TEXT-INDEPENDENT SPEAKER VERIFICATION WITH VERY SHORT UTTERANCES

by

İsmail Rasim Ülgen

B.S., Electrical and Electronics Engineering, Boğaziçi University, 2019

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Electrical and Electronics Engineering
Boğaziçi University

2023

ACKNOWLEDGEMENTS

Firstly, I am extremely grateful to my thesis advisor Prof. Levent Arslan, whose guidance and support are the most valuable and essential parts of my research.

I would like to extend my gratitude to Prof. Murat Saraclar, who has been an inspirational figure for me in the first place, for his very valuable lessons about scientific research.

I am most grateful to my family; my father Şeref, my mother Keziban, and my brother Berkay. I love them deeply.

Lastly, I want to thank my partner Aleyna Habeşoğlu, who has always been by my side with her love and understanding. I can't imagine how hard this process would be without her.

ABSTRACT

TEXT-INDEPENDENT SPEAKER VERIFICATION WITH VERY SHORT UTTERANCES

The accuracy of the text-independent speaker verification suffers greatly when the speech duration is very short. In this thesis, some methods are proposed aiming to compensate for the drastic performance degradation in speaker verification with very short utterances. Firstly, methods that try to leverage the additional information from large-scale speaker datasets are proposed in order to enhance the limited speaker information that is present in the very short speech utterances. Secondly, the problem of short utterances is tackled in a more specific way in terms of the phonetic content of the speech. An analysis of phonetic mismatch between verification utterances is performed, along with experiments of a back-end scoring module that is aware of the phonetic mismatch in speaker verification. Furthermore, contributions to the speaker verification in general, which might be applicable to the very short duration conditions are presented. A novel loss function for back-end scoring module training is introduced. The proposed loss function outperformed the baseline loss function in all cases, including very short duration scenario. Lastly, a novel unsupervised domain adaptation of the discriminative back-end scoring for speaker verification is proposed. The proposed adaptation method improved the performance of the out-of-domain back-end scoring model in the target domain in all cases. The relative improvement of the proposed method, compared to baseline adaptation methods, is highest in short duration conditions.

ÖZET

ÇOK KISA KAYITLARLA METİN BAĞIMSIZ KONUŞMACI DOĞRULAMA

Konuşma süresi çok kısa olduğunda metin-bağımsız konuşmacı doğrulamanın başarısı büyük ölçüde düşmektedir. Bu tezde, çok kısa ifadelerle konuşmacı doğrulamadaki ciddi performans düşüşünü telafi etmeyi amaçlayan bazı yöntemler önerilmektedir. İlk olarak, çok kısa konuşma ifadelerinde mevcut olan sınırlı konuşmacı bilgisini geliştir-mek için büyük ölçekli konuşmacı veri kümelerinden gelen ek bilgilerden yararlanmaya çalışan yöntemler önerilmektedir. İkinci olarak, kısa konuşma ifadeleri sorunu, konuşmanın fonetik içeriği açısından daha spesifik bir şekilde ele alınmaktadır. Doğrulama ifadeleri arasındaki fonetik uyumsuzluğun bir analizi ve konuşmacı doğrulamasında fonetik uyumsuzluğun farkında olan bir arka uç skorlama modülünün deneyleri gerçekleştirilmiştir. Genel olarak konuşmacı doğrulamasına yönelik, ancak çok kısa süreli koşullara uygulanabilecek katkılar da sunulmuştur. Arka uç puanlama modülü eğitimi için yeni bir kayıp fonksiyonu önerilmiştir. Önerilen kayıp fonksiyonu, çok kısa süreli senaryo da dahil olmak üzere tüm durumlarda referans temel kayıp fonksiyonundan daha iyi performans göstermiştir. Son olarak, konuşmacı doğrulamada kullanılan ayrıştırıcı arka uç skorlarması için yeni bir denetimsiz alan uyarlaması önerilmiştir. Önerilen uyarlama yöntemi, tüm durumlarda alan dışı arka uç puanlama modelinin hedef alandaki performansını iyileştirmiştir. Önerilen yöntemin temel uyarlama yöntemlerine kıyasla göreceli iyileşmesi, kısa süreli koşullarda en yüksektir.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

A	Loading matrix
C_k	Utterance set for a given speaker
D	Discriminator
E_y	Expected value for discriminator loss
E_z	Expected value for generator loss
g_w	1-Lipschitz function
G	Generator
l_n	Binary label for target/nontarget trials
$LGAN$	GAN loss
N_{pC}	Number of occurrences for a given phone
$P_{FA}(\theta)$	False alarm probability
$P_{FR}(\theta)$	False reject probability
$P_{nontarget}$	Prior probability of nontarget trials
P_{target}	Prior probability of target trials
S_b	Between-class covariance for LDA
S_w	Within-class covariance for LDA
x_a	Anchor example
x_n	Negative example
x_p	Positive example
z	Generator input
α	Weight of the original embedding in merging
β_n	Weight of the auxiliary embeddings in merging
θ	Threshold value for verification
λ	Gradient penalty weight
μ_k	Class-center of x-vectors
π_k	Class probability for PLDA speakers
σ	Sigmoid Function

ϕ	Margin value for triplet loss
Φ_b	Between-class covariance matrix for PLDA
Φ_w	Within-class covariance matrix for PLDA

LIST OF ACRONYMS/ABBREVIATIONS

2-D	2 Dimensional
BN	Batch Normalization
CNN	Convolutional Neural Network
CQCC	Constant Q Cepstral Coefficients
DCF	Detection Cost Function
DCT	Discrete Cosine Transform
DNN	Deep Neural Network
EER	Equal Error Rate
EM	Expectation Maximization
FA	False Alarm
FFNN	Feed-Forward Neural Network
FR	False Reject
GAN	Generative Adversarial Network
GMM-UBM	Gaussian Mixture Model - Universal Background Model
GPLDA	Gaussian Probabilistic Linear Discriminant Analysis
IN	Instance Normalization
LDA	Linear Discriminant Analysis
LPCC	Linear Prediction Cepstral Coefficients
MFCC	Mel-Frequency Cepstral Coefficients
MRTF	Multi-Resolution Time Frequency
NIST	National Institute of Standards and Technology
NPLDA	Neural Probabilistic Linear Discriminant Analysis
OOD	Out of Domain
PLAR	Perceptual Log Area Ratio
PLDA	Probabilistic Linear Discriminant Analysis
ReLu	Rectified Linear Unit
SDC	Soft Detection Cost
SN-WLDA	Source Normalized Weighted Linear Discriminant analysis

SOTA	State of The Art
SRE	Speaker Recognition Evaluation
SRE	Speaker Recognition Evaluation
TDNN	Time-Delay Neural Network
TFC	Time-Frequency Cepstral Coefficients
WCCN	Within-Class Covariance Normalization
WGAN	Wasserstein Generative Adversarial Network
WGAN-GP	Wasserstein Generative Adversarial Network with Gradient Penalty

1. INTRODUCTION

The task of speaker verification is to verify a person's identity by using the biometric characteristics of the voice. Verification with voice might be employed as a practical and secure automatic verification approach that can be integrated into applications using speech as an input, such as call-centers, smart home assistants, and mobile apps. It might be used not only as a single point of verification but also as an additional measure for increased security. The speaker verification task can be classified into two sub-problems in terms of the content of the speech: text-dependent and text-independent. In text-dependent case, the users a limited to utter a fixed phrase, while there is no limitation on the content in text-independent case. This thesis focuses on the text-independent case which can be easily integrated into applications as a background verification method without the need for an additional verification prompt.

Since the verification is usually an initial step for most applications to proceed, speaker verification with short utterances of speech becomes crucial to make the process faster and more practical as an applicable verification method. However, the accuracy of text-independent speaker verification degrades drastically with the decreasing duration of input speech. In this thesis, some methods are experimented for compensating the performance degradation that occurs when the utterance duration is very short.

