



UNIVERSITÀ POLITECNICA DELLE MARCHE
Repository ISTITUZIONALE

End-to-end integration of speech separation and voice activity detection for low-latency diarization of telephone conversations

This is a pre print version of the following article:

Original

End-to-end integration of speech separation and voice activity detection for low-latency diarization of telephone conversations / Morrone, Giovanni; Cornell, Samuele; Serafini, Luca; Zovato, Enrico; Brutti, Alessio; Squartini, Stefano. - In: SPEECH COMMUNICATION. - ISSN 0167-6393. - ELETTRONICO. - 161:(2024), pp. 103081.1-103081.13. [10.1016/j.specom.2024.103081]

Availability:

This version is available at: 11566/330096 since: 2024-07-01T08:35:07Z

Publisher:

Published

DOI:10.1016/j.specom.2024.103081

Terms of use:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. The use of copyrighted works requires the consent of the rights' holder (author or publisher). Works made available under a Creative Commons license or a Publisher's custom-made license can be used according to the terms and conditions contained therein. See editor's website for further information and terms and conditions.

This item was downloaded from IRIS Università Politecnica delle Marche (<https://iris.univpm.it>). When citing, please refer to the published version.

(Article begins on next page)

End-to-End Integration of Speech Separation and Voice Activity Detection for Low-Latency Diarization of Telephone Conversations

Giovanni Morrone^{a,b,*}, Samuele Cornell^a, Luca Serafini^a, Enrico Zovato^b, Alessio Brutti^c and Stefano Squartini^a

^aUniversità Politecnica delle Marche, Ancona, Italy

^bAlmawave S.p.A., Rome, Italy

^cFondazione Bruno Kessler, Trento, Italy

ARTICLE INFO

Keywords:

online speaker diarization
speech separation
end-to-end learning
overlapped speech
conversational telephone speech

ABSTRACT

Recent works show that speech separation guided diarization (SSGD) is an increasingly promising direction, mainly thanks to the recent progress in speech separation. It performs diarization by first separating the speakers and then applying voice activity detection (VAD) on each separated stream. In this work we conduct an in-depth study of SSGD in the conversational telephone speech (CTS) domain, focusing mainly on low-latency streaming diarization applications. We consider three state-of-the-art speech separation (SSep) algorithms and study their performance both in online and offline scenarios, considering non-causal and causal implementations as well as continuous SSep (CSS) windowed inference. We compare different SSGD algorithms on two widely used CTS datasets: CALLHOME and Fisher Corpus (Part 1 and 2) and evaluate both separation and diarization performance. To improve performance, a novel, causal and computationally efficient leakage removal algorithm is proposed, which significantly decreases false alarms. We also explore, for the first time, fully end-to-end SSGD integration between SSep and VAD modules. Crucially, this enables fine-tuning on real-world data for which oracle speakers sources are not available. In particular, our best model achieves 8.8% DER on CALLHOME, which outperforms the current state-of-the-art end-to-end neural diarization model, despite being trained on an order of magnitude less data and having significantly lower latency, i.e., 0.1 vs. 1 seconds. Finally, we also show that the separated signals can be readily used also for automatic speech recognition, reaching performance close to using oracle sources in some configurations.

1. Introduction

Speaker diarization consists in identifying “who spoke when” in an input audio, by segmenting it into speaker-attributed regions (Anguera et al., 2012; Park et al., 2022) that correspond to speakers’ utterances. It is an essential pre-processing task in many applications, such as lectures, meetings, live captioning, speaker-based indexing, telephone calls, and doctor-patient conversations.

Historically, diarization has relied on clustering-based methods, which have been widely investigated since the 90s, and have represented the de-facto standard approach to diarization for many years. Recently, the invention of end-to-end neural diarization (EEND) (Fujita et al., 2020) has shown promising improvements in diarization accuracy, especially in the presence of overlapped speech, which may constitute up to 20% of total speech in real conversations (Watanabe et al., 2020). Indeed, classical, clustering-based systems are not able to handle overlapped speech as clustering is usually applied on single-speaker embeddings extracted from short frames. In this case, overlap-aware diarization can be performed using post-processing strategies (Bullock et al., 2020; Raj et al., 2021b). There

also exist methods such as target-speaker voice activity detection (VAD) (Medennikov et al., 2020), region proposal networks (Huang et al., 2020) and speech separation guided diarization (SSGD) (Fang et al., 2021) that do not belong to these two categories. In particular, SSGD performs diarization by combining speech separation (SSep) and VAD. It is particularly appealing as separated sources could be readily fed in input to an automatic speech recognition (ASR) system (see Section 4.6).

The majority of the methods above only work offline and thus are not suitable for streaming processing. The extension from offline to online processing is way simpler for EEND. Clustering-based systems consist of a pipeline of several modules (i.e., voice activity detection, speaker embedding extraction, clustering, etc). Each of these modules has to be updated in order to work online. In contrast, end-to-end methods can be easily adapted for online diarization by employing a speaker-tracing buffer (STB) to store previous input-output pairs that can be exploited for online inference (Xue et al., 2021a,b; Horiguchi et al., 2023). An alternative strategy is to simply replace the neural architecture with another one that allows low-latency streaming processing (Han et al., 2021). A similar approach, which is the basis for this work, is proposed by Morrone et al. (2022a) also for SSGD.

In this preliminary work, SSep and VAD models were trained independently and combined together with no additional training. Despite its simplicity, this approach only

*Corresponding author

✉ g.morrone@almawave.it (G. Morrone); s.cornell@pm.univpm.it (S. Cornell); l.serafini@univpm.it (L. Serafini); enrico.zovato@pervoice.it (E. Zovato); brutti@fbk.eu (A. Brutti); s.squartini@univpm.it (S. Squartini)

ORCID(s): 0000-0003-2163-1779 (G. Morrone)

reaches a sub-optimal diarization performance. Additionally, a major disadvantage is that the SSep module needs a dataset in which oracle sources for the speakers are available. This could lead to mismatched training-inference conditions as in most real-world scenarios oracle speaker sources are difficult to obtain (Subakan et al., 2022). In this paper, we build upon this previous work (Morrone et al., 2022a) and propose a fully end-to-end integration of the SSep and VAD modules. This approach only requires diarization labels during the fine-tuning stage and thus eliminates the need for oracle speaker target sources. We show that the proposed end-to-end strategy provides significant improvements on two widely used datasets, i.e., Fisher Corpus (Cieri et al., 2004) and CALLHOME (Przybicki and Alvin, 2001). These reflect two different training vs. inference conditions: fully matched and unmatched respectively. In particular, our best online model outperforms the current state-of-the-art EEND system on the 2-speaker subset of the CALLHOME dataset despite having an order of magnitude lower latency, i.e., 0.1 vs 1 seconds. Moreover, we review and extend our previous work (Morrone et al., 2022a) with additional analysis. The other contributions with respect to the previous paper are summarized below:

- We consider an additional more performing SSep model, i.e., DPTNet (Chen et al., 2020a).
- We carry out an analysis of the effect of varying model latency on diarization performance.
- Since separated sources are not provided in the CALLHOME dataset, we create a simulated version that enables the adaptation of SSep models that further reduces diarization errors.
- We thoroughly study the effect of the leakage removal algorithm on both disjoint and end-to-end trained SSGD.

In the next subsection we report a brief summary of the diarization state-of-the-art. In Section 2 we provide a description of the SSGD framework. The experimental setup is shown in Section 3. Experiments and results are reported in Section 4. Finally, we draw the conclusions in Section 5.

1.1. Related Works

The conventional clustering-based diarization approach is typically a cascade of three tasks: voice activity detection, speaker embedding extraction from speech segments and clustering of the embeddings. Previous works constantly improve the overall performance by developing better methods for one or more tasks. Several papers focus on the development of better speaker embeddings extractors (Dehak et al., 2010; Garcia-Romero et al., 2017; Desplanques et al., 2020; Xiao et al., 2021; Koluguri et al., 2022). In contrast, in Park et al. (2019), Singh and Ganapathy (2021) and Landini et al. (2022) different clustering algorithms are proposed. Conventional clustering algorithms can only handle correctly single-speaker segments, thus overlapped speech is usually

missed out or incorrectly labeled. To deal with overlapped speech, recent works extend the standard pipeline with overlap assignment techniques (Bullock et al., 2020; Raj et al., 2021b; Jung et al., 2021). These methods need accurate overlap detection, which is often hard to train. Furthermore, embedding extractors trained on single-speaker utterances may not be reliable for overlapping segments, resulting in speaker confusion errors (Raj et al., 2021b).

