ONLINE SPEAKER DIARIZATION OF MEETINGS GUIDED BY SPEECH SEPARATION

Elio Gruttadauria, Mathieu Fontaine, Slim Essid

ONLINE SPEAKER DIARIZATION OF MEETINGS GUIDED BY SPEECH SEPARATION

Elio Gruttadauria[1], *Mathieu Fontaine*[1], *Slim Essid*[1]

[1]LTCI, Télécom Paris, Institut Polytechnique de Paris, France

ABSTRACT

Overlapped speech is notoriously problematic for speaker diarization systems. Consequently, the use of speech separation has recently been proposed to improve their performance. Although promising, speech separation models struggle with realistic data because they are trained on simulated mixtures with a fixed number of speakers. In this work, we introduce a new speech separation-guided diarization scheme suitable for the online speaker diarization of long meeting recordings with a variable number of speakers, as present in the AMI corpus. We envisage ConvTasNet and DPRNN as alternatives for the separation networks, with two or three output sources. To obtain the speaker diarization result, voice activity detection is applied on each estimated source. The final model is fine-tuned end-to-end, after first adapting the separation to real data using AMI. The system operates on short segments, and inference is performed by stitching the local predictions using speaker embeddings and incremental clustering. The results show that our system improves the state-of-the-art on the AMI headset mix, using no oracle information and under full evaluation (no collar and including overlapped speech). Finally, we show the strength of our system particularly on overlapped speech sections.

***Index Terms*—** online speaker diarization, source separation, overlapped speech, AMI, speaker embedding

1. INTRODUCTION

Speaker diarization (SD) aims at answering the question "who spoke when?" by segmenting a recording into speaker-homogeneous regions [1].

Speaker diarization has traditionally been framed as a clustering problem, with systems consisting of a cascade of several steps [2, 3], each individually optimized. As diarization systems become more effective, the inability of clustering-based systems to model overlapped speech directly becomes a non-negligible limiting factor. Indeed, up to 20% of total conversational speech time can be categorized as overlapping speech [4], which naturally calls for a change of paradigm. End-to-end neural diarization (EEND) models [5, 6] reframe the diarization task as a multi-label classification problem. By doing this, the EEND framework inherently considers the issue of overlapping speech. Other examples of non-clustering based systems are target-speaker VAD (TS-VAD) [7] and region proposal networks (RPNs) [8]. Although EEND-based systems have shown state-of-the-art performance over the clustering paradigm, the best models rely on the self-attention mechanism [9] and tend to require a lot of data to be trained properly. In this context, speech separation models (SSep) show potential for better handling overlapped speech, while being computationally more efficient [10]. Currently, the novel speech separation guided diarization (SSGD) paradigm [11, 12] is still limited because of the inability of SSep models to behave well on realistic data: the better performance on overlapped speech sections is counteracted with worse performance on the remainder of the audio. Additionally, no work has been done yet to deal with multiple speakers (*i.e.*, more than 2 speakers), making the SSGD paradigm not ready yet for general settings, even less for online speaker diarization.

Online SD systems make predictions at each time step with information available only up until that point (or slightly in the future). Only a few models are online by nature [13], but offline systems may sometimes be adapted to operate online. The work from Kinoshita et al. [14, 15] introduces an adaptation of the EEND model to handle long recordings. Coria et al. [16] used the same technique to adapt the EEND framework to real-time processing. In their proposal, predictions are made locally on short overlapping windows, and incremental clustering is used to solve the permutation problem.

With this work, we introduce a novel speaker diarization system architecture that expands the SSGD paradigm to accommodate meeting recordings (with more than 2 speakers), and we study its performance in the online diarization setting, focusing on single-microphone scenarios. To the best of our knowledge, this is the first work using SSep for diarization outside the conversational telephone speech (CTS) domain where only 2 speakers are present in the entire recording. As will be discussed in section 5, separation models struggle when the number of speakers active during the testing phase differs from that considered during the training of the separation network. Nevertheless, our solution is suitable for an arbitrary number of speakers. Notably, we are able to improve the state-of-the-art performance on AMI headset mix in the online setting using no oracle information. Our system can also estimate sources for each speaker in addition to the diarization result. Finally, we also show the superiority of our method on the overlapped speech sections in particular. The code to reproduce the results of this work is freely available[1].