In recent years, speaker representations from a deep neural network that is trained with a large amount of data, have become state of the art (SOTA) quickly. Usually, those speaker representations are compared by a back-end scoring module in order to assess similarity/dissimilarity for verification. However, many of those SOTA methods also suffer greatly from performance degradation in very short duration case. In this thesis, the proposed methods are applied to one of those SOTA verification methods, that is called x-vector speaker embeddings [1], in conjunction with probabilistic linear discriminant analysis (PLDA) [2], in order to compensate for the performance degradation.

The information is expected to be very limited in a very short duration of speech. From this point of view, methods that try to leverage additional information that is related to the corresponding speaker are proposed in order to improve the performance of verification. Large-scale speaker data is hypothesized to approximate the speaker-space with a sufficient number of speakers and a sufficient amount of speech from each speaker.

In this thesis, the most similar speaker model from a large-scale data is employed as the auxiliary information to the speaker model obtained from the original short utterance. Firstly, a direct combination of the auxiliary speaker model and original speaker model is experimented using the weighted merge of the speaker embeddings. Secondly, neural networks with a novel loss function are utilized to generate a more robust speaker-specific embedding that contains more speaker information from the original speaker embedding from the short utterance. The contribution of the auxiliary information in the neural network generation process and adversarial training that leverages generator-discriminator structure was also examined.

Phonetic mismatch between enrollment and authentication utterances is also a result of the very short speech utterances in the text-independent speaker verification. It is most likely that enrollment and authentication utterances have different phonetic content since the phrase can be anything. In long speech, the content is usually normalized because of the fact that each phone in the language is likely to occur at some point. Phonetic mismatch primarily affects target cases where low verification scores are produced between two utterances of the same speaker, resulting in false reject errors.

In this thesis, the effect of the phonetic mismatch on the verification performance is analyzed in a more detailed way by controlled experiments. A neural-network based back-end scoring module for x-vector speaker embeddings is experimented for compensating performance degradation. Then, the phonetic information is introduced to back-end scoring module, aiming to compensate for phonetic mismatch specifically.

During this thesis work, some applied methods would be effective not only in short duration conditions but also in general. A novel loss function utilized in generative neural network training for enhancing speaker embeddings turned out to be applied to training a back-end scoring model for speaker verification as well. The proposed triplet PLDA loss function is applied to a back-end scoring module for speaker verification that is called Neural PLDA. The Neural PLDA trained with the proposed loss function is compared to Neural PLDA trained with the loss introduced in [3], in terms of speaker verification performance.

Finally, a novel unsupervised domain adaptation approach is proposed for the discriminate type of back-end scoring modules for speaker verification. The proposed approach leverages strong assumptions about speech datasets and creates pseudo target/nontarget examples from unlabeled data for discriminative adaptation. The target examples are created by sampling non-overlapping chunks from a single utterance that likely contains a single speaker. The nontarget examples are created by sampling random chunks from different utterances where the probability of being from the same speaker is low. The proposed domain adaptation approach is applied to Neural PLDA.

The images created within the scope of this thesis, the copyright of which has been transferred to the publisher, have been used in the thesis book in accordance with the publisher's publication policy on the reuse of graphics produced by the author, which can be found on the publisher's website.

In summary, the contribution of this thesis can be listed as:

- A basic method that merges auxiliary information with the original speaker embedding is proposed in order to improve speaker verification performance
- A neural network based method that generates more robust speaker embeddings from the speaker embedding of a very short utterance is introduced
- Effect of the phonetic mismatch between enrollment and authentication speech on the verification performance is analyzed, and some valuable insights are presented

- A novel loss function for discriminative PLDA training is introduced. The proposed loss function also showed its effectiveness in very short duration conditions
- A novel unsupervised domain adaptation method for discriminative PLDA is proposed. The proposed method compensated for some of the performance degradation resulting from domain shift between training and testing conditions.

The structure of this thesis is as follows: in Chapter 2 review of the proposed method in the literature is presented. The theoretical background of the speaker verification method, which is subject to improvement, is presented in Chapter 3. In Chapter 4, the proposed methods relying on the use of auxiliary information are introduced. In Chapter 5, the analysis of phonetic mismatch is presented along with verification results of the proposed back-end scoring module. In Chapter 6, related contributions to speaker verification in general that can be included in this thesis are introduced. Finally, Chapter 7 summarizes the work and concludes this thesis.

2. RELATED WORK

The task of speaker verification with short utterances has been relatively popular, mainly due to the practical advantage in real-time applications. The performance degradation has been observed as the utterance duration decreases, in almost all popular speaker verification approaches such as GMM-UBM, i-vector and x-vector [4–8]. Other than the obvious fact that short utterances have less information, many works have tried to identify the causes of such degradation in more detail. Experiments showed that performance degradation primarily comes from a drastic increase in intra-speaker variability when dealing with very short utterances [9–11]. Verification performance was also reported to be hurt by decreasing inter-speaker variability [9, 10].

Apart from the insufficiency of speaker information in very short utterances, in unconstrained text-independent speaker verification, linguistic content of an utterance varies, especially in very short duration scenario. In the long duration scenario, this is less of a concern since the linguistic content is averaged over many samples. For this reason, the variation in linguistic content, which is reported to be largely responsible for the high intra-speaker variability, constitutes a crucial challenge for this task [9–12].

In order to mitigate performance degradation, many methods have been introduced with different perspectives on the problem. Those works can roughly be grouped into methods focusing on: feature extraction, speaker model estimation, classification, and score normalization depending on the nature of the speaker verification approach. [9, 12]. Additionally, another group can be formed from works on speaker verification in very short utterances, not necessarily focused on very short utterances but rather on speaker verification generally, including all practical durations.

Among the feature extraction methods aiming to improve verification in very short utterances, the main approaches are to extract complementary and/or more detailed information from the limited speech data. In [13], the authors introduced time-

frequency cepstral features (TFC), which applies discrete cosine transform (DCT) on cepstrum-time matrix and selects the most important elements of the transformation. Those features are said to be extracting temporal information from speech signals in a better way than the most popular approach which is concatenating mel-frequency ceptsral coeffients (MFCC) with its deltas and improving verification performance, especially in very short time scenarios. Inspired by this work, in [14] the authors applied 2-D DCT on time-frequency matrix in different scales and introduced multi-resolution time-frequency (MRTF) feature. MRTF is reported to improve speaker verification in short utterances thanks to the ability to focus on different details in temporal and frequency information. In [15], constant-q cepstral coefficients (CQCC) are introduced. CQCC enables the features to have more frequency resolution in lower frequency bins and more time resolution in high-frequency bins than the classical MFCC features and improves speaker verification in very short duration conditions [15].

The work in [16], combines mostly used MFCC features with complementary features such as linear predictive cepstral coefficients (LPCC) and perceptual log area ratio (PLAR) in the Fischer voice speaker verification approach. The complementary LPCC and PLAR features have a small effect in long duration conditions because the MFCC feature contains sufficient speaker information, but they are proved to be effective in very short duration conditions in which speaker information is very limited. In [17], the authors tried to increase feature density by creating artificial features by adding Gaussian noise to the utterance in a controlled manner. Experiments showed that artificial feature provides useful information different from the original feature if the added noise is dominant enough to change the spectral details but not dominant to the extent of changing the speaker characteristic [17].

Some analytical studies on speech features proposed using vocal source characteristics rather than vocal tract characteristics. In [18,19], vocal source characteristics, which is expected to contain essential speaker information and be less sensitive to the phonetic content than vocal tract characteristics, proved to be very effective when combined with conventional speech features in very short duration conditions. Later,

incorporating vocal source characteristics has been further experimented with different feature sets and more recent verification approaches, and a range of features is shown effectiveness in improving speaker verification with very short utterances [20–22].