Recently, deep learning methods are employed for end-to-end neural diarization (EEND) approaches. A major advantage of EEND is that they are able to deal with overlapped speech without any modification. The first EEND methods perform diarization as a simple multi-speaker voice activity detection (VAD) problem, in which each output represents a different speaker's speech activity. EEND systems can employ different neural architectures, such as bidirectional long-short memory (Fujita et al., 2019a) and self-attention (SA-EEND) (Fujita et al., 2019b). EEND-based systems are trained directly to perform diarization using permutation invariant training (PIT) (Kolbæk et al., 2017) as the diarization problem is inherently permutation-invariant without any a-priori information. With enough training data, end-to-end approaches have been shown to outperform current state-of-the-art clustering-based systems (Horiguchi et al., 2021c). Contrary to clustering-based approaches, the maximum number of total speakers is fixed in the aforementioned EEND architectures. Additionally, end-to-end systems need to process the entire input signal during inference, resulting in significant memory consumption for long recordings (e.g., >10 minutes). Chunk-wise processing can help but is not viable as it leads to speaker label permutation across chunks due to PIT. Horiguchi et al. (2020) solves the first problem by extending the basic EEND with an auto-regressive encoder-decoder (EEND-EDA) architecture. The second issue is generally addressed by combining clustering and EEND. In Horiguchi et al. (2021a) EEND is exploited to refine the results of a clustering-based algorithm. Bredin and Laurent (2021) proposes a speaker segmentation model inspired by EEND to perform overlap-aware resegmentation in a conventional diarization pipeline. However, these approaches result in more complex systems in contrast to the simplicity of fully end-to-end architectures. A tighter integration of EEND and clustering, i.e., EEND-vector clustering (EEND-VC), is proposed in Kinoshita et al. (2021a,b). It performs chunk-wise processing to ensure that a given maximum number of active speakers can be present in each chunk (e.g., 2 or 3). The original EEND is modified to output global speaker embeddings that are aggregated across chunks using a constrained clustering algorithm. This method can both deal with an arbitrary number of speakers and solve the inter-chunk speaker permutation problem. An approach similar to EEND-VC, named EEND with global and local attractors (EEND-GLA), is proposed in Horiguchi et al. (2021b) which combines EEND-EDA and unsupervised clustering to deal with cases where the number of speakers appearing during inference is higher than that during training. Another system (Zeghidour et al., 2021) employs a different architecture that

iteratively builds embeddings for each speaker which are exploited to condition a VAD module.

An alternative framework to deal with overlapped speech is continuous speech separation (CSS) (Chen et al., 2020b; Morrone et al., 2022b). CSS extends PIT-based SSep to long recording scenarios, by applying separation in a chunk-wise manner, where each chunk is assumed to contain a fixed number of speakers (usually 2-3). Since the underlying separator is trained via a PIT objective, output permutation consistency between chunks is not guaranteed. CSS solves this problem by performing overlapping inference (i.e., using strides shorter than chunk sizes) and reordering adjacent chunks based on a similarity measure over the portion in which they overlap. Several recent works have proposed diarization systems inspired by CSS. In Raj et al. (2021a) and Xiao et al. (2021) separation is done in windowed segments. In this case, speaker permutation is addressed by applying a diarization method (e.g., clustering-based) across the separated audio streams. On the other hand, Fang et al. (2021) proposes speech separation guided diarization (SSGD), where diarization is performed by first separating the input mixture and applying a conventional VAD to detect speech segments in each channel. In Morrone et al. (2022a), we improve the SSGD architecture using a neural-based VAD and a novel post-processing algorithm that removes the channel leakage generated by separation. In addition, SSGD is adapted to allow online inference employing causal SSep and VAD models.

All the aforementioned approaches, with the exception of Morrone et al. (2022a), are not suitable for streaming applications (e.g., live captioning) as they only work offline. To deal with online processing, Xue et al. (2021b) extends the SA-EEND with an STB mechanism. Inference is performed on short chunks and the speaker permutation information is selected from input-output pairs of previous frames and stored in a buffer. The permutation ambiguity is solved using the buffered frames to condition the output of the current chunk. Xue et al. (2021a) and Horiguchi et al. (2023) propose similar STB-based extensions for EEND-EDA and EEND-GLA, respectively. Instead, Han et al. (2021) improves the online EEND-EDA using chunk-level recurrence to process the chunk hidden states making the model complexity linear in time. Coria et al. (2021) designs a different approach that combines the use of EEND with an x-vector extractor and online clustering. The EEND model is used to gate the representation before the x-vector statistical pooling layer, to extract per-speaker embeddings even in overlap regions.

2. SSGD Framework

Our SSGD pipeline is shown in Fig. 1 and consists of three modules: speech separation, voice activity detection and leakage removal. The system is fed with a single-channel mixed audio input, denoted $\mathbf{Y} \in \mathbb{R}^{1 \times T}$, where T is the number of audio samples.

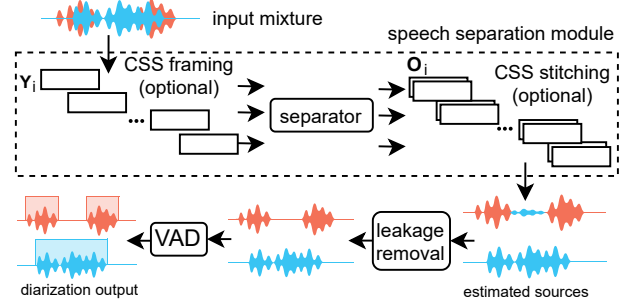


Figure 1: General diagram of the SSGD method.

2.1. Speech Separation Module

In our experiments we employ causal separation models as SSep modules (i.e., Conv-TasNet (Luo and Mesgarani, 2019), DPTNet (Chen et al., 2020a) and DPRNN (Luo et al., 2020)). A streaming modified version of DPTNet is attained in this work by simply masking the future frames in each self-attention layer in the inter-processing blocks, with the rest being equal to causal DPRNN (DPTNet is identical to DPRNN except for self-attention layers).

To compare the proposed SSGD with both clustering-based and EEND state-of-the-art offline systems, we also experiment with non-causal SSep models. Additionally, we consider the use of CSS with non-causal SSep models. In such a configuration, the latency of these models is tied to the CSS window size and thus making offline SSep models suitable for streaming inference. We do not apply CSS in causal SSep models as they do not require chunk-wise processing to work online.

CSS is composed of three stages: framing, separation and stitching. In the framing stage, a windowing operation splits \mathbf{Y} into I overlapped frames $\mathbf{Y}_i \in \mathbb{R}^{1 \times W}$, $i = 1, \dots, I$, with $I = \lceil \frac{T}{H} \rceil$, where W and H are the window and hop sizes, respectively. Each frame \mathbf{Y}_i is fed to the separator, which generates C separated output frames $\mathbf{O}_i \in \mathbb{R}^{C \times W}$. C is the maximum number of speakers in each frame. In this work, C is fixed to 2 as it is a common assumption for telephone conversations. The output channels could be misaligned due to permutation-free training. The stitching module solves the permutation ambiguity by aligning the channels of two consecutive separated outputs \mathbf{O}_i and \mathbf{O}_{i+1} . The correct alignment is estimated according to the cross-correlation between the overlapped portion of the consecutive frames. Finally, the aligned outputs are merged into the final output stream $\mathbf{X} \in \mathbb{R}^{C \times T}$ by using an overlap-add method with Hanning window.

2.2. Voice Activity Detection Module

The VAD module is fed with the estimated separated sources and detects active speech segments. VAD is applied on each separated source $\tilde{\mathbf{X}}_\ell$ independently and the outputs are combined to produce the diarization output. We investigate the use of two different models: an energy-based

conventional VAD (Landini et al., 2021) and a neural-based VAD which employs a temporal convolutional network (TCN) (Bai et al., 2018), as proposed in Cornell et al. (2022). The first method does not require additional training, whilst the latter is data driven.

2.3. Leakage Removal Module

State-of-the-art SSep models have reached impressive performance when tested on short fully overlapped utterances (Wang and Chen, 2018). However, such models could generate channel leakage in sparsely overlapped conversational speech when only one speaker is active (Xiao et al., 2021). Thus, the VAD module can detect as speech the "leaked" segments. This negatively affects the diarization output by introducing false alarm errors. We propose a lightweight post-processing algorithm to mitigate this problem. It does not introduce additional latency and has very little impact on separation quality, missed speech and speaker confusion errors.

The leakage removal post-processing algorithm is summarized in Algorithm 1. The algorithm is fed with an input mixture \mathbf{Y} and two estimated sources $\mathbf{X}^1, \mathbf{X}^2$ which are split into disjoint segments $\mathbf{Y}_\ell, \mathbf{X}_\ell^1, \mathbf{X}_\ell^2$ of length L . For each segment, we compute the Scale-Invariant Signal-to-Distortion Ratios (SI-SDR) (Le Roux et al., 2019) s_ℓ^1, s_ℓ^2 between segments of every source $\mathbf{X}_\ell^1, \mathbf{X}_\ell^2$ with the associated segment \mathbf{Y}_ℓ of input mixture. A leaked segment is detected when both s_ℓ^1, s_ℓ^2 are above a threshold $t_{\ell r}$. In the SSGD framework leakage is removed by filling with zeros the segments with lower SI-SDR. Then, the post-processed estimated sources $\tilde{\mathbf{X}}_\ell$ are passed as input to the following VAD module.

