Traitement multi-microphone pour la segmentation automatique de la parole en réunion

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

16ème Congrès Français d'Acoustique
11-15 Avril 2022, Marseille

Traitement Multi-Microphone pour la Segmentation Automatique de la Parole en Réunion

Théo Mariotte [a,b], Anthony Larcher [b], Jean-Hugh Thomas [a], Silvio Montrésor [a]
[a] LAUM - IAGS - UMR 6613 - Le Mans Université, 72085 Le Mans
[b] LIUM - Le Mans Université, 72085 Le Mans,

"Qui a parlé quand ?" C'est la question à laquelle répond la segmentation et le regroupement de locuteurs. Cette tâche de traitement automatique de la parole consiste à identifier les locuteurs et à déterminer les instants où chacun s'exprime dans un enregistrement audio. Les performances des algorithmes de segmentation et de regroupement des locuteurs sont conditionnées par un ensemble de tâches préalables telles que la détection d'activité vocale (Voice Activity Detection, VAD). La généralisation de l'utilisation des réseaux de neurones profonds et de l'apprentissage automatique a permis d'améliorer les performances des modèles au cours des dernières années. La qualité des résultats tend cependant à se dégrader en conditions acoustiques difficiles (faible rapport signal-à-bruit, réverbération...). L'utilisation d'antennes de microphones est un axe pour permettre l'amélioration des performances dans ce contexte. Ces dispositifs, composés de plusieurs capteurs placés à différentes positions, permettent l'acquisition d'informations spatiales sur le champ acoustique. Des méthodes de traitement des signaux telles que la formation de voies permettent de combiner les canaux afin de filtrer le signal dans une certaine direction spatiale. Ces approches requièrent cependant la localisation explicite des locuteurs, parfois délicate à estimer. Récemment introduits dans la littérature, les mécanismes d'attention permettent aux modèles neuronaux de se focaliser automatiquement sur une partie des données d'entrée. Les travaux présentés s'intéressent à l'utilisation de ce type de mécanisme pour pondérer et combiner automatiquement les signaux issus de chaque microphone. Les expériences sont menées sur les données du corpus AMI, enregistrées au cours de réunions en conditions réelles. Deux méthodes de combinaison des canaux sont mises en œuvre pour la tâche de VAD. Les poids de combinaison des canaux sont également analysés, montrant que le modèle localise intrinsèquement le locuteur.

1 Introduction

Le développement de systèmes de traitement automatique de la parole spontanée robustes reste aujourd'hui un défi. La segmentation et le regroupement de locuteurs (également nommée diarisation), [1, 2] en font partie. Cette tâche consiste à déterminer "qui a parlé quand ?" dans un signal de parole. Il est alors nécessaire de déterminer les frontières temporelles entre les locuteurs afin de construire des segments, puis de regrouper ces derniers par locuteur.

La détection d'activité vocale (Voice Activity Detection, VAD) consiste à déterminer les segments temporels contenant de la parole. Il s'agit de la première étape de segmentation dans un système de diarisation. Historiquement, la détection de parole était réalisée à l'aide de techniques de traitement des signaux [3,4]. Récemment, la généralisation des réseaux de neurones et de l'apprentissage automatique a permis d'améliorer les performances des systèmes de VAD. Les réseaux de neurones récurrents, en particulier les LSTM (Long Short-Term Memory), sont utilisés dans de nombreux systèmes de l'état de l'art [5, 6].

