INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Online Speaker Diarization Equipped with Discriminative Modeling and Guided Inference

Xucheng Wan, Kai Liu, Huan Zhou

Artificial Intelligence Application Research Center, Huawei Technologies
Shenzhen, PRC

wanxucheng@huawei.com, liukai89@huawei.com, zhou.huan@huawei.com

Abstract

Despite considerable efforts, online speaker diarization remains an ongoing challenge. In this study, we propose to tackle the challenge from two perspectives, to endow diarization model with discriminability and to rectify less-reliable online inference with guidance. Specifically, based on the current prior art, UIS-RNN, two enhancement approaches are proposed to concretize our motivations. The effectiveness of our proposals is experimentally validated by results on the AMI evaluation set. With substantial relative improvement of 48.7%, our online speaker diarization system significantly outperformed its baseline. More impressively, its performance in terms of diarization error rate is better than most state-of-the-art offline systems.

Index Terms: online speaker diarization, discriminative training, guided inference, recurrent neural networks

1. Introduction

Speaker diarization (SD) is a process of partitioning an audio recording with multiple speakers into homogeneous segments according to the speaker identity. That is, it addresses "who spoke when"problem, without any prior knowledge of speakers. As an important part of speech recognition systems, SD is valuable for a variety of downstream applications, such as, to provide speaker identities in field of conversational AI, to enrich meeting transcription with speaker meta information or to retrieve audio/video recording for a target speaker.

Traditionally, a speaker diarization system consists of three core modules: 1) segmentation: to separate audio recording into single speaker segments; 2) speaker embedding: to represent segments with embedding vectors and 3) clustering: to group those embeddings into classes. Apart from them, some optional modules, like front-end processing (to suppress acoustic artifacts), speech activity detection (SAD) (to remove non-speech portion) and post-processing (to refine clustering results), could be integrated to further promote the system performance.

With the advent of deep learning technology, there have been various efforts towards enhancing aforementioned modules, either individually or jointly. For example, speaker embeddings are evolved from traditional i-vector[1] to d-vector[2], x-vector[3] or other deep embeddings[4]; traditional k-means clustering, agglomerative hierarchical clustering (AHC) are replaced by spectral clustering (SC) [5, 6]. Regarding joint module optimization, [7] adopts joint segmentation and clustering, [8] jointly performs segmentation, embedding extraction and re-segmentation, and [9] proposes an end-to-end neural solution.

Although large performance gains are achieved by the neural SD approaches, most of them can only work in an offline manner, with an assumption that entire audio recording is available beforehand. While, in realistic meeting or human-computer conversation scenario, to provide diarization output on audio stream (i.e. online mode) is highly desirable, especially for case of long audio recordings.

Nevertheless, it is a non-trivial issue to perform online SD. Firstly, clustering approach primarily is designed for offline processing, which makes those clustering-based SD approaches not directly applicable to online scenario. An intuitive solution to re-run the clustering with every new coming audio segment would be inefficient and, to make matters worse, bring in an issue of temporal discontinuity among segments (due to label ambiguity). Secondly, the online requirement poses additional challenges, including, but not limited to: 1) unavailable future segment information; 2) low latency constraint for result delivery and 3) inapplicable result refinement. All these lead to noteworthy performance gap between online and offline SD.

Recently, a handful of techniques have been developed to tackle these challenges to enable online SD. These prior works could be roughly categorized into following lines of research.

- re-clustering: As an alternative speaker clustering approach, X-means and an online re-clustering approach is introduced [10] to handle variable number of speakers.
- blockwise online processing: [11] applies cosine similarity threshold to assign cluster labels. An incremental Transformer encoder [12] is utilized that attends to only its left contexts to generate attractors block by block. [13] incrementally performs various encoding and decoding operations per audio block and uses speaker-tracing buffer to circumvent the speaker inconsistency issue. By iteratively updating speaker embedding and speech mask, an all-neural approach [14] can handle variable number of speakers and jointly perform speech separation, speaker counting and diarization.
- supervised clustering: [7] performs joint segmentation and clustering, and it models diarization by a sequence generation process in a supervised manner. Its enhanced version [15] adopts new loss and better modeling of the speaker turns. A discriminative neural clustering model based on transformer is recently proposed [16].

Despite the growing research interests, online SD remains a challenging problem, working only relatively well on a small range of public datasets. In real-time meetings with natural spontaneous conversations, there still have much room for performance improvement [17].