For i-vector speaker verification, which had been state-of-the-art for a long time, Gaussian PLDA is introduced in [23], and improved performance substantially in short duration conditions [24] with its ability to model intra-speaker variability in i-vector space. By further experiments with i-vector speaker verification, the authors of [5] showed optimal performance is reached when the model parameters are estimated from training data consisting of short utterances. In addition to matched duration conditions of training and testing, in [4], the authors applied cosine kernel normalization to i-vector space in order to compensate mismatch conditions further for very short duration conditions. In [11], authors introduced utterance variance compensations such as within-class covariance normalization (WCCN) and source normalized linear discriminant analysis (SN-LDA) to the i-vector + GPLDA framework in order to compensate high intra-speaker variability of very short utterances and extended their work in [25] by introducing source normalized weighted linear discriminant analysis (SN-WLDA). In [26], the authors estimated i-vector model parameters with a min-max strategy rather than a classical maximum-likelihood objective in order to deal with extreme conditions resulting from very short utterance durations. Apart from compensation in the modeling phase, many methods also tried compensation at the scoring level. The majority of those methods utilized different quality metrics that include utterance duration as a factor [27–30].

The rising popularity of deep learning approaches also directed researchers to apply neural networks to conventional speaker verification approaches as a remedy for the very short utterance problem. In [31], authors trained a deep neural network (DNN) that tries to generate a residual vector defined as the difference between long-utterance and short-utterance i-vectors. Then the authors supplied original i-vectors obtained from very short utterances with generated residual vectors in order to improve verification performance [31]. In [32], recently introduced generative adversarial network

(GAN) [33] is proposed to generate reliable i-vectors as they would be extracted from long utterances, from short-utterance i-vectors. In [34], a similar approach to [32] is introduced as authors applied GAN to g-vector speaker embeddings with additional objective functions to have more reliable generated g-vectors thus a better speaker verification performance. In [35], authors proposed a non-linear form of PLDA by formulating PLDA as a variational autoencoder where the latent variables of the PLDA are integrated into an encoder-decoder framework.

Some works focused on insufficient phonetic content or possible phonetic mismatch between enrollment and authentication utterances of very short duration, in particular. In [36], authors used a vowel-like region (vowel, semi-vowel, and diphthong sounds) detector to select the more reliable features in terms of speaker information in order to improve speaker verification in extreme conditions. In [37], authors take phonetic content into consideration in score-normalization in the i-vector speaker verification framework and obtain substantial improvement. In [38], authors hypothesized that the short utterance verification problem is largely due to a mismatched prior distribution of speech in limited data conditions. In long duration scenario, speech distribution tends to match the training conditions, and enrollment/test also matches among themselves. The authors proposed to model speakers in acoustic subregions defined by similar speech unit classes to be able to match distributions in an easier way and improve verification performance in very short duration conditions [38]. A similar approach is applied to a more recent speaker verification method: x-vector speaker embeddings in [39]. The authors trained a different time-delay-neural network for each phoneme class and claim that each of the TDNN deals with a relatively constant input distribution, and they are less affected by phonetic variation or mismatch. Thus, they learn more discriminative information for speaker verification [39].

More recent works with a phonetic focus on speaker verification examined the effect of phonetic information in deep learning based methods in which it is relatively hard to explain factors of learning. A phonetic attention mechanism is proposed for speaker verification with convolutional neural networks in [40]. Usually, frame-level outputs are

pooled by statistical or self-attention mechanism, but authors of [40], applied attention relying on phonetic information stating that phonetic attention not only selects more discriminative frame-level features but also normalizes the phonetic variability in short duration conditions. In [41], authors directly fed the phonetic information vectors from an automatic speaker recognition network to the speaker embedding network in the frame-level part. Authors also experimented with a multi-task learning paradigm with two network sharing layers for speaker recognition and phoneme recognition in order to have more generalized phoneme information embedded in frame-level speaker features [41]. Authors stated that incorporating phonetic information in the speaker embedding network improves verification performance, but their experimental setup does not include very short duration conditions [41].

The authors in [42], with the improved verification performance by introducing phonetic information and intuition that phonetic information should be suppressed since it is an additional source of variation. The authors experimented with a multi-task learning paradigm of speaker recognition and phoneme recognition by performing standard learning with shared layers to promote phonetic information in speaker recognition and adversarial learning with a gradient reversal layer to remove phonetic information from speaker representations [42]. They suggest that phonetic information should be promoted in frame-level features of the network while it should be suppressed at the segment level. The work in [43], performed similar experiments with [42] but focused on very short duration conditions and stated that phonetic information should be removed in frame-level speaker features in order to decrease variation.

The works that aim to improve speaker verification in general, focused on applying advancements in deep learning in order to obtain a better representation of speaker information from an utterance. The majority of those methods, largely due to memory requirements of deep learning, are trained with short duration utterances and try to tackle short-duration verification problem in essence. The first and most influential method was x-vector embeddings [1] which has been the baseline method for many years, where embeddings are obtained from one of the last layers of a deep

neural network trained with 3 seconds speech chunks in order to classify between a large number of speakers. Convolutional Neural Networks (CNNs) are also employed in speaker verification along with metric losses and improve verification performance significantly, including very short duration conditions [44]. Channel attention and squeeze-and-excitation mechanisms also turned out to be very effective in improving speaker verification in general and in very short conditions [45]. The effectiveness of those methods might be informative when trying to solve the problem of speaker verification with very short utterances.

3. SPEAKER VERIFICATION WITH SPEAKER EMBEDDINGS

The task of speaker verification is to verify/reject a speaker's identity by comparing speaker enrollment and authentication representations. In a speaker verification system, firstly, a speaker enrolls in the system, and the identity is verified/rejected each time when the speaker utters an authentication utterance by a speaker similarity analysis between enrollment and authentication.

With the advancement of deep learning in recent years, deep neural network based speaker verification methods are proposed and quickly employed as state-of-the-art for this task. Deep learning methods are usually trained to encode speaker information into fixed-size vector representations that are referred to as speaker embeddings. The verification is performed by the similarity analysis between enrollment embedding and authentication embedding that are obtained from speech signals.

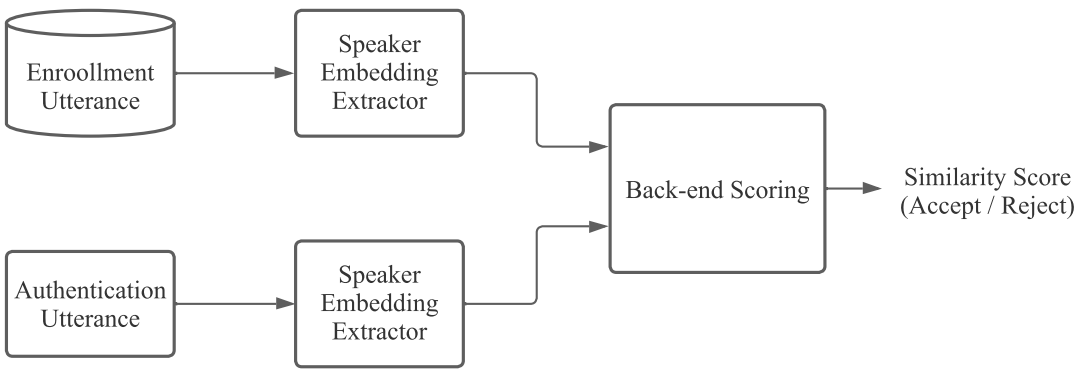

Figure 3.1. Block diagram of a speaker verification system.

In this thesis, we have applied short-duration compensation approaches to speaker verification framework that is based on x-vector + PLDA [1] which uses x-vector as the speaker embedding and PLDA as the back-end scoring module. X-vector + PLDA is one of the state-of-the-art methods for speaker verification.

3.1. X-vector Speaker Embeddings

X-vector speaker embeddings are introduced in [1] and have been state-of-the-art for many years, later becoming a strong baseline for more recent methods. It is based on a deep neural network that is trained to classify between a large number of speakers. After the training phase, output from one of the last layers of the DNN constitutes the fixed-size speaker embedding vector.