La majorité des approches développées pour la détection de parole sont cependant destinées à des applications en champ proche (ex : émission télévisée, livres audio...). Dans un contexte de parole spontanée, le système de captation est souvent placé à distance des locuteurs. Cela présente des intérêts pratiques, en évitant aux participants de s'équiper d'un microphone individuel. La parole lointaine tend cependant à dégrader les performances des systèmes de traitement de la parole, le signal de parole étant sujet à un rapport signal-à-bruit (RSB) plus faible [6]. L'utilisation d'un ensemble de microphones (antennes de capteurs) permet d'acquérir des signaux multicanaux, contenant intrinsèquement des informations sur la répartition spatiale du champ acoustique. Les données multicanaux permettent l'utilisation d'algorithmes d'amélioration des signaux [7] ou de filtrage spatial [8] afin d'en augmenter le RSB. L'utilisation de signaux multicanaux peut donc permettre d'améliorer les performances de VAD. Récemment, Cornell et al. [6] ont proposé l'utilisation de caractéristiques spatiales pour la détection d'activité vocale et de parole superposée. L'utilisation de ces dernières permet de rendre la détection plus robuste, mais nécessite l'extraction manuelle de caractéristiques.

D'autre part, Gong et al. [9] ont proposé une méthode de combinaison des canaux basée sur les mécanismes d'auto-attention (Self-Attention Channel Combinator, SACC). Cette approche permet d'estimer automatiquement les poids appliqués à chaque canal avant de les combiner. Les auteurs montrent que la méthode permet d'améliorer les performances de transcription de la parole lointaine.

À partir de l'état de l'art ci-dessus, nous proposons deux méthodes de détection de parole distante à l'aide d'une antenne de microphones. Le modèle proposé intègre l'approche SACC [9] comme pré-traitement afin d'extraire des caractéristiques du signal. Un réseau de neurones récurrent LSTM est utilisé pour modéliser ces caractéristiques et détecter la présence de parole. Les poids de combinaison appris par le système SACC sont cependant délicats à interpréter. Cette approche est donc étendue dans l'espace des complexes (cSACC) afin d'apprendre des poids plus proches des algorithmes de formation de voies [8]. Les résultats montrent que la combinaison auto-attentive des canaux permet d'atteindre des performances similaires aux algorithmes de formation de voies. Une analyse des poids du système cSACC est également menée et montre qu'ils pourraient permettre la localisation du locuteur actif.

L'article est organisé comme suit. En Section 2 le principe de la VAD et l'architecture du modèle sont présentés. La Section 3 présente les approches SACC et cSACC. Les expériences et les résultats sont présentés en sections 4 et 5 avant analyse détaillée du modèle cSACC en section 6.

2 Détection d'activité vocale

2.1 Principe

La tâche de VAD consiste à détecter la présence de parole dans un signal. Le problème est formulé comme une tâche de classification binaire supervisée. L'apprentissage automatique est utilisé.

Soit un modèle paramétrique $\hat{\boldsymbol{y}} = f(\boldsymbol{X};\theta)$, avec θ les paramètres du modèle, prenant en entrée une séquence de caractéristiques $\boldsymbol{X} \in \mathbb{R}^{F\times N}$ extraites d'un signal temporel. F représente le nombre de caractéristiques extraites et N le nombre de trames de la séquence. Pour chaque entrée, le modèle retourne une séquence $\hat{\boldsymbol{y}} \in \mathbb{R}^{C\times N}$, où C désigne le nombre de classes. La représentation $\hat{\boldsymbol{y}} = [\hat{y}_1, ...\hat{y}_n, ...\hat{y}_N]$ contient la probabilité pour chaque trame d'appartenir à chacune des classes. Dans le cas d'une classification binaire ($C = 2$), une trame n de $\hat{\boldsymbol{y}}$ est associée à $\hat{y}_n = \{p(c = 0|\boldsymbol{X}_n), p(c = 1|\boldsymbol{X}_n)\}$.
À chaque séquence $\hat{\boldsymbol{y}}$ sont associées des étiquettes $\boldsymbol{y} = [y_1, ...y_n, ...y_N] \in \mathbb{R}^{1\times N}$ alignées aux trames de la séquence d'entrée et définies telles que $y_n \in \{0, 1\}$; $y_n = 1$ indique la présence de parole. 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 avec le pré-traitement SACC [9]. Les trois étapes principales du modèle sont présentées ci-dessous.

