# Détection d'activité vocale Multi-flux pour la Diarisation du locuteur

Yannis TEVISSEN[1,2], Jérôme BOUDY[1], Gérard CHOLLET[1], Frédéric PETITPONT[2]

[1] SAMOVAR, Télécom SudParis, Institut Polytechnique de Paris, 91200 Palaiseau, France
[2] Newsbridge, 92100 Boulogne-Billancourt, France

**Résumé** – La diarisation du locuteur, ou la tâche de déterminer « qui parle, quand ? », a récemment connu des avancées majeures, mais la plupart des recherches sont axées sur les représentations vectorielles de la parole et les méthodes de regroupement. Dans cet article, nous étudions l'impact du choix de la détection d'activité vocale sur les performances de diarisation du locuteur. Nous présentons également une nouvelle méthode de détection d'activité vocale multi-flux basée sur une fusion de trois systèmes selon leurs entropies. Celle-ci s'est déjà avérée compétitive lors du challenge de diarisation VoxSRC 2022. Enfin, nous discutons des prochaines étapes pour obtenir des résultats au niveau de l'état de l'art dans le cas général avec cette méthode.

**Abstract** – Speaker diarization, or the task of "who spoke, when?", has recently seen some major breakthroughs but most research focus on speaker embeddings and clustering methods. In this article we study the impact of the choice of the voice activity detection preprocessing on speaker diarization performances. We also present a new multi-stream voice activity method that fuses three systems based on their entropies. This method already has proven to be competitive during the VoxSRC 2022 speaker diarization challenge. Finally, we discuss the next step for achieving state-of-the-art results with this method.

## 1 Introduction

Obtenir la diarisation du locuteur équivaut à répondre à la question « qui parle, quand ? » pour un enregistrement audio donné. La diarisation est un élément clé de presque toutes les applications modernes de traitement de la parole, en particulier celles qui utilisent des algorithmes de reconnaissance vocale pour transcrire des conversations.

Cependant, ces nouvelles approches tirent souvent profit des capacités de calcul grandissantes notamment des processeurs graphiques, souvent onéreux et consommateurs en énergie.

Les approches de la diarisation sont diverses, comme le montrent les auteurs de [1] et [2], mais le principe reste souvent le même. Des segments de parole homogènes sont d'abord isolés dans un fichier audio, puis regroupés afin de déterminer quels segments ont été prononcés par quel locuteur. Au cours des dix dernières années, les performances et la robustesse de la diarisation ont été considérablement améliorées, en partie grâce au développement de réseaux de neurones profonds qui sont désormais capables de traiter de grandes quantités de données et surtout de générer de nouvelles représentations vectorielles robustes de la parole [3]-[5].

Par ailleurs, la plupart des méthodes récentes de diarisation ont un point commun : la détection de l'activité vocale (*voice activity detection* – VAD). Utilisée en prétraitement, elle permet de s'assurer que les algorithmes de diarisation travaillent bien uniquement sur des audios contenant de la parole.

La détection d'activité vocale a l'avantage d'être une des étapes les plus rapides et les moins consommatrices de la diarisation du locuteur. C'est pourquoi, dans le cadre de cet objectif d'amélioration des performances sans une augmentation drastique des ressources nécessaires, nous avons décidé de tester l'impact des performances de différents algorithmes de détection d'activité vocale sur les résultats obtenus via un système de diarisation à l'état de l'art.

Nous expliquons tout d'abord le contexte de cette étude, puis nous proposons notre propre système tirant parti de plusieurs méthodes de VAD sur le principe d'une fusion basée sur l'entropie. Enfin nous discutons des résultats obtenus avec cette nouvelle approche de VAD multi-flux (*multi-stream*) et analysons les prochaines étapes pour améliorer encore ce système.

## 2 Diarisation robuste du locuteur

La reconnaissance vocale, la diarisation et, par conséquent, la détection de l'activité vocale sont désormais utilisées dans des domaines très variés, de la santé [6] aux médias [7].

