Détection de parole superposée distante à l'aide d'une antenne de microphones

Théo Mariotte, Anthony Larcher, Jean-Hugh Thomas, Silvio Montrésor

Détection de parole superposée distante à l'aide d'une antenne de microphones

Théo Mariotte[1, 2] Anthony Larcher[2] Silvio Montrésor[1] Jean-Hugh Thomas[1]
(1) LAUM, UMR 6613, IA-GS, CNRS, Le Mans Université, Avenue Olivier Messiaen, 72085 Le Mans, France
(2) LIUM, Le Mans Université, Avenue René Laennec, 72085 Le Mans, France
theo.mariotte@univ-lemans.fr

RÉSUMÉ

La parole superposée correspond à l'activité simultanée de plusieurs locuteurs dans un enregistrement audio. Elle est source de dégradation des performances des modèles de traitement automatique de la parole. C'est notamment le cas dans le contexte de la segmentation et du regroupement en locuteurs. La détection de parole superposée détermine les instants où ces évènements interviennent. La généralisation de l'utilisation des réseaux de neurones a permis un gain significatif en performances sur cette tâche. La détection tend cependant à se dégrader en conditions de parole distante. Les travaux présentés étudient l'utilisation de mécanismes d'auto-attention pour combiner les canaux issus des différents microphones d'une antenne. Cette approche est mise en œuvre dans le contexte de la détection de parole superposée distante et permet de s'approcher des performances obtenues en champ proche.

ABSTRACT

Distant Overlapped Speech Detection Using a Microphone Array

Overlapped speech occurs when multiple speakers are simultaneously active. This may lead to severe performance degradation in automatic speech processing tasks like diarisation. Overlapped speech detection aims at detecting time segments where several speakers are simultaneously active. Recently, deep neural networks have been extensively used to solve this task. However, performance tends to deteriorate in the far-field scenario. To tackle this problem, microphone arrays can be used to record audio signals. This paper investigates the use of self-attention channel combinators as features extractor for overlapped speech detection. Results show that self-attentive models can help detection and get closer to the close-talk scenario performances.

MOTS-CLÉS : Détection de parole superposée, multicanal, auto-attention, parole lointaine.

KEYWORDS: Overlapped Speech Detection, multichannel audio, self-attention, far field.

1 Introduction

1.1 Contexte

La parole superposée correspond à l'activité simultanée de plusieurs locuteurs. Lorsqu'un microphone enregistre la scène acoustique, le signal résultant est une somme des contributions issues de chacun d'entre eux. La parole superposée peut entraîner une dégradation des performances des systèmes de traitement automatique de la parole (Garcia Perera *et al.*, 2020). Afin de permettre un traitement distinct des segments de parole superposée, il est nécessaire de les détecter. Historiquement, la

détection de parole superposée est réalisée en amont de la tâche de traitement de la parole envisagée, de façon similaire à la détection d'activité vocale. Les performances des modèles de détection de parole superposée tendent à se dégrader dans un contexte de parole distante, lorsque les locuteurs sont éloignés du système de captation. Ce cas de figure présente pourtant des intérêts pratiques. Les participants ne sont plus équipés de microphones individualisés, ce qui permet un traitement indépendant du nombre de locuteurs et limite le matériel de captation. Pour réduire l'influence des perturbations dans ce contexte (faible rapport signal-à-bruit, réverbération), plusieurs microphones peuvent être utilisés simultanément lors de l'acquisition des signaux de parole. La répartition des capteurs dans l'espace permet l'utilisation d'algorithmes s'appuyant sur l'information spatiale intrinsèque aux signaux pour rendre la détection de parole superposée plus robuste.
Les travaux présentés portent sur l'intégration de méthodes neuronales auto-attentives pour combiner automatiquement les signaux issus d'une antenne de microphones pour la détection de parole superposée. Des expériences sont menées sur les données du corpus AMI (Mccowan *et al.*, 2005), enregistrées en réunion, afin d'évaluer l'influence de ces approches sur les performances de détection de parole superposée distante.