Algorithm 1 Leakage Removal

Input: $\mathbf{Y}, \mathbf{X}^1, \mathbf{X}^2, T, L, t_{\ell r}$

Output: $\tilde{\mathbf{X}}_\ell^1, \tilde{\mathbf{X}}_\ell^2$

$\tilde{\mathbf{X}}_\ell^1 \leftarrow \mathbf{X}_\ell^1; \tilde{\mathbf{X}}_\ell^2 \leftarrow \mathbf{X}_\ell^2$

for $i \leftarrow 0$ **to** T **by** L **do**

$s_\ell^1 \leftarrow \text{SI-SDR}(\mathbf{Y}[i:i+L], \mathbf{X}^1[i:i+L])$

$s_\ell^2 \leftarrow \text{SI-SDR}(\mathbf{Y}[i:i+L], \mathbf{X}^2[i:i+L])$

if $s_\ell^1 > t_{\ell r}$ **and** $s_\ell^2 > t_{\ell r}$ **then**

if $s_\ell^1 > s_\ell^2$ **then**

$\tilde{\mathbf{X}}_\ell^2[i:i+L] \leftarrow 0$

else

$\tilde{\mathbf{X}}_\ell^1[i:i+L] \leftarrow 0$

2.4. End-to-End Training

In our previous work (Morrone et al., 2022a) diarization is performed by cascading the SSep, leakage removal and VAD modules. In such a case, the SSep and VAD models are trained independently. We refer to this configuration as *disjoint SSGD*.

In the *end-to-end SSGD* SSep and VAD are jointly optimized, while removing the leakage removal during training. The SSep and VAD modules are initialized with the

parameters of the disjoint SSGD and fine-tuned following two methods. In the first approach, i.e., *VAD fine-tuning*, we freeze the SSep model and only optimize the VAD parameters. Instead, in the *SSep+VAD fine-tuning* method all parameters are optimized jointly.

Also for end-to-end fine-tuned SSGD systems we apply the proposed leakage removal post-processing algorithm. However, for end-to-end SSGD we apply the masking directly on the VAD output to avoid changing the distribution of estimated sources. This is because we found that doing so could degrade the VAD performance after end-to-end adaptation, since it was fine-tuned on the output of the SSep module.

3. Experimental Setup

3.1. Datasets

Because the focus of our work is on the CTS scenario, we use the *Fisher Corpus Part 1* and *Part 2* (Cieri et al., 2004) for both training and test purposes. The whole Fisher consists of 11699 telephone conversations between two participants, totaling approximately 1960 hours of English speech sampled at 8 kHz. The amount of overlapped speech is around 14% of the total speech duration. Since separated signals for the two speaker are provided, we can use this dataset to train and evaluate a separation model using common metrics as the SI-SDR improvement (SI-SDRi) (Le Roux et al., 2019). We split the data in 11577, 61 and 61 conversations for training, validation and test sets respectively, assuring that there is no overlap between speaker identities.

In addition, we generate a simulated fully-overlapped version of Fisher for the purpose of pre-training the SSep models, as done in Morrone et al. (2022b). This portion is derived from the training set and amounts to 30000 mixtures for a total of 44 hours.

To compare with state-of-the-art EEND methods, we also evaluate the proposed algorithms on the portion of the 2000 NIST SRE (Przybocki and Alvin, 2001) denoted as *CALLHOME*. It consists of real telephone conversations in multiple languages. We use the same 2-speaker subset of CALLHOME and the adaptation/test split proposed in Fujita et al. (2019a). In this case, the amount of overlapped speech is around 13% of the total speech duration.

Since separated sources are not available for CALLHOME, we create simulated conversations which are used to adapt SSep models to CALLHOME data. We exploit the provided annotations to extract single-speaker segments from recordings taken by the original adaptation set, discarding segments shorter than 0.1 s. Then, the extracted segments are combined to mimic real-world conversational scenarios similar to SparseLibriMix (Cosentino et al., 2020). Each conversation is created by alternately picking utterances from the two speakers, until a total minimum length of 30 s is reached. To increase variability, we also mix speakers belonging to different recordings. With this procedure, we generate an additional training and validation sets of 3000

(27.3 h) and 500 (4.5 h) examples, respectively. In this case, speakers overlap approximately 16% of the time.

3.2. Architecture, Training and Inference Details

We experiment with SSep architectures both in online and offline settings. In particular, we consider 3 different SSep models: Conv-TasNet, DPTNet and DPRNN. For Conv-TasNet we employ the best hyperparameter configuration proposed by Luo and Mesgarani (2019). Instead, for DPRNN and DPTNet we use the hyperparameters proposed in Luo et al. (2020) and Chen et al. (2020a), respectively, with some changes. The kernel size for encoder/decoder is increased to 16 to reduce memory consumption. The chunk and hop sizes are set to 100 and 50 (i.e., 50% overlap). In addition, the global layer normalization is replaced with standard layer normalization for online/causal models. We use the implementations available through the Asteroid toolkit (Pariante et al., 2020). We pre-train SSep models on the simulated fully overlapped Fisher maximizing the SI-SDR function. We observed that the SSep pre-training converged much faster compared to training from scratch on real data. We use Adam optimizer (Kingma and Ba, 2015), batch size 4 and learning rate 10^{-3} . We clip gradients with l_2 norm greater than 5. Then, we fine-tune each SSep model on 60 s long random segments from real Fisher recording using a batch size of 1 and reducing the learning rate to 10^{-4} . For Conv-TasNet and DPRNN, learning rate is halved whether SI-SDR does not improve on the validation set for 10 consecutive epochs. For DPTNet, we employ the learning rate scheduler used in Chen et al. (2020a) with $k_1 = 1$, $k_2 = 4 \cdot 10^{-4}$, $d_{model} = 64$, and $warmup_n = 5n_{epoch}$, where n_{epoch} is the number of training steps performed in a single epoch. If no improvement is observed for 20 epochs on validation, training is stopped. For the CALLHOME dataset, the separators are adapted on the simulated CALLHOME using the same hyperparameters of the fine-tuned models, except for the length of training segments which is set to 30 s.

We employ the TCN-based causal VAD proposed by Cornell et al. (2022). The latency of the VAD is set to 0.1 s, as the hop size of the log-Mel filterbanks input features. This model is trained on the original Fisher Part 1, using each speaker source separately, to classify speech vs. non-speech for each input frame. We use the following weighted binary cross-entropy loss:

$$\mathcal{L}_{vad} = -\frac{1}{N} \sum_{n=1}^N \lambda_s \cdot s_n \cdot \log(\hat{s}_n) + (1 - s_n) \cdot \log(1 - \hat{s}_n), \quad (1)$$

where N , n , s_n and \hat{s}_n are the total number of frames, the frame index, the target speech/non-speech label and the estimated speech probability, respectively. λ_s is a parameter that is adjusted according to training data distribution. We set it to 0.9 for all experiments.

The VAD is trained on 2 s long segments and the batch size is set to 256. We employ the same optimizer, learning rate scheduler, gradient clipping and early stopping policy used for Conv-TasNet and DPRNN. At inference the VAD is applied on estimated separated sources independently.

For each frame, the VAD predictions above a threshold t_v are labeled as speech. Then, the thresholded predictions are smoothed using a median filter and segments shorter than a threshold t_s are removed to mitigate false alarm errors. The median filter length, t_v and t_s parameters are tuned on the Fisher validation set for each SSGD architecture. Likewise, these parameters are tuned on the CALLHOME adaptation set for CALLHOME models. The threshold $t_{\ell r}$ and the segment length L of the leakage removal algorithm are tuned manually and set to 3 and 0.1 s, respectively, to not affect the latency of the VAD.

The end-to-end models (i.e., VAD and SSep+VAD fine-tuning) are trained with the same hyperparameters above, except for the initial learning rate which is set to 10^{-5} . The SSep module is initialized with the model trained on real Fisher and simulated CALLHOME when fine-tuned with Fisher and CALLHOME, respectively. Instead, we initialize the VAD with the model pre-trained on Fisher for both datasets. Finally, all models are fine-tuned using the real datasets to minimize the loss \mathcal{L}_{vad} (cfr. Equation 1).

4. Experiments and Results

Diariation performance is evaluated in terms of diariation error rate (DER) on Fisher and CALLHOME test sets. Following Fujita et al. (2019b), we consider overlapped speech and use a start and end-point collar tolerance of 0.25 s (i.e., *fair evaluation* setup). Moreover, we also report the DERs without collar tolerance (i.e., *full evaluation* setup) obtained by end-to-end training strategies. The evaluation is carried out using the standard NIST *md-eval*¹ (version 22) scoring tool. Since oracle sources are available for the Fisher test set, we also measure the separation capability using the SI-SDRi metric (Le Roux et al., 2019).

4.1. Online Diarization

4.1.1. Disjoint SSGD

The results for online disjoint SSGD (i.e., SSep and VAD are trained separately) diariation models on Fisher are reported in Table 1. Results for CALLHOME are shown in Table 2. In this case, we do not fine-tune neither SSep nor VAD on CALLHOME data. Oracle sources refer to SSGD with oracle SSep, thus with error coming only from the VAD module. Note that oracle evaluation is missing for CALLHOME, as separated sources are not available. We also show the results of EEND on the CALLHOME test set, as reported in their respective original works.

