Neural Networks 140 (2021) 65–99

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

2021 Special Issue

Speaker recognition based on deep learning: An overview

Zhongxin Bai, Xiao-Lei Zhang *

Center of Intelligent Acoustics and Immersive Communications (CIAIC) and the School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an Shaanxi 710072, China

ARTICLE INFO

Article history:
Available online 17 March 2021

Keywords:
Speaker recognition
Speaker verification
Speaker identification
Speaker diarization
Robust speaker recognition
Deep learning

ABSTRACT

Speaker recognition is a task of identifying persons from their voices. Recently, deep learning has dramatically revolutionized speaker recognition. However, there is lack of comprehensive reviews on the exciting progress. In this paper, we review several major subtasks of speaker recognition, including speaker verification, identification, diarization, and robust speaker recognition, with a focus on deep-learning-based methods. Because the major advantage of deep learning over conventional methods is its representation ability, which is able to produce highly abstract embedding features from utterances, we first pay close attention to deep-learning-based speaker feature extraction, including the inputs, network structures, temporal pooling strategies, and objective functions respectively, which are the fundamental components of many speaker recognition subtasks. Then, we make an overview of speaker diarization, with an emphasis of recent supervised, end-to-end, and online diarization. Finally, we survey robust speaker recognition from the perspectives of domain adaptation and speech enhancement, which are two major approaches of dealing with domain mismatch and noise problems. Popular and recently released corpora are listed at the end of the paper.

1. Introduction

It is known that a speaker's voice contains personal traits of the speaker, given the unique pronunciation organs and speaking manner of the speaker, e.g. the unique vocal tract shape, larynx size, accent, and rhythm (Kinnunen & Li, 2010). Therefore, it is possible to identify a speaker from his/her voice automatically via a computer. This technology is termed as *automatic speaker recognition*, which is the core topic of this paper. We do not discuss speaker recognition by humans. Speaker recognition is a fundamental task of speech processing, and finds its wide applications in real-world scenarios. For example, it is used for the voice-based authentication of personal smart devices, such as cellular phones, vehicles, and laptops. It guarantees the transaction security of bank trading and remote payment. It has been widely applied to forensics for investigating a suspect to be guilty or non-guilty (Campbell et al., 2009; Champod & Meuwly, 2000; Kinnunen & Li, 2010), or surveillance and automatic identity tagging (Togneri & Pullella, 2011). It is important in audio-based information retrieval for broadcast news, meeting recordings and telephone calls. It can also serve as a frontend of automatic speech recognition (ASR) for improving the transcription performance of multi-speaker conversations.

The research on speaker recognition can be dated back to at least 1960s (Pruzansky & Mathews, 1964). In the following forty years, many advanced technologies promoted the development of speaker recognition. For example, a number of acoustic features (e.g. the linear predictive cepstral coefficients, the perceptual linear prediction coefficient, and the mel-frequency cepstral coefficients) and template models (e.g. vector quantization, and dynamic time warping) have been applied, see Kinnunen and Li (2010) for the details. Later on, Reynolds, Quatieri, and Dunn (2000) proposed the Gaussian mixture model based universal background model (GMM-UBM), which has been the foundation of speaker recognition for more than ten years since year 2000. Several representative models based on GMM-UBM have been developed, including the applications of support vector machines (Campbell, Sturim, & Reynolds, 2006) and joint factor analysis (Kenny, Boulianne, Ouellet, & Dumouchel, 2007). Among the models, the GMM-UBM/i-vector frontend (Dehak, Kenny, Dehak, Dumouchel, & Ouellet, 2010) with probabilistic linear discriminant analysis (PLDA) backend (Garcia-Romero & Espy-Wilson, 2011; Kenny, 2010) provided the state-of-the-art performance for several years, until the new era of deep learning based speaker recognition.

Recently, motivated by the powerful feature extraction capability of deep neural networks (DNNs), a lot of deep learning based speaker recognition methods were proposed (Lei, Scheffer, Ferrer, & McLaren, 2014; Snyder, Garcia-Romero, Sell, Povey, & Khudanpur, 2018; Variani, Lei, McDermott, Moreno, & Gonzalez-Dominguez, 2014) right after the great success of deep learning

* Corresponding author.
E-mail addresses: zxbai@mail.nwpu.edu.cn (Z. Bai), xiaolei.zhang@nwpu.edu.cn (X.-L. Zhang).

https://doi.org/10.1016/j.neunet.2021.03.004
0893-6080/

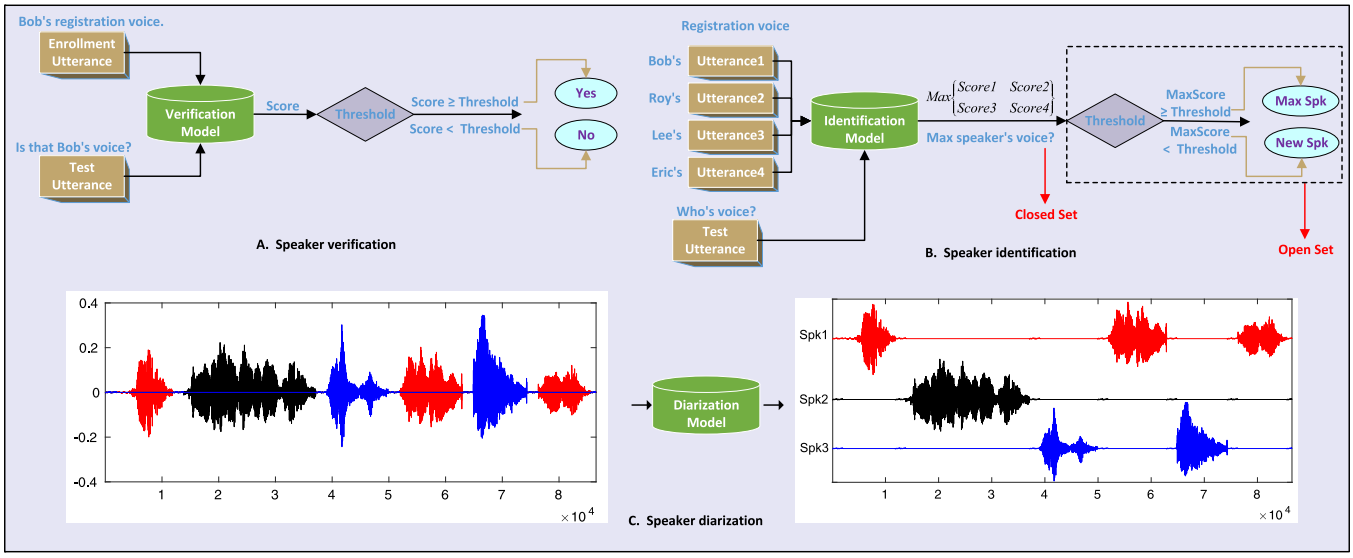

Fig. 1. Flowcharts of speaker verification, speaker identification, and speaker diarization. Fig. A describes speaker verification, which is a task of verifying whether a test utterance and an enrollment utterance are uttered by the same speaker via comparing the similarity score of the utterances with a pre-defined threshold. Fig. B describes speaker identification, which is a task of determining the speaker identity of a test utterance from a set of speakers. If the utterance must be produced from the set of the speakers, then it is a closed set identification problem; otherwise, it is an open set problem. Fig. C describes speaker diarization, which addresses the problem of "who spoke when", i.e., partitioning a conversation recording into several speech recordings, each of which belongs to a single speaker.

based speech recognition, which significantly boosts the performance of speaker recognition to a new level, even in wild environments (McLaren, Ferrer, Castan, & Lawson, 2016; Nagrani, Chung, & Zisserman, 2017).

In this survey article, we give a comprehensive overview to the deep learning based speaker recognition methods in terms of the vital subtasks and research topics, including speaker verification, identification, diarization, and robust speaker recognition. By doing the survey, we hope to provide a useful resource for the speaker recognition community. The main contributions of this paper are summarized as follows:

- We summarize deep learning based speaker feature extraction techniques for speaker verification and identification, from the aspects of inputs, network structures, temporal pooling strategies, and objective functions which are also the fundamental components of many other speaker recognition subtasks beyond speaker verification and identification.
- We make an overview to the deep learning based speaker diarization, with an emphasis of recent supervised, end-to-end, and online diarization.
- We survey robust speaker recognition from the perspectives of domain adaptation and speech enhancement, which are two major approaches to deal with domain mismatch and noise problems.

In the last two decades, many excellent overviews on speaker recognition have been published. This paper is fundamentally different from previous overviews. First, this paper focuses on the recently development of deep learning based speaker recognition techniques, while most previous overviews are based on traditional speaker recognition methods Anguera et al. (2012), Fazel and Chakrabartty (2011), Hansen and Hasan (2015), Kinnunen and Li (2010), Reynolds (2002), Togneri and Pullella (2011), Wu et al. (2015). Although Das, Tian, Kinnunen, and Li (2020), Irum and Salman (2019) summarized deep learning based speaker recognition methods in certain aspects, our paper summarizes different subtasks and topics from new perspectives. Specifically, Das et al. (2020) present an overview to the potential threats of adversarial attacks to speaker verification as well as the spoofing countermeasures, which is not the focus of this overview. We provide a broad and comprehensive overview to a wide aspect of speaker verification, speaker diarization, domain adaptation, most of which have not been mentioned in Irum and Salman (2019).

This article is targeted at three categories of readers: The beginners who wish to study speaker recognition, the researchers who want to learn the whole picture of speaker recognition based on deep learning, and the engineers who need to understand or implement specific algorithms for their speaker recognition related products. In addition, we assume that the readers have basic knowledge of speech signal processing, machine leaning and pattern recognition.

The rest of the survey is organized as follows. In Section 2, we give a general overview and define some notations. In Sections 3 to 10, we survey the deep learning based speaker recognition methods in various aspects. In Section 11, we summarize some speaker recognition challenges and publicly available data. Finally, we conclude this article in Section 12.

2. Overview and scope

This overview summarizes four major research branches of speaker recognition, which are speaker verification, identification, diarization, and robust speaker recognition respectively. The flowcharts of the first three branches are described in Fig. 1, while robust speaker recognition deals with the challenges of noise and domain mismatch problems. The contents of the overview are organized in Fig. 2, which are described briefly as follows.

Speaker verification aims at verifying whether an utterance is pronounced by a hypothesized speaker based on his/her pre-recorded utterances. Speaker verification algorithms can be categorized into *stage-wise* and *end-to-end* ones. A stage-wise speaker verification system usually consists of a front-end for the extraction of speaker features and a back-end for the similarity calculation of speaker features. The front-end transforms an utterance in time domain or time–frequency domain into a high-dimensional feature vector. It accounts for the recent advantage of the deep learning based speaker recognition. We survey the research on the front-end comprehensively in Sections 3 to 7. The

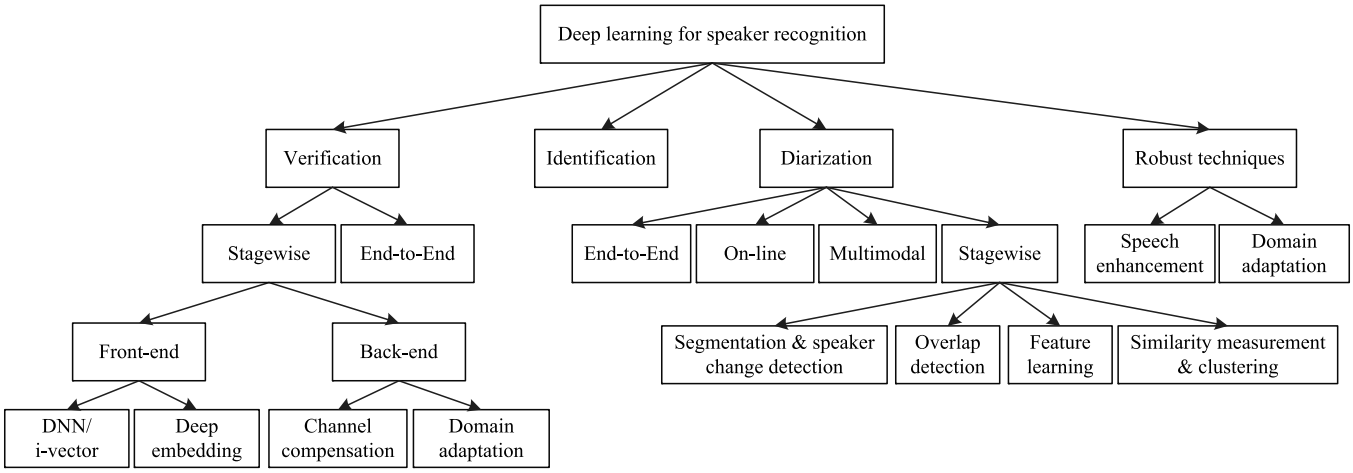

Fig. 2. Overview of deep learning based speaker recognition.

back-end first calculates a similarity score between enrollment and test speaker features and then compare the score with a threshold:

$$f(\mathbf{x}^e, \mathbf{x}^t; \mathbf{w}) \underset{H_1}{\overset{H_0}{\gtrless}} \xi \tag{1}$$

where $f(\cdot)$ denotes a function for calculating the similarity, $\mathbf{w}$ stands for the parameters of the back-end, $\mathbf{x}^e$ and $\mathbf{x}^t$ are the enrollment and test speaker features respectively, ξ is the threshold, H_0 represents the hypothesis of $\mathbf{x}^e$ and $\mathbf{x}^t$ belonging to the same speaker, and H_1 is the opposite hypothesis of H_0. One of the major responsibilities of the back-end is to compensate the channel variability and reduce interferences, e.g. language mismatch. Because most back-ends aim at alleviating the interferences, which belongs to the problem of robust speaker recognition, we put the overview of the back-ends in Section 10.

In contrast to the stage-wise techniques, end-to-end speaker verification takes a pair of speech utterances as the input, and produces their similarity score directly. Because a fundamental difference between the end-to-end speaker verification and the deep embedding techniques in the stage-wise speaker verification is the loss function, we mainly summarize the loss functions of the end-to-end speaker verification in Section 8.

Speaker identification aims at detecting the speaker identity of a test utterance $\mathbf{x}^t$ from an enrollment database $\{\mathbf{x}_k^e | k = 1, 2, \ldots, K\}$ by[1]:

$$k^* = \arg\max_k \{f(\mathbf{x}_1^e, \mathbf{x}^t; \mathbf{w}), f(\mathbf{x}_2^e, \mathbf{x}^t; \mathbf{w}), \ldots, f(\mathbf{x}_K^e, \mathbf{x}^t; \mathbf{w})\} \tag{2}$$

where $K > 1$ denotes the number of the enrollment speakers. If $\mathbf{x}^t$ can never be out of the K registered speakers, then the speaker identification problem is a closed set problem; otherwise, it is an open set problem. Comparing (1) with (2), we see that speaker verification is a special case of the open set speaker identification problem with $K = 1$, therefore, it is possible that the fundamental techniques of speaker identification and verification are similar, as what we have observed in Flemotomos and Dimitriadis (2020), Hong, Wu, Wang, and Huang (2020a), Ji, Cai, and Bo (2018), Wang, Wang, Law, Rudzicz, and Brudno (2019) and Yadav and Rai (2018). Taking this point into consideration, we make a joint overview to speaker verification and identification with an emphasis on the former.

[1] Although some work used all speakers in a given database for both training and test which is essentially regarded as a close-set speaker classification problem (Nagrani et al., 2017), most real world speaker recognition systems must be able to "enroll" and "test" new speakers dynamically.

Table 1
Organization of the contents of this paper.

Sections	Contents
1, 2	Introduction and brief overview.
3, 4, 5, 6, 7	Speaker feature extraction.
8	The loss functions of the end-to-end speaker verification.
9	Speaker diarization.
10	Robust speaker recognition.
11	Benchmark corpora.
12	Conclusions and discussions.

Speaker diarization addresses the problem of "who spoke when", which is a process of partitioning a conversation recording into several speech recordings, each of which belongs to a single speaker. As shown in Fig. 2, a conventional framework of speaker diarization is stage-wise with multiple modules. Although the stage-wise speaker verification and diarization share some common modules, e.g. voice activity detection and speaker feature extraction, they have many differences. First, speaker verification assumes that each utterance belongs to a single speaker, while the number of speakers of a conversation in speaker diarization changes case by case. Moreover, speaker verification has an explicit registration/enrollment procedure, while speaker diarization intends to detect speakers on-the-fly without an enrollment procedure. At last, overlapped speech is one of the biggest challenges of speaker diarization, while speaker verification usually assumes that the enrollment or test utterance contains a single speaker only. Therefore, we focus on reviewing the work on the above distinguished properties of the stage-wise speaker diarization in Section 9. Recently, end-to-end speaker diarization, which outputs the diarization result directly, attracted much attention. Online speaker diarization, which meets the requirement of real-world applications, is also an emerging direction. Furthermore, multimodal speaker diarization, which integrates speech with video or text signals, was also studied extensively. We review the aforementioned end-to-end, online, and multimodal speaker diarization techniques in Section 9.

Besides, speech is easily contaminated by additive noise, reverberation, channel distortions. Therefore, robust speaker recognition is also one of the main topics. It mainly includes speech enhancement and domain adaptation techniques, which will be summarized in detail in Section 10. At last, we survey benchmark corpora in Section 11.

To summarize, the aforementioned contents will be organized as listed in Table 1. The notations are summarized in Table 2.

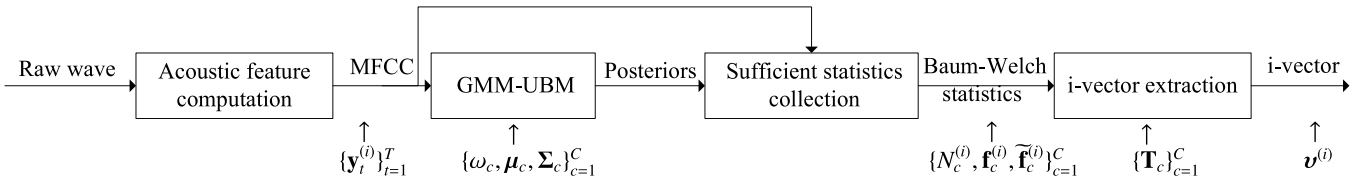

Fig. 3. The traditional GMM/i-vector framework. The term MFCC denotes Mel-frequency cepstral coefficient.

Table 2
Summary of the notations in this paper.

Notation	Description
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^d$	Set of d-dimensional real-valued vectors
$\mathbb{R}^{d_1\times d_2}$	Set of $d_1 \times d_2$ real-valued matrices
$\mathcal{Y}$	Set of acoustic features
$\mathcal{H}$	Set of the last frame-level hidden layer's outputs
$\mathcal{E}$	Set of embeddings
$\mathcal{X}$	Set of the inputs to loss functions
$\mathcal{L}$	The symbol of loss functions
$\mathbf{y}$	A frame acoustic feature
$\mathbf{h}$	A hidden feature of the last frame-level layer's output
$\mathbf{e}$	An embedding feature of the embedding layer's output
$\mathbf{x}$	An input feature to loss functions
$\mathbf{u}$	An output of the temporal pooling layer
t, T	Index and total number of frames in an utterance
i, I	Index and total number of the utterance
j, J	Index and total number of speakers in the training set
$\|\cdot\|$	The ℓ_2 norm
$\odot$	The Hadamard product
$\delta(\cdot)$	Indicator function
$(\cdot)^T$	The transform of matrix or vector

3. Speaker feature extraction with DNN/i-vector

In this section, we first introduce two main streams of the deep learning based improvement to the i-vector framework in Section 3.1, and then comprehensively review the two streams in Sections 3.2 and 3.3 respectively. Finally, we make some discussions to the DNN/i-vector in Section 3.4.

3.1. From GMM/i-vector to DNN/i-vector

The performance of the conventional GMM-UBM based speaker recognition is largely affected by the speaker and channel variations of utterances. To address this issue, Dehak et al. (2010) proposed to reduce the high-dimensional GMM-UBM supervectors into low-dimensional vectors, named *i-vectors* by factor analysis. The GMM/i-vector system eliminates the within-speaker and channel variabilities effectively, which leads to significant performance improvement.

The GMM/i-vector system is shown in Fig. 3. We assume that $\mathcal{Y} = \{\mathbf{y}_t^{(i)} \in \mathbb{R}^{d_1} | t = 1, 2, \ldots, T\}$ represents the ith $(i = 1, 2, \ldots, I)$ utterance of T successive Mel-frequency cepstral coefficient (MFCC) frames, and $\Omega = \{\omega_c \in \mathbb{R}, \boldsymbol{\mu}_c \in \mathbb{R}^{d_1}, \boldsymbol{\Sigma}_c \in \mathbb{R}^{d_1\times d_1} | c = 1, 2, \ldots, C\}$ $(\omega_c \geq 0$ for all c, and $\sum_{c=1}^C \omega_c = 1)$ denotes a GMM-UBM model where C is the total number of components and ω_c, $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the weight, mean, and covariance matrix of the cth component respectively. Then, $\mathbf{y}_t^{(i)}$ is assumed to be generated by the following distribution (Lei, Scheffer, Ferrer, & McLaren, 2014; Snyder, 2020):

$$\mathbf{y}_t^{(i)} \sim \sum_{c=1}^{C} \omega_c N(\boldsymbol{\mu}_c + \mathbf{T}_c \boldsymbol{\upsilon}^{(i)}, \boldsymbol{\Sigma}_c) \tag{3}$$

where $\{\mathbf{T}_c\}_{c=1}^C$ is a so called total variability subspace, $\boldsymbol{\upsilon}^{(i)}$ is a segment-specific standard normal-distributed latent vector. The i-vector used to represent the speech signal is the maximum a posterior (MAP) point estimate of the latent vector $\boldsymbol{\upsilon}^{(i)}$, and it can be regarded as a kind of "*speaker embedding*".[2]

Given a speech segment, the following sufficient statistics can be accumulated from the GMM-UBM:

$$N_c^{(i)} = \sum_{t=1}^{T} p(c|\mathbf{y}_t^{(i)}) \tag{4}$$

$$\mathbf{f}_c^{(i)} = \sum_{t=1}^{T} p(c|\mathbf{y}_t^{(i)}) \mathbf{y}_t^{(i)} \tag{5}$$

$$\mathbf{S}_c^{(i)} = \sum_{t=1}^{T} p(c|\mathbf{y}_t^{(i)}) \mathbf{y}_t^{(i)} (\mathbf{y}_t^{(i)})^T \tag{6}$$

where $p(c|\mathbf{y}_t^{(i)}) = \frac{\omega_c N(\mathbf{y}_t^{(i)}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{\sum_{c'=1}^C \omega_c N(\mathbf{y}_t^{(i)}; \boldsymbol{\mu}_{c'}, \boldsymbol{\Sigma}_{c'})}$ denotes the posterior probability of $\mathbf{y}_t^{(i)}$ against the cth Gaussian component. These sufficient statistics are all that are needed to train the subspace $\{\mathbf{T}_c\}_{c=1}^C$ and extract the i-vector $\boldsymbol{\upsilon}^{(i)}$ (Lei, Scheffer, Ferrer, & McLaren, 2014). See Dehak et al. (2010) and Kenny, Ouellet, Dehak, Gupta, and Dumouchel (2008) for the details of training $\mathbf{T}_c$ and estimating the i-vectors.

Motivated by the success of deep learning for speech recognition, many efforts have been made to replace the GMM-UBM module of the GMM/i-vector system by DNN, which can be categorized to two main streams—DNN-UBM/i-vector and DNN based bottleneck feature (DNN-BNF)/i-vector. The two main streams will be presented in detail in the following two subsections, with selected references summarized in Table 3.

3.2. DNN-UBM/i-vector

From (4), (5), and (6), one can see that only the posteriors of speech frames are needed to collect sufficient statistics for producing the i-vectors. Thus, we can use any probabilistic models beyond GMM-UBM to produce the posteriors theoretically (Lei, Scheffer, Ferrer, & McLaren, 2014). Motivated by this insight, Lei, Scheffer, Ferrer, and McLaren (2014) proposed the DNN-UBM/i-vector framework (Fig. 4) which takes a DNN acoustic model trained for ASR, denoted as DNN-UBM, to generate the posterior probabilities instead of GMM-UBM.

Specifically, DNN-UBM uses a set of senones $\mathcal{Q} = \{Q_c | c = 1, 2, \ldots, C\}$, e.g., the tied-triphone states, to mimic the mixture components of the GMM-UBM. It first trains a DNN-based ASR acoustic model to align each training frame with a senone, and then generates the posterior probabilities of each frame over the senones from the softmax output layer of the DNN acoustic model. The posteriors can be directly applied to (4)–(6) to extract the DNN-UBM based i-vector. Due to the strong representation ability of DNN over GMM, DNN-UBM/i-vector yields 30% relative equal error rate (EER) reduction over GMM/i-vector on

[2] In this paper, the 'embedding' denotes the problem of learning a vector space where speakers are "embedded". The i-vectors, d-vectors (introduced in Section 4.1.1), and x-vectors (introduced in Section 4.1.2) are different embedding models for learning the vector spaces.

Table 3
Two main streams of the DNN/i-vector techniques.

Approaches	References
DNN-UBM/i-vector	Chen et al. (2015), Dey, Madikeri, Ferras, and Motlicek (2016), Dey, Motlicek, Madikeri, and Ferras (2017), Garcia-Romero and McCree (2015), Kenny, Stafylakis, Ouellet, Gupta, and Alam (2014), Lei, Ferrer, McLaren, and Scheffer (2014), Lei, Scheffer, Ferrer, and McLaren (2014), McLaren, Lei, and Ferrer (2015), McLaren, Lei, Scheffer, and Ferrer (2014), Richardson, Reynolds, and Dehak (2015a, 2015b), Sadjadi, Ganapathy, and Pelecanos (2016), Snyder, Garcia-Romero, and Povey (2015), Zeinali, Burget, Sameti, Glembek, and Plchot (2016), Zeinali, Sameti, Burget, et al. (2017), Zheng, Zhang, and Liu (2015)
DNN-BNF/i-vector	Do, Barras, Le, and Sarkar (2013), Ghalehjegh and Rose (2015), Lozano-Diez et al. (2016), McLaren, Ferrer, and Lawson (2016), McLaren et al. (2015), Richardson et al. (2015a, 2015b), Sarkar, Do, Le, and Barras (2014), Zeinali et al. (2016)

the telephone condition of the 2012 NIST speaker recognition evaluation (SRE) (Lei, Scheffer, Ferrer, & McLaren, 2014). Later on, the authors in Lei, Ferrer, McLaren, and Scheffer (2014) and McLaren et al. (2015, 2014) further analyzed the performance of the DNN-UBM/i-vector in microphone and noisy conditions.

A lot of further studies bloomed the DNN-UBM/i-vector related techniques. For example, Richardson et al. (2015a, 2015b) proposed to use a single ASR-DNN for both the speaker and language recognition tasks simultaneously. Additionally, Snyder et al. (2015) employed a time delay deep neural network (TDNN), which was originally applied to speech recognition, to compute the posteriors. It achieved the state-of-the-art performance on the NIST SRE10 corpus at the time. As a third instance, Zheng et al. (2015) replaced the feedforward DNN by a long short-term memory (LSTM) recurrent neural network (RNN). The last but not all, Garcia-Romero and McCree (2015) studied a number of open issues relating to performance, computational complexity, and applicability of different types of DNNs.

The advantage of the DNN acoustic model may be brought by its strong ability in modeling content-related phonetic states explicitly, which not only generates highly compact representation of data but also provides precise frame alignment. This advantage is particularly apparent in text-dependent speaker verification (Chen et al., 2015; Dey et al., 2016, 2017; Zeinali et al., 2016, 2017). However, this comes at the cost of greatly increased computational complexity over the traditional GMM-UBM/i-vector systems (Snyder et al., 2015; Snyder, Garcia-Romero, Povey, & Khudanpur, 2017), since that a DNN usually has more parameters than GMM. In addition, the training of the DNN based acoustic model requires a large number of labeled training data.

To overcome the computational complexity, a supervised GMM-UBM was also investigated based on the DNN acoustic model (Snyder et al., 2015). In specific, a GMM is obtained by:

$$\begin{aligned}\gamma_{ct}^{(i)} &= p(c|\bar{\mathbf{y}}_t^{(i)}) \\ \omega_c &= \sum_{i,t} \gamma_{ct}^{(i)} \\ \boldsymbol{\mu}_c &= \frac{\sum_{i,t} \gamma_{ct}^{(i)} \mathbf{y}_t^{(i)}}{\sum_{i,t} \gamma_{ct}^{(i)}} \\ \boldsymbol{\Sigma}_c &= \frac{\sum_{i,t} \gamma_{ct}^{(i)} \mathbf{y}_t^{(i)} (\mathbf{y}_t^{(i)})^T}{\sum_{i,t} \gamma_{ct}^{(i)}} - \boldsymbol{\mu}_c \boldsymbol{\mu}_c^T \end{aligned} \tag{7}$$

where $\bar{\mathbf{y}}_t^{(i)}$ and $\mathbf{y}_t^{(i)}$ denote the acoustic features for ASR and speaker recognition respectively, and $p(c|\bar{\mathbf{y}}_t^{(i)})$ is the posterior probability corresponding to the cth senones. By this way, the supervised-GMM maintains the training computational complexity of the traditional unsupervised-GMM, with a 20% relative EER reduction on the NIST SRE10 corpus (Snyder et al., 2015). Similar idea was also studied in Lei, Scheffer, Ferrer, and McLaren (2014), though no performance improvement over the baseline is observed. Although the supervised-GMM reduces the training computational complexity, training the DNN acoustic model still needs a large amount of labeled training data.

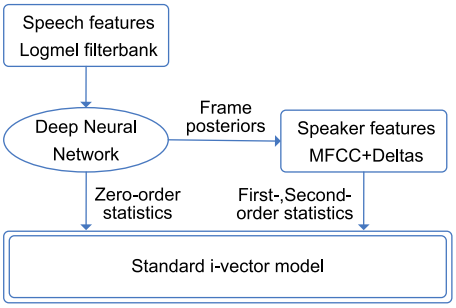

Fig. 4. DNN-UBM/i-vector. The posteriors are produced from the DNN acoustic model of an automatic speech recognition (ASR) system that is trained with, e.g. Logmel filterbank features. On the contrary, the sufficient statistics are computed from a speaker verification (SV) system that is trained with, e.g. MFCC which is not necessarily the same as the features for ASR. That is to say, one does not have to find a feature that works well for both ASR and SV in this framework.
Source: From Lei, Scheffer, Ferrer, and McLaren (2014).

3.3. DNN-BNF/i-vector

The fundamental idea of DNN-BNF/i-vector is to extract a compact feature from the bottleneck layer of a DNN as the input of the factor analysis, where the bottleneck layer is a special hidden layer of the DNN that has much less hidden units than the other hidden layers. In practice, DNN-BNF/i-vector has many variants, as we have summarized in Fig. 5. Like DNN-UBM/i-vector, the deep model in DNN-BNF/i-vector is mainly trained to discriminate senones (Lozano-Diez et al., 2016; McLaren, Ferrer, & Lawson, 2016; McLaren et al., 2015; Richardson et al., 2015a, 2015b; Zeinali et al., 2016) or phonemes (Do et al., 2013; Sarkar et al., 2014).