Lorsque la diarisation du locuteur est appliquée sur des données récoltées en conditions réelles, plusieurs perturbations peuvent subvenir. Elles empêchent les algorithmes les plus performants de traiter les voix avec précision. Ces altérations peuvent être des bruits de fond, des rires, des applaudissements, mais aussi des chevauchements entre locuteurs.

À l'origine, les techniques de détection d'activité vocale fonctionnaient grâce à des seuils appliqués à l'énergie du signal vocal [8]. Cette méthode, certes rapide et peu consommatrice, manque de robustesse, notamment lorsqu'elle est exposée à des niveaux variables de bruit.

Plus récemment, avec l'essor de l'apprentissage profond, de nouvelles techniques de VAD sont apparues [9]–[11]. Elles s'appuient sur des réseaux neuronaux récurrents et convolutifs, et prennent souvent comme paramètres d'entrée les coefficients cepstraux de

fréquence Mel [12], particulièrement adaptés à la représentation de la voix humaine.

Plusieurs concours sont régulièrement organisés pour comparer les meilleurs systèmes de diarisation du locuteur. L'un des derniers en date, le challenge VoxSRC 2022 [13] a mis en lumière plusieurs méthodes qui mettaient l'accent sur l'utilisation en parallèle de plusieurs VADs. Parmi elles, la méthode arrivée première lors du challenge [14] et notre méthode [15] de détection d'activité vocale multi-flux (*multi-stream voice activity detection* – MSVAD) que nous présentons formellement ici.

## 3 Système proposé

Dans cette section, nous décrivons les différents composants que nous avons utilisés pour notre étude de l'impact de la détection d'activité vocale (VAD) sur les performances de la diarisation du locuteur. Nous présentons également notre architecture de détection d'activité vocale multi-flux et le protocole de décision basé sur l'entropie choisi.

### 3.1 Méthode de diarisation

Après une évaluation préliminaire, nous avons sélectionné l'algorithme de diarisation VBx présenté en [16] et utilisant un extracteur de x-vecteurs [4] entraîné sur VoxCeleb 1 & 2 [17], suivi d'un regroupement basé sur des modèles de Markov cachés (HMM) bayésiens. L'extraction de x-vecteurs est basée sur 64 bandes de filtres Mel alimentant un réseau ResNet pour créer des représentations vectorielles de la parole de 256 valeurs discriminantes au niveau du locuteur. Le HMM utilisé est initialisé grâce à un premier regroupement hiérarchique agglomératif.

Pour chaque méthode de VAD, les seuils d'activation permettant d'extraire des segments de parole ont été optimisés sur la base de données de développement VoxConverse.

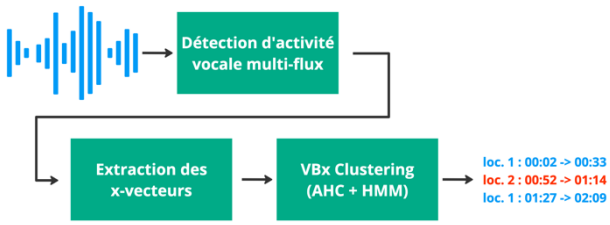


Figure 1 : Schéma d'ensemble du système de diarisation du locuteur proposé

#### 3.1.1 Détection d'activité vocale par calcul de seuil d'énergie

Pour notre expérience de référence, nous avons utilisé une détection d'activité vocale basée sur l'énergie avec un seuil d'activation. Cette méthode a été choisie en premier lieu car elle était recommandée par les auteurs de la méthode VBx dans [19]. C'est d'ailleurs la méthode incluse dans l'implémentation *open-source*[1] de cet algorithme de diarisation.