3.1.1. Neural Network Structure

The structure of the DNN [46] can be seen in Table 3.1. The neural network takes 24-dimensional Mel Frequency Ceptsral Coefficients(MFCC) obtained from short frames of speech with 25 ms duration by sliding the time-frame by 10 ms over time at each step.

Table 3.1. X-vector neural network architecture.

Layer	**Layer Type**	**Context**	**Size**
1	TDNN-ReLU	t-2;t+2	512
2	Dense-ReLU	t	512
3	TDNN-ReLU	t-2,t,t+2	512
4	Dense-ReLU	t	512
5	TDNN-ReLU	t-3,t,t+3	512
6	Dense-ReLU	t	512
7	TDNN-ReLU	t-4,t,t+4	512
8	Dense-ReLU	t	512
9	Dense-ReLU	t	512
11	Pooling	Full seq	2x1500
12	Dense(Embedding)-ReLU		512
13	Dense-ReLU		512
14	Dense-Softmax		Num. spks.

The layers, up until the pooling operation, act on frame-level. The frame-level layers consist of Time Delay Neural Network layers that act like 1-dimensional CNN layers that stride on time-dimension and dense layers following each TDNN layer. Rectified Linear Unit(ReLU) activation is applied in each layer for nonlinearity, and statistical pooling is applied to combine frame-level activations at the segment-level representation. .Statistical pooling is performed by concatenating the mean and standard deviation of frame-level features. After statistical pooling, the segment-level features are fed into two linear dense layers and an output softmax layer that produces classification outputs. The neural network is trained to classify among N speakers (The number of speakers in the training set). After training, the output from 12^{th} layer is used as the speaker embedding vector.

3.1.2. Data Augmentation

Data augmentation is a simple yet very effective method for creating additional data with more challenging conditions for a neural network to generalize well and have robust speaker representations. In the task of speaker verification with x-vectors, data augmentation is a crucial part of improving the verification performance. In x-vector training, data augmentation is applied to speech signals by incorporating additive noises and reverberation.

Reverberation is added to speech signal by convolving it to various room impulse responses from RIR dataset [47]. Usually, room impulse responses are recordings of impulsive noise in a certain environment [48], but since it is hard to obtain such real room impulse responses, the simulated RIRS that is described in [47], are used in x-vector training.

For the noise addition, noise samples from the MUSAN dataset [49] are used. The MUSAN dataset consists of around 900 types of noise samples, 60 hours of speech data from 12 different languages, and 42 hours of music data from different genres [49].

In total, five types of augmentation are applied to training data.

- Noise: Randomly sampled noise signal is added to speech at one-second intervals over the signal assuring an SNR range of 5-15 dB
- Babble: Speech samples of different speakers from the MUSAN dataset are combined to create babble-noise samples, then added to the original signal assuring an SNR range of 13-20 dB
- Music: Randomly selected music sample is trimmed and added to original speech signal assuring SNR range of 0-15 dB
- Reverb: The original signal is convoluted with randomly sampled artificial room impulse response

3.2. Probabilistic Linear Discriminant Analysis

Probabilistic linear discriminant analysis (PLDA) [2] is a powerful back-end scoring tool for speaker verification methods, and it is also employed in verification with x-vector embeddings. PLDA essentially allows extracting more discriminative and robust latent features by a generative probabilistic model. From the latent features that model inter-class and intra-class variability, a verification hypothesis test can be derived as if the enrollment and authentication utterance belong to the same class or not. The main advantage of PLDA is that it allows building powerful models of unseen classes since it is a probabilistic approach.

PLDA is often used in conjunction with linear discriminant analysis (LDA) which is a dimensionality reduction and feature selection method, and it is also incorporated in the x-vector + PLDA(LDA/PLDA) framework. LDA is used to reduce the dimension of x-vector embeddings from 512 to 200 while selecting the most discriminative elements in the x-vector space. PLDA is applied to LDA-transformed x-vectors for speaker verification.

3.2.1. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a linear transformation that aims to select feature dimensions that maximize between-class variability and minimize within-class variability. For LDA, firstly within-class (S_w) and between class (S_b) covariances are calculated as

$$S_w = \frac{\sum_k \sum_{i \in C_k} (x_i - \tilde{x_k})(x_i - \tilde{x_k})^T}{N} \tag{3.1}$$

$$S_b = \frac{\sum_k \sum_{i \in C_k} (\tilde{x_k} - \tilde{x})(\tilde{x_k} - \tilde{x})^T}{N} \tag{3.2}$$

$$x = \frac{\sum_i x_i}{N} \tag{3.3}$$

$$x_k = \frac{\sum_{i \in C_k} x_i}{n_k}, \tag{3.4}$$

where C_k is utterance set for k^{th} speaker, x_i is the x-vector from i^{th} utterance, $\tilde{x}$ is global mean of x-vectors, $\tilde{x_k}$ is the mean of the x-vectors from utterances belonging to k^{th} speaker. The aim of the LDA is to find a transformation $x \rightarrow W^T x$ such that maximizes the between-class variance with reference to within-class variance. W consists of eigenvectors of the equation $S_b \mathbf{w} = \lambda S_w \mathbf{w}$. By this approach, the between-class variance is maximized with reference to within-class variance.

3.2.2. Generative Probabilistic Linear Discriminant Analysis

In generative PLDA, the probabilistic model can be thought of as a Gaussian Mixture Model. For each example x, shares a latent class variable y with other examples in that class. The probabilistic relation can be expressed as

$$P(x|y) = \mathcal{N}(x|y, \Phi_w), \tag{3.5}$$

where class members share a common within-class covariance matrix Φ_w. Usually, the class variable y is estimated from a finite number of classes according to $P(y) = \sum_k \pi_k \delta(y - \mu_k)$ where μ_k is class-center, and required parameters π_k, μ_k, Φ_w are estimated by maximum-likelihood objective.

The probabilistic approach of PLDA comes in when modeling class variable y also as a continuous random variable. The class probability then expressed as

$$P(y) = \mathcal{N}(y|m, \Phi_b), \tag{3.6}$$

where m is the global mean of the class variable, Φ_w is between class covariance that models the speaker variability. This kind of probabilistic modeling can be expressed with a loading matrix A being $\Phi_w = AA^T$, and a class covariance matrix Σ being $\Phi_w = A\Sigma A^T$. For a given utterance, PLDA can model LDA transformed x-vector y with latent variables such as

$$y = Au + \epsilon_r \tag{3.7}$$

$$u \sim \mathcal{N}(.|\epsilon_r, I) \tag{3.8}$$

$$\epsilon_r \sim \mathcal{N}(.|0, \Sigma). \tag{3.9}$$

In PLDA training, the parameters m, A, Σ are estimated from speaker-labeled data with the expectation maximization (EM) algorithm given the conditional probability expressions.