In view of this, this study aims to bridge the performance gap. For a given online SD system, our motivation is to boost its performance by either increasing its discriminativity or control the online inference process based on some global variable. To verify effectiveness of our ideas, we choose UIS-RNN [7] as a backbone, and propose to enhance it from two perspectives: to improve the model during training and to offer guidance for potential unreliable input during inference. Such a baseline choice

http://dx.doi.org/10.21437/Interspeech.2021-261

is due to two-fold considerations: its promising performance and inherent ability of handling variable number of speakers.

The rest of the paper is organized as follows. In section 2, a short overview of UIS-RNN is provided. In section 3, we describe our proposed methods in detail. Experimental results are presented in section 4 and section 5 concludes the paper.

2. Our baseline of UIS-RNN

Before elaborating our proposed methods, we briefly review our baseline system, Unbounded interleaved-state recurrent neural networks (UIS-RNN) [7]. In this work, a trainable system is proposed as a general solution to segment and cluster temporal data in a fully supervised manner. Its key idea is to replace clustering module by an unbounded interleaved-state RNN.

In detail, given an input sequence of speaker embeddings $X = (x_t \in \mathbb{R}^d | t = 1, \cdots, T)$ (provided from other pre-trained embedding extraction module) and corresponding speaker labels $Y = (y_t \in \mathbb{N} | t = 1, \cdots, T)$, the diarization problem is converted into a joint probabilistic framework. Using a conditional speaker chain rule and introducing a latent binary variable z_t, at a given time frame, the conditional probability is decomposed into three product terms, expressed as:

$$p\left(x_t, y_t, z_t | x_{[t-1]}, y_{[t-1]}, z_{[t-1]}\right) = \underbrace{p(x_t|x_{[t-1]}, y_t)}_{\text{sequence generation}} \cdot \underbrace{p(y_t|y_{[t-1]}, z_t)}_{\text{speaker assignment}} \cdot \underbrace{p(z_t|z_{[t-1]})}_{\text{speaker change}} \quad (1)$$

The three probability terms above are respectively modeled by gated recurrent units (GRU)-based RNN, a distance dependent Chinese restaurant process(ddCRP) and Bernoulli distribution.

During training phase, a RNN network aims to learn the mean of distribution of X. Specifically, 1) the network inputs is constructed by speaker embedding sequence X^k from the k-th speaker; 2) the corresponding RNN memory sequence form a so-called speaker profile $\mathcal{M}^k$ (updated by simple moving average); 3) for each input batch, the training loss is accumulated over mean square errors (MSE) between new embedding input x_t^k and historical speaker profile $\mathcal{M}_{t-1}^k$.

The inference process is quite distinct from training. It is performed in an online fashion, by seeking speaker labels that maximize $logP(X, Y)$ based on MAP decoding and beam search algorithm. Experimental results on telephone data show that as an online diarization solution, UIS-RNN outperforms the previous offline state of the art.

3. Proposed Methods

3.1. Motivation

Although UIS-RNN works remarkably well on telephone dataset, this approach has some potential problems.

Firstly, as we can see from the above description, the original UIS-RNN approach makes some specific assumptions on both occurrence of speakers and speaker changing behavior. As pointed out before [15, 16], those assumptions limit the system generalization in unseen testing scenarios where they do not necessarily hold.

Secondly, as a general rule for a system with supervised training, its performance benefits from learning from more examples (as explicitly verified in [7] as well). This suggests that deployment of UIS-RNN might be impeded in low-resource SD application scenarios.

To tackle these problems, our idea is to modify the training loss function with objective of discriminability. Differentiated from prior arts[15], we introduce two additional loss items, incorporated with the original MSE loss (cf., Section 3.2).

Lastly, not all speaker embeddings within an utterance are equally reliable. Some of them may suffer distortions, introduced from non-ideal (noisy background or overlapping speech) acoustic frames. In the context of online SD, such less reliable embeddings could deteriorate performance and they are hard to detect. Bearing that in mind, we circumvent the problem by proposing a heuristic method to yield a so-called confidence factor for each input embedding (cf., Section 3.4). By leveraging these confidence factors for inference, the resulting SD system is expected to be less affected by errors stemmed from potential embedding distortions.