#### 3.1.2 pyannote

Cette approche est purement basée sur des réseaux de neurones, car elle repose sur l'architecture PyanNet décrite dans [9]. Les MFCCs sont utilisés comme caractéristiques d'entrée pour un classificateur binaire. Pour cette expérience, le modèle pyannote pré-entraîné sur le corpus AMI [20] a été utilisé.

#### 3.1.3 VAD par détection de son (WSSED)

Pour obtenir une détection d'activité vocale plus robuste, les auteurs de [10] ont proposé une approche basée sur un réseau de neurones conçu avec une combinaison de couches de convolution et récurrentes. Cette méthode est basée sur un schéma d'entraînement faiblement supervisé (*weakly supervised sound event detection* – WSSED). Entraînée sur le jeu de données Audioset, cette approche apprend à détecter 517 classes de sons différents parmi lesquelles la classe parole. La grande variété de classes prédites a été radicalement réduite pour obtenir un classificateur binaire *Parole/Non-parole*.

#### 3.1.4 speechbrain

La troisième méthode de VAD que nous essayons de fusionner est celle proposée par l'outil *open-source* speechbrain [11].

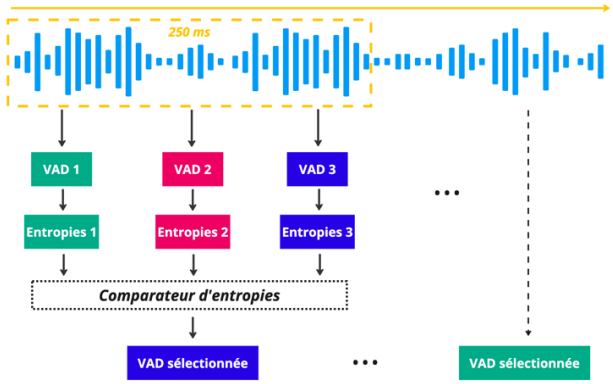


Figure 2 : Schéma du fonctionnement de la détection d'activité vocale multi-flux

### 3.2 Données d'évaluation

La base de données VoxConverse [18] a été également choisi pour les évaluations. C'est un ensemble de données très adapté mais également difficile pour la diarisation en situation réelle. Il contient des médias très variés, allant des conférences de presse aux interviews en studio, avec différents rapports signal sur bruit.

### 3.3 Approche multi-flux proposée

Dans cette section, comme illustré dans la figure 2, nous proposons une approche de détection d'activité vocale multi-flux (*multi-stream voice activity detection*

[1] https://github.com/BUTSpeechFIT/VBx

– MSVAD). Nous utilisons l'entropie comme critère pour choisir automatiquement entre les trois méthodes de VAD. L'objectif est d'utiliser dans chaque contexte la meilleure approche de détection d'activité vocale sur le plan des performances.

Les algorithmes pyannote, WSSED et speechbrain ont été sélectionnés et exécutés en parallèle. Pour chacun, l'entropie locale $h_{k,i}$ du classificateur binaire a été calculée avec la formule suivante (1).

$$h_{k,i} = -P(Speech|o_{k,i})\, log_2\, P(Speech|o_{k,i}) - P(Non\ Speech|o_{k,i})\, log_2\, P(Non\ Speech|o_{k,i}) \quad (1)$$

Avec *i* indice de la fenêtre temporelle, et *k* l'indice du classifieur pour l'observation notée $o_{k,i}$.

On choisit alors, pour chaque fenêtre de 250 ms, l'approche avec l'entropie locale minimale.

## 4 Résultats

Les résultats présentés pour illustrer les performances de la diarisation du locuteur sont ceux relatifs au DER (*Diarization Error Rate*). Pour le calculer, on évalue les durées pendant lesquelles on n'a pas su détecter la parole (PM), les durées où de la parole a été détectée à tort (FA) et enfin les durées où le système s'est trompé de locuteur actif (CF). On applique ensuite la formule suivante pour trouver le DER :

$$DER = \frac{PM+FA+CF}{Durée\ totale\ de\ parole} \quad (2)$$

On s'accorde ici une tolérance de 250ms par rapport à la vérité terrain.