2. RELATED WORK

Fang et al. [11] introduced the speech separation guided diarization (SSGD) approach, refining the work from [12]. Their system employs dynamic selection between conventional clustering-based diarization which is effective for single-speaker segments and SSGD which excels in handling overlapped speech. However, they note occasional SSGD instability and SSep model failures, resulting in speaker confusion and false alarms due to channel leakage and artifacts in the estimated sources. In the context of SSep models, leakage is defined as the presence of one or more other speakers in an estimated source. In [17], a leakage removal algorithm is proposed,

This work was supported by the Audible project, funded by French BPI and partly supported by ANR Project SAROUMANE (ANR-22-CE23-0011). Also, it was performed using HPC resources from GENCI-IDRIS.

[1]egruttadauria98/SSpaVAlDo

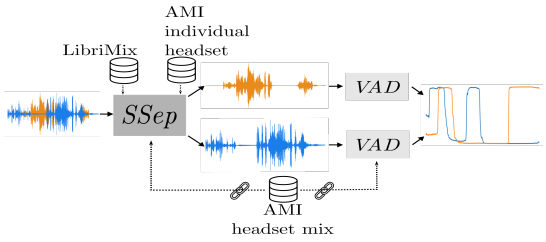

Fig. 1. Diagram of the inference process for local predictions on 5-s windows. The dataset used for the end-to-end finetuning is symbolized with chains.

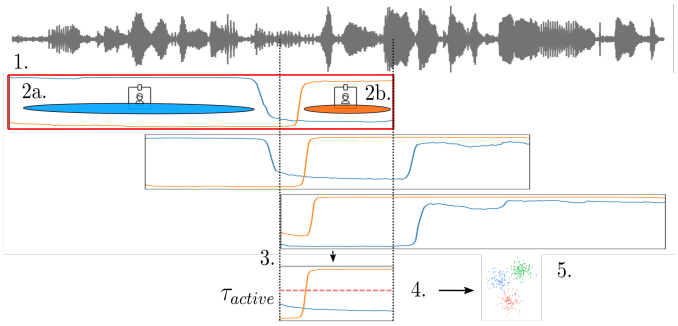

Fig. 2. Diagram of a single step of the stitching of local predictions.

based on the SI-SDR metric [18]. Recent advancements [19] show that fine-tuning the model is an effective approach to mitigate the source leakage problem. Notably, adapting the VAD to the estimated sources reduces false alarms, but the best results arise from jointly fine-tuning the separation model and VAD in an end-to-end manner.

The previous work on SSGD research focuses on the conversational telephone speech (CTS) domain, involving at most two speakers. This limitation simplifies the SSGD problem enabling a single-pass inference to be performed. Alternatively, for longer recordings or SSep models with a restricted receptive field, the system can operate on short overlapping windows whose predictions are then stitched together using the correlation between overlapping sections of consecutive windows. This approach is known as Continuous Speech Separation (CSS) [20]. When working with meeting conversations, where more than two speakers are present, the CSS approach is no longer feasible, as it implies that each local prediction must have as many outputs as the total number of speakers. In fact, as discussed in Section 5, SSep models like ConvTasNet [21] or DPRNN [22] see a drastic degradation in performance when increasing the number of output sources they consider. As an alternative approach, we use the speaker embedding-based stitching method proposed by [16]. Speaker embeddings can be used to solve the permutation problem between different local predictions, but also to distinguish between new and already seen speakers.

3. PROPOSED SYSTEM

The system proposed in this work is composed of 3 components: speech separation (SSep), voice activity detection (VAD) and a speaker embedding-based stitching mechanism.

Speech separation is performed on sliding 5-s windows to obtain "local" predictions. The overlap between subsequent windows is 90%, meaning the step is 500 ms. For each window, the active speakers for the incremental clustering are searched only in the last 500 ms, while the rest of the window is used as context to better estimate the speaker embeddings.