3.2. Discriminative training objective

To alleviate its dependencies on a large amount of training data, the original training objective of UIS-RNN is modified. Motivated by discriminative training, our proposal is to introduce a pair of additional loss items: heterogeneous loss (hetero-loss) and homogeneous loss (homo-loss).

The hetero-loss aims to make the output of GRU-based RNN (i.e. speaker profile) discriminative. However, as aforementioned, the training strategy of UIS-RNN is to develop the RNN based on a set of single speaker sequences. To break the limitation, during RNN training for a single speaker sequence, a synthetic profile is introduced as a negative sample (i.e. profile from different speaker). Such a synthetic profile is conveniently yielded by a random generator. In addition, the proposed hereto-loss simulates the idea of large margin softmax loss [18] for a two-class classification issue, formulated as:

$$L_{\text{hetero}} = -\log\left(\frac{e^{\cos(\beta_1 \theta + \beta_2) - \beta_3}}{e^{\cos(\beta_1 \theta + \beta_2) - \beta_3} + e^{cos(\varphi)} + \epsilon}\right) \quad (2)$$

where θ denotes the angle between embedding X^k and target profile $\mathcal{M}^k$ and; φ denotes the angle between X^k and the synthetic profile $\mathcal{M}^{syn}$; $(\beta_1, \beta_2, \beta_3)$ are manually selected parameters; and ϵ is a small value to avoid a divide-by-zero error.

The homo-loss is designed to add constraint on profile discriminativity, as well. For the target k-th speaker, home-loss minimizes the accumulated angular distances between current observations and historical speaker profile, expressed as:

$$L_{\text{homo}} = < \mathcal{M}^k, X^k > \quad (3)$$

$$= \arccos\left(\frac{\mathcal{M}^k}{\|\mathcal{M}^k\|} \cdot \frac{X^k}{\|X^k\|}\right) \quad (4)$$

With incorporation of the proposed hetero-loss and homo-loss, the overall optimizing loss of our proposed system is shown as following:

$$L_{\text{full}} = L_{\text{ori}} + \alpha_1 L_{\text{hetero}} + \alpha_2 L_{\text{homo}} \quad (5)$$

where L_{ori} represents the loss function inherited from UIS-RNN and α_1 and α_2 are hyper-parameters.

3.3. Global embedding profile $\mathcal{E}$

To support the guided inference, an auxiliary module, called global embedding profile $\mathcal{E}$, is introduced first. It serves to aggregate all rich historical embeddings and produce the most reliable ones, as global embedding profile, which could facilitate the generation of confidence-related cues.

Here the reliability is measured by similarity score. Furthermore, to be free of reliability threshold, scores associated with top P embeddings are also kept as part of a profile.

For a given frame embedding x_t in an audio flow, once the inference system outputs speaker ID as $y_t = k$, then the ID associated specific part, $\mathcal{E}_t^k$ is updated. Here $\mathcal{E}_t^k = \{\bar{E}^k, \bar{D}^k\}$ consists of top P embeddings $\bar{E}^k = \{E_1^k, \cdots, E_P^k\}$ and ranking scores $\bar{D}^k = \{D_1^k, \cdots, D_P^k\}$. The detailed updating algorithm is described by the pseudo codes below.

Algorithm 1 update $\mathcal{E}_t^k$ for given x_t and $y_t = k$

if k is a new speaker **then**
  do initialize:
  $\bar{D}^k = \{D_1^k\}$ where $D_1^k = v$
  $\bar{E}^k = \{E_1^k\}$ where $E_1^k = x_t$
else
  if E_P^k is available **then**
    $D_t = \text{scoring}(x_t, \text{mean}(\bar{E}^k))$
    if $D_t > D_j^k = \min(\bar{D}^k)$ **then**
      do update:
      $\bar{D}^k = \{D_1^k, \cdots, D_{j-1}^k, D_t, D_{j+1}^k, \cdots, D_P^k\}$
      $\bar{E}^k = \{E_1^k, \cdots, E_{j-1}^k, x_t, E_{j+1}^k, \cdots, E_P^k\}$
    end if
  else
    $D_t = \text{scoring}(x_t, \text{mean}(\bar{E}^k))$
    do initialize (or do update):
    $\bar{D}^k = \{\bar{D}^k, D_t\}$
    $\bar{E}^k = \{\bar{E}^k, x_t\}$
  end if
end if

Here v is a pre-defined value closing to 1.0, and scoring is defined as cosine similarity with the following expression:

$$\text{scoring}(a, b) = \frac{a \cdot b}{\|a\| \cdot \|b\|} \tag{6}$$

3.4. Guided Inference

As stated in Section 3.1, the embedding distribution may be subject to distortions from less reliable embedding inputs. To solve this problem, our idea is to conceive a frame-wise confidence factor and apply the factor to guide the frame inference process.