1.2 État de l'art

La détection de parole superposée apparaît dans la littérature depuis plus d'une décennie (Boakye *et al.*, 2011). Les méthodes de l'état de l'art définissent cette tâche comme la classification des éléments d'une séquence, appelés trames. Des caractéristiques prédéfinies du signal temporel sont extraites de chaque trame. Les travaux récents privilégient les réseaux de neurones profonds afin de modéliser cette séquence de caractéristiques. Nombre d'entre eux intègrent des réseaux récurrents de type LSTM (Long Short-Term Memory). Geiger *et al.* (2013) intègrent pour la première fois ce type d'architecture pour la détection de parole superposée, montrant un gain de performances au sein d'un modèle statistique. Les travaux de Bredin *et al.* (2020) et Bullock *et al.* (2020) rapportent de bonnes performances de détection sur plusieurs jeux de données à l'aide d'architectures LSTM. Dans chacune de ces approches, la classification des trames est binaire : chacune appartient ou non à la classe *Parole superposée*.
Dans le contexte de la parole lointaine, les antennes de microphones permettent d'acquérir un signal multicanal, chacun étant issu d'un microphone à une position fixe. Les données intègrent ainsi des informations spatiales sur le champ acoustique. Intuitivement, celles-ci peuvent favoriser la détection de parole superposée. Cornell *et al.* (2020, 2022) proposent d'extraire des caractéristiques spatiales en plus de celles extraites du signal temporel issu d'un des microphones d'une antenne. Les données spatiales permettent d'améliorer les performances de classification, mais nécessitent l'extraction de caractéristiques spatiales pré-définies. Zheng *et al.* (2021) proposent de fusionner les signaux de formation de voies focalisés dans plusieurs directions à l'aide de Transformers (Vaswani *et al.*, 2017). Cette approche permet de bonnes performances de détection, mais le modèle contient un grand nombre de paramètres et nécessite beaucoup de ressources pour l'apprentissage.
Des méthodes d'extraction de caractéristiques à partir de signaux issus d'antennes de microphones ont également été proposées pour d'autres tâches de traitement automatique de la parole. Gong *et al.* (2021) proposent l'utilisation de mécanismes d'auto-attention afin de combiner les canaux dans le domaine temps-fréquence (Self-Attentive Channel Combinator, SACC). Ces mécanismes, introduits par Vaswani *et al.* (2017), permettent aux modèles neuronaux de se focaliser automatiquement sur une partie des données. Gong *et al.* (2021) montrent que le système SACC améliore les performances de transcription automatique en champ lointain par rapport aux approches classiques de traitement du signal multicanal.

1.3 Contributions

Les travaux présentés intègrent l'architecture SACC (Gong *et al.*, 2021) pour l'extraction de caractéristiques au sein d'un système de détection de parole superposée. D'après la littérature étudiée, cette approche n'a pas été appliquée pour cette tâche. Le modèle SACC utilise uniquement le module de la Transformée de Fourier à Court Terme (TFCT) pour déterminer la pondération des canaux. Pour s'approcher des algorithmes de formation de voies traditionnels (Benesty *et al.*, 2008), le modèle est étendu dans le domaine complexe (cSACC) afin de porter l'attention sur l'intégralité de la STFT. Ces approches, mises en œuvre en conditions de captation distante, doivent permettre d'approcher les performances de détection obtenues en champ proche.
L'article est organisé comme suit : une première partie présente la tâche de détection de parole superposée et l'architecture utilisée. Une deuxième section présente les méthodes développées avant de présenter le protocole expérimental et les résultats. L'abréviation *OSD* (Overlapped Speech Detection) est parfois utilisée pour désigner la détection de parole superposée.

