Self-Supervised Learning for Online Speaker Diarization

Jen-Tzung Chien and Sixun Luo
Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

***Abstract*—Speaker diarization deals with the issue of "who spoke when" which is tackled through splitting an utterance into homogeneous segments with individual speakers. Traditional methods were implemented in an offline supervised strategy which constrained the usefulness of a practical system. Real-time processing and self-supervised learning are required. This paper deals with speaker diarization by relaxing the needs of reading the whole utterance and collecting the speaker label. The online pipeline components including feature extraction, voice activity detection, speech segmentation and speaker clustering is implemented. Importantly, an efficient end-to-end speech feature extraction is implemented by an unsupervised or self-supervised method, and then combined with online clustering to carry out online speaker diarization. This feature extractor is implemented by merging a bidirectional long short-term memory and a time-delayed neural network to capture the global and local features, respectively. The contrastive learning is introduced to improve initial speaker clusters. The augmentation invariance is imposed to assure model robustness. The online clustering based on autoregressive and fast-match clustering is investigated. The experiments on speaker diarization over NIST Speaker Recognition Evaluation show the merits of the proposed methods.**

I. Introduction

Speaker diarization aims to distinguish the speaker identities and detect the corresponding speech boundaries from an audio stream. Conventionally, an offline supervised strategy was developed to handle this problem over a batch collection of audio signals, e.g. broadcast shows, meeting recordings or telephone calls where a set of transcribed streams are provided [1], [2], [3]. Such a strategy could not really meet the demands of practical systems including artificial intelligence glasses, real-time summarization, etc, where unsupervised learning and online processing [4], [5] are required. Without knowing the prior identity and number of enrolled speakers, the speaker diarization process, consisting of voice activation, speech segmentation, feature extractor and speaker clustering, is sometimes more challenging than speech recognition. This challenge is even demanding when an immediate decision is required for a very short speech segment given by the newly-detected speaker clusters. To relax the demand of speaker labeling, it is crucial to develop the unsupervised or self-supervised approach [6], [7] to online speaker diarization where additional labeled data are avoided. Self-supervised learning is a specific type of unsupervised learning which was first proposed for image representation. This learning style is helpful for speaker diarization due to twofold reasons. First, self-supervised learning disregards the need of audio streams with speaker labels which considerably save the manpower for labeling or transcription. Without human in the loop, the performance of speaker diarization is robust for various amounts of data. Second, traditional speaker diarization closely depends on the feature extraction which affects the performance of speaker recognition based on supervised training as well as speaker clustering using test data. However, speaker recognition and clustering are not always positively correlated because of the issues of overfitting and domain mismatch. Self-supervised method is beneficial to build a robust feature extraction which consistently improves speaker recognition and clustering for online diarization.

Recently, self-supervised learning was proposed for image clustering [7]. The solution to image clustering is here extended for speaker representation which is used to predict initial speaker cluster labels for various segments of an audio stream. The contrastive loss [8] is minimized to fulfill self-supervised learning for a robust speaker feature extraction as well as a reliable feature augmentation where the augmentation invariance is pursued for the features estimated from original and augmented audio streams. The augmentation method based on SpecAugment [9] is applied. In addition, the online clustering is performed to obtain final clustering result based on those initial speaker cluster labels. In the literature, several online clustering methods have been developed in [10], [11], [12]. The proposed online clustering is based on the initial cluster labels of speakers which are predicted according to a self-supervised learning for feature extraction. The autoregressive clustering and fast-match clustering are investigated. There are threefold ideas or novelties presented in this paper. First, an end-to-end speech feature extraction is presented for fast preprocessing. Second, the self-supervised learning is carried out for online speaker diarization. Third, the on-line speaker clustering methods are investigated for comparison. A set of experiments are evaluated to illustrate the efficiency and usefulness of the proposed methods.

