des filtres rectangulaires de fréquences centrales linéairement espacées et une bande passante de largeur Δf = 100 Hz. N_f : nombre de filtres appris.

que les filtres sont restreints dans la bande [0,600] Hz. Il est supposé que la dégradation est issue du manque de filtres dans les bandes de fréquence supérieures. Afin de vérifier cette hypothèse, une procédure d'initialisation est explorée. Elle consiste à initialiser les poids des couches convolutives (filtres) par la réponse impulsionnelle d'un filtre rectangulaire. Cette approche ne permet pas de combler l'écart de performance avec le modèle TFCT. Elle permet cependant une meilleure robustesse des performances en fonction du nombre de filtres.

L'apprentissage de filtres analytiques pour la tâche d'OSD ne permet pas d'apprendre une grande diversité de filtres. Ces derniers convergent vers les basses fréquences. L'ajout d'une étape de reconstruction du signal d'entrée pourrait permettre d'améliorer la bande passante du banc de filtres. Il pourrait également être envisagé de reconstruire le signal issu du *headset-mix* afin de réaliser une étape de réhaussement du signal en parallèle de l'OSD.

La combinaison auto-attentive de canaux ne semble pas adaptée au traitement des signaux dans le domaine temporel avec l'approche proposée. L'intégration de la phase à l'aide de filtres fréquentiels optimisés ne permet donc pas d'améliorer les performances de segmentation de la

parole distante multicanale. La section suivante propose une seconde approche permettant de modéliser la phase du signal en estimant des poids de combinaison dans le domaine complexe.

5.4 Intégration de la phase de la transformée de Fourier

Le modèle SACC original ne prend pas en compte la phase du signal. Ce paramètre permet cependant de modéliser le retard entre les signaux issus de chaque microphone et pourrait être bénéfique pour la combinaison de canaux dans le contexte de la segmentation de la parole distante. La section précédente montre que l'estimation de poids de combinaison dans le domaine temporel ne permet pas d'obtenir des performances satisfaisantes. Nous proposons une seconde approche d'intégration de la phase, en conservant les parties réelle et imaginaire (respectivement le module et la phase) de la TFCT pour l'estimation des poids de combinaison. Les poids de combinaison sont donc complexes et se rapprochent des algorithmes de formation de voies classiques (ex : MVDR).

5.4.1 Formulation complexe du modèle SACC

Les extensions complexes du modèle SACC sont dénotées cSACC. Deux formulations sont proposées pour prendre en compte la phase de la TFCT. La première modélise séparément le module et la phase (respectivement la partie réelle et la partie imaginaire) de la transformée de Fourier. Ce modèle, décrit comme explicite, est noté EcSACC. La seconde modélise implicitement le module et la phase. Il est nommé IcSACC. Le choix de la formulation du modèle (parties réelles/imaginaires ou module/phase) a été réalisé en fonction des performances obtenues sur la tâche de détection de parole superposée. Cette étude est présentée en annexe A.

Formulation explicite (EcSACC)

Le modèle EcSACC calcule la partie réelle et la partie imaginaire des poids de combinaison indépendamment à partir de la TFCT. Le module et la phase peuvent également être utilisés. Soit $||\mathbf{X}_{tfct}||$ et $\angle\mathbf{X}_{tfct}$ le module et la phase de la TFCT du signal d'entrée. Les poids de combinaison associés à chaque partie sont calculés par la relation suivante :

$$\begin{aligned} |\mathbf{w}| &= \mathrm{SA}_{||}\left(|\mathbf{X}_{tfct}|^2\right), \\ \mathbf{w}_\phi &= \mathrm{SA}_\phi\left(\sin(\angle\mathbf{X}_{tfct})\right), \end{aligned} \tag{5.7}$$

avec $\mathrm{SA}_{||}$ et SA_ϕ les modules d'auto-attention définis en équation (5.2) appliqués respectivement au module et à la phase de la TFCT. La phase est représentée sur un cercle. Cette propriété rend cette fonction continue par morceaux. La discontinuité de la phase peut être difficile à modéliser par un RNA. Une fonction sinusoïdale permet d'encoder la phase afin d'en obtenir une

représentation continue. Un jeu de poids de combinaison $|\mathbf{w}|$ est donc appliqué au module et un autre, $\mathbf{w}_\phi$, à la phase.

Chaque jeu de poids est appliqué à la partie de la TFCT associée puis les canaux sont combinés :

$$\mathbf{Y}_{att} = \sum_{m=1}^{M} \text{softmax}\left(|\mathbf{w}|\right) \odot |\mathbf{X}_{TFCT}| e^{j\text{softmax}(\mathbf{w}_\phi) \odot \angle \mathbf{X}_{tfct}}. \tag{5.8}$$

avec $\odot$ le produit terme-à-terme tel qu'un poids $|w|_t$ à la trame t est appliqué à toutes les fréquences du spectrogramme à la même trame $|\mathbf{X}_t|$. La fonction softmax est appliquée sur la dimension des canaux et permet de garantir $|w|_{t,m}, w_{\phi,t,m} \in [0, 1]$. La somme est appliquée sur la dimension des canaux et permet d'obtenir le spectrogramme moyenné $\mathbf{Y}_{att} \in \mathbb{C}^{T \times F}$.

Après combinaison des canaux, seul le module de la TFCT est conservé puis converti en échelle Mel à l'aide de $F = 64$ filtres. L'architecture pour le calcul des poids EcSACC est présentée en figure 5.6. Le module et la phase peuvent être remplacés par les parties réelle et imaginaire de la TFCT (*cf.* Annexe A).

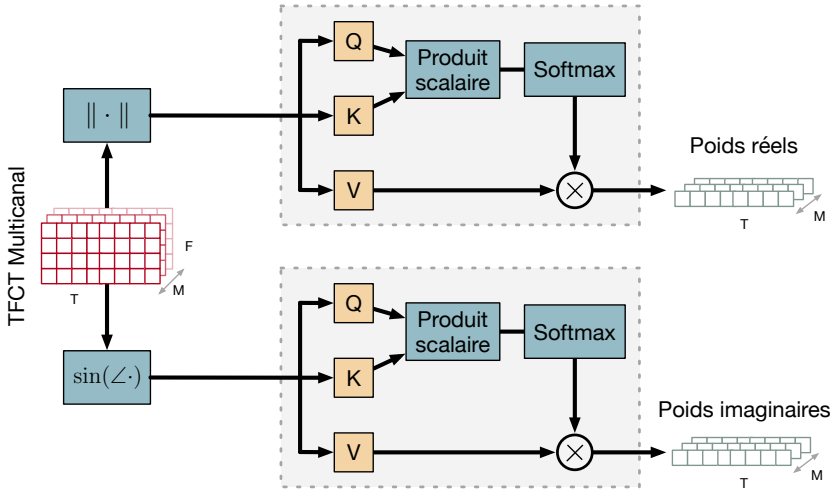


FIGURE 5.6 – Calcul des poids de combinaison à l'aide de l'algorithme SACC complexe explicite. Le module et la phase peuvent être remplacés par les parties réelle et imaginaire.

Formulation implicite (IcSACC)

Le modèle EcSACC ne permet pas d'apprendre une représentation croisée entre le module et la phase. De plus, il nécessite l'apprentissage de deux modules d'auto-attention et double ainsi le nombre de paramètres du module d'extraction de caractéristiques. Le modèle implicite IcSACC est proposé pour permettre au modèle d'apprendre une représentation croisée sans doubler le nombre de paramètres. Pour cela, un module d'auto-attention unique prend en entrée la concaténation du module et de la phase de la TFCT sur la dimension des canaux :

$$\mathbf{w} = \text{SA}\left(\text{cat}(|\mathbf{X}_{TFCT}|^2, \sin(\angle \mathbf{X}_{TFCT}))\right), \tag{5.9}$$

avec $\mathrm{cat}(\cdot,\cdot)$ l'opérateur de concaténation. Le vecteur de sortie $\mathbf{w}$ contient la concaténation du module et de la phase des poids de combinaison. La combinaison des canaux est ensuite réalisée à l'aide de la relation (5.8). De façon similaire à EcSACC, seul le module de la TFCT après combinaison est conservé puis converti à l'échelle Mel avec $F = 64$ filtres. L'architecture du modèle IcSACC est présentée en figure 5.7. Comme pour le modèle EcSACC, le module et la phase peuvent être remplacés par les parties réelle et imaginaire (*cf.* annexe A).

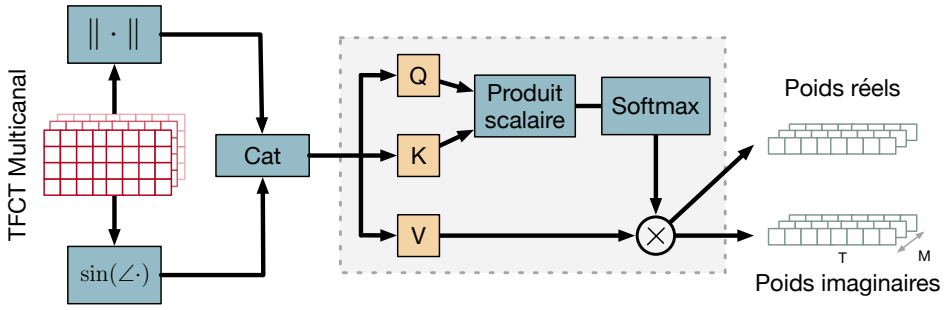


FIGURE 5.7 – Calcul des poids de combinaison à l'aide de l'algorithme SACC complexe implicite. Le module et la phase peuvent être remplacés par les parties réelle et imaginaire.

5.4.2 Performances de segmentation

Les extensions complexes du modèle SACC sont évaluées sur les tâches de VAD et d'OSD. Les modèles IcSACC et EcSACC proposés sont comparés aux modèles de référence et à l'approche SACC originale.

VAD

Les résultats des extensions complexes sur la tâche de VAD sont présentés dans la table 5.7. La première partie de la table présente les performances de chaque modèle au sein de l'architecture BLSTM. L'approche IcSACC échoue à améliorer les performances de détection avec un SER de 6,55% et 7,10% respectivement sur les données de développement et d'évaluation. Le modèle EcSACC permet cependant l'obtention de performances similaires au modèle SACC sur les données de développement (6,00%) et d'évaluation (6,58%).

La seconde partie de la table 5.7 présente les performances de VAD obtenues à l'aide de chaque algorithme d'extraction de caractéristiques au sein d'un système TCN. Les résultats obtenus à l'aide du modèle IcSACC (Dev : 6,19%, Eval : 6,78%) montrent que cette approche n'améliore pas la détection en conditions distantes par rapport au MDU (Dev : 6,57%, Eval : 6,95%). L'approche explicite EcSACC permet cependant une amélioration significative (Dev : 5,87%, Eval : 6,39%) en atteignant des performances similaires aux caractéristiques MVDR et SACC.

TABLE 5.7 – Détection d'activité vocale à l'aide des modèles IcSACC et EcSACC sur les données de développement et d'évaluation du corpus AMI. Les scores en gras indiquent les meilleurs systèmes distants pour chaque modélisation de séquence. Les valeurs en italiques correspondent à la référence en champ proche.

		FA$_{\%\downarrow}$		**Miss**$_{\%\downarrow}$		**SER**$_{\%\downarrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
BLSTM	CT MFCC	*3.32*	*3.12*	*2.36*	*2.61*	*5.68*	*5.70*
	MDU MFCC	4.07	2.08	2.80	6.29	6.88	8.38
	MVDR	3.65	3.07	2.45	3.66	**6.09**	**6.73**
	SACC TFCT	3.52	3.07	2.43	3.55	**5.95**	**6.62**
	IcSACC	4.14	3.45	2.41	3.65	6.55	7.10
	EcSACC	3.75	3.19	2.25	3.39	**6.00**	**6.58**
TCN	CT MFCC	*3.14*	*2.41*	*2.69*	*3.4*	*5.83*	*5.85*
	MDU MFCC	4.76	2.95	1.82	3.99	6.57	6.95
	MVDR	3.15	2.65	2.59	3.78	**5.74**	**6.43**
	SACC TFCT	3.51	3.08	2.30	3.39	**5.81**	**6.47**
	IcSACC	3.96	3.70	2.23	3.09	6.19	6.78
	EcSACC	3.78	3.23	2.09	3.16	**5.87**	**6.39**

OSD

Les résultats des extensions complexes sur la tâche d'OSD sont présentés dans la table 5.8. La première partie de la table présente les résultats obtenus au sein de l'architecture BLSTM. Ici encore, le modèle IcSACC ne permet pas d'améliorer les performances de détection, avec un F1-score de 63,9% et de 59,0% sur les données de développement et d'évaluation respectivement. Ces résultats sont équivalents au microphone distant (Dev : 63,7%, Eval : 59,5%). D'autre part, le modèle EcSACC permet une légère amélioration des performances (Dev : 66,8%, Eval : 63,7%) par rapport au MDU. Ces scores de détection restent cependant inférieurs à ceux obtenus à l'aide de l'algorithme original SACC.

La seconde partie de la table 5.8 présente les performances d'OSD de chaque méthode d'extraction de caractéristiques au sein d'un système TCN. Les résultats montrent que le modèle IcSACC (Dev : 69,1 ; Eval : 64,8) n'améliore pas les performances de détection. Celles-ci sont même dégradées par rapport au MDU. Seule une légère amélioration est visible sur les données de développement, toujours par rapport au MDU.

Le modèle EcSACC s'approche du modèle SACC sur les données de développement avec un F1-score de 71,7%. Cela représente quand même une dégradation absolue de -0,7% par rapport à SACC. Il est cependant moins robuste sur les données d'évaluation sur lesquelles un F1-score de 66,7% est obtenu, soit une dégradation absolue de -2,1% par rapport à SACC. Cette baisse de performance est liée à un écart important entre la précision (64,0%) et le rappel (70,0%) pour les données de développement. Le modèle est donc plus sensible aux seuils de détection choisis.

TABLE 5.8 – Détection de parole superposée à l'aide des modèles IcSACC et EcSACC sur les données de développement et d'évaluation du corpus AMI. Les scores en gras indiquent les meilleurs systèmes distants pour chaque modélisation de séquence. Les valeurs en italiques correspondent à la référence en champ proche.

		Précision$_{\%\uparrow}$		**Rappel**$_{\%\uparrow}$		**F1-score**$_{\%\uparrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
BLSTM	CT MFCC	*68.9*	*76.9*	*71.8*	*60.9*	*70.3*	*67.9*
	MDU MFCC	64.4	65.7	63.0	54.3	63.7	59.5
	MVDR	68.0	71.9	69.6	59.9	**68.8**	**65.3**
	SACC TFCT	68.5	69.9	69.9	61.4	**69.2**	**65.3**
	IcSACC	61.9	70.2	66.0	50.9	63.9	59.0
	EcSACC	68.8	72.5	64.9	56.8	66.8	63.7
TCN	CT MFCC	*73.8*	*81.4*	*74.1*	*61.8*	*73.9*	*70.2*
	MDU MFCC	69.5	72.8	68.1	59.5	68.8	65.5
	MVDR	74.3	73.1	70.0	66.2	**72.1**	**69.5**
	SACC TFCT	72.5	72.9	72.3	65.2	**72.4**	**68.8**
	IcSACC	69.4	74.5	68.6	57.4	69.1	64.8
	EcSACC	72.2	64.0	71.2	70.0	**71.7**	66.7

Conclusions

Cette section introduit deux extensions du modèle SACC dans le domaine complexe. Elles consistent à utiliser l'intégralité de la TFCT pour calculer des poids de combinaison complexes afin de conserver l'information de phase. La première extension explorée, IcSACC, modélise le module et la phase de la TFCT à l'aide d'un unique module d'auto-attention. Celui-ci est alimenté par la concaténation des module et phase de la TFCT. La modélisation implicite ne permet d'améliorer ni les performances de détection d'activité vocale, ni de parole superposée en conditions distantes.

La seconde extension proposée, EcSACC, modélise les module et phase séparément, à l'aide de deux modules d'auto-attention. Ce modèle permet d'améliorer les détections d'activité vocale et de parole superposée en conditions distantes en offrant des performances proches de SACC.

Remarque importante

En plus d'intégrer l'information de la phase, les poids de combinaison complexes doivent permettre une meilleure interprétation de l'extraction de caractéristiques. Ils peuvent être considérés comme un filtre spatial linéaire dont la réponse spatiale se calcule. Cependant, l'équation (5.7) ne correspond pas à un produit complexe. De fait, la transformation appliquée à la TFCT par les poids de combinaison n'est pas linéaire. Dans ce contexte, les poids ne peuvent donc pas être analysés comme un filtre spatial linéaire, réduisant ainsi les possibilités d'interprétation des modèles EcSACC et IcSACC.

La section suivante propose une nouvelle formulation du modèle EcSACC en réalisant un

produit complexe entre les poids de combinaison et la TFCT. Dans ce cas, les poids peuvent être interprétés comme un filtre spatial. Une analyse de ces derniers est donc proposée.

5.5 Extension complexe linéaire pour l'interprétation des poids de combinaison

Cette section présente une autre formulation de cSACC, appliquant un produit complexe entre les poids appris par le modèle et la TFCT multicanale du signal. Ce modèle, dénoté LcSACC, permet d'analyser les poids de combinaison comme ceux d'un filtre spatial et donne accès à une meilleure interprétation du système. Le formalisme est d'abord introduit (section 5.5.1) avant de présenter les performances de segmentation (section 5.5.2) puis l'analyse de la réponse spatiale du modèle (section 5.5.3).

5.5.1 Formalisation du modèle LcSACC

La procédure d'estimation des poids est identique à l'architecture EcSACC proposée en section 5.4. Deux formulations sont considérées. La première utilise les parties réelle et imaginaire ($\mathfrak{R}/\mathfrak{I}$) de la TFCT. La seconde exploite la magnitude et la phase (Mag/ϕ). La phase peut également être encodée à l'aide d'une fonction sinus. Ce type de modèle est dénoté $\sin\phi$.

Formulation $\mathfrak{R}/\mathfrak{I}$

Pour la formulation $\mathfrak{R}/\mathfrak{I}$, le modèle utilise deux mécanismes d'auto-attention SA traitant séparément les parties réelle et imaginaire de la TFCT :

$$\begin{aligned} \mathbf{w}_{\mathfrak{R}} &= \mathrm{SA}_{\mathfrak{R}}(\mathbf{X}_{\mathfrak{R}}), \\ \mathbf{w}_{\mathfrak{I}} &= \mathrm{SA}_{\mathfrak{I}}(\mathbf{X}_{\mathfrak{I}}). \end{aligned} \tag{5.10}$$

avec $\mathbf{w}_{\mathfrak{R}}, \mathbf{w}_{\mathfrak{I}} \in \mathbb{R}^{M\times 1\times T}$ les vecteurs de poids de combinaison estimés sur les parties réelle et imaginaire respectivement. Le vecteur de poids complexes $\mathbf{w}$ obtenu en sortie de ce modèle peut donc s'écrire

$$\mathbf{w} = \mathbf{w}_{\mathfrak{R}} + j\mathbf{w}_{\mathfrak{I}}. \tag{5.11}$$

La combinaison des canaux est réalisée par le produit complexe entre les poids $\mathbf{w}$ et la TFCT. Pour une trame t donnée, ce produit s'écrit :

$$\begin{aligned}\mathbf{X}_{att} &= \mathbf{w}_t \mathbf{X}_t \\ &= \sum_{m=1}^{M} \left(w_{\Re,m} X_{\Re,m} - w_{\Im,m} X_{\Im,m}\right) + j\left(w_{\Re,m} X_{\Im,m} + w_{\Im,m} X_{\Re,m}\right).\end{aligned} \tag{5.12}$$

L'indice t est omis sur la dernière ligne pour alléger les notations. La version linéaire du modèle cSACC combine les canaux à l'aide de l'équation (5.12).

Formulation Mag/ϕ

La seconde formulation du système, modélisant la magnitude et la phase de la TFCT, est similaire à la précédente. Dans ce cas, les modules d'attention estiment les poids de combinaison complexes à partir de la magnitude $|\mathbf{X}|$ et de la phase $\angle\mathbf{X}$ de la TFCT :

$$\begin{aligned}|\mathbf{w}| &= \mathrm{SA}_{||}(|\mathbf{X}|), \\ \mathbf{w}_\phi &= \mathrm{SA}_\phi(\angle\mathbf{X}).\end{aligned} \tag{5.13}$$

Les poids $\mathbf{w}_\phi$ peuvent également être estimés à partir du sinus de la phase de la TFCT $\sin\angle\mathbf{X}$. Les poids de combinaison de la formulation magnitude/phase s'écrivent

$$\mathbf{w} = \mathbf{w}_{|||} e^{j\mathbf{w}_\phi}. \tag{5.14}$$

Dans ce cas de figure, la pondération et la combinaison des canaux s'expriment par la relation suivante :

$$\mathbf{X}_{att} = \sum_{m=1}^{M} |w_m||X_m| e^{j(2\pi w_{\phi,m} + \angle X_m)}. \tag{5.15}$$

La fonction softmax est appliquée sur la dimension des canaux et permet de garantir $|w|_{t,m}, w_{\phi,t,m} \in [0,1]$. La somme est appliquée sur la dimension des canaux et permet d'obtenir le spectrogramme moyenné $\mathbf{X}_{att} \in \mathbb{C}^{T\times F}$. Les poids vérifient $\mathbf{w}_\phi \in [0,1]$ suite à la normalisation par la fonction softmax au sein du module d'auto-attention (équation (5.3)). Un facteur 2π est donc appliqué à la phase des poids afin de lui permettre de parcourir l'intégralité du cercle trigonométrique. Notons que l'équation (5.15) n'est pas équivalente à la première formulation du modèle EcSACC donnée en équation (5.7). Ici, le produit entre les poids est linéaire et permet l'analyse des poids à l'aide de la réponse spatiale. Cette étude est présentée en section 5.5.3.

Dépendance des poids à la fréquence

Les poids estimés dans la première version du modèle cSACC (section 5.4) sont indépendants de la fréquence. La même pondération est donc appliquée à toutes les fréquences de la TFCT. Dans le cas magnitude/phase, cela implique que la même correction est appliquée à chaque fréquence. Il peut cependant être intéressant de rendre cette correction dépendante de la fréquence, comme dans le cas d'un filtre spatial classique (Benesty et al. 2008).

Afin de rendre les poids dépendants de la fréquence, une légère modification est appliquée au modèle d'auto-attention. La requête et la clef conservent les mêmes dimensions $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{T\times M\times D}$. La valeur est modifiée et conserve la dimension des fréquences telle que $\mathbf{V} \in \mathbb{R}^{T\times M\times F}$. Les poids calculés à l'aide de l'auto-attention (Eq. (5.2)) conservent alors la dimension des fréquences tels que $\mathbf{w} \in \mathbb{R}^{T\times M\times F}$.

5.5.2 Performances de segmentation

Cette sous-section évalue les performances de l'approche LcSACC sur les tâches de VAD et d'OSD. Les différentes formulations possibles du modèle sont étudiées sur chaque tâche afin d'identifier la meilleure approche.

Détection d'activité vocale

Influence de la formulation La table 5.9 présente les résultats obtenus sur la tâche de VAD pour le système LcSACC suivi d'un TCN. Les résultats sont obtenus pour chaque formulation du modèle (*cf.* équations (5.12) et (5.15)). Chaque modèle est évalué avec et sans conservation des fréquences dans le module d'attention. En cas de conservation, $F = 257$ fréquences sont considérées.

Table 5.9 – Résultats obtenus sur la tâche de VAD à l'aide du modèle LcSACC pour chaque configuration sur les données de développement et d'évaluation du corpus AMI. (✗) les poids de combinaison ne dépendent pas de la fréquence. (✓) les poids dépendent de la fréquence.

		FA$_{\%\downarrow}$		**Miss**$_{\%\downarrow}$		**SER**$_{\%\downarrow}$	
Formulation	Fréquence	Dev	Eval	Dev	Eval	Dev	Eval
$\mathfrak{R}/\mathfrak{I}$	✗	3,84	2,74	2,06	3,85	5,91	6,59
$\mathfrak{R}/\mathfrak{I}$	✓	3,56	2,83	2,42	3,88	5,97	6,71
Mag/ϕ	✗	3,33	3,34	2,44	2,91	5,77	**6,26**
Mag/ϕ	✓	3,73	3,26	2,11	3,12	5,84	6,38
Mag/sin ϕ	✗	3,50	3,69	2,18	2,72	**5,68**	6,41
Mag/sin ϕ	✓	3,65	2,67	2,16	3,95	5,81	6,61

Dans un premier temps, la table 5.9 montre que la conservation des fréquences au sein du module d'auto-attention n'améliore pas les performances. Par exemple, pour la formulation

Mag/ϕ, le système atteint un SER de 6,26% dans le cas (✗) contre 6,38% dans le cas (✓) sur le sous-ensemble d'évaluation. Cette formulation permet d'obtenir les meilleures performances de VAD avec ce système sur le sous-ensemble d'évaluation. L'approche $\mathfrak{R}/\mathfrak{I}$ obtient un SER de 6,59%, légèrement surpassée par la formulation Mag/sin ϕ obtenant 6,41%. L'approche Mag/sin ϕ offre cependant les meilleurs résultats sur les données de développement, mais présente des difficultés de généralisation.

Comparaison aux autres méthodes La table 5.10 présente les résultats de VAD du modèle LcSACC par rapport aux approches préalablement présentées. Le système LcSACC permet d'obtenir des performances similaires. Il permet notamment d'atteindre les meilleures performances en conditions de parole distante sur les données d'évaluation avec un SER de 6,26%. Sur les données de développement, LcSACC (5,77%) obtient des performances proches du modèle de référence MVDR (5,74%). Ce nouveau formalisme permet également une meilleure généralisation entre les sous-ensembles de développement et d'évaluation avec une dégradation relative de -8,5%. À titre de comparaison, le modèle SACC présente une dégradation relative de -11,4% entre les deux sous-ensembles. LcSACC semble donc être un candidat intéressant pour la VAD en conditions distantes.

Table 5.10 – Comparaison des performances de VAD du système LcSACC avec les approches préalablement étudiées. Les résultats sont obtenus sur les données de développement et d'évaluation du corpus AMI. Les scores en gras indiquent les meilleures performances. Les valeurs en italiques correspondent à la référence en champ proche.

		FA$_{\%\downarrow}$		**Miss**$_{\%\downarrow}$		**SER**$_{\%\downarrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
TCN	CT MFCC	*3.14*	*2.41*	*2.69*	*3.4*	*5.83*	*5.85*
	MDU MFCC	4.76	2.95	1.82	3.99	6.57	6.95
	MVDR	3.15	2.65	2.59	3.78	**5.74**	6.43
	SACC TFCT	3.51	3.08	2.30	3.39	5.81	6.47
	EcSACC	3.78	3.23	2.09	3.16	5.87	6.39
	LcSACC Mag/ϕ	3,33	3,34	2,44	2,91	5,77	**6,26**

Conclusion Le modèle LcSACC offre des performances de VAD intéressantes, notamment avec la formulation Mag/ϕ qui offre les meilleures performances. La dépendance du module d'attention à la fréquence n'est pas favorable à la tâche de VAD. Cette dernière atteint également les meilleurs scores de VAD sur le sous-ensemble d'évaluation du corpus AMI.

Détection de parole superposée

Influence de la formulation La table 5.11 présente les performances de chaque formulation du modèle LcSACC sur la tâche d'OSD. La précision, le rappel et le F1-score sont reportés

pour les sous-ensembles de développement et d'évaluation du corpus AMI. Chaque système est entraîné et évalué avec (✓) et sans (✗) prise en compte des fréquences dans le calcul des poids de combinaison. En cas de prise en compte, nous fixons $F = 257$.

La première partie du tableau présente les performances du système $\mathfrak{R}/\mathfrak{I}$. Sans considérer les fréquences (✗), le modèle obtient un F1-score de 70,8% sur les données de développement et 65,0% en évaluation. Ce modèle limite la généralisation, avec une dégradation absolue de -5,8% entre le développement et l'évaluation. Cette dégradation est principalement liée à une chute du rappel (57,4%). L'ajout de la dépendance aux fréquences (✓) offre un F1-score similaire sur le Dev (70,6%) et limite la dégradation sur l'Eval avec 66,1%. L'ajout des fréquences semble donc améliorer légèrement la robustesse du modèle.

La deuxième partie du tableau présente les résultats de la même expérience avec le système Mag/ϕ. Dans le cas où les fréquences ne sont pas prises en compte (✗), le modèle offre les performances les plus basses avec 69,8% (Dev) et 64,3% (Eval). L'ajout des fréquences (✓) améliore significativement les performances, notamment sur les données d'évaluation avec 67,9%, soit un gain absolu de +3,6% par rapport au modèle précédent (✗). Cette configuration offre les meilleures performances d'OSD.

Table 5.11 – Résultats obtenus pour la tâche d'OSD à l'aide du modèle LcSACC pour chaque configuration sur les données de développement et d'évaluation du corpus AMI.

		Précision$_{\%\uparrow}$		**Rappel**$_{\%\uparrow}$		**F1-score**$_{\%\uparrow}$	
Formulation	Fréquence	Dev	Eval	Dev	Eval	Dev	Eval
$\mathfrak{R}/\mathfrak{I}$	✗	71,7	75,0	70,0	57,4	70,8	65,0
$\mathfrak{R}/\mathfrak{I}$	✓	71,7	73,9	69,6	59,8	70,6	66,1
Mag/ϕ	✗	72,5	73,1	67,4	57,4	69,8	64,3
Mag/ϕ	✓	69,1	73,0	72,1	63,5	70,6	**67,9**
Mag/sin ϕ	✗	71,0	75,8	71,7	59,9	71,3	66,9
Mag/sin ϕ	✓	70,3	74,0	71,8	59,8	71,1	66,2

La troisième partie du tableau présente les performances du modèle avec encodage de la phase Mag/sin ϕ. Avec cette approche, les performances sont similaires sans (✗) et avec (✓) l'utilisation des fréquences avec 66,9% et 66,2% de F1-score d'évaluation respectivement. Ici encore, les performances sont fortement dégradées entre le développement et l'évaluation.

Comparaison aux autres méthodes La table 5.12 présente les performances du modèle LcSACC sur la tâche d'OSD par rapport aux modèles préalablement présentés. LcSACC améliore les performances par rapport au MDU avec un gain absolu de +2,4% sur les données d'évaluation. Ce modèle est également un peu plus performant que EcSACC avec un gain absolu de +1,2% sur les mêmes données. Cependant, LcSACC ne permet pas de surpasser la formation de voies MVDR, ni SACC TFCT.

TABLE 5.12 – Performances d'OSD du modèle LcSACC par rapport aux autres approches. Les résultats sont obtenus sur les sous-ensembles de développement et d'évaluation du corpus AMI. Les scores en gras indiquent les meilleures performances avec chaque modélisation de séquence. Les valeurs en italiques correspondent à la référence en champ proche.

		Précision$_{\%\uparrow}$		**Rappel**$_{\%\uparrow}$		**F1-score**$_{\%\uparrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
TCN	CT MFCC	*73,8*	*81,4*	*74,1*	*61,8*	*73,9*	*70,2*
	MDU MFCC	69,5	72,8	68,1	59,5	68,8	65,5
	MVDR	74,3	73,1	70,0	66,2	**72,1**	**69,5**
	SACC TFCT	72,5	72,9	72,3	65,2	**72,4**	**68,8**
	EcSACC	72,2	64,0	71,2	70,0	**71,7**	66,7
	LcSACC	69,1	73,0	72,1	63,5	70,6	**67,9**

Conclusion Le modèle LcSACC applique un produit complexe linéaire entre un jeu de poids de combinaison, appris à l'aide de l'auto-attention, et la TFCT du signal. L'intégration des composantes réelles et imaginaires de la TFCT ne permet pas d'améliorer la détection de parole superposée par rapport à l'approche originale SACC. Les résultats sont cependant meilleurs que l'approche MDU utilisant uniquement les MFCC. Sur la tâche de VAD, le modèle LcSACC obtient les meilleures performances sur les signaux distants avec un SER de 6,26% sur les données d'évaluation.

Bien que l'extension du modèle SACC dans le domaine complexe n'améliore pas toujours la détection, cette approche permet de visualiser la réponse spatiale du système. Les poids LcSACC sont en effet utilisés comme un filtre spatial linéaire dont la réponse spatiale peut donc être calculée. La section suivante présente l'analyse de la réponse spatiale du modèle LcSACC sur les données de développement du corpus AMI.

5.5.3 Visualisation de la réponse spatiale

L'extension complexe du modèle SACC avec LcSACC permet d'analyser les poids de combinaison comme ceux d'un filtre spatial classique. Cette sous-section présente l'analyse de la réponse spatiale du modèle LcSACC calculée à partir des poids de combinaison complexes. L'analyse est réalisée sur les données de développement du corpus AMI.

Réponse spatiale

La réponse spatiale est un outil d'analyse des algorithmes de formation de voies. Elle consiste à évaluer la réponse d'un filtre spatial en fonction de la direction d'arrivée d'une onde acoustique.

Soient $\mathbf{w} \in \mathbb{R}^{M \times F}$ les coefficients d'un filtre spatial appliqués à chaque canal $m = 1, ..., M$ d'une antenne. Dans le cas d'une ACU, la réponse spatiale du filtre à une onde plane arrivant dans la direction θ s'exprime (BENESTY et al. 2015) :

$$\mathcal{B}_{t,f}[\theta] = \sum_{m=1}^{M} w_{m,t} e^{j\bar{\omega}\cos(\theta-\psi_m)}, \tag{5.16}$$

avec $\psi_m \in [0, 2\pi)$ l'angle du microphone m (*cf.* figure 3.2) et $\bar{\omega} = 2\pi r f / v_s$ où v_s représente la vitesse du son, r le rayon de l'antenne, f la fréquence et t l'indice de la trame [1]. Choisir un ensemble de directions d'arrivée $\theta \in [0, 2\pi)$ permet d'obtenir la réponse spatiale du filtre spatial. Dans le cas où les coefficients (ou poids de combinaison) sont appris par LcSACC, la réponse spatiale définie en équation (5.16) est calculée à chaque trame. Afin d'analyser les directions de focalisation globales du système, nous calculons la moyenne temporelle du module de la réponse spatiale :

$$\hat{\mathcal{B}}_f[\theta] = \frac{1}{T} \sum_{t=1}^{T} \left\| \mathcal{B}_{t,f}[\theta] \right\|^2. \tag{5.17}$$

La réponse spatiale permet d'analyser les directions angulaires sélectionnées par les mécanismes d'attention du modèle LcSACC. La sous-section suivante propose des exemples de visualisation sur des données réelles.

Visualisation des poids par rapport à l'énergie acoustique

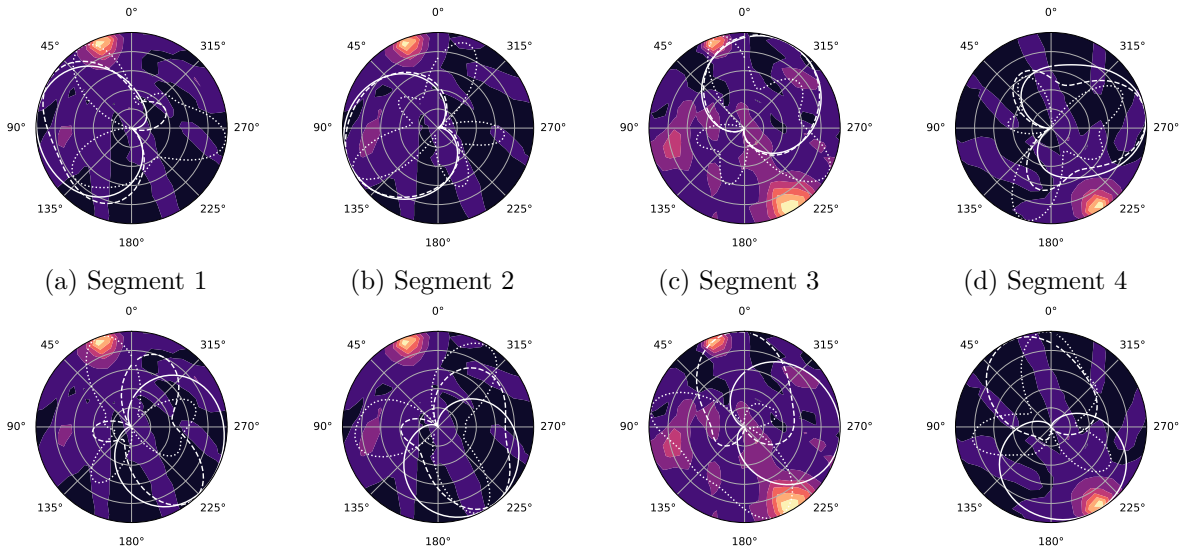
La figure 5.8 présente la réponse spatiale moyennée, calculée avec l'équation (5.17), obtenue à partir des poids des modèles LcSACC $\mathfrak{R}/\mathfrak{I}$, avec et sans dépendance en fréquence. La réponse est calculée sur des segments de deux secondes, sans recouvrement. Elle est représentée pour trois fréquences : 450 Hz, 750 Hz et 1500 Hz. La carte d'énergie acoustique est également représentée. Celle-ci est obtenue avec l'algorithme SRP-PHAT et permet de visualiser la concordance entre les directions de focalisation du modèle et les maxima d'énergie.