2 Détection de parole superposée

2.1 Principe

La détection de parole superposée consiste à détecter les instants où plusieurs locuteurs sont actifs simultanément. Il s'agit d'une tâche de classification binaire des éléments d'une séquence temporelle. Les méthodes développées utilisent un modèle séquence-à-séquence optimisé à l'aide de l'apprentissage automatique supervisé.
Soit un modèle paramétrique $\hat{\boldsymbol{y}} = f(\boldsymbol{X}; \theta)$, avec θ les paramètres du modèle, prenant en entrée une séquence de caractéristiques $\boldsymbol{X}\left[\boldsymbol{X}_1, ...\boldsymbol{X}_t, ...\boldsymbol{X}_T\right]$ où $\boldsymbol{X}_t$ représente le vecteur de caractéristiques extrait pour une trame t. Pour chaque entrée, le modèle retourne une séquence $\hat{\boldsymbol{y}} \in \mathbb{R}^{C \times T}$. La représentation $\hat{\boldsymbol{y}} = [\hat{y}_1, ...\hat{y}_t, ...\hat{y}_T]$ contient la probabilité pour chaque trame d'appartenir à chacune des classes. Dans le cas d'une classification binaire ($C = 2$), une trame t de $\hat{\boldsymbol{y}}$ est associée à $\hat{y}_t = \{p(c = 0|\boldsymbol{X}_t), p(c = 1|\boldsymbol{X}_t)\}$.
À chaque séquence sont associées des étiquettes $\boldsymbol{y} = [y_1, ...y_t, ...y_T] \in \mathbb{R}^{1 \times T}$ alignées aux trames de la séquence d'entrée et définies telles que $y_t \in \{0, 1\}$; $y_t = 1$ indique l'évènement *parole superposée*. Ces annotations binaires servent de référence pour l'apprentissage supervisé.

2.2 Architecture du modèle

Le système utilisé pour la détection de parole superposée est composé de trois parties distinctes : l'extraction de caractéristiques, la modélisation de séquence et la classification. La figure 1 présente l'architecture utilisée.

Extraction de caractéristiques : deux approches différentes sont étudiées pour extraire les caractéristiques du signal. La première concerne le cas du canal unique où les coefficients cepstraux à échelle Mel (MFCC) sont extraits du signal. En pratique, 20 coefficients sont extraits ainsi que les Δ et $\Delta\Delta$ sur des fenêtres de 25 ms avec un pas de 10 ms. La seconde intègre la méthode SACC (Gong *et al.*, 2021), basée sur des mécanismes d'auto-attention pour combiner les canaux à partir du module de la Transformée de Fourier à Court Terme (TFCT). L'extension de cette approche dans le domaine complexe est également évaluée (cSACC). La section 3 détaille les techniques d'extraction

de caractéristiques utilisées. Chaque méthode retourne une séquence de caractéristiques $\boldsymbol{X} \in \mathbb{R}^{F \times T}$ où F représente le nombre de caractéristiques extraites à chaque trame.

Modélisation de séquence : cette partie du modèle permet de prendre en compte les relations temporelles entre les trames de la séquence $\boldsymbol{X}$. Des couches neuronales récurrentes de type LSTM bi-directionnelles sont utilisées de façon similaire aux travaux de Bredin *et al.* (2020) et Bullock *et al.* (2020). Le modèle utilisé est composé de deux couches BLSTM de $K = 128$ cellules chacune. Une fonction d'activation $\tanh$ est appliquée à la sortie de chacune d'elles. Une nouvelle représentation $\boldsymbol{s} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_{2K}\}$ de la séquence $\boldsymbol{X}$ est obtenue en sortie des BLSTM. Elle contient les sorties des couches récurrentes à chaque instant. Le modèle étant bi-directionnel, la sortie contient $2K$ éléments modélisant la séquence d'entrée dans les deux directions.

Classification : la classification des trames de la séquence est réalisée à l'aide d'un perceptron multi-couches (MLP) composé de trois couches linéaires de dimensions (256;128), (128;128) et (128;2). Chaque couche linéaire est suivie d'une fonction d'activation $\tanh$ à l'exception de la dernière. Une fonction $\mathrm{softmax}$, appliquée à la sortie de la dernière couche, permet d'obtenir la séquence de probabilités $\hat{\boldsymbol{y}}$ pour chaque trame d'appartenir à chacune des classes.