Separation and VAD. The 5-s input segments $x \in \mathbb{R}^{1\times T}$ are first fed to the SSep model, which estimates the sources $\hat{s}_i \in \mathbb{R}^{1\times T}$ for each output of the model, where T is the number of samples in the segments. VAD is then applied independently to each $\hat{s}_i$ to estimate the speech activities $\hat{a}_i \in [0,1]^{1\times F}$, where F is the number of frames. The SSep model takes a single-channel audio as input and outputs a fixed number of estimated sources. In this work, we test models with 2 or 3 output sources. To bridge the domain shift between real data and estimated sources, the VAD needs to be fine-tuned. Similarly to [19], we consider two types of finetuning. The first variant is to adapt only the VAD on the estimates of the SSep model. The second finetuning strategy consists of jointly adapting both the SSep and the VAD in an end-to-end fashion.

Speaker embedding-based stitching. To combine local predictions across time, a permutation problem needs to be solved between consecutive windows. Additionally, new speakers can appear as well. In this work, we rely on the use of speaker embeddings for the sake of stitching together the predictions on the 5-s sliding windows, following the approach from [16].

Figure 2 summarizes the logic of the stitching process. At each step: 1) start from the current window (top, outline red), 2a/2b) speaker embeddings are estimated, 3) the predictions of the active speakers are aggregated with a delay (bottom) if the latency is above the minimum of 500ms, 4) the activities are binarized using τ_{active} to get speaker segments, 5) incremental clustering is performed on the speaker embeddings to find the best match between the speaker segments and the existing centroids. All windows in the figure are shown with the estimated activities already computed. To improve the statistic pooling layer, the estimated activities are used to inform the weights of the frames when computing the embeddings, as detailed in [16].

The clustering is governed by other two hyperparameters: δ_{new} and ρ_{update}. The first parameter, δ_{new}, defines the threshold distance between a new embedding and the closest centroid to define a new speaker. The distance metric used is the cosine similarity, as the speaker embedding is an implementation of the X-vector architecture [23, 24] trained with additive angular margin loss [25]. The latter parameter, ρ_{update}, prevents embeddings estimated from short speech segments from updating the centroids of their cluster. The rationale is to prevent noisy embeddings from damaging the speaker representation of the centroids.

4. EXPERIMENTAL SETUP

Dataset. As this work focuses on meeting conversations, we evaluate our models on the AMI dataset [26]. Specifically, the evaluation of all our proposed models is performed on the headset mix, our focus being on single-microphone settings. The other datasets used in the training and finetuning stage of the SSep models are LibriMix type "mix_both" [27] and the individual (speaker-focused) headset recordings from AMI. In order to compare our results to previous works, we have used the AMI evaluation protocol proposed in [28].

Architecture configuration and training details. We consider two different separation architectures: ConvTasNet [21] and DPRNN [22], with a view to gaining insight on the behaviour of our system, especially its robustness, when considering different SSep architectures. Both separation models are first trained on fully overlapped

mixtures from LibriMix type "mix_both"[27], using 3-second segments. For both models, we have used the same configuration as in the Asteroid toolkit [29]. However, for DPRNN, to reduce the computational burden, the kernel size and the stride have been set, respectively, to 32 and 16. The chunk size has been increased to 300 to reduce the length of the inter-RNN processing. The hop size was increased to 150 to maintain it at 50% of the chunk size.

After the training on LibriMix, the SSep models are finetuned on real data using the AMI train set, lowering the learning rate by a factor of 10 to 0.0001. Given that the isolated sources are not available, we have resorted to using the individual headset microphones of the active speakers as the ground truth. Note that these recordings may not be optimal as sources because they include other speakers in the vicinity. Finally, as a last finetuning step, we have joined the VAD to each output source of the SSep model. In our experiments, we have used the pretrained VAD from Pyannote [30] [2]. We have tried two combinations: freezing the SSep model to finetune only the VAD, and finetuning the entire system end-to-end. While both approaches showed remarkable improvements over the use of a pre-trained VAD, the end-to-end approach has shown the best performance, as shown in Table 1 and discussed in Section 5

5. RESULTS

All results presented rely on the AMI protocol presented in [28]. The inference is carried out under full evaluation, meaning with no collar and evaluating also overlapped speech.