Les figures 5.8a à 5.8d présentent la réponse spatiale moyenne du modèle ne dépendant pas de la fréquence ($F = 1$). Bien que les directions de focalisation du modèle varient d'un segment à l'autre, celles-ci ne semblent pas corrélées à la position des maxima d'énergie. Ce comportement est observé pour chaque segment.

Les figures 5.8e à 5.8h présentent la réponse spatiale moyenne du modèle ne dépendant pas de la fréquence ($F = 257$). Les observations sont similaires dans ce cas de figure. La corrélation entre la position des sources et les directions choisies semble cependant plus marquée. Par exemple, en figure 5.8e, la réponse spatiale est orientée en direction de la source à fréquence $f = 1500$ Hz. En figure 5.8g, la direction de la première source est sélectionnée à $f = 750$ Hz et le lobe à $f = 450$ Hz semble s'approcher de la seconde source.

Bien que la corrélation entre la position des sources et les directions de focalisation du modèle LcSACC semble faible, les poids complexes permettent une visualisation des poids de combinaison ayant un sens physique. Dans le cas étudié, les directions sélectionnées varient en fonction de la

1. Dans l'équation proposée par Benesty et al. (2015), les poids du filtre sont conjugués. Dans notre cas, les poids ne sont pas conjugués, car le modèle estime directement les poids appliqués à la TFCT (*cf.* équation 5.12).



(a) Segment 1 (b) Segment 2 (c) Segment 3 (d) Segment 4

(e) Segment 1 (f) Segment 2 (g) Segment 3 (h) Segment 4

FIGURE 5.8 – Module moyen de la réponse spatiale pour quatre segments successifs de la session IS1008a du sous-ensemble de développement du corpus AMI. Les réponses de deux modèles sont comparées : (a)-(d) LcSACC sans dépendance en fréquence, (e)-(h) LcSACC avec dépendance en fréquence. Pour chaque représentation, la réponse spatiale est représentée pour trois fréquences : (–) f = 450 Hz, (- -) f = 750 Hz et (•••) f = 1500 Hz.

position des sources, bien que les maxima ne correspondent pas aux positions des sources. Cela semble montrer que le modèle LcSACC exploite l'information spatiale.

5.5.4 Conclusions et perspectives

Conclusions Cette section présente une extension du modèle SACC dans le domaine complexe. Les poids appris par le modèle sont composés d'une partie réelle et d'une partie imaginaire, estimées par deux modules d'attention. La pondération et la combinaison des canaux est réalisée de façon similaire à un filtre spatial (formation de voies). Cette considération permet notamment d'analyser les poids de combinaison en fonction de la direction d'arrivée des sources, et étudier ainsi les directions de focalisation du système.

Le modèle LcSACC proposé offre les meilleures performances de VAD sur les données d'évaluation. Sur la tâche d'OSD, les performances sont légèrement supérieures au cas MDU sans surpasser le modèle original SACC. Une étude est également menée sur la conservation des fréquences au sein du mécanisme d'attention. L'ajout des fréquences permet un léger gain sur la tâche d'OSD mais dégrade la VAD.

D'autre part, cette section montre que la réponse spatiale est un outil efficace pour visualiser les poids appris par le modèle. Cependant, la comparaison de la réponse spatiale des modèles LcSACC montre que les directions utilisées par le modèle semblent décorrélées des maxima d'énergie acoustique, c'est-à-dire de la position des sources actives. Une évaluation à plus grande

échelle pourrait permettre de quantifier le degré de corrélation entre les directions de focalisation et la position des sources.

Perspectives D'une part, le modèle LcSACC offre des performances limitées sur la tâche d'OSD. D'autre part, les poids appris par le système ne semblent pas être corrélés à la position des sources, limitant fortement l'interprétabilité du système.

Un premier axe pourrait être d'entraîner le module LcSACC à améliorer le signal de parole, tout en étant optimisé pour la segmentation. Le corpus AMI offre des signaux enregistrés en champ proche, pouvant servir de cible pour l'amélioration du signal. En contraignant LcSACC à améliorer le signal, il est possible que les directions de focalisation concordent avec la position des sources. LcSACC agirait ainsi comme un filtre spatial adaptatif.

Un second axe d'amélioration serait d'ajouter un mécanisme d'attention sur la dimension des fréquences (Horiguchi et al. 2022). Conserver les fréquences semblent en effet améliorer les performances de segmentation. Cependant, dans ces travaux, la transformation appliquée aux fréquences est une simple transformation linéaire. Une fois le modèle entraîné, la transformation reste la même et ne s'adapte pas au signal rencontré. L'ajout d'attention permettrait au modèle de sélectionner les fréquences utiles en fonction du signal observé.

5.6 Sélection de filtres spatiaux

La section 5.2 montre que le modèle SACC combine efficacement les canaux dans le domaine des fréquences. Bien que les poids soient calculés explicitement, il est difficile de les relier à des paramètres physiques tels que la position des sources. Les sections 5.4 et 5.5 proposent deux types d'extension du modèle SACC dans le domaine complexe. Les poids de combinaison appartiennent au domaine de la TFCT et permettent, dans le cas du modèle LcSACC, de visualiser les directions de focalisation du modèle. Ces représentations visuelles restent cependant difficiles à corréler à la position des sources actives, ce qui limite leur interprétation.

Cette section présente une nouvelle variante, appelée Beamforming Self-Attention Channel Combinator (BFSACC), permettant de sélectionner les sorties d'un ensemble de filtres spatiaux. Ces filtres extraient des signaux dans différentes directions angulaires. Un nouveau signal, dont chaque canal correspond à une direction, est obtenu. Le modèle SACC est appliqué à ce signal afin de sélectionner les signaux contenant le plus d'information pour la tâche visée. Les poids estimés par le modèle pour la sélection de filtre spatial sont alors directement liés aux directions angulaires et offrent un axe intéressant pour l'interprétation du modèle.

Cette section présente l'approche proposée en sous-section 5.6.1 et les résultats sur la tâche de segmentation sont donnés, illustrés en sous-section 5.6.2. Un protocole est défini en sous-section 5.6.3 pour interpréter les poids du modèle. Un axe d'amélioration du modèle, via la régularisation de la formation de voies, est étudié en sous-section 5.6.4.

5.6.1 Formulation du modèle BFSACC

Cette section présente les éléments constituant le système de sélection de filtre spatial BFSACC. L'architecture du modèle est présentée en figure 5.10.

Conception d'un banc de filtres spatiaux

Le modèle BFSACC consiste d'abord à appliquer un banc de P filtres spatiaux $\mathcal{S} = \{\mathbf{w}_p\}_{p=1}^{P}$ à la TFCT d'un signal multicanal $\mathbf{X} \in \mathbb{C}^{T\times M\times F}$. Chaque filtre p possède une direction de focalisation $\theta_p \in [0, 2\pi)$. Les poids de chaque filtre $\mathbf{w}_p \in \mathbb{C}^{M\times F}$ sont calculés à l'aide de la formation de voies *super-directive* à l'aide de la relation suivante (Minhua et al. 2019 ; Wölfel et al. 2009) :

$$\mathbf{w}_p^H(f) = \frac{\mathbf{v}_p^H(f)\mathbf{\Sigma}_N^{-1}(f)}{\mathbf{v}_p^H(f)\mathbf{\Sigma}_N^{-1}(f)\mathbf{v}_p(f)}, \tag{5.18}$$

avec f la fréquence, $\mathbf{v}_p(f) \in \mathbb{C}^{M\times 1}$ le vecteur de focalisation et $\mathbf{\Sigma}_N(f) \in \mathbb{R}^{M\times M}$ la matrice de covariance du bruit. Sous l'hypothèse d'un champ de bruit isotrope, considérée pour la formation de voies *super-directive* (Wölfel et al. 2009), un élément $\mathbf{\Sigma}_{Nm,n}(f)$ de cette matrice s'exprime :

$$\mathbf{\Sigma}_{Nm,n}(f) = \mathrm{sinc}(2\pi f d_{m,n}/c), \tag{5.19}$$

avec $d_{m,n}$ la distance entre deux microphones m et n. Comme ces travaux utilisent principalement une antenne de microphones circulaire uniforme (ACU), un élément $v_{p,m}$ du vecteur de focalisation de direction θ_p s'exprime (Benesty et al. 2015) :

$$v_{p,m}(f) = e^{j2\pi f r c^{-1}\cos(\theta_p - \psi_m)}, \tag{5.20}$$

avec m l'indice du microphone d'angle ψ_m, c la vitesse du son et r le rayon de l'antenne.

La figure 5.9 présente le module de la réponse spatiale, calculée à l'aide de l'équation (5.16), d'un banc de huit filtres spatiaux super-directifs. Ces réponses sont obtenues pour une ACU de géométrie identique au corpus AMI. Elle montre que le banc de filtres sélectionne efficacement la direction cible. Le faible nombre de microphones implique un repliement spatial important, visible dans les bandes de fréquences supérieures.

La sortie du filtre p à la fréquence f est obtenue par la relation suivante :

$$\mathbf{Y}_p(t, f) = \mathbf{w}_p^H(f)\mathbf{X}(t, f). \tag{5.21}$$

Le signal de sortie $\mathbf{Y}_p \in \mathbb{C}^{T\times F}$ obtenu en équation (5.21) correspond au signal focalisé dans la direction θ_p. Après filtrage du signal $\mathbf{X}$ par tous les filtres de $\mathcal{S}$, un nouveau signal multicanal $\mathbf{Y} \in \mathbb{R}^{T\times P\times F}$ est obtenu. La combinaison de canaux à l'aide du modèle SACC semble plus efficace lorsque seul le module de la TFCT est considéré (*cf.* sections 5.4 et 5.5). Seul le module des

signaux issus des filtres spatiaux sont donc conservés pour la sélection à l'aide de SACC :

$$\mathbf{Y} = \left[\|\mathbf{Y}_1\|^2, \cdots, \|\mathbf{Y}_p\|^2, \cdots, \|\mathbf{Y}_P\|^2 \right]. \tag{5.22}$$

Sélection de filtres spatiaux pour la tâche en aval

Le banc de filtres spatiaux $\mathcal{S}$ permet d'extraire plusieurs versions du signal enregistré par l'antenne. Chaque canal du signal $\mathbf{Y}$ encode l'information spatiale en étant associé à une direction spatiale θ_p.

D'autre part, le modèle SACC présente des capacités intéressantes pour sélectionner les canaux contenant le plus d'information pour une tâche aval spécifique (Gong et al. 2021 ; Mariotte et al. 2022 ; Sharma et al. 2022). Ce modèle est utilisé ici pour sélectionner le signal filtré contenant le maximum d'information. En notant $\mathbf{Q}, \mathbf{K} \in \mathbb{R}^{T \times P \times D}$ la requête et la clef, et $\mathbf{V} \in \mathbb{R}^{T \times P \times 1}$ la valeur, extraites du signal $\mathbf{Y}$, les poids d'auto-attention $\mathbf{w}_{att} \in \mathbb{R}^{T \times P \times P}$ sont estimés par la relation suivante :

$$\mathbf{w}_{att} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{D}} \right). \tag{5.23}$$

Les poids de combinaison $\mathbf{w} \in \mathbb{R}^{T \times P \times 1}$, appliqués au signal afin de sélectionner le signal filtré, sont ensuite calculés :

$$\mathbf{w} = \text{softmax}\left(\mathbf{w}_{att}\mathbf{V} \right). \tag{5.24}$$

La sélection du filtre spatial est réalisée par la somme pondérée des canaux. Un nouveau signal $\hat{\mathbf{Y}} \in \mathbb{R}^{T \times F}$ est ainsi obtenu :

$$\hat{\mathbf{Y}} = \sum_{p=1}^{P} \mathbf{w} \odot \mathbf{Y}. \tag{5.25}$$

La multiplication point-à-point $\odot$ est réalisée à la trame. Un poids w_t à une trame t est appliqué à toutes les fréquences $\mathbf{Y}_t$. La somme est appliquée sur la dimensions des canaux tels que $\hat{\mathbf{Y}} \in \mathbb{R}^{T \times F}$. Le signal $\hat{\mathbf{Y}}$ est ensuite converti à l'échelle Mel à l'aide d'un banc de 64 filtres triangulaires comme pour l'approche SACC. Le spectrogramme à échelle Mel est utilisé comme caractéristique d'entrée du modèle de segmentation.

5.6.2 Performance de segmentation

Cette section présente une étude expérimentale sur les tâches de VAD et d'OSD. Les performances du modèle BFSACC sont comparées à celles des modèles précédemment développés. Pour toutes les expériences, le nombre de filtres est fixé à $P = 8$. Quelques détails sur l'influence du nombre de filtres sont présentés en annexe E.

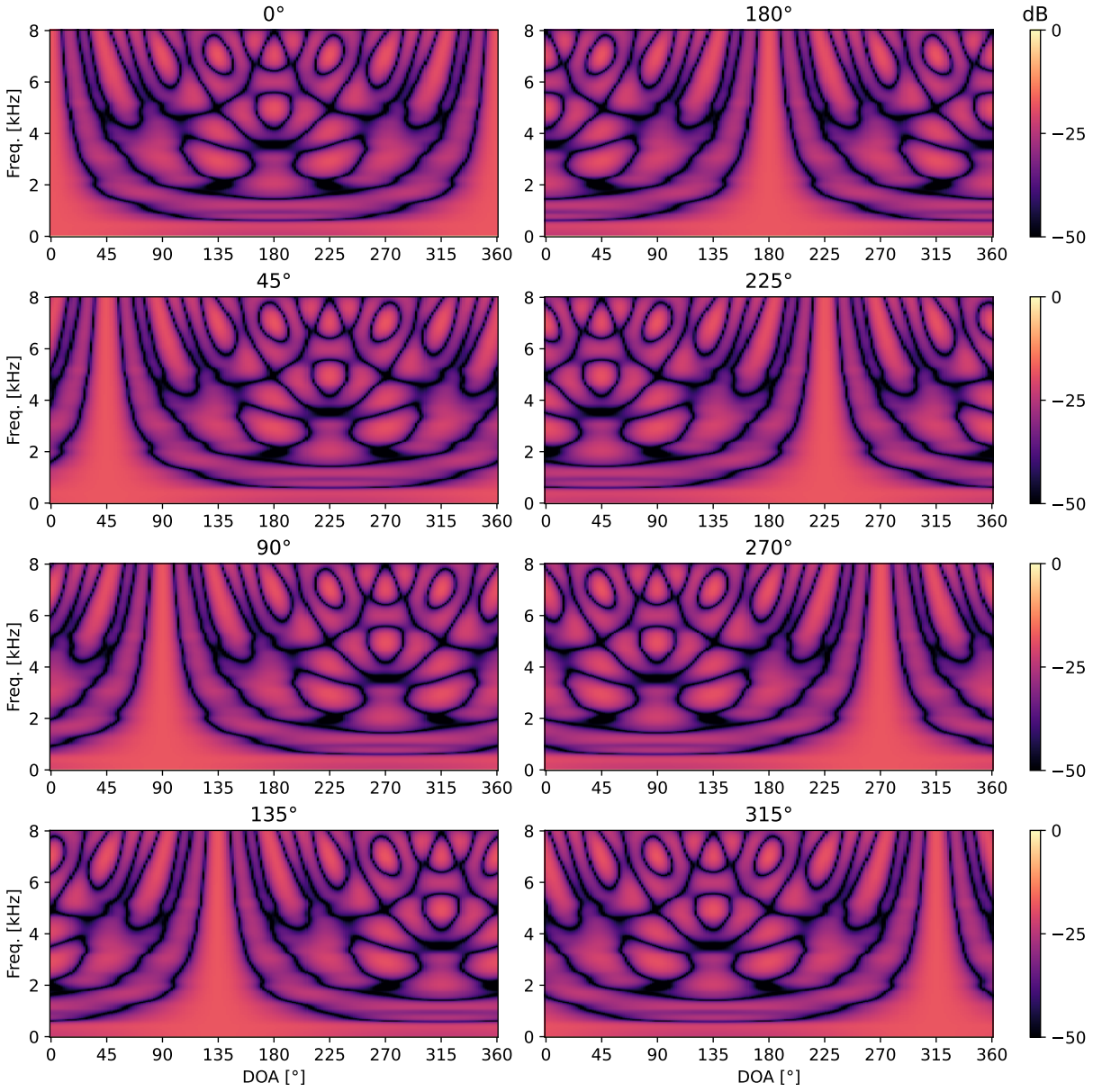


FIGURE 5.9 – Réponse spatiale d'un banc de filtres spatiaux *super-directifs* orientés dans huit directions. Les réponses spatiales sont obtenues pour une ACU composée de huit microphones et de rayon $r = 0,1$ m. Le calcul est réalisé avec une fréquence d'échantillonnage de 16 kHz, 200 DOA et 257 fréquences. La valeur au-dessus des axes indique la direction de focalisation θ_p.

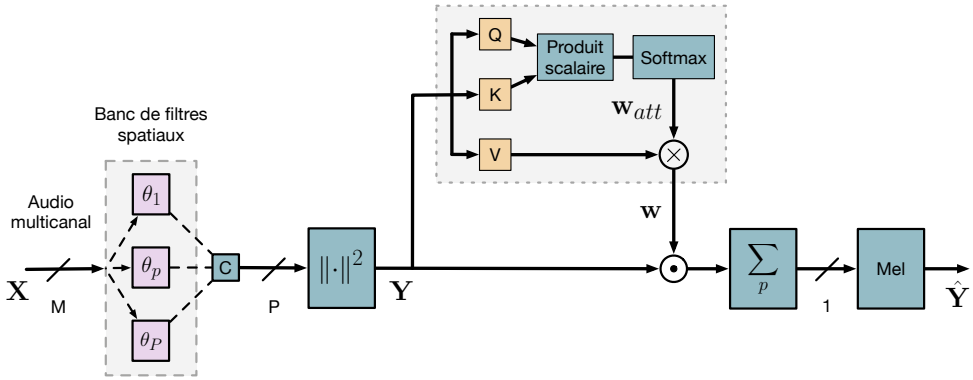


FIGURE 5.10 – Diagramme du modèle BFSACC pour la sélection de filtres spatiaux à l'aide de l'auto-attention. $\boxed{c}$ indique l'opération de concaténation des signaux issus des filtres spatiaux.

Détection d'activité vocale

La table 5.13 présente les performances de VAD du modèle BFSACC couplé avec un TCN pour la modélisation de séquence. Sur les données de développement, ce système obtient un SER de 6,21%, soit un gain absolu de +0,36% par rapport au MDU. À titre de comparaison, le modèle SACC TFCT apporte un gain absolu de +0,76%. Les observations sont similaires sur les données d'évaluation où BFSACC obtient un SER de 6,73% contre 6,95% pour le MDU.

TABLE 5.13 – Détection d'activité vocale à l'aide du modèle BFSACC sur les données de développement et d'évaluation du corpus AMI par rapport aux autres approches.

		FA$_{\%\downarrow}$		**Miss**$_{\%\downarrow}$		**SER**$_{\%\downarrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
TCN	CT MFCC	*3.14*	*2.41*	*2.69*	*3.4*	*5.83*	*5.85*
	MDU MFCC	4.76	2.95	1.82	3.99	6.57	6.95
	MVDR	3.15	2.65	2.59	3.78	**5.74**	**6.43**
	SACC TFCT	3.51	3.08	2.30	3.39	**5.81**	**6.47**
	BFSACC	3,87	3,30	2,35	3,43	6,21	6,73

Détection de parole superposée

La table 5.14 présente les performances d'OSD du système BFSACC suivi d'un TCN par rapport aux autres extracteurs de caractéristiques. Elle montre que la sélection de formation de voies surpasse difficilement le modèle MDU avec un F1-score de 66,8% en phase d'évaluation. BFSACC tend à dégrader les performances par rapport aux autres approches de combinaison de canaux. Par exemple, il présente une dégradation absolue de -2% par rapport au modèle SACC TFCT.

Bien que les performances du modèle BFSACC soient limitées, ce dernier offre des perspectives intéressantes en matière d'interprétabilité des décisions. La pondération et la combinaison des

filtres spatiaux à l'aide de l'auto-attention permet une visualisation explicite des filtres sélectionnés à chaque trame. Lorsque le nombre de filtres utilisé est élevé (ex : $P = 8$), cette approche peut permettre de réaliser une pseudo-localisation des sources en identifiant les filtres sélectionnés. Une étude est proposée en sous-section 5.6.3 et démontre les capacités de sélection et de localisation du modèle.

TABLE 5.14 – Détection de parole superposée (OSD) à l'aide du modèle BFSACC sur les données de développement et d'évaluation du corpus AMI par rapport aux autres approches. Les scores en gras indiquent les meilleures performances. Les valeurs en italiques correspondent à la référence en champ proche.

		Précision$_{\%\uparrow}$		**Rappel**$_{\%\uparrow}$		**F1-score**$_{\%\uparrow}$	
Arch.	Caractéristiques	Dev	Eval	Dev	Eval	Dev	Eval
TCN	CT MFCC	*73.8*	*81.4*	*74.1*	*61.8*	*73.9*	*70.2*
	MDU MFCC	69.5	72.8	68.1	59.5	68.8	65.5
	MVDR	74.3	73.1	70.0	66.2	**72.1**	**69.5**
	SACC TFCT	72.5	72.9	72.3	65.2	**72.4**	**68.8**
	BFSACC	72,5	66,2	66,8	67,4	69,6	66,8

5.6.3 Analyse des poids de combinaison

La sélection de filtres spatiaux, réalisée par le modèle SACC, doit permettre d'activer les directions dans lesquelles le signal contient le plus d'information pour la tâche aval. Théoriquement, un signal contient le maximum d'information lorsque son rapport signal-à-bruit (RSB) est haut (WÖLFEL et al. 2009). Le filtre dont la direction de focalisation est proche de la direction angulaire de la source doit, en théorie, présenter le RSB maximal parmi les signaux disponibles (BENESTY et al. 2008 ; WÖLFEL et al. 2009). L'hypothèse suivante est avancée : le modèle SACC sélectionne les filtres contenant le maximum d'information, et donc les signaux issus des filtres alignés sur la ou les sources actives. Cette sous-section présente le protocole d'évaluation proposé pour vérifier cette hypothèse.

Données et simulation de signaux spatialisés

Les données utilisées pour l'apprentissage du modèle (corpus AMI) ne contiennent pas l'annotation des positions des locuteurs. L'analyse du modèle est donc réalisée sur des données simulées. Pour cela, les données du sous-ensemble de développement des protocoles Libri2Mix et Libri3Mix du corpus LibriMix (COSENTINO et al. 2020) sont utilisées. Ils permettent de créer des signaux de parole superposée contenant respectivement deux et trois locuteurs. Les signaux audio sont issus du corpus Librispeech (PANAYOTOV et al. 2015) obtenus à partir de livres audio. Les données du corpus LibriMix ne contiennent donc pas de parole distante multicanale. L'évaluation des poids de combinaison du modèle est réalisée sur des signaux spatialisés à l'aide d'une simulation, réalisée avec la librairie gpuRIR (DIAZ-GUERRA et al. 2021). Celle-ci permet

de simuler la réponse impulsionnelle d'une salle (RIS) entre une source l de position $\mathbf{s}_l$ et un récepteur m de position $\mathbf{r}_m$. La configuration des microphones est similaire à celle utilisée dans le corpus AMI avec une ACU de $M = 8$ microphones et un rayon de $r = 0,1$ m. En notant x_l le signal temporel associé au locuteur l dans les données LibriMix, le signal spatialisé capté par le microphone m s'exprime :

$$\hat{\mathbf{x}}_{l,m} = x_l * h(\mathbf{s}_l, \mathbf{r}_m), \tag{5.26}$$

où $h(\mathbf{s}_l, \mathbf{r}_m)$ représente la RIS simulée entre une source l et un microphone m, et $*$ le produit de convolution.

Le modèle BFSACC analysé est entraîné comme l'extracteur de caractéristiques d'un système de détection de parole superposée. Les poids de combinaison sont donc analysés dans un cas où plusieurs locuteurs sont actifs simultanément. Dans ce cas, le signal simulé s'exprime :

$$\hat{\mathbf{x}}_m = \sum_{l=1}^{L} x_l * h(\mathbf{s}_l, \mathbf{r}_m). \tag{5.27}$$

La RIS h dépend de la géométrie de la salle, de la position des sources et des récepteurs ainsi que des propriétés acoustiques des matériaux placés sur les parois. Le choix de ces paramètres est décrit ci-après.

Paramètres de la salle Pour l'analyse de BFSACC, différentes RIS sont générées en tirant aléatoirement les dimensions et les paramètres de la salle. Ses dimensions sont représentées par un vecteur $\mathbf{r} = [r_x, r_y, r_z]$ dans lequel chaque composante représente la longueur d'un côté selon un axe du repère. Le tirage des dimensions est réalisé de la façon suivante :

$$\begin{aligned} r_x &\sim \mathcal{U}(3; 10) \\ r_y &\sim \mathcal{U}(3; 10) \\ r_z &\sim \mathcal{U}(2,5; 4), \end{aligned} \tag{5.28}$$

où $\mathcal{U}(a; b)$ représente une distribution uniforme définie sur l'intervalle $[a, b]$.

La RIS dépend également de nombreux paramètres physiques, tels que le coefficient d'absorption des parois et des obstacles. Tous ces paramètres peuvent être englobés dans le T_{60}, correspondant au temps de réverbération de la salle. Ce paramètre est également tiré aléatoirement en suivant une loi uniforme telle que $T_{60} \sim \mathcal{U}(0,3; 1,0)$

Position des sources et des microphones La position des sources est également tirée aléatoirement. Cependant, afin de simplifier l'évaluation des poids appris par le modèle, la position angulaire est supposée appartenir à l'un des P secteurs angulaires définis par les filtres. Un filtre spatial correspond à un secteur angulaire centré sur la direction de focalisation. Les étapes de sélection de la position des sources sont les suivantes :

- Choix du nombre de sources actives L,
- Tirage des indices des secteurs angulaires activés (sans remise) : $\mathbf{q} \sim \mathbb{U}(1; P)$ avec $\mathbf{q} \in \mathbb{R}^L$,
- Tirage de la distance de chaque source par rapport à l'antenne : $\mathbf{d} \sim \mathcal{U}(d_{min}; d_{max})$ avec $\mathbf{d} \in \mathbb{R}^L$.

d_{min} et d_{max} représentent respectivement les distances minimale et maximale de la source par rapport au centre de l'antenne. $\mathbb{U}(a; b)$ représente une distribution uniforme discrète sur l'intervalle $[a, b]$. La position angulaire des sources reste à être déterminée. Deux scénarios de sélection sont considérés :

***Easy* :** la position des sources correspond au centre des secteurs angulaires. Elles sont donc alignées aux filtres spatiaux.

***Hard* :** pour chaque secteur angulaire q préalablement sélectionné, la position de la source est tirée aléatoirement au sein de ce dernier : $\theta_s \sim \mathcal{U}(\theta^q_{inf} + \gamma; \theta^q_{sup} - \gamma)$.

θ^q_{inf} représente l'angle inférieur du secteur angulaire, θ^q_{sup} l'angle supérieur et γ une marge afin d'éviter deux sources de secteurs angulaires adjacents d'être trop proches. Nous fixons $\gamma = 5^\circ$ pour les expériences à suivre. L'angle des secteurs angulaires est défini par rapport au centre de l'antenne.

Le placement des microphones est conditionné par la géométrie de l'antenne. Une antenne circulaire de $M = 8$ microphones est considérée ici. La position du centre de l'antenne est tirée aléatoirement dans un carré de 1×1 m^2 dont le centre correspond au centre de la salle.

Sélection des directions angulaires L'évaluation des directions sélectionnées est réalisée comme pour une tâche de classification. Les positions de référence (secteurs angulaires) sont encodées dans un vecteur binaire $\boldsymbol{\theta} \in \mathbb{R}^P$ dont les éléments activés correspondent aux indices des secteurs angulaires $\mathbf{q}$.

Les directions estimées sont obtenues à partir des poids de sélection $\mathbf{w} \in \mathbb{R}^{T \times P}$ estimés par le modèle SACC pour le segment simulé courant. Les directions sélectionnées par le modèle sont obtenues en calculant la moyenne temporelle de $\mathbf{w}$:

$$\bar{\mathbf{w}} = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t. \tag{5.29}$$

Comme les poids $\mathbf{w}$ sont obtenus en appliquant une fonction softmax sur la dimension des canaux, la moyenne des poids vérifie $\bar{\mathbf{w}} \in [0, 1]$. Un seuil $\tau \in [0, 1]$ est appliqué pour sélectionner les directions les plus activées par le modèle :

$$\hat{\theta}_p = \begin{cases} 1 & \text{si } \bar{w}_p > \tau \\ 0 & \text{sinon.} \end{cases} \tag{5.30}$$

La sélection des directions est évaluée avec l'accuracy et le F1-score entre $\boldsymbol{\theta}$ et $\hat{\boldsymbol{\theta}} = \left[\hat{\theta}_1, \cdots, \hat{\theta}_p\right]$.

Évaluation des directions sélectionnées

Cette section évalue des directions sélectionnées par le modèle BFSACC par rapport aux positions réelles des sources. Les directions sont estimées à l'aide de l'équation (5.30). Les résultats sont présentés dans la table 5.15 dans les cas $L = 2$ et $L = 3$ pour chaque scénario. Un scénario supplémentaire, *random*, présente les scores dans le cas où les directions sélectionnées sont aléatoires.

TABLE 5.15 – Performance de localisation de source du modèle BFSACC en fonction du nombre de sources et du scénario considérés. P : précision, R : rappel, F1 : F1-score, Acc : accuracy.

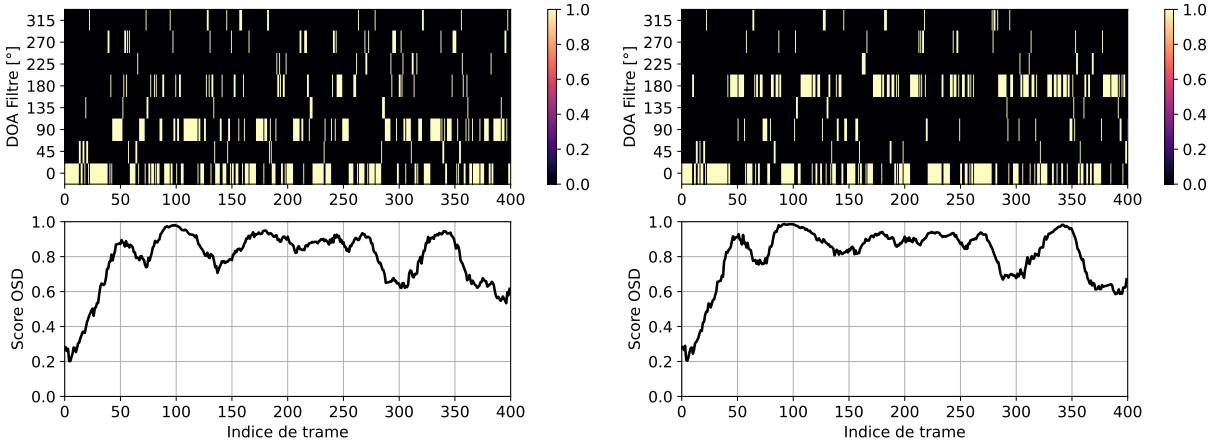
L	**Scenario**	**P** (%)	**R** (%)	**F1** (%)	**Acc** (%)
2	*Random*	49,9	49,9	49,9	62,4
	Easy	89,4	82,1	83,2	88,1
	Hard	82,5	76,8	77,1	83,3
3	*Random*	49,7	49,6	48,7	56,0
	Easy	82,5	76,1	75,9	79,0
	Hard	76,8	71,6	70,8	74,2

Dans le cas $L = 2$, les poids BFSACC permettent de localiser les sources actives avec un F1-score de 83,2% et une accuracy de 88,1% dans le cas *easy*. En conditions *hard*, les sources sont localisées avec un F1-score de 76,8% et une accuracy de 83,3%. Le modèle sélectionne donc moins fréquemment le secteur angulaire de la source active lorsque celle-ci n'est pas strictement alignée sur le filtre. Cependant, dans le cas aléatoire, le F1-score est de seulement 49,9% et l'accuracy de 62,4%. Cela montre que le modèle BFSACC sélectionne principalement les directions des sources actives dans chaque scénario.

Les observations sont similaires dans le cas $L = 3$. Dans le scénario *easy*, le modèle obtient un F1-score de 75,9% et une accuracy de 79,0%. Dans le scénario *hard*, les scores sont dégradés avec un F1-score de 70,8% et une accuracy de 74,2%. Les valeurs sont inférieures au cas $L = 2$, la tâche étant plus difficile lorsque trois sources sont actives simultanément. De plus, le modèle n'a pas observé beaucoup de situations où au moins trois locuteurs sont actifs simultanément lors de la phase d'apprentissage. Cela peut expliquer la différence entre les cas $L = 2$ et $L = 3$. Les scores restent cependant largement supérieurs au cas *random* (F1-score : 48,7%, Accuracy : 56,0%).

La figure 5.11 présente les poids de combinaison obtenus dans le scénario $L = 2$ pour deux directions de sources et la prédiction d'OSD obtenue. En figure 5.11a, les deux directions principalement activées sont 0° et 90°. Celles-ci correspondent aux positions angulaires des sources. La figure 5.11b présente le même comportement. Le modèle sélectionne majoritairement des directions 0° et 180°, correspondant à celles des sources. Les scores de détection de parole superposée sont élevés. Cela montre que le modèle la détecte correctement et que la simulation est cohérente par rapport aux données observées lors de l'apprentissage.

Sur chacune de ces figures, les poids sont quasiment binaires. Cela signifie que les valeurs des



(a) $\theta_{s1} = 0^\circ$, $\theta_{s2} = 90^\circ$ (b) $\theta_{s1} = 0^\circ$, $\theta_{s2} = 180^\circ$

FIGURE 5.11 – (Haut) Poids de combinaison appliqués à chaque canal pour deux sources de directions θ_{s1} et θ_{s2}. (Bas) Score de détection de parole superposée associé. Ces figures sont obtenues sur un segment de 4 secondes du fichier *911-128684-0056_8975-270782-0072_2159-179157-0026.wav* du protocole Libri3Mix avec les locuteurs *s1* et *s2* dont les positions sont fixes.

poids avant d'appliquer le softmax en équation (5.24) sont très contrastées. Ce fort contraste est principalement lié à de grandes valeurs dans les poids de la formation de voies. Ces grandes valeurs sont liées à une singularité pouvant intervenir lors de l'inversion de la matrice $\mathbf{\Sigma}_N$ en équation (5.18).