The input of the factor analysis can be either the bottleneck feature (BNF) produced from the bottleneck layer, a concatenation of BNF with other acoustic feature (Do et al., 2013; Sarkar et al., 2014), or a post-processed feature by principal components analysis (PCA) or linear discriminant analysis (LDA) (Do et al., 2013; Sarkar et al., 2014). One can find that no matter whether we apply BNF alone (Richardson et al., 2015a) or concatenate it with other acoustic features (McLaren et al., 2015), DNN-BNF/i-vector can significantly outperform the conventional GMM/i-vector, which indicates the effectiveness of the framework (Richardson et al., 2015b).

However, it is unclear why a deep model trained to discriminate phonemes or senones can produce speaker-sensitive BNF.

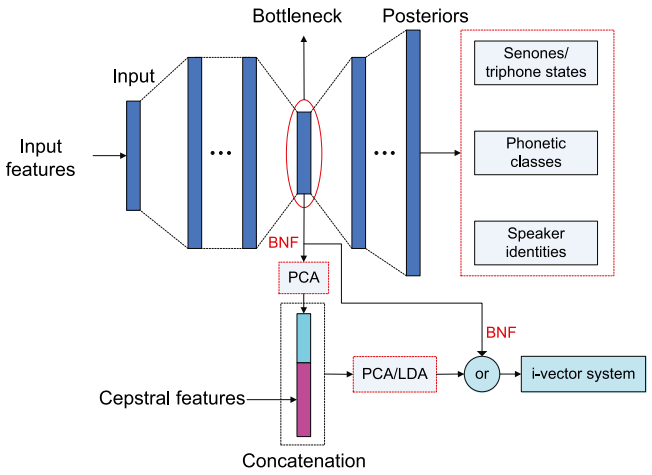

Fig. 5. The DNN-BNF/i-vector framework. The framework is a summarization of related work. The terms "PCA", "LDA", and "BNF" are short for principal components analysis, linear discriminant analysis, and bottleneck feature respectively.

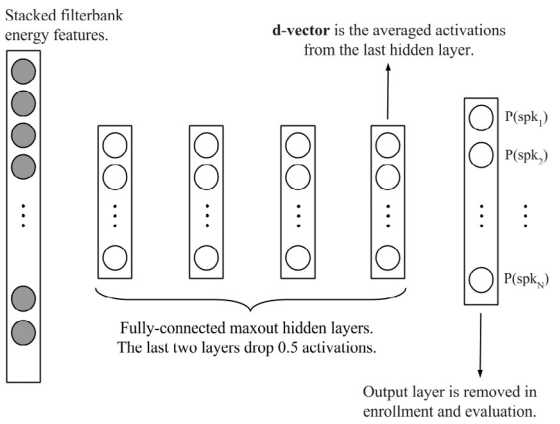

Fig. 6. Diagram of the d-vector framework. *Source:* From Variani et al. (2014).

To address this issue, the authors of McLaren, Ferrer, and Lawson (2016) assumed that speaker information is traded for dense phonetic information when the bottleneck layer moves towards the DNN output layer. Under this hypothesis, they experimentally analyzed the role of BNF by placing the bottleneck layer at different depths of the DNN. They found that, if the training and test conditions match, the closer the bottleneck layer is to the output layer, the better the performance is; otherwise, the bottleneck layer should be placed around the middle of the DNN. The authors of Lozano-Diez et al. (2016) explored whether weakening the accuracy of the acoustic model on speech recognition yields better BNF for speaker recognition. They analyzed the speaker recognition performance in different respects of the acoustic model, including under-trained DNN, different inputs, and different feature normalization strategies. Results indicate that high speech recognition performance in terms of phonetic accuracy does not necessarily imply increased speaker recognition accuracy. In addition, Ghalehjegh and Rose (2015) proposed to take speaker identity as the training target, under the conjecture that this training target should be able to improve the robustness of the phonetic variability of BNF.

3.4. Discussion to the DNN/i-vector

It is known that a major difference between DNN-UBM and GMM-UBM is that DNN-UBM is a discriminant model, while GMM-UBM is a generative one. DNN-UBM is more powerful than GMM-UBM in modeling a complicated data distribution (Lei, Scheffer, Ferrer, & McLaren, 2014; Snyder et al., 2015). Moreover, the DNN acoustic model is trained to align each speech frame to its corresponding senone in a supervised fashion. Its output nodes have a clear physical explanation. It mines the pronunciation characteristics of speakers. On the contrary, GMM-UBM is trained by the expectation-maximum algorithm in an unsupervised manner. Its mixtures have no inherent meaning. Although DNN-UBM/i-vector needs labeled training data and heavier computation power than GMM-UBM, it does yield excellent performance. In addition, many corpora are also developed for the demand of training strong DNN, which will be reviewed in Section 11.

To demonstrate general performance differences of DNN/i-vector and GMM/i-vector, some carefully selected experimental results from literatures are listed in Table 4. Compared to the GMM-UBM/i-vector baseline, one can find that DNN-UBM/i-vector achieves more than 20% relative EER reduction over GMM-UBM/i-vector. In addition, the supervised GMM-UBM in (7) can also get 20% relative improvement according to the fourth row. Finally, from the last two rows, one can see that, when taking DNN-BNF and MFCC as the input features of the GMM-UBM/i-vector respectively, the former achieves better performance than the latter.

It should be noted that, as far as we know, different test conditions may yield slightly different conclusions from those in Table 4. However, to our knowledge, the results in the table can be a representative of the research trend.

4. Speaker feature extraction with deep embedding

In this section, we first introduce two representative deep embeddings—d-vector and x-vector in Section 4.1 with some discussions in Section 4.2, and then identify their key components in Section 4.3, which provide a taxonomy to existing algorithms.

4.1. Two seminal work of deep embeddings

4.1.1. Frame-level embedding—d-vector

D-vector is one of the earliest DNN-based embeddings (Variani et al., 2014). The core idea of d-vector is to assign the ground-truth speaker identity of a training utterance as the labels of the training frames belonging to the utterance in the training stage, which transforms the model training as a classification problem. As shown in Fig. 6, d-vector expands each training frame with its context, and employs a *maxout* DNN to classify the frames of a training utterance to the speaker identity of the utterance, where the DNN takes softmax as the output layer to minimize the cross-entropy loss between the ground-truth labels of the frames and the network output.

In the test stage, d-vector takes the output activation of each frame from the last hidden layer of the DNN as the deep embedding feature of the frame, and averages the deep embedding features of all frames of an utterance as a new compact representation of the utterance, named *d-vector*. An underlying hypothesis of d-vector is that the compact representation space produced from a development set may generalize well to unseen speakers in the test stage.

Table 4
Comparison results between DNN/i-vector and GMM/i-vector. Each row denotes a comparison. The last three columns list the EER of the main models, the EER of the baselines, and the relative EER reductions, respectively. **The results across rows are not comparable**, since they are collected from different references, and their comparisons are not apple-to-apple comparisons.

Comparisons		Test dataset [condition]	EER		
Main models	Baselines		Main	Baseline	Relative reduction
DNN-UBM (Lei, Scheffer, Ferrer, & McLaren, 2014)	GMM-UBM	NIST SRE12 C2	1.39%	1.81%	**23%**
DNN-UBM (Lei, Scheffer, Ferrer, & McLaren, 2014)	GMM-UBM	NIST SRE12 C5	1.92%	2.55%	**25%**
TDNN-UBM (Snyder et al., 2015)	GMM-UBM	NIST SRE10 C5	1.20%	2.42%	**50%**
Sup-GMM-UBM (Snyder et al., 2015)	GMM-UBM	NIST SRE10 C5	1.94%	2.42%	**20%**
BNF (Richardson et al., 2015b)	MFCC	In-domain DAC13	2.00%	2.71%	**26%**
BNF (Richardson et al., 2015b)	MFCC	Out-domain DAC13	2.79%	6.18%	**55%**

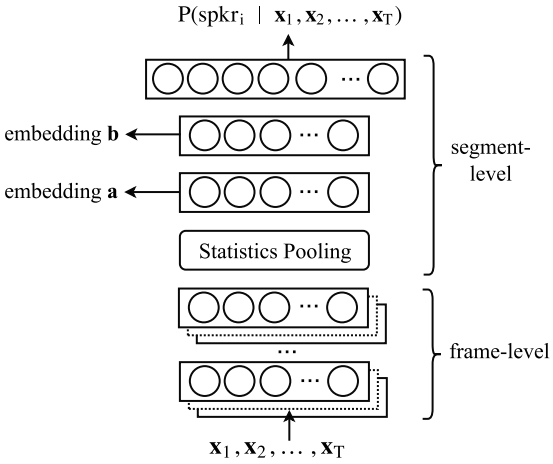

Fig. 7. Diagram of the DNN model for extracting x-vectors. Note that segment-level embeddings (e.g., **a** or **b**) can be extracted from any layer of the network after the statistics pooling layer (Snyder et al., 2017). Snyder et al. (2018) where the name "x-vector" comes from uses of the embedding **a** as the speaker feature. *Source:* From Snyder et al. (2017).

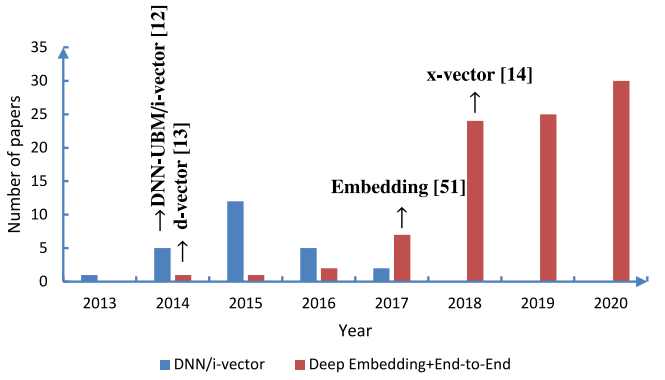

Fig. 8. Statistics of the published papers on DNN/i-vector and deep embedding cited by this article.

4.1.2. Segment-level embedding—x-vector

X-vector (Snyder et al., 2017, 2018) is an important evolution of d-vector that evolves speaker recognition from frame-by-frame speaker labels to utterance-level speaker labels with an aggregation process. The network structure of x-vector is shown in Fig. 7. It first extracts frame-level embeddings of speech frames by time-delay layers, then concatenates the mean and standard deviation of the frame-level embeddings of an utterance as a segment-level (a.k.a., utterance-level) feature by a statistical pooling layer, and finally classifies the segment-level feature to its speaker by a standard feedforward network. The time-delay layers, statistical pooling layer, and feedforward network are jointly trained. *X-vector* is defined as the segment-level speaker embedding produced from the second to last hidden layer of the feedforward network, i.e. the variable **a** in Fig. 7.

The authors in Snyder et al. (2018) found that data augmentation is important in improving the performance of x-vector. We will introduce the data augmentation techniques in Section 10.3.

4.2. Discussion to the speaker embedding

Similar to the i-vector, the d-vector and x-vector are also a kind of speaker embedding, which discriminatively embeds speakers into a vector space by using DNNs. We call this type of speaker embedding as *deep speaker embedding*, or *deep embedding* for short. The main characteristics between different speaker embeddings are summarized in Table 5. Compared to the traditional GMM-UBM/i-vector, the deep embedding is a discriminant model and trained in a supervised fashion. Compared to DNN-UBM, its training data does not need phonetic-level labels. Therefore, the training of the deep embedding is much simpler than that of DNN-UBM and DNN-BNF. In addition, the deep embedding is a new framework, while DNN-UBM/i-vector and DNN-BNF/i-vector are hybrid ones.

Some experimental results on deep embedding are listed in Table 6. From the table, one can find that, the d-vector alone yields higher EER than the i-vector. When fusing the d-vector and i-vector, the combined system achieves 14% and 25% relative EER reduction in clean and noisy test conditions respectively over the i-vector. The "embedding a+b" model, which is the predecessor of the x-vector, achieves lower EER than the GMM-UBM/i-vector baseline on the 10-second short utterances of NIST SRE10, and higher EER than the latter on the 60-second long utterances of NIST SRE10. With enlarged training data and data augmentation, the x-vector achieves significant performance improvement over the GMM-UBM/i-vector.

Fig. 8 shows the number of the related papers. We observe the following phenomena. First, the d-vector and DNN-UBM/i-vector was proposed both in 2014, where the former achieved better performance at the time. Second, the research on DNN/i-vector was mainly conducted in the first few years after its appearance, and then became less studied. Third, the research on deep embedding becomes bloom along with its performance improvement after that the x-vector achieved the state-of-the-art performance. At present, the deep embedding is the trend of speaker recognition, which has been developed in several aspects as summarized in the following subsection.

4.3. Four key components of deep embedding

Motivated by the seminal work d-vector and x-vector, many deep embedding techniques were proposed, most of which are composed of four key components—network input, network structure, temporal pooling, and training objective. These components include but not limited to the following contents:

Table 5
Characteristics of different speaker embeddings and their favorite back-ends.

Model name	Type	Training strategy	Label for model training	Back-end name	Label for back-end training
GMM-UBM/i-vector	Generative/Generative	Unsupervised/Unsupervised	✗/✗	PLDA	Speaker identity
DNN-UBM/i-vector	Discriminative/Generative	Supervised /Unsupervised	Phonetic labels/✗	PLDA	Speaker identity
DNN-BNF/i-vector	Discriminative/Generative	Supervised/Unsupervised	Phonetic labels/✗	PLDA	Speaker identity
D-vector	Discriminative	Supervised	Speaker identity	Cosine	✗
X-vector	Discriminative	Supervised	Speaker identity	PLDA	Speaker identity

Table 6
Selected results on deep embedding in literature. Each row represents a comparison. The results across rows are not comparable.

Comparison methods		Test dataset [condition]	EER		
Deep embedding	Baseline		Deep embedding	Baseline	Relative reduction
d-vector (Variani et al., 2014)	GMM-UBM/i-vector	Google data	4.54%	2.83%	**−37%**
d-vector+i-vector (Variani et al., 2014)	GMM-UBM/i-vector	Google data [clean, noisy]	–	–	**[14%, 25%]**
embedding **a+b** (in Fig. 7) (Snyder et al., 2017)	GMM-UBM/i-vector	NIST SRE10 [10s-10s, 60s]	[7.9%, 2.9%]	[11.0%, 2.3%]	**[28%, −21%]**
embedding **a+b** (in Fig. 7) (Snyder et al., 2017)	GMM-UBM/i-vector	NIST SRE16 [Cantonese, Tagalog]	[6.5%, 16.3%]	[8.3%, 17.6%]	**[22%, 7%]**
x-vector (embedding **a**) (Snyder et al., 2018)	GMM-UBM/i-vector	SITW Core [PLDA and extractor aug., Incl. VoxCeleb]	[6.00%, 4.16%]	[8.04%, 7.45%]	**[25%, 44%]**
x-vector (embedding **a**) (Snyder et al., 2018)	GMM-UBM/i-vector	SRE16 Cantonese [PLDA and extractor aug., Incl. VoxCeleb]	[5.86%, 5.71%]	[8.95%, 9.23%]	**[34%, 38%]**

- *Network inputs and structures*: The network input can be categorized into two classes—raw wave signals in time domain and acoustic features in time–frequency domain, including spectrogram, Mel-filterbanks (f-bank), and MFCC. The network structure is diverse, which is rooted essentially at DNN, RNN/LSTM, and CNN. Because the network input and structure were jointly designed case by case in practice, we will jointly summarize them in Section 5.
- *Temporal pooling*: Temporal pooling represents the transition layer of a neural network that transforms frame-level embedding features to utterance-level embedding features. The temporal pooling strategies consist of two classes—statistical pooling and learning based pooling. We will introduce them in Section 6.
- *Objective functions*: Objective functions affect the effectiveness of speaker recognition much. Both d-vector and x-vector adopt softmax as the output layer and take the cross-entropy minimization as the objective function, which may not be optimal. Recently, many objectives were designed to further improve the performance. We will survey the objective functions in Section 7.

5. Deep embedding: network structures and inputs

Although deep neural networks can be divided roughly into DNN, CNN, and RNN/LSTM structures, the network structure and input for speaker recognition are quite flexible. Each component of a network has many candidates. For example, the hidden layer of a neural network may be a standard convolutional layer (Bhattacharya et al., 2017), a dilated convolution layer (Gao et al., 2018), a LSTM layer (Jung et al., 2018a), a gated recurrent unit (GRU) layer (Jung, Heo, Kim, Shim, & Yu, 2019), a multi-head attention layer, a fully-connected layer, and even a combination of these different layers (Jung, Heo, Kim, Shim, & Yu, 2019; Jung et al., 2018a), etc. The activation functions can be Sigmoid, Rectified Linear Unit (ReLU), Leaky ReLU, or Parametric Rectified Linear Unit (PReLU) etc. Besides, the topology of a network and connection mode between layers are all variables. Even the number of layers and number of hidden units at a layer can also affect the performance. To prevent enumerating the networks case by case, here we first review some commonly used networks for the speaker feature extraction, and then briefly review their inputs.

Time delay neural network (TDNN) (Snyder et al., 2017): TDNN takes a one-dimensional convolution structure along the time axis as a feature extractor (Peddinti, Povey, & Khudanpur, 2015). It is adopted by the well known x-vector, as shown in Fig. 7. Due to the success of the x-vector (Snyder et al., 2019, 2018), TDNN becomes one of the most popular structures for speaker recognition. For example, Liu et al. (2018) introduced phonetic information to the TDNN architecture based embedding extractor. Stafylakis, Rohdin, Plchot, Mizera, and Burget (2019) trained a TDNN embedding extractor without speaker labels via self-supervised training. Zhu and Mak (2020a, 2020b) explored the effectiveness of the orthogonality regularization by TDNN. Generally, TDNN has been frequently used as a framework to study other key components of the deep embedding models, such as the temporal pooling layers (Okabe et al., 2018; Zhu et al., 2018) and objective functions (Bai et al., 2020a; Li, Tang, Shi, & Wang, 2019; Xiang et al., 2019).

The TDNN structure has also been intensively improved. For instance, an extended TDNN architecture (E-TDNN) was introduced in Snyder et al. (2019), which greatly outperforms the x-vector baseline (Snyder et al., 2018). It adopts a slightly wider temporal context than TDNN, and interleaves affine layers in between the convolutional layers (Garcia-Romero, McCree, Snyder, & Sell, 2020). Povey et al. (2018) developed a factorized TDNN (F-TDNN) to reduce the number of parameters. It factorizes the weight matrix of each TDNN layer into the product of two low-rank matrices. It further constrains the first low-rank matrix to be semi-orthogonal under the assumption that the semi-orthogonal constraint prevents information loss. The application of F-TDNN to deep embedding was also investigated (Garcia-Romero, McCree, Snyder, & Sell, 2020; Snyder et al., 2019; Villalba et al., 2019). Some other parameter reduction works can be found in Georges, Huang, and Bocklet (2020) and Yu and Li (2020). Recently, Hong et al. (2020b) integrated TDNN with statistics pooling at each layer for compensating the variation of temporal context in the frame-level transforms. Similarly, Chen et al. (2019) and Tang et al. (2019) inserted LSTM layers into TDNN to capture the temporal information for remedying the weakness of TDNN whose time delay layers focus on local patterns only. Li et al.

Table 7
A brief summary of the inputs and neural network structures in deep speaker feature extraction.

Inputs	CNN	LSTM	Hybrid structures
Wave	Others (Muckenhirn, Doss, & Marcell, 2018; Ravanelli & Bengio, 2018).	–	CNN-LSTM (Jung, Heo, Yang, Shim, & Yu, 2018a, 2018b); CNN-GRU (Jung, Heo, Kim, Shim, & Yu, 2019; Jung, Heo, Shim, & Yu, 2019).
Spectrogram	ResNet (Chung, Nagrani, & Zisserman, 2018; Xie, Nagrani, Chung, & Zisserman, 2019; Yadav & Rai, 2020; Yu, Fan, & Li, 2019); VGGNet (Nagrani et al., 2017; Yadav & Rai, 2018); Inception-resnet-v1 (Zhang & Koishida, 2017; Zhang, Koishida, & Hansen, 2018).	–	CNN-GRU (Zhang et al., 2019)
F-bank	TDNN (Garcia-Romero, McCree, Snyder, & Sell, 2020; Snyder et al., 2018; Zhu & Mak, 2020a, 2020b); ResNet (Garcia-Romero, Sell, & Mccree, 2020; Kim, Kim, Kim, & Choi, 2019; Li et al., 2017; Wang, Yao, Li, & Fang, 2020c); VGGNet (Bhattacharya, Alam, & Kenny, 2017); Inception-resnet-v1 (Li et al., 2019; Li, Tuo, Su, Li, & Yu, 2018; Zhang, Koishida, & Hansen, 2018); Others (Li, Chen, Shi, Tang, & Wang, 2017; Torfi, Dawson, & Nasrabadi, 2018).	rahman Chowdhury, Wang, Moreno, and Wan (2018), Heigold, Moreno, Bengio, and Shazeer (2016), Wan, Wang, Papir, and Moreno (2018).	BLSTM-ResNet (Zhao, Zhou, Chen, & Wu, 2020), TDNN-LSTM (Tang, Ding, Huang, He, & Zhou, 2019)
MFCC	TDNN (Bai, Zhang, & Chen, 2020a; Garcia-Romero et al., 2020; Hong, Wu, Wang, & Huang, 2020b; Li, Tang, Shi, & Wang, 2019; Li et al., 2020; Liu, He, Liu, & Johnson, 2018; Okabe, Koshinaka, & Shinoda, 2018; Snyder et al., 2017, 2019; Villalba et al., 2019; Xiang, Wang, Huang, Qian, & Yu, 2019; Zhu, Ko, Snyder, Mak, & Povey, 2018); ResNet (Zhou, Jiang, Li, Li, & Hong, 2019); Others (Gao, Song, McLoughlin, Guo, & Dai, 2018; Jiang, Song, McLoughlin, Gao, & Dai, 2019).	–	TDNN-LSTM (Chen et al., 2019)

(2020) alleviated the mismatch problem between training and evaluation by incorporating Bayesian neural networks into TDNN.

Residual networks (ResNet) (He, Zhang, Ren, & Sun, 2016): it is another popular structure in speaker embedding. Its trunk architecture is a 2-dimensional CNN with convolutions in both the time and frequency domains. Some work directly used the standard ResNet as their speaker feature extractors (Chung et al., 2018; Li et al., 2017; Wang et al., 2020c; Yu et al., 2019). Some other work employed ResNet as a backbone and modified it for specific purposes or applications (Garcia-Romero et al., 2020; Kim et al., 2019; Xie et al., 2019; Yadav & Rai, 2020; Zhao et al., 2020; Zhou, Jiang, Li, Li, & Hong, 2019). For example, to reduce the number of parameters, Xie et al. (2019) modified the standard ResNet-34 to a *thin* ResNet by cutting down the number of channels in each residual block. The authors in Zhao et al. (2020) combined bi-directional LSTM (BLSTM) and ResNet into a unified architecture, where the BLSTM is used to model long temporal contexts. The authors in Zhou, Jiang, Li, Li, and Hong (2019) incorporated a so-called "squeeze-and-excitation" block into ResNet.

Raw wave neural networks (Jung, Heo, Kim, Shim, & Yu, 2019; Jung, Heo, Shim, & Yu, 2019; Jung et al., 2018a, 2018b; weon Jung, bin Kim, jin Shim, ho Kim, & Yu, 2020; Lin & Mak, 2020; Muckenhirn et al., 2018; Ravanelli & Bengio, 2018): some work takes raw waves in the time domain as the input, which aims to extract learnable acoustic features instead of handcrafted features. For example, Muckenhirn et al. (2018) applied CNN to capture raw speech signal. The experimental results indicate that the filters of the first convolution layer give emphasis to speaker information in low frequency regions. The authors in Ravanelli and Bengio (2018) believed that the first layer is critical for the waveform-based CNNs, since it not only deals with high-dimensional inputs, but suffers more from the gradient vanishing problem than the other layers. Therefore, they proposed a SincNet architecture based on parametrized sinc functions, where only low and high cutoff frequencies of band-pass filters are learned from data (Ravanelli & Bengio, 2018). In Jung et al. (2018b), the authors thought that the difficulty of processing raw audio signals by DNN is mainly caused by the fluctuating scales of the signals. To stabilize the scales, they employed a convolutional layer, named pre-emphasis layer, to mimic the well-known signal pre-emphasis technique $p(t) = s(t) - \alpha s(t-1)$. They also made several improvements to the original raw wave network (Jung, Heo, Kim, Shim, & Yu, 2019; Jung, Heo, Shim, & Yu, 2019; weon Jung et al., 2020) which results in excellent performance. Lin and Mak (2020) designed a *Wav2Spk* architecture to learn speaker embeddings from waveforms, where the traditional MFCC extraction, voice activity detection, and cepstral mean and variance normalization are replaced by a feature encoder, a temporal gating unit and an instance normalization scheme respectively. Wav2Spk performs better than the convention x-vector network.

Other neural networks: in addition to TDNN and ResNet, many other well-known neural network architectures have also been applied to speaker recognition, including VGGNet (Bhattacharya et al., 2017; Nagrani et al., 2017; Yadav & Rai, 2018), Inception-resnet-v1 (Li et al., 2019, 2018; Zhang & Koishida, 2017; Zhang, Koishida, & Hansen, 2018), BERT (Ling, Salazar, Liu, & Kirchhoff, 2020), and Transformer (Safari, India, & Hernando, 2020). Besides, recurrent neural networks, such as LSTM and gated recurrent units, are often used for text-dependent speaker verification (rahman Chowdhury et al., 2018; Heigold et al., 2016; Wan et al., 2018). The CNN models can also be improved by inserting LSTM or gated recurrent units into the backbone networks (Chen et al., 2019; Jung, Heo, Kim, Shim, & Yu, 2019; Jung, Heo, Shim, & Yu, 2019; Jung et al., 2018a, 2018b; Tang et al., 2019; Zhang et al., 2019; Zhao et al., 2020). Finally, apart from the above handcrafted neural architectures, neural architecture search was also recently applied to speaker recognition (Ding, Chen, Gong, Zha, & Wang, 2020; Qu, Wang, & Xiao, 2020).

Neural network inputs: Table 7 provides a summary to the common inputs and neural networks for the deep embedding based speaker feature extraction. From the table, one can see that CNN-based neural networks and f-bank/MFCC acoustic features are becoming popular, while some 2-dimensional convolution structures, e.g. ResNet, use spectrogram as the input feature. In addition to the above common inputs, such as MFCC, spectrum and mel-filterbanks, Liu, Sahidullah, and Kinnunen (2020) recently presented an extensive re-assessment of 14 acoustic feature extractors. They found that the acoustic features equipped

with the techniques of spectral centroids, group delay function, and integrated noise suppression provide promising alternatives to MFCC.

5.1. Discussion to the networks

The network structure plays a key role on performance. For example, as shown in Table 8, E-TDNN and F-TDNN significantly reduced EER on the SITW dataset (McLaren et al., 2016), where F-TDNN achieves more than 40% relative EER reduction over the original TDNN. Although this promotion is not consistent across all datasets (Villalba et al., 2020), it demonstrates the importance of the network structure on performance.

For the acoustic features, the delta and double-delta features are helpful in statistical model based speaker recognition, e.g. the GMM-UBM/i-vector. However, they are not very effective in convolution and time-delay neural networks. This may be caused by that, the statistical model needs the delta and double-delta operations to capture the time dependency between frames, while the neural networks are able to achieve this goal intrinsically.

Although the deep embedding networks have achieved a great success, in our view, the following aspects can be further studied. First, the raw wave networks did not attract much attention. The mainstream of speaker recognition still adopts handcrafted features, which may lose useful information, e.g. the phase information, and finally may result in suboptimal performance as what we have observed in speech separation. Second, the model size and inference efficiency, which is important for the devices with limited computation source, e.g. edge or mobile devices, have not been fully studied. The topic was just recently investigated in Georges et al. (2020), Nunes, Macêdo, and Zanchettin (2020) and Safari et al. (2020).

6. Deep embedding: Temporal pooling layers

As shown in Fig. 7, the temporal pooling layer is a bridge between the frame-level and utterance-level hidden layers. Given a speech segment, we assume that the input and output of the temporal pooling layer are $\mathcal{H} = \{\mathbf{h}_t \in \mathbb{R}^{d_2} | t = 1, 2, \ldots, T\}$ and $\mathbf{u}$, respectively, where $\mathbf{h}_t$ denotes the tth frame-level speaker feature produced from the frame-level hidden layers. In this section, we introduce a number of temporal pooling functions.

6.1. Average pooling

Average pooling (Li et al., 2017, 2018; Yadav & Rai, 2018; Zhang & Koishida, 2017) is the most common pooling function:

$$\mathbf{u} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t \tag{8}$$

6.2. Statistics pooling

Statistics pooling (Snyder et al., 2017, 2018) calculates both the statistic mean $\mathbf{m}$ and standard deviation $\mathbf{d}$ of $\mathcal{H}$:

$$\mathbf{m} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t \tag{9}$$

$$\mathbf{d} = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t \odot \mathbf{h}_t - \mathbf{m}\odot\mathbf{m}} \tag{10}$$

where $\odot$ denotes the Hadamard product. The output of the statistics pooling layer is a concatenation of $\mathbf{m}$ and $\mathbf{d}$, i.e. $\mathbf{u} = [\mathbf{m}^T, \mathbf{d}^T]^T$.

6.3. Self-attention-based pooling

Obviously, (8), (9), and (10) assume that all elements of $\mathcal{H}$ contribute equally to $\mathbf{u}$. However, the assumption may not be true, since that the frames may not provide equal speaker-discriminative information. To address this issue, many works applied self attention mechanisms for weighted statistics pooling layers. Specifically, the attention can be broadly interpreted as a vector of importance weights,[3] which allows a neural network to focus on a specific portion of its input. Furthermore, the self attention computes attentive weights within a single sequence.

In the following two subsections, we first present a general self attention framework which produces weighted means and standard deviations of the input from a self-attentive scoring function in Section 6.3.1, and then list a number of specific self-attention-based pooling methods under the framework in Section 6.3.2.

6.3.1. A self attention pooling framework

Without loss of generality, self-attentive scoring is defined as:

$$\left\{f_{\text{Att}}^{(k)}(\cdot) | k = 1, 2, \ldots, K\right\} \tag{11}$$

where $f_{\text{Att}}^{(k)}(\cdot)$ is usually referred as one-head, and K is the total number of heads. If $K \geq 2$, the self attention mechanism is usually called multi-head self attention which allows the model to jointly attend to information from different representation subspaces (Vaswani et al., 2017); otherwise, it degenerates into a single-head one. Although $f_{\text{Att}}^{(k)}(\cdot)$ has many different implementations, many of the implementations share similar forms with the structured self-attentive function (Lin et al., 2017) which obtains the importance weights by:

$$f_{\text{Att}}^{(k)}(\mathbf{h}_t) = \mathbf{v}^{(k)^T}\tanh(\mathbf{W}^{(k)}\mathbf{h}_t + \mathbf{g}^{(k)}) + b^{(k)}, \quad k = 1, 2, \ldots, K \tag{12}$$

where $\mathbf{W}^{(k)} \in \mathbb{R}^{d_3\times d_2}$, $\mathbf{g}^{(k)} \in \mathbb{R}^{d_3}$, $\mathbf{v}^{(k)} \in \mathbb{R}^{d_3}$ and $b^{(k)} \in \mathbb{R}$ are learnable parameters of the kth scoring function. Suppose $s_t^{(k)} = f_{\text{Att}}^{(k)}(\mathbf{h}_t), \quad k = 1, 2, \ldots, K$, then the importance weights for the frame-level feature $\mathbf{h}_t$ are obtained by normalizing $s_t^{(k)}$ with a softmax function:

$$\alpha_t^{(k)} = \frac{\exp\left(s_t^{(k)}\right)}{\sum_{t'}^{T}\exp\left(s_{t'}^{(k)}\right)}, \quad k = 1, 2, \ldots, K \tag{13}$$

where the normalization guarantees that the weights satisfy $0 \leq \alpha_t^{(k)} \leq 1$ and $\sum_{t=1}^{T}\alpha_t^{(k)} = 1$. Finally, the weighted mean and standard deviation produced from the kth self-attentive scoring function can be derived as follows:

$$\widetilde{\mathbf{m}}^{(k)} = \sum_{t=1}^{T}\alpha_t^{(k)}\mathbf{h}_t, \quad k = 1, 2, \ldots, K \tag{14}$$

$$\widetilde{\mathbf{d}}^{(k)} = \sqrt{\sum_{t=1}^{T}\alpha_t^{(k)}\mathbf{h}_t \odot \mathbf{h}_t - \widetilde{\mathbf{m}}^{(k)}\odot\widetilde{\mathbf{m}}^{(k)}}, \quad k = 1, 2, \ldots, K \tag{15}$$

Finally, $\widetilde{\mathbf{m}}^{(k)}$ and $\widetilde{\mathbf{d}}^{(k)}$ are used to calculate an utterance-level representation as described in the following subsection.