Extraction de caractéristiques : deux approches différentes sont étudiées pour extraire les caractéristiques du signal d'entrée $x \in \mathbb{R}^{M\times T}$. M représente le nombre de microphones et T le nombre d'échantillons temporels. Pour les signaux monocanal ($M = 1$), les coefficients cepstraux à échelle Mel (MFCC) sont utilisés comme caractéristiques. En pratique, 20 MFCC sont extraits ainsi que leurs dérivées sur des fenêtres de 25 ms avec un pas de 10 ms. Dans le cas des signaux multicanaux ($M > 1$), trois méthodes sont utilisées. La formation de voies Minimum Variance Distortion-less (MVDR) [10] est comparée aux approches auto-attentives SACC [9] et cSACC. Les détails de ces méthodes sont présentés en Section 3. Chaque méthode retourne une séquence $\boldsymbol{X} = [\boldsymbol{X}_1, ...\boldsymbol{X}_n, ...\boldsymbol{X}_T]$ où $\boldsymbol{X}_n$ représente les caractéristiques extraites pour une trame n.

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 (BLSTM) sont utilisées de façon similaire aux travaux en [5]. Le modèle utilisé est composé de deux couches BLSTM de $K = 32$ cellules chacune. La moitié d'entre-elles modélise la séquence en avant et l'autre moitié à l'envers, en fonction du temps. Une fonction d'activation tanh est appliquée à la sortie de chacune d'elles. Une nouvelle représentation $\boldsymbol{s} = \{\boldsymbol{s}_1, ..., \boldsymbol{s}_K\}$ de la séquence $\boldsymbol{X}$ est obtenue en sortie des BLSTM.

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 (32;16), (16;16) et (16;2). Chaque couche linéaire est suivie d'une fonction d'activation tanh à l'exception de la dernière où une fonction softmax est appliquée. Cela permet d'obtenir la séquence des probabilités $\hat{\boldsymbol{y}}$ pour chaque trame d'appartenir à chacune des classes.

2.3 Inférence et évaluation

L'inférence est réalisée en appliquant une fenêtre glissante avec un certain recouvrement au signal d'entrée. Pour chaque fenêtre, la présence de parole superposée est estimée. Seule la sortie associée à la classe positive $p(\boldsymbol{c} = 1|\boldsymbol{x})$ est conservée. Elle contient la probabilité de chaque trame de contenir de la parole. Les prédictions obtenues sur les trames se recouvrant entre deux fenêtres successives sont combinées.

Deux seuils de détection th_{in} et th_{out} sont ensuite appliqués afin d'obtenir une séquence binaire. Le premier définit le passage de la classe négative vers la classe positive. Le second permet l'opération inverse. L'application des seuils de détection est schématisée en Figure 1. La séquence binaire ainsi obtenue peut être comparée à la référence issue des annotations manuelles.

3 Extraction de caractéristiques multicanal

3.1 Motivations

Lorsque plusieurs microphones sont utilisés pour acquérir des signaux, ces derniers intègrent intrinsèquement des informations spatiales. Celles-ci peuvent être utilisées explicitement pour traiter les signaux [7, 8, 10, 11]. Ces approches nécessitent cependant des hypothèses fortes (ex : bruit additif, algèbre linéaire...) ou l'extraction explicite d'information telle que la position du locuteur. L'information extraite n'est donc pas optimale pour la tâche de parole réalisée. Les approches neuronales de traitement du signal permettent, via l'apprentissage automatique supervisé, de réaliser un traitement optimal pour la tâche visée. Les deux sous-sections suivantes présentent donc deux méthodes de fusion des canaux issus d'une antenne de microphones. Ces deux approches consistent à pondérer les canaux puis à les sommer entre eux afin d'obtenir une nouvelle représentation dépendante de l'information spatiale. Ces deux approches sont implémentées dans le domaine temps-fréquence (TFCT).

