Interspeech 2024
1-5 September 2024, Kos, Greece

Investigating Confidence Estimation Measures for Speaker Diarization

Anurag Chowdhury, Abhinav Misra, Mark C. Fuhs, Monika Woszczyna

Solventum, USA

achowdhury2@solventum.com, amisra2@solventum.com, mark.fuhs@solventum.com, mwoszczyna@solventum.com

Abstract

Speaker diarization systems segment a conversation recording based on the speakers' identity. Such systems can misclassify the speaker of a portion of audio due to a variety of factors, such as speech pattern variation, background noise, and overlapping speech. These errors propagate to, and can adversely affect, downstream systems that rely on the speaker's identity, such as speaker-adapted speech recognition. One of the ways to mitigate these errors is to provide segment-level diarization confidence scores to downstream systems. In this work, we investigate multiple methods for generating diarization confidence scores, including those derived from the original diarization system and those derived from an external model. Our experiments across multiple datasets and diarization systems demonstrate that the most competitive confidence score methods can isolate ∼30% of the diarization errors within segments with the lowest ∼10% of confidence scores.

Index Terms: Speaker Diarization, Confidence Estimation, Speaker Recognition, Speaker Clustering

1. Introduction

Speaker diarization is the task of determining "who spoke when" in speech audio. Traditionally, we perform speaker diarization by segmenting speech audio into short segments and then clustering them by their perceived speaker identity. However, such an approach inherits the challenges of speaker recognition, such as low inter-speaker variability and high intra-speaker variability [1], in addition to the challenges of speaker clustering, such as overlapping and imbalanced amount of speech from an unknown number of speakers [2]. While several techniques have been developed to address these challenges, most can only deal with a small subset of the challenges within their limitations. For example, the recently developed End-to-End (E2E) speaker diarization systems combine speech activity detection (SAD), speaker recognition, and clustering into one E2E system [2] and are adept at diarizing overlapping speech segments. However, their performance worsens with an increasing number of speakers [3]. On the other hand, system combination techniques [4] are gaining popularity to create an ensemble of several diarization methods that exceed the performance of its constituent methods. However, such ensemble systems are error-prone in portions of the speech audios where the constituent systems do not agree on a decision.

One of the main applications of speaker diarization systems is to segment conversational speech audio into short homogenous speech chunks based on the speaker's identity before feeding them to downstream tasks. For example, automatic conversational transcription systems use speaker diarization together with automatic speech recognition (ASR) to generate turn-by-turn speaker-labeled text transcripts of audio conversations [5]. However, long conversations often feature nuisance factors, such as low signal-to-noise ratio, background noise, short vocal responses, and overlapping speech, leading to decreased diarization accuracy.

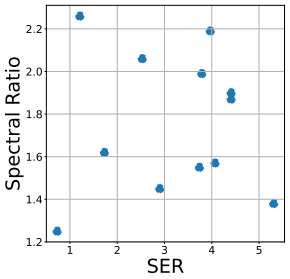


Figure 1: *A plot of spectral ratio (SR) vs. speaker error rate (SER) was computed using ECAPA-TDNN speaker embeddings on the eval set of the AMI dataset. No correlation is observed between the SR and SER, indicating that SR is a poor predictor of diarization confidence.*

These diarization inaccuracies can propagate to and compound in downstream tasks. For example, automatic speech recognition systems can benefit from adapting models to individual speakers [6, 7]; however, while such models improve accuracy on the target speaker, performance degrades when one speaker's model is applied to a different speaker's speech. For segments where the identity of the speaker is unclear, alternate strategies can be employed, such as using a speaker-independent ASR model or selecting from the output of multiple speaker-adapted models. Similarly, semi-supervised adaptation of an ASR model to a particular speaker depends on identifying speech from that speaker with high precision. Spoken language understanding tasks can also be affected. The statement "I'd like to try $\langle medication \rangle$ to see if it helps" in a doctor-patient conversation could imply a suggestion if spoken by the patient, or an actionable item in a plan of care if spoken by the doctor. Actionable items may require additional user confirmation where the speaker identity or transcription is low confidence.