6.3.2. Attention pooling methods

Under the above attention framework, this subsection categorizes existing self-attention based pooling layers into the following six classes, where all methods take (12) as the self-attentive scoring function and take (13) as the normalization function, unless otherwise stated.

[3] https://lilianweng.github.io/lil-log/2018/06/24/attention-attention.html

Table 8
Selected examples of the effect of neural network structures on performance.
Source: From Villalba et al. (2020).

Comparison methods		Test dataset [condition]	EER		
Main models	Baselines		Main	Baseline	Relative reduction
E-TDNN (10M)	TDNN (8.5M)	SITW EVAL CORE (*16 kHz systems*)	2.74%	3.40%	**19%**
F-TDNN (9M)	TDNN (8.5M)	SITW EVAL CORE (*16 kHz systems*)	2.39%	3.40%	**30%**
F-TDNN (17M)	TDNN (8.5M)	SITW EVAL CORE (*16 kHz systems*)	1.89%	3.40%	**44%**
ResNet (8M)	TDNN (8.5M)	SITW EVAL CORE (*16 kHz systems*)	3.01%	3.40%	**11%**

- **Single-head attentive average pooling** (Bhattacharya, Alam, Gupta, & Kenny, 2018; Bhattacharya et al., 2017; rahman Chowdhury et al., 2018): Bhattacharya et al. (2017) takes a fully-connected layer as $f_{\text{Att}}(\cdot)$. Bhattacharya et al. (2018) adopts the cosine function to compute attention scores:

$$s_t = f_{\text{Att}}(\mathbf{h}_t, \mathbf{r}) = \frac{\mathbf{h}_t^T \mathbf{r}}{\|\mathbf{h}_t\|_2 \|\mathbf{r}\|_2} \tag{16}$$

where $\mathbf{r}$ is a nonlinearly transformed i-vector from the same utterance as $\mathbf{h}$. Obviously, the attention weights in (16) are determined by both the frame-level $\mathbf{h}_t$ and the utterance-level information $\mathbf{r}$. In rahman Chowdhury et al. (2018), several attentive functions similar to (12) are investigated. The output of the single-head attentive average pooling is set to the weighted mean:

$$\mathbf{u} = \widetilde{\mathbf{m}}^{(1)}. \tag{17}$$

- **Single-head attentive statistics pooling** (Okabe et al., 2018): It uses a single-head attention function, i.e. $K = 1$. Its output is a concatenation of both the weighted mean and weighted standard deviation:

$$\mathbf{u} = [\widetilde{\mathbf{m}}^{(1)^T}, \widetilde{\mathbf{d}}^{(1)^T}]^T. \tag{18}$$

- **Single-head Baum–Welch statistics attention mechanism based statistics pooling** (Gu, Guo, Dai, & Du, 2020): To overcome the weakness of (12) which cannot fully mine the inner relationship between an utterance and its frames, Gu et al. (2020) integrated the Baum–Welch statistics into the attention mechanism:

$$s_t = \mathbf{v}^T \tanh(\mathbf{K}\mathbf{q}_t + \mathbf{g}) \tag{19}$$

where $\mathbf{K}$ is named the key matrix and $\mathbf{q}_t$ is a query vector calculated by:

$$\mathbf{q}_t = f(\mathbf{h}_t^{(-1)}) \tag{20}$$

where $f(\cdot)$ is a nonlinear function, and $\mathbf{h}_t^{(-1)}$ denotes the output of a penultimate frame-level hidden layer. The key matrix $\mathbf{K}$ is calculated from the Baum–Welch statistics. Specifically, Gu et al. (2020) first calculates the normalized first order statistics $\mathbf{f}_c$ from the cth component of a GMM-UBM model Ω (see (5)), and then conducts the following nonlinear transform:

$$\mathbf{f}'_c = \mathbf{V}_2 \tanh(\mathbf{V}_1 \mathbf{f}_c + \mathbf{g}), \quad \forall c = 1, \ldots, C \tag{21}$$

where $\mathbf{V}_1$, $\mathbf{V}_2$ and $\mathbf{g}$ are the parameters of DNN. Finally, it concatenates $\mathbf{F}' = [\mathbf{f}'_1, \mathbf{f}'_2, \ldots, \mathbf{f}'_C]$ and the trainable matrix $\mathbf{W}$ as the key matrix:

$$\mathbf{K} = [\mathbf{F}', \mathbf{W}]^T \tag{22}$$

After obtaining s_t, $\mathbf{u}$ is obtained in the same way as (18).

- **Global multi-head attentive average pooling**: It first applies a K-head ($K \geq 2$) attention function to $\mathcal{H}$ by (12). Then, the attentive weights and weighted means are calculated by (13) and (14) respectively (Wang et al., 2020c). Finally, the output of the pooling layer $\mathbf{u}$ is the concatenation of the weighted means:

$$\mathbf{u} = [\widetilde{\mathbf{m}}^{(1)^T}, \widetilde{\mathbf{m}}^{(2)^T}, \ldots, \widetilde{\mathbf{m}}^{(K)^T}]^T \tag{23}$$

It can be seen that $\mathbf{u} \in \mathbb{R}^{Kd_2}$. Similar ideas can also be found in Zhou, Zhao, Li, Gong, and Wu (2019) and Zhu et al. (2018). Zhu et al. (2018) also added an additional penalty term into the objective function to enlarge the diversity between the heads.

- **Sub-vectors based multi-head attentive average pooling** (India, Safari, & Hernando, 2019): It first splits $\mathbf{h}_t$ into K ($K \geq 2$) non-overlapping homogeneous sub-vectors $\mathbf{h}_t = [\mathbf{h}_t^{(1)^T}, \mathbf{h}_t^{(2)^T}, \ldots, \mathbf{h}_t^{(K)^T}]^T$, where $\mathbf{h}_t^{(k)} \in \mathbb{R}^{d_2/K}$. Then, it applies single-head attention to each of the sub-vectors $\mathcal{H}^{(k)} = \{\mathbf{h}_t^{(k)} \in \mathbb{R}^{d_2/K} | t = 1, 2, \ldots, T\}$. Finally, it obtains the sub-pooling outputs by:

$$\mathbf{u}^{(k)} = \sum_{t=1}^{T} \alpha_t^{(k)} \mathbf{h}_t^{(k)}, \quad k = 1, 2, \ldots, K \tag{24}$$

It can be seen that $\mathbf{u}^{(k)} \in \mathbb{R}^{d_2/K}$. The output of the pooling layer is a concatenation of the sub-pooling outputs:

$$\mathbf{u} = [\mathbf{u}^{(1)^T}, \mathbf{u}^{(2)^T}, \ldots, \mathbf{u}^{(K)^T}]^T. \tag{25}$$

- **Multi-resolution multi-head attentive average pooling** (Wang et al., 2020c): Because the speaker characteristics are obtained through the aggregation of the attentive weights reweighted frame-level features, Wang et al. (2020c) proposed to control the resolution of the attentive weights with a temperature parameter. They modify the softmax function as:

$$\alpha_t = \frac{\exp(s_t/E)}{\sum_{t'=1}^{T} \exp(s_{t'}/E)} \tag{26}$$

where E is the temperature parameter. It is obvious that increasing E makes the distribution of α_t less sharp, i.e. lower resolution. By incorporating the above intuition, the weighting equation (13) is changed to:

$$\alpha_t^{(k)} = \frac{\exp\left(s_t^{(k)}/E_k\right)}{\sum_{t'}^{T} \exp\left(s_{t'}^{(k)}/E_k\right)} \tag{27}$$

where $E_k \geq 1$ is a temperature hyperparameter of the kth head. Finally, the output $\mathbf{u}$ is calculated in a similar way with that of the global multi-head attentive average pooling except that $\alpha_t^{(k)}$ is replaced by (27).

It is clear that the above attentive pooling methods all employ scalar attention weights for each frame-level vector. Wu, Guo, Gao, Hou, and Xu (2020) further proposed a vector-based attentive pooling method, which adopts vectorial attention weights for each frame-level vector.

6.4. NetVLAD & GhostVLAD pooling

In Xie et al. (2019), the authors applied a dictionary-based NetVLAD layer to aggregate features across time, which can be

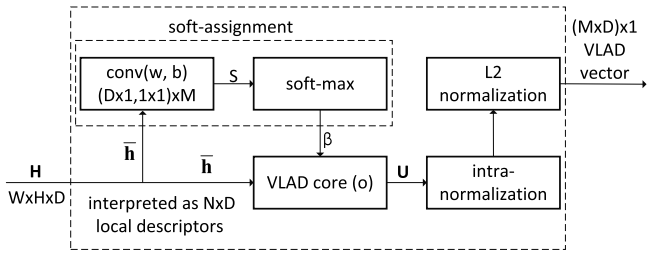

Fig. 9. Diagram of the NetVLAD pooling layer. *Source:* From Arandjelovic, Gronat, Torii, Pajdla, and Sivic (2016).

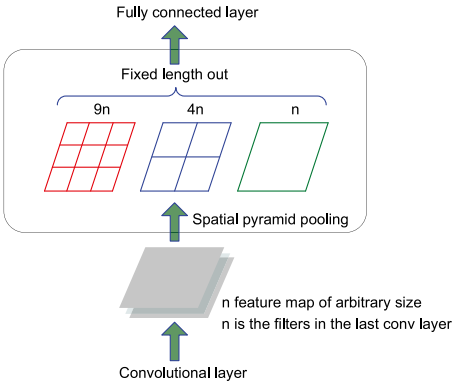

Fig. 10. Spatial pyramid pooling. *Source:* From Zhang, Koishida, and Hansen (2018).

intuitively regarded as trainable discriminative clustering: every frame-level descriptor will be softly assigned to different clusters, making the residuals encoded as the output feature (Xie et al., 2019).

Specifically, as shown in Fig. 9, suppose that the input of the NetVLAD layer is a three-dimensional tensor $\mathbf{H}^{(W\times H)\times D}$, where W, H and D depend on the speech length, the dimensions of the spectrum frequency bins, and the number of convolution kernels respectively. By only retaining the third dimension, $\mathbf{H}$ can be converted to N one-dimensional tensors, i.e. $\overline{\mathcal{H}} = \{\overline{\mathbf{h}}_n \in \mathbb{R}^D | n = 1, 2, \ldots, N\}$ where $N = W \times H$. As shown in Fig. 9, the NetVLAD pooling layer consists of the following four steps (Arandjelovic et al., 2016):

(1) Calculate a matrix $\mathbf{U} \in \mathbb{R}^{D\times M}$ from $\overline{\mathcal{H}}$ by:

$$\mathbf{U}(:, m) = \sum_{n=1}^{N} \beta_m(\overline{\mathbf{h}}_n)(\overline{\mathbf{h}}_n - \mathbf{o}_m) \tag{28}$$

where M is the number of the chosen clusters $\mathcal{O} = \{\mathbf{o}_m \in \mathbb{R}^D | m = 1, 2, \ldots, M\}$, and $\beta_m(\overline{\mathbf{h}}_n)$ is an assignment weight calculated by:

$$\beta_m(\overline{\mathbf{h}}_n) = \frac{\exp\left(\mathbf{w}_m^T \overline{\mathbf{h}}_n + b_m\right)}{\sum_{m'=1}^{M} \exp\left(\mathbf{w}_{m'}^T \overline{\mathbf{h}}_n + b_{m'}\right)} \tag{29}$$

with $\{\mathbf{w}_m\}$, $\{b_m\}$ and $\{\mathbf{o}_m\}$ as the parameters of the network.

(2) Normalize $\mathbf{U}$ by ℓ_2-norm column-wisely. This step is termed as the intra-normalization.

(3) Convert the normalized $\mathbf{U}$ into a vector:

$$\mathbf{u} = [\mathbf{U}(:, 1)^T, \mathbf{U}(:, 2)^T, \ldots, \mathbf{U}(:, M)^T]^T \tag{30}$$

(4) Normalize $\mathbf{u}$ by ℓ_2-norm to generate an $M \times D$ dimensional output vector. This step is termed as the ℓ_2-normalization.

In addition, Xie et al. (2019) also applied a variant of NetVLAD, named GhostVLAD. The main difference between them is that some of the clusters in the GhostVLAD layer, named *"ghost clusters"*, are not included in the final concatenation, and hence do not contribute to the final representation. When aggregating the frame-level features, the contribution of the noisy and undesirable sections of a speech segment to the normal VLAD clusters will be effectively down-weighted, since that larger weights are assigned to the "ghost cluster". See Zhong, Arandjelović, and Zisserman (2018) for the details.

6.5. Learnable dictionary encoding pooling

Motivated by GMM-UBM, Cai, Chen, and Li (2018) proposed a learnable dictionary encoding (LDE) pooling layer which models the distribution of the frame-level features $\mathcal{H}$ by a dictionary. The dictionary learns a set of dictionary component centers $\overline{\mathcal{O}} = \{\overline{\mathbf{o}}_m \in \mathbb{R}^{d_2} | m = 1, 2, \ldots, M\}$, and assigns weights to the frame-level features by:

$$\overline{\beta}_{tm} = \frac{\exp(-\tau_m \|\mathbf{h}_t - \overline{\mathbf{o}}_m\|^2)}{\sum_{m'=1}^{M} \exp(-\tau_{m'} \|\mathbf{h}_t - \overline{\mathbf{o}}_{m'}\|^2)} \tag{31}$$

where the smoothing factor τ_m for each dictionary center $\overline{\mathbf{o}}_m$ is learnable. The aggregated output of the pooling layer with respect to the center $\overline{\mathbf{o}}_m$ is:

$$\mathbf{u}_m = \frac{\sum_{t=1}^{T} \overline{\beta}_{tm}(\mathbf{h}_t - \overline{\mathbf{o}}_m)}{\sum_{t=1}^{T} \overline{\beta}_{tm}} \tag{32}$$

In order to facilitate the derivation, (32) is simplified to:

$$\mathbf{u}_m = \frac{\sum_{t=1}^{T} \overline{\beta}_{tm}(\mathbf{h}_t - \overline{\mathbf{o}}_m)}{T} \tag{33}$$

Finally, the output of the pooling layer is $\mathbf{u} = [\mathbf{u}_1^T, \mathbf{u}_2^T, \ldots, \mathbf{u}_M^T]^T$.

6.6. Spatial pyramid pooling

In order to handle variable-length utterances, Zhang, Koishida, and Hansen (2018) incorporated a Spatial Pyramid Pooling operation into a CNN-based network, which can directly produce fixed-length feature vectors from variable-length utterances.

As shown in Fig. 10, the spatial pyramid pooling layer outputs a fixed length vector by first dividing the input feature maps into 1×1, 2×2, and 3×3 small patches and then performing average pooling over these patches. An exceptional advantage of the spatial pyramid pooling layer is that it maintains spatial information of the last frame-level feature maps by making average pooling in each local small patches.

Jung, Kim, Lim, Choi, and Kim (2019) further extracted embeddings from the divided small patches via a parameter-sharing LDE layer instead of applying the averaging pooling on them.

6.7. Other temporal pooling functions

There are many other successful pooling methods. For example, Gao et al. (2018) proposed a cross-convolutional-layer pooling method to capture the first-order statistics for modeling long-term speaker characteristics. Travadi and Narayanan (2019) reported a total variability model based pooling layer. Heigold et al. (2016) connected the last output of LSTM to the loss function for an utterance-level speaker representation. Apart from the single-scale aggregation methods in Sections 6.1 to 6.6 which generate the pooling output from the last frame-level layer, multiscale aggregation methods have also been proposed (Gao et al.,

2019; Hajavi & Etemad, 2019; Jung, Kye, Choi, Jung, & Kim, 2020b; Seo et al., 2019; Tang et al., 2019) which utilize multiscale features from different frame-level layers to generate the pooling output.

6.8. Discussion to the temporal pooling layers

Because temporal pooling layers behave fundamentally different in different datasets, network structures, or loss functions, it is difficult to conclude which one is the best. To our knowledge, temporal pooling functions with learnable parameters achieved better (at least competitive) results than the simple pooling layers such as the average pooling and statistical pooling in most cases, with a weakness of higher computational complexity than the latter. Some examples are listed in Table 9.

7. Deep embedding: Classification-based objective functions

The objective function largely determines the performance of a neural network. Deep-embedding-based speaker recognition systems usually adopt classification-based objective functions. Before reviewing the objective functions, we first summarize the deep-embedding-based speaker recognition as the following multi-class classification problem.

Let $\mathcal{X} = \{(\mathbf{x}_n, l_n)|n = 1, 2, \ldots, N\}$ denote the training samples in a mini-batch,[4] where $\mathbf{x}_n \in \mathbb{R}^{d_4}$ represents the input of the last fully connected layer, $l_n \in \{1, 2, \ldots, J\}$ is the class label of $\mathbf{x}_n$ with J as the number of speakers in the training set, and N is the batch size. In addition, $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_J]$ and $\mathbf{b} = [b_1, b_2, \ldots, b_J]$ denote the weight matrix and bias vector of the last fully connected layer respectively.

In this section, we comprehensively summarize the classification based objective functions. Without loss of generality, the "loss function", "cost function" and "objective function" are equivalent in this article.

7.1. The variants of softmax loss

As shown in Section 4, both the d-vector and x-vector extractors take the minimum cross entropy as the objective function, and take softmax as the output layer. For short, we denote the objective function as the *Softmax loss*.[5] For a multiclass classification problem, the cross-entropy error function over $\mathcal{X}$ can be calculated as:

$$L = -\frac{1}{N}\sum_{n=1}^{N}\sum_{j=1}^{J} t_{nj}\log p_{nj} \tag{34}$$

where $[t_{n1}, t_{n2}, \ldots, t_{nJ}]$ is a one-hot vector encoded from the label l_n, in other words, t_{nj} equals 1 if and only if sample $\mathbf{x}_n$ belongs to class j. p_{nj} is the posterior probability of $\mathbf{x}_n$ belonging to class j. It is produced from neural networks with the following Softmax function:

$$p_{nj} = \frac{\exp(\mathbf{w}_j^T\mathbf{x}_n + b_j)}{\sum_{j=1}^{J}\exp(\mathbf{w}_j^T\mathbf{x}_n + b_j)} \tag{35}$$

Combining (34) and (35) derives an equivalent form of the Softmax loss:

$$\mathcal{L}_S = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp(\mathbf{w}_{l_n}^T\mathbf{x}_n + b_{l_n})}{\sum_{j=1}^{J}\exp(\mathbf{w}_j^T\mathbf{x}_n + b_j)} \tag{36}$$

[4] DNN is often trained using a mini-batch data in an iteration.

[5] Following Liu et al. (2017), we define the softmax loss as a combination of the last fully connected layer, softmax function, and cross-entropy loss function.

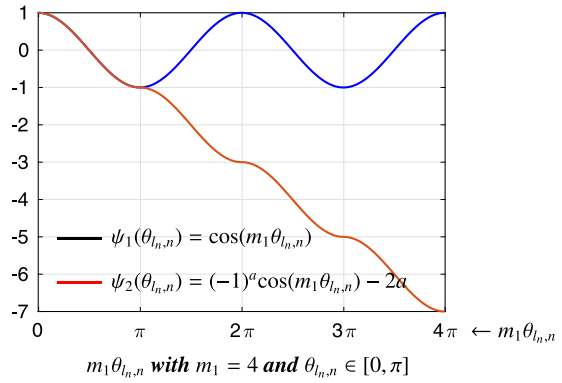

Fig. 11. Illustration of the angle function of the ASoftmax loss. Obviously, the angle $\theta_{l_n,n}$ in the training stage is in $[0, \pi]$. If we simply multiply an integer margin m_1 to $\theta_{l_n,n}$, then the angle function $\psi_1(\cdot)$ is monotonic when $\theta_{l_n,n} \in [0, \frac{\pi}{m_1}]$ only. Therefore, in practice, $\psi_1(\cdot)$ is generalized to $\psi_2(\cdot)$ to ensure that the angle function is monotonically decreasing when $\theta \in [0, \pi]$.

Softmax loss is the most common objective function for deep embedding. However, from (36), one can see that Softmax loss is only good at maximizing the between-class distance, but does not have an explicit constraint on minimizing the within-class variance. Therefore, the performance of deep embedding has much room of improvement. Here we present some representative variants of Softmax loss as follows.

- **Angular softmax (ASoftmax) loss** (Cai et al., 2018; Huang, Wang, & Yu, 2018; Novoselov, Shulipa, Kremnev, Kozlov, & Shchemelinin, 2018): Because the inner product between $\mathbf{w}_j$ and $\mathbf{x}_n$ in (36) can be rewritten as:

$$\mathbf{w}_j^T\mathbf{x}_n = \|\mathbf{w}_j\|\,\|\mathbf{x}_n\|\cos(\theta_{j,n}) \tag{37}$$

where $\theta_{j,n}(0 \le \theta_{j,n} \le \pi)$ denotes the angle between $\mathbf{w}_j$ and $\mathbf{x}_n$, Softmax loss can be rewritten as:

$$\mathcal{L}_S = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp\left(\|\mathbf{w}_{l_n}\|\,\|\mathbf{x}_n\|\cos(\theta_{l_n,n}) + b_{l_n}\right)}{\sum_{j=1}^{J}\exp\left(\|\mathbf{w}_j\|\,\|\mathbf{x}_n\|\cos(\theta_{j,n}) + b_j\right)} \tag{38}$$

If we further set the bias terms to zero, normalize the weights at the forward propagation stage, and add a margin to the angle:

$$b_j = 0, \quad \|\mathbf{w}_j\| = 1, \quad \psi(\theta_{l_n,n}) = (-1)^a\cos(m_1\theta_{l_n,n}) - 2a \tag{39}$$

then, we explicitly constrain the learned features to have a small intra-speaker variation, where $m_1 \ge 1$ is an integer margin hyperparameter, $\theta_{l_n,n} \in [\frac{a\pi}{m_1}, \frac{(a+1)\pi}{m_1}]$, and $a \in \{0, 1, \ldots, m_1 - 1\}$. The intuition behind the angle function $\psi(\theta_{l_n,n})$ is illustrated in Fig. 11. Then, we obtain ASoftmax loss as follows:

$$\mathcal{L}_{AS} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp\left(\|\mathbf{x}_n\|\psi(\theta_{l_n,n})\right)}{\exp\left(\|\mathbf{x}_n\|\psi(\theta_{l_n,n})\right) + \sum_{j=1,j\neq l_n}^{J}\exp\left(\|\mathbf{x}_n\|\cos(\theta_{j,n})\right)} \tag{40}$$

Note that, because m_1 is limited to a positive integer instead of a real number, the margin is not flexible enough.

- **Additive margin softmax (AMSoftmax) loss** (Hajibabaei & Dai, 2018; Xie et al., 2019; Yu et al., 2019): It is a revision of ASoftmax loss by replacing $\psi(\theta_{l_n,n})$ in (40) with $(\cos(\theta_{l_n,n}) -$

Table 9
Experimental results of different temporal pooling functions selected from literature. Each row represents a comparison. The results across rows are not comparable.

Comparison methods		Test dataset [condition]	EER		
Main models	Baselines		Main	Baseline	Relative reduction
Attention (Zhu et al., 2018)	Average	SRE16 Cantonese	5.81%	7.33%	**21%**
NetVLAD (Xie et al., 2019)	Average	VoxCeleb1 test set	3.57%	10.48%	**66%**
GhostVLAD (Xie et al., 2019)	Average	VoxCeleb1 test set	3.22%	10.48%	**69%**
LDE (Villalba et al., 2020)	Statistics	SITW	2.50%	3.01%	**17%**

m_2), and normalizing $\|\mathbf{x}_n\| = 1$:

$$\mathcal{L}_{\text{AMS}} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp\left(\tau(\cos(\theta_{l_n,n}) - m_2)\right)}{\exp\left(\tau(\cos(\theta_{l_n,n}) - m_2)\right) + \sum_{j=1,j\neq l_n}^{J}\exp\left(\tau(\cos(\theta_{j,n}))\right)} \tag{41}$$

where τ is a scaling factor for preventing gradients too small during the training process (Xiang et al., 2019). In addition, Zhou et al. (2020) also proposed a dynamic-additive margin softmax, where m_2 is replaced by a dynamic margin for each training sample.

- **Additive angular margin softmax (AAMSoftmax)** (Liu, He, & Liu, 2019; Xiang et al., 2019): It replaces $(\cos(\theta_{l_n,n}) - m_2)$ in (41) by $\cos(\theta_{l_n,n} + m_3)$:

$$\mathcal{L}_{\text{AAMS}} = -\frac{1}{N}\sum_{n=1}^{N}\log\frac{\exp\left(\tau(\cos(\theta_{l_n,n} + m_3))\right)}{\exp\left(\tau(\cos(\theta_{l_n,n} + m_3))\right) + \sum_{j=1,j\neq l_n}^{J}\exp\left(\tau(\cos(\theta_{j,n}))\right)} \tag{42}$$

To further improve the convergence speed and accuracy, Rybicka and Kowalczyk (2020) recently proposed a parameter adaptation method which adapts the scaling factor τ and margin m_3 at each iteration.

Compared to Softmax, both ASoftmax, AMSoftmax, and AAMSoftmax benefit from the following two aspects: first, the learned features are angularly distributed, which matches with the cosine similarity scoring back-end; second, they introduce an angle, i.e. a cosine margin, to quantitatively control the decision boundary between training speakers for minimizing the within-class variance. More information can be found in Deng, Guo, Xue, and Zafeiriou (2019), Liu et al. (2017), Wang, Cheng, Liu, and Liu (2018) and Wang et al. (2018).

7.2. Regularization for Softmax loss and its variants

As illustrated in Section 7.1, the learned feature by the Softmax loss is not discriminative enough. To address this issue, an alternative way is to combine the Softmax loss with some regularizers (Liu et al., 2017):

$$\mathcal{L} = \mathcal{L}_{\text{S}} + \lambda\mathcal{L}_{\text{Regular}} \tag{43}$$

where λ is a hyperparameter for balancing the Softmax loss $\mathcal{L}_{\text{S}}$ and the regularizer $\mathcal{L}_{\text{Regular}}$. Besides, the regularizer is also applicable to other Softmax variants.

Because the embedding layer that produces the embedding speaker features is not always the last hidden layer, e.g. the x-vector in Fig. 7, the regularizer was sometimes added to the embedding layer. For clarity, we define the output of the embedding layer as $\mathcal{E} = \{(\mathbf{e}_n, l_n)|n = 1, 2, \ldots, N\}$, where $\mathbf{e}_n \in \mathbb{R}^{d_5}$. Here we introduce some regularizers as follows:

- **Center loss** (Cai et al., 2018; Li et al., 2018; Wang, Huang, Qian, & Yu, 2019): It is a typical regularizer for Softmax loss. It explicitly minimizes the within-class variance by:

$$\mathcal{L}_{\text{C}} = \frac{1}{2}\sum_{n=1}^{N}\|\mathbf{e}_n - \mathbf{c}_{l_n}\|^2 \tag{44}$$

where $\mathbf{c}_{l_n} \in \mathbb{R}^{d_5}$ denotes the l_nth class center of the elements in $\mathcal{E}$. At each training iteration, the centers are updated as follows (Li et al., 2018):

$$\mathbf{c}_j^{t+1} = \mathbf{c}_j^t - \epsilon\Delta\mathbf{c}_j^t \tag{45}$$

$$\Delta\mathbf{c}_j = \frac{\sum_{n=1}^{N}\delta(l_n = j)\cdot(\mathbf{c}_j - \mathbf{e}_n)}{1 + \sum_{n=1}^{N}\delta(l_n = j)} \tag{46}$$

where $\epsilon \in [0, 1]$ controls the learning rate of the centers, the superscript "t" represents the number of iterations, and $\delta(\cdot)$ is an indicator function. If the condition of the indicator function, i.e. $l_n = j$, is satisfied, then $\delta = 1$; otherwise, $\delta = 0$. The center loss is usually combined with Softmax loss:

$$\mathcal{L} = \mathcal{L}_{\text{S}} + \lambda\mathcal{L}_{\text{C}} \tag{47}$$

See Wen, Zhang, Li, and Qiao (2016) for more information about the center loss.

- **Ring loss** (Liu et al., 2019): It restricts $\|\mathbf{e}_n\|$ to be close to a target value R for AMSoftmax loss:

$$\mathcal{L} = \mathcal{L}_{\text{AMS}} + \lambda \times \frac{1}{N}\sum_{n=1}^{N}(\|\mathbf{e}_n\| - R)^2 \tag{48}$$

where the target norm R is optimized during the network training. Eq. (48) essentially applies a normalization constraint to the features.

- **Minimum hyperspherical energy criterion** (Liu et al., 2019): It enforces the weights of the output layer to distribute evenly on a hypersphere:

$$\mathcal{L} = \mathcal{L}_{\text{AMS}} + \frac{\lambda}{N(J-1)}\sum_{n=1}^{N}\sum_{j=1,j\neq l_n}^{J}h(\|\hat{\mathbf{w}}_{l_n} - \hat{\mathbf{w}}_j\|) \tag{49}$$

where $\hat{\mathbf{w}}_{l_n}$ and $\hat{\mathbf{w}}_j$ are the ℓ_2-normalized $\mathbf{w}_{l_n}$ and $\mathbf{w}_j$ respectively, and $h(z) = \frac{1}{z^2}$ is a decreasing function. Intuitively, the minimum hyperspherical energy based regularizer enlarges the inter-class separability.

- **Gaussian prior** (Li, Tang, Shi, & Wang, 2019): To reduce information leak, Li, Tang, Shi, and Wang (2019) introduced a Gaussian prior to the output of the embedding layer, which results in the following objective function:

$$\mathcal{L} = \mathcal{L}_{\text{S}} + \lambda\sum_{j}\sum_{\mathbf{e}_n\in\varepsilon(j)}\|\mathbf{e}_n - \mathbf{w}_j\| \tag{50}$$

where $\varepsilon(j)$ is the set of the utterances belonging to the jth speaker, $\mathbf{e}_n$ represents the x-vector, $\mathbf{w}_j$ represents the parameters of the last layer corresponding to the output unit of speaker j.

- **Triplet loss** (Jati et al., 2019): Because Softmax loss does not explicitly reduce intra-class variance, triplet loss was introduced to directly bring samples from the same class

closer than the samples from different classes. Formally, the triplet loss weighted Softmax loss is written as:

$$\mathcal{L} = \lambda\mathcal{L}_S + (1-\lambda)\mathcal{L}_{\text{Triplet}} \tag{51}$$

where $\mathcal{L}_{\text{Triplet}}$ denotes the triplet loss, which will be introduced in Section 8.