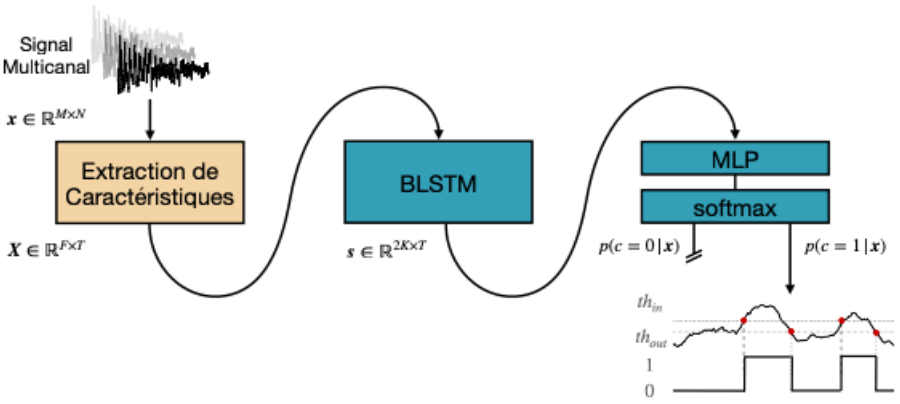

FIGURE 1 – Architecture utilisée pour la détection de parole superposée. Le modèle peut prendre en entrée un signal issu d'un ou plusieurs microphones en fonction du cas de figure considéré. La phase de détection est également représentée : seule la probabilité de la classe positive $p(c = 1|\boldsymbol{x})$ est conservée, et deux seuils de détection sont appliqués.

2.3 Apprentissage et évaluation

L'apprentissage d'un modèle de détection de parole superposée consiste à déterminer les paramètres $\hat{\theta}$ minimisant une fonction de coût $\mathcal{L}_{\hat{\theta}}(\hat{\boldsymbol{y}}, \boldsymbol{y})$. L'entropie croisée est ici utilisée comme objectif d'optimisation. Les paramètres du modèle θ sont ensuite optimisés par l'algorithme de descente de gradient stochastique. Après l'optimisation, le modèle $f(\boldsymbol{X}; \hat{\theta})$ doit permettre de détecter la présence de parole superposée au sein d'une séquence de caractéristiques $\boldsymbol{X}$ n'ayant jamais été observée par le modèle au cours de l'apprentissage.
Lors de la phase d'évaluation du modèle, seule la séquence de probabilités associée à la classe positive $p(c = 1|\boldsymbol{x})$ est conservée. Celle-ci étant continue sur l'intervalle $[0, 1]$, deux seuils de détection th_{in} et th_{out} sont appliqués afin de déterminer si une trame appartient à la classe positive ou non. Le premier seuil indique la probabilité à partir de laquelle une trame passe de la classe négative à la classe positive ; le second permet l'opération inverse. La figure 1 schématise ce processus.

3 Extraction de caractéristiques d'un signal multicanal

L'extraction de caractéristiques permet de déterminer une nouvelle représentation des données d'entrée avant la modélisation de la séquence et sa classification. Soit $\boldsymbol{x} \in \mathbb{R}^{M \times N}$ un signal de N échantillons temporels acquis par M microphones. La séquence de caractéristiques $\boldsymbol{X} \in \mathbb{R}^{F \times T}$ est obtenue par une transformation g telle que $\boldsymbol{X} = g(\boldsymbol{x})$. L'ensemble des échantillons n'étant pas nécessaire à la classification (Bredin *et al.*, 2020; Cornell *et al.*, 2022), la transformation g applique une décimation aux données d'entrée, en passant de N échantillons à T trames ($N > T$). Trois approches d'extraction de caractéristiques sont étudiées pour la détection de parole superposée.

3.1 Coefficients Cepstraux à Echelle Mel

De nombreux modèles de détection de parole superposée utilisent les MFCC pour représenter le signal de parole (Geiger *et al.*, 2013; Bredin *et al.*, 2020; Cornell *et al.*, 2020, 2022). Un spectrogramme en échelle Mel est d'abord extrait du signal temporel. Un banc de filtres triangulaires est ensuite appliqué à cette représentation. Les coefficients sont obtenus à l'aide d'une transformée en cosinus inverse. Les MFCC ne permettent pas d'intégrer d'information spatiale dans la représentation prise en entrée du modèle de détection de parole superposée. Deux méthodes prenant en compte l'ensemble des canaux issus de l'antenne sont donc mises en œuvre pour la tâche d'OSD et détaillées dans les deux sections suivantes.