The probability of two different speaker representations coming from the same speaker is calculated from PLDA parameters (m A, Σ) and used as a similarity score for verification. PLDA score $s(y_e, y_a)$ between LDA transformed enrollment and authentication x-vector, y_e and y_a respectively, can be expressed by derived matrices Q and P by the equation

$$s(y_e, y_a) = y_e^T Q y_e + y_a^T Q y_a + 2 y_e^T P y_a, \tag{3.10}$$

where Q and P are derived from estimated PLDA parameters (A,Σ) as in the equations

$$\mathbf{Q} = \mathbf{\Sigma_{tot}^{-1}} - (\mathbf{\Sigma_{tot}} - \mathbf{\Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac}})^{-\mathbf{1}} \tag{3.11}$$

$$\mathbf{P} = \mathbf{\Sigma_{tot}^{-1}\Sigma_{ac}}(\mathbf{\Sigma_{tot}} - \mathbf{\Sigma_{ac}\Sigma_{tot}^{-1}\Sigma_{ac}})^{-\mathbf{1}} \tag{3.12}$$

$$\mathbf{\Sigma_{ac}} = \mathbf{AA^T} \tag{3.13}$$

$$\mathbf{\Sigma_{tot}} = \mathbf{AA^T} + \mathbf{\Sigma}. \tag{3.14}$$

3.2.3. Neural Probabilistic Linear Discriminant Analysis

Generative PLDA is the standard and most popular form of PLDA, but many variants are introduced upon GPLDA. Neural PLDA is a recently introduced, discriminative variant of GPLDA, modeling the PLDA paradigm as a neural network [3]. The block diagram of Neural PLDA can be seen in Figure 3.2

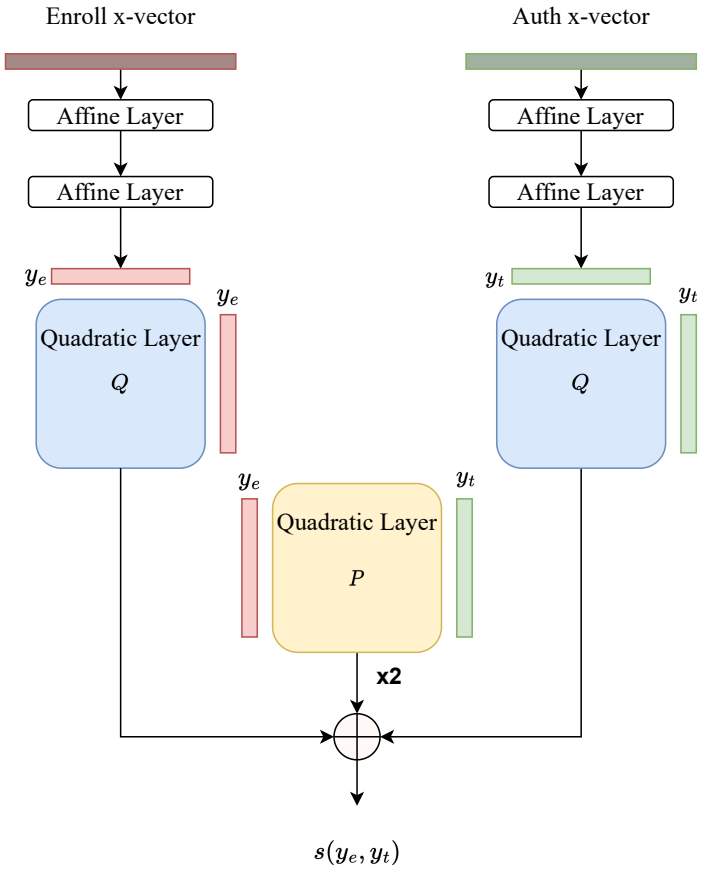

Figure 3.2. Block diagram of neural PLDA.

Neural PLDA (NPLDA) models LDA transformation and length normalization as affine layers, and derived matrices (Q, P) in the Equation 3.10, and directly produces the PLDA score $s(y_e, y_a)$ from enrollment and authentication embedding pair. NPLDA is trained in a supervised, discriminative way by target/non-target embedding pairs and labels.

In training, NPLDA is initialized with Generative PLDA parameters, and it is trained under the objective of Soft Detection Cost Function [3]. Although NPLDA can be trained with binary cross entropy(BCE) loss by target/non-target labels, the authors state that BCE tends to overfit and soft DCF not only helps the model to generalize but also it is a more direct loss for practical scenarios where false alarm errors(FA) are more costly.

Detection cost Function (DCF) is a measure for speaker verification systems to evaluate the system in a proper way for real-life applications. DCF can be expressed as

$$DCF(\beta, \theta) = P_{FR}(\theta) + \beta P_{FA}(\theta) \tag{3.15}$$

$$\beta = \frac{C_{FA}(1 - P_{target})}{C_{FR} P_{nontarget}}, \tag{3.16}$$

where Θ is the threshold value for verification, β is the cost factor for false alarm (FA) errors, C_{FA}, C_{FR} are corresponding costs of false alarm and reject errors, respectively. $P_{target}, P_{non-target}$ are prior probabilities of target/non-target trials. Usually, a larger cost is associated with FA errors, since security is the main concern for speaker verification.

In NPLDA, a soft estimation of DCF is employed in order to have a differentiable loss function for neural network training. DCF is not differentiable because of its discrete nature of $P_{FA}(\Theta)$ and $P_{FR}(\Theta)$. Soft versions of $P_{FA}(\Theta)$ and $P_{FR}(\Theta)$ are derived, such as

$$P_{FA}^{soft} = \frac{\sum_{n=1}^{N} l_n [1 - \sigma(\alpha(s_n - \theta))]}{\sum_{n=1}^{N} l_n} \tag{3.17}$$

$$P_{FR}^{soft} = \frac{\sum_{n=1}^{N} (1 - l_n)[\sigma(\alpha(s_n - \theta))]}{\sum_{n=1}^{N} (1 - l_n)}, \tag{3.18}$$

where l_n is 1 if the sample pair is a target trial and 0 otherwise, s_n is the output score of a sample trial, σ is the sigmoid function, and α is a warping factor. A large value for α enables the approximation of the actual hard versions of P_{FA} and P_{FR} and functions being continuous.

The loss function L becomes

$$L = P_{FR}^{soft}(\theta) + \beta P_{FA}^{soft}(\theta). \tag{3.19}$$

The aim of the NPLDA is to minimize the soft detection cost function, and parameters are directly optimized to have a minimum detection cost function which is one of the most popular evaluation metrics for speaker verification. The final objective of the training can be expressed as

$$minDCF \approx \min_{\theta} L = \min_{\theta} DCF^{soft}(\beta, \theta). \tag{3.20}$$

In this thesis, we have used either GPLDA or NPLDA as back-end scoring module depending on the focus of the methodology, as some introduced methods rely on generative modeling of speaker embeddings while other introduced methods rely on discriminative modeling.

4. LEVERAGING AUXILIARY INFORMATION FROM LARGE SCALE DATA

In this chapter, we present different methods that use auxiliary information from large-scale, closed-set speaker data. Our main motivation for those methods is that information from large-scale data with many speakers and many utterances from each speaker in the dataset can be used to augment the speaker information from a very short utterance. The large-scale data should be a good-enough approximation of the general speaker space. The well-modeled speakers in the closed-set data that are closely related to the actual speaker of the very short utterance will probably contain valuable information about the identity of that speaker. The auxiliary information, such as the most similar speaker model in the closed-set data to the very short utterance, can be used to directly increase the limited speaker information from the very short utterance or guide the compensation process.

In the following sections, a basic approach that merges original embedding from the very short utterance with auxiliary embedding from the closed-set data by weighted addition is introduced first. Later, a more sophisticated approach that utilizes different type of neural networks is introduced with a novel objective function for those networks.

4.1. Enhancing Speaker Embeddings from Very Short Utterances by Weighted Merge with Auxiliary Information

In this approach, we have experimented with the weighted summation of speaker embedding from the closed-set speaker data and the embedding from the very short utterance. Since the embeddings are scattered in Euclidean space in a speaker-aware fashion, the weighted summation of them will result in meaningful representations in terms of speaker information. In other words, a speaker embedding would be guided to a more desirable place in the embedding space, such as the class center of the corresponding speaker, by the help of another speaker embedding for better similarity.

4.1.1. System Overview

In this work, most similar N speaker models from the closed-set speaker data are incorporated as the auxiliary information for the speaker embedding from the very short utterance. The speaker models are obtained by averaging x-vector speaker embeddings from utterances of a given speaker in the dataset. Similarity analysis in order to determine top-N similar speaker models for a given speaker embedding from a very short utterance, performed by PLDA comparison with every speaker model in the closed-set data. Speaker models that have the highest PLDA scores are used as the auxiliary information for the given speaker embedding. The block diagram of the proposed approach can be seen in Figure 4.1.