5.6.4 Régularisation de la formation de voies

Comme évoqué dans la sous-section précédente, le module d'auto-attention génère des poids de combinaison très contrastés suite à des valeurs d'entrée élevées. Une approche pour limiter les grandes valeurs en sortie de la formation de voies consiste à régulariser l'expression (5.18). G. HUANG et al. (2016) présentent une méthode simple de régularisation :

$$\mathbf{w}_p^H(f) = \frac{\mathbf{v}_p^H(f)\left(\mathbf{\Sigma}_N(f) + \lambda \mathbf{I}_P\right)^{-1}}{\mathbf{v}_p^H(f)\left(\mathbf{\Sigma}_N(f) + \lambda \mathbf{I}_P\right)^{-1}\mathbf{v}_p(f)}, \tag{5.31}$$

avec $\mathbf{I}_P$ la matrice identité de taille P et λ un paramètre de régularisation. Les erreurs commises par la formation de voies *super directive* sont principalement liées à un écart entre la géométrie théorique, permettant le calcul de la matrice $\mathbf{\Sigma}_N$, et la géométrie réelle. La régularisation permet de limiter ces erreurs.

Impact sur les performances de segmentation

La table 5.16 présente les performances de VAD obtenues avec la régression. Les résultats montrent que le calcul des poids régularisé avec l'équation (5.31) améliore la détection avec

respectivement 6,06% et 6,52% de SER en développement et en évaluation. Par exemple, la régularisation permet un gain relatif de +3,1% en évaluation par rapport au cas où $\lambda = 0$.

La table 5.16 présente l'impact de la régularisation sur l'OSD. Les résultats montrent que la régularisation des poids de formation de voies a un impact positif sur les performances d'OSD. Elle permet un gain absolu de +1,4% sur le F1-score de développement et de +0,9% en évaluation.

TABLE 5.16 – Détection de parole superposée (OSD) à l'aide du modèle BFSACC sur les données de développement et d'évaluation du corpus AMI en fonction du paramètre de régularisation λ.

VAD	**FA**$_{\%}$ ↓		**Miss**$_{\%}$ ↓		**SER**$_{\%}$ ↓	
	Dev	Eval	Dev	Eval	Dev	Eval
$\lambda = 0$	3,87	3,30	2,35	3,43	6,21	6,73
$\lambda = 10^{-4}$	3,61	3,07	2,45	3,44	**6,06**	**6,52**
OSD	**Précision**$_{\%}$ ↑		**Rappel**$_{\%}$ ↑		**F1-score**$_{\%}$ ↑	
	Dev	Eval	Dev	Eval	Dev	Eval
$\lambda = 0$	72,5	66,2	66,8	67,4	69,6	66,8
$\lambda = 10^{-4}$	70,5	69,5	71,6	65,9	**71,0**	**67,7**

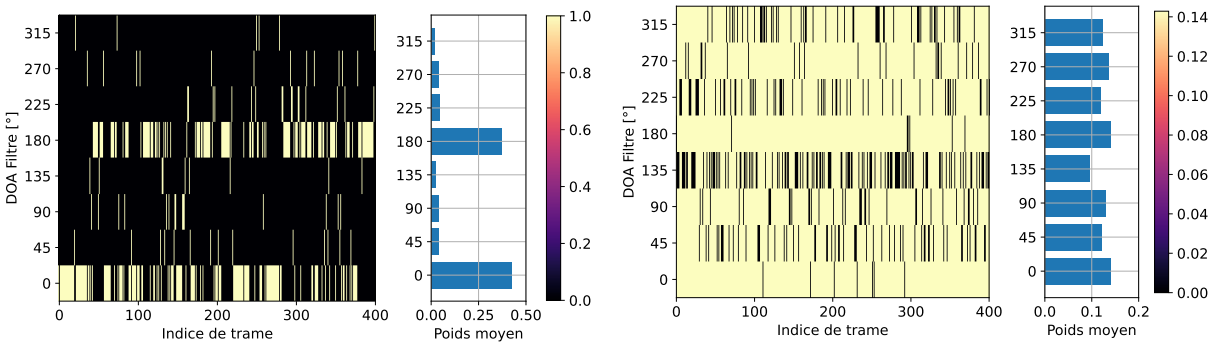
Impact sur l'interprétation

La figure 5.12 présente les poids de combinaison obtenus pour les deux modèles étudiés. Les poids moyens, calculés à l'aide de l'équation (5.29), sont également représentés sous forme d'histogramme afin d'évaluer l'activation moyenne des filtres au cours du temps. Les visualisations sont obtenues pour une simulation faisant intervenir deux locuteurs positionnés dans les directions $\theta_{s1} = 0^\circ$ et $\theta_{s2} = 180^\circ$.

La figure 5.12a présente les poids de combinaison obtenus sans régularisation ($\lambda = 0$). Elle montre un fort contraste dans les valeurs des poids et le modèle sélectionne clairement les deux directions associées aux sources. La visualisation des poids moyens illustre nettement ce comportement avec les deux directions des sources majoritairement activées.

La figure 5.12b présente la même analyse dans le cas régularisé avec $\lambda = 10^{-4}$. Ici, les poids sont nettement moins contrastés. Bien que certaines directions soient majoritairement désactivées (ex : 135°), la sélection des filtres est moins marquée. Sur cet exemple, le modèle active principalement les directions des sources actives (0° et 180°), mais pas exclusivement.

Lorsque la régularisation est considérée, la sélection des filtres semble plus floue. Cela permet potentiellement d'améliorer les performances de l'OSD. En effet, dans le cas $\lambda = 0$, le modèle n'est capable de sélectionner qu'une direction à chaque trame. Le signal résultant de la sélection contient seulement le locuteur actif dans la direction choisie, masquant ainsi le second. La détection de la parole superposée dans ce contexte semble alors délicate. Au contraire, dans le cas $\lambda = 10^{-4}$, le modèle active au moins les deux directions des sources à toutes les trames. Il active donc les deux locuteurs, permettant ainsi au modèle d'observer un signal correspondant à



(a) $\lambda = 0$ (b) $\lambda = 10^{-4}$

FIGURE 5.12 – Poids de combinaison appliqués à chaque canal et leur moyenne temporelle dans les cas (gauche) $\lambda = 0$, (droite) $\lambda = 10^{-4}$. Ces figures sont obtenues sur un segment spatialisé de 4 secondes issu du fichier *911-128684-0056_8975-270782-0072_2159-179157-0026.wav* du protocole Libri3Mix avec les locuteurs *s1* et *s2*. $\theta_{s1} = 0^\circ$ et $\theta_{s2} = 180^\circ$.

la parole superposée.

5.7 VAD et OSD pour la diarisation en locuteurs

Cette section évalue l'influence de la segmentation sur la tâche de diarisation. Pour cela, des expériences sont menées à l'aide du modèle VBx (LANDINI et al. 2022b). La section 5.7.1 introduit le protocole expérimental pour les expériences de diarisation en locuteur. Les résultats sont présentés en section 5.7.2 avant de les discuter et de conclure en section 5.7.3.

5.7.1 Protocole expérimental

L'obtention de la diarisation en locuteur à partir de la segmentation consiste en trois étapes :

- extraction d'embeddings de locuteur sur les segments de parole issus de la VAD,
- regroupement des embeddings à l'aide de l'algorithme VB-HMM (LANDINI et al. 2022a),
- assignation des segments de parole superposés détectés par l'OSD pour affiner la diarisation.

Les détails d'implémentation de ces trois étapes et l'évaluation sont présentés dans cette sous-section.

Intégration de la segmentation

L'objectif de cette section est d'étudier l'influence des modèles de VAD et d'OSD sur la tâche de segmentation et regroupement en locuteur. La VAD permet d'obtenir les segments de parole pour chaque fichier traité. Les embeddings de locuteurs sont ensuite extraits sur ces segments de parole. Cette extraction est réalisée à l'aide d'un réseau convolutif résiduel ResNet101 (K. HE et al. 2016) utilisant un spectrogramme à échelle Mel comme caractéristiques acoustiques. Le regroupement des embeddings par locuteurs et l'affinement de la segmentation sont ensuite

réalisés à l'aide du modèle VB-HMM (Landini et al. 2022b). Cette étape permet d'obtenir une première diarisation.

La diarisation obtenue peut être affinée à partir des segments détectés par l'OSD. La majorité des segments de parole superposée contiennent deux locuteurs. Par exemple, seulement 3,1% des trames contiennent plus de deux locuteurs actifs simultanément sur les données de développement du corpus AMI et 4,1% sur les données d'évaluation (Cornell et al. 2022a). Sous l'hypothèse que la majorité des segments de parole superposée contiennent deux locuteurs, une étape d'assignation est ajoutée à la diarisation. Elle consiste à assigner un second locuteur aux segments détectés par l'OSD à l'aide d'une approche heuristique (Otterson et al. 2007). Le locuteur du segment le plus proche dans le temps est assigné comme second locuteur.

L'étude suivante compare l'influence de la segmentation obtenue avec les modèles MDU, SACC+TFCT, SACC+$\mathcal{A}_{32}$, LcSACC et BFSACC sur la tâche de diarisation en locuteur. Pour chaque système, la segmentation est réalisée avec le modèle obtenant les meilleurs scores sur le sous-ensemble de développement. L'extraction d'embeddings, leur regroupement et l'assignation de la parole superposée sont réalisés à l'aide du code développé par Landini et al. (2021) et disponible en ligne [2].

Évaluation

Les modèles de diarisation du locuteur sont évalués à l'aide du DER et du JER. Pour le calcul du DER, deux méthodes d'évaluation sont considérées et suivent le protocole proposé par le NIST (T. J. Park et al. 2022) :

- δ_c = 25 : une tolérance de 25 ms est appliquée sur les frontières de segments lors de la phase d'évaluation,
- δ_c = 0 : aucune tolérance n'est considérée.

L'évaluation de la diarisation est réalisée sur le sous-ensemble d'évaluation du corpus AMI. Les performances Oracle, avec la VAD et l'OSD de référence, sont également reportées. Elles permettent de montrer les performances espérées si la segmentation est parfaite et donnent ainsi une limite supérieure en matière de diarisation.

5.7.2 Résultats

La table 5.17 présente les performances de diarisation en locuteur obtenues pour chaque système de segmentation. Par la suite, et sauf mention contraire, les DER discutés sont ceux obtenus dans le cas δ_c = 0. Les gains ou dégradations reportés entre modèles sont relatifs.

Dans le cas Oracle, le modèle atteint un DER 22,47% et un JER de 30,52%. Il s'agit donc des meilleures performances pouvant être atteintes avec le système actuel. L'assignation de la parole superposée permet un gain de +36,3% sur le DER et de +15,5% sur le JER. L'influence de l'OSD sur la diarisation en locuteur n'est donc pas négligeable.

2. https://github.com/BUTSpeechFIT/VBx/tree/v1.1_VoxConverse2020 consulté le 12 octobre 2023

Dans le cas où la VAD est obtenue à partir du MDU, le système atteint un DER de 27,45% et un JER de 33,82%. Cette segmentation présente une dégradation relative du DER de -22,2% par rapport aux performances Oracle. La VAD a donc un fort impact sur la diarisation. L'assignation de la parole superposée améliore les performances avec un gain de +9,0%. Ici encore, la prise en compte de la parole superposée améliore les performances, bien que le gain soit limité par rapport à l'Oracle.

TABLE 5.17 – Performances de diarisation en locuteur pour chaque modèle de segmentation en considérant (VAD+OSD) ou non (VAD) l'assignation des segments de parole superposée. Les scores sont obtenus sur le sous-ensemble de développement du corpus AMI.

Système	**Segmentation**	**DER**$_{\%\downarrow}$		**JER**$_{\%\downarrow}$
		δ_c = 25	δ_c = 0	
Oracle	VAD	15,63	22,47	30,52
	VAD+OSD	10,12	14,32	25,78
MDU	VAD	19,38	27,45	33,82
	VAD+OSD	17,60	24,97	32,15
SACC+TFCT	VAD	18,30	26,63	33,18
	VAD+OSD	16,57	**24,02**	31,62
SACC+$\mathcal{A}_{32}$	VAD	19,45	27,80	34,73
	VAD+OSD	17,66	25,30	33,28
LcSACC	VAD	18,34	26,72	33,10
	VAD+OSD	**16,51**	24,13	31,53
BFSACC	VAD	18,41	26,59	32,84
	VAD+OSD	16,96	24,30	**31,35**

Les observations sont similaires pour tous les modèles de segmentation en conditions distantes. SACC+TFCT et LcSACC permettent les meilleures performances en DER. Le premier atteint un DER de 24,02% en prenant en compte la parole superposée. Cela représente un gain de 3,8% par rapport au MDU. Le second, LcSACC, obtient le meilleur DER pour δ_c = 0 avec 16,51%.

Le système BFSACC obtient également des performances similaires, et atteint le meilleur JER avec la VAD (32,84%) et les VAD+OSD (31,35%). Dans le second scénario, cela représente un gain de +2,5% par rapport au MDU. Le système SACC+$\mathcal{A}_{32}$ n'améliore par la diarisation par rapport au MDU. Comme l'ont montré les expériences de la section 5.3, ce système n'améliore ni la VAD, ni l'OSD. Il était ainsi attendu que ce système dégrade les performances de diarisation.

5.7.3 Discussions

Cette section étudie l'impact de chaque système de segmentation (VAD et OSD) sur la tâche de diarisation en locuteur en conditions de parole distante. Les performances Oracle montre que la parole superposée ne doit pas être négligée. L'OSD apporte en effet un gain remarquable sur les

performances. Cependant, malgré ce gain, le DER reste élevé par rapport aux performances de la littérature sur les données AMI en champ proche. Par exemple, Landini et al. (2022b) rapportent un DER de 18,99% et un JER de 24,57% avec le VAD Oracle et sans prise en compte de la parole superposée. L'écart de performance entre champ proche et parole distante est principalement lié à l'extracteur d'embeddings qui n'est pas entraîné sur ce type de données. Remplacer l'extracteur d'embeddings par un modèle basé, par exemple, sur WavLM (S. Chen et al. 2022), permettrait probablement une amélioration.

Les résultats obtenus montrent également un faible écart de performance entre les modèles de segmentation distants. Certes, les modèles SACC+TFCT, LcSACC et BFSACC permettent un gain sur la diarisation par rapport au MDU, mais il reste mineur. Les erreurs commises par les modèles de détection de parole superposée semblent réduire l'intérêt de l'assignation de ces segments. Là où cette assignation apporte +36,3% de gain dans le cas Oracle, il se retrouve réduit à +9,0% avec le système MDU et +9,8% avec SACC+TFCT. L'utilisation du F1-score comme métrique d'évaluation de l'OSD n'est donc pas représentative du gain obtenu sur la tâche finale.

Cependant, au-delà du gain sur le DER, l'OSD permet d'identifier des zones sur lesquelles le système de diarisation est susceptible de commettre des erreurs. Par exemple, la présence de locuteurs en simultané ne permet pas d'extraire des embeddings homogènes et peut ajouter de la confusion lors de l'étape de regroupement. L'OSD peut alors permettre d'éliminer ces segments lors de l'extraction des embeddings. Elle peut également servir pour estimer la confiance dans les décisions du système de diarisation : une zone contenant de la parole superposée est plus susceptible d'être mal assignée.

5.8 Conclusions

Ce chapitre présente un ensemble de modèles permettant de pondérer et de combiner les canaux issus d'une antenne de microphones à l'aide de mécanismes d'auto-attention. Ces approches s'inspirent du modèle SACC proposé par Gong et al. (2021). Le modèle original permet d'obtenir parmi les meilleures performances de détection d'activité vocale et de parole superposée en conditions distantes. Pour répondre à la question de recherche **Q1**, la combinaison auto-attentive permet d'améliorer la détection d'activité vocale et de parole superposée.

Deux extensions sont proposées. La première consiste à utiliser des filtres analytiques optimisés. Ceux-ci peuvent être initialisés aléatoirement, ou à l'aide d'un banc de filtres rectangulaires. Les filtres analytiques optimisés simultanément avec la tâche de détection ne sont pas bénéfiques à la segmentation (**Q2**).

Une seconde approche est proposée, permettant de prendre en compte la phase de la transformée de Fourier. Deux méthodes explicites (EcSACC) et implicite (IcSACC) sont proposées. La première offre des performances de détection se rapprochant du modèle original. La seconde tend à dégrader la qualité de la détection.

Le formalisme des modèles EcSACC et IcSACC ne permet pas d'interpréter les directions

spatiales sélectionnées par le système. Cela réduit l'intérêt de la formulation complexe par rapport au modèle SACC, plus performant. Une seconde formulation, LcSACC, réalise un produit complexe linéaire entre les poids et la TFCT. Cette considération permet d'analyser les poids de combinaison comme un filtre spatial classique. Le modèle LcSACC offre les meilleurs scores de VAD et des performances d'OSD proches de SACC. Pour répondre à la question **Q3**, les modèles complexes peuvent obtenir des performances de détection similaires au modèle SACC original, sans amélioration significative des performances.

L'analyse des poids montre que LcSACC extrait différentes représentations en fonction de la position de la source. Il encode donc l'information spatiale. L'analyse sur des données réelles ne montre cependant pas une corrélation claire entre les maxima d'énergie acoustique (positions potentielles des locuteurs) et la direction de focalisation du système. Une évaluation quantitative permettrait une analyse plus fine pour analyser le comportement du modèle (**Q4**).

Les modèles complexes permettent une meilleure compréhension des poids de combinaison. Leur comportement est cependant difficile à analyser par rapport à la position des locuteurs actifs. Une approche hybride, utilisant SACC pour sélectionner les sorties d'un banc de filtres spatiaux, est donc proposée. Ce système (BFSACC) montre un potentiel intéressant en matière d'interprétabilité. Le modèle sélectionne les filtres spatiaux les plus proches des sources actives, permettant une pseudo-localisation des locuteurs. Les performances de ce modèle peuvent être améliorées à l'aide d'une régularisation, rendant cependant l'interprétation des poids plus difficile. L'auto-attention permet donc de sélectionner des filtres spatiaux en offrant une perspective intéressante pour l'interprétation du modèle (**Q5**).

L'influence de la segmentation est finalement évaluée sur la tâche de diarisation en locuteur. Le modèle VBx est utilisé en utilisant les segments résultants des systèmes de VAD et d'OSD. Dans cette configuration, les modèles SACC+TFCT, LcSACC et BFSACC permettent d'obtenir des performances similaires, meilleures que le microphone distant unique. L'impact de l'assignation de la parole superposée est également évalué. Cette étape permet d'améliorer significativement les performances de diarisation du locuteur. Le gain reste cependant limité par la qualité de la détection. L'utilisation de méthodes de combinaison de canaux pour la segmentation a donc un impact positif sur la diarisation en locuteur (**Q6**).

NB. : le tableau présenté ci-dessous est uniquement proposé à titre d'exemples dédié à justifier les directions choisies par la suite. Le protocole n'est donc pas détaillé ici.

Les modèles de combinaison de canaux permettent d'améliorer la qualité de la VAD et de l'OSD en conditions distantes. Dans l'objectif d'obtenir la segmentation complète du signal, la détection de changements de locuteurs (SCD) doit également être réalisée. Le tableau 5.18 présente un exemple de performances obtenues sur la tâche de SCD avec les modèles MDU MFCC et SACC TFCT. Le protocole utilisé est celui décrit en section 4.3, avec l'encodage des étiquettes proposé dans le chapitre 6.1. Les résultats montrent que la combinaison de canaux

ne permet pas d'améliorer la détection de changements de locuteur par rapport aux MFCC. D'autres approches sont donc requises afin d'améliorer la qualité de la détection sur cette tâche.

TABLE 5.18 – Performances de SCD obtenues à l'aide de deux types de caractéristiques acoustiques sur les données distantes du corpus AMI (développement).

Arch.	Caractéristiques	Pureté (%) ↑	Couverture (%) ↑	S-score (%) ↑
TCN	MDU MFCC	**79,7**	81,5	**80,6**
	SACC TFCT	74,1	**88,1**	**80,9**

Le chapitre suivant propose d'utiliser des caractéristiques spatiales extraites du signal multicanal afin d'apporter de l'information au système de détection. L'influence de ces caractéristiques est évaluée sur les tâches de VAD, OSD et SCD. Un nouveau jeu de caractéristiques spatiales, basé sur les harmoniques circulaires, est proposé afin d'améliorer la robustesse du système au nombre de capteurs.

Chapitre 6

Caractéristiques spatiales pour la segmentation de la parole

Les travaux menés sur la segmentation et le comptage de locuteurs à partir de signaux de parole distante montrent que l'information spatiale est bénéfique (Cornell et al. 2022a). Elle apporte une connaissance sur la répartition du champ acoustique dans l'espace et, explicitement ou non, sur la position des sources. L'extraction de caractéristiques spatiales à partir d'un signal audio multicanal permet d'enrichir les caractéristiques acoustiques (ex : MFCC). Les travaux de Cornell et al. (2022a) montrent notamment que les IPD améliorent la VAD, l'OSD et le comptage de locuteurs à la trame. Cependant, ces caractéristiques spatiales dépendent fortement du nombre de paires de microphones considérées et peuvent manquer de robustesse en cas de modification de la géométrie d'antenne en phase d'évaluation.

Le formalisme des harmoniques circulaires (CH) permet de représenter le champ acoustique comme une combinaison linéaire de fonctions spatiales (Torres et al. 2016). Elles sont l'équivalent à deux dimensions des harmoniques circulaires et cylindriques, plus couramment utilisées (Williams et al. 2000). Les harmoniques circulaires permettent notamment de représenter le signal mesuré à partir d'un faible nombre de capteurs. Elles sont donc un candidat intéressant pour améliorer la robustesse des systèmes au nombre de microphones disponibles. Ce formalisme a été utilisé pour la localisation de sources acoustiques (SongGong et al. 2021) ou l'analyse de champ acoustique dans les volumes fermés (Torres et al. 2016). Les harmoniques sphériques ont également été exploitées pour la localisation de sources à l'aide de méthodes neuronales (Grumiaux et al. 2021). L'étude bibliographique menée ne nous a pas permis d'identifier des travaux ayant utilisé les harmoniques circulaires dans le contexte du traitement automatique de la parole. Ce chapitre propose une première approche pour intégrer ce type de représentation pour la segmentation de la parole pour la diarisation en locuteur.

Le chapitre est organisé comme suit. La section 6.1 présente les systèmes, le protocole d'apprentissage et les données utilisées. La section 6.2 décrit les caractéristiques basées sur les harmoniques circulaires. Les résultats sur les tâches de VAD, d'OSD et de SCD sont abordés en section 6.3. La section 6.4 évalue la sensibilité de ces caractéristiques en fonction du nombre de capteurs.

Questions de recherche

Les questions de recherche étudiées dans ce chapitre sont les suivantes :

Q1 Les caractéristiques spatiales IPD et CSIPD permettent-elles d'améliorer la détection de changements de locuteur ? (Section 6.3)

Q2 Les caractéristiques proposées (CH-DOA) apportent-elles un gain pour la segmentation de la parole distante ? (Section 6.3)

Q3 Quelle est la robustesse des modèles CH-DOA par rapport au nombre de microphones actifs ? (Section 6.4)

Q4 Quel est l'influence des caractéristiques spatiales sur la diarisation en locuteur ? (Section 6.5)

6.1 Méthodologie

Cette section présente les modèles et le protocole d'apprentissage pour l'utilisation de caractéristiques spatiales dans le contexte de la segmentation automatique de la parole multicanal.

6.1.1 Formalisme et système

Tâches visées

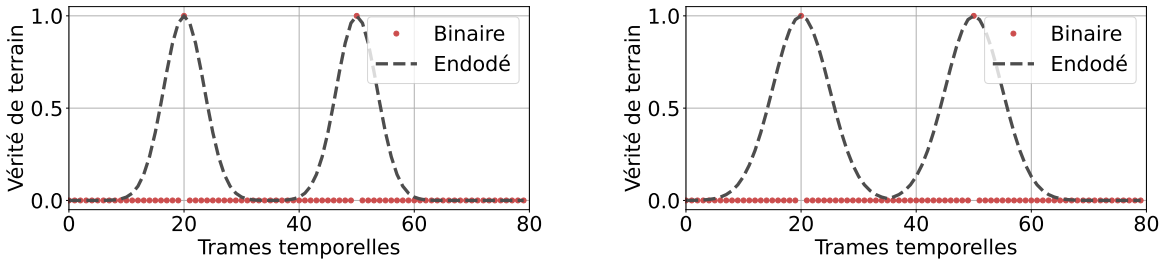
Les caractéristiques spatiales sont utilisées pour résoudre trois tâches de segmentation de la parole en conditions distantes : VAD, OSD et SCD. Les deux premières tâches sont formulées comme la classification binaire d'une séquence de caractéristiques extraites du signal (*cf.* section 5.1.1). La SCD est définie comme une tache de régression (Hrúz et al. 2017). Elle peut également être formulée comme une tâche de classification binaire, au même tire que la VAD et l'OSD (Yin et al. 2017). L'annexe B étudie ces deux formulations et montre que la régression offre des performances légèrement supérieures. La régression évite également le recouvrement entre les annotations, celles-ci pouvant apparaître dans le cas de la classification (Annexe B).

Dans notre cas, le modèle apprend à prédire un ensemble de Gaussiennes centrées sur les frontières des segments. Cette approche a été appliquée dans le contexte de la localisation de sources (W. He et al. 2018). En notant N_+ le nombre de changements de locuteur dans l'annotation, la fonction à prédire est obtenue par la relation suivante :

$$y_{e,i} = \begin{cases} \max_{j=1}^{N_+} \left(e^{-d(y_i,y_j)^2/\sigma^2} \right) & \text{si } N_+ > 0, \\ 0 & \text{sinon,} \end{cases} \tag{6.1}$$

où $d(\cdot,\cdot)$ représente la distance entre deux éléments, y_i un élément i et y_j un élément positif j parmi les N_+ existants et $y_{e,i}$ le nouvel élément encodé. σ représente l'écart-type de la Gaussienne considérée. La figure 6.1 présente deux exemples de référence d'apprentissage pour deux valeurs

de σ. En pratique, l'écart-type de l'équation (6.1) est rendu aléatoire sur l'intervalle $[2, 7]$ durant l'apprentissage en suivant une loi uniforme telle que $\sigma \sim \mathcal{U}_{[2,7]}$.



(a) σ = 5 (b) σ = 7

FIGURE 6.1 – Exemples de références pour la tâche de SCD. L'annotation binaire des frontières de segments est encodée par des Gaussiennes centrées sur cette frontière.

Architecture

Les expériences de segmentation automatique de la parole à l'aide de caractéristiques spatiales sont conduites à l'aide d'une architecture TCN identique à celle utilisée section 5.2.1. Les caractéristiques acoustiques $\mathbf{A} \in \mathbb{R}^{F_a \times T}$ et spatiales $\mathbf{S} \in \mathbb{R}^{F_s \times T}$ sont extraites à partir du signal temporel $\mathbf{x} \in \mathbb{R}^{M \times L}$. F_a et F_s représentent respectivement le nombre de caractéristiques acoustiques et spatiales extraites. Les caractéristiques acoustiques sont concaténées avec les caractéristiques spatiales avant d'alimenter le TCN pour obtenir la séquence prédite $\hat{\mathbf{y}} \in \mathbb{R}^{C \times T}$. Le nombre de sorties C du système dépend de la tâche considérée. Pour une tâche de classification (ex : OSD), le modèle est composée de $C = 2$ sorties, chacune étant associée à une classe spécifique. Dans le cas de la régression, une seule sortie $C = 1$ permet de prédire les valeurs de la fonction à chaque trame. L'architecture complète est présentée en figure 6.2.

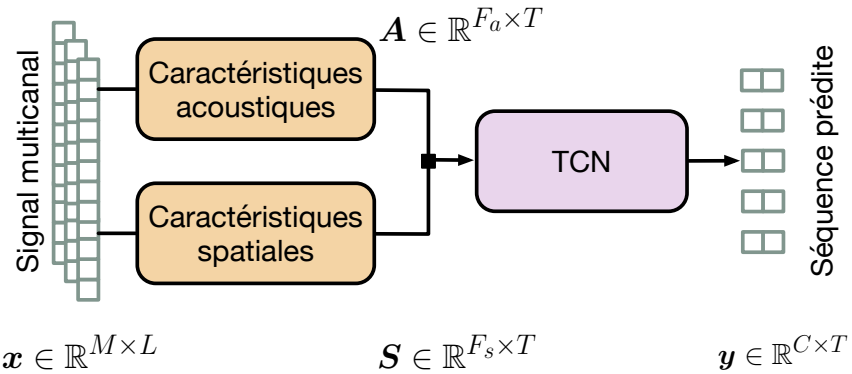


FIGURE 6.2 – Architecture considérée pour la segmentation automatique de la parole à l'aide de caractéristiques acoustiques et spatiales extraites à partir du signal multicanal.

6.1.2 Protocole d'évaluation

Le protocole d'apprentissage est identique à celui présenté en section 4.3. Comme pour les expériences précédentes (section 5.1.2), les poids du modèle sont optimisés à l'aide de l'algorithme Adam (Kingma et al. 2014), paramétré avec un taux d'apprentissage $l_r = 10^{-3}$ fixe. Ce paramètre n'est pas modifié au cours de l'apprentissage [1].

Les modèles de VAD sont évalués à l'aide des taux de fausse alarme (FA), de détection manquée (Miss) et d'erreur de segmentation (SER). Les modèles d'OSD sont évalués à l'aide du F1-score et de la précision moyenne (AP). Enfin, la tâche de SCD est évaluée à l'aide de la pureté (P), de la couverture (C) et du S-score. Ce dernier consiste à calculer la moyenne harmonique des deux précédentes métriques. Le S-score peut être considéré comme un F1-score calculé avec la couverture et la pureté à la place de la précision et du rappel.

6.1.3 Caractéristiques de référence

Dans l'approche considérée (figure 6.2), les caractéristiques acoustiques sont concaténées avec différentes caractéristiques spatiales. Cette sous-section présente les représentations de référence utilisées dans chaque catégorie.

Caractéristiques acoustiques

Deux catégories de caractéristiques sont utilisées pour évaluer l'influence de l'information spatiale sur la segmentation automatique. Les MFCC sont d'abord considérés. Comme dans les précédentes expériences (*cf.* section 5.1.2), 20 coefficients sont extraits sur des fenêtres de 25 ms avec un pas de 10 ms. Le premier coefficient est supprimé. Les dérivées première Δ et seconde $\Delta\Delta$ des coefficients sont également calculées. Un vecteur de $F_a = 59$ caractéristiques est ainsi obtenu.

Le spectrogramme à échelle Mel est considéré comme second type de caractéristique acoustique. Le spectrogramme est calculé sur des fenêtres de 25 ms avec un pas de 10 ms. Il est ensuite converti à l'échelle Mel par $F_a = 128$ filtres triangulaires. Le logarithme de cette représentation est utilisé comme caractéristique (noté Log-Mel).

Les caractéristiques acoustiques sont monocanales. Elles sont donc extraites du premier microphone de l'antenne.

Caractéristiques spatiales

Les caractéristiques spatiales permettent d'obtenir une représentation de la répartition du champ acoustique dans l'espace. Elles sont extraites à partir de plusieurs microphones. Les caractéristiques de référence utilisées sont celles proposées par Cornell et al. (2022a) pour la détection de parole superposée et le comptage de locuteurs.

1. L'utilisation de *scheduler* pour modifier le taux d'apprentissage a été explorée. Cependant, cela a mené à une dégradation des performances.

Les IPD sont d'abord considérées (*cf.* section 3.3.1). Elles sont extraites à partir des $\Pi = 4$ paires de microphones opposés de l'antenne du corpus AMI. Pour garantir l'alignement entre les caractéristiques acoustiques et spatiales, les IPD sont calculées sur des fenêtres de 25 ms avec un pas de 10 ms. La TFCT est calculée sur chaque fenêtre en considérant $n_{fft} = 512$ fréquences. Le vecteur de caractéristiques spatiales résultant est donc composé de $F_s = \Pi \times (n_{fft}/2 + 1)$ caractéristiques. Dans notre cas, cela représente $F_s = 1028$ éléments.

Les CSIPD sont également considérées comme second type de caractéristiques. Elles sont obtenues en calculant le cosinus et le sinus des IPD. Les mêmes paramètres sont conservés pour l'extraction de ces dernières. Le vecteur de caractéristiques ainsi obtenu est composé de $F_s = 2056$ éléments.

6.2 Caractéristiques proposées

Les caractéristiques spatiales telles que les IPD et CSIPD reposent sur le nombre de paires de microphones considérées. Cette représentation est donc sensible au nombre de capteurs disponibles au cours de l'apprentissage.

Dans cette section, nous proposons d'extraire des caractéristiques spatiales en suivant le formalisme des harmoniques circulaires. Il s'agit d'une base de fonctions permettant de décrire le comportement du champ acoustique dans l'espace. Elles permettent d'obtenir des représentations du signal dépendant faiblement du nombre de microphones disponibles. Le jeu de caractéristiques proposé exploite cette propriété. Cette sous-section décrit ce formalisme ainsi que le jeu de caractéristiques utilisé pour la segmentation automatique de la parole.

6.2.1 Formalisme des harmoniques circulaires

Les harmoniques circulaires sont une base de fonctions permettant de décrire la composante angulaire azimutale d'un champ scalaire (ex : champ de pression acoustique). Elles sont similaires aux harmoniques cylindriques ou sphériques représentant l'azimut et l'élévation du champ (Williams et al. 2000). Un signal acoustique peut être représenté en un point de l'espace par une combinaison linéaire d'harmoniques circulaires :

En conditions distantes, la composante radiale du champ est négligeable. Le champ acoustique $X(f, t, \phi)$ est représenté par l'expression suivante :

$$X(f, t, \phi) = \sum_{n=-\infty}^{\infty} C_n(f, t) e^{jn\phi}, \tag{6.2}$$

où f désigne la fréquence et t l'indice d'une trame. Dans l'équation (6.2), $j = \sqrt{-1}$ et $C_n(f, t)$ est le coefficient de l'harmonique circulaire d'ordre n. Les fonctions $e^{jn\phi}$, appelées harmoniques circulaires, forment un ensemble de fonctions orthogonales :

$$\frac{1}{2\pi}\int_0^{2\pi} e^{jn\phi}(e^{ju\phi})^* d\phi = \delta_{nu}, \tag{6.3}$$

avec δ_{nu} l'indice de Kronecker valant l'unité si $n = u$ et zéro sinon et * le conjugué d'un nombre complexe. Cette propriété permet d'exprimer les coefficients :

$$C_n(f,t) = \frac{1}{2\pi}\int_0^{2\pi} X(f,t)e^{-jn\phi} d\phi. \tag{6.4}$$

En pratique, une ACU réalise un échantillonnage spatial discret du champ acoustique. L'intégrale de l'équation (6.4) est donc discrétisée en fonction du nombre de microphones M disponibles. Les coefficients d'harmoniques circulaires sont alors estimés par la relation suivante :

$$\tilde{C}_n(f,t) = \frac{1}{M}\sum_{m=1}^{M} X_m(f,t)e^{-jn\psi_m}, \tag{6.5}$$

où $\tilde{C}_n(f,t)$ est le coefficient estimé et $\psi_m = (m-1)\frac{2\pi}{M}$ représente l'angle du m-ième microphone. Les coefficients sont donc une représentation du signal mesuré $\mathbf{X} = [X_1, \ldots, X_M]$ projeté sur une base d'harmoniques circulaires. Une opération de filtrage spatial peut être réalisée dans le domaine des harmoniques circulaires. Les caractéristiques spatiales utilisées exploitent cette propriété et sont présentées dans la sous-section suivante.

6.2.2 Caractéristiques spatiales proposées

Une opération de filtrage spatial peut être réalisée dans le domaine des harmoniques circulaires (Torres et al. 2016). Cette opération, également appelée formation de voies modale, est définie par la relation :

$$B_N(f,t,\theta) = \sum_{n=-N}^{N} \frac{\tilde{C}_n(f,t)}{j^n J_n(kr)} e^{jn\theta}, \tag{6.6}$$

avec $B_N(f,t,\theta)$ le signal filtré à l'ordre N et $k = 2\pi f/v_s$ le nombre d'onde où v_s représente la vitesse du son. $J_n(kr)$ représente la fonction de Bessel du premier type et d'ordre n avec r le rayon de l'antenne. θ indique ici la direction de focalisation du filtre spatial.