II. Related Work

Traditional speaker diarization was developed as an offline method which was performed after an entire audio recording was collected. A practical solution to online or real-time calculation for speaker diarization is required in many applications. In recent years, online speaker diarization has been extensively studied [13], [14]. The previous online learning methods were developed by considering traditional speaker recognition system under the Gaussian mixture model (GMM) combined with universal background model (UBM) [15], [16], [17]. In

the implementation, two individual gender-dependent UBMs were estimated and used as the background seed models. Each incoming audio segment was evaluated with current speaker models which were used to identify the corresponding speaker. The speaker models were updated or created at runtime with a predefined threshold. Maximum *a posteriori* (MAP) estimation was applied to adapt GMMs to update speaker models. In [17], it was shown that the pre-enrolled speakers in a meeting played an important role for an online system. Without the speaker enrollment, the performance of online solution was much worse than that of offline system. In [18], an online speaker identification was merged with an offline speaker diarization to carry out a hybrid approach to online speaker diarization. A low-latency decision on current speaker was made. In [19], an online speaker recognition system was used to identify word boundaries which were then applied to detect change points of different speakers. Gaussian distributions were assumed and adopted to implement the Bayesian information criterion for model selection or equivalently change point detection while the i-vectors [20] were calculated for speaker clustering.

In addition to online computation, another strategy to improve the usefulness of speaker diarization is to enhance speaker representation without extra training data and speaker labels. This study introduces the self-supervised learning as a new type of unsupervised learning which is employed to enrich feature expression for speaker diarization. Traditionally, there were two categories of self-supervised learning methods which were developed for image representation. The first one included the generative methods [21], [22], [23], [24] while the second one contained the contrastive methods [25], [26], [27], [28], [29]. Typically, the generative methods were estimated by minimizing the reconstruction error in pixel space, or equivalently the loss of pixel labels. Autoencoder was used to conduct learning representation with an encoder and a decoder. Encoder was designed to find a sufficient latent code to represent original data while decoder was used to reconstruct input image using this latent code. On the other hand, the contrastive methods were proposed to learn the general features by teaching the model which input points are similar or dissimilar. In general, contrastive methods are more advanced and systematically robust than generative models. In [30], [31], [32], the end-to-end processing was introduced for speaker diarization where the algorithms for online computation and varying number of speakers were proposed. This study presents the contrastive learning for online self-supervised end-to-end speaker diarization.

III. Proposed Method

The proposed online speaker diarization consists of three pipeline components which are end-to-end speech feature extractor, self-supervised speaker feature extractor, and online speaker clustering as depicted in Figure 1.

A. System Overview

The first component of this system is formed as an end-to-end speech feature extractor which is used to calculate

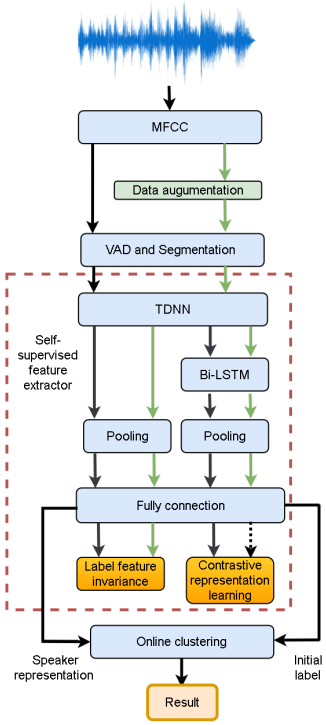

Fig. 1: System architecture for online self-supervised speaker diarization with the augmented data shown in green flow.

speech features from input signals. The feature extractor of Mel-frequency cepstral coefficients (MFCCs) is performed for each audio frame of 25ms with 10ms in frame shift. A voice activity detection is applied to classify each frame into the class of speech or non-speech, and a speech segmentation module is merged to record time stamps of various speech/non-speech boundaries. Only those speech segments are considered to calculate speech features. Non-speech segments are disregarded from feature extraction. Notably, the data augmentation using SpecAugment [9], consisting of warping features, masking blocks of frequency channels, and masking blocks of time steps, is applied. Second, the self-supervised speaker feature extractor is proposed to extract the independent speaker features which are distinguishable for original audio inputs and augmented audio inputs in feature space. This module is constructed as a layer-wise network which is composed of the time-delayed neural network (TDNN) layers with a pooling layer, the bidirectional long short-term memory (Bi-LSTM) layers with another pooling layer, and the fully-connected layers. Such a network is trained to build self-supervised feature extractor where the augmentation invariance is enforced and the contrastive learning is performed. The third component is configured by an online clustering module where the autoregressive clustering and fast-match clustering are examined. Autoregressive clustering is implemented to determine the speaker label at current time t by following the speaker probabilities in a range of history time $[t-\tau, t]$ with length τ