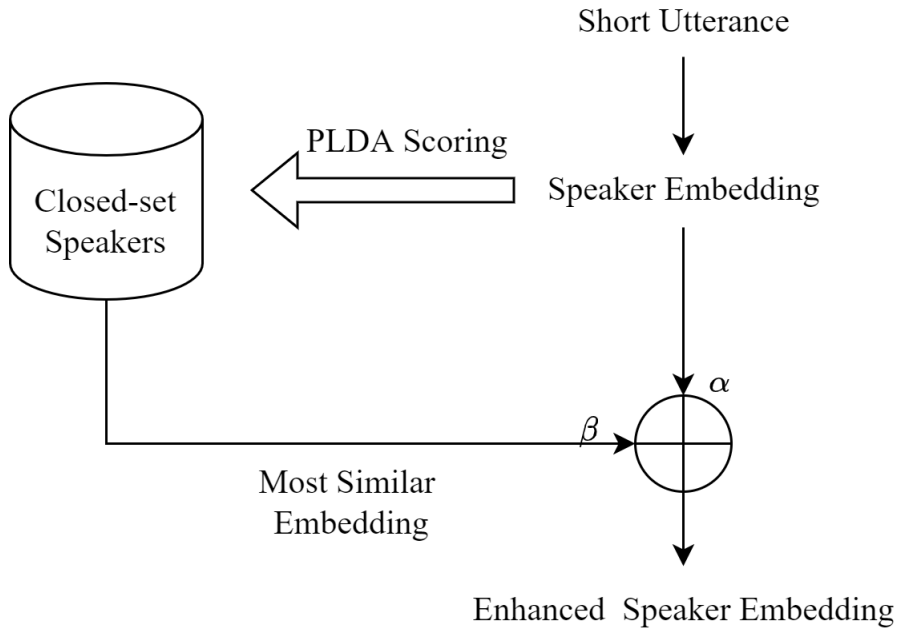

Figure 4.1. Block diagram of weighted merge.

The mathematical expression can be derived for the system such that

$$y_{enhanced} = \alpha y_{orig} + \sum \beta_n y_{aux(n)} \tag{4.1}$$

$$\alpha + \sum \beta_n = 1, \tag{4.2}$$

where y_{orig} is the embedding from the short utterance, $y_{aux(n)}$ are top N speaker model embeddings from the closed-set, α, β_n are experimental weights for merge. In this thesis, enhanced speaker embedding $y_{enhanced}$ is used as enrollment embedding instead

of the original enrollment embedding in our verification experiments in order to improve performance. Due to the computational requirements of comparison with a large set of speakers in the enhancement process, we do not apply the proposed approach to authentication embedding. In the experiments, the x-vector + PLDA framework is used as the verification method. Both for comparing the original very short utterance to the closed-set speaker models and for the verification, the same x-vector + PLDA system is used for the sake of consistency of the proposed approach.

4.1.2. Dataset

We have experimented with two scenarios based on the transmission channel of the speech signals: telephone (8 kHz sampling rate) and microphone (16 kHz sampling rate). For this reason, we utilized two different datasets for speaker embedding model training and verification experiments for both conditions. For the training of x-vector + PLDA for telephone recordings(8 kHz) following datasets are used:

- NIST Speaker Recognition Evaluation(SRE) Data: SRE challenge data from 2004 to 2010 containing around 60k utterances from 4400 speakers
- Switchboard-1 Data (SWBD): SWBD data containing around 2400 utterances from 532 speakers
- voxceleb2 (8kHz): Downsampled version of voxceleb2 dataset [50] containing around 1 million utterances from 5.9k speakers

In the verification experiments for telephony speech, SRE10 evaluation subset of the NIST SRE data is used as the evaluation data. SRE10 eval data consists of around 14K utterances from 440 speakers. The verification trial list is obtained by cross-matching each utterance in the subset. NIST SRE (without SRE10) and SWBD are used as the large-scale, closed-set data for the auxiliary embeddings. NIST SRE and SWBD consist of around 5000 speakers and many utterances for each speaker.

For the experiments with microphone speech, x-vector + PLDA is trained with

the voxceleb2 dataset [50]. Verification trials are created from voxceleb1 dataset [51], which is an earlier version of voxceleb2 and contains 148,642 utterances from 1211 speakers. It is stated that there is no speaker overlap between voxceleb2 and voxceleb1 datasets. In the embedding enhancement process, voxceleb2 is also used as the large-scale, closed-set auxiliary data.

4.1.3. Experimental Setup

In the experiments, verification utterances are truncated to 3 seconds for both enrollment and utterance in order to have a very short duration condition. The verification condition with 10 seconds of enrollment and authentication has also been experimented with in order to see the effect of our method on longer duration conditions. Different number of speaker models N (1,3 and 5) from auxiliary data has been experimented. Weights for merging (α, β_n) are determined experimentally by grid search algorithm for optimum performance. We have measured the equal error rate(EER), which is the error rate at the operation point where false alarm(FA) and false reject (FR) error rates are almost equal.

4.1.4. Results

In Table 4.1, we have compared different enrollment conditions while authentication duration is fixed to 3 seconds for verification trials. The condition with a long duration of enrollment from the original speaker is denoted as full-3s; the condition with a very short duration of enrollment utterance is denoted as 3s-3s; using only auxiliary embedding from closed-set data is denoted as emb(aux)-3s, merging the embedding from very short enrollment with auxiliary embeddings is denoted as emb(orig) + emb(aux).

Table 4.1. Verification results with 3 seconds of authentication.

Method	**Enrollment**	**Enroll Duration**	**EER(%)**
x-vector + PLDA	full	full	9.01
x-vector + PLDA	emb(orig)	3s	17.46
x-vector + PLDA	emb(aux)	-	24.98
x-vector + PLDA	emb(orig) + emb(aux)	3s	16.96
x-vector + PLDA	emb(orig) + Top 3 emb(aux)	3s	16.98
x-vector + PLDA	emb(orig) + Top 5 emb(aux)	3s	17.01

Firstly, performance degradation in very short duration conditions is seen in the results when comparing full-3s and emb(orig)-3s results as EER is increased from 9.01% to 17.46%. Secondly, if we use only the auxiliary embedding from the closed-set, EER rises to 24.98%, as expected since the auxiliary embedding belongs to another speaker. On the other hand, EER is not very high for only emb(aux), meaning that it is still far from 50% EER upper-bound and would indicate that emb(aux), in fact has information about the original speaker's identity. By weighted merging original embedding (emb(orig)) from 3 seconds of enrollment utterance with the auxiliary embedding, we have seen that the EER is reduced to 16.96% from %17.46. These results would indicate that auxiliary embedding has some complementary information to some extent and improves the verification performance. The results of the different number of auxiliary speaker models from the speaker data show that only the most similar model is effective for improving performance as further speaker models have very small weights in the merging and do not improve performance. In Table 4.2, a detailed analysis of the merge weights with only one speaker model is presented.

Table 4.2. Verification results with fifferent merge weights.

Enrollment	**Orig Weight**	**Aux Weight**	**EER(%)**
emb(orig) + emb(aux)	1	0	17.46
emb(orig) + emb(aux)	0.9	0.1	17.24
emb(orig) + emb(aux)	0.8	0.2	17.04
emb(orig) + emb(aux)	0.7	0.3	16.96
emb(orig) + emb(aux)	0.6	0.4	17.03
emb(orig) + emb(aux)	0.5	0.5	17.25
emb(orig) + emb(aux)	0.4	0.6	17.78
emb(orig) + emb(aux)	0.3	0.7	18.60
emb(orig) + emb(aux)	0.2	0.8	19.92
emb(orig) + emb(aux)	0.1	0.9	21.98
emb(orig) + emb(aux)	0	1	24.98

Weight analysis shows that auxiliary embedding has complementary information to some extent and should have small weights. The main source of speaker information is still the original embedding from the very short enrollment utterance, and as the weight of auxiliary embedding, which contains rich information about another speaker, is increased, the performance might be hurt significantly.