Existing confidence assessment methods, such as in [8], use the spectral ratio of eigenvalues to estimate conversation-level confidence scores; they do not perform confidence assessment at the segment level. In our experiments on the AMI dataset, we evaluated the effectiveness of spectral ratio as a predictor of diarization performance. As shown in Fig. 1, the spectral ratio of ECAPA-TDNN-speaker embeddings [9] does not correlate with the Speaker Error Rate (SER) of the diarization system.

10.21437/Interspeech.2024-1044

Spectral ratio, therefore, cannot be used to estimate diarization confidence.

In this work, we investigate methods for assigning segment-level diarization confidence scores to several different types of diarization systems, including xVector- and ECAPA-TDNN-based systems using spectral clustering, an end-to-end diarization system, and output from a multi-system combination. We consider both white-box and black-box scenarios, where confidence is assigned either based on the model of the original diarization system or using a secondary system, the latter being particularly important in the multi-system case.

To analyze the effectiveness of the proposed confidence assessment method, we study the proportion of incorrectly diarized segments isolated in the low-confidence segments. In our experiments, across multiple diarization systems and datasets, we isolated almost 30% of the diarization errors in the 10% of segments with the lowest confidence. This demonstrates the proposed method's efficacy in identifying incorrect diarization outputs without prior knowledge or access to the diarization system. We then explore the distributions of speaker embeddings and their confidence scores to add insight into confidence score performance.

2. Methods

For all of the diarization systems we consider, conversational audio is first segmented into continuous speech segments by a speech activity detector (SAD). The SAD model comprises five TDNN layers interleaved with two LSTM layers, and speech / non-speech output posteriors are smoothed with median filtering [10]. We then apply one (or combination of) diarization systems to the speech portions to segment them into single-speaker segments and label each segment with a speaker index. Confidence scores are then computed for each segment using each relevant method.

2.1. Confidence Assessment

In this work, we investigate the following different methods for computing the confidence scores. In each case, we rely on a speaker embedding per segment that is generated by the original diarization system's model or by a secondary system. Speaker centroids are the average of the embeddings associated with all of the segments assigned to a speaker.

- *Cosine Similarity Score*: This method computes confidence scores as the mean cosine similarity between a segment's speaker embeddings and the speaker's centroid.
- *Local Confidence Score*: This method uses the cosine similarity of speech segment embeddings to their predicted speaker centroids to compute the initial confidence scores. Next, we drop the speech embeddings two standard deviations away from the centroid for each speaker cluster and use the remaining embeddings in each cluster to re-estimate their centroids. We repeat this step until the centroid stabilizes for each speaker cluster. We then use the cosine similarity of the updated speaker centroids with their cluster members to compute the final confidence scores.
- *Silhouette Score*: This method uses the silhouette score [11], a clustering validation metric to compare embeddings with the predicted speaker and also the closest other speaker centroid:

$$s(x) = \frac{b(x) - a(x)}{max(a(x), b(x))} \quad (1)$$

Here, $s(x)$ is the silhouette score of a speech embedding x. $b(x)$ is the cosine distance of x to the centroid of the closest cluster that it is not a part of and $a(x)$ is the cosine distance of x to the centroid of its own cluster. Silhouette score computation needs at least two clusters; therefore, we revert to using the cosine similarity metric for single-speaker audios (such as dictations of notes) as our confidence measure.
- *Spectral Clustering Score*: A variant of the Cosine Similarity Score using the eigenbasis from spectral clustering.

2.2. Spectral Clustering

Briefly, spectral clustering is a three-step process. First, given N D-dim embeddings, the NxD embedding matrix is unit-normalized and the outer product is computed to create an NxN cross-correlation matrix. Then, eigendecomposition of the cross-correlation matrix reveals S large and many small eigenvalues, separated by a so-called eigen-gap [12], corresponding to S different speakers in the conversation; the smaller eigenvalues are discarded. Each large eigenvalue's N-dim eigenvector encodes the embeddings for one speaker with higher values, whereas embeddings for other speakers are near 0. Finally, the cross-correlation vector for each embedding is then represented as an S-dim basis vector over the S retained eigenvectors, and the argmax of this vector indicates the speaker to which the embedding is assigned.