There are also many other regularization approaches. For example, Yu et al. (2019) added a Hilbert–Schmidt independence criterion based constraint to the embedding layer for regularizing AMSoftmax loss $\mathcal{L}_{\text{AMS}}$. See Yu et al. (2019) for the details.

7.3. Multi-task learning for deep embedding

Phonetic information is important in improving the performance of speaker recognition. As illustrated in Section 3, one way to incorporate phonetic information into the i-vector-based systems is to employ an ASR acoustic model, e.g. the DNN-UBM/i-vector or the DNN-BNF/i-vector. As for the deep-embedding-based speaker recognition, the phonetic information was usually incorporated by multi-task learning. For example, Chen, Qian, and Yu (2015b) and Liu et al. (2015) trained a deep embedding network to discriminate the speaker identity and text phrases simultaneously. The training objective is to minimize:

$$\text{CE}([\boldsymbol{l}_1, \boldsymbol{l}_2], [\boldsymbol{l}'_1, \boldsymbol{l}'_2]) = \text{CE}_1(\boldsymbol{l}_1, \boldsymbol{l}'_1) + \text{CE}_2(\boldsymbol{l}_2, \boldsymbol{l}'_2) \tag{52}$$

where CE_1 and CE_2 are two cross-entropy criteria for speaker and text phrase respectively. $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ indicate the true labels for speakers and text individually, and $\boldsymbol{l}'_1$ and $\boldsymbol{l}'_2$ are the two outputs of the network respectively. Some similar ideas can also be found in Dey, Koshinaka, Motlicek, and Madikeri (2018).

Although the text content may be a harmful source to text-independent speaker recognition, some positive results were observed with the multi-task learning. The authors in Liu et al. (2018) added phonetic information to the frame-level layers of the x-vector extractor with an auxiliary ASR acoustic model by multi-task learning. The authors in Wang et al. (2019) conjectured that the phonetic information is helpful for frame-level feature learning, however, it is useless in utterance-level speaker embeddings. They experimentally verified their assumptions by multitask learning and adversarial training, where the phonetic information was used as positive and negative effects respectively.

Besides the phonetic information mining, some multi-task learning approaches intend to improve the performance of the auxiliary and main tasks together. For example, Tang, Li, Wang, and Vipperla (2016) proposed a collaborative learning approach based on multi-task recurrent neural model to improve the performance of both speech and speaker recognition. Yao and Mak (2018) proposed a multitask DNN structure to denoise i-vectors and classify speakers simultaneously. Considering that the acoustic and speaker domains are complementary, Jung, Jung, Goo, and Kim (2020a) recently proposed a multi-task network that performs keyword spotting and speaker verification simultaneously to fully utilize the interrelated domain information.

7.4. Discussion to the classification-based loss functions

Because speaker verification is an open set recognition task, the deep embedding space produced from a training dataset with a limited number of speakers is required to generalize well to unseen test speakers. Therefore, it is the speaker discriminative ability of the embedding rather than the classification accuracy that is important, which accounts for the motivation why many classification-based loss functions are designed to minimize the within-class variance by adding a margin or a regularizer into the Softmax loss.

From the experimental results in literature, one can concluded that the design of loss functions is very important to performance. At present, nearly all state-of-the-art deep embedding systems replaced the traditional Softmax by its variants, especially AMSoftmax and AAMSoftmax. In addition, the Softmax loss, its variants and regularizers are not mutually exclusive. For instance, the regularization terms (48) and (49) were originally added to AMSoftmax (Liu et al., 2019).

8. End-to-end speaker verification: Verification-based objective functions

An emerging direction of speaker recognition is end-to-end speaker verification. It is able to produce the similarity score of a pair of utterances in a test trial directly. The main difference between deep embedding and end-to-end speaker verification is the objective function. Therefore, in this section, we mainly review the verification-based loss functions, and skip the other components that are similar to deep embedding, e.g. the network structures or temporal pooling layers.

Here we emphasize that the borderline between the classification-based deep embedding and verification-based end-to-end speaker verification is unclear in literature. Some work also called the end-to-end speaker verification systems as deep embedding extractors. The main reason for this confusion is that, although the end-to-end speaker verification systems have different objective functions and training strategies from the deep embedding extractors, they need to extract utterance-level speaker embeddings from the hidden layers as the input of some independent back-ends, e.g. PLDA, in the test stage, so as to achieve the state-of-the-art performance. Despite the confusion usage of the terms in literature, here we clearly regard the speaker verification systems whose loss functions yield similarity scores from training trials as end-to-end speaker verification.

In this section, we focus on summarizing verification-based objective functions, each of which needs to address the following three core issues:

- How to design a **training loss** that pushes DNN towards our desired direction: As shown in Fig. 1, speaker verification can be viewed as a binary classification problem of whether a pair of utterances are from the same speaker. A natural solution to this problem is to train a binary classifier in an end-to-end fashion from a large number of manually constructed pairs of training utterances, i.e. training trials. The training loss of the binary classifier largely determines the effectiveness of the classifier.
- How to define a **similarity metric** between a pair of utterances: The similarity of a pair of utterances is calculated from the embeddings of the utterances at the output layer where a proper similarity metric for evaluating the similarity between the embeddings boosts the performance.
- How to **select and construct training trials** from an exponentially large number of training trials: Because the number of all possible training trials is at least the square of the number of training utterances, and also because many of the training trials are less informative, we need to select or even construct some informative training trials instead of using all training trials.

8.1. Pairwise loss

Pairwise loss is a kind of training loss of the end-to-end speaker verification where each training trial contributes to the

accumulation of the training objective value independently. Suppose there is a set of pairwise training trials as $\mathcal{X}_{\text{pair}} = \{(\mathbf{x}_n^e, \mathbf{x}_n^t; l_n) | n = 1, 2, \ldots, N\}$ where $\mathbf{x}_n^e$ and $\mathbf{x}_n^t$ denote a pair of speaker embedding features at the output layer, and $l_n \in \{0, 1\}$ is the ground-truth label. If $\mathbf{x}_n^e$ and $\mathbf{x}_n^t$ belong to the same speaker, then $l_n = 1$; otherwise, $l_n = 0$.

Binary cross-entropy loss is the most common pairwise loss (rahman Chowdhury et al., 2018; Heigold et al., 2016; Rohdin et al., 2018; Snyder et al., 2016; Zhang, Chen, Zhao, Li, & Gong, 2016; Zhang et al., 2019):

$$\mathcal{L}_{\text{BCE}} = -\sum_{n=1}^{N}\left[l_n \ln\left(p(\mathbf{x}_n^e, \mathbf{x}_n^t)\right) + \eta(1-l_n)\ln\left(1-p(\mathbf{x}_n^e, \mathbf{x}_n^t)\right)\right] \quad (53)$$

where η is a balance factor between positive ($l_n = 1$) and negative ($l_n = 0$) trials, and $p(\mathbf{x}_n^e, \mathbf{x}_n^t)$ denotes the acceptance probability, i.e. the probability of $\mathbf{x}_n^e$ and $\mathbf{x}_n^t$ belonging to the same speaker. The reason why there needs a balance factor is that the number of the negative trials is usually much larger than that of positive trials. The difference between the variants of the binary cross-entropy loss is on the calculation method of $p(\mathbf{x}_n^e, \mathbf{x}_n^t)$ which is summarized as follows:

- In Heigold et al. (2016) and rahman Chowdhury et al. (2018), the authors applied sigmoid function to cosine similarity:

$$\begin{aligned} p(\mathbf{x}_n^e, \mathbf{x}_n^t) &= \frac{1}{1+\exp\left(-wS(\mathbf{x}_n^e, \mathbf{x}_n^t) - b\right)} \\ S(\mathbf{x}_n^e, \mathbf{x}_n^t) &= \frac{\mathbf{x}_n^{e^T}\mathbf{x}_n^t}{\|\mathbf{x}_n^e\|\,\|\mathbf{x}_n^t\|} \end{aligned} \quad (54)$$

  where w and b are two learnable parameters, and $-b/w$ corresponds to the verification threshold. Some similar idea can also be found in Zhang et al. (2016).
- In Snyder et al. (2016), the authors further introduced a PLDA-like similarity metric:

$$\begin{aligned} p(\mathbf{x}_n^e, \mathbf{x}_n^t) &= \frac{1}{1+\exp\left(-S(\mathbf{x}_n^e, \mathbf{x}_n^t)\right)} \\ S(\mathbf{x}_n^e, \mathbf{x}_n^t) &= (\mathbf{x}_n^e)^T\mathbf{x}_n^t - (\mathbf{x}_n^e)^T\mathbf{S}\mathbf{x}_n^e - (\mathbf{x}_n^t)^T\mathbf{S}\mathbf{x}_n^t + b \end{aligned} \quad (55)$$

  where $\mathbf{S}$ and b are learnable parameters. Another similar PLDA-based similarity metric was proposed in Rohdin et al. (2018).
- The third calculation method is to learn a score from a joint vector $\mathbf{x}_n^{e,t}$ by Zhang et al. (2019):

$$\begin{aligned} p(\mathbf{x}_n^e, \mathbf{x}_n^t) &= \frac{1}{1+\exp\left(-s_n^{e,t}\right)} \\ s_n^{e,t} &= S(\mathbf{x}_n^{e,t}) \end{aligned} \quad (56)$$

  where $s_n^{e,t}$ is a scalar produced from a fully-connected feed-forward neural network $S(\cdot)$, and $\mathbf{x}_n^{e,t}$ is a joint vector of $\mathbf{x}_n^e$ and $\mathbf{x}_n^t$ produced by a sequence-to-sequence attention mechanism (Zhang et al., 2019). Similar ideas can also be found in Heo, Jung, Yang, Yoon, and Yu (2017).

The training trials of the aforementioned end-to-end speaker verification are constructed from two utterances. To reduce the variability of the training trials, some work (rahman Chowdhury et al., 2018; Heigold et al., 2016; Zhang et al., 2016) obtains the embedding of the enrollment speech $\mathbf{x}_n^e$ from an average of a small amount of utterances.

Contrastive loss (Chung et al., 2018; Yu et al., 2019) is another commonly used pairwise loss:

$$\mathcal{L}_{\text{C}} = \frac{1}{2N}\sum_{n=1}^{N}\left(l_n d_n^2 + (1-l_n)\max(\rho - d_n, 0)^2\right) \quad (57)$$

where d_n denotes the Euclidean distance between $\mathbf{x}_n^e$ and $\mathbf{x}_n^t$, and ρ is a manually-defined margin. Unfortunately, training an end-to-end network with the contrastive loss is notoriously difficult. In order to avoid bad local minima in the early training stage, Chung et al. (2018) proposed to first pre-train a speaker embedding system using Softmax loss, and then fine-tune the system with the contrastive loss. Wan et al. (2018) proposed a generalization of the contrastive loss as follows:

$$L_{\text{GC}} = \sum_{n=1}^{N}\left(l_n\left(1-p(\mathbf{x}_n^e, \mathbf{x}_n^t)\right) + (1-l_n)\max_{\mathbf{x}_n^e \in \{\mathbf{c}_{j'}\}_{j'=1}^{J'}} p(\mathbf{x}_n^e, \mathbf{x}_n^t)\right) \quad (58)$$

where $p(\mathbf{x}_n^e, \mathbf{x}_n^t)$ is the same as (54), and $\mathbf{c}_{j'}$ with $j' = 1, 2, \ldots, J'$ is the speaker centroid of the jth speaker in a mini-batch which is obtained by averaging the utterances that belong to the j'th speaker.

Besides the above two common training losses, some other loss functions are as follows. In Gao et al. (2019), the authors proposed a *discriminant analysis loss* $\mathcal{L}_{\text{DALoss}}$ to learn discriminative embeddings:

$$\mathcal{L}_{\text{DALoss}} = \eta_1\mathcal{L}_{\text{intra}} + \eta_2\mathcal{L}_{\text{inter}} \quad (59)$$

where η_1 and η_2 are the weights of the two loss items, and $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{inter}}$ are described as follows. $\mathcal{L}_{\text{intra}}$ represents the intra-speaker variabilities which is defined as:

$$\mathcal{L}_{\text{intra}} = \sum_{j'=1}^{J'}\frac{C^{j'}}{\sum_{k=1}^{C^{j'}}\frac{1}{d_k(\mathbf{x}_{n_1}^{j'}, \mathbf{x}_{n_2}^{j'})}} \quad (60)$$

where $j' = 1, 2, \ldots, J'$ denotes the index of the training speaker in each mini-batch, and $d_k(\mathbf{x}_{n_1}^{j'}, \mathbf{x}_{n_2}^{j'})$ denotes the kth largest squared Euclidean distance between the embeddings of the j'th speaker. The overall cost is the mean of the first $C^{j'}$th largest distances within each speaker. $\mathcal{L}_{\text{inter}}$ represents the inter-speaker variabilities:

$$\mathcal{L}_{\text{inter}} = \max(0, \zeta - \min(d(\widetilde{\mathbf{x}}^{j'_1}, \widetilde{\mathbf{x}}^{j'_2}))) \quad (61)$$

where $\widetilde{\mathbf{x}}^{j'_1}$ and $\widetilde{\mathbf{x}}^{j'_2}$ denote the centers of the feature vectors of the j'_1th and j'_2th speakers respectively with $j'_1 \neq j'_2 \in \{1, 2, \ldots, J'\}$, $d(\cdot)$ denotes the distance (e.g. the squared Euclidean distance), and ζ denotes a margin. Thus, minimizing $\mathcal{L}_{\text{inter}}$ is equivalent to maximizing the distances between the centers to be larger than the minimum margin ζ.

In Mingote, Miguel, Ribas, Giménez, and Lleida (2019), the authors proposed to minimize both the empirical false alarm rate P_{fa} and miss detection rate P_{miss}:

$$\mathcal{L} = \eta_1 \cdot P_{\text{fa}}(\xi) + \eta_2 \cdot P_{\text{miss}}(\xi) \quad (62)$$

$$P_{\text{miss}}(\xi) = \frac{\sum_{n=1}^{N} l_n\,\delta\left(S(\mathbf{x}_n^e, \mathbf{x}_n^t) < \xi\right)}{\sum_{n=1}^{N}\delta(l_n = 1)} \quad (63)$$

$$P_{\text{fa}}(\xi) = \frac{\sum_{n=1}^{N}(1-l_n)\,\delta\left(S(\mathbf{x}_n^e, \mathbf{x}_n^t) > \xi\right)}{\sum_{n=1}^{N}\delta(l_n = 0)} \quad (64)$$

where $\delta(\cdot)$ denotes an indicator function, ξ denotes a decision threshold which is optimized with the neural network, and η_1 and η_2 are two tunable hyperparameters. The score $S(\mathbf{x}_n^e, \mathbf{x}_n^t)$ is obtained from the output linear layer of the neural network, where the number of units of the output layer equals the number of the speakers in the training data. Specifically, it uses a batch of

input vectors $\{\mathbf{x}_i\}_{i=1}^{I}$ and the parameters $\{\mathbf{w}_j, b_j\}_{j=1}^{J}$ of the output linear layer to construct training trials in a mini-batch:

$$S(\mathbf{x}_n^e, \mathbf{x}_n^t) = (\mathbf{x}_n^e)^T \mathbf{x}_n^t + b_n^e, \tag{65}$$

where $(\mathbf{x}_n^e, b_n^e) \in \{\mathbf{w}_j, b_j\}_{j=1}^{J}$ and $\mathbf{x}_n^t \in \{\mathbf{x}_i\}_{i=1}^{I}$. To make (63) and (64) differentiable, the indicator function $\delta(z > 0)$ is relaxed to a sigmoid function $\sigma(z) = 1/(1 + \exp(-z))$.

8.2. Triplet loss

Triplet loss is a kind of training loss that each training sample that contributes to the accumulation of the training objective value independently is constructed from three utterances. A triplet training sample consists of three utterances, including an anchor utterance, a positive utterance that is produced from the same speaker as the anchor utterance, and a negative utterance from a different speaker. Suppose the speaker features of a training sample produced from the top hidden layer are $\mathbf{x}^a$ (anchor), $\mathbf{x}^p$ (positive), and $\mathbf{x}^n$ (negative), respectively. We denote the training set as $\mathcal{X}_{\text{trip}} = \{(\mathbf{x}_n^a, \mathbf{x}_n^p, \mathbf{x}_n^n) | n = 1, 2, \ldots, N\}$.

Triplet loss designs a margin-based loss to push the positive utterance $\mathbf{x}_n^p$ closer to the anchor $\mathbf{x}_n^a$ than the negative utterance $\mathbf{x}_n^n$ in a trial as shown in Fig. 12. For any training sample in $\mathcal{X}_{\text{trip}}$, we require:

$$s_n^{an} - s_n^{ap} + \zeta \leq 0 \tag{66}$$

where, without loss of generality, s_n^{an} denotes the cosine similarity between $\mathbf{x}^a$ and $\mathbf{x}^n$, s_n^{ap} denotes the cosine similarity between $\mathbf{x}^a$ and $\mathbf{x}^p$, and $\zeta \in \mathbb{R}^+$ is a manually-defined *safety* margin between positive and negative pairs. Note that s_n^{an} and s_n^{ap} could be the scores of any similarity measurement instead of merely the cosine similarity. Given (66), the triplet loss is defined as:

$$\mathcal{L}_{\text{trip}} = \sum_{n=1}^{N} \max(0, s_n^{an} - s_n^{ap} + \zeta) \tag{67}$$

Cosine similarity (Li et al., 2017) and squared Euclidean distance (Bredin, 2017; Huang, Wang, & Qian, 2018; Zhang & Koishida, 2017) are the most common similarity metric for the triplet loss. Before calculating the similarities, each speaker embedding in the training samples needs to be length-normalized. It is easy to prove that the two similarity metrics are equivalent (Bai, Zhang, & Chen, 2020b; Zhang, Koishida, & Hansen, 2018) after the length normalization. Besides the two similarity metrics, the authors in Dey, Madikeri, and Motlicek (2018) proposed several distance functions to explore phonetic information for text-dependent speaker verification. They first compute the Euclidean distance between any pair of frame-level hidden representations of the two input utterances $\mathcal{H}_1 = \{\mathbf{h}_{1,t_1} | t_1 = 1, 2, \ldots, T_1\}$ and $\mathcal{H}_2 = \{\mathbf{h}_{2,t_2} | t_2 = 1, 2, \ldots, T_2\}$ via $\mathbf{D} = \{d(\mathbf{h}_{1,t_1}, \mathbf{h}_{2,t_2}) | t_1 = 1, 2, \ldots, T_1, t_2 = 1, 2, \ldots, T_2\} \in \mathbb{R}^{T_1, T_2}$. Then, they integrate the $T_1 \times T_2$ frame-level Euclidean distances into an utterance-level similarity score of $\mathcal{H}_1$ and $\mathcal{H}_2$ by, e.g. the attention mechanism.

Given a training set, we can see that the number of all possible triplet training samples is cubically larger than the number of training utterances. It is neither efficient nor effective to enumerate all possible triplets (Bredin, 2017), and only those that violate the constraint of $s_n^{an} - s_n^{ap} + \zeta \leq 0$ contributes to the training process. Therefore, how to select informative triplet training samples is fundamental to the effectiveness of the model training. In practice, the *"hard negative"* sampling strategy is popular (Bredin, 2017). It consists of the following two steps at each epoch:

(1) Randomly sample m utterances from each of the M speakers of the training set, which constructs $Mm(m-1)/2$ anchor-positive pairs.

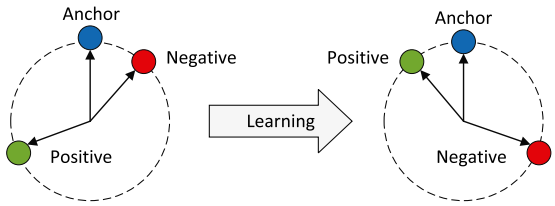

Fig. 12. Triplet loss based on cosine similarity.
Source: From Li et al. (2017).

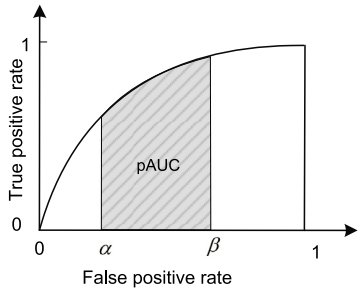

Fig. 13. Illustration of the ROC curve, AUC, and pAUC.
Source: From Bai et al. (2020a).

(2) For each of the anchor-positive pairs, randomly choose one negative utterance that satisfies $s_n^{an} - s_n^{ap} + \zeta > 0$ from the $(M-1)m$ negative candidates.

Several variants of the *"hard negative"* sampling were proposed as well. For example, Zhang and Koishida (2017) changed the first step by randomly selecting a small number of speakers from the speaker pool instead of from all speakers. In Huang, Wang, and Qian (2018), the authors divided training speakers into different groups and constructed each triplet training sample from a single group. Besides the *"hard negative"* sampling, the "semi-hard" negative sample selection (Jati et al., 2019; Schroff, Kalenichenko, & Philbin, 2015) and softmax pre-training (Li et al., 2017) are all used to stabilize the training process of the triplet loss.

8.3. Quadruplet loss

Quadruplet loss is a kind of training loss for end-to-end speaker verification where each training sample that contributes to the accumulation of the training objective value independently is constructed from four utterances. Suppose there are a positive pairwise training set $\mathcal{X}_{\text{same}} = \{(\mathbf{x}_{n_1}^e, \mathbf{x}_{n_1}^t) | n_1 = 1, 2, \ldots, N_1\}$ and a negative pairwise training set $\mathcal{X}_{\text{diff}} = \{(\mathbf{x}_{n_2}^e, \mathbf{x}_{n_2}^t) | n_2 = 1, 2, \ldots, N_2\}$ respectively, where $\mathbf{x}_{n_1}^e$ and $\mathbf{x}_{n_1}^t$ are from the same speaker while $\mathbf{x}_{n_2}^e$ and $\mathbf{x}_{n_2}^t$ are from different speakers. We have $\mathcal{X}_{\text{same}} \cup \mathcal{X}_{\text{diff}} = \mathcal{X}_{\text{pair}}$. Currently, the quadruplet loss is formulated as the maximization of the partial interested area under the ROC curve (pAUC) (Bai et al., 2020a).

The maximization of pAUC is to maximize an interested gray area of Fig. 13 that is defined by two hyperparameters α and β. It has the following three steps:

(1) Rank the similarity scores of all pairwise trials in $\mathcal{X}_{\text{diff}}$ in descending order, and selecting those elements that rank between the $(\lceil N_2 \times \alpha \rceil + 1)$th to $\lfloor N_2 \times \beta \rfloor$th positions to construct $\mathcal{X}'_{\text{diff}} = \{(\mathbf{x}_{n_3}^e, \mathbf{x}_{n_3}^t) | n_3 = 1, 2, \ldots, N_3\}$ where $N_3 = \lfloor N_2 \times \beta \rfloor - (\lceil N_2 \times \alpha \rceil + 1)$.
(2) Calculate pAUC on $\mathcal{X}_{\text{same}}$ and $\mathcal{X}'_{\text{diff}}$:

$$\text{pAUC} = 1 - \frac{1}{N_1 N_3} \sum_{n_1=1}^{N_1} \sum_{n_3=1}^{N_3} \left(\delta(s_{n_1} < s_{n_3}) + \frac{1}{2} \delta(s_{n_1} = s_{n_3}) \right) \tag{68}$$

where $\delta(\cdot)$ denotes the indicator function, and s_{n_1} and s_{n_3} denote the cosine similarity of the pairwise trials in $\mathcal{X}_{\text{same}}$ and $\mathcal{X}'_{\text{diff}}$ respectively.

(3) Relax the indicator function by the hinge loss which reformulates (68) to:

$$\mathcal{L}_{\text{pAUC}} = \frac{1}{N_1 N_3} \sum_{n_1=1}^{N_1} \sum_{n_3=1}^{N_3} \max\Big(0, \zeta - \big(s_{n_1} - s_{n_3}\big)\Big)^2 \tag{69}$$

where ζ is a margin hyperparameter.

It can be seen clearly that (69) is a quadruplet loss, since that $(s_{n_1} - s_{n_3})$ is calculated from four utterances.

Bai et al. (2020a) proposed two training sample construction methods. The first one is named random sampling. For a mini-batch, it first randomly chooses a mini-batch number of speakers, then randomly selects two utterances for each of the selected speaker, and finally generates the training trials of the batch by pairing all the selected utterances. The second one is named class-center learning. Before training, it first assigns a class center to each speaker in the training set. Then, for each training iteration, it generates the training trials of a mini-batch by pairing each of the class centers with each of the utterances in the batch, where the class centers are updated together with the DNN parameters.

The pAUC based loss has several advantages: (i) it directly optimizes the detection error tradeoff (DET) curve which is the major evaluation metric of speaker verification (Bai et al., 2020b); (ii) it naturally overcomes the class-imbalanced problem; (iii) it is able to select difficult quadruplet training samples by setting $\alpha = 0$ and β to a small value, e.g. 0.01; (iv) triplet training samples is a subset of quadruplet training samples when given the same training utterances (Bai et al., 2020b). Actually the pAUC should be calculated on the entire dataset, however, due to the limited computation resource, (68) is an empirical approximation to it within a mini-batch. Therefore, a large batch size is usually set to reduce the approximation error as much as possible. Fortunately, experimental results demonstrate that a good approximation can be obtained with a batch size of no larger than 512.

8.4. Prototypical network loss

In Wang, Wang, Law, Rudzicz, and Brudno (2019), the *prototypical network loss* (Snell, Swersky, & Zemel, 2017), which was originally proposed for few-shot learning, was applied to speaker embedding models. Suppose that a mini-batch contains a support set of N labeled samples $\mathcal{S} = \{(\mathbf{x}_n, l_n)|n = 1, 2, \ldots, N\}$ where $l_n \in \{1, 2, \ldots, J\}$ is the label of the sample $\mathbf{x}_n$, and $\mathcal{S}_j$ denotes the set of all samples of class j. Then, the *prototype* of each class is the mean vector of the support points belonging to the class:

$$\mathbf{c}_j = \frac{1}{|\mathcal{S}_j|} \sum_{(\mathbf{x}_n, l_n) \in \mathcal{S}_j} \mathbf{x}_n, \quad j = 1, 2, \ldots, J \tag{70}$$

Given a query set $\mathcal{Q} = \{(\mathbf{x}_q, l_q)|q = 1, 2, \ldots, Q\}$ with $l_q \in \{1, 2, \ldots, J\}$, the prototypical network loss classifies each query point $\mathbf{x}_q$ against J prototypes $\{\mathbf{c}_j|j = 1, 2, \ldots, J\}$ via a softmax function:

$$\mathcal{L}_{\text{PNL}} = -\sum_{(\mathbf{x}_q, l_q) \in Q} \log \frac{\exp\Big(-d\big(\mathbf{x}_q, \mathbf{c}_{l_q}\big)\Big)}{\sum_{j'=1}^{J} \exp\Big(-d\big(\mathbf{x}_q, \mathbf{c}_{j'}\big)\Big)} \tag{71}$$

where $d(\cdot)$ denotes the squared Euclidean distance.

For each mini-batch, Wang, Wang, Law, Rudzicz, and Brudno (2019) first randomly select a number of speakers from the training speaker pool, and then randomly choose a support set and a query set for each of the selected speakers, where the samples of the support set and query set do not overlap. Similar works were also conducted in Anand, Singh, Srivastava, and Lall (2019), Chung et al. (2020) and Kye, Jung, Lee, Hwang, and Kim (2020).

Before Wang, Wang, Law, Rudzicz, and Brudno (2019), Wan et al. (2018) proposed a generalized end-to-end loss based on the softmax function, which shares a similar idea with the prototypical network loss except that it uses a single set as both the support and query sets. In addition, Wei, Du, and Liu (2020) recently proposed an AM-Centroid loss which replaced the weights of the AAMSoftmax loss function with speaker centroids proposed in Wan et al. (2018). This loss function aims to overcome the weakness of the AAMSoftmax loss based deep networks whose number of parameters at the output layer grows linearly with the number of training speakers.

8.5. Other end-to-end loss functions

Some loss functions cannot be categorized to the above categories. For example, given learnable speaker bases $\{\mathbf{w}_j\}_j^J$ and a mini-batch of utterances $\{(\mathbf{x}_n, l_n)\}_n^N$ where $l_n \in \{1, 2, \ldots, J\}$, Heo et al. (2019) proposed a between-class variation based loss $\mathcal{L}_{\text{BC}}$,

$$\mathcal{L}_{\text{BC}} = \sum_{j_2=1}^{J} \sum_{j_1=1, j_1 \neq j_2}^{J} \frac{\mathbf{w}_{j_1}^T \mathbf{w}_{j_2}}{\|\mathbf{w}_{j_1}\| \, \|\mathbf{w}_{j_2}\|} \tag{72}$$

and a hard negative mining loss $\mathcal{L}_{\text{H}}$,

$$\mathcal{L}_{\text{H}} = \sum_{n=1}^{N} \sum_{\mathbf{w}_h \in \text{Hard}_n} \log\Big(1 + \exp\big(S(\mathbf{w}_h, \mathbf{x}_n) - S(\mathbf{w}_{l_n}, \mathbf{x}_n)\big)\Big) \tag{73}$$

where $\mathbf{x}_n$ denotes the nth utterance, $\mathbf{w}_{l_n}$ denotes the basis that $\mathbf{x}_n$ belongs to, $S(\cdot)$ is the cosine similarity, and Hard_n is a set of so-called hard negative speaker bases of $\mathbf{x}_n$ which correspond to the top H largest values in $\{S(\mathbf{w}_j, \mathbf{x}_n)|j \neq l_n, j = 1, 2, \ldots, J\}$.

8.6. Discussion to the verification-based loss functions

The verification-based loss functions are fundamentally different from the classification-based loss functions in at least the following aspects. First, speaker verification is essentially an open-set metric learning problem instead of a closed set classification problem. The verification-based losses are consistent with the test pipeline, which directly outputs verification scores.

Second, the output layers of the verification-based losses are very small and irrelevant to the number of training speakers, which is an important advantage of the verification-based methods over classification-based methods. Specifically, the number of parameters of a classification-based network at the output layer grows linearly with the increase of the number of training speakers, which make the network large-scale and easily overfit to the training data. For example, if a training set consists of 50000 speakers and if the top hidden layer of a classification network has 512 hidden units, then the output layer of the network contains 25.6 million parameters. On the contrary, the verification-based systems do not suffer the aforementioned weakness. Aware of this issue, Wei et al. (2020) tried to solve the parameter explosion problem by drawing lessons from the prototypical network loss.

The main weakness of the verification-based systems is that they are harder to train than the classification-based systems, since that they need to construct a large number of training trials and then select those that contributes significantly to the effectiveness of the systems, while the classification-based systems just classify each training utterance to its corresponding speaker. To overcome this weakness, many sample selection strategies for selecting highly-informative trials have been developed, such as

the hard negative sampling in the triplet loss and the pAUC optimization in the quadruplet loss. Because the highly-informative trials are dynamically changing during the training process, the optimization process is not very stable and consistent. Some unstable examples include the triplet loss or pAUC maximization with the random sampling strategy. Fortunately, this weakness can be alleviated by constructing trials with speaker centroids via, e.g. the class-center learning (Bai et al., 2020a).

In respect of the performance, although the classification-based systems outperformed the verification-based systems once, recently results shown that the latter can achieve competitive performance with the former (Bai et al., 2020a; Kye et al., 2020).