On the whole, the guided online inference process is graphically illustrated in Figure.1. Note that the upper part above the green dash-line is the original diagram of UIS-RNN (where embedding x' and RNN state h' denote new speaker information for inference). By making use of the auxiliary global embedding profile (described in Section 3.3), the original joint likelihood, that feeds to the MAP decoding, is reinforced with our proposed confidence factors.

To be specific, for a given testing utterance $X^{\text{test}} = (x_1, x_2, \cdots, x_T)$, three input branches are fed into proposed inference system sequentially over time frames. Namely, at t-th time frame, the proposed inference procedure takes input of embedding x_t, historical speaker profile $\mathcal{M}_{t-1}$ and global embedding profile $\mathcal{E}_{t-1}$. Then following operations are conducted:

- calculate the log joint likelihood $\mathbf{S}_{\text{L}}(x_t)$ for existing speaker and new speaker (like the original UIS-RNN inference);
- calculate the proposed confidence factor $\mathbf{S}_{\text{CS}}(x_t)$ for existing speaker and new speaker;
- scale the MSE score by element wise product with a hyper-parameter γ shown as following:

$$\mathbf{S}_{\text{f}(x_t)} = \mathbf{S}_{\text{L}}(x_t) * (1 - \gamma * \mathbf{S}_{\text{CS}}(x_t)) \tag{7}$$

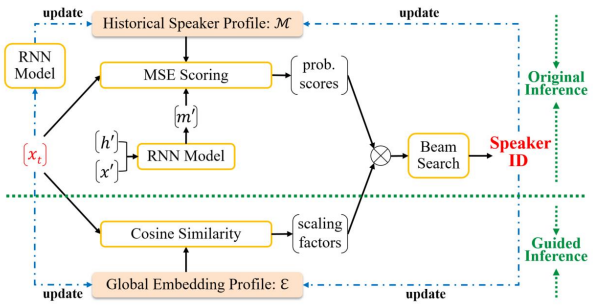

Figure 1: *Block diagram of proposed online inference process.*

- apply MAP decoding and predict the speaker ID $y_t = k$;
- once k-th speaker is predicted, update both historical speaker profile $\mathcal{M}_{t-1}^k$ (by simple moving average) and global embedding profile $\mathcal{E}_{t-1}^k$ (cf. Section 3.3).

Here in Eq.7, for an embedding x_t, its confidence factor $\mathbf{S}_{\text{CS}}(x_t)$ is designed based on its similarity with embedding information within profile $\mathcal{E}^k$, formulated as:

$$S_{\text{CS}}^k(x_t) = \begin{cases} \dfrac{\text{mean}(\bar{E}^k) \cdot x_t}{\|\text{mean}(\bar{E}^k)\| \cdot \|x_t\|} & k = 1, \cdots, K_{t-1} \\ \sigma & k = K_{t-1} + 1 \end{cases} \tag{8}$$

where K_{t-1} denotes the total number of unique speakers up to the $(t-1)$-th frame embedding entry and σ is hyper-parameter.

Although the method is heuristic, it is not only beneficial to prevent imperfect x_t from hurting the prediction of y_t, but also implicitly favorable to the update of $\mathcal{M}$. As verified in Section 4, this method works considerately well on meeting dataset.

4. Experiments

4.1. Dataset Preparation

The proposed methods are evaluated on AMI dataset [1], a widely used public corpus with 100-hour meeting recordings and 3 to 5 speakers per meeting. AMI dataset is splited into train, dev and eval sets and the division details are listed in Table.1. Following common procedure reported in prior arts, TNO meetings are excluded from both Dev and Eval sets.

Table 1: *Data Structure of AMI meeting corpus*

	#Meeting	#Speaker
Train	135	155
Dev	18	21 (2 presented in Train)
Eval	12	12 (unseen speakers)

To validate the ability of our system, we use the same data pre-processing as in [6], which consists of using Oracle SAD for online segmentation and excluding overlapping speech during evaluation. In the online setting, we apply a sliding window to break each utterance into short segments, with duration less than or equal to 2 seconds. The above procedures make diarization error rate (DER) equal to speaker error rate (SER), and thus in our case the collar tolerance does not apply.