A spectral variant of cosine similarity can be derived using the spectral basis. Each speaker's centroid is estimated from the S-dimensional embedding basis vectors assigned to that speaker. Confidence scores are then computed as the cosine similarity between each embedding and it's assigned speaker centroid, both in the spectral basis. Where different portions of a segment are covered by different embeddings, the cosine similarity scores are averaged.

2.3. Confidence Estimation Metrics

Confidence scores are evaluated in a rank-order context: we partition the segments of each conversation into low- and high-confidence subsets, containing the segments with the lowest (1-Cov)% and highest Cov% of confidence scores, respectively. The following metrics, computed on the high-confidence subset, are then reported:

- Coverage (Cov): This metric indicates the proportion of the diarized audio length in the entire conversation that attains a high-confidence value. This is an independent variable that we evaluate at two operating points: 70% coverage and 90% coverage.
- Covered Diarization Error Rate (cDER): This metric estimates the diarization error rate (DER) within the covered region of the diarization output. In order to prevent double counting the false negatives, the low-confidence speech segments are reported by the coverage metric and are not counted as a 'Miss.' If confidence scores were random, then cDER would approximate DER. Informative confidence scores should shift high-error segments out of the high-confidence subset, lowering the cDER.

2.4. Global Thresholding for Data Selection

To directly compare cDERs at the same operating point, we evaluated at fixed coverage percentages. However, for downstream tasks relying on semi-supervised data selection, one may wish to automatically select a subset of the data that maximizes the amount of data selected while minimizing the DER of the selected data (cDER). Here, a validation set is employed to find the minimum of the ratio of cDER to coverage, which varied between 50% and 90% coverage on our datasets. The exact

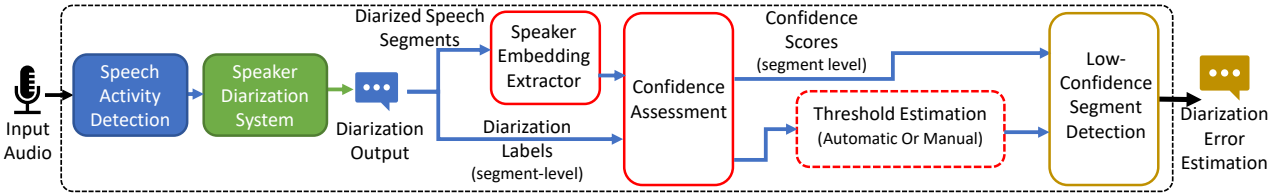


Figure 2: *A visual representation of the proposed speaker diarization confidence assessment framework.*

minimization criterion could be adjusted depending on the sensitivity of the down-stream task to precision vs. recall errors.

Table 1: *Diarization confidence assessment results on the DoPaCo and AMI datasets, given at fixed coverage values (Cov.) of 70% and 90%. The best performing confidence assessment method (T.M.) for each diarization method (D.M.) is marked in bold. Note the results reported in bold for the T1 method (no thresholding) are given at a coverage of 100%. The T2 method gives baseline (spectral clustering-based) results. The T3 to T5 methods give results of the proposed confidence assessment methods*