3.2 Combinaison de canaux à l'aide d'un mécanisme d'attention

Lorsque des signaux issus de plusieurs microphones sont utilisés pour le traitement automatique de la parole, ils intègrent des informations spatiales sur le champ acoustique. Des techniques de filtrage spatial telles que la formation de voies permettent de minimiser la contribution du bruit et de maximiser celle d'un locuteur en pondérant et en combinant les canaux issus de l'antenne (Benesty *et al.*, 2008). Ces approches nécessitent cependant une estimation explicite de la position des locuteurs actifs, cette tâche pouvant s'avérer délicate.
Afin de combiner les canaux sans déterminer la position des locuteurs, Gong *et al.* (2021) proposent la méthode SAAC, développée initialement pour la reconnaissance de la parole. Celle-ci intègre des mécanismes d'auto-attention (Vaswani *et al.*, 2017) pour pondérer automatiquement les canaux à partir de la TFCT des signaux en entrée. Le système se focalise ainsi sur une direction particulière à chaque trame. Le processus d'extraction des caractéristiques à partir d'un signal multicanal est présenté en figure 2.

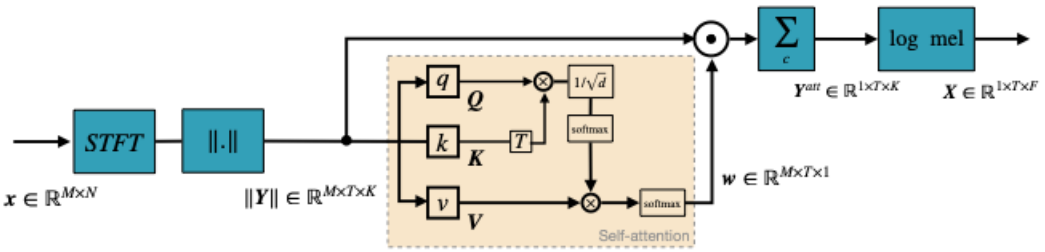

FIGURE 2 – Modèle de combinaison des canaux auto-attentive proposé par Gong *et al.* (2021). L'attention est portée sur le module de la TFCT.

La méthode de combinaison auto-attentive des canaux consiste à estimer les poids $\boldsymbol{w} \in \mathbb{R}^{T \times M \times 1}$ à appliquer à chaque canal de l'antenne en portant l'attention sur le module TFCT $\|\boldsymbol{Y}\|$. Pour cela, trois vecteurs sont estimés : la *requête* $\boldsymbol{Q} = q(\|\boldsymbol{Y}\|) \in \mathbb{R}^{T \times M \times d}$, la *clef* $\boldsymbol{K} = k(\|\boldsymbol{Y}\|) \in \mathbb{R}^{T \times M \times d}$ et la *valeur* $\boldsymbol{V} = v(\|\boldsymbol{Y}\|) \in \mathbb{R}^{T \times M \times 1}$. Les fonctions q,k et v sont implémentées par des couches neuronales linéaires. La dimension d représente le nombre de sorties des couches q et k. La couche k ne contient qu'une sortie afin de rendre les poids indépendants de la fréquence (Gong *et al.*, 2021). Ces poids sont déterminés par la relation suivante :

$$\boldsymbol{w} = \mathrm{softmax}\left(\frac{\boldsymbol{Q}.\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{d}}\right)\boldsymbol{V}. \tag{1}$$

Le module de la TFCT $\|\boldsymbol{Y}\|$ est pondéré par les poids $\boldsymbol{w}$ et les canaux sont combinés :

$$\boldsymbol{Y}^{\boldsymbol{att}} = \sum_{c} \mathrm{softmax}(\boldsymbol{w}) \odot \|\boldsymbol{Y}\|, \tag{2}$$

avec $\odot$ le produit terme-à-terme entre deux vecteurs. La fonction $\mathrm{softmax}$ permet de garantir $w_i \in [0, 1]$. Le module pondéré de la TFCT $\boldsymbol{Y}^{\boldsymbol{att}}$ est ensuite converti en échelle Mel à l'aide de $F = 64$ filtres afin d'obtenir la séquence de caractéristiques $\boldsymbol{X}$. Pour les expériences menées, la dimension de des clef et requête est fixée à $d = 256$.