3.2 Combinaison Auto-Attentive des Canaux

Récemment, Gong et al. [9] ont proposé la méthode SACC pour combiner les canaux issus des microphones d'une antenne. Celle-ci est basée sur les mécanismes d'auto-

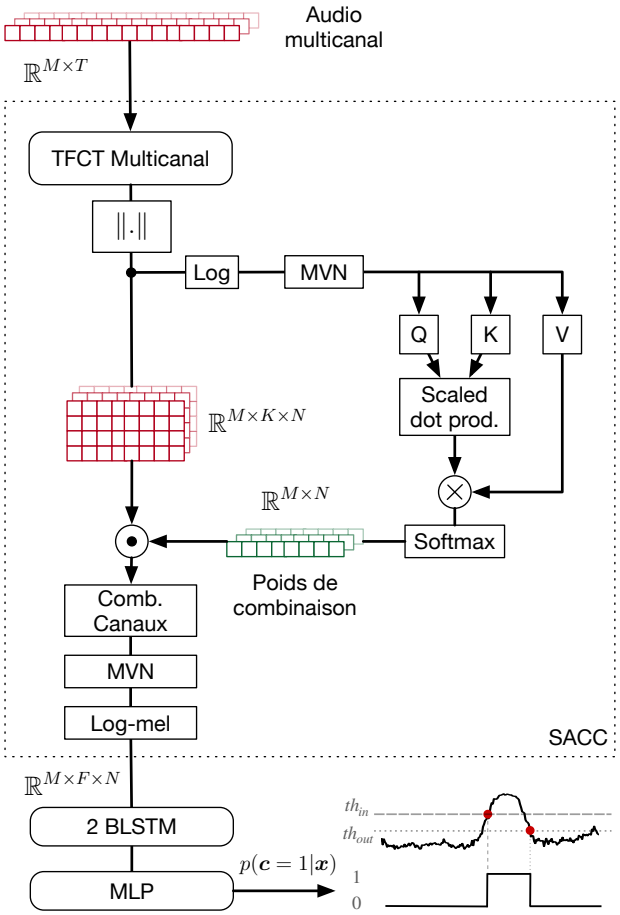


FIGURE 1 – Architecture utilisée pour la détection de parole. Le modèle peut prendre en entrée un signal issu d'un ou plusieurs microphones en fonction du cas de figure considéré. La phase de détection est également représentée : seule la probabilité de la classe positive $p(c = 1|\boldsymbol{x})$ est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

attention [12] permettant de porter l'attention sur les canaux utiles à chaque trame.

La méthode SACC consiste à estimer les poids $\boldsymbol{w}$ à appliquer à chaque canal de l'antenne en portant l'attention sur le module de la TFCT $\|\boldsymbol{Y}\| \in \mathbb{R}^{M\times K\times N}$, où K représente le nombre de fréquences dans la STFT. Pour cela, trois vecteurs sont estimés : la *requête* $\boldsymbol{Q} = q(\|\boldsymbol{Y}\|)$, la *clef* $\boldsymbol{K} = k(\|\boldsymbol{Y}\|)$ et la *valeur* $\boldsymbol{V} = v(\|\boldsymbol{Y}\|)$. Les fonctions q, k et v sont implémentées par des couches neuronales linéaires. Les poids sont déterminés par la relation suivante :

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

Avec d la dimension des couches linéaires q et k. Le module de la TFCT $\|\boldsymbol{Y}\|$ est pondéré par les poids $\boldsymbol{w}$ et les canaux sont combinés :

$$\boldsymbol{Y}^{\boldsymbol{att}} = \sum_c \text{softmax}(\boldsymbol{w}) \odot \|\boldsymbol{Y}\|, \tag{2}$$

avec $\odot$ le produit terme-à-terme entre deux vecteurs. La fonction softmax permet de garantir $w_i \in [0, 1]$. Une normalisation de la moyenne et la variance (MVN) est appliquée avant l'estimation des poids et l'entrée dans le modèle de détection. 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 [9]. Le spectrogramme $\boldsymbol{Y}^{\boldsymbol{att}}$ est converti en échelle Mel à l'aide de $F = 64$ filtres afin d'obtenir la séquence de caractéristiques $\boldsymbol{X}$.