The comparison between the TCN VAD and the energy-based one highlights that, as expected, the former performs overall better. The difference however is not major, highlighting the fact that, if the separator performs well a simple VAD may be sufficient in some instances.

The separation capability of the Conv-TasNet model was very poor, generating large false alarm errors. Indeed, it is limited by its short receptive fields, i.e., ~ 1.5 s, which prevents it to learn long term dependencies. Such a problem

¹<https://github.com/usnistgov/SCTK>

Method	VAD	Leakage Removal	Latency (s)	SI-SDRi	MS	FA	SC	DER
<i>Oracle sources</i>	Energy	\times	0.01	∞	7.3	1.8	0.1	9.2
Conv-TasNet				8.9	9.5	26.9	1.6	38.0
DPTNet				21.6	7.5	2.6	0.8	10.6
DPRNN				22.7	7.5	1.7	0.8	10.0
<i>Oracle sources</i>	TCN	\times	0.01	∞	3.5	1.8	0.1	5.3
Conv-TasNet				8.9	7.4	30.9	5.3	43.6
DPTNet				21.6	4.3	3.2	0.6	8.0
DPRNN				22.7	3.3	3.5	0.7	7.5
Conv-TasNet	TCN	\checkmark	0.01	5.9	7.7	4.3	13.3	25.3
DPTNet			0.1	21.2	4.1	2.2	1.1	7.3
DPRNN			0.1	22.3	3.7	2.6	0.8	7.1

Table 1

Disjoint SSGD: speech separation and diarization results on the Fisher test set in the **online** scenario. Separation is assessed using the SI-SDR (dB) improvements over the input mixtures. Diarization is assessed using diarization error rate (DER), missed speech (MS), false alarm (FA) and speaker confusion errors (SC). Algorithmic latency is reported in seconds. The best results among the proposed techniques are shown in **bold**.

Method	VAD	Leakage Removal	Latency (s)	MS	FA	SC	DER
SA-EEND w/STB (Xue et al., 2021b)	n.a.	n.a.	1				12.5
BW-EDA-EEND (Han et al., 2021)			10				11.8
SA-EEND-EDA w/STB (Xue et al., 2021a)			10				10.0
EEND-GLA w/BW-STB (Horiguchi et al., 2023)			1				<u>9.0</u>
Conv-TasNet	Energy	\times	0.01	7.4	35.3	3.0	45.7
DPTNet			0.1	5.5	6.3	0.8	12.6
DPRNN			0.1	5.7	5.2	1.8	12.7
Conv-TasNet	TCN	\times	0.01	12.2	36.2	5.1	53.4
DPTNet			0.1	6.6	4.5	0.6	11.7
DPRNN			0.1	5.9	3.8	1.7	11.6
Conv-TasNet	TCN	\checkmark	0.01	13.2	6.3	13.0	32.5
DPTNet			0.1	6.3	2.6	1.2	10.0
DPRNN			0.1	6.8	2.2	2.2	11.2

Table 2

Disjoint SSGD: diarization results on the CALLHOME test set in the **online** scenario. Diarization is assessed using diarization error rate (DER), missed speech (MS), false alarm (FA) and speaker confusion errors (SC). Latency of the system is reported in seconds. The best results among proposed techniques are shown in **bold**, and among EEND methods are underlined.

is solved by the DPRNN and DPTNet which can track the speakers for much longer due to LSTMs (coupled with self-attention in DPTNet) in the inter-block. This leads to much better diarization results. These two models reached similar performance on Fisher, whereas DPTNet performed better on CALLHOME when used in conjunction with leakage removal.

Crucially, the proposed leakage removal post-processing consistently improved diarization performance for all SSeg models. We observed the major benefits when leakage removal is used with the TCN-based VAD. Indeed, since it is trained on real Fisher data and not on the output of the separators (disjoint training), it is prone to classify channel leakage as active speech. For Conv-TasNet, the DER was reduced by 42.0% and 39.1% on Fisher and CALLHOME, respectively. However, the diarization accuracy remained relatively low due to poor separation capability. For DPTNet

and DPRNN, the leakage removal almost halved the false alarm errors. We observed a larger improvement with DPTNet as the DER is reduced by 8.8% and 14.5%, as opposed to DPRNN for which the DER improved by 5.3% and 3.4% on Fisher and CALLHOME, respectively.

As a comparison, the current best performing online system on the CALLHOME dataset (i.e., EEND-GLA with block-wise speaker tracing buffer (Horiguchi et al., 2023)) obtains 9.0% DER, which is slightly better than ours but is obtained with higher latency of 1 s. Our approach works with a latency of 0.1 s, making it appealing for applications where strict real-time requirements are very important (e.g., real-time captioning).

4.1.2. End-to-End SSGD

We focus on the DPRNN-based architecture as it performs best on the Fisher dataset. Additionally, we observed

Method	Online	SI-SDRi	Fair Eval.				Full Eval.			
			MS	FA	SC	DER	MS	FA	SC	DER
<i>Oracle sources</i>	n.a.	∞	3.5	1.8	0.1	5.3	7.0	4.1	0.3	11.5
Disjoint SSGD w/LR	✓	22.3	3.7	2.6	0.8	7.1	7.3	5.1	1.1	13.5
VAD fine-tuning		22.7	3.9	1.6	0.7	6.2	7.7	3.6	0.8	12.1
VAD fine-tuning w/LR		22.7	5.3	0.9	0.9	7.1	10.8	2.3	1.0	14.0
SSep+VAD fine-tuning		14.9	1.9	1.6	0.7	4.2	4.3	3.5	0.7	8.5
SSep+VAD fine-tuning w/LR		14.9	2.1	1.4	0.7	4.2	3.9	4.1	0.7	8.7
Disjoint SSGD w/LR	✗	22.8	3.9	2.0	0.2	6.1	7.7	4.1	0.5	12.2
VAD fine-tuning		23.2	4.0	1.3	0.1	5.4	8.0	3.3	0.3	11.7
VAD fine-tuning w/LR		23.2	5.7	0.8	0.1	6.7	11.8	2.2	0.4	14.4
SSep+VAD fine-tuning		17.7	3.1	0.9	0.2	4.1	5.5	2.7	0.2	8.4
SSep+VAD fine-tuning w/LR		17.7	3.2	0.8	0.2	4.2	5.8	2.5	0.2	8.6

Table 3

End-to-end SSGD: separation and diarization results on the Fisher test set. The best results among proposed techniques are shown in **bold**.

Method	Online	Fair Eval.				Full Eval.			
		MS	FA	SC	DER	MS	FA	SC	DER
Disjoint SSGD w/LR	✓	6.8	2.2	2.2	11.2	9.8	10.6	3.1	23.6
Disjoint SSGD w/LR (sim-CH adapt.)		6.7	2.2	0.6	9.5	6.5	11.2	1.5	22.1
VAD fine-tuning		6.8	2.3	0.6	9.8	10.9	6.9	1.3	19.1
VAD fine-tuning w/LR		7.6	1.5	0.8	9.8	12.5	5.3	1.5	19.3
SSep+VAD fine-tuning		6.3	2.1	0.4	8.8	8.8	7.0	0.7	16.5
SSep+VAD fine-tuning w/LR		6.5	1.9	0.4	8.8	9.3	6.5	0.8	16.6
Disjoint SSGD w/LR	✗	6.2	2.6	1.5	10.2	8.5	12.1	2.3	22.9
Disjoint SSGD w/LR (sim-CH adapt.)		5.6	2.9	0.8	9.3	7.7	12.9	1.5	22.1
VAD fine-tuning		7.2	2.1	1.0	10.2	11.3	6.2	1.6	19.1
VAD fine-tuning w/LR		7.6	1.3	1.2	10.0	12.3	5.0	1.8	19.1
SSep+VAD fine-tuning		6.5	2.0	0.7	9.2	9.2	6.2	1.0	16.4
SSep+VAD fine-tuning w/LR		6.7	1.8	0.7	9.2	9.8	5.7	1.0	16.6

Table 4

End-to-end SSGD: diarization results on the CALLHOME test set. The best results among proposed techniques are shown in **bold**.

that the adaptation of DPRNN on simulated CALLHOME is more stable compared to other separators (i.e., Conv-TasNet and DPTNet) leading to better final performance on the real CALLHOME test set.

We experiment with different adaptation and end-to-end training strategies. Contrary to what we have reported for disjoint SSGD models, we also report here the *full evaluation* (no collar evaluation) as it can provide a better estimation of segmentation accuracy.