Les caractéristiques spatiales utilisées dans le contexte de la segmentation automatique de la parole consistent à estimer la direction d'arrivée $\tilde{\phi}$ des locuteurs. Cette direction est estimée à partir du vecteur de pseudo-intensité acoustique. Celui-ci n'utilise que les filtres spatiaux d'ordres 0 et 1, rendant son calcul peu sensible au nombre de microphones disponibles. Le filtre spatial à l'ordre 0 est obtenu à l'aide de l'équation (6.6) en posant $N = 0$:

$$B_0(f,t) = \frac{\tilde{C}_0(f,t)}{J_0(kr)}. \tag{6.7}$$

Pour $N = 1$, deux filtres orthogonaux sont respectivement orientés selon les directions $\theta_x = 0$

et $\theta_y = \pi/2$. Le filtre $B_{1x}(f,t,\theta_x)$ (respectivement B_{1y} dans la direction θ_y) s'exprime :

$$B_{1x}(f,t,\theta_x) = \sum_{-1}^{1} \frac{\tilde{C}_n(f,t)}{j^n J_n(kr)} e^{jn\theta_x}. \tag{6.8}$$

Le vecteur de pseudo-intensité acoustique $\mathbf{I}(f,t) = [I_x(f,t), I_y(f,t)]$ est ensuite obtenu par la relation suivante :

$$\begin{bmatrix} I_x(f,t) \\ I_y(f,t) \end{bmatrix} = \frac{1}{2}\Re\left\{ B_0^*(f,t) \begin{bmatrix} B_{1x}(f,t,\theta_x) \\ B_{1y}(f,t,\theta_y) \end{bmatrix} \right\} \tag{6.9}$$

où $\Re$ représente la partie réelle. Le vecteur $\mathbf{I}$ est supposé être orienté dans la direction d'arrivée de l'onde acoustique incidente. La direction angulaire de ce vecteur dans le repère de l'antenne est donc une estimation de la direction $\tilde{\phi}$ de la source active (SongGong et al. 2021) :

$$\tilde{\phi}(f,t) = \arctan\left(\frac{I_y(f,t)}{I_x(f,t)}\right). \tag{6.10}$$

L'estimation de la direction d'arrivée $\tilde{\phi}$ définie dans l'équation (6.10) sert de caractéristique spatiale pour la segmentation automatique de la parole. Ces caractéristiques sont ensuite nommées CH-DOA.

6.2.3 Exemple de représentation obtenue

La figure 6.3 présente un exemple de caractéristiques CH-DOA pour un segment issu des données d'apprentissage du corpus AMI. Les lignes verticales indiquent les frontières entre les différents segments. Cinq zones sont repérées dans le signal. (1) indique un silence, (2) un premier locuteur est actif, suivi d'un second en (3), puis d'un silence (4) et d'une nouvelle activité du locuteur 2 (5). Les caractéristiques CH-DOA présentent un contenu différent entre les segments associés au locuteur 1 (2) et ceux associés au locuteur deux (3 et 5). Dans le premier cas, un plus grand nombre de fréquences sont associées à la direction proche de $-\pi$. Dans le second cas, une majorité de fréquences sont assignées aux directions proches de $+\pi/2$. Les caractéristiques CH-DOA semblent donc adaptées à la segmentation de la parole, notamment pour les détections d'activité vocale et de changement de locuteur.

6.3 Résultats de segmentation

Cette section présente les performances de segmentation automatique de la parole obtenues à l'aide des caractéristiques spatiales basées sur les harmoniques circulaires (CH-DOA). Elles sont comparées aux caractéristiques spatiales de références et combinées avec deux types de caractéristiques acoustiques.

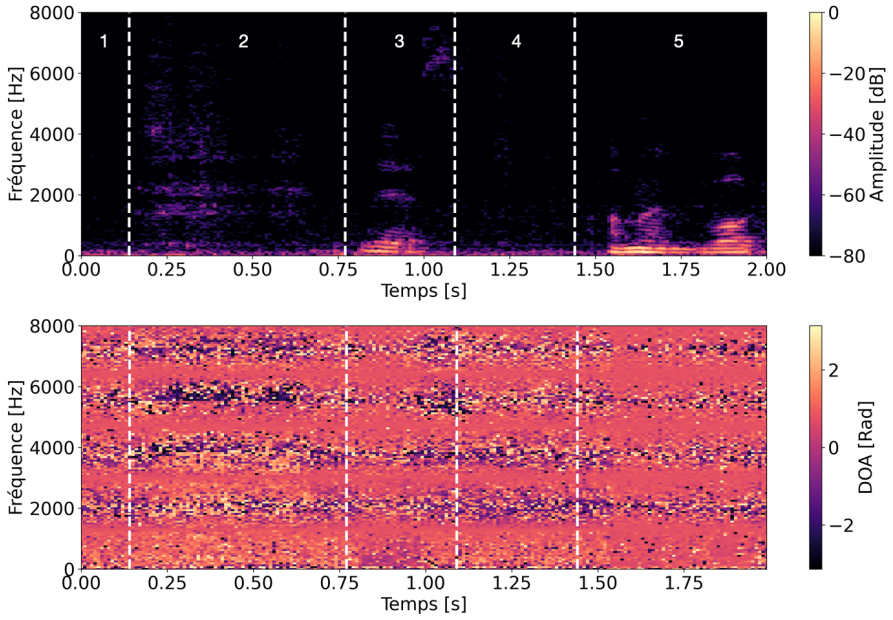


FIGURE 6.3 – Exemple d'une séquence de caractéristiques CH-DOA sur un segment de deux secondes extrait des données d'apprentissage du corpus AMI. (Haut) Spectrogramme du segment de parole, (Bas) caractéristiques CH-DOA. La carte de couleurs représente la direction d'arrivée en fonction de la fréquence et du temps. Les lignes discontinues blanches indiquent des frontières entre les segments de parole et de silence. Les numéros indiquent différents évènements : (1) silence, (2) locuteur 1, (3) locuteur 2, (4) silence, (5) locuteur 2.

6.3.1 Détection d'activité vocale

Le tableau 6.1 présente les performances de chaque système sur la tâche de VAD. La première partie du tableau présente les résultats obtenus en combinant les MFCC avec diverses caractéristiques spatiales. Les MFCC seuls permettent d'obtenir un SER de 6,5% avec des taux de fausse alarme et de détection manquée équilibrés (respectivement 3,5% et 3,0%). L'ajout de caractéristiques spatiales IPD tend à dégrader les performances avec 7,2% de SER. Ce modèle présente notamment un taux de fausse alarme élevé (4,2%). Les performances sont similaires avec les caractéristiques CSIPD obtenant un SER de 7,1%. Ce modèle présente cependant un taux de détection manquée élevé (4,1%), montrant qu'il tend à manquer certaines trames annotées comme contenant de la parole. Les caractéristiques CH-DOA permettent d'obtenir de meilleurs résultats que les IPD ou CSIPD avec un SER de 6,5%, mais sans améliorer le modèle utilisant uniquement les MFCC.

La seconde partie du tableau 6.1 présente les performances de VAD en combinant le spectrogramme à échelle Mel avec les caractéristiques spatiales. Le modèle utilisant uniquement les Log-Mel obtient un SER de 6,7% avec des taux FA et Miss équilibrés (respectivement 3,2% et 3,5%). L'ajout de caractéristiques IPD dégrade les performances de détection avec un SER de 7,3%. Les caractéristiques CSIPD atteignent un score légèrement meilleur qu'avec les MFCC

TABLE 6.1 – Performances de VAD pour chaque configuration sur les données d'évaluation du corpus AMI. FA : taux de fausse alarme, Miss : taux de détection manquée, SER : taux d'erreur de segmentation. Le nombre de paramètres est indiqué en millions (M). Les valeurs en gras indiquent les meilleures performances pour chaque type de caractéristiques acoustiques.

Caractéristiques	# param.	**FA**$_{\%\downarrow}$	**Miss**$_{\%\downarrow}$	**SER**$_{\%\downarrow}$
MFCC	0.26M	**3.5**	**3.0**	**6,5**
+ IPD	0.33M	4.2	**3.0**	7,2
+ CSIPD	0.40M	**3.0**	4.1	7,1
+ CH-DOA	0.28M	**3.4**	**3.1**	**6,5**
Log-Mel	0.27M	**3.2**	**3.5**	6,7
+ IPD	0.33M	**3.0**	4.3	7,3
+ CSIPD	0.40M	3.7	**3.1**	6,8
+ CH-DOA	0.28M	**3.2**	**3.3**	**6,5**

avec 6,8% de SER. L'utilisation des CH-DOA permet de maintenir des performances identiques aux caractéristiques acoustiques seules avec un SER de 6,5%. Elles ne permettent cependant pas d'amélioration des performances de détection.

En conditions distantes, il semble difficile de descendre sous le seuil des 6,5% de SER sur la tâche de VAD. Cette difficulté pourrait être liée à la qualité des annotations dans les données d'évaluation du corpus AMI. Une perspective pour contourner ce problème peut être l'utilisation de pseudo-annotations obtenues à l'aide d'un modèle de VAD pré-entraîné. Ainsi, les étiquettes (annotées manuellement) peuvent être enrichies par les prédictions du modèle (pseudo-annotations). L'utilisation de données artificielles, générées à l'aide d'algorithmes de synthèse de la parole est également une voie d'amélioration envisageable.

6.3.2 Détection de parole superposée

Les performances de chaque système sur la tâche de détection de parole superposée sont présentés dans le tableau 6.2. Comme pour la tâche précédente, la première partie du tableau présente les résultats obtenus à l'aide des caractéristiques MFCC. Ces caractéristiques seules atteignent un F1-score de 64,5% et une AP de 65,1%. En concaténant les MFCC avec les caractéristiques spatiales IPD, les performances de détection sont dégradées avec un F1-score de 60,7% et une AP de 62,5%. Cela représente une dégradation absolue de -3,8% du F1-score par rapport aux MFCC seuls. Les IPD ne permettent donc pas d'améliorer les performances de détection dans cette configuration. A l'inverse, la concaténation avec des caractéristiques CSIPD permet un gain de performance significatif avec un F1-score de 71,7% et une AP de 75,9%. Ce modèle surpasse de +7,2% le F1-score et de +10,8% l'AP obtenus avec les MFCC seuls. Les caractéristiques proposées (CH-DOA) permettent également d'améliorer les performances d'OSD par rapport aux MFCC seuls avec un F1-score de 69,3% et une AP de 73,0%. Le gain n'est cependant pas aussi important qu'avec les CSIPD. Ces caractéristiques permettent cependant de

réduire le nombre de paramètres en passant de 0,4M à 0,28M tout en conservant des performances supérieures au modèle monocanal.

TABLE 6.2 – Performances sur la tâche d'OSD pour chaque type de caractéristique. AP : précision moyenne. Le nombre de paramètres est indiqué en millions (M). Les valeurs en gras indiquent les meilleures performances pour chaque type de caractéristiques acoustiques.

Caractéristiques	# param.	**F1-score**$_{\%\uparrow}$	**AP**$_{\%\uparrow}$
MFCC	0,26M	64.5	65.1
+ IPD	0,33M	60.7	62.5
+ CSIPD	0,40M	**71.7**	**75.9**
+ CH-DOA	0,28M	**69.3**	**73.0**
Log-Mel	0,27M	66.1	68.9
+ IPD	0,33M	65.2	65.9
+ CSIPD	0,40M	**73.4**	**75,6**
+ CH-DOA	0,28M	67.3	68.3

La seconde partie du tableau 6.2 présente la même étude avec un spectrogramme à échelle Mel comme caractéristique spatiale. Le modèle Log-Mel obtient un F1-score de 66,1% et une AP de 68,9%. Cela représente un gain absolu de +1,6% sur le F1-score par rapport aux MFCC. Comme dans le cas des MFCC, l'ajout d'IPD dégrade les performances de détection avec un F1-score de 65,2% et une AP de 65,9%. La concaténation avec les CSIPD permet un gain significatif d'OSD avec un F1-score de 73,4%, soit un gain absolu de +7,3% par rapport au Mel-spectrogramme seul. Les CH-DOA présentent une amélioration par rapport au Log-Mel seul, sans surpasser les CSIPD, avec un F1-score de 67,3%.

Les caractéristiques spatiales sont bénéfiques à la détection de parole superposée. Les CSIPD permettent notamment un gain significatif en performance en atteignant 73,4% de F1-score en les concaténant aux Log-Mel. Les caractéristiques CH-DOA permettent également d'améliorer la détection (AP : 73,0% avec les MFCC), sans atteindre les mêmes performances que les CSIPD.

6.3.3 Détection de changements de locuteur

L'influence des différentes caractéristiques spatiales est évaluée sur la tâche de détection de changement de locuteur. Les résultats sont présentés dans le tableau 6.3. La première partie du tableau présente les résultats obtenus avec les MFCC. En considérant uniquement les caractéristiques acoustiques, le modèle atteint un S-score de 80,7%. Comme pour l'OSD, l'ajout de caractéristiques IPD dégrade la détection avec un S-score de 75,6%, soit une dégradation absolue de -5,1%. Les CSIPD permettent d'améliorer la détection avec un S-score de 83,9% et notamment une couverture de 85,9%. Cela représente un gain absolu de +3,2% sur le S-score par rapport aux MFCC seuls. Les caractéristiques CH-DOA permettent d'atteindre des performances similaires aux CSIPD avec un S-score de 84,4%. La pureté et la couverture sont également équilibrées (P : 84,6%, C : 84,3%). Le modèle semble donc moins sensible au choix du seuil

effectué sur les données de développement.

TABLE 6.3 – Performances sur la tâche de SCD pour chaque type de caractéristique. P : pureté, C : couverture. Les valeurs en gras indiquent les meilleures performances pour chaque type de caractéristiques acoustiques.

Caractéristiques	$\mathbf{P}_{\%\uparrow}$	$\mathbf{C}_{\%\uparrow}$	S-score$_{\%\uparrow}$
MFCC	82.2	79.2	80.7
+ IPD	74.8	76.4	75.6
+ CSIPD	81.9	**85.9**	**83.9**
+ CH-DOA	**84.6**	**84.3**	**84.4**
Log-Mel	83.9	80.4	82.1
+ IPD	79.5	76.8	78.1
+ CSIPD	85.5	**83.9**	**84.7**
+ CH-DOA	**87.2**	**85.6**	**86.4**

La seconde partie du tableau 6.3 présente les résultats de SCD obtenus à l'aide des caractéristiques Log-Mel. Utilisées seules, elles mènent à un S-score de 82,1%. Cela représente un gain absolu de +1,4% par rapport aux MFCC. L'ajout d'IPD dégrade significativement les performances avec S-score de 78,1%, soit une dégradation absolue de -4% par rapport au Log-Mel. Les CSIPD permettent une amélioration de la détection avec un S-score de 84,7%. Les CH-DOA apportent un gain significatif à la SCD avec un S-score de 86,4% avec une pureté de 87,2%. Cela représente des gains absolus de +4,3% et +1,7% sur le S-score par rapport au Log-Mel et aux CSIPD respectivement.

Bien que peu explorées dans la littérature, les caractéristiques spatiales permettent une amélioration significative des performances de détection de changements de locuteur en conditions de parole distante. En particulier, les CH-DOA proposées permettent d'atteindre les meilleures performances avec un S-score de 86,4% quand elles sont combinées au Log-Mel. Elles permettent également une réduction significative du nombre de paramètres du modèle (tableau 6.1) sans perte de performance.

6.4 Robustesse au nombre de canaux

D'après l'équation (6.9), les caractéristiques spatiales CH-DOA utilisent uniquement les harmoniques circulaires aux ordres $n = 0$ et $n = 1$. En théorie, deux microphones suffisent pour représenter le champ acoustique à ces ordres. Par construction, les caractéristiques CH-DOA sont peu sensibles au nombre de capteurs utilisés tant que l'antenne est circulaire. Cette sous-section propose d'évaluer les modèles d'OSD et de SCD basés sur les CH-DOA en faisant varier le nombre de microphones. Ils sont comparés au modèle CSIPD et aux caractéristiques acoustiques seules.

6.4.1 Protocole

Pour évaluer l'impact du nombre de microphones sur les performances de segmentation, les modèles préalablement entraînés dans la section 6.3 sont évalués en supprimant des canaux dans les données d'évaluation du corpus AMI. Trois cas de figure sont considérés :

- $M = 4$: les microphones $\{0, 2, 4, 6\}$ sont conservés,
- $M = 2$: les microphones $\{0, 4\}$ sont conservés,
- *Aléatoire* : les microphones sont sélectionnés aléatoirement au cours de l'évaluation. La même graine aléatoire est utilisée pour chaque modèle. La sélection des canaux est donc identique pour chaque cas.

Dans le cas $M = 4$, l'antenne résultante conserve un espacement uniforme entre les microphones. Cette propriété est importante pour le formalisme des harmoniques circulaires qui requièrent un échantillonnage spatial uniforme pour estimer les coefficients de l'équation (6.5). Dans le cas *Aléatoire*, le nombre de microphones actifs est tiré aléatoirement pour chaque exemple en suivant une loi uniforme $\mathcal{U}(2, M)$. Cette situation ne garantit pas que la géométrie de l'antenne soit circulaire uniforme.

L'évaluation de la robustesse des caractéristiques CSIPD par rapport aux CH-DOA est cependant difficile. Une fois le premier modèle entraîné avec un nombre de paires de microphones donné, le nombre de canaux en entrée du modèle de segmentation ne peut pas être modifié. Deux protocoles d'évaluation de la robustesse sont donc proposés et sont présentés ci-dessous.

Protocole A : masquage des canaux Ce protocole s'adresse aux modèles utilisant les caractéristiques spatiales CSIPD, pour lesquels les canaux ne peuvent pas être retirés du signal. La robustesse des modèles est alors évaluée en désactivant certains canaux, c'est-à-dire en plaçant leurs valeurs à zéro. Les modèles présentés en section 6.3 ne sont cependant pas entraînés sur des données masquées. Deux types de modèles sont alors entraînés afin d'évaluer la robustesse des CSIPD au nombre de canaux actifs :

- le premier est entraîné avec tous les canaux disponibles comme ceux présentés en section 6.3. Les canaux sont ensuite masqués lors de l'évaluation ;
- le second est entraîné en masquant les canaux aléatoirement au cours de l'apprentissage. Les canaux sont ensuite masqués au cours de l'évaluation.

Protocole B : suppression des canaux Ce protocole s'adresse aux modèles utilisant les caractéristiques spatiales CH-DOA. Les caractéristiques proposées permettent de retirer les canaux du signal. Aucun masquage n'est donc appliqué dans ce cas. Tous les canaux disponibles sont utilisés au cours de l'apprentissage du modèle. Seuls les canaux utiles sont conservés lors de son évaluation. Notons cependant que la position angulaire des microphones de l'antenne est supposée inconnue lors de l'évaluation. Lorsque certains capteurs sont retirés, il n'est plus garanti que l'espacement entre les microphones restant soit uniforme. Dans le cas *Aléatoire*,

nous supposons cependant que les microphones sont uniformément espacés. Cette considération introduit donc des erreurs de géométrie, sources d'erreur dans le calcul des CH-DOA. Le rayon de l'antenne est supposé identique à celui de l'apprentissage.

Les protocoles **A** et **B** sont appliqués pour évaluer l'influence du nombre de microphones sur les tâches d'OSD et de SCD en considérant les deux types de caractéristiques acoustiques préalablement étudiées : Log-Mel et MFCC. Les performances des modèles n'utilisant qu'un seul microphone (MFCC et Log-Mel, $M = 1$) sont également reportées. Elles permettent d'évaluer la robustesse des systèmes multicanaux par rapport au cas monocanal en fonction du nombre de microphones actifs.

6.4.2 Résultats

Cette sous-section présente l'influence du nombre de microphones sur les performances de détection de parole superposée et de changement de locuteur. Les évaluations sont réalisées avec les deux types de caractéristiques acoustiques. Dans les tableaux, les scores en gras indiquent les meilleurs scores pour chaque configuration d'antenne. Cependant, les protocoles d'évaluation **A** et **B** sont différents. Les comparaison des scores entre les deux protocoles doit être considérée avec prudence.

Détection de parole superposée

La table 6.4 présente les performances d'OSD en fonction du nombre de microphones sélectionnés. La première partie du tableau présente l'influence de la suppression de canaux lorsque que les caractéristiques spatiales CSIPD sont utilisées. L'évaluation est donc réalisée à l'aide du protocole **A**.

Lorsque les CSIPD sont concaténées avec les MFCC, le masquage aléatoire des canaux améliore légèrement les performances dans le cas où tous les microphones sont activés ($M = 8$). Le modèle avec masquage atteint un F1-score de 72,3% contre 71,7% pour le modèle sans masquage. Le gain reste cependant minime, comme le montre également la précision moyenne avec seulement 0,2% d'écart absolu. Le masquage des canaux apporte un gain en cas de désactivation des canaux. Par exemple, lorsque $M = 4$ canaux sont conservés, le masquage permet un gain absolu de +17,3% sur la précision moyenne. Les performances restent cependant fortement dégradées quand $M = 2$ (AP : 38,8%). Dans le cas où le masquage des canaux est aléatoire, le modèle conserve des performances décentes (F1-score : 63,8%), contrairement au modèle sans masquage.

En cas d'utilisation du spectrogramme à échelle Mel, les observations sont similaires. Le masquage permet de limiter la dégradation jusqu'à $M = 4$ (+6,3% sur l'AP) et dans le cas d'un masquage aléatoire (gain absolu de +28,1% sur l'AP). Les performances sont cependant dégradées pour $M = 2$.

Pour chaque cas où le nombre de microphones disponible est réduit, les performances des

modèles multi-microphones sont dégradées par rapport au cas monocanal. Par exemple, le modèle CSIPD+Log-Mel avec masquage présente une dégradation absolue de -3,1% sur l'AP pour $M = 4$ par rapport au cas monocanal. Seul le cas CSIPD+MFCC avec masquage conserve des performances légèrement supérieures pour $M = 4$ avec un gain absolu de +3,7% sur l'AP. Il semble donc préférable d'utiliser les CSIPD dans un cas où la configuration des capteurs est identique à celle de l'apprentissage, sans être susceptible de varier.

TABLE 6.4 – Performance de détection de parole superposée en fonction du nombre de microphones disponibles. Performances dans le cas $M = 1$: **MFCC** F1-score 64,5%, AP 65,1% - **Log-Mel** F1-score 66,1%, AP 68,9% (*cf.* table 6.2). Les valeurs en gras indiquent les meilleures performances.

			$M = 8$		$M = 4$		$M = 2$		*Aléatoire*	
Spat.	Acous.	Masquage	**F1**$_{\%\uparrow}$	**AP**$_{\%\uparrow}$	**F1**$_{\%\uparrow}$	**AP**$_{\%\uparrow}$	**F1**$_{\%\uparrow}$	**AP**$_{\%\uparrow}$	**F1**$_{\%\uparrow}$	**AP**$_{\%\uparrow}$
Protocole **A**										
CSIPD	MFCC	✗	71,7	**75,9**	54,2	51,5	n.d.	n.d.	6,61	38,8
		✓	72,3	75,7	63,7	68,8	n.d.	n.d.	63,8	66,9
	Log-Mel	✗	**73,4**	75,6	46,9	59,5	26,0	47,1	29,6	31,0
		✓	69,5	70,7	55,3	65,8	n.d	n.d.	61,3	59,1
Protocole **B**										
CH	MFCC	✗	69,3	73,3	**69,6**	**71,4**	64,7	65,5	53,2	55,7
	Log-Mel	✗	67,3	68,3	64,3	68,2	**65,5**	**67,5**	65,8	69,0

La seconde partie du tableau présente les performances obtenues à l'aide des caractéristiques CH-DOA (notées CH ici) en suivant le protocole **B**. Dans le cas où les CH-DOA sont concaténées avec les MFCC, la suppression de certains canaux dégrade faiblement les performances. Dans le cas $M = 4$, le modèle obtient une AP 71,4% et un F1-score de 69,6%. Les performances se dégradent dans le cas $M = 2$ avec un F1-score de 64,7% et une AP de 65,5%. La dégradation est encore plus marquée dans le cas où le nombre de canaux est aléatoire avec une AP de 55,7%. Jusqu'à $M = 2$, le modèle multicanal présente des performances supérieures ou similaires au cas monocanal (AP 65,1%). Les caractéristiques CH-DOA améliorent donc la robustesse au nombre de capteurs avec les MFCC. Cette robustesse est cependant contrainte par la géométrie de l'antenne. En effet, dans le cas où le nombre de capteur est aléatoire, les performances se dégradent significativement. L'antenne, supposée circulaire uniforme pour le calcul des caractéristiques, ne respecte plus ces conditions.

Dans le cas du spectrogramme à échelle Mel, les CH-DOA obtiennent un F1-score de 64,3% et une AP de 68,2% dans le cas $M = 4$. Cela représente une dégradation de -3% sur la première métrique par rapport au cas $M = 8$. Les scores de détection restent stables dans la cas $M = 2$ avec un F1-score de 65,5% et une AP de 67,5%. Avec ce système, les performances pour un nombre de microphones aléatoires sont proches du cas $M = 8$ et équivalentes au cas monocanal lorsque l'AP est considérée (69,0%). Le F1-score présente une dégradation entre les cas $M = 8$ et

aléatoire. Quelle que soit la métrique considérée, les résultats montrent que ce système n'utilise pas efficacement les données spatiales.

L'influence de la désactivation des canaux est également évaluée sur la tâche de SCD dans la sous section suivante.

Détection de changements de locuteur

La table 6.5 présente le S-score de chaque modèle en fonction du nombre de microphones utilisés. La première partie du tableau présente les résultats obtenus avec les CSIPD et chaque type de caractéristique acoustique avec le protocole **A**.

Dans le cas des MFCC, le modèle sans masquage de canaux présente une dégradation entre le cas M = 8 (83,9%) et les cas M = 4 (78,9%) et M = 2 (61,2%). Une dégradation est également observée dans le cas où les canaux sont masqués aléatoirement (77,1%). Le masquage des canaux au cours de l'apprentissage tend à dégrader les performances du modèle, avec un S-score de 80,2% dans le cas M = 8. Le masquage n'améliore pas la robustesse dans le cas où certains microphones sont désactivés (ex : 56,3% pour M = 2).

Les observations sont différentes avec les caractéristiques acoustiques Log-Mel. Sans masquage des canaux, le modèle atteint 84,7% de S-score avec une dégradation à 79,6% pour M = 4 et 72,3% dans le cas M = 2. L'ajout de masquage des canaux dégrade légèrement les performances dans le cas M = 8 avec 81,3% soit -3,4% de dégradation absolue. Dans les cas M = 4 et *Aléatoire*, les performances sont améliorées avec un gain absolu de +2% et +6,1% respectivement par rapport au modèle sans masquage.

L'ajout de caractéristiques spatiales CSIPD améliore la SCD lorsque tous les microphones sont actifs, sans masquage de canaux au cours de l'apprentissage. Dans les autres cas, les caractéristiques spatiales CSIPD difficilement avec le modèle monocanal obtenant un S-score de 80,7% avec les MFCC et 82,1% avec les Log-Mel.

La seconde partie du tableau présente les résultats obtenus en supprimant des canaux avec les caractéristiques CH-DOA. Le protocole **B** est donc utilisé. Comme sur la tâche d'OSD, ce modèle offre une meilleure robustesse en cas de suppression de canaux. Avec les MFCC, les performances restent stables dans le cas M = 4 (85,2%) par rapport au cas M = 8 (84,4%). Elles se dégradent cependant pour M = 2 (75,5%). Dans ce dernier cas de figure, le cas monocanal est meilleur (80,7%). Les performances restent cependant supérieures à tous les modèles basés sur les CSIPD dans la même configuration. Des performances similaires sont obtenues en cas de suppression aléatoire des canaux avec 75,4%. Comme pour l'OSD, le modèle MFCC manque de robustesse dans le cas *Aléatoire*. Cela provient probablement d'un décalage entre la géométrie réelle de l'antenne et celle utilisée dans le calcul.

Dans le cas où les Log-Mel sont utilisés, le modèle offre des performances similaires dans chaque configuration. Il atteint notamment 86,2% de S-score dans le cas M = 4. Les performances se dégradent cependant dans le cas *Aléatoire* avec 78,3%. Le modèle basé sur les Log-Mel offre cependant une meilleure robustesse que les MFCC et offre des performances supérieures aux

TABLE 6.5 – Performance de détection de changements de locuteur en fonction du nombre de microphones disponibles. Seul le S-score$_{\%\uparrow}$ est présenté. Performances dans le cas $M = 1$: **MFCC** S-score 80,7% - **Log-Mel** S-score 82,1% (*cf.* table 6.3). Les valeurs en gras indiquent les meilleures performances.

Spat.	Acous.	Masquage	$M = 8$	$M = 4$	$M = 2$	*Aléatoire*
Protocole **A**						
CSIPD	MFCC	✗	83,9	78,9	61,2	77,1
		✓	80,2	70,8	56,3	75,6
	Log-Mel	✗	84,7	79,6	72,3	72,3
		✓	81,3	81,6	70,7	**78,4**
Protocole **B**						
CH	MFCC	✗	84,4	85,2	75,5	75,4
	Log-Mel	✗	**86,4**	**86,2**	**84,0**	**78,3**

Log-Mel monocanal (82,1%) dans tous les cas à l'exception de l'*Aléatoire*. Contrairement à la tâche d'OSD, le modèle de SCD semble utiliser les caractéristiques spatiales efficacement pour la détection.

6.5 Évaluation sur la diarisation en locuteurs

Cette section évalue les performances de diarisation en locuteur obtenues à partir de la segmentation (VAD et OSD) obtenue avec les caractéristiques spatiales CSIPD et CH-DOA combinées aux MFCC. La diarisation en locuteur est réalisée avec le système VBx (LANDINI et al. 2022b). Le protocole d'évaluation est identique à celui utilisé en section 5.7. Sauf mention contraire, les DER présentés sont ceux obtenus dans le cas $\delta_c = 0$ et les écarts entre les scores sont relatifs.

6.5.1 Résultats

La table 6.6 présente le DER et le JER obtenus pour chaque modèle sur le sous-ensemble d'évaluation du corpus AMI. Les résultats montrent que le modèle CSIPD atteint les meilleures performances en DER dans les cas VAD et VAD+OSD. Dans le premier cas, ce modèle obtient un DER de 27,01%, soit un gain de +1,6% par rapport au MDU. Dans le second cas, l'assignation de la parole superposée permet d'atteindre 23,63%, ce qui représente un gain de +5,4% par rapport au MDU.

Le modèle CH-DOA obtient des performances intermédiaires entre le MDU et les CSIPD. Dans le scénario selon lequel seule la VAD est considérée, il atteint un DER de 27,42% et un JER de 33,73%. L'assignation des segments de parole superposée permet d'améliorer le DER de +10,8%. Ce modèle atteint également le meilleur JER avec 31,91%.

TABLE 6.6 – Performances de diarisation en locuteur pour chaque système de segmentation de la parole distante. Les résultats sont obtenus sur le sous-ensemble de développement du corpus AMI. Les caractéristiques spatiales sont concaténées aux caractéristiques acoustiques MFCC.

Système	**Segmentation**	**DER**$_{\%\downarrow}$		**JER**$_{\%\downarrow}$
		$\delta_c = 25$	$\delta_c = 0$	
Oracle	VAD	15,63	22,47	30,52
	VAD+OSD	10,12	14,32	25,78
MDU	VAD	19,38	27,45	33,82
	VAD+OSD	17,60	24,97	32,15
CSIPD	VAD	18,93	27,01	34,15
	VAD+OSD	**16,55**	**23,63**	32,35
CH-DOA	VAD	19,05	27,42	33,73
	VAD+OSD	16,91	24,45	**31,91**

6.5.2 Discussions

Les caractéristiques spatiales permettent d'améliorer la diarisation en cas de parole distante. Les CSIPD améliorent nettement les performances, notamment lorsque les segments de parole superposés sont assignés. Les CH-DOA permettent également une légère amélioration.

Bien que le système CH-DOA obtienne un SER inférieur aux CSIPD sur la tâche de VAD (cf. tableau 6.1), le DER obtenu avec les CH-DOA est plus élevé. Le modèle CSIPD présente cependant un taux de fausse alarme inférieur, ce qui peut expliquer l'écart de performances. D'autre part, le système CH-DOA obtient les meilleurs JER. Cette métrique évalue les erreurs de segmentation par locuteur. Les CH-DOA semblent donc réduire les erreurs de segmentation intra-locuteur.

6.6 Conclusions

Dans ce chapitre, nous proposons l'utilisation de caractéristiques spatiales pour différentes tâches de segmentation : VAD, OSD and SCD. Plusieurs types de caractéristiques sont explorés. Pour répondre à la première question de recherche (**Q1**), les travaux menés ont montré que les CSIPD améliorent significativement la détection de changements de locuteurs, en atteignant un S-score de 84,7% avec le meilleur modèle.

De nouvelles caractéristiques spatiales sont également proposées, en utilisant le formalisme des harmoniques circulaires (CH-DOA). Elles permettent d'obtenir une estimation de la direction angulaire de la source en ne dépendant que des ordres $n = 0$ et $n = 1$ des harmoniques circulaires, réduisant ainsi la dépendance au nombre de microphones disponibles. Les CH-DOA améliorent la VAD (SER : 6,5%) par rapport aux IPD et CSIPD. Elles permettent également d'améliorer les performances en OSD (AP : 73,0%) et SCD (S-score : 86,4%) par rapport

aux caractéristiques acoustiques seules (monocanal). Sur cette dernière tâche, elles surpassent également les caractéristiques spatiales de référence (CSIPD). Pour répondre à la question **Q2**, les caractéristiques CH-DOA améliorent les performances de segmentation sur toutes les tâches considérées par rapport au cas monocanal. Elles surpassent également les CSIPD sur les tâches de VAD et de SCD.

Par construction, les harmoniques circulaires dépendent faiblement du nombre de microphones. Elles améliorent donc la robustesse des systèmes lorsque le nombre de capteurs varie. Une analyse de leur robustesse est conduite et montre que les performances restent stables pour $M = 4$ et $M = 2$. Elles se dégradent cependant lorsque le nombre de capteurs est aléatoire. Dans ce dernier cas, l'antenne n'est plus circulaire et limite les capacités de modélisation du champ acoustique. Les caractéristiques CH-DOA proposées sont donc robustes au nombre de microphones disponibles (**Q3**) à condition de conserver une géométrie circulaire. Une étude annexe évalue la robustesse des caractéristiques CSIPD au nombre de capteurs à l'aide d'un protocole différent (masquage des canaux). Les résultats montrent que les CSIPD ne permettent pas d'atteindre la même robustesse que les CH-DOA dans la plupart des situations.

Comme dans le cas de la combinaison de canaux (*cf.* chapitre 5), l'amélioration de la segmentation à l'aide des caractéristiques spatiales apporte un gain pour la diarisation en locuteur. Les CSIPD conduisent au meilleur DER en conditions de parole distante. L'assignation de la parole superposée est particulièrement efficace pour ce système. Les CH-DOA apportent un gain mitigé, mais permettent d'obtenir le meilleur JER. L'utilisation de caractéristiques spatiales en conditions distantes a donc un impact positif sur la diarisation en locuteur (**Q4**).

Dans l'objectif de rendre les modèles de segmentation robustes quel que soit le nombre de canaux disponibles, le chapitre suivant introduit une nouvelle méthode d'apprentissage. Celle-ci contraint les caractéristiques extraites à l'aide d'un modèle de type SACC (combinaison de canaux auto-attentive) à être similaires quel que soit le nombre de microphones disponibles. De plus, les modèles pré-entrainés par apprentissage autosupervisé permettent des gains de performance remarquables sur de nombreuses tâches de traitement automatique de la parole en champ proche (S. Chen et al. 2022 ; Feng et al. 2023). Le chapitre suivant propose d'utiliser les caractéristiques pré-entrainées Wavlm (S. Chen et al. 2022) dans le contexte de l'OSD en conditions distantes. Le traitement, réalisé sur un seul canal de l'antenne, devient monocanal et ne dépend plus de la géométrie de l'antenne.