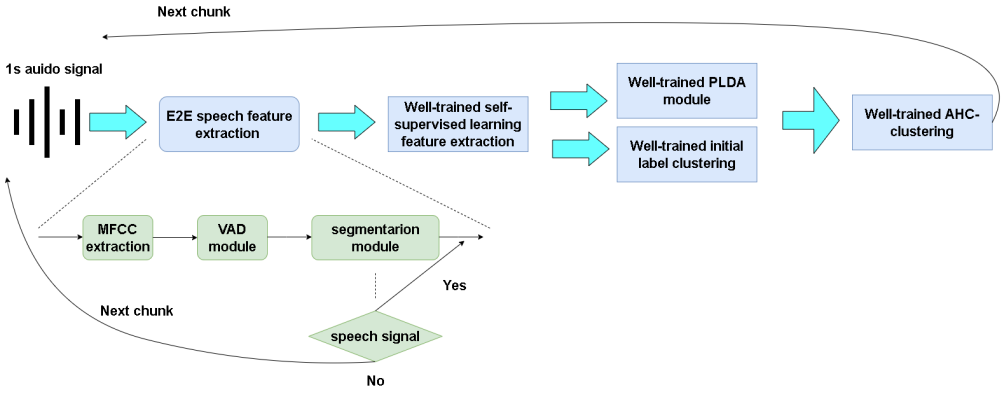

Fig. 2: Online processing in E2E speech feature extractor and self-supervised speaker feature extractor for speaker diarization.

using the speaker features with high confidence of predicted labels in the past time $[0, t)$. Also, the fast-match clustering is performed in a way of either low-cost clustering for saving computation cost or high-cost clustering for obtaining precise representation. Typically, high-cost clustering is based on the probabilistic linear discriminant analysis (PLDA) [33], [34], [35], [36], [37] with the agglomerative hierarchical clustering (AHC) [38]. The final result on speaker diarization is obtained after online clustering. In what follows, the individual pipeline processing units or learning components are addressed.

B. Data Augmentation

This paper adopts the SpecAugment [9] as data augmentation scheme which is performed directly over MFCC features. Augmentation method is based on two kinds of masking blocks. One is for frequency channels, and the other is for time steps. The first kind of augmentation is to implement the frequency masking where a frequency band f is first chosen from a uniform distribution between 0 and F. F is a parameter for frequency masking. Let v denote the number of Mel-frequency channels. Then, f_0 is randomly chosen from the interval $[0, v - f)$. The consecutive Mel-frequency channels $[f_0, f_0 + f)$ are masked. In addition, the second kind of augmentation is based on time masking which aims to ignore time information. Similar to frequency masking over frequency channels, time masking is performed to mask those values in time interval $[t_0, t_0 + t)$. Let T denote the parameter for time masking. Thus, time width t is first randomly chosen from the interval between 0 and T. Then, t_0 is chosen from the interval $[0, T - t)$. Those augmented data based on frequency masking and time masking are then merged to train the self-supervised feature extractor. Such a treatment basically enhance the robustness of speech recognition or speaker diarization especially in presence of low-quality and low-resource speech data.

C. End-to-End Speech Feature Extractor

Figure 2 shows the online processing for two parts. One is the end-to-end (E2E) speech feature extraction (shown in green components), and the other is the self-supervised speaker feature extraction. The first part is constructed by cascading the components of MFCC extraction, voice action detection (VAD) and speech segmentation. Assuming that an audio chunk of one second is observed as an input signal, the first step is to extract MFCC features for each frame which is then classified to either speech signal or non-speech signal via VAD module. This classification is based on a threshold for the likelihood ratio of speech segment with respect to non-speech segment in 1 second. This threshold controls the sensitivity of VAD module. The VAD result is used to judge if the chunk with MFCCs belongs to speech segment. Only the chunk of speech segment is forwarded to the next step for self-supervised learning. Non-speech chunk is discarded immediately. Such an end-to-end speech feature extractor is especially crucial since traditional development tool using Kaldi [39] was time-consuming in the implementation due to offline processing overhead for data reading and writing. The original processing time in Kaldi was too long. Therefore, this study integrates three processing units to construct an end-to-end structure to save computation time for speech feature extraction. VAD method is performed according to the frame-based energy function. If the chunk belongs to non-speech segment, this chunk is totally ignored and next chunk of audio signal is then targeted in the online processing. To handle this cascaded processing, a sliding window is applied to alleviate the deviation caused by short-term chunks.