Table 3 reports the results on the Fisher test set. The two fine-tuning strategies provided meaningful improvements over the online disjoint SSGD with leakage removal. In particular, the VAD and SSep+VAD fine-tuning strategies reduced the *fair/full* DER by 12.7/10.3% and 40.8/37.0%, respectively. Both approaches helped a lot to reduce FA errors. In these cases, the use of the leakage removal algorithm was not useful. It even degraded the performances when used with the VAD fine-tuning. Note that the SSep+VAD fine-tuning strategies also outperformed the evaluation with

oracle sources, meaning that the model was able to estimate “custom” separated sources which resulted in better diarization outputs. The significant improvement in diarization performance was obtained at the cost of lowering separation performance which however still remains acceptable.

The results on the CALLHOME test set are shown in Table 4. The adaptation of the separator on the simulated CALLHOME was very effective to reduce SC errors and then the overall DER. As such, we apply all the end-to-end strategies starting from this SSGD adapted on the simulated CALLHOME. The VAD fine-tuning only improved with the full evaluation. Instead, the SSep+VAD fine-tuning reduced the *fair/full* DER by 7.3/24.9% over the adapted disjoint SSGD.

To the best of our knowledge, the SSep+VAD approach *fair* evaluation outperformed all online state-of-the-art EEND methods on the 2-speaker CALLHOME test set with significantly lower latency, i.e., 0.1 vs 1 or 10 s. Additionally, the SSGD is trained using a dataset of ~1900

Method	VAD	Leakage Removal	SI-SDRi	MS	FA	SC	DER
VBx (Landini et al., 2022)	TCN	n.a.	n.a.	10.0	0.3	0.5	10.8
VBx (Landini et al., 2022)	Kaldi			8.9	0.4	0.9	10.2
+ Overlap assignment (Bullock et al., 2020)	Kaldi			<u>4.4</u>	2.1	0.9	<u>7.4</u>
Spectral clustering (Park et al., 2019)	Kaldi			8.9	0.4	<u>0.2</u>	9.5
+ Overlap assignment (Raj et al., 2021b)	Kaldi			5.2	2.0	<u>0.2</u>	<u>7.4</u>
Oracle sources	Energy	✗	∞	7.3	1.8	0.1	9.2
Conv-TasNet			20.1	7.6	4.1	0.9	12.7
DPTNet			22.2	7.5	2.1	0.4	10.0
DPRNN			23.2	7.5	1.6	0.2	9.3
Oracle sources	TCN	✗	∞	3.5	1.8	0.1	5.3
Conv-TasNet			20.1	4.1	5.2	0.7	10.0
DPTNet			22.2	4.8	2.4	0.1	7.3
DPRNN			23.2	3.4	3.1	0.1	6.5
Conv-TasNet	TCN	✓	19.7	4.9	1.3	1.6	7.9
DPTNet			21.9	5.5	1.4	0.1	7.0
DPRNN			22.8	3.9	2.0	0.2	6.1

Table 5

Disjoint SSGD: speech separation and diarization results on the Fisher test set in the **offline** scenario. The best results among proposed techniques are shown in **bold**, and among baselines are underlined.

Method	VAD	Leakage Removal	MS	FA	SC	DER
VBx (Landini et al., 2022)	TCN	n.a.	7.3	1.9	3.1	12.3
VBx (Landini et al., 2022)	Kaldi		8.3	<u>0.9</u>	2.6	11.7
+ Overlap assignment (Bullock et al., 2020)	Kaldi		5.3	2.5	2.4	10.3
Spectral clustering (Park et al., 2019)	Kaldi		8.3	<u>0.9</u>	5.3	14.5
+ Overlap assignment (Raj et al., 2021b)	Kaldi		5.7	2.7	5.8	14.1
SA-EEND (Fujita et al., 2019b)	n.a.	n.a.	<u>4.0</u>	2.4	<u>0.5</u>	9.5
SA-EEND-EDA (Horiguchi et al., 2020)	n.a.					8.1
EEND + VC (Kinoshita et al., 2021b)	n.a.					7.0
EEND-GLA (Horiguchi et al., 2021b)	n.a.					6.9
DIVE (Zeghidour et al., 2021)	n.a.					<u>6.7</u>
Conv-TasNet	Energy	✗	5.5	4.7	0.7	10.9
DPTNet			5.4	5.3	0.4	11.1
DPRNN			5.5	5.2	0.9	11.6
Conv-TasNet	TCN	✗	6.5	4.4	0.5	11.4
DPTNet			7.0	3.3	0.4	10.7
DPRNN			6.5	4.0	0.7	11.2
Conv-TasNet	TCN	✓	6.1	2.6	0.9	9.6
DPTNet			6.3	2.4	0.8	9.5
DPRNN			6.2	2.6	1.5	10.2

Table 6

Disjoint SSGD: diarization results on the CALLHOME test set in the **offline** scenario. The best results among proposed techniques are shown in **bold**, and among baselines/EEND methods are underlined.

hours of speech, which is about 5x smaller than the ones used to train the state-of-the-art EEND models (i.e., ~10000 hours).

4.2. Offline Diarization

We compare the offline SSGD with both clustering-based and EEND methods. As clustering-based baselines we use VBx (Landini et al., 2022) and spectral clustering (Park et al., 2019), along with their overlap-aware counterparts

(Bullock et al., 2020; Raj et al., 2021b). In this case, we use the publicly available Kaldi ASPIRE VAD (Peddinti et al., 2015) model². For overlap detection, we fine-tune the Pyannote (Bredin et al., 2020) segmentation model³ on the full CALLHOME adaptation set. We tune the hyperparameters for each aforementioned module on the associated validation sets. To perform a fair comparison, we also tested the use

²<https://kaldi-asr.org/models/m4>

³<https://huggingface.co/pyannote/segmentation>

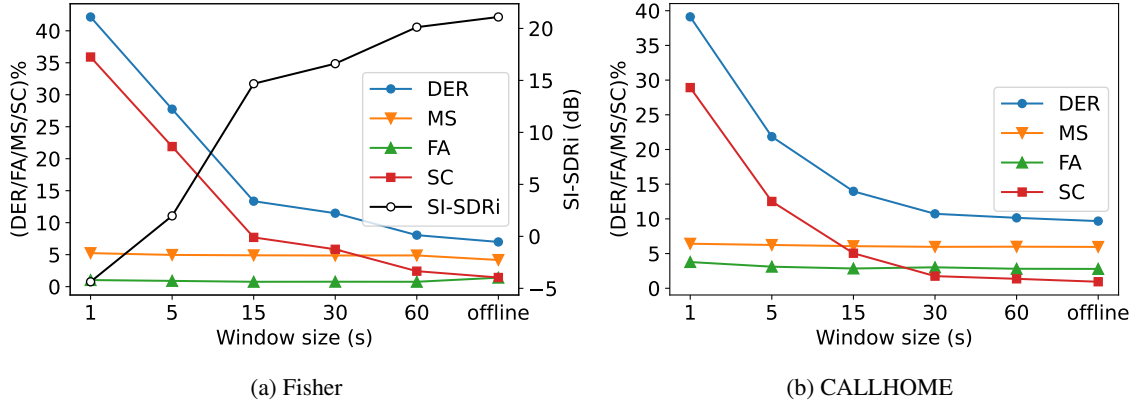


Figure 2: Separation and diarization results on the test sets with different CSS windows. The overlap between windows is set to 50%. The results are obtained with the disjoint SSGD model (DPRNN+TCN+Leakage removal).

of the TCN-based VAD in the VBx system, which however led to lower performance. For CALLHOME, we also report diarization errors of state-of-the-art EEND systems, as done in Table 2.

The results for the offline setting are reported in Tables 5 and 6 for Fisher and CALLHOME, respectively.

Regarding separation performance (SI-SDRi), we can see that the offline DPTNet and DPRNN slightly outperformed their online counterparts. The offline Conv-TasNet was able to obtain good separation capability, resulting in diarization performances similar the DPTNet and DPRNN for CALLHOME. Here, the use of non-causal convolutions and global layer normalization allows the model to track correctly the speakers. However, the dual-path SSep methods outperformed Conv-TasNet on the Fisher test set on all metrics. The overlap-aware VBx resulted the best among the clustering-based baselines, which were all surpassed by our best SSGD systems on the two test sets. Regarding separation performance (SI-SDRi), the offline DPRNN and DPTNet slightly outperformed their online counterparts which confirmed that both dual-path SSep online models are very effective when only past context is available.

As in the online setting, the TCN VAD outperformed the energy-based one and the proposed leakage removal algorithm continued to be useful.

Regarding the proposed adaptation and end-to-end training methods, we can make similar considerations with respect to the online scenario (cfr. Section 4.1.2).

For the CALLHOME data, the best performing offline model is comparable with SA-EEND (Fujita et al., 2019b), although it is not competitive with the current best performing approaches (Kinoshita et al., 2021b; Zeghidour et al., 2021; Horiguchi et al., 2021b), making it less attractive for offline applications. However, it can be a cost-effective solution as the separated signals can be readily used in downstream applications such as ASR. We will show this in Section 4.6.

4.3. CSS Window Analysis

The CSS framework allows the processing of arbitrarily long inputs using chunk-wise processing. We can also exploit CSS to reduce latency of a non-causal SSep model to the CSS window length. In this way, we implement an alternative approach to perform online diarization which employs an offline SSep model in an SSGD framework.

