Détection d'activité vocale multi-flux pour la diarisation du locuteur

Yannis Tevissen, Jérôme Boudy, Gérard Chollet, Frédéric Petitpont

DÉTECTION D'ACTIVITÉ VOCALE MULTI-FLUX POUR LA DIARISATION DU LOCUTEUR

Auteurs

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

[1] Newsbridge
[2] Institut Polytechnique de Paris, Télécom SudParis

Partenaire

newsbridge

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 orientées vers la creation de nouvelles 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**.

LA DIARISATION DU LOCUTEUR

UNE TÂCHE MODULAIRE

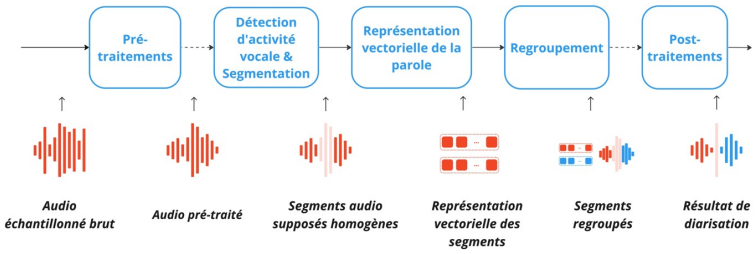

Fig. 1 : Schéma général de la tâche de diarisation du locuteur

LA DÉTECTION D'ACTIVITÉ VOCALE

En raison de la difficulté d'annoter pour la tâche de diarisation, il est également **très difficile de choisir une bonne détection d'activité vocale** (VAD).

Nous présentons ici les résultats d'un système de diarisation du locuteur de l'état de l'art (VBx) en ne faisant varier que sa VAD.

Nous introduisons également une **détection d'activité vocale multi-flux (MSVAD)** qui choisit dynamiquement parmi 3 méthodes de VAD, celle avec l'entropie la plus faible.

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

Les différents systèmes sont évalués sur la base de données **VoxConverse**, composée principalement de contenus télévisés (émissions en studio, meeting politique, interviews, etc.).

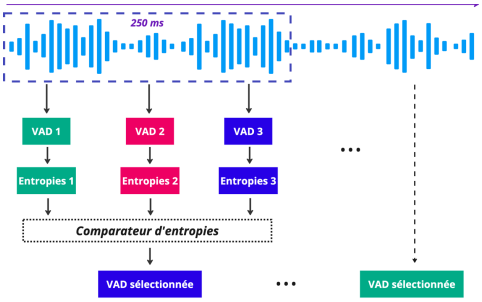

Fig. 2 : Schéma de la détection d'activité vocale multi-flux

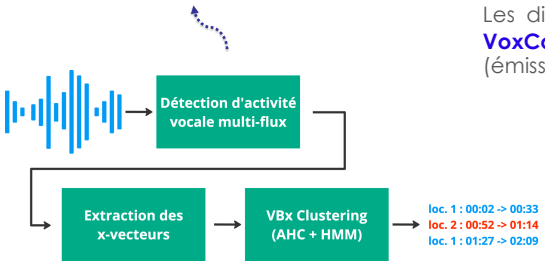

Fig. 3 : Schéma complet du système de diarisation utilisé

Tab. 1: Résultats de diarisation sur VoxConverse (50h)

VAD utilisée	DER	PM	FA	CF
Énergie	22.58	10.34	7.30	4.94
GP-VAD	9.76	3.78	2.30	3.68
speechbrain	9.94	**2.43**	3.74	3.76
pyannote 2.1	**6.66**	3.09	**0.79**	**2.78**
MSVAD	7.76	2.51	1.88	3.36

Tab. 2 & 3 : Comparaison de la MSVAD et de pyannote sur un sous-ensemble de VoxConverse

VAD utilisée	DER	PM	FA	CF
pyannote 2.1	8.21	3.92	**0.66**	3.63
MSVAD	**6.20**	**2.30**	1.50	**2.40**

Nombre de locuteurs	1	2	3	4	5	>5
pyannote 2.1	**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**

RÉSULTATS ET CONCLUSION

Les résultats obtenus montrent l'impact de la VAD sur **toutes les composantes du DER** (parole manquée, fausses alarmes et confusion de locuteurs).

Notre système s'est également montré plus performant que pyannote sur certains enregistrements audios, en particulier lorsqu'il s'agit de déterminer le **nombre exact de locuteurs actifs**.

Une prochaine étape consistera à déterminer précisément sur quelles catégories d'enregistrement la MSVAD est meilleure (domaine, SNR, etc.).

Bibliographie

Park, T. J., Kanda, N., Dimitriadis, D., Han, K. J., Watanabe, S., and Narayanan, S. (2022) A Review of Speaker Diarization: Recent Advances with Deep Learning. *Computer Speech & Language* 72(C).

Landini, F., Profant, J., Diez, M., and 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.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020) pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020 - IEEE International Conference on Acoustics, Speech and Signal Processing*.

Dinkel, H., Chen, Y., Wu, M., and Yu, K. (2020) Voice activity detection in the wild via weakly supervised sound event detection. In *Interspeech 2020*.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., et al. (2021, June 8) SpeechBrain: A General-Purpose Speech Toolkit. https://arxiv.org/abs/2106.04624.

Tevissen, Y., Boudy, J., and Petitpont, F. (2022) The Newsbridge - Telecom SudParis VoxCeleb Speaker Recognition Challenge 2022 System Description. In *VoxCeleb Speaker Recognition Challenge 2022*.

GRETSI 2023
Août 2023

Contact yannis.tevissen@telecom-sudparis.eu **newsbridge.io**