3.3 Formulation dans le domaine complexe

La méthode de combinaison des canaux proposée par Gong *et al.* (2021) calcule des poids uniquement à partir du module de la TFCT. L'information de la phase est donc perdue dès le début du traitement. Cette information renseigne sur le retard temporel entre chaque microphone de l'antenne et donc intrinsèquement sur la position des locuteurs. De plus, les algorithmes de formation de voies traditionnels utilisent une pondération dans le domaine complexe. Il semble donc pertinent que les poids appliqués à la TFCT soient complexes afin de ne pas perdre d'information lors de la pondération. La méthode de Gong *et al.* (2021) est donc étendue dans le domaine complexe, nommée cSACC. Pour cela, deux mécanismes d'auto-attention sont utilisés. Le premier calcule les poids $\boldsymbol{w}^{\Re}$ sur la partie réelle de la TFCT, le second les poids $\boldsymbol{w}^{\Im}$ sur la partie imaginaire. De façon similaire à l'équation (2), la TFCT pondérée est calculée par la relation suivante :

$$\boldsymbol{Y}^{att} = \left\| \sum_{c} \mathrm{softmax}(\boldsymbol{w}^{\Re}) \odot \boldsymbol{Y}^{\Re} + j \times \mathrm{softmax}(\boldsymbol{w}^{\Im}) \odot \boldsymbol{Y}^{\Im} \right\|. \tag{3}$$

La TFCT pondérée est ainsi complexe $\boldsymbol{Y}^{att} \in \mathbb{C}^{N \times K}$ avec K le nombre de fréquences. Les caractéristiques $\boldsymbol{X}$ sont extraites à partir de $\boldsymbol{Y}^{att}$ en appliquant $F = 64$ filtres Mel au module de la TFCT pondérée.

4 Résultats

4.1 Corpus AMI

Les expériences de détection de parole superposée sont menées sur le corpus AMI (Mccowan *et al.*, 2005) contenant des données multimodales enregistrées en réunion. Deux types d'enregistrement audio sont utilisés : les signaux enregistrés par une antenne circulaire uniforme de 8 microphones

array 1, placée au centre des tables au cours des réunions et le *headset-mix*, contenant un mixage des signaux acquis par les microphones-casques des participants. Les seconds permettent d'obtenir des performances d'OSD de référence en champ proche. Le corpus est divisé en trois sous-ensembles *Train*, *Dev* et *Test* contenant respectivement 80 h, 10 h et 10 h de données annotées. Le contenu des sous-ensembles suit le protocole proposé par Landini *et al.* (2022), garantissant des locuteurs différents pour chacun d'entre eux.
Pour l'apprentissage, des séquences d'une durée $L = 2.0$ s sont tirées aléatoirement dans les données de *Train*. La parole superposée est un évènement minoritaire (13% du contenu du corpus). Pour pallier le déséquilibre entre les classes, 50% des segments d'apprentissage sont augmentés en les sommant à un second segment tiré aléatoirement (Bredin *et al.*, 2020). Aucune autre technique d'augmentation de données n'est utilisée afin de rendre les résultats comparables entre les différentes approches. En effet, les méthodes d'augmentation de données diffèrent entre les cas monocanal et multicanal. Les performances de classification des modèles sont évaluées à l'aide du F1-score, de la précision et du rappel (Bredin *et al.*, 2020; Bullock *et al.*, 2020). La précision moyenne est également reportée comme le proposent Cornell *et al.* (2020, 2022) afin que les résultats ne dépendent pas des seuils de détection utilisés.