For our purposes, we consider the offline DPRNN-based disjoint SSGD with leakage removal from Table 5 to analyze how varying CSS windows size affects diarization performance.

The results are shown in Fig. 2 for both datasets. Generally, we observed that the main source of error came from speaker confusion which consistently decreased for larger CSS windows, while missed speech and false alarm error rates remained approximately constant. In the presence of longer input recordings the CSS benefited from using longer windows to reduce errors in computing the correct permutation across frames during the stitching stage. Indeed, the correct permutation is computed on the overlapped portion using the cross-correlation, which is more reliable when computed on larger segments. The performances were close to the offline ones for windows larger than 60 and 30 s for Fisher and CALLHOME, respectively. The different slope of DER curves in Fig. 2a and 2b is due to the different average recording duration, which is 10 minutes and 72 s for Fisher and CALLHOME, respectively. This finding also suggests a parallelization strategy for offline CSS. Indeed, using shorter windows (e.g., 30 or 60 s) results in a lower memory footprint and higher inference speed-ups without affecting separation and diarization capabilities.

Compared to the online SSGD approach with causal SSep (Sec. 4.1), the CSS framework requires higher latency to obtain the same performance (i.e. 0.1 vs 30/60 s). On the other hand, it could be an appropriate option in applications in which a lower memory footprint and better ASR accuracy are important requirements rather than low latency, especially for very long recordings (e.g., > 10 minutes).

End-to-End Integration of Speech Separation and Voice Activity
Detection for Low-Latency Diarization of Telephone Conversations

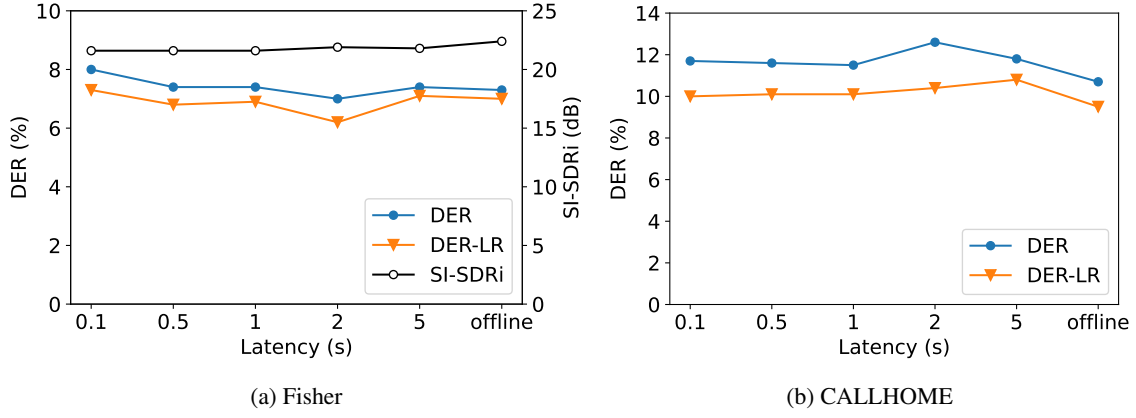


Figure 3: Separation and diarization results on the test sets with the online DPTNet-based SSGD by varying latency.

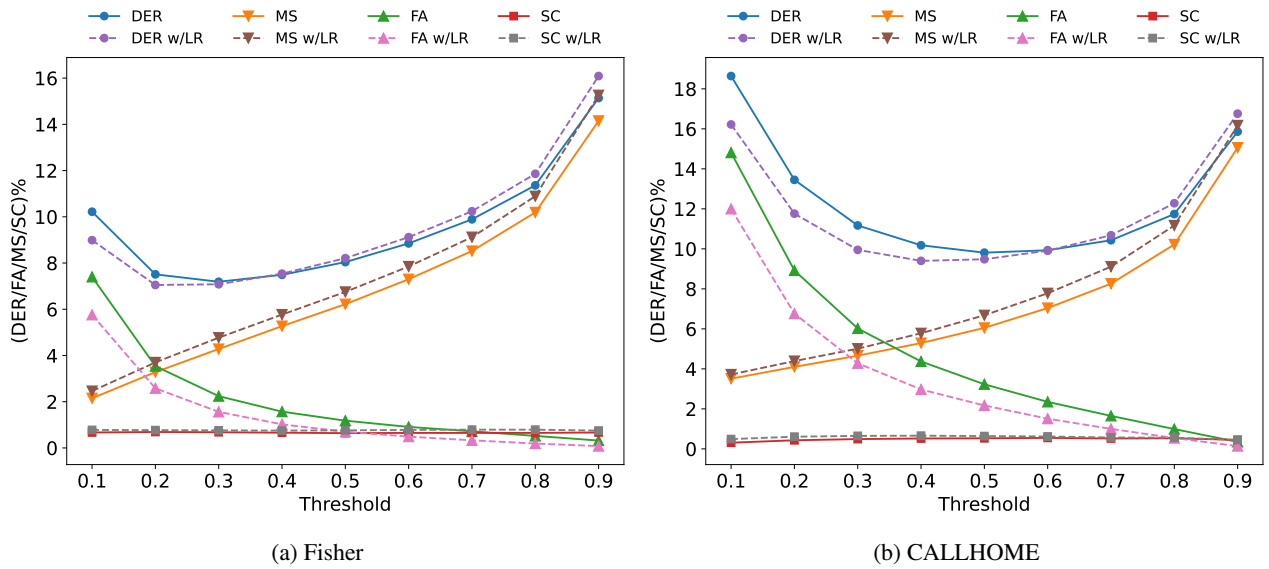


Figure 4: Diarization results of the disjoint SSGD with and without leakage removal on the test sets by varying VAD threshold.

4.4. Latency Analysis with DPTNet

We employ the online DPTNet-based SSGD to study how diarization performance changes by varying model latency from 0.1 to 5 s using the inter-chunk self-attention layers. In this way, the additional latency allows the separator to exploit more future context which may improve separation and diarization capabilities. We do not experiment with the online DPRNN-based model as it is not possible to modify latency while keeping the same hyper-parameters (e.g., intra-chunk size) thus performing a fair analysis. The results are shown in Fig. 3. For both Fisher and CALLHOME test sets, there is not a clear trend meaning that DPTNet is not able to take much advantage of more future information at least till 5 s lookahead. For this reason, it is always convenient to use the model with the lowest latency (i.e., 0.1 s) or the full offline system when streaming processing is not required.

4.5. Leakage Removal Analysis

We perform a study in order to analyze the effect of the leakage removal algorithm on disjoint and end-to-end SSGD. In particular, we employ the same DPRNN-based model by varying the VAD threshold t_v .

Fig. 4 shows the results of the disjoint SSGD on the Fisher and CALLHOME datasets. As expected, lower thresholds generated higher FA and lower MS errors, whilst higher thresholds produced the opposite behavior (i.e., lower FA and higher MS). On the other hand, SC remained roughly constant. The use of leakage removal consistently reduced DER and FA when the VAD threshold is below a value depending on the dataset. When the threshold is above that value, FA errors are very close to zero and the higher degradation of MS resulted in higher DER. On the CALLHOME dataset, the leakage removal algorithm was way more effective compared to Fisher. Indeed, the VAD is only trained on Fisher data, and leakage removal was able to

End-to-End Integration of Speech Separation and Voice Activity
Detection for Low-Latency Diarization of Telephone Conversations

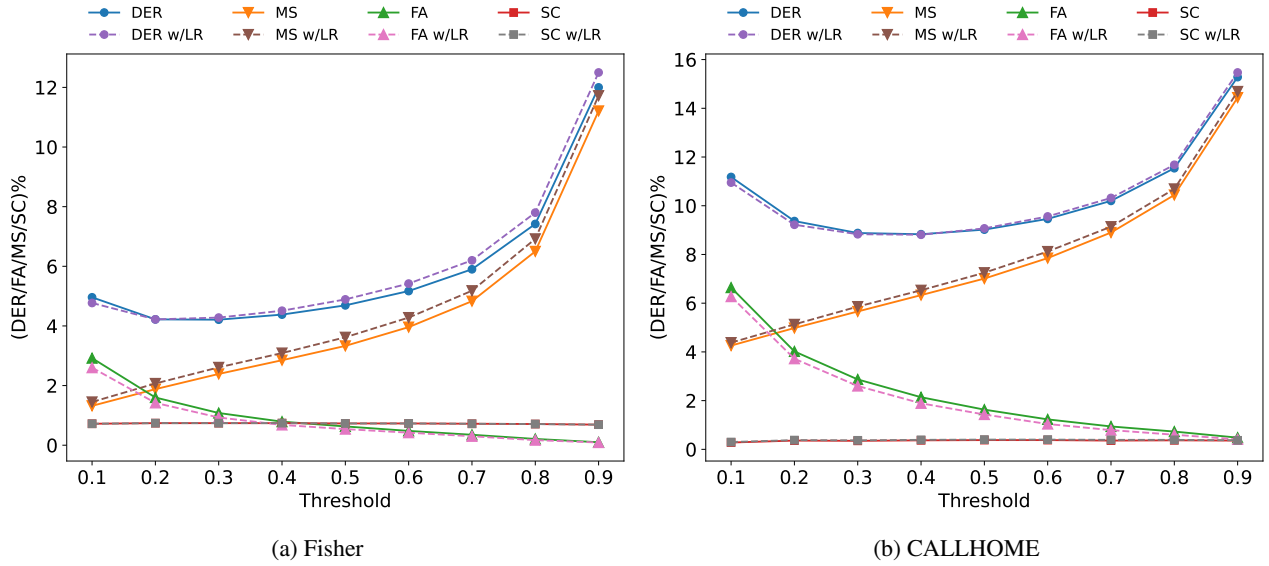


Figure 5: Diarization results of the end-to-end SSGD with and without leakage removal on the test sets by varying VAD threshold.

mitigate the mismatch between the data distributions of the two datasets.