D. Self-Supervised Speaker Feature Extractor

Augmentation invariance and contrastic learning are implemented for self-supervised speaker feature extraction.

1) Augmentation invariance: Augmentation invariance aims to make sure that the augmented audio data have the same speaker feature representation as the original audio data. The augmentation is based on frequency masking and time masking on MFCCs of original data. To maximize the similarity between original and augmented features, the label features $\mathbf{l}_i = f(\mathbf{x}_i; \theta_a)$ of an original audio input $\mathbf{x}_i$ are constrained to be close to those features $\widehat{\mathbf{l}}_i$ corresponding

to the augmented audio data $A(\mathbf{x}_i)$. Here, A denotes an augmentation method and $f(\cdot)$ denotes the feature extractor with parameter θ_a for original and augmented data under two class labels. This study pursues the label feature invariance by estimating θ_a for label features $\mathbf{l}_i$ and $\widehat{\mathbf{l}}_i$ over audio inputs $\mathbf{x}_i$ within speech segments where the summation of individual negative cosine similarities

$$\mathcal{L}_a(\mathbf{l}_i, \widehat{\mathbf{l}}_i) = -\frac{\mathbf{l}_i}{\| \mathbf{l}_i \|_2} \cdot \frac{\widehat{\mathbf{l}}_i}{\| \widehat{\mathbf{l}}_i \|_2} \triangleq d(\mathbf{l}_i, \widehat{\mathbf{l}}_i) \tag{1}$$

as distances $d(\mathbf{l}_i, \widehat{\mathbf{l}}_i)$ is minimized. By preserving the property of label feature invariance, the speaker feature extraction is able to compensate the degraded speaker information in the augmented audio data due to time and frequency maskings.

2) Contrastive representation learning: In addition, the contrastive learning is developed to implement self-supervised speaker representation for online speaker diarization. The goal is to enhance feature representation which can distinguish utterances from the same or different speakers. This goal is achieved by minimizing within-class variance and simultaneously maximizing between-class distance. The separability maximization is pursued. In short, speaker representation is enhanced such that the vectors from the same speaker are close together, and those from different speakers are separate from each other. To fulfill this goal, a contrastive loss was constructed as [8]

$$\mathcal{L}_c(\mathbf{l}_i, \mathbf{l}_j, y) = \frac{y}{2} d^2 + \frac{1-y}{2} (\max(0, m-d))^2 \tag{2}$$

where a neural network with parameters θ_c is used to extract the individual features $\{\mathbf{l}_i, \mathbf{l}_j\}$ for the paired samples $\{\mathbf{x}_i, \mathbf{x}_j\}$ and y is a binary label assigned to the paired samples. $y = 0$ means two samples $\{\mathbf{x}_i, \mathbf{x}_j\}$ belong to different classes while $y = 1$ means two samples are from the same class. $d(\mathbf{l}_i, \mathbf{l}_j)$ is the Euclidean distance between $\mathbf{l}_i$ and $\mathbf{l}_j$ which is minimized by adjusting the network parameters θ_c for those similar samples when $y = 1$. Dissimilar samples, i.e. $y = 0$, contributes to the loss function if their distance is within a radius or margin m. In this study, the label y_{ij} for each pair of similar features $\{\mathbf{l}_i, \mathbf{l}_j\}$ is introduced. m is set to 1. Distance measure between two label features is set as $d(\mathbf{l}_i, \mathbf{l}_j)$. Contrastive learning is treated as a binary classification. The binary cross-entropy loss in terms of distance measure $d(\mathbf{l}_i, \mathbf{l}_j)$ is measured for individual paired samples $\{\mathbf{x}_i, \mathbf{x}_j\}$ by

$$\begin{aligned} & - y_{ij} \log \left(d \left(\mathbf{l}_i, \mathbf{l}_j \right) \right) - \left(1 - y_{ij} \right) \log \left(1 - d \left(\mathbf{l}_i, \mathbf{l}_j \right) \right) \\ & = -y_{ij} \log \left(\frac{\mathbf{l}_i}{\|\mathbf{l}_i\|_2} \cdot \frac{\mathbf{l}_j}{\|\mathbf{l}_j\|_2} \right) \\ & \quad - \left(1 - y_{ij} \right) \log \left(1 - \frac{\mathbf{l}_i}{\|\mathbf{l}_i\|_2} \cdot \frac{\mathbf{l}_j}{\|\mathbf{l}_j\|_2} \right) \triangleq \mathcal{L}_c(\mathbf{l}_i, \mathbf{l}_j, y_{ij}) \end{aligned} \tag{3}$$