3.3 Combinaison des canaux dans le domaine complexe

La méthode de combinaison des canaux proposée en [9] calcule des poids uniquement à partir du module de la TFCT. L'information de la phase est donc perdue dès le début du traitement. De plus, les algorithmes de formation de voies traditionnels [8] utilisent des poids définis dans l'espace des complexes. Il semble donc pertinent que les poids appliqués à la TFCT appartiennent à cet espace afin de ne pas perdre d'information lors de la pondération. La méthode SACC est donc étendue dans l'espace des complexes (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 :

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

Les caractéristiques $\boldsymbol{X}$ à partir de $\boldsymbol{Y}^{att}$ sont extraites en appliquant $F = 64$ filtres Mel au module de la TFCT pondérée.

4 Etude expérimentale

4.1 Corpus AMI

Les expériences de détection d'activité vocale sont menées sur le corpus AMI [13] 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. Le corpus est divisé en trois sous-ensembles *Train*, *Dev* et *Eval* contenant respectivement environ 80 h, 10 h et 10 h de données annotées. Le contenu des sous-ensembles suit le protocole proposé par [2], garantissant des locuteurs différents au sein de chacun d'entre eux.

4.2 Apprentissage et évaluation

Pour l'apprentissage, des séquences d'une durée $L_s = 2$ s sont tirées aléatoirement dans les données de *Train*. Aucune technique d'augmentation de données n'est utilisée afin

de rendre les résultats comparables entre les différentes approches. En effet, les méthodes peuvent différer en fonction des cas monocanal ou multicanal. De plus, l'ajout de réverbération ou de masquage temps-fréquence n'a pas permis d'améliorer les performances.

Les performances de classification des modèles sont évaluées à l'aide du F1-score. La précision et le rappel sont également reportés [5]. La première informe sur le nombre de trames correctement classées sur le nombre total de détectées. Le rappel évalue le nombre de trames contenant de la parole détectés sur le nombre total contenu dans les données. Une précision et un rappel élevés signifient que le modèle est précis et fiable.

En phase d'évaluation, les performances de chaque modèle sont évaluées sur les données de développement pour différents seuils de détection. Les seuils permettant les meilleurs scores sont conservés. Les systèmes de VAD sont ensuite évalués sur les données d'évaluation à l'aide des seuils préalablement sélectionnés.

4.3 Modèles de référence

Trois modèles de référence sont entraînés pour évaluer les performances des approches SACC et cSACC. Le premier est un modèle monocanal, entraîné sur des données du corpus AMI enregistrées en champ proche (*headset-mix*). Elles permettent d'avoir une référence de détection en conditions acoustiques favorables. Le deuxième modèle est entraîné sur le signal issu du premier microphone de l'antenne. Cela correspond au cas du microphone unique distant (Single Distant Microphone, SDM), l'approche la plus simple pour acquérir des données de parole distante. Il permet d'obtenir une référence en champ lointain. La troisième référence consiste à combiner les canaux à l'aide de l'algorithme de formation de voies MVDR (Minimum Variance Distortion-less) [10]. L'implémentation proposée dans la librairie Torchaudio [1] est utilisée. Notre modèle étant entraîné sur des données réelles, les masques temps-fréquence du signal et du bruit doivent être estimés. Pour cela, la méthode [14] est utilisée pour calculer les masques à partir du rapport cohérence-à-diffusion. L'estimateur indépendant de la position de la source, sous hypothèse de bruit diffus, est utilisé ([14], eq. 25).

5 Résultats

5.1 Évaluation des différents modèles

Les résultats obtenus en VAD à l'aide de chaque système sont présentés dans le tableau 1. L'utilisation d'un microphone distant unique entraîne une dégradation de la détection lié à un manque de précision du modèle. Les résultats montrent que l'utilisation de méthodes de combinaison de canaux auto-attentives permet un gain de performances d'environ 5% sur le F1-score par rapport au microphone distant. Les performances obtenues à l'aide des méthodes SACC et cSACC sont similaires et s'approchent de celles obtenues en conditions de champ proche. Les approches auto-attentives permettent également des performances de détection similaires à la formation de voies MVDR sans nécessiter ni localisation des locuteurs ni masquage temps-fréquence.

