Multi-stream voice activity detection for robust speaker diarization

Yannis Tevissen, Jérôme Boudy, Gérard Chollet

MULTI-STREAM VOICE ACTIVITY DETECTION

FOR ROBUST SPEAKER DIARIZATION

Authors

Yannis TEVISSEN
Jérôme BOUDY
Gérard CHOLLET

Partners

newsbridge

Related works

Y. Tevissen, J. Boudy, F. Petitpont, "**The Newsbridge - Telecom SudParis VoxCeleb Speaker Recognition Challenge 2022 System Description**", VoxSRC Workshop 2022.

[1] H. Bredin et al., "**Pyannote.Audio: Neural Building Blocks for Speaker Diarization**," ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc., vol. 2020-May, pp. 7124–7128, 2020.

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

[3] M. Diez, L. Burget, S. Wang, J. Rohdin, and H. Černocký, "**Bayesian HMM based x-vector clustering for speaker diarization**," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 2019.

VOICE ACTIVITY DETECTION

AS SPEAKER DIARIZATION PRE-PROCESSING

Speaker diarization is the task of determining « **who spoke, when ?** »:

➔ Starts by identifying **parts of the audio where there were speech**.

Several robust voice activity detection (VAD) methods have been introduced in the past years (Bredin2020[1], Chen2020[2]).

➔ **Complementarity** between Pyannote 2.0 VAD (constant good results) and GPVAD (good in noisy environments)

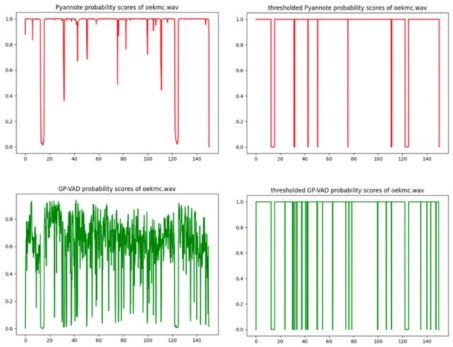

Figure 2. *Voice activity detection raw and thresholded scores obtained with pyannote 2.0 (in red) and GPVAD (in green)*

EXPERIMENTAL RESULTS

Our experiments highlighted the **importance of voice activity detection for speaker diarization**.

➔ Ranked 13th for DER and 5th for JER during VoxSRC 2022 challenge (see Table 2).

Nevertheless, when applied on VoxConverse dataset, made of YouTube videos, MSVAD does not outperform a system made with pyannote 2.0 VAD (see Table 1).

Method	MS	FA	SC	DER	JER
pyannote diarization	2.29	1.31	4.36	7.96	43.38
pyannote VAD + VBx	3.09	0.79	**2.78**	6.66	**29.61**
w/ resegmentation	**1.55**	**1.65**	2.98	**6.18**	29.73
GP-VAD + VBx	3.78	2.30	3.68	9.76	31.82
w/ resegmentation	**1.55**	**1.65**	3.5&	6.88	30.44
Multi-Stream VAD + VBx	3.05	1.41	3.09	7.54	30.22
w/ resegmentation	**1.55**	**1.65**	3.04	6.39	29.80

Table 1. *Results obtained on VoxConverse test set*

newsbridge

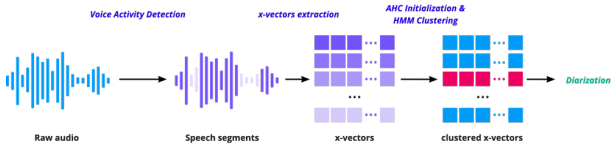

Figure 1. *Speaker diarization overall system description*

ENTROPY DECISION PROTOCOL

FOR MULTI-STREAM VAD

To get the best of both VADs for our speaker diarization system (based on Diez2019[3]) we designed an entropy-based decision protocol.

1. For every 0.5s time window, we compute the entropy of each VAD classifier.
2. We compare the entropies and select dynamically **the classifier with the lowest entropy** (cf. Fig. 3).
3. We apply VBx diarization with the obtained speech segments.

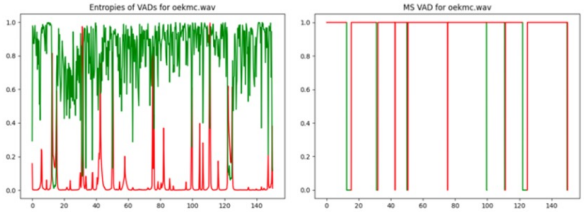

Figure 3. *Local entropies of each VAD classifier (left) and multi-stream output (right)*

PERSPECTIVES

Our upcoming research will focus on determining **under which conditions does MSVAD produce the best results**.

➔ SNR study of both the datasets we used and acoustic domain identification.

Finally we will work on speaker diarization biases with a particular focus on finding biases within the most used algorithms.

Method	DER	JER
Pyannote 2.0 VAD + VBx	8.30	31.14
Pyannote 2.0 VAD + VBx + resegmentation	7.32	30.12
Multi-Stream VAD + VBx + resegmentation	**6.62**	**29.01**

Table 2. *Results obtained on blind test set for VoxSRC 2022*

GDR ISIS - Traitement du signal pour la voix (Action Audio)
Oct. 2022

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