The results of the end-to-end SSGD are depicted in Fig. 5. Contrary to the disjoint SSGD, the leakage removal did not improve overall diarization performance not even for low thresholds. This demonstrated that the end-to-end training was able to mitigate channel leakage as the VAD is updated with the separator output. In this case, the use of leakage removal would be a convenient choice only when having lower FA is desirable.

4.6. ASR Evaluation

The SSGD framework outputs both separated sources and segmentation which can be readily fed in input to a back-end ASR system. It is a great advantage over other diarization methods. We investigate the ASR performance when estimated sources from DPRNN models are fed to a downstream ASR, considering or not leakage removal. We also analyze the effect of the TCN-based VAD compared to oracle segmentation. We use the pre-trained Kaldi ASiPiRE ASR (Povey et al., 2011) model⁴ and report the performance in terms of word error rate (WER). Additionally, we also report the results obtained with input mixture and oracle sources, which represent the upper and lower bound for ASR evaluation.

The results are reported in Table 7. The degradation of all SSGD systems, except for SSep+VAD fine-tuning, was very small compared to the evaluation with oracle sources. Additionally, the SSGD obtained large improvements over the mixtures. These findings confirm the effectiveness of the proposed SSep methods. In general, the significant gap between the proposed model with the fully oracle system (i.e., oracle VAD + oracle sources) demonstrated that the VAD segmentation represents the main source of error. This

⁴<https://kaldi-asr.org/models/m1>

Method	Online	VAD	
		TCN	Oracle
Mixture	n.a.	38.74	30.69
Oracle sources		25.44	19.50
DPRNN	✓	26.28	20.63
+ Leakage removal		26.67	21.12
+ Leakage removal (seg-only)		26.23	20.63
+ VAD fine-tuning		26.04	20.63
+ SSep+VAD fine-tuning		31.82	29.29
DPRNN	✗	25.77	19.98
+ Leakage removal		26.22	20.34
+ Leakage removal (seg-only)		25.79	19.98
+ VAD fine-tuning		25.65	19.98
+ SSep+VAD fine-tuning		27.48	21.79

Table 7

WER evaluation on the Fisher test set. The best online/offline non-oracle results are reported in **bold**.

finding is consistent with diarization results where the errors mainly came from MS and FA.

Although the leakage removal post-processing generally improves diarization, the filled zeros could produce less natural utterances which negatively affect the WER. On the other hand, in the proposed framework it could be only exploited for obtaining the segmentation using the non-processed estimated sources (+ *Leakage removal (seg-only)*). In this latter case the DER is slightly reduced, as well for the VAD fine-tuning. Indeed, better segmentation improves performance when used with the same separated sources. The lower separation capability of models (cfr. Table 3) trained with the SSep+VAD fine-tuning resulted in higher WER as, during fine-tuning, there are no guarantees the output of the separator will be distortion-free, and even

subtle distortions, while minimal in terms of SI-SDR, are known to affect ASR models significantly (von Neumann et al., 2020). Adding distortion-free constraints while fine-tuning, such as multi-frame minimum variance distortion-less response (Tammen and Doclo, 2021) is a possible future direction.

5. Conclusion and Future Work

In this paper, we have explored end-to-end integration of SSGD components, i.e. the speech separator and VAD. We have focused on low-latency models which allow diarization for arbitrarily long inputs in a frame-online fashion. The proposed fine-tuning strategies showed significant improvements on both Fisher and CALLHOME datasets compared to the system without fine-tuning. In particular, our best online model, i.e., SSep+VAD fine-tuning, outperformed the current state-of-the-art methods based on EEND on the CALLHOME dataset with an order of magnitude lower latency (i.e. 0.1 s vs. 1 s).

Additionally, we have extended our previous work by performing a more comprehensive analysis of SSGD in real-world telephone conversation scenarios. We have experimented with another SSep model, i.e., DPTNet, which can be easily adapted to perform online inference with desired latency. Experimental results demonstrated that additional latency did not improve diarization accuracy, thus it was always convenient to use the model with the lowest latency. We have also generated a simulated dataset for the purpose of adaptation of SSep models to CALLHOME data. Adaptation on such data provided significant improvements, especially for DPRNN. The use of the proposed leakage removal algorithm was investigated both for disjoint and SSep+VAD fine-tuned SSGD. Experiments clearly showed that the post-processing algorithm was very effective to reduce both FA and DER in disjoint SSGD. However, when used in conjunction with fine-tuned models it did not improve the overall performance as the models learned to handle leaked segments directly from data. Finally, we have demonstrated that estimated sources generated by SSGD could be readily fed in input to an ASR system. The ASR performance largely improved over the input mixture and in some instance was even close to the one obtained with oracle sources. However, the end-to-end integration led to significant ASR performance degradation.

In future works, the SSGD framework will be extended to domains in which a higher number of speakers is typically involved (e.g., meeting scenarios). This will need likely the development of new techniques for speech separation. One of the major issues of having more than two speakers is that the speaker tracking for arbitrarily long inputs is significantly more difficult, potentially leading to confusion errors. Another issue is that the channel leakage problem also becomes more challenging as the number of output streams grows (e.g. 4 channels for a max of 4 local speakers). This latter can be addressed by adopting separation methods which can handle a variable number of output channels Nachmani et al. (2020). At the same time, novel strategies should

be investigated to mitigate the separation/ASR degradation when using end-to-end fine-tuned SSGD models. This includes distortion-less constraints and/or continual learning approaches.

6. Acknowledgements

This work has been supported by the AGEVOLA project (SIME code 2019.0227), funded by the Fondazione CAR-ITRO.

References

- Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., Vinyals, O., 2012. Speaker diarization: A review of recent research. *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing* 20, 356–370.
- Bai, S., Kolter, J.Z., Koltun, V., 2018. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.
- Bredin, H., Laurent, A., 2021. End-to-end speaker segmentation for overlap-aware resegmentation, in: *Proc. of Interspeech, ISCA*. pp. 3111–3115.
- Bredin, H., Yin, R., Coria, J.M., Gelly, G., Korshunov, P., et al., 2020. Pyannote.audio: neural building blocks for speaker diarization, in: *Proc. of ICASSP, IEEE*. pp. 7124–7128.
- Bullock, L., Bredin, H., Garcia-Perera, L.P., 2020. Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection, in: *Proc. of ICASSP, IEEE*. pp. 7114–7118.
- Chen, J., Mao, Q., Liu, D., 2020a. Dual-Path Transformer Network: Direct Context-Aware Modeling for End-to-End Monaural Speech Separation, in: *Proc. of Interspeech, ISCA*. pp. 2642–2646.
- Chen, Z., Yoshioka, T., Lu, L., Zhou, T., Meng, Z., et al., 2020b. Continuous speech separation: Dataset and analysis, in: *Proc. of ICASSP, IEEE*. pp. 7284–7288.
- Cieri, C., Miller, D., Walker, K., 2004. The Fisher Corpus: A resource for the next generations of speech-to-text., in: *LREC*. pp. 69–71.
- Coria, J.M., Bredin, H., Ghannay, S., Rosset, S., 2021. Overlap-aware low-latency online speaker diarization based on end-to-end local segmentation, in: *Proc. of ASRU, IEEE*. pp. 1139–1146.
- Cornell, S., Omologo, M., Squartini, S., Vincent, E., 2022. Overlapped speech detection and speaker counting using distant microphone arrays. *Computer Speech & Language* 72, 101306.
- Cosentino, J., Pariente, M., Cornell, S., Deleforge, A., Vincent, E., 2020. Librimix: An open-source dataset for generalizable speech separation. *arXiv preprint arXiv:2005.11262*.
- Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P., 2010. Front-end factor analysis for speaker verification. *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing* 19, 788–798.
- Desplanques, B., Thienpondt, J., Demuynck, K., 2020. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification, in: *Proc. of Interspeech, ISCA*. pp. 3830–3834.
- Fang, X., Ling, Z.H., Sun, L., Niu, S.T., Du, J., et al., 2021. A deep analysis of speech separation guided diarization under realistic conditions, in: *Proc. of APSIPA ASC*. pp. 667–671.
- Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., Watanabe, S., 2019a. End-to-end neural speaker diarization with permutation-free objectives, in: *Proc. of Interspeech, ISCA*. pp. 4300–4304.
- Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., Watanabe, S., 2019b. End-to-end neural speaker diarization with self-attention, in: *Proc. of ASRU, IEEE*. pp. 296–303.
- Fujita, Y., Watanabe, S., Horiguchi, S., Xue, Y., Nagamatsu, K., 2020. End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification. *arXiv preprint arXiv:2003.02966*.
- Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., McCree, A., 2017. Speaker diarization using deep neural network embeddings, in: *Proc. of ICASSP, IEEE*. pp. 4930–4934.