	F1-score (%)		Précision (%)		Rappel (%)	
AMI	Dev	Eval	Dev	Eval	Dev	Eval
Ref MFCC	96,3	96,3	95,4	96,2	97,1	96,4
SDM MFCC	93,5	93,2	89,8	89,4	97,5	97,3
MVDR	96,0	95,9	95,3	96,7	96,8	95,2
SACC	96,2	**96,3**	95,4	96,2	97,4	96,3
cSACC	**96,3**	96,1	95,4	96,2	97,2	96,1

Tableau 1 – Performances de détection pour la tâche de VAD avec l'architecture BLSTM sur les données de développement et d'évaluation du corpus AMI pour chaque configuration. Les scores en gras indiquent le meilleur modèle selon le F1-score.

5.2 Comparaison des modèles en champ lointain

Les figures 2a et 2b présentent la sortie des différents modèles de VAD en parole distante évalués (SDM, MVDR, SACC et cSACC). Les prédictions sont comparées à la référence annotée. Elles sont obtenues sur des segments de $L_s = 4$ s issus du sous-ensemble d'évaluation du corpus AMI. Ces figures présentent les sorties brutes du modèle, aucun seuil de détection n'a donc été appliqué ici.

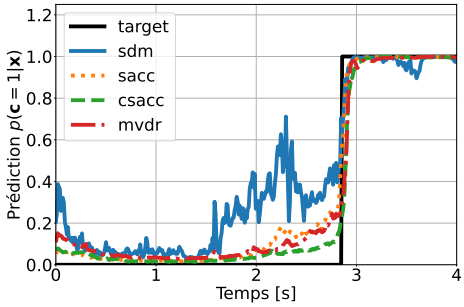
La figure 2a montre que les quatre pré-traitements permettent une détection de la parole robuste. Le microphone distant unique (SDM) est cependant moins robuste dans ce cas de figure. La prédiction de ce dernier est également moins stable et présente plus de variations. Cela peut mener à des erreurs de détections lorsque les seuils sont appliqués (*cf* Tableau 1). La figure 2b présente un cas où aucun des modèles n'identifie le silence entre les deux segments de parole. Cela peut être source d'erreurs de segmentation au sein d'un modèle de diarisation.

6 Analyse

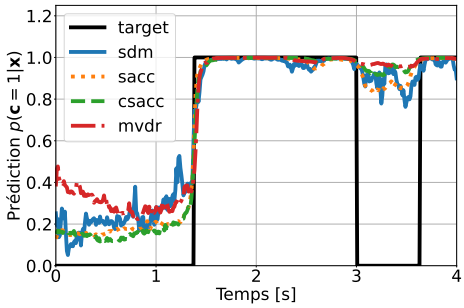
6.1 Réponse spatiale

Les poids de combinaison des canaux appris par le modèle cSACC appartiennent à l'ensemble des complexes. Ils peuvent alors être analysés de façon similaire aux poids issus des algorithmes de formation de voies. La réponse spatiale permet de calculer la réponse d'un filtre spatiale pour une géométrie d'antenne donnée [11]. Elle consiste à déterminer la transformation appliquée par le filtre à une onde plane provenant d'une direction azimutale θ donnée. Le module de la réponse spatiale informe sur le gain appliqué par le filtre dans chaque direction à une pulsation ω donnée.

1. https://github.com/pytorch/audio



(a)



(b)

FIGURE 2 – Exemples de prédictions réalisées par chaque modèle en champ lointain. La référence (target) est issue des annotations manuelles du corpus AMI. Les segments (a) et (b) sont issus du même fichier à des instants différents.