Chapitre 7

Vers une généralisation à la géométrie de l'antenne

Développer des méthodes de traitement automatique de la parole exploitant plusieurs microphones requiert de garantir leur robustesse au nombre de capteurs disponibles. En conditions d'utilisation réelles, un système est susceptible d'être utilisé avec un dispositif de captation différent de celui ayant acquis les données d'apprentissage. Certains microphones peuvent également cesser de fonctionner au cours de l'utilisation. Les performances des modèles doivent alors être conservées malgré ces aléas. Ce chapitre présente deux approches permettant de limiter la dégradation des performances lorsque que l'antenne est différente ou que certains microphones sont désactivés. La première consiste en une méthode d'apprentissage contraignant le modèle SACC à être robuste au nombre de canaux disponibles. Cette contrainte est permise par l'ajout d'une fonction de perte maximisant la similarité entre les caractéristiques obtenues avec tous les microphones et des caractéristiques dégradées suite au masquage de certains canaux. La seconde exploite les capacités de modélisation du modèle pré-entraîné Wavlm pour réaliser la segmentation sur un canal unique. Cette dernière approche n'exploite donc plus l'information spatiale, mais permet d'évaluer les capacités de modélisation des modèles pré-entraînés sur des données distantes. Ces deux approches sont respectivement détaillées et évaluées respectivement en sections 7.1 et 7.2.

Questions de recherche

Ce chapitre vise à répondre aux questions de recherche suivantes :

Q1 Une approche par mesure de similarité entre représentations permet-elle d'améliorer la robustesse en cas de microphones désactivés ? (Section 7.1)

Q2 Quelle mesure de similarité permet de contraindre le modèle à être invariant au nombre de microphones ? (Section 7.1)

Q3 Les systèmes invariants au nombre de capteurs obtiennent-ils de meilleures performances en cas de captation par une antenne non conforme ? (Section 7.1)

Q4 Les caractéristiques Wavlm permettent-elles d'obtenir de bonnes performances de segmentation à partir d'un microphone distant unique ? (Section 7.2)

7.1 Représentation indépendante du nombre de canaux pour la segmentation

Dans les jeux de données acquis au cours de réunions et disponibles librement (Carletta et al. 2006 ; Fu et al. 2021), les signaux sont enregistrés à l'aide d'une même antenne, quels que soient les sous-ensembles considérés (apprentissage ou évaluation). Les données présentent alors peu de variabilité dans les dispositifs d'acquisition utilisés. En cas d'utilisation d'un système de traitement automatique de la parole en conditions réelles, le nombre de microphones et la géométrie de l'antenne peuvent varier. Cependant, les méthodes de combinaison auto-attentive de canaux (ex : SACC) présentées dans le chapitre 5 reposent sur un nombre fixe de microphones. Les performances de ces algorithmes sont donc susceptibles de se dégrader lorsque certains capteurs sont désactivés ou que l'antenne change.

Les caractéristiques spatiales CH-DOA proposées en section 6.2 améliorent déjà la robustesse au nombre de microphones. Cependant, les performances de détection de parole superposée obtenues avec ces caractéristiques n'égalent pas les approches de combinaison de canaux SACC. Un autre axe est donc proposé afin de rendre les caractéristiques SACC robustes au nombre de capteurs actifs.

Cette section introduit une méthode d'apprentissage contraignant le modèle à apprendre une représentation du signal unique, quel que soit le nombre de microphones disponibles.

7.1.1 Formulation

Soit $g : \mathbf{x}, \boldsymbol{\Theta} \rightarrow \mathbf{X}$ un extracteur de caractéristiques multicanales à partir du signal audio $\mathbf{x}$ acquis par Mcanaux disponibles dans les données d'apprentissage. L'apprentissage invariant au nombre de canaux détermine un ensemble de paramètres $\boldsymbol{\Theta}$ permettant d'extraire un vecteur de caractéristiques similaire quel que soit le nombre de canaux actifs. La procédure d'apprentissage est décrite ci-après et schématisée en figure 7.1. Les paramètres des modèles SACC représentés sur la figure sont partagés et permettent d'extraire différentes représentations d'un même signal d'entrée dont certains canaux ont été masqués.

Le segment audio d'entrée, $\mathbf{x}$, est dupliqué P fois. Les versions dupliquées du signal sont notées $\bar{\mathbf{x}}_p$, avec $p = 1, \cdots, P$ l'indice de la duplication. Une partie des canaux des segments dupliqués $\bar{\mathbf{x}}_p$ est ensuite masquée aléatoirement. Le nombre de canaux conservés du p-ième segment est noté M_p. Ce paramètre suit une distribution uniforme discrète $M_p \sim \mathbb{U}(2; M)$. Le masquage consiste à remplacer les valeurs des canaux masqués par des zéros. Les caractéristiques $\bar{\mathbf{X}}_p$ sont ensuite extraites du segment masqué $\bar{\mathbf{x}}_p$ à l'aide de la fonction g. Les caractéristiques issues des segments masqués sont rassemblées dans un ensemble $\mathcal{D} = \{\bar{\mathbf{X}}_p\}_{p=1}^{P}$. Un masquage aléatoire est également appliqué au segment de référence $\mathbf{x}$. Une fonction de perte $\mathcal{L}_{inv}(\mathbf{X}, \mathcal{D})$ mesure la distance entre les caractéristiques $\mathbf{X}$ extraites du segment original $\mathbf{x}$ et chaque caractéristique de $\mathcal{D}$. Le modèle de segmentation apprend à minimiser la distance entre les caractéristiques de référence et les caractéristiques masquées afin d'obtenir une représentation commune, invariante au nombre de

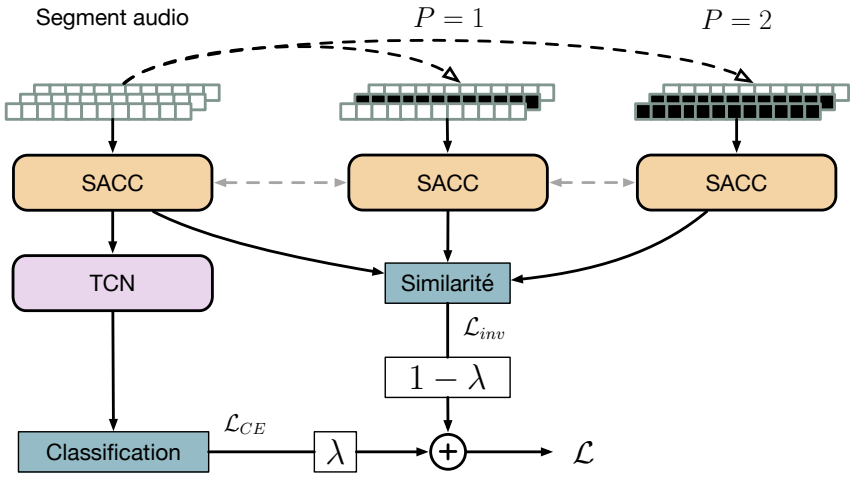


FIGURE 7.1 – Diagramme de la méthode d'apprentissage du modèle de segmentation invariant au nombre de canaux. Les traits pointillés grisés indiquent que les poids sont partagés entre chaque instance du modèle SACC. Les flèches creuses avec une ligne discontinue indiquent les opérations de duplication et de masquage du segment temporel.

microphones actifs.

Deux fonctions de perte sont explorées afin d'apprendre une représentation indépendante du nombre de canaux. La première fonction consiste à minimiser la distance entre la représentation de référence $\mathbf{X}$ et chaque élément de $\mathcal{D}$:

$$\mathcal{L}_{inv}^{F}(\mathbf{X}, \mathcal{D}) = \frac{1}{P} \sum_{p=1}^{P} \frac{\|\mathbf{X} - \bar{\mathbf{X}}_p\|_F^2}{\|\mathbf{X}\|_F^2 \cdot \|\bar{\mathbf{X}}_p\|_F^2}, \tag{7.1}$$

où $\|\cdot\|_F$ représente la norme de Frobenius d'une matrice. La seconde approche consiste à maximiser la similarité cosinus entre les caractéristiques de référence et les caractéristiques masquées :

$$\mathcal{L}_{inv}^{\cos}(\mathbf{X}, \mathcal{D}) = -\frac{1}{PT} \sum_{p=1}^{P} \sum_{t=1}^{T} \frac{\langle \mathbf{X}[t], \bar{\mathbf{X}}_p[t] \rangle}{\max\left(\|\mathbf{X}[t]\|, \|\bar{\mathbf{X}}_p[t]\|\right)}, \tag{7.2}$$

avec $\langle \cdot, \cdot \rangle$ le produit scalaire et t l'indice de la trame courante.

L'apprentissage du modèle de classification et de la représentation invariante est réalisé en suivant le formalisme de l'apprentissage multitâche. En notant $\mathcal{L}_{CE}(\mathbf{y}, \tilde{\mathbf{y}})$ la fonction de perte de classification (entropie croisée), l'objectif d'apprentissage final du modèle s'exprime :

$$\mathcal{L} = \lambda \mathcal{L}_{CE}(\mathbf{y}, \tilde{\mathbf{y}}) + (1 - \lambda) \mathcal{L}_{inv}(\mathbf{X}, \mathcal{D}), \tag{7.3}$$

avec $\lambda \in [0, 1]$ un paramètre de pondération entre l'objectif de classification et celui d'une représentation invariante. $\mathcal{L}_{inv}$ représente une fonction de perte d'invariance quelconque.

L'ajout d'une permutation aléatoire des canaux peut être considéré lors de l'apprentissage invariant. Cette considération peut renforcer l'apprentissage d'une représentation invariante au nombre de microphones et à l'antenne considérés. Cependant, l'étude menée en annexe C montre que la permutation n'améliore pas la robustesse du modèle. Elle n'est donc pas considérée par la suite.

7.1.2 Protocole d'évaluation

L'apprentissage invariant est appliqué à la tâche de détection de parole superposée OSD. Les deux fonctions de perte proposées en équations (7.1) et (7.2) sont étudiées pour l'apprentissage avec plusieurs modèles SACC. Les protocoles d'apprentissage et d'évaluation sont présentés dans cette sous-section.

Apprentissage

Trois modèles d'OSD sont entraînés dans le contexte de l'invariance au nombre de canaux : SACC+TFCT, SACC+$\mathcal{A}_{32}$ et SACC+$\mathcal{A}_{128}$. Les expériences sont menées uniquement avec l'architecture TCN utilisée dans les chapitres 5 et 6. Le protocole d'apprentissage est identique à celui décrit en section 4.3.

Modèles de référence Les performances des modèles entraînés à l'aide de l'apprentissage invariant sont comparées à celles obtenues selon deux protocoles d'apprentissage. Le premier consiste à entraîner chacun des modèles avec tous les microphones disponibles. Ce modèle n'a donc jamais observé de canaux masqués. Ces modèles sont nommés uniquement par la représentation fréquentielle utilisée (ex : $\mathcal{A}_{32}$).

Le second consiste à masquer les canaux aléatoirement au cours de l'apprentissage, sans considérer l'apprentissage multitâche défini en équation (7.3). Dans ce cas, les canaux conservés sont tirés aléatoirement en suivant une distribution uniforme discrète. Les autres canaux sont masqués. Ce modèle observe donc le masquage des canaux. Il est espéré qu'il soit plus robuste lorsque certains canaux sont masqués dans les données d'entrée. Ces modèles sont nommés par la représentation fréquentielle utilisée suivie d'une étoile (ex : $\mathcal{A}^*_{32}$).

Apprentissage invariant L'apprentissage des modèles invariants utilise la fonction de coût (7.3). Le calcul des fonctions de perte définies en équations (7.1) et (7.2) requiert de dupliquer les segments audio afin d'appliquer un masquage aléatoire des canaux. Pour les expériences menées, les segments sont dupliqués $P = 2$ fois pour permettre l'apprentissage d'une représentation invariante. Le facteur de pondération pour l'apprentissage multitâche est fixé à $\lambda = 0.7$ afin de favoriser la classification par rapport à l'apprentissage invariant. Les modèles invariants sont présentés avec la notation *représentation fréquentielle+fonction d'invariance* (ex : $\mathcal{A}_{32} + \mathcal{L}^F_{inv}$).

Évaluation

L'évaluation est réalisée sur le sous-ensemble de test du corpus AMI. Les résultats sur les données de développement sont également présentés pour évaluer les capacités de généralisation du système. Tous les modèles sont évalués en suivant ce protocole. Une première évaluation est menée sur les signaux issus de l'antenne 1. Il s'agit d'une ACU composée de 8 microphones avec un rayon r = 10 cm. Afin de mesurer l'influence de l'apprentissage invariant, chaque modèle est évalué avec un nombre variable de canaux activés. Les configurations suivantes sont considérées :

- M = 8 : tous les microphones de l'antenne sont activés,
- M = 4 : les microphones $\{0, 2, 4, 6\}$ sont conservés,
- M = 2 : les microphones $\{1, 4\}$ sont conservés.

Cette analyse vise à évaluer la robustesse de la méthode en conservant une configuration proche de l'apprentissage (antenne 1). Elle correspond au cas où certains microphones de l'antenne cesseraient de fonctionner. De plus, la désactivation des microphones revient à modifier la géométrie de l'antenne.

Afin d'évaluer les modèles en considérant des antennes non conformes, ces derniers sont également évalués sur les signaux issus de l'antenne 2. Il s'agit d'une antenne circulaire ou linéaire en fonction des réunions considérées [1]. Quelle que soit sa géométrie, cette antenne est toujours composée de 4 microphones. Elle est placée à une extrémité de la table. La distance entre les locuteurs et l'antenne est donc accrue par rapport à l'antenne 1. La répartition des locuteurs autour du dispositif de captation est également modifiée par rapport aux données issues de l'autre antenne.

Outre les points présentés ci-dessus, les métriques et le protocole d'évaluation sont identiques à ceux de la section 4.3.

7.1.3 Impact de la désactivation des canaux

Cette section présente les performances de détection de parole superposée (OSD) des systèmes SACC+TFCT, SACC+$\mathcal{A}_{32}$ et SACC+$\mathcal{A}_{128}$ dans chaque configuration d'apprentissage. L'impact de l'invariance sur les performances est évalué sur les données AMI issues de l'antenne 1 en suivant le protocole précédemment décrit.

Activation de tous les microphones

La table 7.1 présente les résultats sur la tâche d'OSD avec M = 8 obtenus pour chaque configuration. Dans le cas du modèle TFCT*, le masquage des canaux améliore légèrement les performances, notamment sur les données d'évaluation (69,5%). L'ajout d'une fonction d'invariance améliore également les performances lorsque tous les microphones sont actifs. La

1. La géométrie d'antenne pour chaque réunion n'est pas annotée.

fonction $\mathcal{L}_{inv}^{F}$ permet d'obtenir un F1-score de 73,3% et 69,5% respectivement sur les données de développement et d'évaluation. Le modèle d'origine atteint 68,8% sur ce dernier sous-ensemble. La similarité cosine ($\mathcal{L}_{inv}^{cos}$) permet des performances similaires avec un F1-score de 69,2% en évaluation.

Le simple masquage des canaux $\mathcal{A}_{32}^{*}$ améliore la détection avec un gain de +2,1% en évaluation. Cette augmentation des données semble permettre l'apprentissage de filtres plus adaptés pour l'OSD.

L'ajout de la fonction de perte $\mathcal{L}_{inv}^{F}$ permet un gain absolu de +1,5% sur le F1-score d'évaluation avec un score de 66,9%. Elle n'améliore donc pas la détection par rapport au masquage des canaux $\mathcal{A}_{32}^{*}$. La fonction $\mathcal{L}_{inv}^{cos}$ permet une amélioration des performances plus importante sur les données d'évaluation avec un F1-score de 68,3%, soit un gain absolu de +2,9% par rapport au modèle original. Elle renforce la généralisation entre développement et évaluation avec une dégradation de seulement 1,6% entre les deux jeux de données. Cette dernière fonction permet d'obtenir des résultats similaires au modèle TFCT.

Dans le cas des filtres $\mathcal{A}_{128}$, le masquage des canaux * dégrade la détection. La fonction de perte $\mathcal{L}_{inv}^{F}$ permet cependant d'atteindre les meilleures performances avec 66,8% de F1-score sur les données d'évaluation. Cela représente un gain absolu de +7,4% par rapport au modèle original (59,4%). La fonction $\mathcal{L}_{inv}^{cos}$ apporte également une amélioration avec un F1-score de 66,2% sur les données d'évaluation, soit un gain absolu de +6,8% par rapport au modèle original.

Le simple masquage aléatoire des canaux renforce les modèles d'OSD lorsque tous les microphones sont actifs. Dans certains cas, notamment avec les modèles analytiques, les fonctions de perte imposant une contrainte d'invariance au nombre de microphones améliorent davantage les performances. Cela est particulièrement visible avec le modèle $\mathcal{A}_{128}$ dont les performances sont nettement améliorées (jusqu'à +7,4% de gain sur le F1-score d'évaluation). L'apprentissage d'une représentation indépendante du nombre de canaux permet probablement de régulariser le modèle au cours de l'apprentissage en limitant le sur-apprentissage sur la tâche de classification.

Activation de quatre microphones

Le tableau 7.2 présente les performances d'OSD avec chaque approche SACC et chaque fonction d'invariance dans le cas où $M = 4$ microphones sont activés. Le modèle d'origine basé sur la TFCT présente une forte dégradation des performances avec un F1-score de 41,4% sur les données d'évaluation. Cela représente une dégradation absolue de -27,4% par rapport au cas où tous les microphones sont actifs. Ce résultat était attendu, car ce système n'est pas entraîné à être robuste aux variations du nombre de capteurs disponibles. Le masquage des canaux TFCT* permet une amélioration significative avec un F1-score de 68,2% en évaluation. L'ajout d'une fonction de perte d'invariance limite également la dégradation. La fonction de perte $\mathcal{L}_{inv}^{cos}$ permet d'obtenir un F1-score de 64,5% sur les données d'évaluation. Cela représente un gain absolu de +23,1% par rapport au modèle original, mais une dégradation de -3,7% par rapport à TFCT*. La fonction $\mathcal{L}_{inv}^{F}$

TABLE 7.1 – Performances des modèles SACC sur la tâche d'OSD avec les différentes fonctions de perte d'invariance dans le cas où tous les microphones (M = 8) sont activés. Les valeurs en gras indiquent les meilleures performances pour chaque type de caractéristiques acoustiques.

AMI Antenne 1	**Précision** (%) ↑		**Rappel** (%) ↑		**F1-score** (%) ↑	
M = 8	Dev	Eval	Dev	Eval	Dev	Eval
TFCT	72,5	72,8	72,3	65,2	72,4	68,8
TFCT*	74,4	78,3	71,0	62,4	72,7	**69,5**
TFCT + $\mathcal{L}_{inv}^{F}$	75.0	77.1	71.6	63.2	**73.3**	**69.5**
TFCT + $\mathcal{L}_{inv}^{cos}$	71.9	72,6	73.2	66,2	72.6	**69,2**
$\mathcal{A}_{32}$	71,4	73,9	67,3	58.6	69.3	65.4
$\mathcal{A}_{32}^{*}$	67,7	69,4	72,8	65,6	**70,2**	**67,5**
$\mathcal{A}_{32}$ + $\mathcal{L}_{inv}^{F}$	70,8	73,8	69,6	61,1	**70,2**	66,9
$\mathcal{A}_{32}$ + $\mathcal{L}_{inv}^{cos}$	66,4	68,4	73,9	68,1	**69,9**	**68,3**
$\mathcal{A}_{128}$	66.4	73.9	69.0	49.8	67.7	59.4
$\mathcal{A}_{128}^{*}$	58,9	60,4	61,7	55,6	60,3	57,9
$\mathcal{A}_{128}$ + $\mathcal{L}_{inv}^{F}$	70,9	73,9	69,4	60,1	**70,2**	**66,8**
$\mathcal{A}_{128}$ + $\mathcal{L}_{inv}^{cos}$	71,3	73,3	68,3	60,3	**69,8**	**66,2**

permet de meilleures performances avec un F1-score de 70,1% et de 67,8% sur les données d'évaluation et de développement. Ce modèle obtient les meilleures performances d'OSD dans le cas M = 4 sur les données de développement. Ce modèle généralise cependant moins bien que TFCT*.

TABLE 7.2 – Performances des modèles SACC sur la tâche d'OSD avec les différentes fonctions de perte d'invariance dans le cas où M = 4 sont activés. Les valeurs en gras indiquent les meilleures performances pour chaque type de caractéristiques acoustiques.

AMI Antenne 1	**Précision** (%) ↑		**Rappel** (%) ↑		**F1-score** (%) ↑	
M = 4	Dev	Eval	Dev	Eval	Dev	Eval
TFCT	41.2	36.5	49.6	47.7	45.0	41.4
TFCT*	69,6	75,6	68,3	62,2	68,9	68,2
TFCT + $\mathcal{L}_{inv}^{F}$	70.1	72.0	68.9	64.0	**70.1**	**67.8**
TFCT + $\mathcal{L}_{inv}^{cos}$	64.6	63,9	62.3	65,1	63.4	64,5
$\mathcal{A}_{32}$	75,5	77,2	62,6	53,3	68,5	63,1
$\mathcal{A}_{32}^{*}$	67,1	68,9	73,2	66,2	**70,0**	**67,5**
$\mathcal{A}_{32}$ + $\mathcal{L}_{inv}^{F}$	70,9	73,9	69,4	60,1	**70,2**	66,8
$\mathcal{A}_{32}$ + $\mathcal{L}_{inv}^{cos}$	66,5	68,4	73,7	67,6	**69,9**	**68,0**
$\mathcal{A}_{128}$	30.9	29.8	75.8	75.9	43.8	42.8
$\mathcal{A}_{128}^{*}$	58,5	60,0	62,1	55,3	60,3	57,6
$\mathcal{A}_{128}$ + $\mathcal{L}_{inv}^{F}$	71.4	74.0	68.7	58.0	**70.0**	65.1
$\mathcal{A}_{128}$ + $\mathcal{L}_{inv}^{cos}$	71,0	73,2	68,3	60,8	**69,6**	**66,4**

Le modèle $\mathcal{A}_{32}$ obtient un F1-score de 68,5% et de 63,1% respectivement sur les données de développement et d'évaluation. Cela représente une dégradation absolue de -2,3% à l'évaluation du modèle dans le cas $M = 8$. Le modèle présente également des difficultés à généraliser entre les données de développement et d'évaluation. Le masquage des canaux $\mathcal{A}_{32}^{*}$ améliore nettement des performances avec un F1-score de 67,5% sur l'Eval. L'ajout d'une fonction d'invariance permet une légère amélioration des performances. Dans le cas $\mathcal{L}_{inv}^{F}$, un F1-score de 70,2% et 66,8% est obtenu sur les données de développement et d'évaluation respectivement. La fonction de perte $\mathcal{L}_{inv}^{cos}$ obtient le meilleur score sur les données d'évaluation (68,0%). Les fonctions d'invariance offrent cependant une amélioration mitigée par rapport au simple masquage des canaux.

Comme dans le cas TFCT, le modèle $\mathcal{A}_{128}$ présente une forte dégradation lorsque quatre canaux sont désactivés avec un F1-score de 42,8% sur les données d'évaluation, soit une baisse absolue de -16,6% par rapport au cas $M = 8$. $\mathcal{A}_{128}^{*}$ renforce le modèle avec un F1-score de 57,6% en évaluation. L'ajout de contraintes d'invariance sur les caractéristiques permet de conserver les performances de classification obtenues dans le cas $M = 8$. Les fonctions de perte $\mathcal{L}_{inv}^{F}$ et $\mathcal{L}_{inv}^{cos}$ permettent d'obtenir un F1-score de 65,1% et 66,4% respectivement sur les données d'évaluation.

L'ajout de fonctions d'invariance renforce les systèmes d'OSD dans le cas où seuls quatre microphones sont actifs. Le simple masquage des canaux est également efficace mais présente des limites sur le modèle $\mathcal{A}_{128}$.

Activation de deux microphones

La table 7.3 présente les résultats obtenus sur la tâche d'OSD lorsque $M = 2$ microphones sont activés. Dans le cas des modèles TFCT, les performances du modèle original sont fortement dégradées avec des F1-scores respectifs de 27,7% et 33,2% sur les jeux de développement et d'évaluation. Le modèle TFCT* offre des performances remarquables avec 70,2% en évaluation. Ce modèle présente cependant un comportement peu commun, avec un score plus faible (65,4%) sur les données de développement. Les fonctions d'invariance améliorent également la robustesse du système. Le modèle $\mathcal{L}_{inv}^{cos}$ atteint des F1-scores respectifs de 60,1% et 56,7% sur les jeux de développement et d'évaluation. Une dégradation absolue de -12,5% est observée sur l'ensemble d'évaluation par rapport au cas $M = 8$. La fonction $\mathcal{L}_{inv}^{F}$ permet d'obtenir de meilleures performances avec un F1-score de 70,6% sur le Dev et de 67,6% sur l'Eval. Une dégradation de seulement -1,9% est observée en évaluation par rapport au cas $M = 8$.

Dans le cas $M = 2$, le modèle $\mathcal{A}_{32}$ présente une faible dégradation par rapport au cas $M = 8$ avec un F1-score de 63,1% sur les données d'évaluation. Le système $\mathcal{A}_{32}^{*}$ conserve les mêmes performances que dans le cas $M = 8$ (67,2%). Les fonctions d'invariance améliorent aussi les performances d'OSD avec un F1-score de 66,9% pour $\mathcal{L}_{inv}^{F}$ et 68,0% pour $\mathcal{L}_{inv}^{cos}$ sur les données d'évaluation.

Le modèle $\mathcal{A}_{128}$ n'est pas robuste au cas $M = 2$ au point qu'il est difficile de calculer les scores, car le modèle ne détecte plus la parole superposée. La prédiction du système pour cette

classe est nulle. Le masquage $\mathcal{A}^{*}_{128}$ améliore la robustesse avec 57,6% en évaluation. L'ajout de la fonction de perte $\mathcal{L}^{F}_{inv}$ permet de conserver les performances de classification obtenues dans le cas $M = 8$ (65,4%). La fonction $\mathcal{L}^{cos}_{inv}$ permet d'atteindre les meilleures performances avec 66,0%. Cette dernière permet également une meilleure généralisation entre les jeux de développement et d'évaluation.

TABLE 7.3 – Performances des modèles SACC sur la tâche d'OSD avec les différentes fonctions de perte d'invariance dans le cas où $M = 2$ sont activés. Les valeurs en gras indiquent les meilleures performances pour chaque type de caractéristiques acoustiques.

AMI Antenne 1	**Précision** (%) ↑		**Rappel** (%) ↑		**F1-score** (%) ↑	
$M = 2$	Dev	Eval	Dev	Eval	Dev	Eval
TFCT	46,0	47,7	19,9	25,5	27,7	33,2
TFCT*	71,0	70,7	60,5	69,7	65,4	**70,2**
TFCT + $\mathcal{L}^{F}_{inv}$	71.1	69.1	70.1	66.0	**70.6**	67.6
TFCT + $\mathcal{L}^{cos}_{inv}$	60.9	62,8	59.2	51,7	60.1	56,7
$\mathcal{A}_{32}$	75,2	77,0	62,1	53,4	68,0	63,1
$\mathcal{A}^{*}_{32}$	66,8	68,0	73,1	66,5	**69,8**	**67,2**
$\mathcal{A}_{32}$ + $\mathcal{L}^{F}_{inv}$	70,8	73,7	68,9	61,2	**69,8**	66,9
$\mathcal{A}_{32}$ + $\mathcal{L}^{cos}_{inv}$	66,2	68,3	73,0	67,6	**69,5**	**68,0**
$\mathcal{A}_{128}$	0.12	-	100.0	-	0.18	-
$\mathcal{A}^{*}_{128}$	58,3	59,6	61,9	55,7	60,0	57,6
$\mathcal{A}_{128}$ + $\mathcal{L}^{F}_{inv}$	71.1	73.4	68.5	59.1	**69.7**	**65.4**
$\mathcal{A}_{128}$ + $\mathcal{L}^{cos}_{inv}$	70,0	71,5	68,6	61,3	**69,2**	**66,0**

Afin d'illustrer la robustesse des systèmes invariants, la figure 7.2 présente des exemples de prédiction du système $\mathcal{A}_{128}$ pour deux segments issus des données de développement du corpus AMI. Le modèle original est comparé au modèle invariant entraîné avec la fonction $\mathcal{L}^{F}_{inv}$. La prédiction du modèle est présentée dans les cas $M = 8$ et $M = 2$. Cette figure montre que les prédictions obtenues par le modèle original sont très différentes entre les cas $M = 8$ et $M = 2$. Le modèle ne permet pas de détecter la parole superposée de façon suffisamment robuste. L'apprentissage invariant permet de régulariser le modèle en obtenant des prédictions similaires entre les cas $M = 8$ et $M = 2$. Sur l'exemple de gauche, l'apprentissage invariant permet également une meilleure détection des segments de parole superposée quand tous les microphones sont actifs.

Discussions

Cette section montre que l'apprentissage de caractéristiques invariantes au nombre de microphones actifs permet une meilleure robustesse des modèles d'OSD si des microphones sont désactivés. Les deux fonctions de perte étudiées, basées respectivement sur la norme de Frobenius et la similarité cosine, renforcent les performances dans ces conditions. Les expériences menées

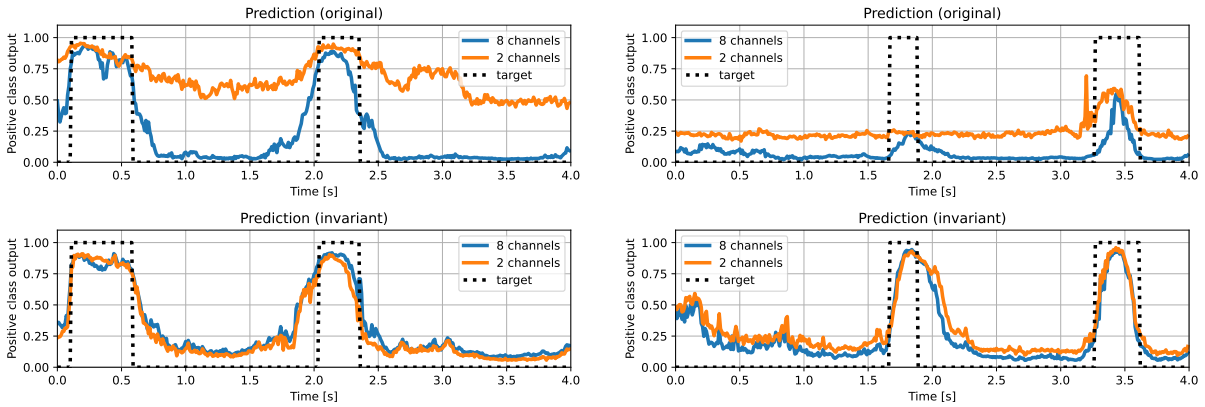


FIGURE 7.2 – Deux exemples de prédiction du modèle SACC+$\mathcal{A}_{128}$ sans (haut) et avec (bas) l'apprentissage invariant.

montrent qu'un simple masquage des canaux (ex : TFCT*) est parfois suffisant pour obtenir un modèle robuste. Ce type d'approche a cependant montré des limites dans le cas du modèle $\mathcal{A}_{128}$. La fonction $\mathcal{L}_{inv}^{cos}$ semble également renforcer significativement la généralisation du modèle entre les jeux de données de développement et d'évaluation.

Une seconde évaluation est réalisée afin d'évaluer la robustesse des systèmes dans le cas d'une antenne de microphones non conforme à celle de l'apprentissage. Cette étude est présentée dans la sous-section suivante.

7.1.4 Évaluation avec une antenne non conforme

Les modèles invariants au nombre de microphones disponibles ont été évalués sur des données acquises à l'aide du même dispositif que celui des données d'apprentissage. Cette section propose d'évaluer ces mêmes modèles sur les données issues de la seconde antenne du corpus AMI. Les résultats sont présentés dans la sous-section suivante.

Résultats

La table 7.4 présente les résultats obtenus sur les données issues de l'antenne 2 pour la tâche d'OSD. Le modèle TFCT atteint des performances d'OSD mitigées avec un F1-score de 49,0% en développement et 43,8% en évaluation. Cela correspond à une dégradation absolue de -25,0% par rapport à l'antenne 1 sur ce dernier jeu de données (table 7.1). Ce modèle présente donc une forte sensibilité à la configuration de l'antenne.

Le masquage des canaux TFCT* améliore nettement la robustesse du modèle avec un F1-score de 66,7% sur l'ensemble de développement et 65,1% en évaluation. Il obtient les meilleurs scores sur ce dernier jeu de données. Le modèle généralise cependant moins bien que lors des expériences menées sur l'antenne 1 (*cf.* table 7.3).

L'ajout de la fonction de perte $\mathcal{L}_{inv}^{cos}$ permet d'améliorer la détection de parole superposée avec un gain absolu de +14,8% sur les données d'évaluation (F1-score Dev : 57,7%, Eval : 58,6%). Cette fonction de perte permet également d'améliorer la généralisation du modèle en offrant des performances similaires sur les deux jeux de données. L'apprentissage d'une représentation invariante au nombre de canaux permet d'améliorer la robustesse du système en cas d'antenne non conforme. La fonction de perte $\mathcal{L}_{inv}^{F}$ permet d'obtenir les meilleures performances d'OSD sur les données de développement avec un F1-score de 68,8%. Les performances sont légèrement inférieures en évaluation avec 64,2%. Ce modèle présente une dégradation absolue de seulement -5,3% sur les données d'évaluation par rapport au cas $M = 8$ sur les données issues de l'antenne 1. La fonction $\mathcal{L}_{inv}^{F}$ et le masquage (TFCT*) permettent donc une meilleure généralisation du système SACC+TFCT pour la tâche d'OSD.

Table 7.4 – Performances sur la tâche d'OSD pour chaque modèle et chaque méthode d'apprentissage invariant sur les signaux captés par l'antenne 2 du corpus AMI, composée de $M = 4$ microphones. Les valeurs en gras indiquent les meilleures performances pour chaque type de caractéristiques acoustiques.

AMI Antenne 2	**Précision** (%)↑		**Rappel** (%)↑		**F1-score** (%) ↑	
$M = 4$	Dev	Eval	Dev	Eval	Dev	Eval
TFCT	39,0	34,3	65,9	60,4	49,0	43,8
TFCT*	66,6	63,5	66,8	66,8	66,7 (+17,7)	**65,1** (+21,3)
TFCT + $\mathcal{L}_{inv}^{F}$	71,4	72,1	66,3	58,0	**68,8** (+19,8)	**64,2** (+20,4)
TFCT + $\mathcal{L}_{inv}^{cos}$	49,6	54,0	69,0	64,0	57,7 (+8,7)	58,6 (+14,8)
$\mathcal{A}_{32}$	63,1	59,1	63,2	60,6	63,2	59,8
$\mathcal{A}_{32}^{*}$	53,6	48,2	74,5	73,3	62,3 (-0.9)	58,1 (-1,7)
$\mathcal{A}_{32}$ + $\mathcal{L}_{inv}^{F}$	62.9	57.8	65,8	62,2	**64.3** (+1,1)	59.9 (+0,1)
$\mathcal{A}_{32}$ + $\mathcal{L}_{inv}^{cos}$	65.7	61,5	65,6	61,9	**65.6** (+2,4)	**61.7** (+1,9)
$\mathcal{A}_{128}$	47,6	57,9	50.3	38,4	48,9	46,2
$\mathcal{A}_{128}^{*}$	49,1	48,5	53,3	53,9	51,1 (+2,2)	51,0 (+4,8)
$\mathcal{A}_{128}$ + $\mathcal{L}_{inv}^{F}$	58,6	54,9	65,9	61,3	62,0 (+13,1)	57,9 (+11,7)
$\mathcal{A}_{128}$ + $\mathcal{L}_{inv}^{cos}$	62,9	58,4	65,0	63,9	**64,0** (+15,1)	**61,0** (+14,8)

Comme avec l'antenne 1, le modèle $\mathcal{A}_{32}$ offre des performances similaires en considérant ou non l'apprentissage invariant. Le modèle initial obtient un F1-score de 59,8% sur les données d'évaluation, soit une dégradation de -5,6% par rapport à l'antenne 1. Le masquage aléatoire des canaux $\mathcal{A}_{32}^{*}$ ne permet pas d'améliorer la détection dans ce cas de figure et dégrade le F1-score de -1,7% en évaluation. L'ajout de la fonction $\mathcal{L}_{inv}^{F}$ permet d'obtenir des performances similaires (59,9%) en phase d'évaluation. Les performances sont légèrement supérieures sur les données de développement avec un F1-score de 64,3%. Ce gain est dû à un meilleur rappel (65,8%) par rapport au modèle initial (63,2%) à précision égale. La fonction d'invariance $\mathcal{L}_{inv}^{cos}$ améliore légèrement les performances sur les deux jeux de données avec des F1-scores respectifs de 65,6%

et 61,7%. Ce modèle atteint les meilleures performances d'OSD à l'aide de filtres analytiques sur les données de l'antenne 2.