Dataset		AMI Dataset		Dopaco Dataset	
D.M.	T.M.	Cov. = 70%	Cov. = 90%	Cov. = 70%	Cov. = 90%
		cDER	cDER	cDER	cDER
M1	T1	*12.16*		*17.48*	
	T2	6.19	10.19	12.59	15.66
	T3	3.85	8.8	8.06	14.27
	T4	4.71	7.75	7.98	13.49
	T5	**3.67**	**6.66**	**7.78**	**12.79**
M2	T1	*5.18*		*13.03*	
	T2	**3.33**	4.17	7.39	10.39
	T3	3.35	4.55	8.37	11.48
	T4	**3.33**	4.17	7.39	10.39
	T5	3.34	**4.01**	**6.67**	**9.25**
	T6	4.31	4.89	N.A	N.A
M3	T1	*6.9*		*8.36*	
	T2	5.28	6.05	5.51	6.98
	T3	3.7	5.37	**4.45**	6.39
	T4	3.68	4.76	4.88	6.39
	T5	**3.67**	**4.43**	4.63	**6.07**
M4	T1	*3.33*		*8.85*	
	T2	1.26	2.53	3.92	**7.52**
	T3	2.51	3.06	6.51	8.08
	T4	1.26	2.53	3.92	**7.52**
	T5	**1.2**	**2.43**	**3.71**	**7.52**

M1	M2	M3	M4
xVector +SC	PyAnnote2.0 (E2E)	ECAPA-TDNN +SC	DOVER-Lap (of M1, M2, M3)

T1	T2	T3	T4	T5	T6
100% Coverage	Spectral Clustering Score	Local Confidence Score	Cosine Similarity Score	Silhouette Score	PyAnnote2.0 (E2E) Score

3. Experiments

3.1. Datasets

We perform our speaker diarization experiments on the following multi-speaker conversation datasets.

- **AMI:** This is the Augmented Multi-party Interaction (AMI) meeting dataset [13]. We use the official "Full ASR corpus" split with TNO meetings excluded from the Dev and Eval set. This dataset provides a point of reference for comparing the baseline diarization methods' performance and the corresponding metrics for confidence scores.
- **DoPaCo:** This internal dataset consists of manually de-identified doctor-patient conversations recorded using near-and far-field microphones in semi-unconstrained indoor settings of doctors' examination rooms. We use this dataset to demonstrate the diarization performance of several publicly available state-of-the-art methods on the challenges offered by a doctor-patient conversation.

3.2. Speaker Diarization Systems

We use the SpeechBrain [14] toolkit's speaker diarization setup using the ECAPA-TDNN [15] and xVector [16] models (pre-trained on the VoxCeleb [17] dataset) paired with spectral clustering (SC) as our first and second speaker diarization systems. The ECAPA-TDNN and xVector-based diarization systems generate embeddings from overlapping windows of length 1.5 secs, shifted by 250ms. PyAnnote 2.0 toolkit's [18] pre-trained E2E speaker diarization system [19] is our third diarization system. We also use the DOVER-Lap [4] method to combine the diarization outputs of the above three diarization systems, thereby instituting a fourth diarization system. Confidence scores for all methods except the E2E Activation Score were computed using embeddings generated by the ECAPA-TDNN system, since it was overall the strongest. DER and cDER are computed using a collar of 250 ms and exclude overlapping speech segments to maintain consistency with other published works.

4. Results

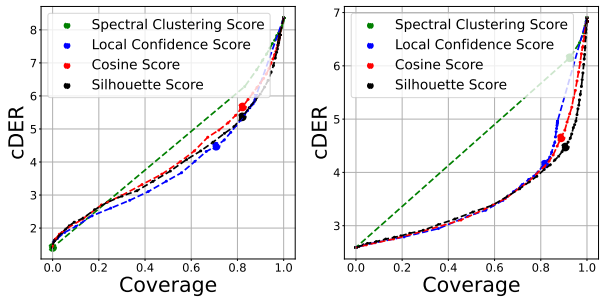
We report the speaker diarization performance (see Table 1 and Fig. 3) using covered Diarization Error Rate (cDER) and coverage metrics. In this set of experiments, we analyze and compare the performance (cDER) of the different confidence assessment methods at fixed coverage values of 70% and 90%.

Overall, silhouette score (T5) was the strongest performer, with the best or a close-second-best cDER across systems and coverages. T3 and T4 were also competitive, whereas using the spectral basis (T2) showed significant degradation in some conditions. On both the datasets and across all the diarization systems, the methods T3, T4, and T5 obtain an average $\sim$55% and $\sim$31% reduction in cDER at the fixed coverage values of 70% and 90% compared to the overall DER (T1).