A remark: the term "end-to-end" in this section intends to make a difference from the embedding systems in Section 7. However, the term in many other speech processing tasks, which takes raw wave signals as the input and directly output decisions, is broader than the concept here. From the broader concept of "end-to-end", an end-to-end speaker verification system needs to further integrate additional procedures, including the voice activity detection, cepstral mean and variance normalization, into the network. It also has to prevent using additional back-ends, such as PLDA (Lin & Mak, 2020).

9. Speaker diarization

In this section, we overview four kinds of speaker diarization technologies—stage-wise diarization, end-to-end diarization, online diarization, and multimodal diarization, where the stage-wise diarization has been studied for a long time, while the last three are emerging directions.

9.1. Stage-wise speaker diarization

Stage-wise speaker diarization is composed of multiple independent modules. As shown in Fig. 14, most stage-wise speaker diarization systems consist of four modules—voice activity detection, speech segmentation, speaker feature extraction, and speaker clustering. Some systems also have an optional re-segmentation module. This subsection briefly reviews deep learning based methods for each module. Voice activity detection detects speech in an audio recording and removes non-speech regions. Although it is an important module, it is usually studied independently. Therefore, we focus on reviewing the other modules.

9.1.1. Speech segmentation

Speech segmentation splits speech into multiple speaker-homogeneous segments where each segment belongs to a single speaker. It usually can be categorized to two classes—*uniform segmentation* and *speaker change detection* (SCD). Uniform segmentation divides a long audio stream into short segments evenly by a sliding window (Garcia-Romero, Snyder, Sell, Povey, & McCree, 2017; Lin et al., 2020; Pal et al., 2020; Sell et al., 2018; Wang, Downey, Wan, Mansfield, & Moreno, 2018). For example, a 1.5 s sliding window with 0.75 s overlap is a common setting of uniform segmentation.

SCD partitions an audio recording according to the detected speaker change points, which results in non-uniform segments. Generally, it first partitions an audio recording into small segments, then computes the similarity between the two adjacent speech segments in terms of the distance between their representations, and finally decides whether the two adjacent segments are produced from the same speaker by thresholding the distance or finding a local extremum in the consecutive distance stream (Bredin, 2017). A common method for segmenting the audio recording into short segments is to use a sliding window (Bredin, 2017; Wang, Gu, Li, Xu, & Zheng, 2017). Recently, an ASR based segmentation (Aronowitz & Zhu, 2020; Sarı, Thomas, Hasegawa-Johnson, & Picheny, 2019) is also employed.

Conventional SCD algorithms usually adopt common handcrafted features, e.g. MFCC, as the acoustic representation (Chen, Gopalakrishnan, et al., 1998; Siegler, Jain, Raj, & Stern, 1997). An important advantage of the conventional methods is that only the step of tuning the threshold needs some experience, while the other parts do not need training (Yin, Bredin, & Barras, 2017). Recently, the deep speaker embedding features are used as the representation of speech segments instead of conventional handcrafted features (Aronowitz & Zhu, 2020; Bredin, 2017; Jati & Georgiou, 2018; Sarı et al., 2019; Wang et al., 2017).

To calculate the similarity between two adjacent segments, Euclidean distance (Bredin, 2017) and cosine similarity (Wang et al., 2017) are two common similarity measurements. Some methods also feed two (Jati & Georgiou, 2018; Sarı et al., 2019) or more (Aronowitz & Zhu, 2020) consecutive embeddings together into a pre-trained DNN to predict the similarity between two adjacent segments. Recently, some work formulates SCD as a sequence labeling task, which directly predicts if there is a change point in a speech segment (Hrúz & Zajíc, 2017; Yin et al., 2017; Zajíc, Hrúz, & Müller, 2017).

To summarize, on one side, the uniform segmentation is simple and works fine in many cases, which is the choice of many real-world diarization systems (Aronowitz & Zhu, 2020; Lin et al., 2020; Sell & Garcia-Romero, 2014); on the other side, the research on SCD is important not only for speaker diarization but also for many other applications, such as the closed captioning of broadcast television for hearing-impaired people (Aronowitz & Zhu, 2020).

9.1.2. Speaker feature extraction

The speaker feature extraction module in diarization shares similar technologies with speaker verification. Both of them map speech segments into speaker embeddings by i-vector (Sell & Garcia-Romero, 2014), DNN-UBM/i-vector (Sell, Garcia-Romero, & McCree, 2015), x-vector (Diez, Burget, Landini, Wang, & Černocký, 2020; Diez, Burget, Wang, Rohdin, & Cernockỳ, 2019; Landini et al., 2020; Sell et al., 2018), or some other deep embedding extractors (Sun, Zhang, & Woodland, 2019; Wang, Downey, Wan, Mansfield, & Moreno, 2018; Yella & Stolcke, 2015). See Sections 3 to 8 for the details. Here we only review some embedding methods that utilize additional information for speaker diarization. The authors in Higuchi, Suzuki, and Kurata (2020) incorporated acoustic conditions, such as the distances between speakers and microphones in a meeting or the channel conditions of different speakers in a telephone conversation, into the speaker embeddings, given the fact that the acoustic conditions provide discriminative information for diarization. Wang et al. (2020) utilized a graph neural network to refine speaker embeddings, where the local structural information between speech segments is utilized as additional information.

9.1.3. Speaker clustering

Given the segment-level embedding features of an audio recording, speaker clustering aims to partition the speech segments into several groups, each of which belongs to a single speaker. It first defines a similarity measurement for evaluating the similarity of two segments, and then conducts clustering according to the similarity scores. Popular similarity measurements include cosine similarity (Wang, Downey, Wan, Mansfield, & Moreno, 2018) and PLDA-bases similarity (Sell & Garcia-Romero, 2014; Sell et al., 2018). Recently, some deep-learning-based similarity measurements were also introduced, such as

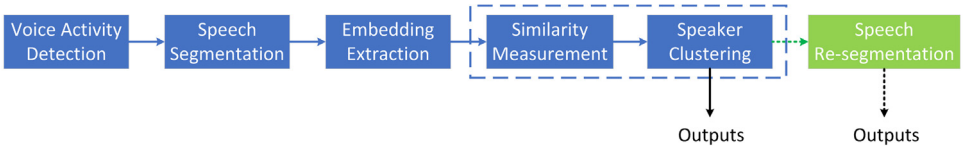

Fig. 14. Diagram of the stage-wise speaker diarization, where speech re-segmentation is an optional module.

the LSTM-based scoring (Lin, Yin, Li, Bredin, & Barras, 2019), self-attentive similarity measurement strategies (Lin, Hou, & Li, 2020b), and joint training of speaker embedding and PLDA scoring (Garcia-Romero et al., 2017). Common clustering algorithms include k-means (Wang, Downey, Wan, Mansfield, & Moreno, 2018), agglomerative hierarchical clustering (Sell et al., 2018), spectral clustering (Lin, Yin, Li, Bredin, & Barras, 2019; Wang, Downey, Wan, Mansfield, & Moreno, 2018), Bayesian Hidden Markov Model based clustering (Diez et al., 2020; Diez et al., 2019; Landini et al., 2020) etc. Recently, Zhang (2018) proposed a non-neural-network deep model, named multilayer bootstrap networks, and applied it to speaker clustering (Li & Zhang, 2018; Zhang, 2016), which demonstrates competitive performance to the common speaker clustering algorithms. However, these clustering algorithms are unsupervised, which is difficult to utilize manually-labeled training data, e.g. the time-stamped speaker ground truth (Zhang, Wang, Zhu, Paisley, & Wang, 2019).

To address the problem, some work formulated speaker clustering as a semi-supervised learning problem (Milner & Hain, 2016; Yu & Hansen, 2017). Specifically, Milner and Hain (2016) built a new DNN iteratively from a pre-trained speaker separation DNN for the speaker clustering of each audio file. In Yu and Hansen (2017), the authors proposed an active-learning-based speaker clustering algorithm, which needs some involvement of human labor during the clustering process.

Some work formulated speaker clustering as a supervised learning problem (Fini & Brutti, 2020; Li, Kreyssig, Zhang, & Woodland, 2019; Zhang, Wang, Zhu, Paisley, & Wang, 2019). Specifically, Li, Kreyssig, Zhang, and Woodland (2019) and Zhang, Wang, Zhu, Paisley, and Wang (2019) defined the speaker labels of training data according to their first appearance in the training data. As shown in Fig. 15, given a sequence of speech segments $\mathcal{E} = \{(\mathbf{e}_n, l_n)|n = 1, 2, \ldots, N\}$ where $l_n \in \{\text{A, B, C, D, E, }\ldots\}$ is the ground truth label of the nth segment with each capital letter representing a speaker, the training label of the sequence $\hat{l}_n \in \{1, 2, 3, 4, 5, \ldots\}$ is tagged with positive integers in the order of the speaker appearance in the sequence. For example, the training labels of the two sequences [E, A, C, A, E, E, C] and [A, C, A, B, B, C, D, B, D] are tagged as [1, 2, 3, 2, 1, 1, 3] and [1, 2, 1, 3, 3, 2, 4, 3, 4] respectively. Under this labeling manner, Zhang, Wang, Zhu, Paisley, and Wang (2019) trained a parameter-sharing RNN clustering model in a supervised way by multiple-instance-learning. They further integrated the RNN with a distance-dependent Chinese restaurant process to address the difficult problem of the unknown number of speakers. An improvement to Zhang, Wang, Zhu, Paisley, and Wang (2019) was further presented in Fini and Brutti (2020). Additionally, the clustering procedure was also modeled by a discriminative sequence-to-sequence neural network (Li, Kreyssig, Zhang, & Woodland, 2019).

9.1.4. Speech re-segmentation

Re-segmentation is an optional step after the speaker clustering. It refines speech boundaries between the speech segments. Variational-Bayesian refinement (Diez, Burget, & Matejka, 2018; Sell & Garcia-Romero, 2015; Sell et al., 2018) is the most famous conventional method. Recently, deep learning based re-segmentation methods were also developed. For example, following the successful application of the RNN to voice activity detection and SCD, Yin, Bredin, and Barras (2018) proposed to address re-segmentation with LSTM.

ground truth sequence	defined training label sequence
E A C A E E C	1 2 3 2 1 1 3
A C A B B C D B D	1 2 1 3 3 2 4 3 4

Fig. 15. Two examples of the training label sequence definition of the supervised clustering in Li, Kreyssig, Zhang, and Woodland (2019).
Source: From Li, Kreyssig, Zhang, and Woodland (2019).

9.1.5. Speech overlap detection

Most stage-wise speaker diarization systems simply assume that only one person is speaking at any time. In other words, they do not consider the speech overlap problem. However, speech overlap is one of the most important factors that hinder the diarization performance (Lin et al., 2020; Ryant et al., 2018, 2019; Sell et al., 2018), since it happens frequently in practice, e.g. a fast multi-speaker conversation. There have been several traditional studies on overlap detection for speaker diarization (Boakye, Trueba-Hornero, Vinyals, & Friedland, 2008; Huijbregts, van Leeuwen, & Jong, 2009; Otterson & Ostendorf, 2007; Yella & Bourlard, 2014). Recently, some deep learning based speech overlap detection methods were also proposed (Bullock, Bredin, & Garcia-Perera, 2020; Huang et al., 2020). Specifically, Bullock et al. (2020) first addressed the overlapped speech detection as a sequence labeling problem by an LSTM-based architecture, and then assigned the detected overlap regions to two speakers by a simple yet effective overlap-aware re-segmentation module. In Huang et al. (2020), a region proposal network was first used to detect overlapped speech, and then removed the highly overlapped segments in the post-processing stage. In addition to the above methods, end-to-end speaker diarization, which will be reviewed in the next section, is another way to deal with the speech overlap problem.

9.2. End-to-end speaker diarization

Because conventional clustering algorithms are unsupervised, it cannot minimize the diarization error rate directly (Fujita, Watanabe, Horiguchi, Xue, & Nagamatsu, 2020) and is difficult to deal with the speech overlap problem (Fujita, Watanabe, Horiguchi, Xue, & Nagamatsu, 2020). Moreover, because each module of the stage-wise speaker diarization in Fig. 14 is optimized independently, the performance is difficult to be boosted. Although several semi-supervised and supervised speaker clustering methods have recently been proposed, the potential of deep neural networks, which are mainly used to extract speaker embeddings in the stage-wise speaker diarization, has not been fully explored yet. To address these problems, Fujita, Kanda, Horiguchi, Nagamatsu, and Watanabe (2019), Fujita et al. (2019) and Fujita, Watanabe, Horiguchi, Xue, and Nagamatsu (2020) proposed an end-to-end diarization method by formulating speaker diarization as a multi-label classification problem.

As shown in Fig. 16, given an acoustic feature sequence of an audio recording $\mathcal{Y} = \{\mathbf{y}_t \in \mathbb{R}^{d_1}|t = 1, 2, \ldots, T\}$, speaker diarization estimates the speaker label sequence $L = \{\boldsymbol{l}_t|t =$

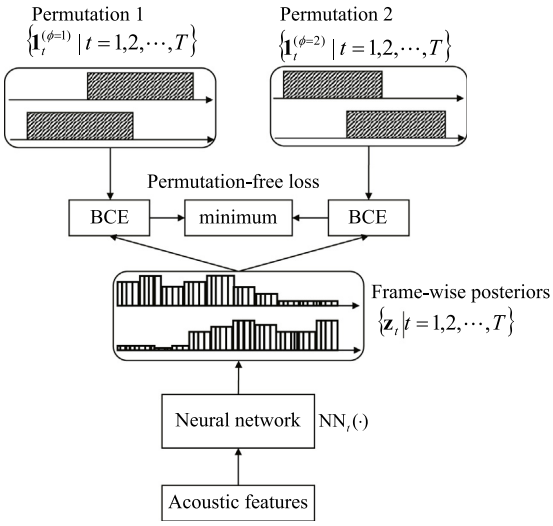

Fig. 16. End-to-end neural diarization with the permutation-free loss for two-speaker diarization.
Source: From Fujita, Watanabe, Horiguchi, Xue, and Nagamatsu (2020).

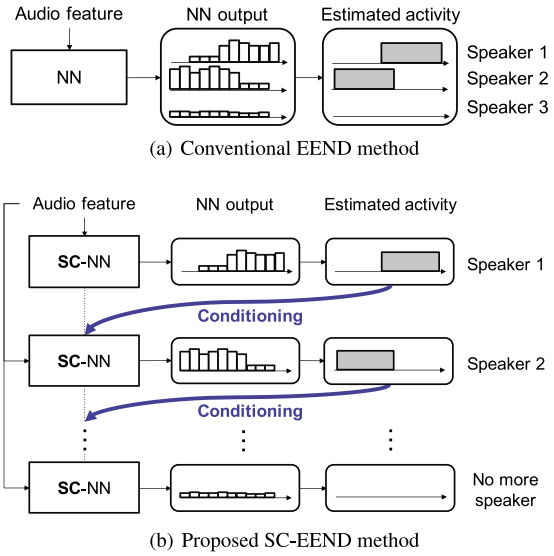

Fig. 17. Diagrams of the conventional end-to-end neural diarization (EEND) method with fixed speaker numbers in Fujita, Watanabe, Horiguchi, Xue, and Nagamatsu (2020), and the speaker-wise conditional end-to-end neural diarization (SC-EEND) method.
Source: From Fujita et al. (2020).

$1, 2, \ldots, T\}$, where $\boldsymbol{l}_t = \{l_{t,j} \in \{0, 1\} | j = 1, 2, \ldots, J\}$ denotes a joint activity of a total number of J speakers at time t. This multi-label formulation can represent speaker overlap regions properly. For example, $l_{t,j} = l_{t,j'} = 1$ with $j \neq j'$ represents that the speech of the speakers j and j' overlaps at time t. With the assumption that each speaker is present independently, the frame-wise posteriors are estimated by a neural network as follows (Fujita, Watanabe, Horiguchi, Xue, & Nagamatsu, 2020):

$$\mathbf{z}_t = [P(l_{t,1}|\mathcal{Y}), P(l_{t,2}|\mathcal{Y}), \ldots, P(l_{t,J}|\mathcal{Y})] = \mathrm{NN}_t(\mathcal{Y}) \in (0, 1)^J \tag{74}$$

where $\mathrm{NN}_t(\cdot)$ denotes a neural network, which was implemented by a BLSTM (Fujita, Kanda, Horiguchi, Nagamatsu, & Watanabe, 2019), or a self-attention based neural network (Fujita et al., 2019).

However, a difficult problem for the end-to-end speaker diarization is the speaker-label permutation ambiguity problem when aligning the ground truth label with the speech recording in preparing the training data. For example, given an audio recording with three speakers A, B and C. If the ground-truth label of the recording is "AAABBC", then the encoded labels "111223" and "222113" are equally correct, making the neural network hard to define a unique training label sequence (Lin, Li, Yang, Wang, & Li, 2020c). To cope with this problem, as shown in Fig. 16, a neural network is trained to minimize the permutation-invariant training (PIT) loss between the output $\mathbf{z}_t$ and the reference speaker label $\mathbf{1}_t \in \{0, 1\}^J$ (Fujita, Kanda, Horiguchi, Nagamatsu, & Watanabe, 2019; Fujita et al., 2019; Fujita, Watanabe, Horiguchi, Xue, & Nagamatsu, 2020):

$$\mathcal{L}^{PF} = \frac{1}{TJ} \min_{\phi \in \mathrm{perm}(J)} \sum_t \mathrm{BCE}(\mathbf{1}_t^{\phi}, \mathbf{z}_t) \tag{75}$$

where $\mathrm{perm}(J)$ is the set of all possible permutations of the speaker identifiers $\{1, 2 \ldots, J\}$, and $\mathbf{1}_t^{\phi}$ is the ϕth permutation of the ground-truth label sequence, and $\mathrm{BCE}(\cdot, \cdot)$ is the binary cross entropy between the label and the network output.

Under the multi-label classification framework (74), the end-to-end diarization system is unable to deal with the test scenario where the number of speakers is larger than the maximum number of speakers in any of the training conversations. Here we bravely call this problem the *fixed speaker capacity* issue for short. To this end, the end-to-end diarization framework is less flexible than the stage-wise methods which can handle any number of speakers in a test conversation (Horiguchi, Fujita, Watanabe, Xue, & Nagamatsu, 2020). To enlarge the speaker capacity, the number of the output nodes of the neural network (74) tends to be set large. Unfortunately, the computational resource for training with the PIT loss will be exponentially increased along with the increase of the number of the output nodes. To deal with this contradiction, Lin et al. (2020c) proposed an optimal mapping loss, which directly computes the best match between the output speaker sequence and the ground-truth speaker sequence through a so-called Hungarian algorithm. It reduces the computational complexity to polynomial time, and meanwhile yields similar performance as the PIT loss. It should be noted that, the fixed speaker capacity issue remains unsolved in the optimal mapping loss.

To address the fixed speaker capacity issue, Fujita et al. (2020) proposed a speaker-wise conditional end-to-end (SC-EEND) speaker diarization method. As shown in Fig. 17, it uses an encoder–decoder architecture to decode each speaker's speech activity iteratively conditioned on the estimated historical speech activities. Although the SC-EEND method achieves increased speaker counting accuracy, it is difficult to handle more than four speakers (Fujita et al., 2020), since it still has to deal with the PIT problem during decoding. Besides, they also proposed an encoder–decoder based attractor calculation method, where a flexible number of speaker attractors are calculated from a speech embedding sequence (Horiguchi et al., 2020). However, the speaker capacity of the method remains limited, due to the PIT loss which determines the assignment of the attractors to the training speakers.

9.3. Online speaker diarization

Most state-of-the-art speaker diarization systems work in an offline manner. Online speaker diarization, which outputs the diarization result right after the audio segment arrives, is not an easy task, since that future information is unavailable when analyzing the current segment (Xue, Horiguchi, Fujita, Watanabe, &

Nagamatsu, 2020). In history, a number of online speaker diarization and speaker tracking solutions have been reported (Dimitriadis & Fousek, 2017; Patino et al., 2018). Here we focus on the deep learning based ones, which can be categorized to stage-wise online diarization and end-to-end online diarization methods.

In respect of the stage-wise online diarization, Zhang, Wang, Zhu, Paisley, and Wang (2019) replaced the commonly used clustering module with a trainable unbounded interleaved-state RNN to make prediction in an online fashion, and it was further improved by introducing a *sample mean loss* (Fini & Brutti, 2020). Additionally, a transformer-based discriminative neural clustering model proposed in Li, Kreyssig, Zhang, and Woodland (2019) can also perform online diarization. However, although it is possible to transfer the online clustering to end-to-end speaker diarization, these methods still suffer from the assumption that only one speaker is present in a single segment (Lin et al., 2020c; Xue et al., 2020).

In respect of the end-to-end online diarization, Xue et al. (2020) extended the self-attention-based end-to-end speaker diarization method in Fig. 16 to an online version, with a speaker-tracing mechanism that is important for the success of the online diarization. Specifically, a straightforward online extension to the framework in Fig. 16 is to perform diarization independently for each chunked recording. However, it is observed that the extension degrades the diarization error rate, due to the speaker permutation inconsistency across the chunk, especially for the short-duration chunks. To overcome this weakness, they applied a speaker-tracing buffer to record the speaker permutation information determined in previous chunks for a correct speaker-tracing in the following chunk. Because this method is limited to two-speaker diarization, more flexible speaker online end-to-end diarization methods need to be further explored (Xue et al., 2020). In addition, von Neumann et al. (2019) presented an all-neural approach to simultaneously conduct speaker counting, diarization, and source separation in a block-online manner, where a speaker-tracing mechanism is also employed to avoid the permutation inconsistency problem across time chunks.

9.4. Multimodal speaker diarization

Although speaker diarization is conventionally an audio-only task, the linguistic content carried by speech signals (Flemotomos, Georgiou, & Narayanan, 2020) and the speaker behaviors, e.g., the movement of lips, recorded by videos (Ding, Xu, Zhang, Cong, & Wang, 2020) offer valuable supplementary cues to the detection of active speakers. To incorporate the aforementioned knowledge, *multimodal speaker diarization* is emerging. Here we summarize some work as follows.

The first class is *audio-linguistic* speaker diarization, Park and Georgiou (2018) integrated lexical cues and acoustic cues together by a gated recurrent unit-based sequence-to-sequence model, which improves the diarization performance by exploring linguistic variability deeply. The effectiveness of using both the linguistic and acoustic cues for diarization has been manifested further in structured scenarios (El Shafey, Soltau, & Shafran, 2019; Flemotomos et al., 2020) where the speakers are assumed to produce distinguishable linguistic patterns. For instance, a teacher is likely to speak in a more didactic style while a student tends to be more inquisitive; a doctor is likely to inquire on symptoms and prescribe while a patient describe symptoms, etc. Another emerging direction is *audio–visual* speaker diarization. In Ding et al. (2020), the authors proposed a self-supervised audio–video synchronization learning method for the scenario where there lacks massive labeled data. The authors in Chung, Lee, and Han (2019) proposed an iterative audio–visual approach which first enrolls speaker models using audio–visual correspondence, and then combines the enrolled models together with the visual information to determine the active speaker. In addition, microphone arrays provide important spatial information as well. For example, Kang, Roy, and Chow (2020) recently combined d-vectors with the spatial information produced from beamforming for the multimodal speaker diarization.

10. Robust speaker recognition

Along with the rapid progress of speaker recognition, the frontier turns to "*recognition in the wild*" (McLaren et al., 2016; Nagrani et al., 2017; Ryant et al., 2018), where lots of domain mismatch and noisy problems arise. To overcome these difficulties, many domain adaptation and noise reduction methods were proposed. In this section, we comprehensively review these robust speaker recognition methods, including domain adaptation in Section 10.1, speech enhancement preprocessing in Section 10.2, and data augmentation techniques in Section 10.3.

10.1. Domain adaptation

Over the past few years, speaker recognition has achieved great success due to the application of deep learning and large amount of labeled speech data. However, because collecting and annotating data for every new application is extremely expensive and time-consuming, sufficient training data may not be always available. For example, although large-scale labeled English databases are publicly available, Cantonese databases may be scarce (Sadjadi et al., 2017). Hence, it is needed to improve the performance of low-resource speaker recognition by using the large amount of auxiliary data. However, there are many distribution mismatch or domain shift problems between the low-resource data and auxiliary data, including different languages, phonemes, recording equipments, etc., which hinder the effectiveness of the auxiliary data. Fortunately, the mismatch problem can be alleviated by domain adaptation techniques.

Without loss of generality, the domain of interest is called the *target* domain, and the domain with sufficient labeled training data is called the *source* domain. The data distributions of the target domain and source domain are denoted as $p_t(\mathbf{x}, y)$ and $p_s(\mathbf{x}, y)$ respectively, where $p_s(\mathbf{x}, y) \neq p_t(\mathbf{x}, y)$. Domain adaptation uses large amount of labeled data in the source domain to solve the problems in the target domain. If the training data in the target domain is manually labeled, then the domain adaptation is supervised; otherwise, it is unsupervised. This paper focuses on the unsupervised domain adaptation, since it is more common and technically more challenging than the supervised domain adaptation in speaker recognition.

Domain adaptation has long been a significant topic in speaker recognition. It received much attention after the domain adaptation challenge in 2013.[6] Over the past years, lots of supervised (Garcia-Romero & McCree, 2014; Wang, Okabe, Lee, & Koshinaka, 2020; Wang, Yamamoto, & Koshinaka, 2016) and unsupervised (Alam, Bhattacharya, & Kenny, 2018; Bousquet & Rouvier, 2019; Garcia-Romero, McCree, Shum, Brummer, & Vaquero, 2014; Glembek et al., 2014; Kanagasundaram, Dean, & Sridharan, 2015; Lee, Wang, & Koshinaka, 2019; Misra & Hansen, 2018; Shum, Reynolds, Garcia-Romero, & McCree, 2014; Villalba & Lleida, 2014; Wang et al., 2020) shallow domain adaptation methods based on the well known i-vector/PLDA pipeline have been developed. Among them, the adaptation is usually accomplished at the back-end, including the methods of compensating the domain mismatch in the i-vector space by an independent

[6] https://www.clsp.jhu.edu/workshops/13-workshop/speaker-and-language-recognition/

Table 10
Summary of adversarial-training-based domain adaptation methods.

	Variable factors	Instantiation	References
1.	Source and target mappings M_s and M_t	Shared parameters $M_s = M_t = M$, or unshared parameters $M_s \neq M_t$.	Chen, Wang, Qian, and Yu (2020), Fang, Zou, Li, Sun, and Ling (2019), Luu, Bell, and Renals (2020), Wang et al. (2018), Xia, Huang, and Hansen (2019)
2.	Domain discriminator loss $\mathcal{L}_{\text{adv}D}$	Binary cross-entropy, multi-class cross-entropy, or Wasserstein distance.	Rohdin et al. (2019), Wang et al. (2018), Zhou et al. (2019)
3.	Adversarial loss $\mathcal{L}_{\text{adv}M}$	The gradient reversal layer, or the GAN loss function.	Bhattacharya, Alam, and Kenny (2019), Bhattacharya, Monteiro, Alam, and Kenny (2019), Luu et al. (2020), Wang et al. (2018), Xia et al. (2019)
4.	Input feature $\mathbf{X}_s$ and $\mathbf{X}_t$	Utterance-level speaker features (e.g. i-vector and x-vector), or frame-level acoustic features (e.g. MFCC and F-bank).	Bhattacharya, Alam, and Kenny (2019), Luu et al. (2020), Tu, Mak, and Chien (2019), Wang et al. (2018)
5.	Adaptation target	Channel invariant, language invariant, phoneme invariant, noise robust, or short utterance compensation.	Bhattacharya, Alam, and Kenny (2019), Tawara, Ogawa, Iwata, Delcroix, and Ogawa (2020), Wang et al. (2018), Zhang, Inoue, and Shinoda (2018), Zhou et al. (2019)

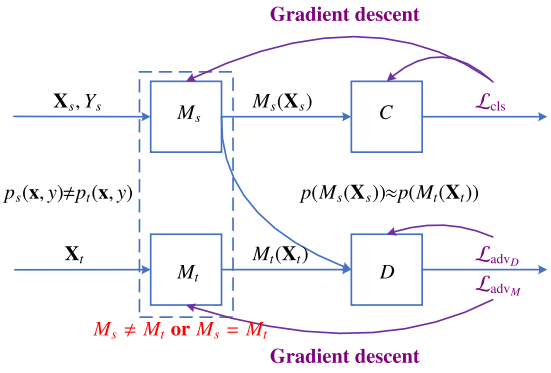

Fig. 18. A unified framework of adversarial-training-based domain adaptation.

module before LDA and PLDA (Alam et al., 2018; Aronowitz, 2014; Kanagasundaram et al., 2015; Misra & Hansen, 2018), and conducting domain adaptation at LDA (Glembek et al., 2014; McLaren & Van Leeuwen, 2011) or PLDA (Garcia-Romero & McCree, 2014; Garcia-Romero et al., 2014; Lee et al., 2019; Shum et al., 2014; Villalba & Lleida, 2014; Wang et al., 2020, 2016). Although the methods are quite effective, here we do not discuss their details, as this article focuses on deep learning based ones.

Recently, many deep learning based domain adaptation methods were proposed. Following the categorization of the domain adaptation techniques (Tzeng, Hoffman, Saenko, & Darrell, 2017; Wang & Deng, 2018), this paper categorizes the deep-learning-based domain adaptation in speaker recognition into the following three classes:

- **Adversarial-training-based domain adaptation:** It seeks to minimize an approximate domain discrepancy distance through an adversarial objective with a domain discriminator.
- **Reconstruction-based domain adaptation:** It assumes that the data reconstruction of the source or target samples can be helpful for improving the performance of domain adaptation.
- **Discrepancy-based domain adaptation:** It aligns the statistical distribution shift between the source and target domains using some mechanisms.

10.1.1. Adversarial-training-based domain adaptation

Before presenting the literature, we first build a unified framework of adversarial-training-based domain adaptation for speaker recognition by drawing lessons from Tzeng et al. (2017). As shown in Fig. 18, the framework consists of:

- $\mathbf{X}_s$ and Y_s drawn from $p_s(\mathbf{x}, y)$ are the source features and speaker labels, and $\mathbf{X}_t$ drawn from $p_t(\mathbf{x}, y)$ are the target features without labels;
- M_s and M_t are the feature mappings for the source and target domains respectively;
- C is a classifier for discriminating speakers in the source domain.
- D is a domain discriminator for discriminating the source domain and target domain.

The adversarial adaptation methods aim at learning M_s and M_t for minimizing the distance between the empirical source and target data distributions in the feature space, i.e. making $p(M_s(\mathbf{X}_s)) \approx p(M_t(\mathbf{X}_t))$. After the adaptation, the recognition models trained on $M_s(\mathbf{X}_s)$ can be directly applied to the target domain.

In the training stage, M_s and C are jointly trained using the standard supervised loss:

$$\begin{aligned}(\hat{M}_s, \hat{C}) &= \arg\min_{M_s, C} \mathcal{L}_{\text{cls}}(\mathbf{X}_s, Y_s; M_s, C) \\ &= -\mathbb{E}_{(\mathbf{x}_s, y_s)\sim(\mathbf{X}_s, Y_s)} \sum_{j=1}^{J} \mathbb{I}_{[j=y_s]} \log[C(M_s(\mathbf{x}_s))]\end{aligned} \tag{76}$$

where J is the number of speakers, and $\mathcal{L}_{\text{cls}}$ is either the standard softmax or its variants described in Section 7, which ensures that the outputs of M_s are speaker-discriminative. To minimize the difference between the source and target representations, the adversarial adaptation methods conduct the following two steps alternatively for D and M_t:

$$\hat{D} = \arg\min_{D} \mathcal{L}_{\text{adv}_D}(\mathbf{X}_s, \mathbf{X}_t; \hat{M}_s, \hat{M}_t, D) \tag{77}$$

$$\hat{M}_t = \arg\min_{M_t} \mathcal{L}_{\text{adv}_M}(\mathbf{X}_s, \mathbf{X}_t; \hat{M}_s, M_t, \hat{D}) \tag{78}$$

where the symbols with hat, such as $\hat{M}_s$, denote that they are fixed during the alternative optimization.