In Table 4.3, we have experimented with a duration of 10 seconds both for enrollment and authentication in order to see the effectiveness of our method in longer duration conditions. For all experiments, the authentication duration is set to 10s, and different enrollment conditions are evaluated. The results for 10 seconds of enrollment and authentication the proposed method is ineffective as by merging the original embedding from the enrollment utterance with the auxiliary embedding, the verification EER increases slightly. The benefit of the auxiliary information from closed-set data seems present in only very short duration conditions. However, relatively reasonable EER of using only auxiliary embedding in enrollment may also indicate that there is information about the original speaker in the most similar auxiliary speaker model in a large dataset.

Table 4.3. Verification results with 10 seconds of authentication.

Method	**Enrollment**	**Enroll Duration**	**EER(%)**
x-vector + PLDA	emb(orig)	10s	10.25
x-vector + PLDA	emb(aux)	-	17.96
x-vector + PLDA	emb(orig) + emb(aux)	10s	10.78

In Table 4.4, the results of the experiment in microphone speech data are presented. In microphone speech experiments, in addition to 3 seconds duration conditions, we have experimented with 2 seconds of duration conditions both for authentication and enrollment.

Table 4.4. Verification results in microphone speech.

Method	**Enrollment**	**Enroll-Auth Duration**	**EER(%)**
x-vector + PLDA	emb(orig)	3s-3s	4.95
x-vector + PLDA	emb(orig) + emb(aux)	3s-3s	4.94
x-vector + PLDA	emb(orig)	2s-2s	8.37
x-vector + PLDA	emb(orig) + emb(aux)	2s-2s	8.08

In microphone speech verification, the proposed method improved the verification in a very slight amount in 3 seconds duration conditions. In 2 seconds duration conditions, the improvement is more significant, indicating that the proposed method is effective in different model/data conditions, but its effectiveness varies in the duration conditions.

4.2. Speaker Embedding Mapping with Deep Neural Networks for Very Short Utterances Using Auxiliary Information

In this section, another proposed method to improve speaker verification with very short utterances is introduced. This method tries to generate more robust speaker embeddings from speaker embeddings of very short utterances and auxiliary information. This embedding generation process is performed by a deep neural network which is a

concept that has been developing rapidly for generative modeling with the introduction of advanced structures, loss functions, optimization methods, and many improvements. Recently introduced in [33], Generative Adversarial Network(GAN) structure especially showed promising results in generative tasks, and it is also employed in this method.

Our proposed method is to train a GAN that takes the speaker embedding from the very short utterance as input, and the auxiliary speaker model embedding from the closed-set data as the conditional input. The network then produces a more robust speaker embedding that would be obtained from an utterance with a longer duration. Our main motivation is that it might be possible for a deep neural network to model and compensate the shift of embeddings between certain duration conditions that causes performance degradation. Moreover, that compensation would be helped by the guidance of the auxiliary information from large-scale data that is proven to be helpful in Section 4.1. From another perspective, it also would be possible to extract complementary information from auxiliary speaker model embeddings in a better way.

4.2.1. Generative Adversarial Networks

Generative Adversarial Network is introduced in [33], and it is a neural network structure consisting of two different subnetworks: generator and discriminator. It is designed to model the generative processes. It is trained by an adversarial objective between the generator and discriminator. In GAN training, the generator tries to fool the discriminator by generating examples resembling real examples as much as possible, while the discriminator tries to successfully discriminate between real and generated examples. During the time generator and discriminator try to beat each other by a min-max game over a loss function, they both improve each other until reaching an equilibrium. The diagram of a basic GAN structure can be seen in the Figure 4.2,

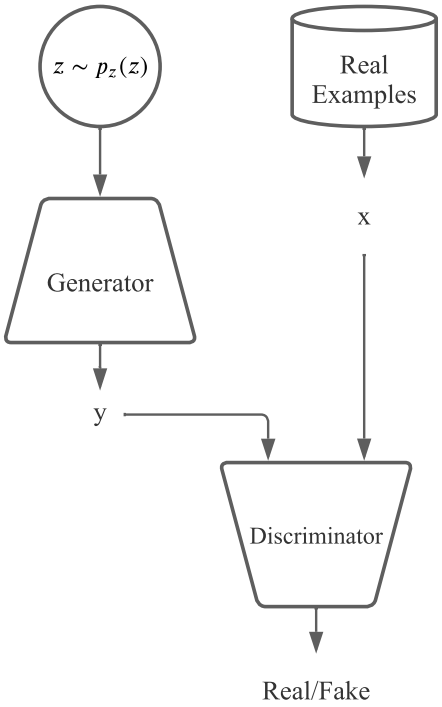

Figure 4.2. Block diagram of a basic GAN structure.

The GAN is usually trained by optimizing standard GAN loss function L_{GAN}, which can be expressed as

$$\min_G \max_D L_{GAN}(G, D) = E_y[logD(y)] + E_z[log(1 - D(G(z)))], \tag{4.3}$$

where G, D are the generator and discriminator, respectively. z is the input to the generator, usually a random vector from a specific prior distribution; y is the real example from the dataset. The loss is updated alternately with respect to the generator and discriminator. The discriminator D tries to minimize the classification objective of real/fake discrimination while the generator maximizes the loss by generating real-like examples. This standard loss for GAN is reported to be hard to optimize in training because of the vanishing gradients. Another loss function is introduced for GAN training that measures the Wasserstein distance between real and fake distributions for smoother gradients [52]. Wasserstein GAN loss uses 1-Lipschitz constrained functions g_w in order to calculate the distance between probability distributions, and the GAN loss becomes

$$WGAN = \max_{g_w \in 1-Lipschitz} E[g_w(y)] - E[g_w(x)]. \tag{4.4}$$

One way to satisfy 1-Lipschitz conditions is having the gradient equal to 1 almost everywhere, and WGAN with gradient penalty is introduced in order to satisfy the constraint by the discriminator function [53]. The loss WGAN-GP can be expressed as

$$WGAN - GP(D, G) = \max_{G,D} E[D(x)] - E[D(G(z))] - \lambda E[(||\nabla D(\hat{x})|| - 1)^2], \quad (4.5)$$

where $\hat{x}$ is the interpolation point between generated examples y and real examples x. WGAN-GP is used in this thesis for the sake of better optimization of GAN models in training.

4.2.2. Proposed System Overview

In this thesis, we have introduced a conditional version of WGAN-GP where the conditional input is the auxiliary speaker embedding from the large-scale data, which is the most similar embedding in the data to the original speaker embedding from the very short utterance.

The input to the generator is the embedding from the very short utterance, and the conditioning is applied by concatenating the auxiliary embedding to the input. In the proposed work, the real examples for the discriminator are embeddings from an utterance of longer duration, and the generator is forced to generate embeddings that would be from long duration utterances. For each long utterance, a very short segment is trimmed, and extracted embedding from that segment is fed to the generator as the input, while embedding from the long utterance is fed to the discriminator as the real example. The block diagram of the proposed GAN structure in training can be seen in Figure 4.3 . The generator consists of linear layers having a dimension of 1024, which is the total dimension of concatenated speaker embeddings with 512 dimensions each. Leaky ReLU non-linearity and batch normalization are applied to all layers of the generator except the output layer. The output layer is a linear layer with 512 dimensions to match the embedding dimension, and a different kind of instance normalization is applied at the output where the $L-2$ norm of output embeddings are set to $\sqrt{512}$ for a proper PLDA scoring.