M3 is a white-box condition, where the embedding model used for diarization is the same as the model used for confidence score estimation. While we expected that M1 and M2 systems might show more cDER improvement, owing to the use of a different embedding model for confidence scores, this was not a consistent trend.

T6 is the white-box condition for the E2E system, where embeddings are taken from within and the cosine similarity method is applied. This is in most direct contrast to method T4: cosine similarity over the ECAPA-TDNN embeddings generated on the E2E system's speaker segmentation. T6 was not competitive with the other methods on the AMI dataset, so we did not pursue it further on the DoPaCo dataset.

The DOVER-Lap (M4) system's performance was hampered by the spread of DERs of the base systems. Competi-



(a) *DoPaCo Dataset* (b) *AMI Dataset*

Figure 3: *cDER vs Coverage plots using the ECAPA-TDNN based diarization method. Large markers in the plots show the operating points at global thresholds.*

tive performance from M2 and M3 on the AMI dataset likely contributed to M4's strong AMI performance. For the DoPaCo dataset, M3 was far ahead of M1 and M2, and the combination did not improve the baseline DER. Interestingly, M4's cDER at 70% coverage was better than the single M3 system, despite higher DER than M3. Overall, we observe a ∼20% and ∼60% (relative) reduction in cDER at 10% and 30% loss of coverage, respectively, using the proposed methods. This demonstrates the efficacy of the methods at identifying low-confidence segments in the combined diarization output of several independent diarization systems.

Figure 3 shows the variation in cDER with coverage values for the proposed and baseline confidence assessment methods. We also mark the operating points corresponding to each technique's optimal threshold obtained by the global thresholding process (Section 2.4). Embedding-space methods T3, T4 and T5 show a much larger reduction in cDER values compared to the spectral method (T2) for a similar amount of coverage. For example, on the AMI dataset, methods T3-T5 demonstrate an average relative reduction of 57% cDER at 80% coverage. In contrast, the spectral method only reduced the cDER by 19% at a similar coverage.

5. Analysis

The spectral method T2 did not perform as well as the other methods in several cases. An analysis of the distribution of scores generated by T2 revealed a strongly skewed distribution, shown in Fig. 4(a), where there was little range over which to distinguish bad segment labels from the good ones. Scores based on embeddings from the ECAPA-TDNN and E2E models covered a broader range, as shown in Figures 4(b) and 4(c).

To better visualize the distribution of embeddings, Figure 5 shows the 2-dim t-SNE representation of the ECAPA-TDNN embeddings of the speech segments of a doctor-patient conversation marked with high or low confidence from the local confidence method (T3), including overlapping speech. We note that the speech segments farther from speaker centroids (possibly due to noise or high intra-speaker variance) are assessed as low-confidence segments. We also note that the proposed method assesses most overlapping speech segments as low-confidence segments, likely due to the speaker recognition system's inability to assign overlapping speech segments to either of the speakers. This demonstrates a possible limitation of the proposed method, as it will remove most overlapping speech segments when used to prune an overlap-aware speaker diarization system. However, in the future, we plan to use this limitation to prune outliers of an overlap speech detector [19] and combine

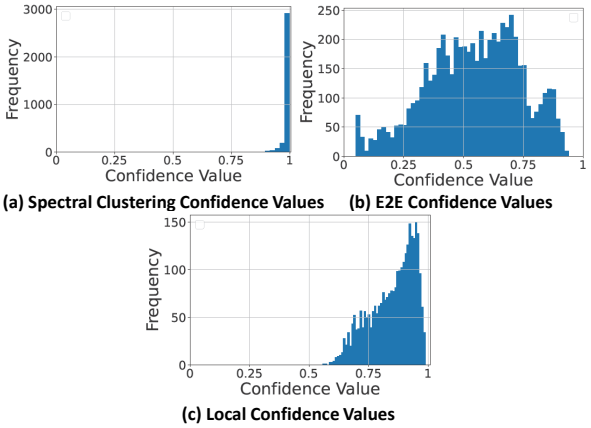


Figure 4: *A comparison of histograms of diarization confidences scores estimated using (a) spectral clustering-based, (b) End-to-End-based and (c) proposed confidence assessment methods on the AMI dataset.*