Model	DER	FA	MS	SC
SSep AMI + VAD E2E	**27.2**	**1.8**	18.4	**7.0**
SSep LibriMix + VAD E2E	28.4	2.0	**18.3**	8.1
SSep LibriMix + VAD finetuned	34.4	1.9	22.6	9.9
SSep LibriMix + VAD	42.8	3.7	19.3	19.8
Coria et al.[3] [16]	28.5	4.4	**12.0**	12.1
Kwon et al. [31]	22.9	n.a.	14.5	8.3
Yue et al. [32]	19.0	-	-	-
Kynych et al. [33]	21.2	-	-	-

Table 1. Comparison of our proposed online diarization system with the literature. The top section of the table presents an ablation study of the training methodology. The SSep used is ConvTasNet with 2 outputs. The bottom section of the table report results which rely at least in part on oracle information.

Model performance and ablations. Table 1 presents the results of our new speaker diarization system based on SSep and VAD finetuned end-to-end, along with a few ablations allowing one to clarify the impact of each component of the system. The results show that our proposed system is competitive with the previous work [16], and that it improves the performance as measured by the overall DER.

Comparing the works from the bottom part of Table 1 with our methods is not straightforward as they use oracle information. The only comparison that we can make is against VBx [28] (on which [32] is based), but in an offline setting. To this end, we use the speaker diarization pipeline from [34], which can be considered as an offline variant of [16]. In this case, our best model achieves a DER of 23.5% without any hyperparameter tuning, while VBx with Pyannote VAD (instead of oracle VAD) achieves 24.1% [30].

[2]available at hf.co/pyannote/embedding/. Note that this is not the same model as the one used in [16], which exploits an improved version instead.

[3]Reproduced results

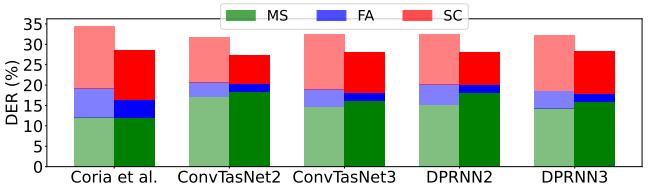

Fig. 3. All proposed online systems compared to Coria et al.'s, tested at minimum (0.5s, left bar) and maximum latency (5s, right bar). The DER is broken down into its constituents: Missed Speech (MS), False Alarm (FA) and Speaker Confusion (SC).

Further, the ablation experiments (upper part of Table 1) show that all the components of our system play a role in improving performance. Removing the AMI finetuning of the SSep model lowers performance even if the model is still finetuned end-to-end. In line with [19], switching from end-to-end finetuning to VAD-only adaptation degrades the performance. Finally, a pre-trained SSep model with a non-finetuned VAD leads to the worst performance overall.

Choice and parametrisation of the SSep model. To test the dependence of the results on the quality of the SSep system used, we repeat the online inference on AMI with two different SSep models, with 2 or 3 output sources, for a total of four model combinations as presented in Table 2. The results show that all the models considered are competitive with [16]. For both ConvTasNet and DPRNN, the 2-output models achieve a better score than the 3-output counterpart. On the other hand, the 2-output models obtain the worst missed speech result, which is expected as the distribution of 5-s segments with more than 2 speakers in AMI is non negligible.

The benchmark of the different SSep models is further explored in Figure 3. Relative to Coria et al [16], our system improves the performance in low algorithmic latency settings. For all SSep models considered, increasing the latency leads to a reduction of false alarms and speaker confusion, but also to an increase in missed detection. This is in contrast with [16], for which missed speech seems not to be affected by the change in latency.

Model	DER	FA	MS	SC
ConvTasNet2	**27.2**	**1.8**	18.4	**7.0**
ConvTasNet3	28.1	2.1	16.0	10.0
DPRNN2	28.0	2.2	18.0	7.8
DPRNN3	28.4	2.1	15.8	10.4

Table 2. Online diarization results for SSep AMI + VAD E2E using different SSep models with 2 or 3 output sources.