Soit une antenne circulaire uniforme de rayon r et dont la répartition angulaire des microphones est définie par $\psi_m = 2\pi(m-1)/M$. La réponse spatiale d'un filtre défini par les poids $\boldsymbol{w}_n(\omega, \theta_s$ et dirigé dans une direction θ_s est définie par :

$$\mathcal{B}_n[\boldsymbol{w}_n(\omega,\theta_s),\theta] = \sum_{m=1}^{M} w^*_{m,n}(\omega,\theta_s) e^{j\bar{\omega}\cos(\theta-\psi_m)}, \qquad (4)$$

avec $\bar{\omega} = \omega r/c$ où $c = 340$ m/s représente la vitesse du son. L'indice n indique la trame considérée, les poids $w_{m,n}$ étant variables au cours du temps dans le cas de la méthode cSACC. La réponse spatiale peut donc être moyennée au cours du temps sur un nombre de trames N_T donné :

$$\hat{\mathcal{B}}(\omega,\theta) = \frac{1}{N_T}\sum_{n=0}^{N_T-1} \mathcal{B}_n[\boldsymbol{w}_n(\omega,\theta_s),\theta]. \qquad (5)$$

Dans le cas de l'approche cSACC, les poids sont définis comme indépendants de la fréquence. De plus, la direction de focalisation θ_s n'est pas connue explicitement, le modèle apprenant les poids à partir des données. Le calcul de la réponse spatiale à partir des poids de combinaison cSACC nous informe donc sur les directions favorisées par le modèle afin de détecter la parole.

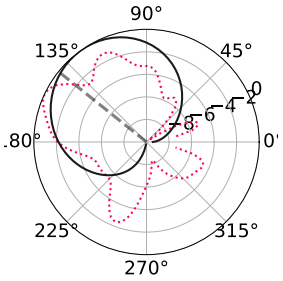
6.2 Simulations

Les annotations du corpus AMI ne contiennent pas la position des locuteurs. Cette donnée manquante ne permet pas d'analyser les poids en fonction de la position précise des locuteurs. Afin d'analyser le modèle dans un environnement contrôlé, une simulation est réalisée à l'aide du jeu de données Librispeech [15]. Ce dernier contient des enregistrements de livres audio. Il s'agit donc de données monocanal faiblement dégradées par le bruit et la réverbération. Pour s'approcher des données réelles distantes, une salle de dimension $L = 5$ m, $l = 4$ m et $h = 3$ m est simulée. Une antenne circulaire uniforme de rayon $r = 10$ cm est simulée. Les réponses impulsionnelles de salle (RIS) sont ensuite calculées entre la position source et chaque microphone à l'aide de l'algorithme de sources-images gpuRIR [16]. La convolution du signal source avec chaque RIS permet de simuler un signal multicanal en champ lointain. Le temps de réverbération à -60 dB est fixé à $T_{60} = 1$ s.

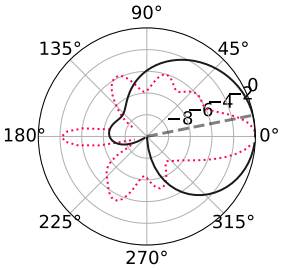
6.3 Analyse du modèle cSACC

Des exemples de réponses spatiales moyennées, calculées à l'aide des équations (4) et (5) sont présentées en figures 3a et 3b. La carte d'énergie obtenue à l'aide de l'algorithme SRP-PHAT [8] sur un cercle de 1 m de rayon est également présentée. L'implémentation proposée dans la librairie Speechbrain [2] est utilisée. Elle permet de comparer la réponse spatiale du modèle cSACC à un algorithme classique de localisation de sources.

La figure 3 montre que le modèle cSACC se focalise en moyenne dans la direction de la source. Cette approche pourrait donc permettre, uniquement en entraînant un modèle de VAD, d'estimer la position des locuteurs.



(a)



(b)

FIGURE 3 – Réponse spatiale moyennée (–) comparée à la carte d'énergie SRP-PHAT calculée sur un cercle de rayon R=1 m (···). La ligne discontinue (- - -) indique la direction du locuteur actif. La réponse spatiale et la carte SRP-PHAT sont normalisées pour permettre une comparaison, les niveaux obtenus pour chaque représentation étant différents. Nombre de trames pour le calcul des moyennes : (a) $N_T = 1358$, (b) $N_T = 1211$.