By alternatively minimizing (77) and (78), D and M_t play an adversarial game: D is optimized to predict the domain labels of $M_s(\mathbf{X}_s)$ and $M_t(\mathbf{X}_t)$, while M_t is trained to make the prediction as incorrect as possible. When the training process converges, we have $p(M_s(\mathbf{X}_s)) \approx p(M_t(\mathbf{X}_t))$ since that D is unable to discriminate $M_s(\mathbf{X}_s)$ and $M_t(\mathbf{X}_t)$. In addition, because the output of M_s is speaker-discriminative which is ensured by the speaker classifier C, the output of M_t is supposed to be speaker-discriminative too. For clarity, the manual label requirement in (76), (77), and (78) is summarized in Table 11. Different implementations of the framework in Fig. 18 are summarized in Table 10 which will be reviewed in detail as follows.

One of the most popular domain-adversarial neural network (DANN) architectures in literature is shown in Fig. 19 (Ganin & Lempitsky, 2015). It consists of an encoder M, a speaker classifier

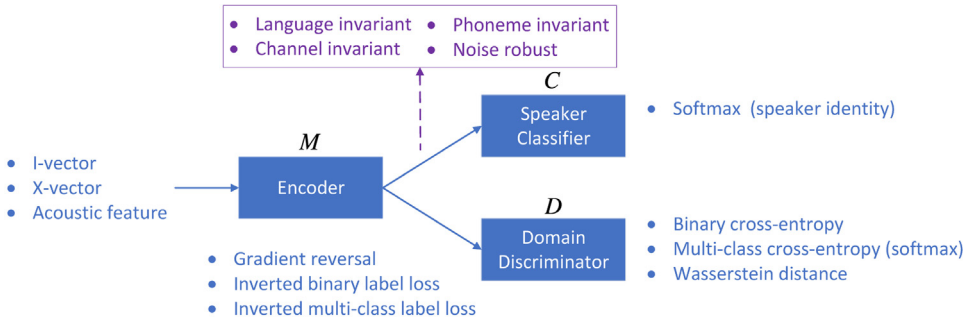

Fig. 19. Domain-adversarial neural network (DANN) architecture with $M_s = M_t = M$.

Table 11
The manual label requirement in the loss functions (76), (77), and (78). The domain labels can be any factors that are needed to be mitigated, including the types of channels, languages, phonemes, noise, etc.

Losses	Target domain speaker labels	Source domain speaker labels	Domain labels
$\mathcal{L}_{\text{cls}}$	✓	✗	✗
$\mathcal{L}_{\text{adv}_D}$	✗	✗	✓
$\mathcal{L}_{\text{adv}_M}$	✗	✗	✓

C, and a domain discriminator D. It follows the framework in Fig. 18 with a constraint $M_s = M_t = M$. Its training alternates the following steps:

(1) The encoder is optimized by merging (76) and (78): $\hat{M} = \arg\min_M [\mathcal{L}_{\text{cls}}(\mathbf{X}_s, Y_s; M, \hat{C}) + \lambda \mathcal{L}_{\text{adv}_M}(\mathbf{X}_s, \mathbf{X}_t; M, \hat{D})]$, where λ is a balance factor.
(2) The speaker classifier is obtained by minimizing (76): $\hat{C} = \arg\min_C \mathcal{L}_{\text{cls}}(\mathbf{X}_s, Y_s; \hat{M}, C)$.
(3) The domain discriminator is obtained by minimizing (77): $\hat{D} = \arg\min_D \mathcal{L}_{\text{adv}_D}(\mathbf{X}_s, \mathbf{X}_t; \hat{M}, D)$.

After training, the encoder is responsible for extracting domain-invariant and speaker-discriminative speaker features, while the speaker classifier and domain discriminator will be discarded.

The domain-adversarial architecture has been used to explore channel-invariant (Chen et al., 2020; Fang et al., 2019; Luu et al., 2020; Wang et al., 2018), language-invariant (Bhattacharya, Alam, & Kenny, 2019; Bhattacharya, Monteiro, Alam, & Kenny, 2019; Rohdin et al., 2019; Tu et al., 2019; Tu, Mak, & Chien, 2020), phoneme-invariant (Tawara et al., 2020; Wang et al., 2019), noise-robust (Meng, Zhao, Li, & Gong, 2019; Peri, Pal, Jati, Somandepalli, & Narayanan, 2020; Zhou et al., 2019) speaker features, etc. Here we present them briefly as follows.

Wang et al. (2018) applied DANN to learn a channel invariant feature extractor from the i-vector subspace. It takes the binary cross-entropy loss as the loss function of $\mathcal{L}_{\text{adv}_D}$ to train the discriminator D. It adds a gradient reversal layer between the encoder and the discriminator to realize a minimax game $\mathcal{L}_{\text{adv}_M} = -\mathcal{L}_{\text{adv}_D}$. Luu et al. (2020) developed a similar framework with Fig. 19 on an x-vector extractor. The framework produces speaker features that are invariant to the granularity of the recording channels. Instead of predicting the concrete domain labels, its domain discriminator predicts whether a pair of speaker embeddings that comes from the same speaker belong to the same recording in a Siamese fashion, see Fig. 20 for the above process. There is also a gradient reversal layer between then x-vector extractor and the domain discriminator. Chen et al. (2020) also proposed a similar work to suppress the channel variability.

In addition to the channel mismatch problem, language mismatch is another challenge in speaker recognition. In Bhattacharya, Alam, and Kenny (2019), the authors applied the domain-adversarial architecture directly to acoustic features for a language invariant feature extractor, where the binary cross-entropy loss and a gradient reversal layer are applied to (77) and (78) respectively. In Bhattacharya, Monteiro, Alam, and Kenny (2019), they further replaced the gradient reversal layer with a generative adversarial network (GAN) loss, a.k.a. inverted-label loss. Specifically, $\mathcal{L}_{\text{adv}_D}$ still adopts the binary cross-entropy loss, while $\mathcal{L}_{\text{adv}_M}$ fools the domain discriminator by inverting the domain labels instead of using a gradient reversal layer, i.e.:

$$\mathcal{L}_{\text{adv}_D}(\mathbf{X}_s, \mathbf{X}_t; \hat{M}, D) = - \mathbb{E}_{\mathbf{x}_s \sim \mathbf{X}_s}[\log(D(\hat{M}(\mathbf{x}_s)))] - \mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t}[\log(1 - D(\hat{M}(\mathbf{x}_t)))] \tag{79}$$

$$\mathcal{L}_{\text{adv}_M}(\mathbf{X}_s, \mathbf{X}_t; M, \hat{D}) = - \mathbb{E}_{\mathbf{x}_s \sim \mathbf{X}_s}[\log(1 - \hat{D}(M(\mathbf{x}_s)))] - \mathbb{E}_{\mathbf{x}_t \sim \mathbf{X}_t}[\log(\hat{D}(M(\mathbf{x}_t)))] \tag{80}$$

where $\mathcal{L}_{\text{adv}_D}$ tags the data from the source domain as "1" and the data from the target domain as "0", and $\mathcal{L}_{\text{adv}_M}$ takes the opposite domain labels. This objective has the same fixed-point properties as the gradient reversal layer but provides stronger gradients to the encoder than the latter (Bhattacharya, Monteiro, Alam, & Kenny, 2019; Tzeng et al., 2017). The authors in Rohdin et al. (2019) trained a language-invariant embedding extractor in an end-to-end fashion, where the embedding extractor is a standard TDNN based x-vector extractor (Fig. 21). They utilized the discriminator to estimate the empirical Wasserstein distance between the source and target samples, and optimized the feature extractor network to minimize the distance in an adversarial manner. Tu et al. (2019) added an additional variational autoencoder (VAE) branch to the standard DANN structure of Fig. 19 for a language-invariant feature extractor as shown in Fig. 22, where the state-of-the-art x-vector was used as the input. The VAE branch performs like a variational regularization which constrains the learned features to be Gaussian. As we know, Gaussian distribution is essential for the effectiveness of the standard PLDA backend. In Tu et al. (2020), they further replaced the VAE with an information-maximized VAE, which not only retains the variational regularization but also inclines to preserve more speaker discriminative information than the VAE.

For text-independent speaker recognition, phonetic information is sometimes harmful, given the fact that it is difficult to ensure shared phonetic coverage across short enrollment and test utterances. However, for long recordings, phonetic information

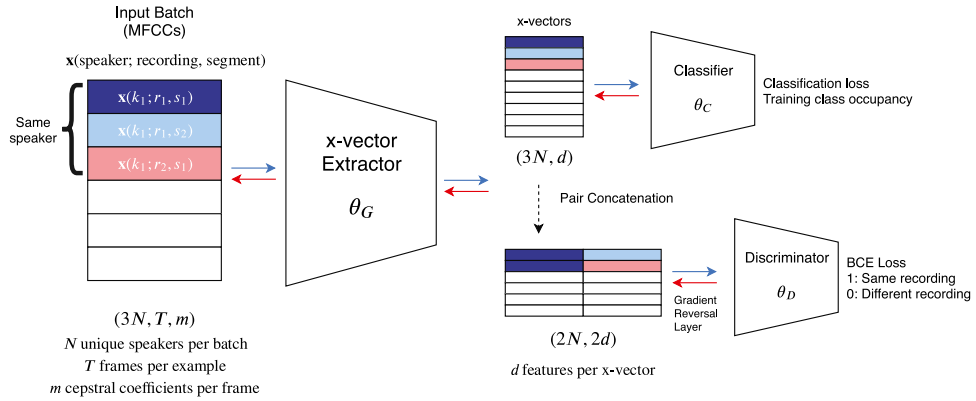

Fig. 20. Channel-invariant domain adaptation. The classifier is trained in the same way as the ordinary x-vector. The discriminator is trained on concatenated pairs of within-speaker pairs. The blue arrows represent the forward propagation. The red arrows represent the backward propagation of gradients. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)
Source: From Luu et al. (2020).

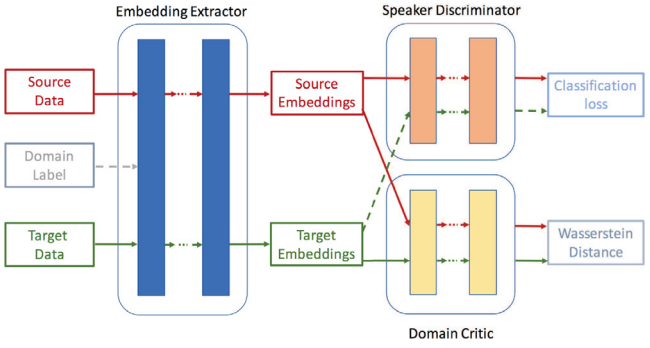

Fig. 21. End-to-end adversarial language adaptation.
Source: From Rohdin et al. (2019).

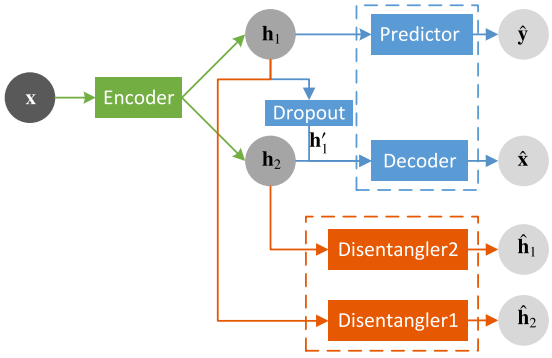

Fig. 23. A brief diagram of unsupervised adversarial invariance.

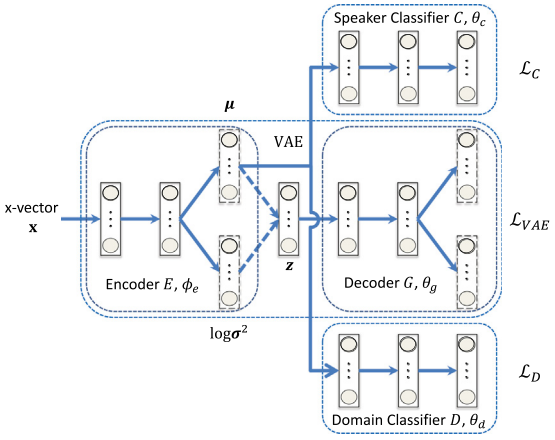

Fig. 22. Variational domain adversarial neural network.
Source: From Tu et al. (2019).

and even more higher level cues, such as word usages, actually differentiates speakers. Quest for such high-level speaker features was an active area of research in about two decades ago. Recently, some studies also investigated its effect on text-independent speaker recognition in the era of deep learning. For example, Wang et al. (2019) applied multi-task learning on frame-level layers to enhance the phonetic information in the frame-level features, and used the adversarial training on segment-level layers to learn phoneme-independent representations. Finally, both operations result in improved performance. Tawara et al. (2020) concluded that phonetic information should be suppressed in text-independent speaker recognition working with frame-wise or extremely short utterances.

Additive noise is one of the most serious interferences of speaker recognition. To explore a noise-robust feature extractor, many works resorted to DANN. For instance, Zhou et al. (2019) used a multi-class cross entropy loss as the loss function of a noise discriminator to train a noise-condition-invariant feature extractor. Meng et al. (2019) applied the adversarial training to learn a feature extractor that are invariant to two kinds of conditions—different environments and different SNRs, where the different environments are represented as a categorical variable and the range of SNRs is formulated as a continuous variable.

Peri et al. (2020) applied an unsupervised adversarial invariance (UAI) architecture to disentangle speaker-discriminative information. As shown in Fig. 23, the encoder generates two latent representations $\mathbf{h}_1$ and $\mathbf{h}_2$ from the x-vector $\mathbf{x}$, where $\mathbf{h}_1$ only contains the speaker-discriminative information, and $\mathbf{h}_2$ contains all other information of $\mathbf{x}$. This was implemented by optimizing the encoder, predictor and decoder together, where the predictor aims to predict speaker labels $\hat{\mathbf{y}}$ from $\mathbf{h}_1$, and the decoder aims to recover $\hat{\mathbf{x}}$ from a concatenation of $\mathbf{h}_2$ and a noise corrupted version of $\mathbf{h}_1$, denoted as $\mathbf{h}'_1$. In order to further encourage the "disentanglement", a minimax game between the disentanglers and the encoder was conducted, where the disentanglers try to reconstruct the two latent representations from each other, while the encoder is optimized against the disentanglers. An important merit of UAI over DANN is that the adversarial game does not need domain labels.

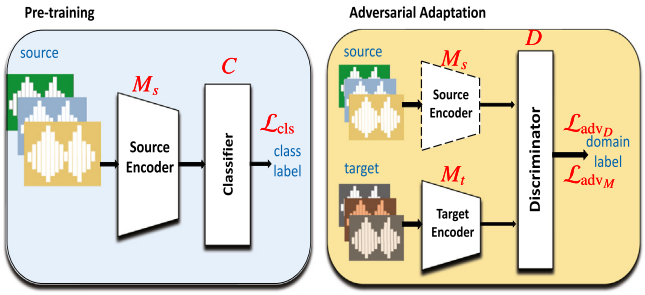

Fig. 24. Adversarial discriminative domain adaptation (ADDA) approach. Note that the source encoder is fixed during the adversarial adaptation stage.
Source: From Xia et al. (2019).

In addition to DANN where $M_s = M_t = M$, another kind of adversarial-training-based domain adaptation methods train an encoder for each domain, i.e. $M_s \neq M_t$. Adversarial discriminative domain adaptation (ADDA) shown in Fig. 24 is such a representative approach (Tzeng et al., 2017). In Xia et al. (2019), the authors applied ADDA to learn an asymmetric mapping that adapts the target domain encoder to the source domain encoder, where the two domains are in different languages.

From the view of the framework in Fig. 18, ADDA is trained by the following two successive steps:

(1) Pre-train a source domain encoder M_s and a speaker classifier C with the labeled source data by (76).
(2) Fix the source domain encoder M_s, and perform adversarial training on the target encoder M_t and domain discriminator D by alternatively minimizing $\mathcal{L}_{\text{adv}_D}$ and $\mathcal{L}_{\text{adv}_M}$ via (77) and (78) respectively. The domain discriminator D minimizes the binary cross-entropy loss, i.e. $\mathcal{L}_{\text{adv}_D} = -\mathbb{E}_{\mathbf{x}_s\sim\mathbf{x}_s}[\log(D(\hat{M}_s(\mathbf{x}_s)))] - \mathbb{E}_{\mathbf{x}_t\sim\mathbf{x}_t}[\log(1 - D(\hat{M}_t(\mathbf{x}_t)))]$. The target encoder M_t minimizes an inverted label loss, i.e. $\mathcal{L}_{\text{adv}_M} = -\mathbb{E}_{\mathbf{x}_t\sim\mathbf{x}_t}[\log(\hat{D}(M_t(\mathbf{x}_t)))]$.[7]

In the test stage, the data from the target domain is first mapped to the shared feature space by the target domain encoder, and then classified by a back-end classifier trained in the source domain, e.g. PLDA.

In order to compensate the unreliability of short utterances, Zhang, Inoue, and Shinoda (2018) proposed to compensate short utterances by long utterances using GAN. Specifically, it uses a generator to generate compensated i-vectors from short-utterance i-vectors, and uses a discriminator to determine whether an i-vector is generated by the generator or extracted from a long utterance. Similarly, Liu and Zhou (2020) addressed the problem by adversely learning a mapping function that maps short embedding features to enhanced embedding features.

Despite of the success of the adversarial-training-based domain adaptation, its training is not easy in practice. For example, DANN learns a common encoder for both domains, which may make the optimization poorly conditioned, since a single encoder has to handle features from two separate domains (Tzeng et al., 2017). Although ADDA is able to learn domain specific encoders, its target domain has no labels. As a result, without shared weights between the encoders, it may quickly fall into a degenerate solution if not properly initialized (Tzeng et al., 2017). To remedy this weakness, a pre-trained source encoder is used to initialize the target encoder, leaving the source encoder fixed during the adversarial training (Tzeng et al., 2017; Xia et al., 2019). Besides, many training difficulties observed in related areas, such as image processing, have been encountered in speaker recognition as well, though seldom discussed in depth.

[7] Different from (80), the constant part $-\mathbb{E}_{\mathbf{x}_s\sim\mathbf{x}_s}[\log(1 - \hat{D}(\hat{M}_s(\mathbf{x}_s)))]$ was removed from $\mathcal{L}_{\text{adv}_M}$, given that M_s is fixed.

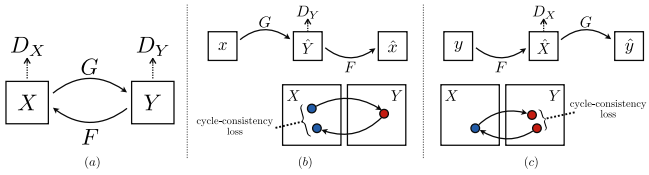

Fig. 25. The cycle GAN architecture.
Source: From Zhu et al. (2017).

10.1.2. Reconstruction-based domain adaptation

In the machine learning community, CycleGAN, which was originally proposed for image-to-image translation (Zhu, Park, Isola, & Efros, 2017), is one of the most common reconstruction-based domain adaptation methods (Wang & Deng, 2018). Recently, it was introduced to speaker recognition (Nidadavolu, Kataria, Villalba, & Dehak, 2019; Nidadavolu, Kataria, Villalba, García-Perera, & Dehak, 2020; Nidadavolu, Villalba, & Dehak, 2019). As shown in Fig. 25, CycleGAN comprises two generators and two discriminators. The generator G transforms the feature X in a domain to the feature Y in another domain, producing an approximation of Y, i.e. $\hat{Y} = G(X)$. The discriminator D_Y, which aligns with G, discriminates between $\hat{Y}$ and Y. The other generator–discriminator pair, i.e. F and D_X, is intended to transfer features from Y to X. The generators and discriminators are trained using a cycle consistency loss and a combination of two adversarial losses, where the cycle consistency loss measures how well the original input is reconstructed after a sequence of two generators, i.e. $F(G(X)) \approx X$ or $G(F(Y)) \approx Y$. Because of the adversarial and cycle reconstruction mechanisms of CycleGAN, it has an outstanding advantage that its training needs neither speaker labels nor paired data between the source and target domains.

At the acoustic feature level, Nidadavolu, Villalba, and Dehak (2019) explored a domain adaptation approach by learning feature mappings between a microphone domain and a telephone domain using CycleGAN. It maps the acoustic features from the target domain (microphone) back to the source domain (telephone), and conducts speaker recognition using the system trained in the source domain. In Nidadavolu, Kataria, Villalba, and Dehak (2019), they further investigated the effectiveness of CycleGAN in low resource scenarios where the target domain only has limited amount of data. They found that, the adaptation system trained on limited amount of target domain data performs slightly better than the adaptation system trained on a larger amount of target domain data, when some noise was added to the data. In Nidadavolu et al. (2020), they developed a CycleGAN-based feature enhancement approach in the log-filter bank space to improve the performance of speaker verification in noisy and reverberant environments.

Besides CycleGAN, the encoder–decoder structure is another popular reconstruction-based domain adaptation method in machine learning (Wang & Deng, 2018). As for speaker recognition, Shon, Mun, Kim, and Ko (2017) combined an autoencoder with a denoising autoencoder to adapt resource-rich source domain data to the target domain.

10.1.3. Discrepancy-based domain adaptation

Discrepancy-based domain adaptation aligns the statistical distribution shift between the source and target domains by using some statistic criteria, including maximum mean discrepancy (MMD), correlation alignment (CORAL), Kullback–Leibler divergence, etc. For speaker verification, those criteria have been widely studied in shallow domain adaptation models (Alam et al.,

2018; Lee et al., 2019). Recently, some of them were also introduced to the deep learning based adaptation approaches. Specifically, Lin, Mak, Li, and Chien (2018) added a MMD based loss to the reconstruction loss of an autoencoder to train a domain-invariant encoder for multi-source adaptation of i-vectors. In Lin, Mak, and Chien (2018) and Lin, Mak, Tu, and Chien (2019), they further proposed a nuisance-attribute autoencoder based on MMD. In Lin, Mak, Li, Su, and Yu (2020), they proposed a multi-level deep neural network adaptation method using MMD and consistency regularization.

10.2. Speech enhancement and de-reverberation preprocessing

Speech is always distorted by noise and reverberation in real-world scenarios. A natural choice of coping with these distortions for speaker recognition is to add a speech enhancement or de-reverberation preprocessing module. Recently, deep learning based speech enhancement and de-reverberation techniques (Wang & Chen, 2018) have been applied to speaker recognition, which can be roughly divided into three categories, i.e. masking-based (Chang & Wang, 2017; Kolbœk, Tan, & Jensen, 2016; Shon, Tang, & Glass, 2019; Zhao, Li, & Zhang, 2019; Zhao, Wang, & Wang, 2014), mapping-based (Novotnỳ, Plchot, Glembek, Burget, et al., 2019; Novotny, Plchot, Matejka, & Glembek, 2018; Oo et al., 2016; Plchot, Burget, Aronowitz, & Matejka, 2016; Sun et al., 2018), and GAN-based (Michelsanti & Tan, 2017) techniques. It should be noted that this section only reviews some general concepts and provides useful cues without considering the details of the speech enhancement techniques, since it is out of the scope of this paper. More details can be found in speech enhancement related papers (Wang & Chen, 2018).

Masking-based speech enhancement has received a lot of attention and shown impressive performance in speech quality and speech intelligibility. It uses a DNN to estimate a time–frequency mask of noisy speech, and then uses the mask to recover the corresponding clean speech. In Chang and Wang (2017), Kolbœk et al. (2016) and Zhao et al. (2014), the authors applied masking-based speech enhancement techniques as an independently noise reduction module for speaker recognition. In Zhao et al. (2019), the authors jointly optimized speech separation and speaker verification networks together. Besides, Shon et al. (2019) designed a *VoiceID loss* which jointly optimize a pre-trained speaker embedding system and a speech enhancement network where the speaker embedding system is fixed during the joint optimization, as shown in Fig. 26.

DNN-based autoencoder, which maps noisy speech directly to its clean counterpart, is another speech enhancement method for speaker recognition. Plchot et al. (2016) trained an autoencoder to map the log magnitude spectrum of noisy speech to its clean counterpart by minimizing the mean squared error, which demonstrated its effectiveness on the text-dependent GMM-MAP and text-independent i-vector systems. The authors of Novotny et al. (2018) explored a similar noise reduction method with Plchot et al. (2016), and applied it to an x-vector system (Novotnỳ et al., 2019). In Oo et al. (2016), the authors enhanced both the amplitude feature and the phase feature, where they used MFCC as the amplitude feature, and modified group delay cepstral coefficients as the phase feature. They concluded that simultaneous enhancing of the amplitude and phase features is more effective than enhancing their individual components alone. Apart from the simplest feedforward DNN, more complicated LSTM based speech enhancement methods were also explored (Sun et al., 2018).

Besides the masking and mapping based speech enhancement, GAN-based speech enhancement methods for speaker recognition were also developed. In specific, Michelsanti and Tan (2017) used conditional GANs to learn a mapping from a noisy spectrum to its enhanced counterpart. The conditional GAN consists of a generator and a discriminator which are trained in an adversarial manner. The generator enhances the noisy spectrum; the discriminator aims to distinguish the enhanced spectrum from their clean counterpart using the noisy spectrum as a condition. In addition, Yu, Tan, Ma, and Guo (2017) proposed a noise robust bottleneck feature extraction method based on adversarial training.

In addition to the above speech enhancement methods for speaker recognition, speech de-reverberation has also been studied in speaker recognition (Guzewich & Zahorian, 2017; Mošner, Matějka, Novotnỳ, & Černockỳ, 2018; sner, rich Plchot, Matějka, rej Novotný, & Černocký, 2018). Recently, far-field and multi-channel speaker recognition also attracted much attention (Cai, Qin, & Li, 2019; Qin, Cai, & Li, 2019; Taherian, Wang, Chang, & Wang, 2020; Taherian, Wang, & Wang, 2019).

10.3. Data augmentation for robust speaker recognition

Large-scale multi-condition training is an effective way to improve the generalization of speaker recognition in noisy environments. Particularly, we have observed that the performance of deep speaker embedding systems appears to be highly dependent on the amount of training data. One way to prepare large-scale noisy training data is *data augmentation*. In Snyder et al. (2018), the authors employed additive noises and reverberation to the original training data for the data augmentation of x-vectors, which has shown to be very effective. Zhu, Ko, and Mak (2019) applied a mixup learning strategy to improve the generalization of x-vector extractors. To improve the performance of speaker verification for children with limited data, Shahnawazuddin, Ahmad, Adiga, and Kumar (2020) made speed and pitch perturbation as well as voice conversion to increase the amount of training data. Besides, Wang et al. (2020) validated the effectiveness of *spectral augmentation*, which was originally proposed for speech recognition, for deep speaker embeddings.

10.4. Other robust methods

Apart from the aforementioned methods, some other robust speaker recognition methods are as follows. For example, Kataria et al. (2020) optimized a deep feature loss for feature-domain enhancement of x-vector extractors. Cai, Cai, and Li (2020) designed a new loss function for noise-robust speaker recognition. Kim et al. (2019) proposed an orthogonal vector pooling strategy to remove unwanted factors. There are also many robust back-ends for speaker verification (Bhattacharya, Alam, Kenn, & Gupta, 2016; Ghahabi & Hernando, 2017; Guo et al., 2018; Mahto, Yamamoto, & Koshinaka, 2017; Yang, Heo, Yoon, & Yu, 2017).

11. Datasets

In this section, we make an overview to existing challenges, and datasets for speaker recognition.

Table 12 summarizes the brief information of most common and some recently developed datasets. The recording methods of the databases are briefly introduced as follows.

- NIST SRE: The National Institute of Standards and Technology (NIST) of America has successfully conducted 15 Speaker Recognition Evaluations (SREs) in the past 20 years. It is the largest and most popular challenge in speaker recognition. More information can be found in Gonzalez-Rodriguez (2014) and Greenberg et al. (2020).

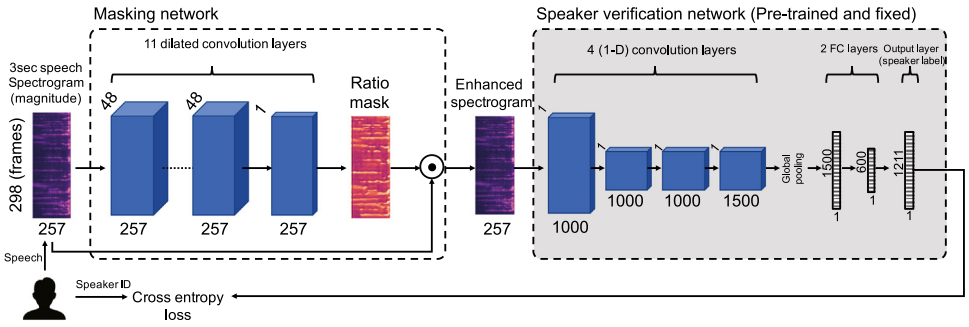

Fig. 26. Flow chart of the VoiceID loss.
Source: From Shon et al. (2019).

Table 12
Popular databases and challenges for speaker recognition. The term "wild" denotes that the audio is acquired across unconstrained conditions. The term "quite" denotes that the data were recorded indoor under a typical office environment. The term "SV", "SI" and "SR" are the abbreviations of "speaker verification", "speaker identification" and "speaker recognition" respectively.