[1] http://groups.inf.ed.ac.uk/ami/corpus/

4.2. Model Specifications

When extracting the speaker embedding, since the state-of-the-art d-vector and its variants that deployed in [7] are publicly unavailable, we have to use our in-house speaker embedding model throughout all our experiments. This model is developed under the framework of ECAPA-TDNN[19], trained by data from public domain[2].

To build optimal sequence generation model, a few attempts are made with combinations of different RNN structure (GRU or LSTM) and layer number. The best model obtained has GRU spanning three layers and 512 nodes per layer. The model is trained with application of Adam optimizer, a batch size of 10 and learning rate of 0.0005. The parameters $(\beta_1, \beta_2, \beta_3)$ in Eq.2 are set as $(1.045, 0.04, 0.05)$, following the recommendations given in [18]. Other hyper-parameters, like $\alpha_1, \alpha_2, \gamma$ and σ are fine-tuned based on the Dev set.

4.3. Results and discussions

Table 2: *DER (%) results on AMI Eval dataset. The best case performance for each mode is given in bold font.*

Ref.	online mode	clustering model	embedding	DER
Ours	✓	UIS-RNN Proposed	in-house embedding	15.95 **7.83**
[6]	x	UIS-RNN	d-vector	12.52
		SC		13.52
		k-means		17.70
		improved DEC		10.66
[20]	x	AHC	x-vector	10.37
		k-means	clusterGAN	11.56
			fusion	8.92
[21]	x	SC	x-vector	6.23
			clusterGAN	8.16
			fusion	**2.87**

Experimental results on AMI dataset are listed in Tab.2. To authors' best knowledge, few online SD approaches have been reported on AMI dataset before. The only relevant prior work [11] reported EER (equal error rate) higher than 15% at the cost of very long speaker latency. Thus, typical offline SD approaches, experimented on the AMI under the same evaluation setup as ours, are listed herein for comparison purpose.

A few valuable insights can be obtained from the experiment results. Firstly, our proposed model significantly boosts the baseline by a substantial margin of 48.7% relative in terms of DER. Secondly, in context of same embedding model, performance gains from different clustering schemes show that UIS-RNN is a promising clustering model, which outperforms AHC, k-means and SC. Thirdly, a similar observation regarding the embedding model, shows a performance trend that fused embedding is strikingly effective than individual embedding. Finally yet importantly, our proposed model, with online operations, even consistently outperforms most offline prior arts, which strongly validated the effectiveness of our approaches.

[2] http://www.openslr.org

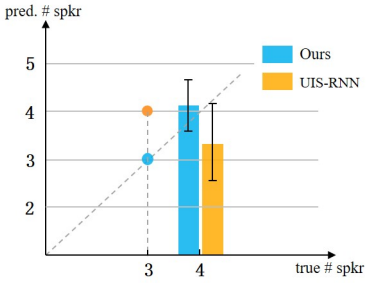

Figure 2: *Statistical analysis on predicted number of speakers. Regarding the 4-speaker meeting cases, for UIS-RNN and our proposed system, the mean and standard deviation are (3.36,0.81) and (4.09,0.54), respectively.*

Furthermore, close observation of Tab.2 reveals that much more prominent performance improvement is provided by embedding fusion approach, rather than clustering schemes.This suggests that research on representative speaker embeddings, to exploit the complementary merits of individual embedding, might be an interesting topic towards our future research.

It is widely known that having a correct estimate of the number of speakers is crucial for the overall performance. Thus, aside from DER, the predicted speaker numbers are statistically analyzed in form of error-bar charts. The results are shown in Figure.2. Among the 12 meetings in Eval set, one has 3 speakers and remaining 11 consistently have 4 speakers. Comparing the means and variations of two approaches, it is clear that our proposed system is much stabler, and the average of predicted speaker number are very close to the ground truths.

Table 3: *Ablation study results*

Models	Eval. DER(%)
our proposal (M)	7.83
M w/o homo-loss	8.73
M w/o Guided Inference	9.78
M w/o hetero-loss	10.28

Lastly, to unveil the effect of each proposed component, ablation studies are conducted by removing some component from our optimal model. The study results are reported in Tab.3. As expected, the hetero-loss provides the most and homo-loss the least contribution; and the guided inference also works well with solid performance contribution.