Dans le cas du modèle $\mathcal{A}_{128}$, les performances d'OSD sont fortement dégradées sur les signaux de l'antenne 2 avec un F1-score de 46,2% sur les données d'évaluation. Le masquage des canaux $\mathcal{A}^{*}_{128}$ améliore légèrement les performances avec un F1-score de 51,0% en évaluation. L'ajout de contraintes lors de l'apprentissage permet une amélioration significative des performances d'OSD. Dans le cas de la fonction $\mathcal{L}^{F}_{inv}$, un gain absolu de +11,7% est observé sur le F1-score d'évaluation (+13,1% sur le jeu de développement). La fonction $\mathcal{L}^{cos}_{inv}$ permet un gain absolu de +14,8% sur cette même métrique en phase d'évaluation (+15,1% sur le jeu de développement). L'ajout d'invariance au nombre de canaux permet une meilleure robustesse à la géométrie de l'antenne pour ce système.

Discussions

Cette sous-section présente les résultats obtenus par les modèles invariants sur des données acquises à l'aide d'une antenne non conforme. Cette antenne de $M = 4$ canaux est placée sur un côté de la table durant les réunions. La géométrie de cette dernière et sa position par rapport aux locuteurs est donc différente de l'antenne 1, utilisée pour l'apprentissage. Les résultats montrent que le simple masquage aléatoire des canaux au cours de l'apprentissage renforce le modèle TFCT. Cette approche n'est cependant pas robuste avec les modèles analytiques. Les modèles invariants $\mathcal{L}^{F}_{inv}$ et $\mathcal{L}^{cos}_{inv}$ sont nettement plus robustes en cas d'antenne non conforme sur ce type de système.

Pour les modèles invariants, une dégradation persiste par rapport à l'antenne 1 et peut être liée à la différence de distribution des locuteurs autour de l'antenne. La distribution spatiale du champ acoustique est différente des données d'apprentissage et peut causer la baisse de performance. D'autre part, l'antenne est placée plus loin des locuteurs. Le rapport signal-à-bruit est donc potentiellement plus faible.

Pour combler l'écart de performance persistant entre les données de l'antenne 1 et celles de l'antenne 2, des méthodes d'augmentation de données pourraient être explorées. Par exemple, le retard entre les différents canaux peut être modulé aléatoirement afin de simuler une géométrie d'antenne différente.

7.1.5 Conclusions

Les performances des modèles de combinaison auto-attentive des canaux dépendent fortement du nombre de microphones actifs. En cas de désactivation de certains microphones, la qualité de la détection de la parole superposée est fortement dégradée. Cette section présente une approche permettant de garantir la robustesse des performances quel que soit le nombre de microphones actifs. La méthode consiste à extraire des caractéristiques de signaux dont les canaux ont été masqués puis à minimiser leur distance par rapport à des caractéristiques de référence. Cette

référence est obtenue en activant tous les microphones. Les résultats obtenus montrent que l'ajout de cette contrainte améliore la qualité dans le cas où tous les microphones sont actifs. L'approche $\mathcal{L}_{inv}^{F}$ obtient notamment les meilleures performances avec un modèle de combinaison de canaux sur la tâche d'OSD (F1-score Dev : 73,3%, Eval : 69,5% avec SACC+TFCT). L'ajout d'une fonction de perte d'invariance semble en effet régulariser l'entraînement du modèle en limitant le sur-apprentissage.

Les fonctions d'invariance permettent également de conserver des performances similaires lorsque certains microphones sont désactivés. Elles sont particulièrement avantageuses sur les modèles basés sur les filtres analytiques pour lesquels un simple masquage aléatoire des canaux ne suffit pas. De plus, cette approche permet une meilleure généralisation dans le cas où le dispositif d'acquisition n'est pas conforme à la configuration d'apprentissage. La méthode proposée améliore donc la robustesse du système d'OSD à un changement de dispositif d'acquisition. Contraindre la représentation du signal obtenue afin qu'elle ne dépende plus du nombre de microphones permet donc de limiter la dépendance au dispositif d'acquisition.

Dans la littérature récente, des modèles pré-entraînés tels que Wavlm (S. Chen et al. 2022) ont montré des performances remarquables sur diverses tâches de traitement automatique de la parole. La section suivante propose d'utiliser ce modèle afin de représenter le signal issu d'un microphone de l'antenne. L'information spatiale contenue dans le signal multicanal n'est donc plus exploitée. L'objectif est d'évaluer la robustesse de ces caractéristiques dans le contexte de la détection de parole superposée distante.

7.2 Caractéristiques monocanal pré-entraînées

La section précédente présente une approche permettant de réduire la dépendance des modèles au nombre de microphones disponibles. Dans la littérature récente, de nombreux modèles pré-entraînés ont été proposés pour le traitement automatique de la parole (Baevski et al. 2020 ; S. Chen et al. 2022). En particulier, l'architecture Wavlm (S. Chen et al. 2022) a montré des performances à l'état de l'art sur de nombreuses tâches de traitement automatique de la parole (Feng et al. 2023).Cette section propose de revenir à l'utilisation d'un microphone seul en utilisant des caractéristiques Wavlm pour la détection de parole superposée.

7.2.1 Architecture basée sur Wavlm

La figure 7.3 présente l'architecture utilisée pour la segmentation de la parole à l'aide de caractéristiques Wavlm. Les caractéristiques Wavlm sont extraites d'un signal audio monocanal. Une représentation $\bar{\mathbf{X}} \in \mathbb{R}^{F \times \bar{T}}$ est obtenue, avec F la dimension d'une trame Wavlm et $\bar{T}$ le nombre de trames extraites du signal. La représentation $\bar{\mathbf{X}}$ ne correspond pas directement à la sortie du modèle Wavlm, mais à la moyenne de la sortie de chacune des couches du modèle. La fréquence d'échantillonnage de la séquence $\bar{\mathbf{X}}$ ne correspond pas à celle souhaitée pour les prédictions. Par exemple, pour un segment audio de $L = 32000$ échantillons, Wavlm extrait une

séquence de $\bar{T} = 99$ trames contre les $T = 200$ souhaitées afin d'obtenir une prédiction toutes les 10 millisecondes. Une couche linéaire est donc ajoutée afin d'interpoler les trames issues de Wavlm pour obtenir la fréquence des trames souhaitée. Il en résulte une séquence de caractéristiques $\mathbf{X} \in \mathbb{R}^{F \times T}$. Les caractéristiques interpolées sont utilisées en entrée d'un modèle TCN afin de prédire la présence de parole superposée dans le signal.

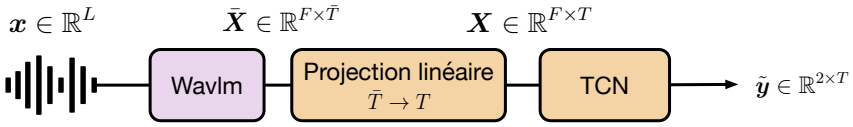


FIGURE 7.3 – Architecture de détection de parole superposée à l'aide de caractéristiques Wavlm. Le nombre de trames $\bar{T}$ extraites par Wavlm ne correspond pas au nombre de trames cibles T. Une couche linéaire permet d'interpoler les trames afin d'obtenir la dimension souhaitée.

7.2.2 Protocole d'évaluation

Le protocole expérimental est identique à celui présenté en section 4.3. Les caractéristiques SACC+TFCT et MFCC sont utilisées respectivement comme références multicanales et monocanal. Notons cependant que les caractéristiques MFCC et SACC sont difficilement comparables à Wavlm. Les premières sont obtenues à partir du domaine de Fourier et conservent un sens du point de vue du signal. Les caractéristiques Wavlm sont extraites à l'aide d'un modèle neuronal composé d'un grand nombre de paramètres (>100 millions) optimisés pour représenter le contenu phonétique et linguistique de la parole. Les résultats MFCC et SACC sont donc présentés à titre d'exemple pour illustrer les performances remarquables obtenues avec Wavlm.

Comme pour les MFCC, les caractéristiques Wavlm sont extraites du signal issu du premier canal de l'antenne. Ce modèle est nommé MDU-Wavlm. Un modèle d'OSD basé sur les caractéristiques Wavlm est également entraîné sur les données issues des microphones placés sur les casques des participants. Ce dernier est dénoté CT-Wavlm et permet de comparer les performances entre des signaux acquis en conditions distantes et en champ proche.

Les modèles sont évalués à l'aide du F1-score et de la précision moyenne (AP). Cette dernière permet d'obtenir un score indépendant du seuil de détection choisi pour l'assignation à la classe.

7.2.3 Résultats

Le tableau 7.5 présente les performances d'OSD de chaque modèle. Le modèle CT-Wavlm atteint un F1-score de 78,9% et 78,6% respectivement sur les données de développement et d'évaluation. Il s'agit des meilleures performances atteintes sur les données AMI en champ proche dans le cadre de ces travaux.

Sur les données distantes, le modèle MDU-MFCC atteint un F1-score de 67,9% et 64,5% sur chaque jeu de données. Comme présenté dans le chapitre 5, le modèle mono-microphone est surpassé par l'approche SACC+TFCT (F1-score Dev : 73,3%, Eval : 69,5%).

L'utilisation de caractéristiques Wavlm sur les signaux issus du premier canal (MDU) permet d'atteindre des performances remarquables avec un F1-score de 78,6% sur les données de développement et de 76,7% sur le jeu d'évaluation. Cela représente, sur ce jeu d'évaluation, un gain absolu de +12,2% par rapport au MDU-MFCC et de +7,2% par rapport au modèle SACC+MFCC. Les observations sont similaires sur la précision moyenne où le modèle MDU-Wavlm surpasse de +12% le modèle SACC+TFCT sur les données d'évaluation. Enfin, les performances obtenues à l'aide des caractéristiques Wavlm sur la parole distante (MDU-Wavlm) sont similaires à celles obtenues sur les signaux acquis en champ proche (CT-Wavlm). Le F1-score est dégradé de seulement -1,9% en conditions distantes sur les données d'évaluation.

TABLE 7.5 – Performance sur la tâche d'OSD avec les caractéristiques Wavlm et une modélisation de séquence TCN sur le corpus AMI. La colonne Qté audio indique le nombre total d'heures de signal utilisées pour l'apprentissage des modèles, en prenant en compte le pré-apprentissage de Wavlm (S. CHEN et al. 2022).

			F1-score (%) ↑		**AP** (%) ↑	
	# Param.	Qté audio (h)	Dev	Eval	Dev	Eval
CT-Wavlm	>100 M	64k~94k + 80	*78,9*	*78,6*	*86,9*	*84,5*
MDU-MFCC	0,26 M	80	67,9	64,5	71,8	65,1
SACC+TFCT	0,40 M	80	73,3	69,5	77,7	70,6
MDU-Wavlm	>100 M	64k~94k + 80	**78,6**	**76,7**	**85,9**	**82,6**

7.2.4 Discussions

L'extraction de caractéristiques acoustiques à l'aide du modèle pré-entraîné Wavlm permet d'atteindre des performances remarquables pour la détection de parole superposée. La différence de qualité pour la détection entre les signaux acquis en champ proche ou distant est quasiment effacée. Les modèles multi-microphones tels que SACC sont largement dépassés par ce système.

Bien qu'ils permettent des performances à l'état de l'art sur la tâche de détection de parole superposée (LEBOURDAIS et al. 2022), ces modèles sont composés d'un grand nombre de paramètres. En fonction de la taille des modèles, Wavlm contient entre 94,7 M et 316 M de paramètres optimisés (S. CHEN et al. 2022). D'autre part, ce modèle requiert une grande quantité de données pour être optimisé (*cf* table 7.5, colonne Qté. audio). Les auteurs de l'article présentant Wavlm indiquent que 94 kh d'audio peuvent être requises pour l'apprentissage du modèle le plus large. Ces modèles nécessitent donc beaucoup de ressources pour être pré-entraînés.

De plus, le grand nombre de paramètres composant ces modèles ainsi que les tâches permettant leur pré-apprentissage rendent la représentation du signal obtenue difficilement interprétable. Contrairement à un spectrogramme ou aux MFCC, il n'est pas possible de déterminer à quoi correspond chaque dimension de la représentation. À l'inverse, les représentations obtenues à l'aide des modèles SACC appartiennent au domaine temps-fréquence et peuvent être facilement

interprétées.

Les modèles pré-entraînés sont donc intéressants du point de vue de la performance. Ils permettent d'ailleurs l'obtention de résultats à l'état de l'art sur de nombreuses tâches de traitement automatique de la parole (S. Chen et al. 2022). Cependant, ces modèles requièrent de larges ressources rendant leur apprentissage coûteux. De plus, les représentations du signal qu'ils génèrent restent opaques, rendant difficile leur interprétation.

7.3 Conclusions

Ce chapitre explore deux approches permettant de réduire la dépendance des systèmes d'OSD au nombre de canaux disponibles pour le traitement de la parole distante et multicanale. La première exploite le modèle d'extraction de caractéristiques basée sur la combinaison auto-attentive des canaux SACC. Elle consiste à apprendre une représentation indépendante du nombre de microphones activés en entrée. Pour cela, les caractéristiques obtenues lorsque tous les microphones sont actifs servent de référence. Le segment audio est ensuite dupliqué en masquant aléatoirement des canaux. Les caractéristiques sont à nouveau extraites pour ces signaux masqués. Le modèle est ensuite entraîné à détecter la parole superposée tout en minimisant la distance entre les caractéristiques masquées et la référence.

Les expériences menées avec trois modèles SACC montrent que la contrainte d'invariance au nombre de microphones améliore nettement la robustesse des systèmes. Les performances des systèmes sont maintenues lorsque certains microphones sont désactivés (**Q1**). Les performances sont cependant similaires au simple masquage aléatoire des canaux. L'apprentissage invariant permet cependant un meilleur contrôle de la représentation apprise. La conception d'autres fonctions de coût pourrait permettre d'améliorer les performances, voire de rendre les caractéristiques extraites plus interprétables en choisissant d'autres contraintes.

Deux mesures de similarité sont proposées afin d'apprendre une représentation invariante. La similarité cosine permet une meilleure généralisation des performances, notamment avec les modèles analytiques. La fonction de perte basée sur la norme de Frobenius apporte un gain significatif pour le modèle utilisant la TFCT. Dans les deux cas, les deux fonctions de perte permettent une meilleure robustesse au nombre de capteurs actifs (**Q2**). Les modèles invariants sont également évalués sur des données acquises à l'aide d'une antenne non conforme aux données d'apprentissage. L'invariance au nombre de capteurs améliore significativement les performances par rapport au modèle original. Le gain par rapport au masquage aléatoire des canaux reste cependant minime. (**Q3**).

La seconde méthode utilise les caractéristiques extraites à l'aide du modèle pré-entraîné Wavlm. Celles-ci sont obtenues uniquement sur le signal du premier microphone de l'antenne. Cette étude s'éloigne de l'utilisation de l'information spatiale pour la segmentation de la parole. Elle montre cependant que ces caractéristiques présentent une forte robustesse au bruit dans les signaux. En effet, les performances obtenues sur les données en champ proche (AP Eval

78,6%) sont très proches de celles obtenues sur les signaux distants (AP Eval 76,7%). De plus, les performances obtenues surpassent largement le meilleur modèle SACC sur la tâche d'OSD. Les caractéristiques Wavlm offrent donc des performances remarquables en conditions distantes (**Q4**). Ces modèles nécessitent cependant beaucoup de ressources (données, calcul...) pour leur apprentissage. Les représentations qu'ils extraient du signal sont également difficiles à interpréter, rendant les modèles opaques. Les caractéristiques Wavlm restent cependant les meilleures candidates du point de vue de la performance.

La première étude menée dans ce chapitre a donc montré que l'ajout de contraintes sur les caractéristiques extraites par combinaison de canaux permettait d'améliorer la robustesse des systèmes. La seconde montre que les représentations pré-entrainées permettent des performances remarquables et robustes au bruit. Il serait intéressant de combiner ces deux approches. Les modèles actuels de représentation du signal audio, obtenues par apprentissage auto-supervisé, ne prennent pas en compte l'information issue de plusieurs microphones. Celle-ci est pourtant nécessaire pour certaines tâches telles que la localisation de sources. L'apprentissage d'une représentation intégrant l'information spatiale permettrait une meilleure discrimination dans l'espace tout en conservant les qualités de modélisation du signal offertes par Wavlm.

Chapitre 8

Collecte de données multicanales en réunion

Les données acquises à l'aide de plusieurs microphones, et permettant l'apprentissage de réseaux de neurones à grande échelle, sont rares. D'autre part, certaines informations pouvant être utiles dans le cas multicanal sont manquantes dans les jeux de données disponibles. Ce chapitre d'ouverture présente le protocole réalisé et les outils développés pour l'acquisition d'une base de données en réunion à l'aide de plusieurs microphones. L'objectif est de fournir un jeu de données conséquent proposant une annotation pour la diarisation en locuteur avec l'information de localisation. À la date de rédaction de ce manuscrit, seuls le dispositif d'acquisition et le logiciel associé ont été développés dans le cadre d'un projet étudiant. L'acquisition de données à grande échelle n'a pas été réalisée.

8.1 Contexte et motivations

Cette section présente les jeux de données existants permettant, soit la localisation de locuteurs, soit la diarisation.

8.1.1 Travaux similaires

Les jeux de données pour le traitement automatique de la parole en réunion ont été identifiés en section 2.4. Le corpus AMI (Carletta et al. 2006) est massivement utilisé pour la diarisation (Bredin et al. 2020 ; Landini et al. 2022b). Plusieurs réunions en anglais, en partie scénarisées, ont été enregistrées en conditions réelles à l'aide de divers dispositifs (caméras, micro-casque, antennes de microphones...). Les annotations fournies contiennent la segmentation et la retranscription de chaque réunion. La position des participants n'est cependant pas annotée.

Précédemment, le corpus ICSI (Janin et al. 2003) a été proposé pour le traitement de la parole en réunion. Les signaux audio ont été enregistrés à l'aide de microphones placés sur les casques et de 6 microphones distants placés sur la table. Les annotations contiennent la retranscription des réunions. Les méta-données apportent des informations sur les locuteurs telles que l'âge, le genre et le numéro de siège. Cependant, les plans de salle ne sont pas disponibles, rendant difficile l'association du numéro de siège et de la position. De plus, seuls les signaux issus

des micro-casques sont disponibles sur le serveur de téléchargement [1].

Récemment, le corpus AISHELL-4 (Fu et al. 2021) a été proposé pour le traitement automatique de la parole en réunion. Il contient environ 100 h de parole en mandarin enregistrée au cours de différentes réunions. La parole est spontanée, seul le sujet des discussions était fixé au début des sessions. Une antenne de 8 microphones a été utilisée pour acquérir les signaux de parole. Les annotations contiennent la segmentation et l'identité des locuteurs ainsi que la transcription de chaque réunion. La position des locuteurs n'est cependant pas annotée.

Quelques corpus ont été proposés pour la localisation du locuteur dans des domaines autres que les réunions. Le corpus LOCATA (Löllmann et al. 2018) a été présenté dans le cadre du challenge éponyme. Plusieurs tâches étaient proposées pour ce challenge telles que la localisation et le suivi d'un ou plusieurs locuteurs. Les données contiennent des signaux de parole enregistrés dans divers environnements à l'aide de plusieurs dispositifs tels qu'une antenne sphérique, une tête de robot et une antenne spécifiquement conçue pour l'acquisition de données. Ce corpus contient une faible quantité de données (~10 h), insuffisante pour l'apprentissage de réseaux de neurones profonds, pour la segmentation ou la diarisation. De plus, les annotations d'activité des locuteurs ne sont pas disponibles.

Des bases de données multimodales ont également été développées pour la localisation audiovisuelle des locuteurs. Par exemple, AV16.3 (Lathoud et al. 2004) se focalise sur le suivi de locuteur à l'aide du signal audio et de la vidéo. Les données du corpus RAVEL (Alameda-Pineda et al. 2013) ont également été acquises pour la localisation et le suivi de locuteur à l'aide d'un robot. W. He et al. (2018) ont développé une base de données pour la localisation simultanée de plusieurs locuteurs. Les auteurs ont enregistré 16h de données du corpus AMI diffusées par des haut-parleurs dans une salle. Des enregistrements réels de 220 secondes sont également proposés pour l'évaluation. La durée de signal disponible n'est cependant pas suffisante pour l'apprentissage d'un système de diarisation du locuteur ou de localisation.

8.1.2 Acquisition envisagée

Parmi les corpus multicanaux présentés, aucun ne propose une quantité suffisante de données enregistrées en conditions de réunion réelles tout en fournissant les annotations de diarisation et de localisation des locuteurs. Ce chapitre propose un protocole pour le développement et l'acquisition d'une telle base de données. Les données seront acquises au cours de réunions à l'aide d'une antenne de microphones. La position des locuteurs ainsi que leur activité seront annotées afin de permettre l'entraînement et l'évaluation de systèmes sur les tâches suivantes :

- Segmentation de la parole (VAD, OSD et SCD),
- Segmentation et regroupement en locuteurs (diarisation),

1. https://groups.inf.ed.ac.uk/ami/icsi/download/ consulté le 18 juillet 2023

- Localisation et suivi des locuteurs dans l'espace,
- Reconnaissance du locuteur.

Les annotations de la position du locuteur permettent également de favoriser les travaux de recherche en prenant en compte la position des participants pour les tâches de traitement de la parole listées ci-dessus.

8.2 Protocole d'acquisition

8.2.1 Choix des salles et caractéristiques des données

Ce projet a pour objectif d'acquérir au minimum 60 h de parole pour permettre l'entraînement et l'évaluation de réseaux de neurones profonds pour la localisation du locuteur et la diarisation. Environ 45 h sont réservées à l'apprentissage, 5 h au développement et 10 h à l'évaluation.

Les données seront acquises au cours de réunions au sein de différentes composantes de l'Université du Mans. Cela permettra d'augmenter la variabilité en locuteurs, l'environnement acoustique et les types de réunions enregistrées. Il est envisagé de faire intervenir au moins 50 locuteurs en garantissant au maximum la parité des genres. Les locuteurs doivent être différents entre les données d'apprentissage, de développement et d'évaluation. Cela permet d'évaluer les modèles de diarisation en conditions ouvertes, où les locuteurs observés en phase d'inférence n'ont pas été enrôlés.

Les réunions seront acquises dans différentes salles afin de faire varier l'environnement acoustique de l'acquisition. La table 8.1 présente des exemples de salles envisagées pour réaliser l'acquisition des réunions. Certaines salles sont dotées de vidéoprojecteurs ou de climatiseurs permettant l'ajout de bruit de fond réaliste. L'acquisition de données dans un environnement ouvert et bruyant telle qu'une cafétéria est également envisagée. Certaines salles sont ouvertes, permettant à des personnes extérieures d'arriver au cours de la réunion. Le nombre et l'identité des locuteurs sont donc variables au cours de la séance.

Salle	Dimensions	Vidéoproj.	Climatisation	Ouverte	Autre
Conseils LIUM	grande	✓			Loin de l'antenne
Réunion ENSIM	petite	✓	✓		-
Master LIUM	grande	✓		✓	-
Cafétéria ENSIM	grande			✓	Bruit de fond élevé

TABLE 8.1 – Exemples de salles envisagées pour l'acquisition de réunions.

Il est prévu que les participants puissent quitter la réunion ou y entrer au cours des séances d'enregistrement. De plus, les participants devront utiliser leurs ordinateurs afin de générer des bruits de fond tels que le son des touches du clavier. Chaque réunion fera intervenir entre 4

et 8 participants. Lorsque deux réunions font intervenir des locuteurs en commun, il leur sera demandé de ne pas se placer dans la même position par rapport au système de captation.

Plusieurs types de réunions seront enregistrés afin de varier les types de signaux de parole (ex : spontanée, scriptée, etc.). Les données seront donc enregistrées au cours de discussions scientifiques (ex : réunion de projet), de présentations techniques (ex : avancement de thèse, projets étudiants...) et de réunions administratives (ex : gestion de projet, organisation). Cela permettra également de faire varier la proportion de parole superposée entre les fichiers.

Les réunions seront enregistrées par une antenne de microphones placée au centre de la table. Si deux réunions différentes interviennent dans la même salle, la position de l'antenne sera modifiée. Cela permettra une acquisition plus réaliste, en considérant que le dispositif est remis en place pour chaque réunion. L'orientation de l'antenne sera également variable, le dispositif d'acquisition permettant de connaître cette orientation par rapport à une position de référence. Des marqueurs peuvent également être placés sur les murs afin de connaître l'orientation de l'antenne dans l'espace à l'aide de la caméra.

8.2.2 Matériel

L'acquisition audio est réalisée à l'aide d'une antenne de microphones. Afin d'affiner l'annotation des positions des locuteurs, une caméra 360° permet une acquisition vidéo de la scène complète.

Antenne de microphones

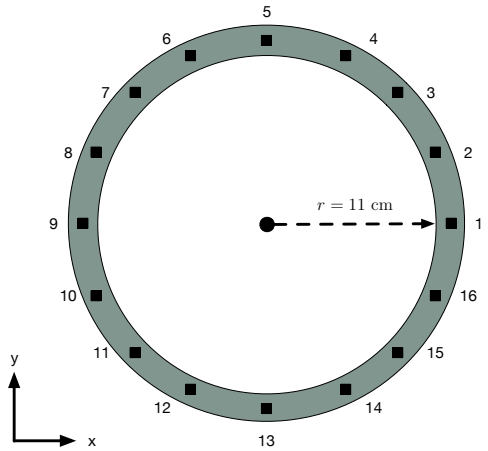


FIGURE 8.1 – Schéma de l'antenne de microphones utilisée pour la collecte de données.

Les données audio seront acquises à l'aide d'une antenne circulaire uniforme de r = 11 cm de diamètre composée de 16 microphones MEMS numériques. La géométrie de l'antenne est détaillée en figure 8.1. Les 16 canaux audio sont acquis à l'aide d'une interface MiniDSP MCHStreamer [2].

2. https://www.minidsp.com/products/usb-audio-interface/mchstreamer consulté le 13 mars 2023

Acquisition vidéo

Afin d'affiner l'annotation de la position des locuteurs après acquisition des données, un enregistrement vidéo 360° est envisagé. Les données vidéo seront acquises simultanément avec les signaux audio à l'aide d'une caméra Kodak 360 PixPro [3]. Le repère de la caméra sera équivalent à celui de l'antenne afin de garantir l'alignement des positions de la vidéo et de l'antenne.

8.2.3 Annotations

Annotation de la position des locuteurs

Dans le contexte des réunions, la position des locuteurs ne nécessite pas d'être connue au degré près. En pratique, l'estimation de la position du locuteur actif peut permettre de diriger un filtre spatial dans sa direction. Les filtres spatiaux n'étant pas idéaux, la direction de focalisation correspond à une zone spatiale et non à un point précis de l'espace. La position du locuteur estimée ne nécessite donc pas d'être exacte.

Les annotations de la position des locuteurs sont donc réalisées par cadrans angulaires. Les cadrans sont indiqués sur la table et la position de chaque participant est annotée au début de la réunion. La position des cadrans sera fixe pour une réunion donnée et définira le repère dans lequel se trouvent les sources et l'antenne de microphones. L'acquisition vidéo permet de corriger les annotations initiales dans le cas où un participant se déplace au cours de la réunion. Les cadrans sont visibles dans la vidéo afin de permettre ces corrections. La position de l'antenne par rapport aux cadrans sera différente en fonction des sessions.

La pré-annotation des positions sera fournie et la correction de ces dernières peut être réalisée par un organisme externe.

Annotation en locuteur

L'annotation de l'activité des locuteurs au cours des réunions sera réalisée par un organisme externe. Ces annotations contiendront les instants de début et de fin de chaque segment de parole et l'identifiant du locuteur associé. Afin de faciliter le processus d'annotation, les segments de parole, les changements de locuteur et les segments de parole superposée seront extraits à l'aide des algorithmes préalablement développés (Chapitres 5 et 6). La tâche d'annotation consistera à corriger les frontières des segments estimés et à assigner l'étiquette du locuteur associée.

8.3 Validation de l'acquisition et pré-annotation

Une réunion a été enregistrée afin de valider le processus d'acquisition développé. Les signaux acquis permettent d'évaluer la possibilité d'utiliser des algorithmes de segmentation pré-entrainés

3. https://kodakpixpro.com/cameras/360-vr/ consulté le 13 mars 2023

sur les données AMI afin de fournir une pré-annotation des données. Trois algorithmes de VAD, OSD et SCD sont considérés. Les trois systèmes sont composés d'un modèle TCN avec des caractéristiques Log-Mel comme décrit en section 6.1. Pour les tâches d'OSD et de SCD, les caractéristiques spatiales CSIPD sont utilisées pour 4 paires de microphones. La figure 8.2 présente les prédictions obtenues par ces systèmes sur un signal acquis à l'aide du dispositif développé. Elle présente également la référence annotée de l'activité des trois locuteurs (N, P, E) présents dans cet exemple.

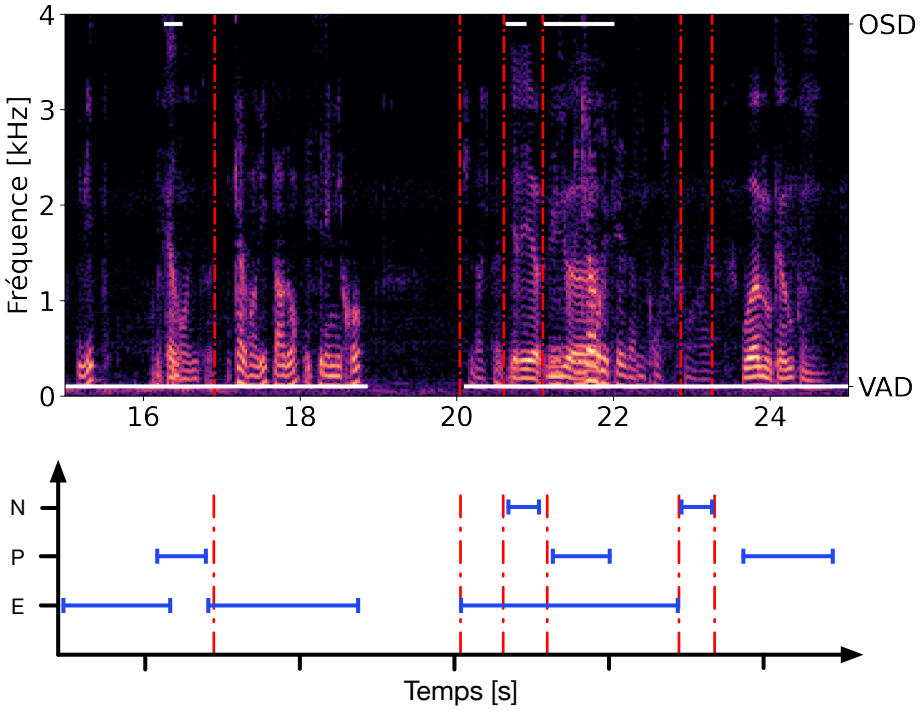


FIGURE 8.2 – Exemple de prédictions obtenues sur un exemple de signal acquis avec le dispositif lors d'une réunion de projet. La seconde partie de la figure présente l'activité des locuteurs annotée. (−·) changements de locuteur détectés.

La partie supérieure de la figure 8.2 présente le spectrogramme d'un segment issu des données acquises. Les lignes tracées sur ce dernier correspondent aux prédictions des modèles de segmentation (VAD, OSD et SCD). La partie inférieure présente la référence annotée à la main. Cette figure montre que les prédictions des systèmes sont en accord avec les activités réelles des locuteurs. Les segments de parole superposée correspondent effectivement à ceux présents dans le signal (également visibles sur le spectrogramme). La détection d'activité vocale permet également d'identifier les segments contenant de la parole. Cela peut faciliter l'annotation en éliminant les tronçons sans signal utile. Enfin, la détection des tours de parole permet de prédire certaines frontières entre les locuteurs. Certaines sont manquées comme lors du premier segment de parole superposée (autour de 16,3 secondes). Afin de faciliter les annotations de la diarisation, plusieurs prédictions de tours de parole peuvent être proposées, obtenues avec différents seuils de

détection. L'annotateur peut ensuite choisir celui qui correspond à la réalité.

8.4 Conclusions et discussions

Ce chapitre présente la première phase d'un projet de collecte de données pour la diarisation en locuteur et leur localisation. L'objectif de ce jeu de données est de fournir des données multimicrophones acquises au cours de réunions avec les annotations permettant la segmentation en locuteur et la localisation de ces derniers. Cela doit permettre de favoriser la recherche pour la diarisation multicanal et l'utilisation de l'information de position pour cette tâche, aujourd'hui peu étudiée. Un dispositif d'acquisition composé d'une antenne de 16 microphones et d'une caméra 360° permet d'acquérir les signaux de parole et l'annotation de la position des locuteurs *a posteriori*. Une pré-annotation des locuteurs est fournie au début de l'acquisition par secteurs angulaires discrets. Ce chapitre montre également que les algorithmes de segmentation préalablement développés permettent de réaliser une pré-annotation des segments de parole dans le but de faciliter la tâche d'annotation pour la diarisation.

La collecte de données à grande échelle n'a pas pu être conduite au cours de cette thèse. Il s'agit d'une tâche coûteuse en temps et demandant des ressources humaines supplémentaires afin de la mener à bien. D'autres possibilités peuvent être considérées afin d'utiliser l'information de position des locuteurs pour la diarisation. La première consiste à simuler des données spatialisées à l'aide de simulations physiques (ex : GpuRIR (Diaz-Guerra et al. 2021)). Bien que le réalisme des données obtenues soit limité, certaines considérations peuvent aider à la généralisation sur des données réelles (Srivastava et al. 2022). La seconde approche consiste à utiliser la spatialisation audio afin de créer une scène acoustique simulée réaliste. Les dômes ambisoniques (Lecomte 2016) permettent notamment de placer une source sonore à une position précise tout en simulant l'environnement acoustique (ex : réverbération). Une antenne de microphones peut ensuite être placée au centre du dispositif afin d'acquérir de nouvelles données (É. Bavu et al. 2022). Cette dernière approche offre des perspectives intéressantes, à la frontière entre la simulation numérique et l'acquisition de données réelles.

Chapitre 9

Conclusion et perspectives

9.1 Conclusion

Les travaux menés au cours de cette thèse se placent dans le contexte de la diarisation en locuteurs, et portent sur la segmentation automatique des signaux de parole capturés par une antenne de microphones dans le contexte des réunions. La segmentation du signal se sépare en trois tâches distinctes :

VAD : détection d'activité vocale, consistant à détecter les segments de parole dans le signal audio,

OSD : détection de parole superposée, visant à détecter les segments dans lesquels plusieurs locuteurs sont simultanément actifs,

SCD : détection de changements de locuteur, consistant à détecter les frontières entre les différents tours de parole des locuteurs dans le signal.

Les méthodes développées consistent à extraire des caractéristiques du signal multicanal afin de prendre en compte l'information spatiale contenue dans ces derniers. Les caractéristiques sont modélisées par un réseau de neurones prenant en compte les dépendances temporelles de la séquence. Le modèle est entraîné à prédire la présence d'un des évènements listés ci-dessus sous forme d'une classification binaire (VAD et OSD) ou de régression (SCD). La détection est réalisée sur des trames extraites toutes les 10 ms sur le signal multicanal.

9.1.1 Combinaison de canaux auto-attentive pour la segmentation de la parole

Les premiers travaux portent sur l'extraction de caractéristiques à l'aide d'algorithmes de combinaison auto-attentive des canaux. Le premier chapitre de contributions propose d'appliquer le modèle SACC, proposé par Gong et al. (2021), à la VAD et l'OSD en conditions distantes. Cette approche permet d'obtenir des performances similaires à la formation de voies sur les tâches de détection d'activité vocale (SER Eval : 6,47%) et de parole superposée (F1-score Eval : 68,8%). Nous proposons ensuite plusieurs extensions de ce système.