Dataset	Year	Condition	Language	Sample rate	Application	Speakers	Data amount	Annotation method
NIST SRE (Greenberg, Mason, Sadjadi, & Reynolds, 2020)	1996 ∼ 2020	Clean, noisy	Multilingual	–	Text-independent SR	–	–	Hand annotated
VoxCeleb1 (Nagrani et al., 2017)	2017	Multi-media (wild)	Mostly English	16 kHZ	Text-independent SV and SI	1251 (690 males)	153,516 utterances, 352 hours	Automated pipeline
VoxCeleb2 (Chung et al., 2018)	2018	Multi-media (wild)	Multilingual Mostly English	–	Text-independent SV and SI	6112 (3761 males)	1,128,246 utterances, 2442 hours	Automated pipeline
SITW (McLaren et al., 2016)	2016	Multi-media (wild)	–	16 kHZ	Text-independent single and multi-speaker SV	299 (203 males)	2800 utterances	Hand annotated
RSR2015 (Larcher, Lee, Ma, & Li, 2014)	2015	Smart-phones and tablets (quite)	English	16 kHZ	Text-dependent SV	300 (157 males)	196,844 files, 151 hours	Hand annotated
RedDots (Lee et al., 2015)	2015	Mobile devices (through internet)	English	–	Fixed phrase, free speech and text-prompted SV	45	–	Manual, automatic, or semi-automatic
VOICES (Nandwana et al., 2019; Richey et al., 2018)	2018	Far-field microphones (noisy room)	English	48 kHZ	Text-independent SV and SI	300	374,688 files, 1440 hours	Hand annotated
Librispeech (Panayotov, Chen, Povey, & Khudanpur, 2015)	2015	–	English	16 kHz	ASR and SR	Over 9000	1000 hours	Manually annotated
CN-CELEB (Fan et al., 2020)	2019	Multi-media (wild)	Chinese	–	Text-independent SV	1000	130,109 utterances, 274 hours	Automated pipeline with human check
BookTube-Speech (Pham, Li, & Whitehill, 2020)	2020	Multi-media	–	–	Text-independent SV	8450	–	Automatic pipeline
Hi-MIA (Qin, Bu, & Li, 2020)	2020	Microphone arrays (rooms, far-field)	Chinese, English	16 kHZ, 44.1 kHz	Text-dependent SV	340 (175 male)	More than 3,936,003 utterance, 1561 hours	Hand annotated
FFSVC 2020 (Qin et al., 2020)	2020	Close-talk cellphone, far-field microphone arrays (far-field)	Mandarin	16 kHZ, 48 kHZ	Text-dependent and text-independent SV	–	–	Hand annotated
DIHARD1 (Ryant et al., 2018)	2018	Single channel (wild)	English (most), Mandarin	16 kHZ	Speaker diarization	–	40 hours	Hand annotated
DIHARD2 (Ryant et al., 2019)	2019	Single channel and multichannel (wild)	English (most), Mandarin	16 kHZ	Speaker diarization	–	503 files, 339.95 hours	Hand annotated
AMI (Carletta et al., 2005)	2005	Multi-modal	English	–	Speaker diarization	–	100 hours	Hand annotated

- VoxCeleb1,2: The VoxCeleb dataset was collected by a fully automated pipeline based on computer vision techniques from open-source media. The pipeline obtains videos from YouTube, performs active speaker verification using a two-stream synchronization CNN, and confirms the identity of the speaker using CNN based facial recognition (Chung et al., 2018; Nagrani et al., 2017).
- SITW: The speakers in the wild (SITW) database contains hand-annotated speech samples from open-source media for the purpose of benchmarking text-independent speaker recognition technology on single and multi-speaker audio acquired across unconstrained conditions (McLaren et al., 2016).
- RSR2015: The RSR2015 database were recorded indoor under a typical office environment with six mobile devices (five smart-phones and one tablet) (Larcher et al., 2014).
- RedDots: The RedDots project used a mobile app as the recording front-end. Speakers recorded their voices offline

and later on uploaded the recordings to an Apache web server when Internet connection was available (Lee et al., 2015).

- VOICES: The voices obscured in complex environmental settings (VOICES) corpus were recorded in furnished rooms with background noise played in conjunction with foreground speech selected from the LibriSpeech corpus. The displayed noises include television, music, or overlapping speech from multiple speakers (referred to as babble) (Nandwana et al., 2019; Richey et al., 2018).
- LibriSpeech: While intended created for speech recognition rather than verification or diarization, LibriSpeech does include labels of speaker identities and is thus useful for speaker recognition (Panayotov et al., 2015; Pham et al., 2020).
- CN-CELEB: CN-CELEB was collected following a two-stage strategy: firstly the authors used an automated pipeline to extract potential segments of the Person of Interest from "bilibili.com", and then they applied a human check to remove incorrect segments (Fan et al., 2020).
- BookTubeSpeech: The BookTubeSpeech was collected by an automatic pipeline from BookTube videos (Pham et al., 2020).
- Hi-MIA: Hi-MIA database was designed for far-field scenarios. Recordings were captured by multiple microphone arrays located in different directions and distance to the speaker and a high-fidelity close-talking microphone (Qin et al., 2020).
- FFSVC 2020: The far-field speaker verification challenge 2020 (FFSVC20) is designed to boost the speaker verification research with special focus on far-field distributed microphone arrays under noisy conditions in real scenarios (Qin et al., 2020, 2020).
- DIHARD1,2: The DIHARD challenge intended to improve the robustness of diarization systems to variation in recording equipment, noise conditions, and conversational domain (Ryant et al., 2018, 2019).
- AMI: The AMI meeting corpus is a multi-modal dataset consisting of 100 h of meeting recordings, and it was recorded using a wide range of devices including close-talking and far-field microphones, individual and room-view video cameras, projection, a whiteboard, and individual pens, all of which produce output signals that are synchronized with each other (Carletta et al., 2005).

12. Conclusions and discussions

This paper has provided a comprehensive overview of the deep learning based speaker recognition. We have analyzed the relationship between different subtasks, including speaker verification, identification, and diarization, and summarized some common difficulties. Based on the analysis, we summarized the subtasks from three widely studied core issues—speaker feature extraction, speaker diarization, and robust speaker recognition. For speaker feature extraction, we reviewed two kinds of hybrid structures, which are DNN-UBM/i-vector and DNN-BNF/i-vector. In addition, the overview of the state-of-the-art deep speaker embedding was made in respect of four key components, which are the inputs, network structures, temporal pooling strategies, and loss functions respectively. Particularly, we reviewed the loss functions of the end-to-end speaker verification for feature learning from the perspective of different training sample construction methods. For speaker diarization, we reviewed stage-wise diarization, supervised end-to-end diarization, online diarization, and multimodal diarization. For robust speaker recognition, we surveyed three kinds of deep learning based domain adaptation methods as well as several speech preprocessing methods, which deal with the domain mismatch and back-ground noise respectively. Some popular and recently developed datasets were summarized as well. To conclude, deep learning has boosted the performance of speaker recognition to a new high level. We make our best to summarize the recent rapid progress of the deep learning based speaker recognition, hopefully this provides a knowledge resource and further blooms the research community.

Although the deep learning based speaker recognition has achieved a great success, many issues remain to be addressed. Here we list some open problems from the perspectives of network training, loss functions, real-world diarization, and domain adaptation respectively. For the network training, most speaker feature extraction methods need handcraft acoustic features as the input, which may not be optimal. The state-of-the-art deep models have a large number of parameters, which are difficult to be applied to portable devices. The network training also needs large amounts of labeled training data and heavy computation resources. For the loss functions, although so many loss functions have been proposed, there is lack of strong theoretical base for the success of the loss functions, nor theoretical guidance that could lead to better loss functions. Although the verification losses for the end-to-end speaker verification meet the verification process tightly, their potentials have not been fully developed yet. For the real-world diarization, from the recent DIHARD challenges, one can see that speaker diarization is still a hard problem. In a real-world conversation, a speech recording may be contaminated by serious speech overlap and strong background noise. Some technically difficult problems, such as the unknown number of speakers and rapid speaker changes, also hinder the performance of speaker diarization in real-world applications severely. Finally, although many domain adaptation algorithms have been proposed, especially those based on adversarial learning, they have not made landmark progress compared to traditional shallow adaptation methods, e.g. the PLDA based adaptation, which needs further efforts.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors are grateful to Prof. DeLiang Wang, the Co-Editor-in-Chief, and the anonymous reviewers for their valuable comments, which helped to greatly improve the quality of the paper. The authors would also like to thank their colleagues Meng-Zhen Li and Rui Wang for their helpful discussions.

This work was supported in part by the National Key Research and Development Program of China under Grant No. 2018AAA0102200, in part by National Science Foundation of China under Grant No. 61761146001, 61831019, and in part by the Open Research Project of the State Key Laboratory of Media Convergence and Communication, Communication University of China, China under Grant No. SKLMCC2020KF009.

References

Alam, M. J., Bhattacharya, G., & Kenny, P. (2018). Speaker verification in mismatched conditions with frustratingly easy domain adaptation. In *Odyssey* (pp. 176–180).

Anand, P., Singh, A. K., Srivastava, S., & Lall, B. (2019). Few shot speaker recognition using deep neural networks. arXiv preprint arXiv:1904.08775.

Anguera, X., Bozonnet, S., Evans, N., Fredouille, C., Friedland, G., & Vinyals, O. (2012). Speaker diarization: A review of recent research. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(2), 356–370.

Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., & Sivic, J. (2016). NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5297–5307).

Aronowitz, H. (2014). Inter dataset variability compensation for speaker recognition. In *2014 IEEE international conference on acoustics, speech and signal processing* (pp. 4002–4006). IEEE.

Aronowitz, H., & Zhu, W. (2020). Context and uncertainty modeling for online speaker change detection. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 8379–8383).

Bai, Z., Zhang, X. -L., & Chen, J. (2020a). Partial AUC optimization based deep speaker embeddings with class-center learning for text-independent speaker verification. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6819–6823). IEEE.

Bai, Z., Zhang, X. -L., & Chen, J. (2020b). Speaker verification by partial AUC optimization with mahalanobis distance metric learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1533–1548.

Bhattacharya, G., Alam, M. J., Gupta, V., & Kenny, P. (2018). Deeply fused speaker embeddings for text-independent speaker verification. In *INTERSPEECH* (pp. 3588–3592).

Bhattacharya, G., Alam, J., Kenn, P., & Gupta, V. (2016). Modelling speaker and channel variability using deep neural networks for robust speaker verification. In *2016 IEEE spoken language technology workshop* (pp. 192–198). IEEE.

Bhattacharya, G., Alam, M. J., & Kenny, P. (2017). Deep speaker embeddings for short-duration speaker verification. In *INTERSPEECH* (pp. 1517–1521).

Bhattacharya, G., Alam, J., & Kenny, P. (2019). Adapting end-to-end neural speaker verification to new languages and recording conditions with adversarial training. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6041–6045). IEEE.

Bhattacharya, G., Monteiro, J., Alam, J., & Kenny, P. (2019). Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6226–6230). IEEE.

Boakye, K., Trueba-Hornero, B., Vinyals, O., & Friedland, G. (2008). Overlapped speech detection for improved speaker diarization in multiparty meetings. In *2008 IEEE international conference on acoustics, speech and signal processing* (pp. 4353–4356). IEEE.

Bousquet, P. -M., & Rouvier, M. (2019). On robustness of unsupervised domain adaptation for speaker recognition. In *INTERSPEECH* (pp. 2958–2962).

Bredin, H. (2017). Tristounet: Triplet loss for speaker turn embedding. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5430–5434). IEEE.

Bullock, L., Bredin, H., & Garcia-Perera, L. P. (2020). Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7114–7118).

Cai, D., Cai, W., & Li, M. (2020). Within-sample variability-invariant loss for robust speaker recognition under noisy environments. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6469–6473).

Cai, W., Chen, J., & Li, M. (2018). Exploring the encoding layer and loss function in end-to-end speaker and language recognition system. In *Proc. Odyssey 2018 the speaker and language recognition workshop* (pp. 74–81).

Cai, D., Qin, X., & Li, M. (2019). Multi-channel training for end-to-end speaker recognition under reverberant and noisy environment. In *INTERSPEECH* (pp. 4365–4369).

Campbell, J. P., Shen, W., Campbell, W. M., Schwartz, R., Bonastre, J. -F., & Matrouf, D. (2009). Forensic speaker recognition. *IEEE Signal Processing Magazine*, *26*(2), 95–103.

Campbell, W. M., Sturim, D. E., & Reynolds, D. A. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, *13*(5), 308–311.

Carletta, J., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., et al. (2005). The AMI meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction* (pp. 28–39). Springer.

Champod, C., & Meuwly, D. (2000). The inference of identity in forensic speaker recognition. *Speech Communication*, *31*(2–3), 193–203.

Chang, J., & Wang, D. (2017). Robust speaker recognition based on DNN/i-Vectors and speech separation. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 5415–5419). IEEE.

Chen, S., Gopalakrishnan, P., et al. (1998). Speaker, environment and channel change detection and clustering via the bayesian information criterion. In *Proc. DARPA broadcast news transcription and understanding workshop*. Vol. 8 (pp. 127–132). Virginia, USA.

Chen, L., Lee, K. A., Ma, B., Guo, W., Li, H., & Dai, L. -R. (2015). Phone-centric local variability vector for text-constrained speaker verification. In *Sixteenth annual conference of the international speech communication association*.

Chen, N., Qian, Y., & Yu, K. (2015). Multi-task learning for text-dependent speaker verification. In *Sixteenth annual conference of the international speech communication association*.

Chen, Z., Wang, S., Qian, Y., & Yu, K. (2020). Channel invariant speaker embedding learning with joint multi-task and adversarial training. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6574–6578).

Chen, C. -P., Zhang, S. -Y., Yeh, C. -T., Wang, J. -C., Wang, T., & Huang, C. -L. (2019). Speaker characterization using TDNN-LSTM based speaker embedding. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6211–6215). IEEE.

rahman Chowdhury, F. R., Wang, Q., Moreno, I. L., & Wan, L. (2018). Attention-based models for text-dependent speaker verification. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5359–5363). IEEE.

Chung, J. S., Huh, J., Mun, S., Lee, M., Heo, H. -S., Choe, S., et al. (2020). In defence of metric learning for speaker recognition. In *Proc. INTERSPEECH 2020* (pp. 2977–2981).

Chung, J. S., Lee, B. -J., & Han, I. (2019). Who said that?: Audio-visual speaker diarisation of real-world meetings. In *Proc. INTERSPEECH 2019* (pp. 371–375).

Chung, J. S., Nagrani, A., & Zisserman, A. (2018). VoxCeleb2: Deep speaker recognition. In *Proc. INTERSPEECH 2018* (pp. 1086–1090).

Das, R. K., Tian, X., Kinnunen, T., & Li, H. (2020). The attacker's perspective on automatic speaker verification: An overview. arXiv preprint arXiv:2004.08849.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2010). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4690–4699).

Dey, S., Koshinaka, T., Motlicek, P., & Madikeri, S. (2018). DNN based speaker embedding using content information for text-dependent speaker verification. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5344–5348). IEEE.

Dey, S., Madikeri, S., Ferras, M., & Motlicek, P. (2016). Deep neural network based posteriors for text-dependent speaker verification. In *2016 IEEE international conference on acoustics, speech and signal processing* (pp. 5050–5054). IEEE.

Dey, S., Madikeri, S. R., & Motlicek, P. (2018). End-to-end text-dependent speaker verification using novel distance measures. In *INTERSPEECH* (pp. 3598–3602).

Dey, S., Motlicek, P., Madikeri, S., & Ferras, M. (2017). Exploiting sequence information for text-dependent speaker verification. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 5370–5374). Ieee.

Diez, M., Burget, L., Landini, F., Wang, S., & Černocký, H. (2020). Optimizing Bayesian HMM based X-vector clustering for the second DIHARD speech diarization challenge. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6519–6523).

Diez, M., Burget, L., & Matejka, P. (2018). Speaker diarization based on Bayesian HMM with eigenvoice priors. In *Odyssey* (pp. 147–154).

Diez, M., Burget, L., Wang, S., Rohdin, J., & Cernockỳ, J. (2019). Bayesian HMM based x-vector clustering for speaker diarization. In *INTERSPEECH* (pp. 346–350).

Dimitriadis, D., & Fousek, P. (2017). Developing on-line speaker diarization system. In *INTERSPEECH* (pp. 2739–2743).

Ding, S., Chen, T., Gong, X., Zha, W., & Wang, Z. (2020). AutoSpeech: Neural architecture search for speaker recognition. In *Proc. INTERSPEECH 2020* (pp. 916–920).

Ding, Y., Xu, Y., Zhang, S., Cong, Y., & Wang, L. (2020). Self-Supervised learning for audio-visual speaker diarization. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 4367–4371).

Do, C. -T., Barras, C., Le, V. -B., & Sarkar, A. (2013). Augmenting short-term cepstral features with long-term discriminative features for speaker verification of telephone data. In *INTERSPEECH* (pp. 3562–3566).

El Shafey, L., Soltau, H., & Shafran, I. (2019). Joint speech recognition and speaker diarization via sequence transduction. In *Proc. INTERSPEECH 2019* (pp. 396–400).

Fan, Y., Kang, J. W., Li, L. T., Li, K. C., Chen, H. L., Cheng, S. T., et al. (2020). CN-Celeb: A challenging Chinese speaker recognition dataset. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7604–7608).

Fang, X., Zou, L., Li, J., Sun, L., & Ling, Z. -H. (2019). Channel adversarial training for cross-channel text-independent speaker recognition. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6221–6225). IEEE.

Fazel, A., & Chakrabartty, S. (2011). An overview of statistical pattern recognition techniques for speaker verification. *IEEE Circuits and Systems Magazine*, *11*(2), 62–81.

Fini, E., & Brutti, A. (2020). Supervised online diarization with sample mean loss for multi-domain data. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7134–7138).

Flemotomos, N., & Dimitriadis, D. (2020). A memory augmented architecture for continuous speaker identification in meetings. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6524–6528).

Flemotomos, N., Georgiou, P., & Narayanan, S. (2020). Linguistically aided speaker diarization using speaker role information. In *Proc. Odyssey 2020 the speaker and language recognition workshop* (pp. 117–124).

Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., & Watanabe, S. (2019). End-to-End neural speaker diarization with permutation-free objectives. In *Proc. INTERSPEECH 2019* (pp. 4300–4304).

Fujita, Y., Kanda, N., Horiguchi, S., Xue, Y., Nagamatsu, K., & Watanabe, S. (2019). End-to-end neural speaker diarization with self-attention. In *2019 IEEE automatic speech recognition and understanding workshop* (pp. 296–303). IEEE.

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.

Fujita, Y., Watanabe, S., Horiguchi, S., Xue, Y., Shi, J., & Nagamatsu, K. (2020). Neural speaker diarization with speaker-wise chain rule. arXiv preprint arXiv:2006.01796.

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *International conference on machine learning* (pp. 1180–1189).

Gao, Z., Song, Y., McLoughlin, I., Guo, W., & Dai, L. (2018). An improved deep embedding learning method for short duration speaker verification. In *Proc. INTERSPEECH 2018* (pp. 3578–3582).

Gao, Z., Song, Y., McLoughlin, I., Li, P., Jiang, Y., & Dai, L. (2019). Improving aggregation and loss function for better embedding learning in end-to-end speaker verification system. In *Proc. INTERSPEECH 2019* (pp. 361–365).

Garcia-Romero, D., & Espy-Wilson, C. Y. (2011). Analysis of i-vector length normalization in speaker recognition systems. In *Twelfth annual conference of the international speech communication association*.

Garcia-Romero, D., & McCree, A. (2014). Supervised domain adaptation for i-vector based speaker recognition. In *2014 IEEE international conference on acoustics, speech and signal processing* (pp. 4047–4051). IEEE.

Garcia-Romero, D., & McCree, A. (2015). Insights into deep neural networks for speaker recognition. In *Sixteenth annual conference of the international speech communication association*.

Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., & Vaquero, C. (2014). Unsupervised domain adaptation for i-vector speaker recognition. In *Proceedings of odyssey: the speaker and language recognition workshop: Vol. 8*.

Garcia-Romero, D., McCree, A., Snyder, D., & Sell, G. (2020). JHU-HLTCOE system for the voxsrc speaker recognition challenge. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7559–7563).

Garcia-Romero, D., McCree, A., Snyder, D., & Sell, G. (2020). JHU-HLTCOE system for the Voxsrc speaker recognition challenge. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7559–7563). IEEE.

Garcia-Romero, D., Sell, G., & Mccree, A. (2020). MagNetO: X-vector magnitude estimation network plus offset for improved speaker recognition. In *Proc. Odyssey 2020 the speaker and language recognition workshop* (pp. 1–8).

Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., & McCree, A. (2017). Speaker diarization using deep neural network embeddings. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 4930–4934). IEEE.

Georges, M., Huang, J., & Bocklet, T. (2020). Compact speaker embedding: Lrx-Vector. In *Proc. INTERSPEECH 2020* (pp. 3236–3240).

Ghahabi, O., & Hernando, J. (2017). Deep learning backend for single and multisession i-vector speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(4), 807–817.

Ghalehjegh, S. H., & Rose, R. C. (2015). Deep bottleneck features for i-vector based text-independent speaker verification. In *2015 IEEE workshop on automatic speech recognition and understanding* (pp. 555–560). IEEE.

Glembek, O., Ma, J., Matějka, P., Zhang, B., Plchot, O., Bürget, L., et al. (2014). Domain adaptation via within-class covariance correction in i-vector based speaker recognition systems. In *2014 IEEE international conference on acoustics, speech and signal processing* (pp. 4032–4036). IEEE.

Gonzalez-Rodriguez, J. (2014). Evaluating automatic speaker recognition systems: An overview of the nist speaker recognition evaluations (1996-2014). *Loquens*.

Greenberg, C. S., Mason, L. P., Sadjadi, S. O., & Reynolds, D. A. (2020). Two decades of speaker recognition evaluation at the national institute of standards and technology. *Computer Speech and Language*, *60*, Article 101032.

Gu, B., Guo, W., Dai, L., & Du, J. (2020). An improved deep neural network for modeling speaker characteristics at different temporal scales. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6814–6818).

Guo, J., Xu, N., Qian, K., Shi, Y., Xu, K., Wu, Y., et al. (2018). Deep neural network based i-vector mapping for speaker verification using short utterances. *Speech Communication*, *105*, 92–102.

Guzewich, P., & Zahorian, S. A. (2017). Improving speaker verification for reverberant conditions with deep neural network dereverberation processing. In *INTERSPEECH* (pp. 171–175).

Hajavi, A., & Etemad, A. (2019). A deep neural network for short-segment speaker recognition. In *Proc. INTERSPEECH 2019* (pp. 2878–2882).

Hajibabaei, M., & Dai, D. (2018). Unified hypersphere embedding for speaker recognition. arXiv preprint arXiv:1807.08312.

Hansen, J. H., & Hasan, T. (2015). Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, *32*(6), 74–99.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heigold, G., Moreno, I., Bengio, S., & Shazeer, N. (2016). End-to-end text-dependent speaker verification. In *2016 IEEE international conference on acoustics, speech and signal processing* (pp. 5115–5119). IEEE.

Heo, H. -S., weon Jung, J., Yang, I. -H., Yoon, S. -H., jin Shim, H., & Yu, H. -J. (2019). End-to-end losses based on speaker basis vectors and all-speaker hard negative mining for speaker verification. In *Proc. INTERSPEECH 2019* (pp. 4035–4039).

Heo, H.-s., Jung, J.-w., Yang, I.-h., Yoon, S.-h., & Yu, H.-j. (2017). Joint training of expanded end-to-end DNN for text-dependent speaker verification. In *INTERSPEECH* (pp. 1532–1536).

Higuchi, Y., Suzuki, M., & Kurata, G. (2020). Speaker embeddings incorporating acoustic conditions for diarization. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7129–7133).

Hong, Q., Wu, C., Wang, H., & Huang, C. (2020). Combining deep embeddings of acoustic and articulatory features for speaker identification. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7589–7593).

Hong, Q., Wu, C., Wang, H., & Huang, C. (2020). Statistics pooling time delay neural network based on x-vector for speaker verification. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6849–6853).

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. INTERSPEECH 2020* (pp. 269–273).

Hrúz, M., & Zajíc, Z. (2017). Convolutional neural network for speaker change detection in telephone speaker diarization system. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 4945–4949). IEEE.

Huang, Z., Wang, S., & Qian, Y. (2018). Joint i-vector with end-to-end system for short duration text-independent speaker verification. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 4869–4873). IEEE.

Huang, Z., Wang, S., & Yu, K. (2018). Angular softmax for short-duration text-independent speaker verification. In *INTERSPEECH* (pp. 3623–3627).

Huang, Z., Watanabe, S., Fujita, Y., García, P., Shao, Y., Povey, D., et al. (2020). Speaker diarization with region proposal network. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6514–6518).

Huijbregts, M., van Leeuwen, D. A., & Jong, F. (2009). *Speech overlap detection in a two-pass speaker diarization system*. Brighton: ISCA.

India, M., Safari, P., & Hernando, J. (2019). Self multi-head attention for speaker recognition. In *Proc. INTERSPEECH 2019* (pp. 4305–4309).

Irum, A., & Salman, A. (2019). Speaker verification using deep neural networks: A review. *International Journal of Machine Learning and Computing*, *9*(1), 20–25.

Jati, A., & Georgiou, P. G. (2018). An unsupervised neural prediction framework for learning speaker embeddings using recurrent neural networks. In *INTERSPEECH* (pp. 1131–1135).

Jati, A., Peri, R., Pal, M., Park, T. J., Kumar, N., Travadi, R., et al. (2019). Multi-Task discriminative training of hybrid DNN-TVM model for speaker verification with noisy and far-field speech. In *INTERSPEECH* (pp. 2463–2467).

Ji, R., Cai, X., & Bo, X. (2018). An end-to-end text-independent speaker identification system on short utterances. In *Proc. INTERSPEECH 2018* (pp. 3628–3632).

Jiang, Y., Song, Y., McLoughlin, I., Gao, Z., & Dai, L. (2019). An effective deep embedding learning architecture for speaker verification. In *Proc. INTERSPEECH 2019* (pp. 4040–4044).

Jung, J.-w., Heo, H. -S., Kim, J.-h., Shim, H.-j., & Yu, H. -J. (2019). RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. In *Proc. INTERSPEECH 2019* (pp. 1268–1272).

Jung, J.-w., Heo, H. -S., Shim, H.-j., & Yu, H. -J. (2019). Short utterance compensation in speaker verification via cosine-based teacher-student learning of speaker embeddings. In *2019 IEEE automatic speech recognition and understanding workshop* (pp. 335–341). IEEE.

Jung, J.-w., Heo, H.-s., Yang, I.-h., Shim, H.-j., & Yu, H.-j. (2018a). Avoiding speaker overfitting in end-to-end DNNs using raw waveform for text-independent speaker verification. In *Proc. INTERSPEECH 2018* (pp. 3583–3587).

Jung, J. -W., Heo, H. -S., Yang, I. -H., Shim, H. -J., & Yu, H. -J. (2018b). A complete end-to-end speaker verification system using deep neural networks: From raw signals to verification result. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5349–5353). IEEE.

Jung, M., Jung, Y., Goo, J., & Kim, H. (2020). Multi-Task network for noise-robust keyword spotting and speaker verification using CTC-based soft VAD and global query attention. In *Proc. INTERSPEECH 2020* (pp. 931–935).

Jung, Y., Kim, Y., Lim, H., Choi, Y., & Kim, H. (2019). Spatial pyramid encoding with convex length normalization for text-independent speaker verification. In *Proc. INTERSPEECH 2019* (pp. 4030–4034).

weon Jung, J., bin Kim, S., jin Shim, H., ho Kim, J., & Yu, H. -J. (2020). Improved rawnet with feature map scaling for text-independent speaker verification using raw waveforms. In *Proc. INTERSPEECH 2020* (pp. 1496–1500).

Jung, Y., Kye, S. M., Choi, Y., Jung, M., & Kim, H. (2020). Improving multi-scale aggregation using feature pyramid module for robust speaker verification of variable-duration utterances. In *Proc. INTERSPEECH 2020* (pp. 1501–1505).

Kanagasundaram, A., Dean, D., & Sridharan, S. (2015). Improving out-domain PLDA speaker verification using unsupervised inter-dataset variability compensation approach. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 4654–4658). IEEE.

Kang, W., Roy, B. C., & Chow, W. (2020). Multimodal speaker diarization of real-world meetings using d-vectors with spatial features. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6509–6513).

Kataria, S., Nidadavolu, P. S., Villalba, J., Chen, N., García-Perera, P., & Dehak, N. (2020). Feature enhancement with deep feature losses for speaker verification. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7584–7588).

Kenny, P. (2010). Bayesian speaker verification with heavy-tailed priors. In *Odyssey*: *Vol. 14*.

Kenny, P., Boulianne, G., Ouellet, P., & Dumouchel, P. (2007). Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(4), 1435–1447.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., & Dumouchel, P. (2008). A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(5), 980–988.

Kenny, P., Stafylakis, T., Ouellet, P., Gupta, V., & Alam, M. J. (2014). Deep neural networks for extracting Baum-Welch statistics for speaker recognition. In *Odyssey*: *Vol. 2014*, (pp. 293–298).

Kim, I., Kim, K., Kim, J., & Choi, C. (2019). Deep speaker representation using orthogonal decomposition and recombination for speaker verification. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6126–6130). IEEE.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, *52*(1), 12–40.

Kolbœk, M., Tan, Z. -H., & Jensen, J. (2016). Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification. In *2016 IEEE spoken language technology workshop* (pp. 305–311). IEEE.

Kye, S. M., Jung, Y., Lee, H. B., Hwang, S. J., & Kim, H. (2020). Meta-learning for short utterance speaker recognition with imbalance length pairs. In *Proc. INTERSPEECH 2020* (pp. 2982–2986).

Landini, F., Wang, S., Diez, M., Burget, L., Matějka, P., Žmolí ková an, K., et al. (2020). BUT system for the second dihard speech diarization challenge. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6529–6533).

Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and RSR2015. *Speech Communication*, *60*, 56–77.

Lee, K. A., Larcher, A., Wang, G., Kenny, P., Brümmer, N., Leeuwen, D. v., et al. (2015). The RedDots data collection for speaker recognition. In *Sixteenth annual conference of the international speech communication association*.

Lee, K. A., Wang, Q., & Koshinaka, T. (2019). The CORAL+ algorithm for unsupervised domain adaptation of PLDA. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 5821–5825). IEEE.

Lei, Y., Ferrer, L., McLaren, M., & Scheffer, N. (2014). A deep neural network speaker verification system targeting microphone speech. In *Fifteenth annual conference of the international speech communication association*.

Lei, Y., Scheffer, N., Ferrer, L., & McLaren, M. (2014). A novel scheme for speaker recognition using a phonetically-aware deep neural network. In *2014 IEEE international conference on acoustics, speech and signal processing* (pp. 1695–1699). IEEE.

Li, L., Chen, Y., Shi, Y., Tang, Z., & Wang, D. (2017). Deep speaker feature learning for text-independent speaker verification. In *Proc. INTERSPEECH 2017* (pp. 1542–1546).

Li, Q., Kreyssig, F. L., Zhang, C., & Woodland, P. C. (2019). Discriminative neural clustering for speaker diarisation. arXiv preprint arXiv:1910.09703.

Li, R., Li, N., Tuo, D., Yu, M., Su, D., & Yu, D. (2019). Boundary discriminative large margin cosine loss for text-independent speaker verification. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6321–6325). IEEE.

Li, C., Ma, X., Jiang, B., Li, X., Zhang, X., Liu, X., et al. (2017). Deep speaker: An end-to-end neural speaker embedding system. arXiv preprint arXiv:1705.02304.

Li, L., Tang, Z., Shi, Y., & Wang, D. (2019). Gaussian-constrained training for speaker verification. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6036–6040). IEEE.

Li, N., Tuo, D., Su, D., Li, Z., & Yu, D. (2018). Deep discriminative embeddings for duration robust speaker verification. In *INTERSPEECH* (pp. 2262–2266).

Li, M. -Z., & Zhang, X. -L. (2018). An investigation of speaker clustering algorithms in adverse acoustic environments. In *2018 Asia-Pacific signal and information processing association annual summit and conference* (pp. 1462–1466). IEEE.

Li, X., Zhong, J., Yu, J., Hu, S., Wu, X., Liu, X., et al. (2020). Bayesian x-vector: Bayesian neural network based x-vector system for speaker verification. In *Proc. Odyssey 2020 the speaker and language recognition workshop* (pp. 365–371).

Lin, Q., Cai, W., Yang, L., Wang, J., Zhang, J., & Li, M. (2020). DIHARD II is still hard: Experimental results and discussions from the DKU-LENOVO team. In *Proc. Odyssey 2020 the speaker and language recognition workshop* (pp. 102–109).

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., et al. (2017). A structured self-attentive sentence embedding. arXiv preprint arXiv:1703.03130.

Lin, Q., Hou, Y., & Li, M. (2020). Self-attentive similarity measurement strategies in speaker diarization. In *Proc. INTERSPEECH 2020* (pp. 284–288).