Local inference and overlapped speech performance. To disentangle the contributions of the local prediction from those of the stitching mechanism, we have evaluated the SSep models on individual segments of 5 seconds, as detailed in Table 3. The only hyperparameter here is the threshold value to convert the continuous prediction into binarized outputs, equivalent to τ_{active}. For each model, we also report the performance when scoring only overlapped speech sections. The baseline for comparison is the segmentation model from [16], an LSTM-based EEND model with 4 outputs trained on multiple datasets including AMI.

The results show that all our proposed models improve on the baseline, both regarding overall test DER and considering only overlapped speech. Interestingly, in contrast with the results in Table 2, here we find DPRNN to perform better that ConvTasNet. Also, 3-

Model	Test DER by number of speakers				Test DER
	1 spk	2 spks	3 spks	4 spks	
ConvTasNet2	6.0 ± 0.3	14.8 ± 0.4	25.3 ± 0.5	34.6 ± 0.8	15.8 ± 0.3
OVL-only scoring	*n.a.*	*16.7 ± 0.4*	*26.7 ± 0.4*	*33.0 ± 0.5*	*24.3 ± 0.3*
ConvTasNet3	5.9 ± 0.4	16.4 ± 3.8	23.9 ± 0.6	30.3 ± 0.1	15.4 ± 0.3
OVL-only scoring	*n.a.*	*21.1 ± 0.4*	*24.2 ± 0.5*	*28.7 ± 0.8*	*24.0 ± 0.3*
DPRNN2	6.8 ± 0.3	15.7 ± 0.5	25.3 ± 0.5	35.3 ± 0.9	16.4 ± 0.3
OVL-only scoring	*n.a.*	*17.7 ± 0.4*	*27.6 ± 0.5*	*33.9 ± 0.6*	*25.2 ± 0.3*
DPRNN3	5.4 ± 0.3	15.0 ± 0.5	24.7 ± 0.4	33.1 ± 0.8	$\mathbf{15.2 \pm 0.3}$
OVL-only scoring	*n.a.*	*18.8 ± 0.4*	*25.6 ± 0.4*	*32.1 ± 0.7*	*24.5 ± 0.3*
Coria et al. [16]	5.9 ± 0.3	17.0 ± 0.5	26.9 ± 0.6	33.5 ± 0.9	16.7 ± 0.2
OVL-only scoring	*n.a.*	*24.1 ± 0.6*	*29.6 ± 0.5*	*32.7 ± 0.6*	*28.2 ± 0.3*

Table 3. Performance on individual segments of 5 seconds. The error on segments with no speakers is not reported because it is null for all the models. The performance scoring only the overlapped portion of the speech is noted as *OVL-only scoring*. For all models $\tau_{active} = 0.5$. The results are reported with a 95% confidence interval.

output models perform better than the 2-output ones. It is important to note that all systems are competitive also on segments with only one speaker, which can be mishandled by SSep models, as discussed in [11, 12]. Additionally, SSep models with 2 and 3 outputs have similar performance on segments with 1 and 2 speakers, which is not to be expected if one considers the results from Section 5. We attribute this generalization to the end-to-end finetuning, as models are trained also on segments with fewer speakers, contrary to training with SI-SDR loss. With these results, we claim that the SSGD framework can be robust enough to be used as a stand-alone approach, without being integrated with other methods like in [11, 12].

Behaviour of SSep models after adaptation on real data. The SSGD framework is appealing also because it performs separation for free. For each step of the training pipeline detailed in Section 4, we show some examples of how the SSep models behave[4]. It is not possible to objectively evaluate the separation on AMI because ground-truth sources are not available. Here we limit ourselves to a few observations on the behaviour of the models.