where $y_{i,j} = 1$ represents that the audio samples $\mathbf{x}_i$ and $\mathbf{x}_j$ belong to the same cluster, and $y_{i,j} = 0$ implies that the audio samples $\mathbf{x}_i$ and $\mathbf{x}_j$ come from different clusters. We develop a learning algorithm that combines loss functions for contrastive learning and augmentation invariance learning which are minimized for independent speaker feature extraction as well as feature label augmentation for invariance in original and augmented audio streams. A self-supervised learning task is formed in an unsupervised manner where no additional labeled data are required. Self-supervised learning is run by minimizing $\mathcal{L} = \mathcal{L}_c + \alpha \mathcal{L}_a$ with a regularization parameters α which is accumulated over different features and augmented features $\{\mathbf{l}_i, \widehat{\mathbf{l}}_j\}$ and different pairs of features $\{\mathbf{l}_i, \mathbf{l}_j\}$.

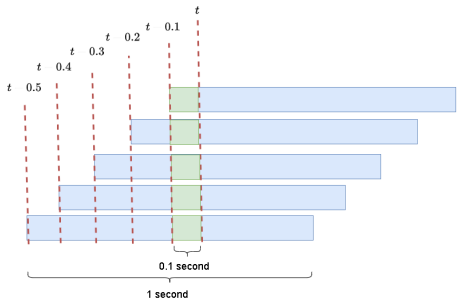

Fig. 3: Illustration for autoregressive clustering.

E. Online Clustering

Autoregressive clustering and fast-match clustering are performed for online speaker diarization. Figure 3 illustrates the autoregressive clustering via voting. This study uses one second of audio signal as a chunk and 0.1 second as the time shift of a frame. The decision delay time is 0.5 second in the beginning. Assume that the speech frames are given by $\{\mathbf{x}_t\}_{t=0}^{T}$. When the initial label is obtained in current time t and $t > 0.5$, the current speaker feature labels are $\{l_t^n\}_{n=1}^{5}$ for five consecutive frames with overlapping. For example, l_t^1 means the speaker feature label early at 0.1 second at $n = 1$ and l_t^5 means the label at current time at $n = 5$. The final speaker feature label of green area at an interval of 0.1 second in current time t is obtained by voting over the labels corresponding to five speaker features. Speaker embeddings are not required by using this method. For comparison, this study also carries out the fast-match clustering which is based on the softmax layer of a self-supervised speaker feature extractor where the number of clusters is set as 256. We didn't use the cluster number of Voxceleb for CALLHOME data to avoid the domain mismatch problem. A kind of clustering over the speaker embeddings with similar characteristics is performed. Based on the initial speaker label from this fast-match scheme and the speaker vector from speaker feature extractor, a well-trained PLDA model for speaker scoring combined with a well-trained AHC for speaker clustering are applied to implement a precise clustering. The clustering result with the largest probability is obtained.

IV. Experiments

A. Data Preprocessing & Zero-Shot Learning

In this speaker diarization task, the evaluation or test data were collected from the disk-8 of 2000 NIST Speaker Recognition Evaluation which was also known as the CALLHOME