La première remplace la TFCT par un banc de filtres analytiques optimisés avec la tâche de segmentation. L'apprentissage de filtres à partir des données n'apporte pas de gain de performance sur les deux tâches de segmentations traitées. L'initialisation de ces filtres ne permet pas non plus d'améliorer les performances. Le meilleur modèle atteint un SER de 7,02% sur le VAD et un F1-score de 65,4% sur l'OSD avec les données d'évaluation.

La deuxième extension proposée modélise toute l'information contenue dans la TFCT afin d'estimer des poids de combinaison complexes. Ils permettent de conserver l'information de la phase fournissant des informations sur le retard temporel entre microphones. Ils se rapprochent ainsi des filtres spatiaux classiques (ex : formation de voies). Deux architectures sont proposées avec deux formulations distinctes :

- implicite (IcSACC) : l'estimation des poids complexes est réalisée à l'aide d'un unique bloc d'attention modélisant le module et la phase,
- explicite (EcSACC) : le module et la phase des poids de combinaison sont modélisés séparément à l'aide de deux modèles d'attention distincts.

Le modèle implicite obtient des performances de segmentation mitigées et ne permet pas un gain significatif par rapport au MDU. Cependant, le modèle explicite permet d'obtenir des performances de détection similaires à l'architecture SACC originale. Il atteint un SER de 6,39% pour la VAD et un F1-score de 66,7% sur la tâche d'OSD sur les données d'évaluation. Les extensions complexes EcSACC et IcSACC offrent un gain limité sur les performances de segmentation. De plus, ces modèles ne réalisent pas un produit complexe linéaire entre les poids et la TFCT. Considérer un produit complexe linéaire permet cependant d'analyser les poids de combinaison comme un filtre spatial linéaire.

Une seconde extension complexe est alors proposée, LcSACC, et garantit un produit complexe entre les poids de combinaison et la TFCT. Cette extension permet des performances similaires à EcSACC sur l'OSD (F1-score : 67,9%)et les meilleures performances de VAD sur les données d'évaluation (SER : 6,26%). Le formalisme LcSACC permet de visualiser la réponse spatiale du modèle et ainsi d'interpréter les directions spatiales sélectionnées par le modèle. La comparaison des cartes d'énergie acoustique et de la réponse spatiale ne semble pas montrer de corrélation entre la position des sources et la focalisation du système. Cette observation limite les capacités d'interprétation des poids.

Une troisième extension, BFSACC, est proposée pour favoriser l'interprétation des directions sélectionnées par le système. Un modèle SACC est entraîné à sélectionner les sorties d'un banc de filtres spatiaux dirigés dans des directions fixes. Les performances de ce système sont mitigées. Cependant, l'analyse des poids de combinaison permet de réaliser une pseudo-localisation des sources et offre des axes intéressants en matière d'interprétabilité. La régularisation du banc de filtres améliore nettement les performances (VAD SER : 6,52%, OSD F1-score : 67,7%), mais limite l'interprétation.

Les modèles de segmentation SACC, LcSACC et BFSACC sont évalués sur la tâche de diarisation en locuteurs. La VAD obtenue est utilisée comme première segmentation pour le modèle VBx (Landini et al. 2022b). La parole superposée est également réassignée à l'aide d'une approche heuristique (Otterson et al. 2007). Les résultats montrent que l'assignation de la parole superposée permet des gains significatifs sur les DER et le JER. Les modèles SACC+TFCT et LcSACC offrent les meilleurs DER (respectivement 24,02% et 24,30%) en conditions distantes. BFSACC offre le meilleur JER (31,35%). Les performances de diarisation sont cependant limitées

par l'extracteur d'embeddings de locuteurs, ce dernier n'étant pas entraîné sur des signaux de parole distante.

9.1.2 Caractéristiques spatiales pour la segmentation de la parole

Le deuxième chapitre des contributions explore l'utilisation de caractéristiques spatiales extraites à partir du signal multicanal. Les caractéristiques spatiales sont combinées avec les caractéristiques acoustiques (ex : MFCC). Un modèle TCN est utilisé pour résoudre les tâches de détection d'activité vocale VAD, de parole superposée OSD et de changements de locuteur SCD. Nous proposons d'abord d'évaluer l'influence des caractéristiques spatiales de l'état de l'art (IPD et CSIPD) sur la détection de changement de locuteurs (SCD). Ces caractéristiques présentent cependant une faible robustesse en cas de désactivation de certains microphones. Un nouveau jeu de caractéristiques basées sur les harmoniques circulaires, appelées CH-DOA, est proposé. Elles estiment la direction d'arrivée des sources actives à l'aide du vecteur de pseudo-intensité acoustique, estimé à l'aide de la formation de voies modale. Ce formalisme permet de réduire la dépendance au nombre de microphones disponibles au sein de l'antenne, car seuls les ordres faibles sont nécessaires.

Les résultats obtenus montrent que les CH-DOA améliorent les performances sur la plupart des tâches de segmentation, particulièrement la SCD avec un S-score de 86,4% pour le meilleur modèle. Elles présentent également une meilleure robustesse lorsque certains microphones de l'antenne sont désactivés. Bien qu'offrant une meilleure robustesse au nombre de microphones disponibles, ces caractéristiques nécessitent l'utilisation d'une antenne circulaire pour que la robustesse soit garantie. Les performances tendent donc à se dégrader lorsque cette condition n'est pas vérifiée.

L'influence de la segmentation (VAD et OSD) obtenue par combinaison de caractéristiques acoustiques et spatiales est évaluée sur la tâche de diarisation en locuteurs. La diarisation est estimée à l'aide du modèle VBx et de l'assignation de parole superposée heuristique. Les CSIPD obtiennent le meilleur DER (23,63%) et les CH-DOA proposées le meilleur JER (31,91%).

9.1.3 Indépendance au nombre de canaux

Le troisième chapitre de contributions présente une méthode d'apprentissage contraignant un module de type SACC à extraire une représentation indépendante du nombre de microphones disponibles. Cette approche vise à rendre les modèles SACC robustes à la géométrie de l'antenne utilisée sans connaissance *a priori* de cette dernière.

La méthode consiste à extraire une représentation du signal avec tous les microphones disponibles. Le segment d'entrée est ensuite dupliqué en masquant une partie des canaux afin de le dégrader. Le même module est utilisé pour extraire une représentation à partir du segment dégradé. Au cours de l'apprentissage, le système de segmentation apprend à minimiser la distance (ou maximiser la similarité) entre les caractéristiques dégradées et les caractéristiques de références.

Cette procédure est réalisée simultanément avec la classification à l'aide de l'apprentissage multi-tâche.

Deux fonctions de perte sont proposées afin d'apprendre la représentation invariante. La première, basées sur la norme de Frobenius, offre les meilleures performances pour le modèle SACC basé sur la TFCT. La seconde, basée sur la similarité cosine, améliore les performances des modèles SACC basés sur les filtres analytiques. Elle permet également une meilleure généralisation du modèle dans la majorité des cas.

Contraindre la représentation à être indépendante du nombre de capteurs améliore les performances des modèles d'OSD par rapport aux modèles entraînés sur tous les canaux. Cette observation est valide en cas de désactivation de microphones ou d'antenne non conforme. La détection est également améliorée lorsque tous les microphones sont actifs. L'apprentissage invariant n'améliore cependant pas les performances par rapport à un simple masquage aléatoire des canaux en phase d'apprentissage. L'approche proposée permet cependant un meilleur contrôle de la représentation invariante, celle-ci étant obtenue avec une fonction de perte définie.

D'autre part, l'utilisation de caractéristiques extraites à l'aide du modèle pré-entraîné Wavlm est explorée. Celles-ci sont obtenues sur le signal extrait du premier canal. Les performances d'OSD atteintes dépassent largement les approches développées au cours de ces travaux (ex : 78% F1-score Eval). De plus, les résultats sont proches entre les données en champ proche et la parole distante. Les caractéristiques pré-entraînées requièrent cependant de larges ressources pour être entraînées. Par exemple, 64 000 h de signal ont été utilisées pour apprendre la version la plus légère de Wavlm. Ces modèles contiennent également de nombreux paramètres et les représentations qu'ils extraient sont difficiles à interpréter.

Acquisition de données de parole distante en réunion

Le dernier chapitre de contributions propose un protocole d'acquisition de données en réunion. L'étude de la bibliographie montre que les jeux de données de parole multicanale sont rares. Ces derniers ne sont pas annotés spécifiquement pour la diarisation, ce qui peut mener à des erreurs dans l'annotation de la segmentation. Enfin, nous pensons que fournir l'information de la position des locuteurs serait utile à la communauté. Elles permettraient d'accentuer les efforts de recherche en diarisation informée par la localisation et l'évaluation d'algorithmes de localisation du locuteur sur des données réelles. Les travaux menés dans cette thèse ont permis de mettre en place un protocole d'acquisition et un système d'acquisition.

9.2 Perspectives

Les travaux présentés se focalisent sur la segmentation du signal de parole en utilisant plusieurs microphones. Différentes caractéristiques et approches ont été développées dans l'objectif d'améliorer les performances sur ces tâches clefs pour la diarisation en locuteurs. Ces travaux montrent cependant certaines limitations. L'utilisation de plusieurs microphones peut trouver

des applications à plus grande échelle dans le contexte de la segmentation et du regroupement en locuteurs. Les perspectives envisagées à la suite de ces travaux sont présentées dans les sous-sections suivantes.

9.2.1 Embedding de locuteur informé par la localisation

La tâche de diarisation en locuteurs repose à la fois sur la segmentation du signal et le regroupement des segments afin de les assigner à leur locuteur. Les locuteurs sont modélisés par des vecteurs de taille fixe (embeddings) supposés séparer les locuteurs dans leur espace de représentation. Les expériences de diarisation ont montré que les performances sont limitées par la qualité des embeddings de locuteur extraits en conditions distantes. Les approches actuelles utilisent des réseaux de neurones pour extraire les vecteurs (Larcher et al. 2021 ; Q. Lin et al. 2019 ; Snyder et al. 2018). Ceux-ci utilisent uniquement l'information contenue dans le signal monocanal pour représenter les locuteurs. Dans le scénario selon lequel les signaux sont acquis à l'aide d'une antenne de microphones, l'information sur la répartition spatiale des sources est intrinsèque au signal et peut faciliter la séparation des locuteurs pour la diarisation.

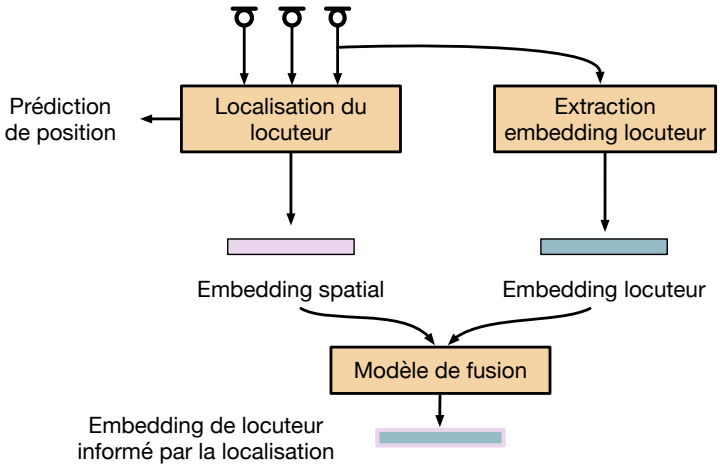


Figure 9.1 – Schéma de principe d'extraction d'embeddings de locuteur informés par la localisation.

La fusion d'informations du locuteur et spatiale a été étudiée par W. He et al. (2021). La localisation du locuteur est réalisée simultanément avec l'extraction des embeddings à l'aide d'un apprentissage multi-tâches. L'information de localisation est également utilisée au sein du réseau lors de la phase de *pooling*, permettant de compresser la dimension temporelle du signal. L'approche proposée présente des gains significatifs sur la tâche de vérification du locuteur, montrant que la localisation permet de faciliter la discrimination. L'exploitation de la localisation pour la représentation des locuteurs n'est cependant pas étudiée dans le contexte de la diarisation. De plus, d'autres schémas de fusion peuvent être explorés, comme l'attention croisée afin de visualiser l'information utilisée par le modèle, et d'identifier les situations dans lesquelles la

localisation est bénéfique. La figure 9.1 présente une vue schématique de la fusion des données de localisation avec un embedding de locuteur.

9.2.2 Apprentissage de représentation d'un signal multicanal

Une seconde perspective de ces travaux repose sur l'apprentissage d'une représentation d'un signal multicanal. D'une part, les caractéristiques spatiales ont montré leurs bénéfices dans le contexte de la segmentation de la parole (*cf.* chapitre 6). Elles présentent cependant des difficultés à généraliser à des géométries d'antennes ou des contextes d'acquisition différents. Des caractéristiques robustes existent (ex : CH-DOA) mais ne permettent pas toujours d'améliorer la segmentation. D'autre part, l'apprentissage d'une représentation invariante au nombre de capteurs a montré son intérêt pour la segmentation de la parole en permettant de conserver les performances lorsque les canaux sont désactivés ou que l'antenne change. L'apprentissage d'une représentation d'un signal multicanal robuste au nombre de capteurs semble être un axe intéressant pour obtenir le meilleur des deux mondes.

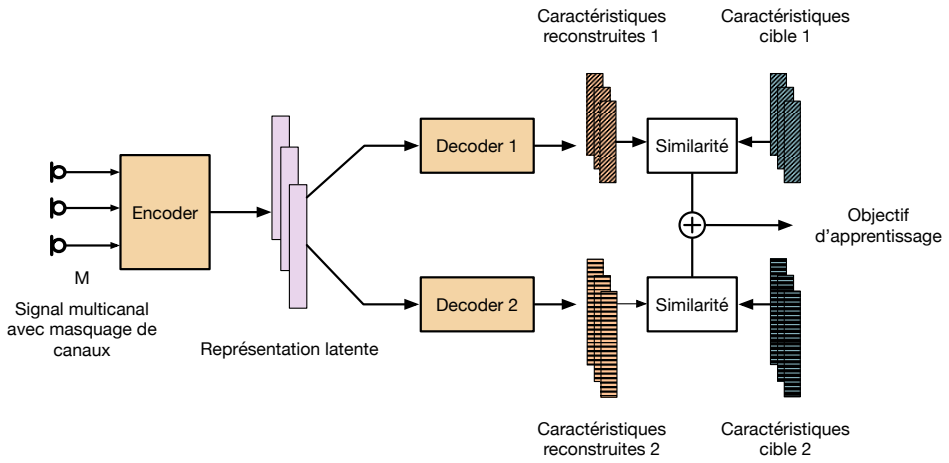


FIGURE 9.2 – Schéma de principe d'apprentissage d'une représentation spatiale à partir d'un signal multicanal.

L'apprentissage auto-supervisé est aujourd'hui largement répandu dans la littérature (JAISWAL et al. 2020) pour l'apprentissage de représentations. Une approche pour l'apprentissage de représentation spatiale à partir d'un signal multicanal est proposée en figure 9.2. Un encodeur permet de projeter le signal brut vers une séquence latente. Les canaux sont masqués aléatoirement ou supprimés pour permettre au modèle d'être robuste au nombre de canaux disponibles. La séquence latente (encodée) est ensuite utilisée pour reconstruire différents types de caractéristiques spatiales à l'aide de différents décodeurs. Par exemple, la séquence latente peut permettre de reconstruire les IPD, prédire la position des locuteurs, etc. La séquence latente doit ainsi contenir suffisamment d'information pour reconstruire chaque type de caractéristique spatiale. Elle peut ensuite être utilisée comme caractéristique spatiale pour le traitement automatique de la parole.

9.2.3 Diarisation multicanal bout-à-bout

Les modèles bout-à-bout se développent pour la diarisation du locuteur. Les approches de la littérature utilisent principalement des données monocanal. L'information spatiale contenue dans les signaux acquis à l'aide d'une antenne peut cependant être bénéfique pour cette tâche (Horiguchi et al. 2022). Par exemple, un modèle de type SACC peut être placé en entrée de ce type de système afin d'améliorer le signal en combinant les canaux. Des caractéristiques spatiales peuvent également être intégrées afin d'améliorer la détection et la discrimination des locuteurs au sein du système.

9.2.4 Segmentation multi-tâche

Dans ces travaux, les tâches de segmentation sont réalisées à l'aide de modèles indépendants. Il est envisageable de construire un modèle permettant de résoudre les trois tâches simultanément. Kunešová et al. (2023) ont montré qu'il était possible d'apprendre un système unique résolvant chaque tâche à partir de caractéristiques Wav2vec 2.0 (Baevski et al. 2020). Les auteurs proposent d'apprendre les trois tâches simultanément à l'aide de trois modèles de détection prenant les mêmes caractéristiques en entrée. Cette approche peut être étendue au domaine multicanal en remplaçant les caractéristiques Wav2vec par SACC ou en ajoutant des caractéristiques spatiales dans le processus. L'apprentissage multi-tâche peut permettre d'améliorer les performances sur chaque tâche en régularisant chaque système de détection.

BIBLIOGRAPHIE

Alameda-Pineda, X., Sanchez-Riera, J., Wienke, J., Franc, V., Čech, J., Kulkarni, K., Deleforge, A., & Horaud, R., (2013), RAVEL : An annotated corpus for training robots with audiovisual abilities, *Journal on Multimodal User Interfaces*, *7 1*, 79-91 (cf. p. 170).

Andrei, V., Cucu, H., & Burileanu, C., (2017), Detecting Overlapped Speech on Short Timeframes Using Deep Learning., *Interspeech*, 1198-1202 (cf. p. 51, 53).

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O., (2012), Speaker diarization : A review of recent research, *IEEE Transactions on audio, speech, and language processing*, *20 2*, 356-370 (cf. p. 20, 22, 56).

Anguera, X., Wooters, C., & Hernando, J., (2007), Acoustic beamforming for speaker diarization of meetings, *IEEE Transactions on Audio, Speech, and Language Processing*, *15 7*, 2011-2022 (cf. p. 74).

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M., (2020), wav2vec 2.0 : A framework for self-supervised learning of speech representations, *Advances in Neural Information Processing Systems*, *33*, 12449-12460 (cf. p. 43, 46, 163, 183).

Bahdanau, D., Cho, K., & Bengio, Y., (2014), Neural machine translation by jointly learning to align and translate, *arXiv preprint arXiv :1409.0473* (cf. p. 40-42).

Bai, S., Kolter, J. Z., & Koltun, V., (2018), An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling, *arXiv :1803.01271 [cs]* (cf. p. 40).

Bavu, E., Ramamonjy, A., Pujol, H., & Garcia, A., (2019), TimeScaleNet : A multiresolution approach for raw audio recognition using learnable biquadratic IIR filters and residual networks of depthwise-separable one-dimensional atrous convolutions, *IEEE Journal of Selected Topics in Signal Processing*, *13 2*, 220-235 (cf. p. 46).

Bavu, É., Pujol, H., Garcia, A., Langrenne, C., Hengy, S., Rassy, O., Thome, N., Karmim, Y., Schertzer, S., & Matwyschuk, A., (2022), Deeplomatics : A deep-learning based multimodal approach for aerial drone detection and localization, *QUIET DRONES Second International e-Symposium on UAV/UAS Noise* (cf. p. 175).

Benesty, J., Chen, J., & Cohen, I., (2015), *Design of circular differential microphone arrays* (T. 12), Springer, (cf. p. 62, 64, 112, 113, 116).

Benesty, J., Chen, J., & Huang, Y., (2008), *Microphone array signal processing* (T. 1), Springer Science & Business Media, (cf. p. 22, 66, 80, 109, 120).

Bengio, Y., Simard, P., & Frasconi, P., (1994), Learning long-term dependencies with gradient descent is difficult, *IEEE transactions on neural networks*, *5 2*, 157-166 (cf. p. 39).

Boakye, K., Trueba-Hornero, B., Vinyals, O., & Friedland, G., (2008), Overlapped speech detection for improved speaker diarization in multiparty meetings, *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4353-4356 (cf. p. 50, 53).

Boakye, K., Vinyals, O., & Friedland, G., (2011), Improved overlapped speech handling for speaker diarization, *Interspeech*, 941-944, https://doi.org/10.21437/Interspeech.2011-382 (cf. p. 50)

Bredin, H., (2017), Tristounet : triplet loss for speaker turn embedding, *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5430-5434 (cf. p. 54, 55).

Bredin, H., & Laurent, A., (2021), End-To-End Speaker Segmentation for Overlap-Aware Resegmentation, *Proc. Interspeech 2021*, 3111-3115, https://doi.org/10.21437/Interspeech.2021-560 (cf. p. 52, 53, 74)

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P., (2020), Pyannote.Audio : Neural Building Blocks for Speaker Diarization, *ICASSP*, 7124-7128, https://doi.org/10.1109/ICASSP40776.2020.9052974 (cf. p. 23, 47, 57, 83, 169)

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P., (2021), Geometric deep learning : Grids, groups, graphs, geodesics, and gauges, *arXiv preprint arXiv :2104.13478* (cf. p. 41).

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., (2020), Language models are few-shot learners, *Advances in neural information processing systems*, *33*, 1877-1901 (cf. p. 46).

Bullock, L., Bredin, H., & Garcia-Perera, L. P., (2020), Overlap-Aware Diarization : Resegmentation Using Neural End-to-End Overlapped Speech Detection, *ICASSP*, 7114-7118, https://doi.org/10.1109/ICASSP40776.2020.9053096 (cf. p. 46, 50-53, 55, 58, 89)

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V., Kraaij, W., Kronenthal, M., et al., (2006), The AMI meeting corpus : A pre-announcement, *Machine Learning for Multimodal Interaction : Second International Workshop, MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2*, 28-39 (cf. p. 59, 81, 82, 152, 169).

Charlet, D., Barras, C., & Lienard, J.-S., (2013), Impact of overlapping speech detection on speaker diarization for broadcast news and debates, *ICASSP*, 7707-7711, https://doi.org/10.1109/ICASSP.2013.6639163 (cf. p. 50, 53)

Chen, S., Wang, C., Chen, Z. [Zhengyang], Wu, Y., Liu, S. [Shujie], Chen, Z. [Zhuo], Li, J., Kanda, N., Yoshioka, T., Xiao, X., et al., (2022), Wavlm : Large-scale self-supervised pre-training for full stack speech processing, *IEEE Journal of Selected Topics in Signal Processing*, *16 6*, 1505-1518 (cf. p. 41, 43, 46, 129, 150, 163, 165, 166).

Chen, S. S., & Gopalakrishnan, P. S., (1998), Speaker, environment and channel change detection and clustering via the bayesian information criterion, *DARPA* (cf. p. 54, 55).

Chengalvarayan, R., (1999), Robust energy normalization using speech/nonspeech discriminator for German connected digit recognition, *Sixth European conference on speech communication and technology* (cf. p. 47, 49).

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y., (2014), Learning phrase representations using RNN encoder-decoder for statistical machine translation, *arXiv preprint arXiv :1406.1078* (cf. p. 39).

Cobos, M., Marti, A., & Lopez, J. J., (2010), A modified SRP-PHAT functional for robust real-time sound source localization with scalable spatial sampling, *IEEE Signal Processing Letters*, *18 1*, 71-74 (cf. p. 71).

Cornell, S., Omologo, M., Squartini, S., & Vincent, E., (2020), Detecting and Counting Overlapping Speakers in Distant Speech Scenarios, *Interspeech*, 3107-3111, https://doi.org/10.21437/Interspeech.2020-2671 (cf. p. 51, 53)

Cornell, S., Omologo, M., Squartini, S., & Vincent, E., (2022a), Overlapped Speech Detection and speaker counting using distant microphone arrays, *Computer Speech & Language*, *72*, 101306, https://doi.org/10.1016/j.csl.2021.101306 (cf. p. 22, 52, 53, 60, 70, 72, 76, 80, 81, 88, 89, 127, 133, 136)

Cornell, S., Pariente, M., Grondin, F., & Squartini, S., (2022b), Learning filterbanks for end-to-end acoustic beamforming, *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6507-6511 (cf. p. 67, 100).

Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., & Vincent, E., (2020), LibriMix : An Open-Source Dataset for Generalizable Speech Separation, (cf. p. 120).

Dawalatabad, N., Ravanelli, M., Grondin, F., Thienpondt, J., Desplanques, B., & Na, H., (2021), ECAPA-TDNN Embeddings for Speaker Diarization, *Proc. Interspeech 2021*, 3560-3564, https://doi.org/10.21437/Interspeech.2021-941 (cf. p. 57)

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P., (2010), Front-end factor analysis for speaker verification, *IEEE Transactions on Audio, Speech, and Language Processing*, *19 4*, 788-798 (cf. p. 56, 75).

Demšar, J., (2006), Statistical comparisons of classifiers over multiple data sets, *The Journal of Machine learning research*, *7*, 1-30 (cf. p. 84, 88).

Deng, S., Bao, C., & Cheng, R., (2020), GEV Beamforming with BAN Integrating LPS Estimation and Post-filtering, *2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 1-5, https://doi.org/10.1109/ICSPCC50002.2020.9259463 (cf. p. 66, 67)

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K., (2018), Bert : Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv :1810.04805* (cf. p. 41, 46).

Diaz-Guerra, D., Miguel, A., & Beltran, J. R., (2020), Robust sound source tracking using SRP-PHAT and 3D convolutional neural networks, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 300-311 (cf. p. 71).

Diaz-Guerra, D., Miguel, A., & Beltran, J. R., (2021), gpuRIR : A python library for room impulse response simulation with GPU acceleration, *Multimedia Tools and Applications*, *80*, 5653-5671 (cf. p. 120, 175).

Doddington, G. R., Przybocki, M. A., Martin, A. F., & Reynolds, D. A., (2000), The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective, *Speech Communication*, *31 2*, 225-254, https://doi.org/https://doi.org/10.1016/S0167-6393(99)00080-1 (cf. p. 49)

Evans, N. W., Fredouille, C., & Bonastre, J.-F., (2009), Speaker diarization using unsupervised discriminant analysis of inter-channel delay features, *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, 4061-4064 (cf. p. 71, 74).

Feng, T.-h., Dong, A., Yeh, C.-F., Yang, S.-w., Lin, T.-Q., Shi, J., Chang, K.-W., Huang, Z., Wu, H., Chang, X., et al., (2023), Superb@ slt 2022 : Challenge on generalization and efficiency of self-supervised speech representation learning, *2022 IEEE Spoken Language Technology Workshop (SLT)*, 1096-1103 (cf. p. 20, 150, 163).

Fiscus, J. G., Ajot, J., Michel, M., & Garofolo, J. S., (2006), The rich transcription 2006 spring meeting recognition evaluation, *Machine Learning for Multimodal Interaction : Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers 3*, 309-322 (cf. p. 57).

Flanagan, J., (1980), Parametric coding of speech spectra, *The Journal of the Acoustical Society of America*, *68 2*, 412-419 (cf. p. 45).

Fu, Y., Cheng, L., Lv, S., Jv, Y., Kong, Y., Chen, Z., Hu, Y., Xie, L., Wu, J., Bu, H., Xu, X., Du, J., & Chen, J., (2021), AISHELL-4 : An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario, *Proc. Interspeech 2021*, 3665-3669, https://doi.org/10.21437/Interspeech.2021-1397 (cf. p. 59, 81, 152, 170)

Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., & Watanabe, S., (2019), End-to-end neural speaker diarization with self-attention, *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 296-303 (cf. p. 56, 75).

Garcia-Perera, L. G., Villalba, J., Bredin, H., Du, J., Castan, D., Cristia, A., Bullock, L., Guo, L., Okabe, K., Nidadavolu, P. S., et al., (2020), Speaker detection in the wild : Lessons learned from JSALT 2019, *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 415-422 (cf. p. 47, 50, 58, 75).

Geiger, J. T., Eyben, F., Schuller, B., & Rigoll, G., (2013), Detecting overlapping speech with long short-term memory recurrent neural networks, *Interspeech* (cf. p. 51, 53).

Gelly, G., & Gauvain, J.-L., (2017), Optimization of RNN-based speech activity detection, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26 3*, 646-656 (cf. p. 48, 49).

Ghaemmaghami, H., Baker, B., Vogt, R., & Sridharan, S., (2010), Noise robust voice activity detection using features extracted from the time-domain autocorrelation function, *11th Annual Conference of the ISCA*, 3118-3121 (cf. p. 47).

Gong, R., Quillen, C., Sharma, D., Goderre, A., Laińez, J., & Milanović, L., (2021), Self-attention channel combinator frontend for end-to-end multichannel far-field speech recognition, *arXiv preprint arXiv :2109.04783* (cf. p. 68, 80, 86, 87, 91, 92, 94, 117, 129, 177).

Goodfellow, I., Bengio, Y., & Courville, A., (2016), *Deep learning*, MIT press, (cf. p. 34, 36, 38, 39).

Gray, R. M., (2010), A History of Realtime Digital Speech on Packet Networks : Part II of Linear Predictive Coding and the Internet Protocol, *Foundations and Trends® in Signal Processing*, *3 4*, 203-303, https://doi.org/10.1561/2000000036 (cf. p. 19)

Grumiaux, P.-A., Kitić, S., Girin, L., & Guérin, A., (2021), High-resolution speaker counting in reverberant rooms using CRNN with ambisonics features, *2020 28th European Signal Processing Conference (EUSIPCO)*, 71-75 (cf. p. 22, 133).

Gu, R., Zhang, S.-X., Chen, L., Xu, Y., Yu, M., Su, D., Zou, Y., & Yu, D., (2020a), Enhancing end-to-end multi-channel speech separation via spatial feature learning, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7319-7323 (cf. p. 72, 81).

Gu, R., Zhang, S.-X., Xu, Y., Chen, L., Zou, Y., & Yu, D., (2020b), Multi-modal multi-channel target speech separation, *IEEE Journal of Selected Topics in Signal Processing*, *14 3*, 530-541 (cf. p. 70).

He, K., Zhang, X., Ren, S., & Sun, J., (2016), Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778 (cf. p. 36, 126).

He, W., Motlicek, P., & Odobez, J.-M., (2018), Deep neural networks for multiple speaker detection and localization, *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 74-79 (cf. p. 134, 170).

He, W., Motlicek, P., & Odobez, J.-M., (2021), Multi-Task Neural Network for Robust Multiple Speaker Embedding Extraction., *Interspeech*, 506-510 (cf. p. 181).

Heymann, J., Drude, L., & Haeb-Umbach, R., (2016), Neural network based spectral mask estimation for acoustic beamforming, *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 196-200 (cf. p. 67).

Hochreiter, S., & Schmidhuber, J., (1997), Long short-term memory, *Neural computation*, *9 8*, 1735-1780 (cf. p. 38).

Horiguchi, S., Takashima, Y., Watanabe, S., & Garcia, P., (2022), Mutual Learning of Single- and Multi-Channel End-to-End Neural Diarization, *arXiv preprint arXiv :2210.03459* (cf. p. 75, 115, 183).

Hornik, K., Stinchcombe, M., & White, H., (1989), Multilayer feedforward networks are universal approximators, *Neural networks*, *2 5*, 359-366 (cf. p. 32, 34).

Hrúz, M., & Zajić, Z., (2017), Convolutional neural network for speaker change detection in telephone speaker diarization system, *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4945-4949 (cf. p. 55, 134).

Hu, M., Sharma, D., Doclo, S., Brookes, M., & Naylor, P. A., (2015), Speaker change detection and speaker diarization using spatial information, *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5743-5747 (cf. p. 71, 75).

Huang, G., Benesty, J., & Chen, J., (2016), Subspace superdirective beamforming with uniform circular microphone arrays, *2016 IEEE International Workshop on Acoustic Signal Enhancement (IWAENC)*, 1-5 (cf. p. 124).

Hughes, T., & Mierle, K., (2013), Recurrent neural networks for voice activity detection, *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7378-7382 (cf. p. 48, 49).

Huijbregts, M., van Leeuwen, D. A., & Wooters, C., (2011), Speaker diarization error analysis using oracle components, *IEEE Transactions on Audio, Speech, and Language Processing*, *20 2*, 393-403 (cf. p. 58).

Jaiswal, A., Babu, A. R., Zadeh, M. Z., Banerjee, D., & Makedon, F., (2020), A survey on contrastive self-supervised learning, *Technologies*, *9 1*, 2 (cf. p. 182).

Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A., et al., (2003), The ICSI meeting corpus, *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, *1*, I-I (cf. p. 169).

Juang, B.-H., & Rabiner, L., (2006), Speech Recognition, Automatic : History, In K. Brown (Éd.), *Encyclopedia of Language & Linguistics (Second Edition)* (Second Edition, p. 806-819), Elsevier, https://doi.org/https://doi.org/10.1016/B0-08-044854-2/00906-8, (cf. p. 19)

Jung, J.-W., Heo, H.-S., Kwon, Y., Chung, J. S., & Lee, B.-J., (2021), Three-Class Overlapped Speech Detection Using a Convolutional Recurrent Neural Network, *Interspeech*, 3086-3090, https://doi.org/10.21437/Interspeech.2021-149 (cf. p. 51-53)

Kalda, J., & Alumäe, T., (2022), Collar-aware Training for Streaming Speaker Change Detection in Broadcast Speech, *arXiv preprint arXiv :2205.07086* (cf. p. 55).

Kang, W., Roy, B. C., & Chow, W., (2020), Multimodal speaker diarization of real-world meetings using d-vectors with spatial features, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6509-6513 (cf. p. 75).

Kingma, D. P., & Ba, J., (2014), Adam : A method for stochastic optimization, *arXiv preprint arXiv :1412.6980* (cf. p. 83, 88, 136).

Kinoshita, K., Delcroix, M., & Tawara, N., (2021), Integrating end-to-end neural and clustering-based diarization : Getting the best of both worlds, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7198-7202 (cf. p. 56, 75).

Knapp, C., & Carter, G., (1976), The generalized correlation method for estimation of time delay, *IEEE transactions on acoustics, speech, and signal processing*, *24 4*, 320-327 (cf. p. 70).

Kristjansson, T., Deligne, S., & Olsen, P., (2005), Voicing features for robust speech detection, *Entropy*, *22.5*, 3 (cf. p. 47, 49).

Kunesová, M., Hrúz, M., Zajić, Z., & Radová, V., (2019), Detection of overlapping speech for the purposes of speaker diarization, *International Conference on Speech and Computer*, 247-257 (cf. p. 51, 53).

Kunešová, M., & Zajić, Z., (2023), Multitask Detection of Speaker Changes, Overlapping Speech and Voice Activity Using Wav2vec 2.0, *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5 (cf. p. 183).

Landini, F., Glembek, O., Matějka, P., Rohdin, J., Burget, L., Diez, M., & Silnova, A., (2021), Analysis of the but Diarization System for Voxconverse Challenge, *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5819-5823, https://doi.org/10.1109/ICASSP39728.2021.9414315 (cf. p. 58, 127)

Landini, F., Lozano-Diez, A., Diez, M., & Burget, L., (2022a), From Simulated Mixtures to Simulated Conversations as Training Data for End-to-End Neural Diarization, *Proc. Interspeech 2022*, 5095-5099, https://doi.org/10.21437/Interspeech.2022-10451 (cf. p. 56, 126)

Landini, F., Profant, J., Diez, M., & Burget, L., (2022b), Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization : theory, implementation and analysis on standard tasks, *Computer Speech & Language*, *71*, 101254 (cf. p. 57, 82, 126, 127, 129, 148, 169, 178).

Larcher, A., Mehrish, A., Tahon, M., Meignier, S., Carrive, J., Doukhan, D., Galibert, O., & Evans, N., (2021), Speaker embeddings for diarization of broadcast data in the allies challenge, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5799-5803 (cf. p. 57, 74, 181).

Lathoud, G., Odobez, J.-M., & Gatica-Perez, D., (2004), AV16. 3 : An audio-visual corpus for speaker localization and tracking, *International Workshop on Machine Learning for Multimodal Interaction*, 182-195 (cf. p. 170).