The SSep models are first pretrained on fully overlapped mixtures from LibriMix, with as many speakers as the number of outputs of the model. For a SSep model trained on fully overlapped mixtures, all recordings with less speakers than the number of outputs are out-of-domain examples. After finetuning the SSep models on real data from AMI, the estimated sources were found to be less affected by phenomena that lead to speaker confusion, such as splitting one speaker into multiple outputs. Nevertheless, because the SSep models are finetuned on individual microphones which contain also speech from nearby speakers, the estimated sources present more leakage than the models just trained on LibriMix. The leaked speakers are always at lower energy than the main speaker, so a finetuned VAD is usually able to distinguish them and avoid false alarms. With the end-to-end finetuning, the SSep models learn to make a few little adjustments to improve the diarization score, but the leakage is still present. As such, finetuning end-to-end alone does not lead to better separation automatically, as the goal is only diarization performance. Our interpretation is that the finetuning end-to-end pushes the model to reduce the leakage at least when it can lead to false alarm, while it is otherwise kept.

Relationship between SSep performance and number of outputs. We have found that increasing the number of outputs of the speech separation model always leads to a loss in performance. Figure 4 shows how the performance of ConvTasNet5 changes when testing it on mixtures with 5 or fewer speakers, as shown on the x-axis.

[4]egruttadauria98/SSpaVAlDo

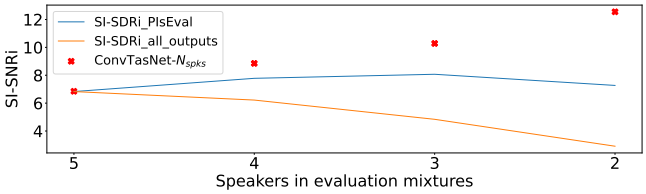

Fig. 4. Performance of ConvTasNet5 trained on Libri5Mix on mixtures with 5, 4, 3 and 2 speakers. The red crosses show the performance of the SSep model with as many outputs as the speakers in the mixtures.

As there is no straightforward way to evaluate a SSep model with a mismatch between the estimated outputs and the actual number of sources, we use both a harsh metric and a forgiving metric. The harsh metric, *all outputs*, used as a reference for the additional sources an all zero-signal [5]. The forgiving metric, *PIsEval*, uses the oracle number of speakers in the mixture, N_{spks}, to score only the estimated sources that best resemble the references. For ConvTasNet5, *PIsEval* improves initially when N_{spks} is reduced first to 4 and then to 3, because the mixtures are easier to separate. Once the N_{spks} reaches 2, the performance worsens, possibly because the out-of-domain factor outweighs the easier separation. Lastly, we also plot with red crosses the performance of ConvTasNet-N_{spks} on mixtures with N_{spks} speakers. At each value of the x-axis, the difference between the red cross and the blue line shows the minimum loss in performance by using a SSep with 5 outputs instead of a SSep with as many outputs as the speakers in the mixtures.

6. CONCLUSIONS

We have presented a novel SSGD system for online speaker diarization that achives state-of-the-art performance on AMI headset mix. Our results show that the limitations of SSep on real data can be overcome, leading to a diarization model that can better handle overlapped speech and estimates sources for each speaker.

[5]A small constant is added to avoid numerical errors in the computation of the SI-SDR metric

7. REFERENCES

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A Review of Speaker Diarization: Recent Advances with Deep Learning," 2021.

[2] M. Sahidullah, J. Patino, S. Cornell, R. Yin, S. Sivasankaran, H. Bredin *et al.*, "The Speed Submission to DIHARD II: Contributions & Lessons Learned," 2019.

[3] F. Landini, O. Glembek, P. Matějka, J. Rohdin, L. Burget, M. Diez *et al.*, "Analysis of the BUT Diarization System for VoxConverse Challenge," 2021.

[4] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang *et al.*, "CHiME-6 Challenge:Tackling Multispeaker Speech Recognition for Unsegmented Recordings," 2020.

[5] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue, and K. Nagamatsu, "End-to-End Neural Diarization: Reformulating Speaker Diarization as Simple Multi-label Classification," 2020.

[6] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and P. Garcia, "Encoder-Decoder Based Attractors for End-to-End Neural Diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 30, pp. 1493–1507, 2022.

[7] I. Medennikov, M. Korenevsky, T. Prisyach, Y. Khokhlov, M. Korenevskaya, I. Sorokin *et al.*, "Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario," in *Proc. Interspeech*, 2020, pp. 274–278.