4.2 Protocole expérimental

Afin d'évaluer l'influence des méthodes de combinaison auto-attentives des canaux SACC et cSACC sur les performances d'OSD, cinq modèles sont entraînés : trois sur des signaux mono-canal utilisés comme référence, et deux pour des signaux multi-canaux. La modélisation de séquence et la classification sont identiques dans chaque cas ; seule l'extraction de caractéristiques diffère parmi les différents systèmes. Un premier modèle est entraîné sur les données *headset-mix* du corpus AMI et fournit une référence en champ proche. Deux autres modèles sont entraînés à partir des signaux issus de l'antenne *array 1*, l'un sur la somme des signaux issus de chaque canal, l'autre sur le signal issu du premier microphone uniquement. Ils servent de référence sur les signaux de parole lointaine. Pour ces trois modèles, les caractéristiques utilisées sont les MFCC. Les deux derniers modèles intègrent respectivement les approches SACC et cSACC pour l'extraction de caractéristiques. Les signaux utilisés sont également issus de l'antenne *array 1*.
Tous les systèmes sont entraînés avec un taux d'apprentissage $lr = 10^{-3}$ durant 200 époques, chacune contenant 2000 lots de 64 segments. Dans le cas de signaux issus de l'antenne, le nombre de canaux est fixé entre l'apprentissage et l'évaluation. Des travaux seront consacrés à l'avenir afin de rendre le modèle invariant au nombre de microphones.
Les résultats rapportés sont obtenus avec les données de *Dev* et de *Test* du corpus AMI. La détection de parole superposée est réalisée sur une fenêtre glissante de $L = 2.0$ s avec 75% de recouvrement. Les seuils de détection sont ajustés en phase de développement. Ceux permettant les meilleures performances sont conservés lors de la phase d'évaluation.

4.3 Résultats

Les performances de chaque système de détection de parole superposée sont reportées dans le tableau 1. Les trois premières lignes présentent les résultats obtenus avec les trois modèles de référence (MFCC). Les meilleures performances sont obtenues en champ proche, sur les données *headset-mix* du corpus AMI. La détection est fortement dégradée lorsque les canaux de l'antenne sont sommés. Les résultats sur le premier canal sont également reportés pour comparaison. L'objectif est de se rapprocher des performances obtenues en champ proche avec l'antenne de microphones. Les deux

dernières lignes du tableau 1 montrent que l'utilisation de mécanismes d'auto-attention permet un gain en performances dans le contexte de la parole distante. Les meilleurs résultats sont obtenus avec le modèle SACC, permettant un gain de 20% sur la précision moyenne en phase d'évaluation par rapport à l'utilisation d'un canal unique. Les résultats sont également proches de ceux obtenus en champ proche, avec des valeurs de F1-score et de précision moyenne similaires en phase d'évaluation.

	F1-score (%)		Précision (%)		Rappel (%)		Pr. Moy (%)	
AMI	Dev	Eval	Dev	Eval	Dev	Eval	Dev	Eval
MFCC Headset-mix	66,4	62,1	65,7	71,9	67,3	54,6	70,4	61,8
MFCC Canal 1	46,7	38,2	46,3	40,2	46,8	36,4	51,3	41,6
MFCC Somme	36,8	23,6	43,7	38,9	31,8	16,9	41,0	32,8
SACC	**64,0**	**62,4**	**66,2**	**66,6**	**63,7**	**58,7**	**68,7**	**61,6**
cSACC	62,0	58,9	62,7	65,2	61,4	53,7	65,0	55,8

TABLE 1 – Performances des différents modèles de détection de parole superposée en phase de développement (AMI *Dev*) et d'évaluation (AMI *Test*). Les scores en gras indiquent les meilleures performances de détection en contexte de parole lointaine. Les seuils de détection sont déterminés en phase de développement puis conservés lors de l'évaluation.

Le système cSACC ne semble pas améliorer la détection de parole superposée par rapport au modèle original SACC. L'apprentissage des poids complexes est probablement délicat pour que les parties réelles et imaginaires soient cohérentes. Cette approche améliore cependant la détection en champ lointain, avec un gain de 14,2% de précision moyenne par rapport au canal unique en évaluation. Les poids obtenus avec les modèles d'auto-attention favorisent également l'interprétabilité en permettant de connaître les directions favorisées lors de la détection. Des travaux d'analyse seront menés à l'avenir sur ces directions.

5 Conclusions et perspectives

Les travaux présentés portent sur l'utilisation de méthodes de combinaison auto-attentives des canaux d'une antenne de microphones pour la détection de parole superposée. Ce dispositif est utilisé pour le traitement de la parole distante. La méthode de combinaison développée par Gong *et al.* (2021) a été appliquée pour la tâche de détection de parole superposée distante. Cette méthode a également été étendue dans le domaine complexe afin d'utiliser toute l'information fournie par la TFCT. Les résultats montrent que les techniques de combinaison auto-attentives des canaux permettent d'obtenir des performances de détection proches de celles obtenues avec des microphones-casques utilisés par chacun des locuteurs. La formulation dans le domaine complexe ne permet pas de gain additionnel, le modèle n'arrivant probablement pas à apprendre des poids d'attention complexes cohérents. Elle permet cependant de meilleures performances que le traitement d'un canal unique de l'antenne.