Lin, Q., Li, T., Yang, L., Wang, J., & Li, M. (2020). Optimal mapping loss: A faster loss for end-to-end speaker diarization. In *Proc. Odyssey 2020 the speaker and language recognition workshop* (pp. 125–131).

Lin, W., & Mak, M. -W. (2020). Wav2Spk: A simple DNN architecture for learning speaker embeddings from waveforms. In *Proc. INTERSPEECH 2020* (pp. 3211–3215).

Lin, W.-w., Mak, M. -W., & Chien, J. -T. (2018). Multisource i-vectors domain adaptation using maximum mean discrepancy based autoencoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(12), 2412–2422.

Lin, W. -W., Mak, M. -W., Li, L., & Chien, J. -T. (2018). Reducing Domain mismatch by maximum mean discrepancy based autoencoders. In *Odyssey* (pp. 162–167).

Lin, W., Mak, M., Li, N., Su, D., & Yu, D. (2020). Multi-Level deep neural network adaptation for speaker verification using MMD and consistency regularization. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6839–6843).

Lin, W., Mak, M. -W., Tu, Y., & Chien, J. -T. (2019). Semi-supervised nuisance-attribute networks for domain adaptation. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6236–6240). IEEE.

Lin, Q., Yin, R., Li, M., Bredin, H., & Barras, C. (2019). LSTM based similarity measurement with spectral clustering for speaker diarization. In *Proc. INTERSPEECH 2019* (pp. 366–370).

Ling, S., Salazar, J., Liu, Y., & Kirchhoff, K. (2020). BERTphone: Phonetically-aware encoder representations for utterance-level speaker and language recognition. In *Proc. Odyssey 2020 the speaker and language recognition workshop* (pp. 9–16).

Liu, Y., He, L., & Liu, J. (2019). Large margin softmax loss for speaker verification. In *Proc. INTERSPEECH 2019* (pp. 2873–2877).

Liu, Y., He, L., Liu, J., & Johnson, M. T. (2018). Speaker embedding extraction with phonetic information. In *Proc. INTERSPEECH 2018* (pp. 2247–2251).

Liu, Y., Qian, Y., Chen, N., Fu, T., Zhang, Y., & Yu, K. (2015). Deep feature for text-dependent speaker verification. *Speech Communication*, *73*, 1–13.

Liu, X., Sahidullah, M., & Kinnunen, T. (2020). A comparative re-assessment of feature extractors for deep speaker embeddings. In *Proc. INTERSPEECH 2020* (pp. 3221–3225).

Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., & Song, L. (2017). Sphereface: Deep hypersphere embedding for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 212–220).

Liu, K., & Zhou, H. (2020). Text-independent speaker verification with adversarial learning on short utterances. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6569–6573).

Lozano-Diez, A., Silnova, A., Matejka, P., Glembek, O., Plchot, O., Pesan, J., et al. (2016). Analysis and optimization of bottleneck features for speaker recognition. In *Odyssey* (pp. 352–357).

Luu, C., Bell, P., & Renals, S. (2020). Channel adversarial training for speaker verification and diarization. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7094–7098).

Mahto, S., Yamamoto, H., & Koshinaka, T. (2017). i-Vector transformation using a novel discriminative denoising autoencoder for noise-robust speaker recognition. In *INTERSPEECH* (pp. 3722–3726).

McLaren, M., Ferrer, L., Castan, D., & Lawson, A. (2016). The speakers in the wild, SITW speaker recognition database. In *INTERSPEECH* (pp. 818–822).

McLaren, M., Ferrer, L., & Lawson, A. (2016). Exploring the role of phonetic bottleneck features for speaker and language recognition. In *2016 IEEE international conference on acoustics, speech and signal processing* (pp. 5575–5579). IEEE.

McLaren, M., Lei, Y., & Ferrer, L. (2015). Advances in deep neural network approaches to speaker recognition. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 4814–4818). IEEE.

McLaren, M., Lei, Y., Scheffer, N., & Ferrer, L. (2014). Application of convolutional neural networks to speaker recognition in noisy conditions. In *Fifteenth annual conference of the international speech communication association*.

McLaren, M., & Van Leeuwen, D. (2011). Source-normalized LDA for robust speaker recognition using i-vectors from multiple speech sources. *IEEE Transactions on Audio, Speech, and Language Processing*, *20*(3), 755–766.

Meng, Z., Zhao, Y., Li, J., & Gong, Y. (2019). Adversarial speaker verification. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6216–6220). IEEE.

Michelsanti, D., & Tan, Z. -H. (2017). Conditional generative adversarial networks for speech enhancement and noise-robust speaker verification. In *Proc. INTERSPEECH 2017* (pp. 2008–2012).

Milner, R., & Hain, T. (2016). DNN-Based speaker clustering for speaker diarisation. In *Proceedings of the annual conference of the international speech communication association* (pp. 2185–2189). Sheffield.

Mingote, V., Miguel, A., Ribas, D., Giménez, A. O., & Lleida, E. (2019). Optimization of false acceptance/rejection rates and decision threshold for end-to-end text-dependent speaker verification systems. In *INTERSPEECH* (pp. 2903–2907).
Misra, A., & Hansen, J. H. (2018). Maximum-likelihood linear transformation for unsupervised domain adaptation in speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(9), 1549–1558.
Mošner, L., Matějka, P., Novotnỳ, O., & Černockỳ, J. H. (2018). Dereverberation and beamforming in far-field speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5254–5258). IEEE.
Muckenhirn, H., Doss, M. M., & Marcell, S. (2018). Towards directly modeling raw speech signal for speaker verification using CNNs. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 4884–4888). IEEE.
Nagrani, A., Chung, J. S., & Zisserman, A. (2017). VoxCeleb: A large-scale speaker identification dataset. In *Proc. INTERSPEECH 2017* (pp. 2616–2620).
Nandwana, M. K., Van Hout, J., McLaren, M., Richey, C., Lawson, A., & Barrios, M. A. (2019). The voices from a distance challenge 2019 evaluation plan. arXiv preprint arXiv:1902.10828.
von Neumann, T., Kinoshita, K., Delcroix, M., Araki, S., Nakatani, T., & Haeb-Umbach, R. (2019). All-neural online source separation, counting, and diarization for meeting analysis. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 91–95). IEEE.
Nidadavolu, P. S., Kataria, S., Villalba, J., & Dehak, N. (2019). Low-resource domain adaptation for speaker recognition using cycle-gans. In *2019 IEEE automatic speech recognition and understanding workshop* (pp. 710–717). IEEE.
Nidadavolu, P., Kataria, S., Villalba, J., García-Perera, P., & Dehak, N. (2020). Unsupervised feature enhancement for speaker verification. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7599–7603).
Nidadavolu, P. S., Villalba, J., & Dehak, N. (2019). Cycle-GANS for domain adaptation of acoustic features for speaker recognition. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6206–6210). IEEE.
Novoselov, S., Shulipa, A., Kremnev, I., Kozlov, A., & Shchemelinin, V. (2018). On deep speaker embeddings for text-independent speaker recognition. In *Proc. Odyssey 2018 the speaker and language recognition workshop* (pp. 378–385).
Novotnỳ, O., Plchot, O., Glembek, O., Burget, L., et al. (2019). Analysis of DNN speech signal enhancement for robust speaker recognition. *Computer Speech and Language*, *58*, 403–421.
Novotny, O., Plchot, O., Matejka, P., & Glembek, O. (2018). On the use of DNN autoencoder for robust speaker recognition. arXiv preprint arXiv:1811.02938.
Nunes, J. A. C., Macêdo, D., & Zanchettin, C. (2020). AM-MobileNet1D: A portable model for speaker recognition. arXiv preprint arXiv:2004.00132.
Okabe, K., Koshinaka, T., & Shinoda, K. (2018). Attentive statistics pooling for deep speaker embedding. In *Proc. INTERSPEECH 2018* (pp. 2252–2256).
Oo, Z., Kawakami, Y., Wang, L., Nakagawa, S., Xiao, X., & Iwahashi, M. (2016). DNN-Based amplitude and phase feature enhancement for noise robust speaker identification. In *INTERSPEECH* (pp. 2204–2208).
Otterson, S., & Ostendorf, M. (2007). Efficient use of overlap information in speaker diarization. In *2007 IEEE workshop on automatic speech recognition & understanding* (pp. 683–686). IEEE.
Pal, M., Kumar, M., Peri, R., Park, T. J., Hyun Kim, S., Lord, C., et al. (2020). Speaker diarization using latent space clustering in generative adversarial network. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6504–6508).
Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 5206–5210). IEEE.
Park, T. J., & Georgiou, P. (2018). Multimodal speaker segmentation and diarization using lexical and acoustic cues via sequence to sequence neural networks. In *Proc. INTERSPEECH 2018* (pp. 1373–1377).
Patino, J., Yin, R., Delgado, H., Bredin, H., Komaty, A., Wisniewski, G., et al. (2018). Low-latency speaker spotting with online diarization and detection. In *Odyssey* (pp. 140–146).
Peddinti, V., Povey, D., & Khudanpur, S. (2015). A time delay neural network architecture for efficient modeling of long temporal contexts. In *Sixteenth annual conference of the international speech communication association*.
Peri, R., Pal, M., Jati, A., Somandepalli, K., & Narayanan, S. (2020). Robust speaker recognition using unsupervised adversarial invariance. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6614–6618).
Pham, M., Li, Z., & Whitehill, J. (2020). Toward better speaker embeddings: Automated collection of speech samples from unknown distinct speakers. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7089–7093).
Plchot, O., Burget, L., Aronowitz, H., & Matejka, P. (2016). Audio enhancing with DNN autoencoder for speaker recognition. In *2016 IEEE international conference on acoustics, speech and signal processing* (pp. 5090–5094). IEEE.
Povey, D., Cheng, G., Wang, Y., Li, K., Xu, H., Yarmohammadi, M., et al. (2018). Semi-orthogonal low-rank matrix factorization for deep neural networks. In *INTERSPEECH* (pp. 3743–3747).
Pruzansky, S., & Mathews, M. V. (1964). Talker-recognition procedure based on analysis of variance. *Journal of the Acoustical Society of America*, *36*(11), 2041–2047.
Qin, X., Bu, H., & Li, M. (2020). HI-MIA: A far-field text-dependent speaker verification database and the baselines. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7609–7613).
Qin, X., Cai, D., & Li, M. (2019). Far-field end-to-end text-dependent speaker verification based on mixed training data with transfer learning and enrollment data augmentation. In *INTERSPEECH* (pp. 4045–4049).
Qin, X., Li, M., Bu, H., Das, R. K., Rao, W., Narayanan, S., et al. (2020). The FFSVC 2020 evaluation plan. arXiv preprint arXiv:2002.00387.
Qin, X., Li, M., Bu, H., Rao, W., Das, R. K., Narayanan, S., et al. (2020). The INTERSPEECH 2020 far-field speaker verification challenge. arXiv preprint arXiv:2005.08046.
Qu, X., Wang, J., & Xiao, J. (2020). Evolutionary algorithm enhanced neural architecture search for text-independent speaker verification. In *Proc. INTERSPEECH 2020* (pp. 961–965).
Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. In *2018 IEEE spoken language technology workshop* (pp. 1021–1028). IEEE.
Reynolds, D. A. (2002). An overview of automatic speaker recognition technology. In *2002 IEEE international conference on acoustics, speech, and signal processing, Vol. 4* (pp. IV–4072). IEEE.
Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*(1–3), 19–41.
Richardson, F., Reynolds, D. A., & Dehak, N. (2015a). A unified deep neural network for speaker and language recognition. In *Sixteenth annual conference of the international speech communication association*.
Richardson, F., Reynolds, D., & Dehak, N. (2015b). Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters*, *22*(10), 1671–1675.
Richey, C., Barrios, M. A., Armstrong, Z., Bartels, C., Franco, H., Graciarena, M., et al. (2018). Voices obscured in complex environmental settings (voices) corpus. In *Proc. INTERSPEECH 2018* (pp. 1566–1570).
Rohdin, J., Silnova, A., Diez, M., Plchot, O., Matějka, P., & Burget, L. (2018). End-to-end DNN based speaker recognition inspired by i-vector and PLDA. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 4874–4878). IEEE.
Rohdin, J., Stafylakis, T., Silnova, A., Zeinali, H., Burget, L., & Plchot, O. (2019). Speaker verification using end-to-end adversarial language adaptation. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6006–6010). IEEE.
Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., et al. (2018). *First DIHARD challenge evaluation plan: Tech. rep.*.
Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., et al. (2019). *Second dihard challenge evaluation plan: Linguistic data consortium, tech. rep.*
Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., et al. (2019). The second DIHARD diarization challenge: Dataset, task, and baselines. In *Proc. INTERSPEECH 2018* (pp. 978–982).
Rybicka, M., & Kowalczyk, K. (2020). On parameter adaptation in softmax-based cross-entropy loss for improved convergence speed and accuracy in DNN-based speaker recognition. In *Proc. INTERSPEECH 2020* (pp. 3805–3809).
Sadjadi, S. O., Ganapathy, S., & Pelecanos, J. W. (2016). The ibm 2016 speaker recognition system. arXiv preprint arXiv:1602.07291.
Sadjadi, S. O., Kheyrkhah, T., Tong, A., Greenberg, C. S., Reynolds, D. A., Singer, E., et al. (2017). The 2016 NIST speaker recognition evaluation. In *INTERSPEECH* (pp. 1353–1357).
Safari, P., India, M., & Hernando, J. (2020). Self-Attention encoding and pooling for speaker recognition. In *Proc. INTERSPEECH 2020* (pp. 941–945).
Sarı, L., Thomas, S., Hasegawa-Johnson, M., & Picheny, M. (2019). Pre-training of speaker embeddings for low-latency speaker change detection in broadcast news. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6286–6290). IEEE.
Sarkar, A. K., Do, C. -T., Le, V. -B., & Barras, C. (2014). Combination of cepstral and phonetically discriminative features for speaker verification. *IEEE Signal Processing Letters*, *21*(9), 1040–1044.
Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 815–823).
Sell, G., & Garcia-Romero, D. (2014). Speaker diarization with PLDA i-vector scoring and unsupervised calibration. In *2014 IEEE spoken language technology workshop* (pp. 413–417). IEEE.
Sell, G., & Garcia-Romero, D. (2015). Diarization resegmentation in the factor analysis subspace. In *2015 IEEE international conference on acoustics, speech and signal processing* (pp. 4794–4798). IEEE.
Sell, G., Garcia-Romero, D., & McCree, A. (2015). Speaker diarization with i-vectors from DNN senone eriors. In *Sixteenth annual conference of the international speech communication association*.
Sell, G., Snyder, D., McCree, A., Garcia-Romero, D., Villalba, J., Maciejewski, M., et al. (2018). Diarization is hard: Some experiences and lessons learned for the JHU team in the inaugural DIHARD challenge. In *Interspeech: Vol. 2018*, (pp. 2808–2812).
Seo, S., Rim, D. J., Lim, M., Lee, D., Park, H., Oh, J., et al. (2019). Shortcut connections based deep speaker embeddings for end-to-end speaker verification system. In *INTERSPEECH* (pp. 2928–2932).

Shahnawazuddin, S., Ahmad, W., Adiga, N., & Kumar, A. (2020). In-Domain and out-of-domain data augmentation to improve children's speaker verification system in limited data scenario. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7554–7558).

Shon, S., Mun, S., Kim, W., & Ko, H. (2017). Autoencoder based domain adaptation for speaker recognition under insufficient channel information. In *Proc. INTERSPEECH 2017* (pp. 1014–1018).

Shon, S., Tang, H., & Glass, J. (2019). VoiceID Loss: Speech enhancement for speaker verification. In *Proc. INTERSPEECH 2019* (pp. 2888–2892).

Shum, S. H., Reynolds, D. A., Garcia-Romero, D., & McCree, A. (2014). Unsupervised clustering approaches for domain adaptation in speaker recognition systems. In *Proceedings of Odyssey: The speaker and language recognition workshop* (pp. 265–272).

Siegler, M. A., Jain, U., Raj, B., & Stern, R. M. (1997). Automatic segmentation, classification and clustering of broadcast news audio. In *Proc. DARPA speech recognition workshop*: *Vol. 1997*.

Snell, J., Swersky, K., & Zemel, R. (2017). Prototypical networks for few-shot learning. In *Advances in neural information processing systems* (pp. 4077–4087).

sner, L. M., rich Plchot, O., Matějka, P., rej Novotný, O., & Černocký, J. (2018). Dereverberation and beamforming in robust far-field speaker recognition. In *Proc. INTERSPEECH 2018* (pp. 1334–1338).

Snyder, D. (2020). *X-vectors: Robust neural embeddings for speaker recognition* (Ph.D. thesis), Johns Hopkins University.

Snyder, D., Garcia-Romero, D., & Povey, D. (2015). Time delay deep neural network-based universal background models for speaker recognition. In *2015 IEEE workshop on automatic speech recognition and understanding* (pp. 92–97). IEEE.

Snyder, D., Garcia-Romero, D., Povey, D., & Khudanpur, S. (2017). Deep neural network embeddings for text-independent speaker verification. In *INTERSPEECH* (pp. 999–1003).

Snyder, D., Garcia-Romero, D., Sell, G., McCree, A., Povey, D., & Khudanpur, S. (2019). Speaker recognition for multi-speaker conversations using x-vectors. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 5796–5800). IEEE.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., & Khudanpur, S. (2018). X-vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5329–5333). IEEE.

Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., & Khudanpur, S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE spoken language technology workshop* (pp. 165–170). IEEE.

Snyder, D., Villalba, J., Chen, N., Povey, D., Sell, G., Dehak, N., et al. (2019). The JHU speaker recognition system for the voices 2019 challenge. In *INTERSPEECH* (pp. 2468–2472).

Stafylakis, T., Rohdin, J., Plchot, O., Mizera, P., & Burget, L. (2019). Self-Supervised speaker embeddings. In *Proc. INTERSPEECH 2019* (pp. 2863–2867).

Sun, L., Du, J., Jiang, C., Zhang, X., He, S., Yin, B., et al. (2018). Speaker diarization with enhancing speech for the first DIHARD challenge. In *INTERSPEECH* (pp. 2793–2797).

Sun, G., Zhang, C., & Woodland, P. C. (2019). Speaker diarisation using 2D self-attentive combination of embeddings. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 5801–5805). IEEE.

Taherian, H., Wang, Z. -Q., Chang, J., & Wang, D. (2020). Robust speaker recognition based on single-channel and multi-channel speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *28*, 1293–1302.

Taherian, H., Wang, Z. -Q., & Wang, D. (2019). Deep learning based multi-channel speaker recognition in noisy and reverberant environment. In *INTERSPEECH* (pp. 4070–4074).

Tang, Y., Ding, G., Huang, J., He, X., & Zhou, B. (2019). Deep speaker embedding learning with multi-level pooling for text-independent speaker verification. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6116–6120). IEEE.

Tang, Z., Li, L., Wang, D., & Vipperla, R. (2016). Collaborative joint training with multitask recurrent model for speech and speaker recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(3), 493–504.

Tawara, N., Ogawa, A., Iwata, T., Delcroix, M., & Ogawa, T. (2020). Frame-level phoneme-invariant speaker embedding for text-independent speaker recognition on extremely short utterances. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6799–6803).

Togneri, R., & Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits and Systems Magazine*, *11*(2), 23–61.

Torfi, A., Dawson, J., & Nasrabadi, N. M. (2018). Text-independent speaker verification using 3d convolutional neural networks. In *2018 IEEE international conference on multimedia and expo* (pp. 1–6). IEEE.

Travadi, R., & Narayanan, S. (2019). Total variability layer in deep neural network embeddings for speaker verification. *IEEE Signal Processing Letters*, *26*(6), 893–897.

Tu, Y., Mak, M. -W., & Chien, J. -T. (2019). Variational domain adversarial learning for speaker verification. In *Proc. INTERSPEECH 2019* (pp. 4315–4319).

Tu, Y., Mak, M., & Chien, J. (2020). Information maximized variational domain adversarial learning for speaker verification. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6449–6453).

Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7167–7176).

Variani, E., Lei, X., McDermott, E., Moreno, I. L., & Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE international conference on acoustics, speech and signal processing* (pp. 4052–4056). IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., et al. (2019). State-of-the-art speaker recognition for telephone and video speech: The JHU-MIT submission for NIST SRE18. In *Proc. INTERSPEECH 2019* (pp. 1488–1492).

Villalba, J., Chen, N., Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., et al. (2020). State-of-the-art speaker recognition with neural network embeddings in NIST SRE18 and speakers in the wild evaluations. *Computer Speech and Language*, *60*, Article 101026.

Villalba, J., & Lleida, E. (2014). Unsupervised adaptation of PLDA by using variational Bayes methods. In *2014 IEEE international conference on acoustics, speech and signal processing* (pp. 744–748). IEEE.

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2018). Generalized end-to-end loss for speaker verification. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 4879–4883). IEEE.

Wang, D., & Chen, J. (2018). Supervised speech separation based on deep learning: An overview. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(10), 1702–1726.

Wang, F., Cheng, J., Liu, W., & Liu, H. (2018). Additive margin softmax for face verification. *IEEE Signal Processing Letters*, *25*(7), 926–930.

Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. *Neurocomputing*, *312*, 135–153.

Wang, Q., Downey, C., Wan, L., Mansfield, P. A., & Moreno, I. L. (2018). Speaker diarization with LSTM. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 5239–5243). IEEE.

Wang, R., Gu, M., Li, L., Xu, M., & Zheng, T. F. (2017). Speaker segmentation using deep speaker vectors for fast speaker change scenarios. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 5420–5424). IEEE.

Wang, S., Huang, Z., Qian, Y., & Yu, K. (2019). Discriminative neural embedding learning for short-duration text-independent speaker verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(11), 1686–1696.

Wang, Q., Okabe, K., Lee, K. A., & Koshinaka, T. (2020). A generalized framework for domain adaptation of plda in speaker recognition. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6619–6623). IEEE.

Wang, Q., Rao, W., Sun, S., Xie, L., Chng, E. S., & Li, H. (2018). Unsupervised domain adaptation via domain adversarial training for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing* (pp. 4889–4893). IEEE.

Wang, S., Rohdin, J., Burget, L., Plchot, O., Qian, Y., & Yu, K. (2019). On the usage of phonetic information for text-independent speaker embedding extraction. In *Proc. INTERSPEECH 2019* (pp. 1148–1152).

Wang, S., Rohdin, J., Plchot, O., Burget, L., Yu, K., & Černocký, J. (2020). Investigation of specaugment for deep speaker embedding learning. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7139–7143).

Wang, J., Wang, K. -C., Law, M. T., Rudzicz, F., & Brudno, M. (2019). Centroid-based deep metric learning for speaker recognition. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 3652–3656). IEEE.

Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., et al. (2018). Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5265–5274).

Wang, J., Xiao, X., Wu, J., Ramamurthy, R., Rudzicz, F., & Brudno, M. (2020). Speaker diarization with session-level speaker embedding refinement using graph neural networks. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 7109–7113).

Wang, Q., Yamamoto, H., & Koshinaka, T. (2016). Domain adaptation using maximum likelihood linear transformation for plda-based speaker verification. In *2016 IEEE international conference on acoustics, speech and signal processing* (pp. 5110–5114). IEEE.

Wang, Z., Yao, K., Li, X., & Fang, S. (2020). Multi-resolution multi-head attention in deep speaker embedding. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6464–6468).

Wei, Y., Du, J., & Liu, H. (2020). Angular margin centroid loss for text-independent speaker recognition. In *Proc. INTERSPEECH 2020* (pp. 3820–3824).

Wen, Y., Zhang, K., Li, Z., & Qiao, Y. (2016). A discriminative feature learning approach for deep face recognition. In *European conference on computer vision* (pp. 499–515). Springer.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., & Li, H. (2015). Spoofing and countermeasures for speaker verification: A survey. *Speech Communication*, *66*, 130–153.
Wu, Y., Guo, C., Gao, H., Hou, X., & Xu, J. (2020). Vector-based attentive pooling for text-independent speaker verification. In *Proc. INTERSPEECH 2020* (pp. 936–940).
Xia, W., Huang, J., & Hansen, J. H. (2019). Cross-lingual text-independent speaker verification using unsupervised adversarial discriminative domain adaptation. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 5816–5820). IEEE.
Xiang, X., Wang, S., Huang, H., Qian, Y., & Yu, K. (2019). Margin matters: Towards more discriminative deep neural network embeddings for speaker recognition. In *2019 Asia-Pacific signal and information processing association annual summit and conference* (pp. 1652–1656). IEEE.
Xie, W., Nagrani, A., Chung, J. S., & Zisserman, A. (2019). Utterance-level aggregation for speaker recognition in the wild. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 5791–5795). IEEE.
Xue, Y., Horiguchi, S., Fujita, Y., Watanabe, S., & Nagamatsu, K. (2020). Online end-to-end neural diarization with speaker-tracing buffer. arXiv preprint arXiv:2006.02616.
Yadav, S., & Rai, A. (2018). Learning discriminative features for speaker identification and verification. In *INTERSPEECH* (pp. 2237–2241).
Yadav, S., & Rai, A. (2020). Frequency and temporal convolutional attention for text-independent speaker recognition. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6794–6798).
Yang, I. -H., Heo, H. -S., Yoon, S. -H., & Yu, H. -J. (2017). Applying compensation techniques on i-vectors extracted from short-test utterances for speaker verification using deep neural network. In *2017 IEEE international conference on acoustics, speech and signal processing* (pp. 5490–5494). IEEE.
Yao, Q., & Mak, M. -W. (2018). SNR-Invariant multitask deep neural networks for robust speaker verification. *IEEE Signal Processing Letters*, *25*(11), 1670–1674.
Yella, S. H., & Bourlard, H. (2014). Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1688–1700.
Yella, S. H., & Stolcke, A. (2015). A comparison of neural network feature transforms for speaker diarization. In *Sixteenth annual conference of the international speech communication association*.
Yin, R., Bredin, H., & Barras, C. (2017). Speaker change detection in broadcast TV using bidirectional long short-term memory networks. In *Proc. INTERSPEECH 2017* (pp. 3827–3831).
Yin, R., Bredin, H., & Barras, C. (2018). Neural speech turn segmentation and affinity propagation for speaker diarization. In *Proc. INTERSPEECH 2018* (pp. 1393–1397).
Yu, Y. -Q., Fan, L., & Li, W. -J. (2019). Ensemble additive margin softmax for speaker verification. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6046–6050). IEEE.
Yu, C., & Hansen, J. H. (2017). Active learning based constrained clustering for speaker diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *25*(11), 2188–2198.
Yu, Y. -Q., & Li, W. -J. (2020). Densely connected time delay neural network for speaker verification. In *Proc. INTERSPEECH 2020* (pp. 921–925).
Yu, H., Tan, Z. -H., Ma, Z., & Guo, J. (2017). Adversarial network bottleneck features for noise robust speaker verification. In *Proc. INTERSPEECH 2017* (pp. 1492–1496).
Zajíc, Z., Hrúz, M., & Müller, L. (2017). Speaker diarization using convolutional neural network for statistics accumulation refinement. In *INTERSPEECH* (pp. 3562–3566).
Zeinali, H., Burget, L., Sameti, H., Glembek, O., & Plchot, O. (2016). Deep neural networks and hidden Markov models in i-vector-based text-dependent speaker verification. In *Odyssey* (pp. 24–30).
Zeinali, H., Sameti, H., Burget, L., et al. (2017). Text-dependent speaker verification based on i-vectors, neural networks and hidden Markov models. *Computer Speech and Language*, *46*, 53–71.
Zhang, X. -L. (2016). Universal background sparse coding and multilayer bootstrap network for speaker clustering. In *INTERSPEECH* (pp. 1858–1862).
Zhang, X. -L. (2018). Multilayer bootstrap networks. *Neural Networks*, *103*, 29–43.
Zhang, S. -X., Chen, Z., Zhao, Y., Li, J., & Gong, Y. (2016). End-to-end attention based text-dependent speaker verification. In *2016 IEEE spoken language technology workshop* (pp. 171–178). IEEE.
Zhang, J., Inoue, N., & Shinoda, K. (2018). I-vector transformation using conditional generative adversarial networks for short utterance speaker verification. In *Proc. INTERSPEECH 2018* (pp. 3613–3617).
Zhang, C., & Koishida, K. (2017). End-to-end text-independent speaker verification with triplet loss on short utterances. In *INTERSPEECH* (pp. 1487–1491).
Zhang, C., Koishida, K., & Hansen, J. H. (2018). Text-independent speaker verification based on triplet convolutional neural network embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(9), 1633–1644.
Zhang, A., Wang, Q., Zhu, Z., Paisley, J., & Wang, C. (2019). Fully supervised speaker diarization. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6301–6305). IEEE.
Zhang, Y., Yu, M., Li, N., Yu, C., Cui, J., & Yu, D. (2019). Seq2seq attentional siamese neural networks for text-dependent speaker verification. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6131–6135). IEEE.
Zhao, F., Li, H., & Zhang, X. (2019). A robust text-independent speaker verification method based on speech separation and deep speaker. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6101–6105). IEEE.
Zhao, X., Wang, Y., & Wang, D. (2014). Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(4), 836–845.
Zhao, Y., Zhou, T., Chen, Z., & Wu, J. (2020). Improving deep CNN networks with long temporal context for text-independent speaker verification. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6834–6838).
Zheng, H., Zhang, S., & Liu, W. (2015). Exploring robustness of DNN/RNN for extracting speaker Baum-Welch statistics in mismatched conditions. In *Sixteenth annual conference of the international speech communication association*.
Zhong, Y., Arandjelović, R., & Zisserman, A. (2018). GhostVLAD for set-based face recognition. In *Asian conference on computer vision* (pp. 35–50). Springer.
Zhou, J., Jiang, T., Li, L., Hong, Q., Wang, Z., & Xia, B. (2019). Training multitask adversarial network for extracting noise-robust speaker embedding. In *ICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing* (pp. 6196–6200). IEEE.
Zhou, J., Jiang, T., Li, Z., Li, L., & Hong, Q. (2019). Deep speaker embedding extraction with channel-wise feature responses and additive supervision softmax loss function. In *Proc. INTERSPEECH 2019* (pp. 2883–2887).
Zhou, D., Wang, L., Lee, K. A., Wu, Y., Liu, M., Dang, J., et al. (2020). Dynamic margin softmax loss for speaker verification. In *Proc. INTERSPEECH 2020* (pp. 3800–3804).
Zhou, T., Zhao, Y., Li, J., Gong, Y., & Wu, J. (2019). CNN with phonetic attention for text-independent speaker verification. In *2019 IEEE automatic speech recognition and understanding workshop* (pp. 718–725). IEEE.
Zhu, Y., Ko, T., & Mak, B. (2019). Mixup learning strategies for text-independent speaker verification. In *Proc. INTERSPEECH 2019* (pp. 4345–4349).
Zhu, Y., Ko, T., Snyder, D., Mak, B., & Povey, D. (2018). Self-attentive speaker embeddings for text-independent speaker verification. In *INTERSPEECH* (pp. 3573–3577).
Zhu, Y., & Mak, B. (2020a). Orthogonal training for text-independent speaker verification. In *ICASSP 2020 - 2020 IEEE international conference on acoustics, speech and signal processing* (pp. 6584–6588).
Zhu, Y., & Mak, B. (2020b). Orthogonality regularizations for end-to-end speaker verification. In *Proc. Odyssey 2020 the speaker and language recognition workshop* (pp. 17–23).
Zhu, J. -Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232).