- Han, E., Lee, C., Stolcke, A., 2021. BW-EDA-EEND: Streaming end-to-end neural speaker diarization for a variable number of speakers, in: Proc. of ICASSP, IEEE. pp. 7193–7197.
- Horiguchi, S., Fujita, Y., Watanabe, S., Xue, Y., Nagamatsu, K., 2020. End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors, in: Proc. of Interspeech, ISCA. pp. 269–273.
- Horiguchi, S., Garcia, P., Fujita, Y., Watanabe, S., Nagamatsu, K., 2021a. End-to-end speaker diarization as post-processing, in: Proc. of ICASSP, IEEE. pp. 7188–7192.
- Horiguchi, S., Watanabe, S., García, P., Xue, Y., Takashima, Y., Kawaguchi, Y., 2021b. Towards neural diarization for unlimited numbers of speakers using global and local attractors, in: Proc. of ASRU, IEEE. pp. 98–105.
- Horiguchi, S., Watanabe, S., García, P., Takashima, Y., Kawaguchi, Y., 2023. Online neural diarization of unlimited numbers of speakers using global and local attractors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31, 706–720.
- Horiguchi, S., Yalta, N., García, P., Takashima, Y., Xue, Y., et al., 2021c. The Hitachi-JHU DIHARD III system: Competitive end-to-end neural diarization and x-vector clustering systems combined by DOVER-Lap. *arXiv preprint arXiv:2102.01363*.
- Huang, Z., Watanabe, S., Fujita, Y., García, P., Shao, Y., et al., 2020. Speaker diarization with region proposal network, in: Proc. of ICASSP, IEEE. pp. 6514–6518.
- Jung, J.w., Heo, H.S., Kwon, Y., Chung, J.S., Lee, B.J., 2021. Three-Class Overlapped Speech Detection Using a Convolutional Recurrent Neural Network, in: Proc. of Interspeech, ISCA. pp. 3086–3090.
- Kingma, D.P., Ba, J., 2015. Adam: A method for stochastic optimization, in: Proc. of ICLR.
- Kinoshita, K., Delcroix, M., Tawara, N., 2021a. Advances in Integration of End-to-End Neural and Clustering-Based Diarization for Real Conversational Speech, in: Proc. of Interspeech, ISCA. pp. 3565–3569.
- Kinoshita, K., Delcroix, M., Tawara, N., 2021b. Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds, in: Proc. of ICASSP, IEEE. pp. 7198–7202.
- Kolbæk, M., Yu, D., Tan, Z., Jensen, J., 2017. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing* 25, 1901–1913.
- Koluguri, N.R., Park, T., Ginsburg, B., 2022. TitaNet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context, in: Proc. of ICASSP, IEEE. pp. 8102–8106.
- Landini, F., Glembek, O., Matějka, P., Rohdin, J., Burget, L.e.a., 2021. Analysis of the BUT diarization system for VoxConverse challenge, in: Proc. of ICASSP, IEEE. pp. 5819–5823.
- Landini, F., Profant, J., Diez, M., 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, 101254.
- Le Roux, J., Wisdom, S., Erdogan, H., Hershey, J.R., 2019. SDR-half-baked or well done?, in: Proc. of ICASSP, IEEE. pp. 626–630.
- Luo, Y., Chen, Z., Yoshioka, T., 2020. Dual-Path RNN: Efficient long sequence modeling for time-domain single-channel speech separation, in: Proc. of ICASSP, IEEE. pp. 46–50.
- Luo, Y., Mesgarani, N., 2019. Conv-TasNet: Surpassing ideal time-frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing* 27, 1256–1266.
- Medennikov, I., Korenevsky, M., Prisyach, T., Khokhlov, Y.Y., Korenevskaya, M., et al., 2020. Target-speaker voice activity detection: a novel approach for multi-speaker diarization in a dinner party scenario, in: Proc. of Interspeech, ISCA. pp. 274–278.
- Morrone, G., Cornell, S., Raj, D., Serafini, L., Zovato, E., Brutti, A., Squartini, S., 2022a. Low-latency speech separation guided diarization for telephone conversations, in: Proc. of SLT, IEEE. pp. 641–646.
- Morrone, G., Cornell, S., Zovato, E., Brutti, A., Squartini, S., 2022b. Conversational speech separation: an evaluation study for streaming applications, in: Proc. of AES Convention 152, AES.
- Nachmani, E., Adi, Y., Wolf, L., 2020. Voice separation with an unknown number of multiple speakers, in: Proc. of ICML, PMLR. pp. 7164–7175.
- von Neumann, T., Kinoshita, K., Drude, L., Boeddeker, C., Delcroix, M., Nakatani, T., Haeb-Umbach, R., 2020. End-to-end training of time domain audio separation and recognition, in: Proc. of ICASSP, IEEE. pp. 7004–7008.
- Pariente, M., Cornell, S., Cosentino, J., Sivasankaran, S., Tzinis, E., et al., 2020. Asteroid: The PyTorch-Based Audio Source Separation Toolkit for Researchers, in: Proc. of Interspeech, ISCA. pp. 2637–2641.
- Park, T.J., Han, K.J., Kumar, M., Narayanan, S., 2019. Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap. *IEEE Signal Processing Letters* 27, 381–385.
- Park, T.J., Kanda, N., Dimitriadis, D., Han, K.J., Watanabe, S., Narayanan, S., 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language* 72, 101317.
- Peddinti, V., Chen, G., Manohar, V., Ko, T., Povey, D., Khudanpur, S., 2015. JHU ASPIRE system: Robust LVCSR with TDNNs, iVector adaptation and RNN-LMS, in: Proc. of ASRU, IEEE. pp. 539–546.
- Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al., 2011. The Kaldi speech recognition toolkit, in: Proc. of ASRU, IEEE.
- Przybocki, M., Alvin, M., 2001. 2000 NIST Speaker Recognition Evaluation LDC2001S9. URL: <https://catalog.ldc.upenn.edu/LDC2001S97>.
- Raj, D., Denisov, P., Chen, Z., Erdogan, H., Huang, Z., et al., 2021a. Integration of speech separation, diarization, and recognition for multi-speaker meetings: System description, comparison, and analysis, in: Proc. of SLT, IEEE. pp. 897–904.
- Raj, D., Huang, Z., Khudanpur, S., 2021b. Multi-class spectral clustering with overlaps for speaker diarization, in: Proc. of SLT, IEEE. pp. 582–589.
- Singh, P., Ganapathy, S., 2021. Self-supervised metric learning with graph clustering for speaker diarization, in: Proc. of ASRU, IEEE. pp. 90–97.
- Subakan, C., Ravanelli, M., Cornell, S., Grondin, F., 2022. REAL-M: Towards speech separation on real mixtures, in: Proc. of ICASSP, IEEE. pp. 6862–6866.
- Tammen, M., Doclo, S., 2021. Deep multi-frame MVDR filtering for single-microphone speech enhancement, in: Proc. of ICASSP, IEEE. pp. 8443–8447.
- Wang, D., Chen, J., 2018. Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Acoustics, Speech, and Signal Processing* 26, 1702–1726.
- Watanabe, S., Mandel, M., Barker, J., Vincent, E., Arora, A., Chang, X., Khudanpur, S., Manohar, V., Povey, D., Raj, D., et al., 2020. CHiME-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings, in: Proc. of 6th International Workshop on Speech Processing in Everyday Environments, pp. 1–7.
- Xiao, X., Kanda, N., Chen, Z., Zhou, T., Yoshioka, T., et al., 2021. Microsoft speaker diarization system for the VoxCeleb speaker recognition challenge 2020, in: Proc. of ICASSP, IEEE. pp. 5824–5828.
- Xue, Y., Horiguchi, S., Fujita, Y., Takashima, Y., Watanabe, S., Garcia, P., Nagamatsu, K., 2021a. Online streaming end-to-end neural diarization handling overlapping speech and flexible numbers of speakers, in: Proc. of Interspeech, ISCA. pp. 3116–3120.
- Xue, Y., Horiguchi, S., Fujita, Y., Watanabe, S., García, P., Nagamatsu, K., 2021b. Online end-to-end neural diarization with speaker-tracing buffer, in: Proc. of SLT, IEEE. pp. 841–848.
- Zeghidour, N., Teboul, O., Grangier, D., 2021. DIVE: End-to-end speech diarization via iterative speaker embedding, in: Proc. of ASRU, IEEE. pp. 702–709.