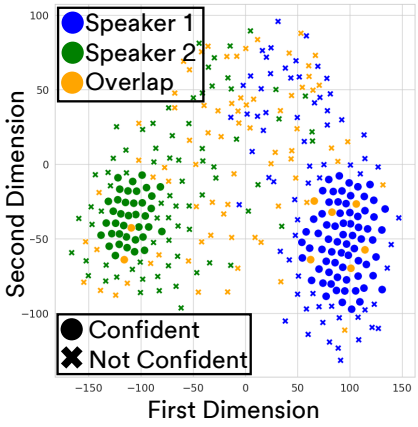


Figure 5: *A visual representation of the Local Confidence estimation results on a doctor-patient conversation from the DoPaCo dataset. The proposed method assigns lower confidence to most overlapping speech segments.*

it with the currently proposed method to develop an overlap-aware confidence assessment and thresholding technique.

6. Conclusion

Speaker diarization systems face many challenges due to the highly volatile nature of multi-talker speech conversations captured in unconstrained environments. While it is essential to develop speaker diarization systems robust to such challenges, downstream tasks that rely on speaker identity may be able to mitigate diarization errors where the system provides informative confidence measures. We conducted experiments across multiple datasets and diarization systems, comparing several different methods for assigning segment-level confidence to diarization systems' outputs. We found that silhouette score was always either the best or a close-second-best method for confidence assessment. Nonetheless, the top three methods were all able to isolate ∼30% of the diarization errors within segments with the lowest ∼10% of confidence scores, and ∼55% of the diarization errors within segments with the lowest ∼30% of confidence scores, indicating their predictive value.

7. References

[1] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, 2010.

[2] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, and Shinji Watanabe, "End-to-end neural speaker diarization with self-attention," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.

[3] Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Kenji Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors," in *INTERSPEECH*, 2021.

[4] Desh Raj, Leibny Paola Garcia-Perera, Zili Huang, Shinji Watanabe, Daniel Povey, Andreas Stolcke, and Sanjeev Khudanpur, "Dover-lap: A method for combining overlap-aware diarization outputs," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2021.

[5] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, 2022.

[6] Yunxin Zhao, "An acoustic-phonetic-based speaker adaptation technique for improving speaker-independent continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, 1994.

[7] Zhong Meng, Yashesh Gaur, Jinyu Li, and Yifan Gong, "Speaker adaptation for attention-based end-to-end speech recognition," *Interspeech*, 2019.

[8] Orith Toledo-Ronen and Hagai Aronowitz, "Confidence for speaker diarization using PCA spectral ratio," in *Annual Conference of the International Speech Communication Association*, 2012.

[9] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification," in *INTERSPEECH*, 2020.

[10] Vijayaditya Peddinti, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, "Low latency acoustic modeling using temporal convolution and lstms," *IEEE Signal Processing Letters*, 2018.

[11] Ketan Rajshekhar Shahapure and Charles Nicholas, "Cluster quality analysis using silhouette score," in *International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2020.

[12] Ulrike Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, 2007.

[13] Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post, "The AMI meeting corpus," 2005.

[14] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, et al., "Speechbrain: A general-purpose speech toolkit," *arXiv preprint arXiv:2106.04624*, 2021.

[15] Nauman Dawalatabad, Mirco Ravanelli, François Grondin, Jenthe Thienpondt, Brecht Desplanques, and Hwidong Na, "ECAPA-TDNN embeddings for speaker diarization," in *INTERSPEECH*, 2021.

[16] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-Vectors: robust DNN embeddings for speaker recognition," in *ICASSP*, 2018.

[17] Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, 2020.

[18] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "pyannote.audio: neural building blocks for speaker diarization," in *IEEE ICASSP*, 2020.

[19] Hervé Bredin and Antoine Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *INTERSPEECH*, 2021.