Les prochains travaux porteront sur l'intégration d'une nouvelle architecture pour la modélisation de séquence. Les résultats obtenus par Cornell *et al.* (2020, 2022) montrent que l'utilisation de Transformers (Vaswani *et al.*, 2017) ou de réseaux convolutifs temporels (TCN) (Bai *et al.*, 2018) apporte un gain en performances par rapport aux BLSTM. Des travaux seront également menés afin de rendre la méthode indépendante du nombre de canaux (Luo *et al.*, 2020). Le modèle SACC pourrait également être étendu en rendant les poids dépendants de la fréquence. L'utilisation de plusieurs têtes d'auto-attention sera également étudiée pour permettre au modèle de se focaliser dans des directions spatiales différentes simultanément.

Références

BAI S., KOLTER J. Z. & KOLTUN V. (2018). An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv :1803.01271 [cs]*. arXiv : 1803.01271.

BENESTY J., CHEN J. & HUANG Y. (2008). *Microphone Array Signal Processing*. Springer.

BOAKYE K., VINYALS O. & FRIEDLAND G. (2011). Improved overlapped speech handling for speaker diarization. In *Interspeech 2011*, p. 941–944.

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. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7124–7128.

BULLOCK L., BREDIN H. & GARCIA-PERERA L. P. (2020). Overlap-Aware Diarization : Resegmentation Using Neural End-to-End Overlapped Speech Detection. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 7114–7118.

CORNELL S., OMOLOGO M., SQUARTINI S. & VINCENT E. (2020). Detecting and Counting Overlapping Speakers in Distant Speech Scenarios. In *Interspeech 2020*, p. 3107–3111.

CORNELL S., OMOLOGO M., SQUARTINI S. & VINCENT E. (2022). Overlapped Speech Detection and speaker counting using distant microphone arrays. *Computer Speech & Language*, **72**, 101306.

GARCIA PERERA L. P., VILLALBA J., BREDIN H., DU J., CASTAN D., CRISTIA A., BULLOCK L., GUO L., OKABE K., NIDADAVOLU P. S., KATARIA S., CHEN S., GALMANT L., LAVECHIN M., SUN L., GILL M.-P., BEN-YAIR B., ABDOLI S., WANG X., BOUAZIZ W., TITEUX H., DUPOUX E., LEE K. A. & DEHAK N. (2020). Speaker Detection in the Wild : Lessons Learned from JSALT 2019. In *The Speaker and Language Recognition Workshop (Odyssey 2020)*, p. 415–422.

GEIGER J. T., EYBEN F., SCHULLER B. & RIGOLL G. (2013). Detecting overlapping speech with long short-term memory recurrent neural networks. In *Proc. Interspeech 2013*, p. 1668–1672.

GONG R., QUILLEN C., SHARMA D., GODERRE A., LAÍNEZ J. & MILANOVIĆ L. (2021). Self-Attention Channel Combinator Frontend for End-to-End Multichannel Far-Field Speech Recognition. In *Interspeech 2021*, p. 3840–3844.

LANDINI F., PROFANT J., DIEZ M. & BURGET L. (2022). Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization : Theory, implementation and analysis on standard tasks. *Computer Speech & Language*, **71**, 101254.

LUO Y., CHEN Z., MESGARANI N. & YOSHIOKA T. (2020). End-to-end microphone permutation and number invariant multi-channel speech separation. In *ICASSP*, p. 6394–6398.

MCCOWAN I., LATHOUD G., LINCOLN M., LISOWSKA A., POST W., REIDSMA D. & WELLNER P. (2005). The ami meeting corpus. In *In : Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen : Noldus Information Technology*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER L. & POLOSUKHIN I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, p. 6000–6010.

ZHENG S., ZHANG S., HUANG W., CHEN Q., SUO H., LEI M., FENG J. & YAN Z. (2021). BeamTransformer : Microphone Array-based Overlapping Speech Detection. *arXiv :2109.04049 [cs, eess]*. arXiv : 2109.04049.