Tab 1 : Résultats détaillés de la diarisation sur le jeu de test de la base VoxConverse

| VAD utilisée | DER | PM | FA | CF |
|---|---|---|---|---|
| Énergie | 22.58 | 10.34 | 7.30 | 4.94 |
| WSSED | 9.76 | 3.78 | 2.30 | 3.68 |
| speechbrain | 9.94 | **2.43** | 3.74 | 3.76 |
| pyannote | **6.66** | 3.09 | **0.79** | **2.78** |
| MSVAD | 7.76 | 2.51 | 1.88 | 3.36 |

La diarisation n'utilisant que la VAD proposée dans l'outil pyannote étant généralement meilleure, on choisit d'étudier les cas où la méthode MSVAD surperforme les autres VADs. Afin d'identifier les points forts de notre méthode, on extrait parmi l'ensemble des résultats les 47 fichiers sur 232 pour lesquelles la MSVAD a produit de meilleurs résultats en termes de DER. Sur ce sous-ensemble de fichiers, on compare les résultats des deux meilleures méthodes.

Tab 2 : Résultats détaillés de la diarisation sur un sous-ensemble du jeu de test de la base VoxConverse

| VAD utilisée | DER | PM | FA | CF |
|---|---|---|---|---|
| pyannote | 8.21 | 3.92 | **0.66** | 3.63 |
| MSVAD | **6.20** | **2.30** | 1.50 | **2.40** |

Enfin, toujours sur ce même sous-ensemble, on étudie la capacité des deux meilleures méthodes à identifier le bon nombre de locuteurs s'exprimant dans l'enregistrement. Cette valeur peut s'apparenter dans certains cas à une étude de la justesse algorithmique de la diarisation [21].

Tab 3 : Pourcentages de bonne détection du nombre de locuteurs présents sur un sous-ensemble du jeu de test de la base VoxConverse

| Nombre de locuteurs | 1 | 2 | 3 | 4 | 5 | >5 |
|---|---|---|---|---|---|---|
| Pyannote | 66.7 | 88.9 | 42.9 | 66.7 | 33.3 | 89.5 |
| MSVAD | **66.7** | **100.0** | **71.4** | **100.0** | **66.7** | **94.7** |

## 5 Discussion et Perspectives

Si nous n'avons pas pu, dans le cas général, améliorer les performances de l'état de l'art avec notre système de détection d'activité vocale multi-flux, cette étude met en lumière le fort impact de la VAD sur les résultats de la diarisation.

On note en particulier que les résultats liés à la confusion entre locuteurs sont impactés par le choix de la VAD. Ceci s'explique par le fait qu'en modifiant, même légèrement, les bornes de détection d'un segment de parole, le système de regroupement utilisé peut être amené à prédire la présence d'un nouveau *cluster*.

On constate par ailleurs qu'à l'image de nos résultats au challenge VoxSRC 2022 [15] on parvient, sur un sous-groupe de la base VoxConverse, à améliorer les performances de la diarisation en ne travaillant que sur la VAD, ainsi que sa capacité à correctement identifier le nombre de locuteurs actifs.

Nos futurs travaux se concentreront sur l'étude de ce sous-groupe pour comprendre dans quels cas précis notre méthode multi-flux présente des avantages. On envisage notamment une étude détaillée des niveaux de bruit. Cela permettra d'optimiser le système de fusion multi-flux pour le cas général. Enfin on pourra étudier les avantages en termes de consommation énergétique à utiliser un tel système.

## 6 Conclusion

En conclusion, cet article a permis de confirmer le fort impact du choix de l'algorithme de détection d'activité vocale sur les performances de la diarisation du locuteur en conditions réelles.

Enfin, nous avons présenté avec précision un nouveau système de détection d'activité vocale multi-flux pour la diarisation du locuteur et démontré son intérêt dans certaines conditions.