5. Conclusions

As an effort to overcome the online speaker diarization challenge, we proposed two methods to enhance a baseline system, UIS-RNN. One is to construct a discriminative training process and the other is to design a guided inference mechanism. Experiments conducted on AMI dataset achieve impressive performances comparable to most offline prior arts. These experimental results strongly validate the efficacy of our proposal.

Future work includes investigating the generalization of the proposed methods, to apply them across multiple datasets (e.g. NIST CALLHOME benchmark). In addition, the results above also reveal an interesting direction to extend the scope of our research by exploring fusion strategy for embedding learning.

6. References

[1] G. Sell and D. Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 413–417.

[2] G. Heigold, I. Moreno, S. Bengio, and N. Shazeer, "End-to-end text-dependent speaker verification," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5115–5119.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[4] G. Sun, C. Zhang, and P. Woodland, "Combination of deep speaker embeddings for diarisation," *Neural networks : the official journal of the International Neural Network Society*, vol. 141, p. 372—384, April 2021. [Online]. Available: https://doi.org/10.1016/j.neunet.2021.04.020

[5] T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, "Auto-tuning spectral clustering for speaker diarization using normalized maximum eigengap," *IEEE Signal Processing Letters*, vol. 27, pp. 381–385, 2020.

[6] D. Dimitriadis, "Enhancements for audio-only diarization systems," *ArXiv*, vol. abs/1909.00082, 2019.

[7] A. Zhang, Q. Wang, Z. Zhu, J. Paisley, and C. Wang, "Fully supervised speaker diarization," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6301–6305.

[8] Z. Huang, S. Watanabe, Y. Fujita, P. García, Y. Shao, D. Povey, and S. Khudanpur, "Speaker diarization with region proposal network," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6514–6518.

[9] Y. Fujita, N. Kanda, S. Horiguchi, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with permutation-free objectives," in *Interspeech 2019*, 2019, pp. 4300–4304.

[10] D. Dimitriadis and P. Fousek, "Developing on-line speaker diarization system," in *Proc. Interspeech*, 2017, pp. 2739–2743.

[11] J. Patino, R. Yin, H. Delgado, H. Bredin, A. Komaty, G. Wisniewski, C. Barras, N. Evans, and S. Marcel, "Low-latency speaker spotting with online diarization and detection," in *The Speaker and Language Recognition Workshop (Odyssey)*, 2018.

[12] E. Han, C. Lee, and A. Stolcke, "Bw-eda-eend: Streaming end-to-end neural speaker diarization for a variable number of speakers," in *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[13] Y. Xue, S. Horiguchi, Y. Fujita, S. Watanabe, P. Garcia, and K. Nagamatsu, "Online end-to-end neural diarization with speaker-tracing buffer," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 841–848.

[14] T. v. Neumann, K. Kinoshita, M. Delcroix, S. Araki, T. Nakatani, and R. Haeb-Umbach, "All-neural online source separation, counting, and diarization for meeting analysis," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 91–95.

[15] E. Fini and A. Brutti, "Supervised online diarization with sample mean loss for multi-domain data," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7134–7138.

[16] Q. Li, F. L. Kreyssig, C. Zhang, and P. C. Woodland, "Discriminative neural clustering for speaker diarisation," in *2021 IEEE Spoken Language Technology Workshop (SLT)*, 2020.

[17] A. Addlesee, Y. Yu, and A. Eshghi, "A comprehensive evaluation of incremental speech recognition and diarization for conversational AI," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3492–3503.

[18] Y. Fathullah, C. Zhang, and P. C. Woodland, "Improved large-margin softmax loss for speaker diarisation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 7104–7108.

[19] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech 2020*, Oct 2020.

[20] M. Pal, M. Kumar, R. Peri, T. J. Park, S. Hyun Kim, C. Lord, S. Bishop, and S. Narayanan, "Speaker diarization using latent space clustering in generative adversarial network," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 6504–6508.

[21] M. Pal, M. Kumar, R. Peri, T. J. Park, S. H. Kim, C. Lord, S. Bishop, and S. S. Narayanan, "Meta-learning with latent space clustering in generative adversarial network for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1204–1219, 2021.