data consisting of 148.9 hours of conversational telephone speech. This dataset provided speaker labels and time information of each speech segments which were used for evaluation of speaker diarization. 5-fold cross validation was performed. This study conducted the system evaluation by using the Voxceleb data (1251 speakers in Voxceleb1 and 6112 speakers in Voxceleb2) from YouTube (153.5K sentences in 351 hours) as the training data which were seen as out-of-domain data. Genders of speakers were balanced. Different types of noises were present. Averaged length of each sentence was 8.2 sec. *Zero-shot learning* was evaluated for online speaker diarization since training and test data were from different domains. Such an evaluation was challenging since the unsupervised learning was performed in two different domains in training and test data. This scenario was different from the previous works [40], [41], [42], [43] where in-domain data SRE2004, SRE2005, SRE2006, SRE2008, SWBD and SWBD2 were adopted. In addition, different from [44] where the speaker labels for training x-vectors were required, this work built an unsupervised speaker diarization [40] which relaxed the need of speaker labels in feature extraction for speaker clustering during training procedure. In particular, the experimental setting based on *online* speaker diarization made the task even harder than the other tasks. In the implementation, the original sampling rates of audio files in Voxceleb and CALLHOME were 16K and 8K, respectively. Downsampling the audio signals of Voxceleb to 8K was performed. Window size of each frame was 25ms, and frame shift was 10ms. The 23-dimensional MFCCs were calculated for each frame in audio feature extraction. The amount of training data was increased by data augmentation where the reverberation, noise, music and babble noises from MUSAN corpus [45] were added to original speech signals. SpecAugment method [9] was implemented to generate the augmented data for self-supervised learning. Diarization error rate (DER) was measured in the comparison.

B. Model Configuration

Model architecture for the proposed self-supervised speaker diarization is addressed. The first five layers were formed by time-delayed nural network (TDNN), where speech frames with a temporal context window centered at current time t were used as the inputs. For example, the input of frame layer 3 was from the spliced output of frame layer 2 at frames $t-3$, t and $t+3$. Frame layer 3 was able to see a total context of 15 frames. On top of TDNN layers, a statistical pooling layer was configured to compute the mean and variance of the outputs of TDNN over the time $[0, T)$. On top of pooling layer, there were two fully connection layers used for speaker-level feature extraction. The outputs of the second speaker layer was used as the inputs to softmax layer with the outputs corresponding to individual speaker clusters. After the model was well trained, the embeddings extracted from the affine component of the second speaker layer were known as x-vectors. The number of Bi-LSTM layers was two, and the output dimension was 1500. The fully-connection module of speaker feature extractor had three layers. There were two layers in speaker label classifier. An empirical threshold was set to determine whether a new cluster was generated or not. The effects of using x-vector and domain mismatch are investigated in what follows.

TABLE I: Comparisons of DER (%) using SRE+SWBD and Voxceleb in offline speaker diarization. Evaluation is conducted whether the number of speakers is given or not.

Model	Is speaker no. given	**DER** (SRE+SWBD)	**DER** (Voxceleb)
i-vector	No	12.1	16.2
x-vector	No	8.4	12.2
x-vector	Yes	7.1	11.0
x-vector+reseg.	No	6.5	9.6
x-vector+SSL	No	5.2	7.4

C. Experimental Results

First of all, Table I reports DERs of different models using different speaker features and processing schemes where offline speaker diarization is investigated. Evaluation is also conducted to see the effects due to the training data using Voxceleb as out-domain data and SRE+SWBD as in-domain data as well as the condition whether the number of speakers is provided or not. Obviously, x-vectors significantly perform better than i-vectors. Using x-vectors, given the number of speakers is beneficial to reduce DER by constructing the reliable speaker clusters. DER is further reduced by applying resegmentation even the number of speakers is unknown. Nevertheless, the lowest DERs are achieved by applying the proposed self-supervised learning (SSL) with augmented data via SpecAugment. In this set of experiments, DERs with supervised in-domain learning using SRE+SWBD are considerably lower than those with unsupervised out-domain using Voxceleb. Such results reflect the influence of domain mismatch and labeled data in speaker diarization.

Next, Table II shows the results of DER and computation time per chunk (in msec) for different models where online speaker diarization is examined under the condition that number of speakers is unknown. Each chunk is set by 1 second. CPU of E5-2620 v4 is used. The results of online speaker diarization using x-vector and self-attention based end-to-end speaker diarization (SA-EEND) [31], where SWBD+SRE was adopted as supervised in-domain learning, are included for comparison. This study implemented the proposed SSL with different online clustering for online speaker diarization where zero-shot learning is performed as an unsupervised out-domain learning using Voxceleb. The efficiency is also evaluated in terms of computation time. From this set of experiments, we find that online speaker diarization obtains higher DERs than offline speaker diarization. Out-domain learning is more difficult than in-domain learning which is reflected from DERs. In such case of zero-shot learning, the proposed SSL performs significantly better than standard setting based on x-vectors. With the ablation study, online method via fast-matching clustering works better than that via autoregressive clustering. DER using SSL is further reduced by merging with

the self attention in [31]. The lowest DER is achieved as 11.9% by applying speaker-tracing buffer (STB) [31] where in-domain training data are used. One interesting result in this work is the evaluation of computation time. The proposed SSL method substantially reduces the computation time when compared with recent work based on SA-EEND-STB [31]. A real-time speaker diarization is built. Source codes in PyTorch are accessible at https://github.com/NCTUMLlab/Si-Xun-Luo.