## 7 Remerciements

Cette recherche a bénéficié d'un soutien financier total de la société Newsbridge (https://newsbridge.io/) que nous tenons à remercier.

## Références


[1] X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, et O. Vinyals, « Speaker Diarization: A Review of Recent Research », IEEE Trans. Audio, Speech Lang. Process., vol. 20, no 2, p. 356-370, 2012.

[2] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, et S. Narayanan, « A Review of Speaker Diarization: Recent Advances with Deep Learning », Comput. Speech Lang., vol. 72, 2022.

[3] E. Variani, X. Lei, E. McDermott, I. L. Moreno, et J. Gonzalez-Dominguez, « Deep neural networks for small footprint text-dependent speaker verification », ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., p. 4052-4056, 2014.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, et S. Khudanpur, « X-Vectors: Robust DNN Embeddings for Speaker Recognition », ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2018-April, p. 5329-5333, 2018.

[5] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, et H. Na, « ECAPA-TDNN embeddings for speaker diarization », Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 4, p. 2528-2532, 2021.

[6] R. Riad et al., « A comparison study on patient-psychologist voice diarization », SLPAT 2022 - 9th Work. Speech Lang. Process. Assist. Technol. Proc. Work., p. 30-36, 2022, doi: 10.18653/v1/2022.slpat-1.4.

[7] D. Charlet, C. Barras, et J. S. Lienard, « Impact of overlapping speech detection on speaker diarization for broadcast news and debates », ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., p. 7707-7711, 2013.

[8] D. K. Freeman, G. Cosier, C. B. Southcott, et I. Boyd, « Voice activity detector for the Pan-European digital cellular mobile telephone service », ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 1, p. 369-372, 1989.

[9] H. Bredin et al., « Pyannote.Audio: Neural Building Blocks for Speaker Diarization », ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, p. 7124-7128, 2020.

[10] H. Dinkel, Y. Chen, M. Wu, et K. Yu, « Voice activity detection in the wild via weakly supervised sound event detection », Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, 2020.

[11] M. Ravanelli et al., « SpeechBrain: A General-Purpose Speech Toolkit », arXiv preprint.

[12] S. B. Davis et P. Mermelstein, « Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences », IEEE Trans. Acoust., p. 357-366, 1980.

[13] A. Brown, J. Huh, J. S. Chung, A. Nagrani, et A. Zisserman, « VoxSRC 2022: The Fourth VoxCeleb Speaker Recognition Challenge », 2023.

[14] Q. Cai, G. Hong, Z. Ye, X. Li, et H. Li, « The Kriston AI System for the VoxCeleb Speaker Recognition Challenge 2022 », p. 2-6, 2022.

[15] Y. Tevissen, J. Boudy, et F. Petitpont, « The Newsbridge-Telecom SudParis VoxCeleb Speaker Recognition Challenge 2022 System Description », p. 2-4, 2022.

[16] M. Diez, L. Burget, S. Wang, J. Rohdin, et H. Černocký, « Bayesian HMM based x-vector clustering for speaker diarization », Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2019-Septe, p. 346-350, 2019.

[17] A. Nagraniy, J. S. Chungy, et A. Zisserman, « VoxCeleb: A large-scale speaker identification dataset », Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2017-Augus, p. 2616-2620, 2017.

[18] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, et A. Zisserman, « Spot the conversation: Speaker diarisation in the wild », Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2020-Octob, p. 299-303, 2020.

[19] F. Landini, J. Profant, M. Diez, et L. Burget, « Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: Theory, implementation and analysis on standard tasks », Comput. Speech Lang., vol. 71, p. 1-39, 2021.

[20] I. McCowan et al., « The AMI meeting corpus », Int'l. Conf. Methods Tech. Behav. Res., 2005.

[21] Y. Tevissen, J. Boudy, G. Chollet, et F. Petitpont, « Towards Measuring and Scoring Speaker Diarization Fairness », arXiv preprint.