2. https://speechbrain.github.io/

Le modèle cSACC ne se focalise cependant pas toujours dans la direction du locuteur actif. Il se peut que les réflexions sur les parois et la réverbération incitent le modèle à porter l'attention dans d'autres directions, comme le montre la figure 4. Des expériences pourront être menées afin d'évaluer qualitativement les performances de localisation du modèle cSACC à partir des poids du modèle.

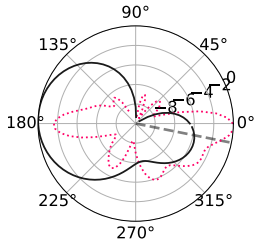


FIGURE 4 – Exemple de réponse spatiale moyennée dirigée dans une direction autre que celle du locuteur actif.

7 Conclusion

La détection d'activité vocale (VAD) est une tâche clef pour le traitement automatique de la parole. Cependant, les performances des algorithmes de détection se dégradent en conditions de parole lointaine. Les travaux présentés montrent que l'utilisation de méthodes de combinaison des canaux basées sur l'auto-attention permet de s'approcher des performances obtenues en champ proche. Les scores de détection s'approchent également de ceux obtenus à l'aide de la formation de voies, sans nécessiter une localisation explicite du locuteur. L'analyse des poids du modèle cSACC montre également que le modèle porte son attention dans la direction du locuteur actif. Des travaux pourront être menés pour localiser le locuteur à partir du modèle de VAD. L'étude des différentes approches sur la tâche complète de diarisation est également envisagée.

Références

[1] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman. The Third DIHARD Diarization Challenge. In *INTERSPEECH*, pages 3570–3574, 2021.

[2] F. Landini, J. Profant, M. Diez, and L. Burget. Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization : Theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71 :101254, January 2022.

[3] S. V. Gerven and F. Xie. A comparative study of speech detection methods. In *Fifth European Conference on Speech Communication and Technology*. Citeseer, 1997.

[4] H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan. Noise robust voice activity detection using features extracted from the time-domain autocorrelation function. In *11th Annual Conference of the ISCA*, pages 3118–3121, 2010.

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

[6] C. Samuele, O. Maurizio, S. Stefano, and E. Vincent. Overlapped speech detection and speaker counting using distant microphone arrays. *Computer Speech & Language*, 72 :101306, March 2022.

[7] E. Vincent, T. Virtanen, and S. Gannot, editors. *Audio Source Separation and Speech Enhancement*. John Wiley & Sons Ltd, September 2018.

[8] J. Benesty, J. Chen, and Y. Huang. *Microphone Array Signal Processing*. Springer, 2008.

[9] R. Gong, C. Quillen, D. Sharma, A. Goderre, J. Laínez, and L. Milanović. Self-Attention Channel Combinator Frontend for End-to-End Multichannel Far-Field Speech Recognition. In *Interspeech 2021*, pages 3840–3844, August 2021.

[10] M. Souden, J. Benesty, and S. Affes. On optimal frequency-domain multichannel linear filtering for noise reduction. *IEEE TASLP*, 18(2) :260–276, 2010.

[11] J. Benesty, J. Chen, and I. Cohen. *Design of Circular Differential Microphone Arrays*, volume 12 of *Springer Topics in Signal Processing*. Springer International Publishing, Cham, 2015.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, 2017.

[13] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, and P. Reidsma, A.and Wellner. The ami meeting corpus. In *In : 5th In- ternational Conference on Methods and Techniques in Behavioral Research*, 2005.

[14] A. Schwarz and W. Kellermann. Coherent-to-Diffuse Power Ratio Estimation for Dereverberation. *IEEE/ACM TASLP*, 23(6) :1006–1018, 2015.

[15] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech : An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015.

[16] D. Diaz-Guerra, A. Miguel, and J. R. Beltran. gpuRIR : A python library for room impulse response simulation with GPU acceleration. *Multimedia Tools and Applications*, 80(4) :5653–5671, February 2021.