Lavechin, M., Gill, M.-P., Bousbib, R., Bredin, H., & Garcia-Perera, L. P., (2020), End-to-End Domain-Adversarial Voice Activity Detection, *Proc. Interspeech 2020*, 3685-3689, https://doi.org/10.21437/Interspeech.2020-2285 (cf. p. 48, 49, 51, 74)

Lebourdais, M., et al., (2022), Overlapped speech and gender detection with WavLM pre-trained features, *Proc. Interspeech 2022*, 5010-5014, https://doi.org/10.21437/Interspeech.2022-10825 (cf. p. 51, 53, 165)

Lecomte, P., (2016), *Ambisonie d'ordre élevé en trois dimensions : captation, transformations et décodage adaptatifs de champs sonores* (thèse de doct.), Conservatoire national des arts et metiers-CNAM, (cf. p. 175).

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D., (1989), Backpropagation applied to handwritten zip code recognition, *Neural computation*, *1 4*, 541-551 (cf. p. 32, 34).

Lee, S., Kim, J., Park, J., & Hahn, M., (2016), Overlapping speech detection with cluster-based HMM framework, *Proceedings of the 8th International Conference on Signal Processing Systems*, 138-141 (cf. p. 50, 53).

Li, B., Sainath, T. N., Weiss, R. J., Wilson, K. W., & Bacchiani, M., (2016), Neural network adaptive beamforming for robust multichannel speech recognition (cf. p. 68).

Lin, Q., Yin, R., Li, M., Bredin, H., & Barras, C., (2019), Lstm based similarity measurement with spectral clustering for speaker diarization, *arXiv preprint arXiv :1907.10393* (cf. p. 181).

Lippmann, R. P., (1997), Speech recognition by machines and humans, *Speech communication*, *22 1*, 1-15 (cf. p. 49).

Liu, D., & Kubala, F., (1999), Fast speaker change detection for broadcast news transcription and indexing, *Sixth European Conference on Speech Communication and Technology* (cf. p. 54).

Liu, S. [Shuying], & Deng, W., (2015), Very deep convolutional neural network based image classification using small training sample size, *2015 3rd IAPR Asian conference on pattern recognition (ACPR)*, 730-734 (cf. p. 36).

Liu, Y., Zhang, P., & Hain, T., (2014), Using neural network front-ends on far field multiple microphones based speech recognition, *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5542-5546 (cf. p. 22).

Löllmann, H. W., Evers, C., Schmidt, A., Mellmann, H., Barfuss, H., Naylor, P. A., & Kellermann, W., (2018), The LOCATA challenge data corpus for acoustic source localization and tracking, *2018 IEEE 10th Sensor Array and Multichannel Signal Processing Workshop (SAM)*, 410-414 (cf. p. 170).

Luo, Y., Chen, Z., Mesgarani, N., & Yoshioka, T., (2020), End-to-end microphone permutation and number invariant multi-channel speech separation, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6394-6398 (cf. p. 72, 80).

Luo, Y., & Mesgarani, N., (2019), Conv-tasnet : Surpassing ideal time–frequency magnitude masking for speech separation, *IEEE/ACM transactions on audio, speech, and language processing*, *27 8*, 1256-1266 (cf. p. 45).

Luong, M.-T., Pham, H., & Manning, C. D., (2015), Effective approaches to attention-based neural machine translation, *arXiv preprint arXiv :1508.04025* (cf. p. 40, 41).

Maciejewski, M., Snyder, D., Manohar, V., Dehak, N., & Khudanpur, S., (2018), Characterizing Performance of Speaker Diarization Systems on Far-Field Speech Using Standard Methods, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5244-5248, https://doi.org/10.1109/ICASSP.2018.8461546 (cf. p. 21, 73)

Mariotte, T., Larcher, A., Montrésor, S., & Thomas, J.-H., (2022), Microphone Array Channel Combination Algorithms for Overlapped Speech Detection, *Proc. Interspeech 2022*, 4636-4640, https://doi.org/10.21437/Interspeech.2022-10758 (cf. p. 117)

Marti, A., Cobos, M., & Lopez, J. J., (2011), Real time speaker localization and detection system for camera steering in multiparticipant videoconferencing environments, *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2592-2595 (cf. p. 71).

McCulloch, W. S., & Pitts, W., (1943), A logical calculus of the ideas immanent in nervous activity, *The bulletin of mathematical biophysics*, *5 4*, 115-133 (cf. p. 32).

Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y., Korenevskaya, M., Sorokin, I., Timofeeva, T., Mitrofanov, A., Andrusenko, A., Podluzhny, I., et al., (2020), Target-speaker voice activity detection : a novel approach for multi-speaker diarization in a dinner party scenario, *arXiv preprint arXiv :2005.07272* (cf. p. 75).

Minhua, W., Kumatani, K., Sundaram, S., Ström, N., & Hoffmeister, B., (2019), Frequency domain multi-channel acoustic modeling for distant speech recognition, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6640-6644 (cf. p. 68, 116).

Nemer, E., Goubran, R., & Mahmoud, S., (2001), Robust voice activity detection using higher-order statistics in the LPC residual domain, *IEEE Transactions on Speech and Audio Processing*, *9 3*, 217-231 (cf. p. 47, 49).

Ng, T., Zhang, B., Nguyen, L., Matsoukas, S., Zhou, X., Mesgarani, N., Veselỳ, K., & Matějka, P., (2012), Developing a speech activity detection system for the DARPA RATS program, *Thirteenth annual conference of the international speech communication association* (cf. p. 48, 49).

Ochiai, T., Watanabe, S., Hori, T., Hershey, J. R., & Xiao, X., (2017), Unified architecture for multichannel end-to-end speech recognition with neural beamforming, *IEEE Journal of Selected Topics in Signal Processing*, *11 8*, 1274-1288 (cf. p. 67).

Otterson, S., & Ostendorf, M., (2007), Efficient use of overlap information in speaker diarization, *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 683-686 (cf. p. 23, 127, 178).

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S., (2015), Librispeech : an asr corpus based on public domain audio books, *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206-5210 (cf. p. 120).

Pariente, M., Cornell, S., Deleforge, A., & Vincent, E., (2020), Filterbank design for end-to-end speech separation, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6364-6368 (cf. p. 45, 80, 95, 100).

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., & Narayanan, S., (2022), A review of speaker diarization : Recent advances with deep learning, *Computer Speech & Language*, *72*, 101317 (cf. p. 22, 47, 56, 57, 127).

Park, T., Kumatani, K., Wu, M., & Sundaram, S., (2020), Robust multi-channel speech recognition using frequency aligned network, *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6859-6863 (cf. p. 68).

Pfau, T., Ellis, D. P., & Stolcke, A., (2001), Multispeaker speech activity detection for the ICSI meeting recorder, *IEEE Workshop on Automatic Speech Recognition and Understanding, 2001. ASRU'01.*, 107-110 (cf. p. 48, 49).

Ravanelli, M., & Bengio, Y., (2018), Speaker recognition from raw waveform with sincnet, *2018 IEEE Spoken Language Technology Workshop (SLT)*, 1021-1028 (cf. p. 45, 99).

Ryant, N., Liberman, M., & Yuan, J., (2013), Speech activity detection on youtube using deep neural networks., *INTERSPEECH*, 728-731 (cf. p. 48, 49).

Sainath, T. N., Weiss, R. J., Wilson, K. W., Li, B., Narayanan, A., Variani, E., Bacchiani, M., Shafran, I., Senior, A., Chin, K., et al., (2017), Multichannel signal processing with deep neural networks for automatic speech recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25 5*, 965-979 (cf. p. 45, 68).

Sajjan, N., Ganesh, S., Sharma, N., Ganapathy, S., & Ryant, N., (2018), Leveraging LSTM models for overlap detection in multi-party meetings, *ICASSP*, 5249-5253 (cf. p. 51, 53).

Schneider, S., Baevski, A., Collobert, R., & Auli, M., (2019), wav2vec : Unsupervised pre-training for speech recognition, *arXiv preprint arXiv :1904.05862* (cf. p. 43).

Schuster, M., & Paliwal, K. K., (1997), Bidirectional recurrent neural networks, *IEEE transactions on Signal Processing*, *45 11*, 2673-2681 (cf. p. 40).

Schwarz, A., & Kellermann, W., (2015), Coherent-to-Diffuse Power Ratio Estimation for Dereverberation, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *23 6*, 1006-1018, https://doi.org/10.1109/TASLP.2015.2418571 (cf. p. 67, 90)

Sharma, D., Gong, R., Fosburgh, J., Kruchinin, S. Y., Naylor, P. A., & Milanović, L., (2022), Spatial processing front-end for distant ASR exploiting self-attention channel combinator, *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7997-8001 (cf. p. 117).

Siegler, M. A., Jain, U., Raj, B., & Stern, R. M., (1997), Automatic segmentation, classification and clustering of broadcast news audio, *Proc. DARPA speech recognition workshop*, *1997* (cf. p. 54, 55).

Sivasankaran, S., (2020), *Séparation de la parole guidée par la localisation* (thèse de doct.), Université de Lorraine, (cf. p. 70).

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S., (2018), X-vectors : Robust dnn embeddings for speaker recognition, *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5329-5333 (cf. p. 23, 57, 74, 181).

Sohn, J., Kim, N. S., & Sung, W., (1999), A statistical model-based voice activity detection, *IEEE signal processing letters*, *6 1*, 1-3 (cf. p. 48, 49).

SongGong, K., & Chen, H., (2021), Robust Indoor Speaker Localization in the Circular Harmonic Domain, *IEEE Transactions on Industrial Electronics*, *68 4*, 3413-3422, https://doi.org/10.1109/TIE.2020.2979556 (cf. p. 81, 133, 139)

Souden, M., Benesty, J., & Affes, S., (2009), On optimal beamforming for noise reduction and interference rejection, *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 109-112, https://doi.org/10.1109/ASPAA.2009.5346525 (cf. p. 66, 90)

Srivastava, P., Deleforge, A., Politis, A., & Vincent, E., (2022), How to (virtually) train your sound source localizer, *arXiv preprint arXiv :2211.16958* (cf. p. 175).

Stevens, S. S., Volkmann, J., & Newman, E. B., (1937), A scale for the measurement of the psychological magnitude pitch, *The journal of the acoustical society of america*, *8 3*, 185-190 (cf. p. 44).

Taherian, H., Eskimez, S. E., Yoshioka, T., Wang, H., Chen, Z., & Huang, X., (2022), One model to enhance them all : array geometry agnostic multi-channel personalized speech enhancement, *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 271-275 (cf. p. 72, 73, 80, 81).

Tan, K., Wang, Z.-Q., & Wang, D., (2022), Neural spectrospatial filtering, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *30*, 605-621 (cf. p. 69).

Thomas, S., Ganapathy, S., Saon, G., & Soltau, H., (2014), Analyzing convolutional neural networks for speech activity detection in mismatched acoustic conditions, *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2519-2523 (cf. p. 48, 49).

Torres, A. M., Mateo, J., & Cobos, M., (2016), Room Acoustics Analysis Using Circular Arrays : A Comparison Between Plane-Wave Decomposition and Modal Beamforming Approaches, *Circuits, Systems, and Signal Processing*, *35 5*, 1625-1642, https://doi.org/10.1007/s00034-015-0133-2 (cf. p. 133, 138)

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I., (2017), Attention is all you need, *Advances in neural information processing systems*, *30* (cf. p. 41-43, 46).

Vijayasenan, D., Valente, F., & Bourlard, H., (2011), An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization, *IEEE Transactions on Audio, Speech, and Language Processing*, *19 2*, 431-438, https://doi.org/10.1109/TASL.2010.2048603 (cf. p. 71, 74)

Vijayasenan, D., Valente, F., & Bourlard, H., (2012), Multistream speaker diarization of meetings recordings beyond MFCC and TDOA features, *Speech Communication*, *54 1*, 55-67 (cf. p. 74).

Vincent, E., Virtanen, T., & Gannot, S., (2018), *Audio source separation and speech enhancement*, John Wiley & Sons, (cf. p. 67, 69, 70).

Vipperla, R., Geiger, J. T., Bozonnet, S., Wang, D., Evans, N., Schuller, B., & Rigoll, G., (2012), Speech overlap detection and attribution using convolutive non-negative sparse coding, *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4181-4184, https://doi.org/10.1109/ICASSP.2012.6288840 (cf. p. 50, 53)

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L., (2018), Speaker diarization with LSTM, *2018 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, 5239-5243 (cf. p. 57, 70, 74, 75).

Wang, Z.-Q., Le Roux, J., & Hershey, J. R., (2018), Multi-channel deep clustering : Discriminative spectral and spatial embeddings for speaker-independent speech separation, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1-5 (cf. p. 71).

Wang, Z.-Q., Wang, P., & Wang, D., (2020), Complex spectral mapping for single-and multi-channel speech enhancement and robust ASR, *IEEE/ACM transactions on audio, speech, and language processing*, *28*, 1778-1787 (cf. p. 68).

Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., et al., (2020), CHiME-6 challenge : Tackling multispeaker speech recognition for unsegmented recordings, *arXiv preprint arXiv :2004.09249* (cf. p. 59, 64, 75).

Williams, E. G., & Mann III, J. A., (2000), *Fourier acoustics : sound radiation and nearfield acoustical holography*, Acoustical Society of America, (cf. p. 133, 137).

Wölfel, M., & McDonough, J., (2009), *Distant speech recognition*, John Wiley & Sons, (cf. p. 116, 120).

Woo, K.-H., Yang, T.-Y., Park, K.-J., & Lee, C., (2000), Robust voice activity detection algorithm for estimating noise spectrum, *Electronics Letters*, *36 2*, 180-181 (cf. p. 47).

Wu, C., Zhou, L., Chen, X., & Chen, L., (2021), Microphone array speech separation algorithm based on dnn, *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 1305-1310 (cf. p. 71).

Yella, S. H., & Bourlard, H., (2014), Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22 12*, 1688-1700 (cf. p. 50, 53).

Yin, R., Bredin, H., & Barras, C., (2017), Speaker change detection in broadcast tv using bidirectional long short-term memory networks, *Interspeech 2017* (cf. p. 55, 134, 202, 203).

Yoshioka, T., Erdogan, H., Chen, Z., & Alleva, F., (2018), Multi-microphone neural speech separation for far-field multi-talker speech recognition, *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5739-5743 (cf. p. 22).

Yu, F. [Fan], Zhang, S., Fu, Y., Xie, L., Zheng, S., Du, Z., Huang, W., Guo, P., Yan, Z., Ma, B., et al., (2022), M2MeT : The ICASSP 2022 multi-channel multi-party meeting transcription challenge, *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6167-6171 (cf. p. 75).

Yu, F. [Fisher], & Koltun, V., (2015), Multi-scale context aggregation by dilated convolutions, *arXiv preprint arXiv :1511.07122* (cf. p. 36).

Zhang, A., Wang, Q., Zhu, Z., Paisley, J., & Wang, C., (2019), Fully supervised speaker diarization, *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6301-6305 (cf. p. 57).

Zheng, S., Huang, W., Wang, X., Suo, H., Feng, J., & Yan, Z., (2021), A real-time speaker diarization system based on spatial spectrum, *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7208-7212 (cf. p. 71, 75).

Troisième partie

Annexes

Annexe A

Formulation des modèles complexes

Contexte

Les modèles cSACC présentés dans le chapitre 5 peuvent exploiter plusieurs formulations de la TFCT. Cette annexe présente les résultats obtenus sur la tâche de détection de parole superposée pour chaque approche. Trois approches sont considérées :

- Parties réelle et imaginaire ($\mathfrak{R}$-$\mathfrak{I}$),
- Magnitude et phase (Mag-ϕ),
- Magnitude et sinus de la phase (Mag-sin ϕ).

Dans le dernier cas, la phase de la TFCT est encodée à l'aide d'une fonction sinus afin d'obtenir une représentation continue de ce paramètre.

Résultats

Le tableau A.1 présente les résultats obtenus avec le modèle complexe IcSACC. Il montre que la formulation choisie impacte faiblement les performances de détection. Le modèle $\mathfrak{R}$-$\mathfrak{I}$ obtient un F1-score de 64,5% sur les données d'évaluation contre 64,4% pour le modèle Mag-ϕ et 64,8% pour Mag-sin ϕ. Ce dernier modèle offre cependant des performances légèrement supérieures sur les données de développement. C'est donc cette formulation qui est retenue par la suite.

Table A.1 – Évaluation des performances sur la tâche d'OSD du modèle IcSACC pour chaque formulation.

AMI	F1-score (%)		Précision (%)		Rappel (%)	
	Dev	Eval	Dev	Eval	Dev	Eval
$\mathfrak{R}$-$\mathfrak{I}$	**67.1**	**64.5**	67.6	63.0	66.6	66.0
Mag-ϕ	**68.4**	**64.4**	**68.6**	75.7	68.2	56.0
Mag-sin ϕ	**69.1**	**64.8**	**69.4**	74.5	68.6	57.4

Le tableau A.2 présente les résultats obtenus avec chaque formulation dans le cas du modèle EcSACC. Les observations sont similaires. L'encodage de la phase à l'aide d'une fonction

sinusoïdale améliore légèrement les performances avec un F1-score de 71,7% sur les données de développement et de 66,7% pour l'évaluation.

TABLE A.2 – Évaluation des performances sur la tâche d'OSD du modèle EcSACC pour chaque formulation.

AMI	F1-score (%)		Precision (%)		Recall (%)	
	Dev	Eval	Dev	Eval	Dev	Eval
$\mathfrak{R}$-$\mathfrak{I}$	70.3	66.1	**72.5**	69.7	68.3	62.8
Mag-ϕ	71.6	65.3	70.7	**73.8**	**72.4**	58.6
Mag-sin ϕ	**71.7**	**66.7**	**72.2**	64.0	71.2	**70.0**

Conclusions

Cette annexe présente les différentes formulations explorées pour la conception des extensions de l'approche SACC dans le domaine complexe. Les résultats obtenus montrent que l'utilisation de la magnitude et du sinus phase de la TFCT permet une légère amélioration des performances de détection. C'est donc cette formulation qui est retenue dans les travaux menés.

Annexe B

Détection de changement de locuteurs : validation de la formulation

---

Contexte

La détection de changement de locuteurs (SCD) est la tâche consistant à détecter les instants des tours de parole dans un signal. Deux approches peuvent être considérées pour la résoudre :

- classification à la trame : chaque trame est classée comme contenant un changement ou n'en contenant pas,
- régression : les changements de locuteurs sont encodés par une fonction continue dont les maxima sont localisés aux instants des changements.

Dans le premier cas, les classes sont fortement déséquilibrées (les tours de parole sont rares au sein d'un signal). Les labels sont donc couramment augmentés en appliquant une fenêtre autour des valeurs positives (Yin et al. 2017). Cette approche présente cependant un inconvénient lorsque les changements sont proches dans le temps. Les fenêtres peuvent se superposer les unes par rapport aux autres. La figure B.1 présente la procédure d'augmentation des annotations ainsi que le problème de recouvrement pouvant intervenir.

Dans le second cas, les tours de parole sont encodés par une fonction continue. Cette

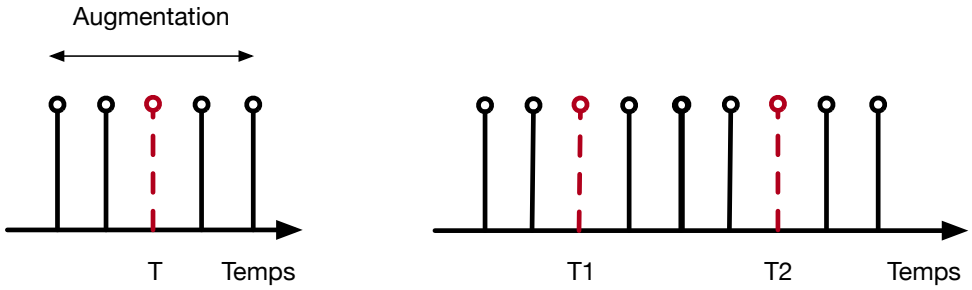


Figure B.1 – Annotation des tours de parole pour la Classification. Le trait discontinu représente un tour de parole. Les traits pleins représentent les annotations augmentées. (Gauche) principe de l'augmentation des annotations pour un tour de parole T. (Droite) Problème de recouvrement intervenant en cas de deux tours de parole T1 et T2 proches.

formulation évite le recouvrement entre les annotations dans le cas où les changements apparaissent fréquemment. Il est cependant nécessaire de vérifier que les performances obtenues avec cette approche sont similaires à celles obtenues à l'aide de la classification.

L'influence de chaque formulation est évaluée dans la section suivante. Le protocole d'évaluation est identique à celui présenté dans le chapitre 4.3. Dans le cas de la classification, les annotations des tours de paroles sont augmentées à l'aide d'une fenêtre de 100 ms centrée sur le changement (Yin et al. 2017). Dans le cas de la régression, nous utilisons la même approche que celle proposée dans la section 6.1. Outre la différence de formulation, le protocole d'apprentissage est identique entre la classification et la régression.

Un modèle de segmentation TCN est entraîné avec des caractéristiques acoustiques Log-Mel dans les deux cas de figure. Ces caractéristiques acoustiques sont également combinées aux caractéristiques spatiales IPD et CSIPD. Les modèles sont évalués sur les données de développement du corpus AMI à l'aide de la pureté, de la couverture et du S-score.

Comparaison des deux approches

La première partie du tableau B.1 présente les performances sur la tâche de SCD formulée en classification. Cette approche permet d'obtenir un S-score de 82,1% dans le cas du microphone distant unique (MDU). Les caractéristiques spatiales IPD atteignent un S-score inférieur avec 78,7%, et les CSIPD améliorent la détection avec 83,3%.

La seconde partie du tableau B.1 présente les performances sur la tâche de SCD formulée en régression. Le MDU obtient un S-score de 81,6%. Les caractéristiques spatiales IPD atteignent un S-score inférieur avec 78,6%, et les CSIPD améliorent la détection avec 85,2%.

La formulation en régression permet d'atteindre des performances similaires à la classification. Dans le cas de l'ajout de caractéristiques spatiales CSIPD, les performances sont même améliorées de +1,9% en absolu.

Table B.1 – Performance de SCD sur les données de développement du corpus AMI dans le cas de la parole distante. Les formulations en classification et en régression sont présentées.

	Modèle	Pureté (%)	Couverture (%)	S-score (%)
Class.	Log-Mel MDU	82.9	81.4	82.1
	Log-Mel + IPD	77.8	79.6	78.7
	Log-Mel + CSIPD	**83.8**	**82.6**	**83.3**
Reg.	Log-Mel MDU	82.3	80.9	81.6
	Log-Mel + IPD	77.6	79.6	78.6
	Log-Mel + CSIPD	**84.0**	**86.4**	**85.2**

Conclusions

Cette annexe étudie deux types de formulation pour la détection de changements de locuteurs (SCD) : la classification à la trame et la régression. Les résultats obtenus pour plusieurs types de caractéristiques pour la segmentation de la parole distante montrent que la régression obtient des performances similaires à la classification. Les performances sont légèrement améliorées lors de l'ajout de caractéristiques CSIPD.

Annexe C

Apprentissage invariant au nombre de canaux : influence de la permutation des canaux

---

Contexte

La section 7.1 présente une méthode d'apprentissage forçant l'invariance des modèles de détection de parole superposée (OSD) au nombre de canaux. Dans la formulation originale de la méthode, aucune permutation n'est appliquée sur l'ordre des canaux. Cette annexe étudie l'impact d'une telle procédure sur les performances et la robustesse.

Le protocole est identique à celui de la section 7.1. Lorsque qu'elle est utilisée, la permutation des canaux est obtenue par un tirage aléatoire parmi les indices des canaux. Ce tirage suit une distribution uniforme discrète.

Résultats

Le tableau C.1 présente les résultats d'un modèle invariant SACC+TFCT dans les cas où la permutation des canaux est considérée ou non au cours de l'apprentissage. Dans le cas où aucune permutation n'est considérée, le modèle obtient un F1-score de 72,3% lorsque tous les microphones sont activés (M = 8). Il atteint un F1-score de 70,0% dans le cas M = 4 et 69,1% pour M = 2. Pour plus de détails concernant les performances des modèles invariants, le lecteur peut se référer à la section 7.1.

Dans le cas où les permutations aléatoires des canaux sont considérées, le modèle atteint un F1-score de 72,3% pour M = 8, 70,2% pour M = 4, et 69,9% pour M = 2. Les performances sont donc très similaires à celles obtenues sans appliquer de permutation.

Conclusions

Dans cette annexe, nous étudions l'influence des permutations de canaux aléatoires pour l'apprentissage des modèles invariants au nombre de canaux. L'étude menée sur le modèle SACC+TFCT montre que les permutations n'améliorent pas les performances. Elles ne sont donc pas considérées pour les travaux présentés dans le chapitre 7.

TABLE C.1 – Performances d'OSD du modèle SACC+TFCT, entraîné avec l'invariance $\mathcal{L}_{inv}^{F}$, avec et sans permutation. Nous considérons les cas où $M = 8$, $M = 4$ et $M = 2$ microphones sont activés. Les résultats sont obtenus sur les données de développement du corpus AMI.

Permutation	# micro.	F1-score (%)	Précision (%)	Rappel (%)
Non	$M = 8$	72.3	72.1	72.5
	$M = 4$	70.0	72.2	67.9
	$M = 2$	69.1	70.5	67.9
Oui	$M = 8$	72.9	72.8	72.9
	$M = 4$	70.2	73.0	67.5
	$M = 2$	69.9	72.0	67.9

Annexe D

Sélection du formalisme LcSACC

Influence du nombre de fréquences

La table D.1 présente les performances obtenues par le modèle LcSACC $\mathfrak{R}/\mathfrak{I}$ sur la tâche d'OSD en fonction du nombre de fréquences considérées. Les résultats sont présentés uniquement sur les données de développement du corpus AMI.

Table D.1 – Influence du nombre de fréquences considérées dans la TFCT sur les performances d'OSD.

# Fréquences	Freq. Att.	Précision	Rappel	F1-score
129	✗	69,5	68,4	69,0
	✓	71,5	66,8	69,1
257	✗	71,7	70,0	70,8
	✓	71,7	69,6	70,6
513	✗	73,9	68,3	71,0
	✓	73,6	66,8	70,0

La table montre que la résolution fréquentielle choisie impacte faiblement les performances du système. Dans le cas où seulement 129 fréquences sont utilisées, le modèle atteint un F1-score de 69,1% sans considérer les fréquences, et de 69,0% lorsqu'elles sont conservées dans le module d'attention. Augmenter le nombre de fréquences à 257 permet un léger gain avec 70,8% et 70,6% respectivement avec et sans dépendance aux fréquences. Enfin, choisir 513 fréquences n'améliore pas les performances avec 71,0% et 70,0% respectivement. La dépendance en fréquence semble cependant dégrader les performances du modèle dans ce cas de figure.

Le nombre de fréquences ayant un impact limité sur les performances, le modèle LcSACC utilisant 257 fréquences, en conservant cette dimension dans le mécanisme d'attention, est utilisé pour la comparaison aux autres modèles. Ces résultats sont présentés dans le paragraphe suivant.

Annexe E

Configurations pour la sélection de filtres spatiaux

Cette annexe présente l'analyse de différentes configurations du modèle BFSACC, introduit en section 5.6. La section E évalue l'influence du nombre de filtres spatiaux choisis sur les performances d'OSD. La section E étudie l'utilisation des poids d'attentions comme caractéristiques supplémentaires pour le modèle de segmentation. La section E présente les performances d'OSD pour plusieurs paramètres de régularisation de la formation de voies (*cf.* équation 5.31).

Influence du nombre de filtres

Le nombre de filtres P sélectionné est un paramètre du modèle BFSACC. La table E.1 présente les performances de détection de parole superposée en fonction de ce paramètre. Les résultats sont obtenus sur les données de développement du corpus AMI. Ils montrent que l'utilisation de $P = 4$ filtres mène aux F1-score le plus faible avec 68,7%. Les performances sont ensuite similaires pour $P > 4$. Par exemple, le modèle $P = 8$ obtient un F1-score de 69,5%. Seul le cas $P = 7$ atteint un score légèrement plus faible (69,0%).

Table E.1 – Influence du nombre de filtres spatiaux considérés sur les performances d'OSD. AMI dev set.

P	Précision$_{\%}$ ↑	Rappel $_{\%}$ ↑	F1-score $_{\%}$ ↑
4	70,6	67,0	68,7
5	70,3	68,7	**69,5**
6	68,4	70,5	**69,5**
7	68,6	69,4	69,0
8	71,0	68,0	**69,5**

Bien que les performances soient similaires à partir de $P = 5$, le choix est fait d'utiliser $P = 8$ dans les travaux. Augmenter le nombre de directions de focalisation permet d'affiner la résolution spatiale du modèle. Il peut ainsi permettre de localiser les sources actives comme l'a montré l'analyse de la section 5.6.3.

Poids d'attention comme caractéristiques additionnelles

Les performances du modèle BFSACC sont limitées par rapport à celles des autres approches de combinaison de canaux (*cf.* section 5.6.2). D'autre part, les poids de combinaison encodent des informations importantes sur l'activation des filtres spatiaux en fonction du temps. Donner cette connaissance au modèle pourrait en améliorer les performances de détection. Deux approches sont étudiées :

- $\mathbf{w}_{att}$: les poids d'attention sont aplatis tels que $\mathbb{R}^{P\times P\times T} \rightarrow \mathbb{R}^{P^2\times T}$ et concaténés au spectrogramme à échelle Mel.
- $\mathbf{w}_{comb}$: les poids de combinaison (Eq. 5.24) sont concaténés au spectrogramme à échelle Mel.

La table E.2 présente les performances d'OSD obtenues pour chaque cas. Elle montre que l'ajout des poids comme caractéristiques n'améliore pas, voire dégrade les performances. Dans le cas $\mathbf{w}_{att}$, le modèle obtient un F1-score de seulement 68,6%. L'ajout des poids de combinaison $\mathbf{w}_{comb}$ n'a aucune influence avec un F1-score de 69,6%, soit un gain absolu de +0,1% par rapport au modèle d'origine.

TABLE E.2 – Impact de l'utilisation des poids d'attention comme caractéristiques additionnelles sur les performances d'OSD dans le cas $P = 8$.

Info. Sup.	Précision$_{\%}$ ↑	Rappel $_{\%}$ ↑	F1-score $_{\%}$ ↑
-	71,0	68,0	**69,5**
$\mathbf{w}_{att}$	68,8	68,4	68,6
$\mathbf{w}_{comb}$	69,4	69,8	**69,6**

L'intégration des poids de combinaison obtenus par le module d'auto-attention comme caractéristiques supplémentaires n'améliore pas les performances de segmentation. Cette architecture a donc été abandonnée par la suite. L'étude d'autres schémas de fusion pourrait permettre de tirer meilleur bénéfice de l'information contenue dans les poids de combinaison.

Régularisation des filtres

La section 5.6.4 a montré que l'ajout d'un terme de régularisation dans le calcul des poids des filtres spatiaux améliore significativement les performances. Cette section étudie l'impact de différentes valeurs de ce paramètre.

La table E.3 montre que le terme de régularisation a un impact non négligeable sur les performances d'OSD. Un paramètre trop faible (ex : $\lambda = 10^{-6}$) n'affecte pas la détection, voire la dégrade légèrement. Un paramètre trop grand ($\lambda = 10^{-2}$) semble également dégrader les performances. Dans ce cas, le terme de régularisation est prépondérant sur les valeurs de la matrice interspectrale. Cette dernière ne décrit donc plus la physique correctement. Le paramètre

TABLE E.3 – Détection de parole superposée à l'aide du modèle BFSACC sur les données de développement du corpus AMI en fonction du paramètre de régularisation λ.

λ	Précision$_{\%}$ ↑	Rappel $_{\%}$ ↑	F1-score $_{\%}$ ↑
0	72,5	66,8	69,6
10^{-6}	69,9	68,2	69,1
10^{-4}	70,5	71,6	**71,0**
10^{-2}	70,3	70,4	70,4

$\lambda = 10^{-4}$ semble être un bon compromis en menant à un F1-score de 71,0% sur les données de développement.

COLLEGE SCIENCES
DOCTORAL DE L'INGENIERIE
PAYS DE LA LOIRE ET DES SYSTEMES

Le Mans Université

Titre : Traitement automatique de la parole en réunion par dissémination de capteurs

Mot clés : parole distante, antennes de microphones, segmentation automatique de la parole, diarisation en locuteurs, apprentissage profond

Résumé : Ces travaux de thèse se concentrent sur le traitement automatique de la parole, et plus particulièrement sur la diarisation en locuteurs. Cette tâche nécessite de segmenter le signal afin d'identifier des évènements tels que la présence de parole, de parole superposée ou de changements de locuteur. Cette recherche se focalise sur le cas où le signal est capté par un dispositif placé au centre d'un groupe de locuteurs, comme lors de réunions. Ces conditions entraînent une dégradation de la qualité des signaux en raison de l'éloignement des sources sonores (parole distante).

Afin de pallier cette dégradation, une approche consiste à enregistrer le signal à l'aide d'un ensemble de microphones formant une antenne acoustique. Le signal multicanal obtenu permet d'obtenir des informations sur la répartition spatiale du champ acoustique. Deux axes de recherche sont explorés pour la segmentation de la parole à l'aide d'antennes de microphones.

Le premier axe introduit une méthode combinant des caractéristiques acoustiques avec des caractéristiques spatiales. Un nouveau jeu de caractéristiques, basé sur le formalisme des harmoniques circulaires, est proposé. Cette approche améliore les performances de segmentation en conditions distantes, tout en réduisant le nombre de paramètres des modèles et en garantissant une certaine robustesse en cas de désactivation de certains microphones.

Le second axe propose plusieurs approches de combinaison des canaux en utilisant des mécanismes d'auto-attention. Différents modèles, inspirés d'une architecture existante, sont développés. La combinaison de canaux améliore également la segmentation en conditions distantes. Deux de ces approches rendent l'extraction de caractéristiques plus interprétable. Les systèmes de segmentation de la parole distante proposés améliorent également la diarisation en locuteurs.

La combinaison de canaux montre une faible robustesse en cas de changement de géométrie de l'antenne en phase d'évaluation. Pour y remédier, une procédure d'apprentissage est proposée, qui améliore la robustesse en présence d'une antenne non conforme.

Finalement, les travaux menés ont permis d'identifier un manque dans les jeux de données publics disponibles pour le traitement automatique de la parole distante. Un protocole d'acquisition est introduit pour l'acquisition de signaux en réunions et intégrant l'annotation de la position des locuteurs en plus de la segmentation.

En somme, ces travaux visent à améliorer la qualité de la segmentation de la parole distante multicanale. Les méthodes proposées exploitent l'information spatiale fournie par les antennes de microphones en garantissant une certaine robustesse au nombre de microphones disponibles.

Title: Automatic Speech Processing in Meetings using Microphone Arrays

Keywords: distant speech, multichannel audio, automatic speech segmentation, speaker diarization, deep learning

Abstract: This thesis work focuses on automatic speech processing, and more specifically on speaker diarization. This task requires the signal to be segmented to identify events such as voice activity, overlapped speech, or speaker changes. This work tackles the scenario where the signal is recorded by a device located in the center of a group of speakers, as in meetings. These conditions lead to a degradation in signal quality due to the distance between the speakers (distant speech).

To mitigate this degradation, one approach is to record the signal using a microphone array. The resulting multichannel signal provides information on the spatial distribution of the acoustic field. Two lines of research are being explored for speech segmentation using microphone arrays.

The first introduces a method combining acoustic features with spatial features. We propose a new set of features based on the circular harmonics expansion. This approach improves segmentation performance under distant speech conditions while reducing the number of model parameters and improving robustness in case of change in the array geometry.

The second proposes several approaches that combine channels using self-attention. Different models, inspired by an existing architecture, are developed. Combining channels also improves segmentation under distant speech conditions. Two of these approaches make feature extraction more interpretable. The proposed distant speech segmentation systems also improve speaker diarization.

Channel combination shows poor robustness to changes in the array geometry during inference. To avoid this behavior, a learning procedure is proposed, which improves the robustness in case of array mismatch.

Finally, we identified a gap in the public datasets available for distant multichannel automatic speech processing. An acquisition protocol is introduced to build a new dataset, integrating speaker position annotation in addition to speaker diarization.

Thus, this work aims to improve the quality of multichannel distant speech segmentation. The proposed methods exploit the spatial information provided by microphone arrays while improving the robustness in case of array mismatch.