[8] Z. Huang, S. Watanabe, Y. Fujita, P. García, Y. Shao, D. Povey *et al.*, "Speaker Diarization with Region Proposal Network," in *Proc. ICASSP*, 2020, pp. 6514–6518.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, "Attention Is All You Need," 2023.

[10] E. Tzinis, Z. Wang, and P. Smaragdis, "Sudo rm -rf: Efficient Networks for Universal Audio Source Separation," in *IEEE MLSP*, Sep. 2020, pp. 1–6.

[11] X. Fang, Z.-H. Ling, L. Sun, S.-T. Niu, J. Du, C. Liu *et al.*, "A Deep Analysis of Speech Separation Guided Diarization Under Realistic Conditions," in *Proc. APSIPA ASC*, 2021, pp. 667–671.

[12] S.-T. Niu, J. Du, L. Sun, and C.-H. Lee, "Separation Guided Speaker Diarization in Realistic Mismatched Conditions," 2021.

[13] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully Supervised Speaker Diarization," in *Proc. ICASSP*, 2019, pp. 6301–6305.

[14] K. Kinoshita, M. Delcroix, and N. Tawara, "Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds," *Proc. ICASSP*, 2021.

[15] ——, "Advances in integration of end-to-end neural and clustering-based diarization for real conversational speech," *Proc. Interspeech*, 2021.

[16] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "Overlap-Aware Low-Latency Online Speaker Diarization Based On End-To-End Local Segmentation," in *IEEE ASRU*, 2021.

[17] G. Morrone, S. Cornell, D. Raj, L. Serafini, E. Zovato, A. Brutti *et al.*, "Low-Latency Speech Separation Guided Diarization for Telephone Conversations," in *IEEE SLT*, 2023, pp. 641–646.

[18] J. L. Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR - half-baked or well done?" in *Proc. ICASSP*, 2019.

[19] G. Morrone, S. Cornell, L. Serafini, E. Zovato, A. Brutti, and S. Squartini, "End-to-End Integration of Speech Separation and Voice Activity Detection for Low-Latency Diarization of Telephone Conversations," 2023.

[20] Z. Chen, T. Yoshioka, L. Lu, T. Zhou, Z. Meng, Y. Luo *et al.*, "Continuous Speech Separation: Dataset and Analysis," in *Proc. ICASSP*. IEEE, 2020, pp. 7284–7288.

[21] Y. Luo and N. Mesgarani, "Conv-TasNet: Surpassing Ideal Time-Frequency Magnitude Masking for Speech Separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 8, pp. 1256–1266, 2019.

[22] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path RNN: efficient long sequence modeling for time-domain single-channel speech separation," in *Proc. ICASSP*, 2020.

[23] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur, "Deep Neural Network Embeddings for Text-Independent Speaker Verification," in *Proc. Interspeech*, 2017.

[24] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.

[25] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," in *Proc. CVPR*, 2019, pp. 4685–4694.

[26] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn *et al.*, "The AMI meeting corpus," 2005.

[27] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, "LibriMix: An Open-Source Dataset for Generalizable Speech Separation," 2020.

[28] 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*, vol. 71, p. 101254, 2022.

[29] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper *et al.*, "Asteroid: the PyTorch-based audio source separation toolkit for researchers," 2020.

[30] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Proc. Interspeech*, 2021.

[31] Y. Kwon, H.-S. Heo, B.-J. Lee, Y. J. Kim, and J.-w. Jung, "Absolute decision corrupts absolutely: conservative online speaker diarisation," 2022.

[32] Y. Yue, J. Du, M.-K. He, Y. Yeung, and R. Wang, "Online Speaker Diarization with Core Samples Selection," in *Proc. Interspeech*, 2022, pp. 1466–1470.

[33] F. Kynych, J. Zdansky, P. Cerva, and L. Mateju, "Online Speaker Diarization Using Optimized SE-ResNet Architecture," in *Text, Speech, and Dialogue*, K. Ekštein, F. Pártl, and M. Konopík, Eds. Springer Nature Switzerland, 2023, vol. 14102, pp. 176–187, series Title: Lecture Notes in Computer Science.

[34] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *INTERSPEECH 2023*. ISCA, Aug. 2023, pp. 1983–1987.