TABLE II: Comparisons of DER (%) and computation time (msec) per chunk in online speaker diarization using different training data. Online methods using autoregressive clustering (AC) and fast-match clustering (FC) are included. Self attention (SA) and speaker-tracing buffer (STB) are evaluated.

Model	Training data	DER	Comp
Online x-vector [31]	SWBD+SRE	26.9	–
Online SA-EEND [31]	SWBD+SRE	36.6	1070
Online SA-EEND-STB [31]	SWBD+SRE	12.8	500
Online x-vector	Voxceleb	42.3	–
Online SSL-AC	Voxceleb	23.4	68
Online SSL-FC	Voxceleb	22.2	70
Online SSL-FC-SA	Voxceleb	20.9	280
Online SSL-FC-STB	SWBD+SRE	11.9	390

V. Conclusions

This study has presented the self-supervised learning with data augmentation for online speaker diarization where model configuration was addressed. Contrastive learning and augmentation invariance for feature labels were proposed to assure a reliable learning representation where the bidirectional LSTM and time-delayed neural network were implemented. Efficiency in online clustering was sufficiently attained by the proposed fast-matching clustering. From the evaluation of online speaker diarization with in-domain and out-domain training data, the results on CALLHOME showed that the performance was considerably improved by applying the self-supervised learning where the feature extraction module related to speaker clustering was benefited.

References

[1] Sue E. Tranter and Douglas A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[3] Man-Wai Mak and Jen-Tzung Chien, *Machine Learning for Speaker Recognition*, Cambridge University Press, 2020.

[4] Jen-Tzung Chien, "Online hierarchical transformation of hidden markov models for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 656–667, 1999.

[5] Hsin-Lung Hsieh and Jen-Tzung Chien, "Online bayesian learning for dynamic source separation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 1950–1953.

[6] Carl Doersch, Abhinav Gupta, and Alexei A Efros, "Unsupervised visual representation learning by context prediction," in *Proc. of IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430.

[7] Chuang Niu, Jun Zhang, Ge Wang, and Jimin Liang, "GATCluster: Self-supervised gaussian-attention network for image clustering," in *Proc. of European Conference on Computer Vision*, 2020, pp. 735–751.

[8] Raia Hadsell, Sumit Chopra, and Yann LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2006, vol. 2, pp. 1735–1742.

[9] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[10] Vincent Cohen-Addad, Benjamin Guedj, Varun Kanade, and Guy Rom, "Online k-means clustering," *arXiv preprint arXiv:1909.06861*, 2019.

[11] Anna Choromanska and Claire Monteleoni, "Online clustering with experts," in *Proc. of Artificial Intelligence and Statistics*, 2012, pp. 227–235.

[12] Le Li, Benjamin Guedj, Sébastien Loustau, et al., "A quasi-Bayesian perspective to online clustering," *Electronic Journal of Statistics*, vol. 12, no. 2, pp. 3071–3113, 2018.

[13] Catherine Breslin, KK Chin, Mark JF Gales, and Kate Knill, "Integrated online speaker clustering and adaptation," in *Proc. of Annual Conference of the International Speech Communication Association*, 2011, pp. 1085–1088.

[14] Weizhong Zhu and Jason Pelecanos, "Online speaker diarization using adapted i-vector transforms," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 5045–5049.

[15] Konstantin Markov and Satoshi Nakamura, "Improved novelty detection for online GMM based speaker diarization," in *Proc. of Annual Conference of International Speech Communication Association*, 2008, pp. 363–366.

[16] Jürgen Geiger, Frank Wallhoff, and Gerhard Rigoll, "GMM-UBM based open-set online speaker diarization," in *Proc. of Annual Conference of International Speech Communication Association*, 2010, pp. 2330–2333.

[17] Giovanni Soldi, Christophe Beaugeant, and Nicholas Evans, "Adaptive and online speaker diarization for meeting data," in *Proc. of European Signal Processing Conference*, 2015, pp. 2112–2116.

[18] Carlos Vaquero, Oriol Vinyals, and Gerald Friedland, "A hybrid approach to online speaker diarization," in *Proc. of Annual Conference of International Speech Communication Association*, 2010.

[19] Dimitrios Dimitriadis and Petr Fousek, "Developing on-line speaker diarization system," in *Proc. of Annual Conference of International Speech Communication Association*, 2017, pp. 2739–2743.

[20] Wei-wei Lin, Man-Wai Mak, and Jen-Tzung Chien, "Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2412–2422, 2018.

[21] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever, "Improving language understanding by generative pre-training," 2018.

[22] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, pp. 9, 2019.

[23] Aaron van den Oord, Nal Kalchbrenner, and Koray Kavukcuoglu, "Pixel recurrent neural networks," *arXiv preprint arXiv:1601.06759*, 2016.

[24] Laurent Dinh, David Krueger, and Yoshua Bengio, "NICE: Non-linear independent components estimation," *arXiv preprint arXiv:1410.8516*, 2014.

[25] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey Hinton, "Big self-supervised models are strong semi-supervised learners," *arXiv preprint arXiv:2006.10029*, 2020.

[26] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of International Conference on Machine Learning*, 2020, pp. 1597–1607.

[27] Yonglong Tian, Dilip Krishnan, and Phillip Isola, "Contrastive representation distillation," *arXiv preprint arXiv:1910.10699*, 2019.

[28] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[29] Abubakar Abid and James Zou, "Contrastive variational autoencoder enhances salient features," *arXiv preprint arXiv:1902.04601*, 2019.

[30] Eunjung Han, Chul Lee, and Andreas Stolcke, "BW-EDA-EEND: streaming end-to-end neural speaker diarization for a variable number of speakers," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 7193–7197.

[31] Yawen Xue, Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Paola García, and Kenji Nagamatsu, "Online end-to-end neural diarization

with speaker-tracing buffer," in *Proc. of IEEE Spoken Language Technology Workshop*, 2020, pp. 841–848.

[32] Weiwei Lin, Man-Wai Mak, and Jen-Tzung Chien, "Strategies for end-to-end text-independent speaker verification.," in *Proc. of Annual Conference of International Speech Communication Association*, 2020, pp. 4308–4312.

[33] Sergey Ioffe, "Probabilistic linear discriminant analysis," in *Proc. of European Conference on Computer Vision*, 2006, pp. 531–542.

[34] Man-Wai Mak, Xiaomin Pang, and Jen-Tzung Chien, "Mixture of PLDA for noise robust i-vector speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 1, pp. 130–142, 2015.

[35] Na Li, Man-Wai Mak, and Jen-Tzung Chien, "Dnn-driven mixture of PLDA for robust speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1371–1383, 2017.

[36] Wei-wei Lin, Man-Wai Mak, and Jen-Tzung Chien, "Fast scoring for PLDA with uncertainty propagation via i-vector grouping," *Computer Speech & Language*, vol. 45, pp. 503–515, 2017.

[37] Jen-Tzung Chien and Kang-Ting Peng, "Neural adversarial learning for speaker recognition," *Computer Speech & Language*, vol. 58, pp. 422–440, 2019.

[38] William H. E. Day and Herbert Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of Classification*, vol. 1, no. 1, pp. 7–24, 1984.

[39] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The Kaldi speech recognition toolkit," in *Proc. of IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.

[40] Stephen H Shum, Najim Dehak, Réda Dehak, and James R Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[41] Youzhi Tu, Man-Wai Mak, and Jen-Tzung Chien, "Variational domain adversarial learning with mutual information maximization for speaker verification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2013–2024, 2020.

[42] Longxin Li, Man-Wai Mak, and Jen-Tzung Chien, "Contrastive adversarial domain adaptation networks for speaker recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[43] Weiwei Lin, Man-Wai Mak, Youzhi Tu, and Jen-Tzung Chien, "Semi-supervised nuisance-attribute networks for domain adaptation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2019, pp. 6236–6240.

[44] Prachi Singh and Sriram Ganapathy, "Deep self-supervised hierarchical clustering for speaker diarization," *Proc. of Annual Conference of International Speech Communication Association*, pp. 294–298, 2020.

[45] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.