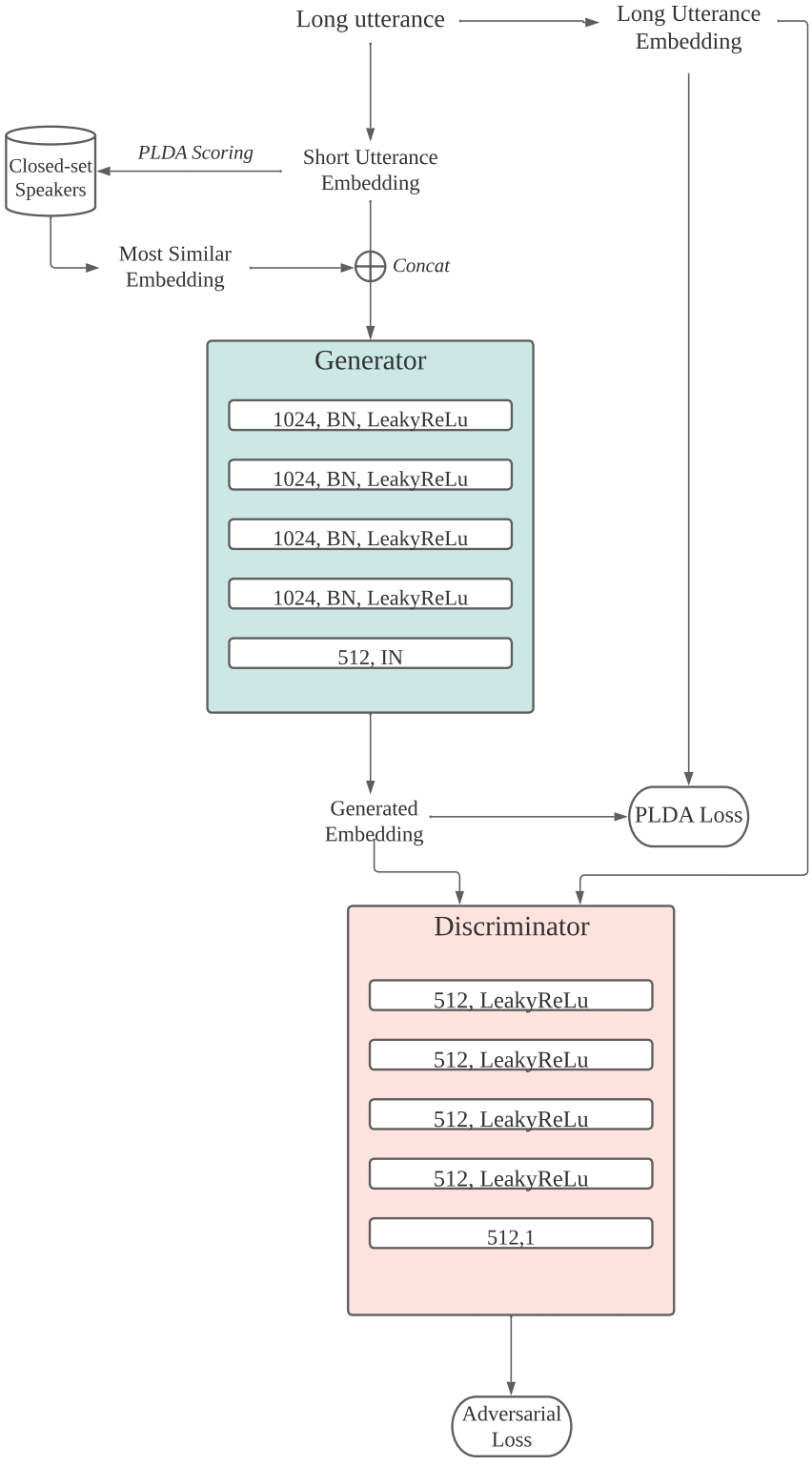

Figure 4.3. Block diagram of proposed GAN training.

The discriminator consists of linear layers having a dimension of 512, which is the speaker embedding dimension. Leaky Relu non-linearity is applied to all layers except the output layer. No activation function is applied at the output layer because the regularization is satisfied by 1-Lipschitz constraint.

The WGAN-GP loss function is modified for conditional training with the auxiliary speaker embeddings. The modified WGAN-GP loss function can be expressed as

$$WGAN - GP(D, G) = \max_{G,D} E[D(x)] - E[D(G(z|u))] - \lambda E[(||\nabla D(\hat{x})|| - 1)^2], \quad (4.6)$$

where x is the long-utterance embedding, z is the short-utterance embedding, and u is the auxiliary speaker model embedding from the closed-set.

Along with the WGAN-GP loss, a speaker discrimination-related cost function is employed in GAN training. In most tasks, the standard GAN loss function would not be sufficient, and additional loss functions are required in order to enforce the model act towards the corresponding task. In this work, a novel triplet PLDA loss is introduced in order to force the generated embeddings to be speaker discriminative.

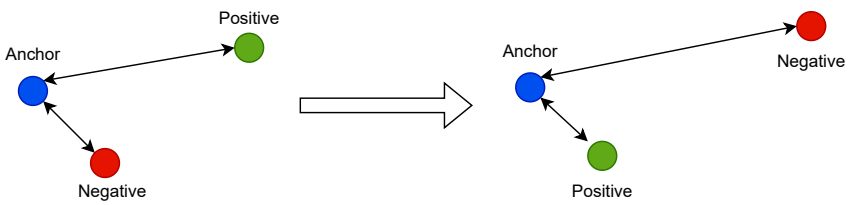

Figure 4.4. Triplet loss objective.

Triplet loss is first introduced in [54], as a metric learning objective. Each example (anchor) is compared to a positive example from the same class, and a negative example from a different class in terms of a specified metric. The objective is to minimize the distance between the anchor-positive pair while maximizing the distance between the anchor-negative pair at the same time. The basic objective of the triplet loss is illustrated in Figure 4.4.

The popular metric for triplet loss is $L-2$ norm, and the triplet loss objective can be expressed as

$$L_{triplet} = \max(||x_a - x_p||_2^2 - ||x_a - x_n||_2^2 + \phi, 0), \tag{4.7}$$

where x_a, x_p, x_n are anchor, positive and negative examples, respectively. ϕ is the desired margin between the positive and negative distances. The loss is updated until the safe distance is obtained by the maximum operator with the 0 value.

In this work, we propose to use PLDA score $s(x_e, x_a)$ in equation 3.10 as the metric between positive and negative examples. Our motivation is to use a more powerful and direct measure for ensuring speaker discrimination objective. The proposed loss function would also allow us to directly integrate the proposed method into the x-vector + PLDA verification framework. PLDA scoring is a linear operation, and with the pre-defined, fixed parameters Q, P it is suitable for backpropagation in neural network training.

We have modified the triplet loss with PLDA score as a metric and introduced triplet PLDA loss that can be expressed as

$$L_{tripletPLDA} = \min(s(G(z|u), x_{target}) - s(G(z|u), x_{nontarget}) - \psi, 0), \tag{4.8}$$

where $G(z|u)$ is the generated vector by the generator, x_{target} is the positive example that is the embedding extracted from the longer duration version of the input utterance. $x_{nontarget}$ is speaker embedding from the long utterance of another speaker, and ψ is the safe margin for positive and negative margins. By maximizing the target PLDA score while minimizing the nontarget PLDA score between embeddings from long utterances, we aim to obtain a robust generated embedding complying with the speaker discrimination ability of long utterance embeddings.

The overall loss function for the GAN training can be expressed as weighted the summation of WGAN-GP loss and Triplet PLDA loss, such as

$$L_{GAN} = WGAN - GP + \beta L_{tripletPLDA} \tag{4.9}$$

where β is the experimental weight for the triplet PLDA loss in training. In addition to the proposed GAN structure, in order to see the effect of adversarial training, we have also experimented with the only generator training with the triplet PLDA loss. The block-diagram of only-generator model can be seen in Figure 4.5. In this method, the same generator structure of the GAN is trained only with the triplet PLDA loss ($L_{tripletPLDA}$) without the discriminator and the WGAN-GP objective.

For evaluating the effect of the auxiliary information, another generator model is trained without conditional input, that is the auxiliary embedding from the dataset. The structure of the generator without condition can be seen in Figure 4.6. The generator, without condition, only uses embeddings from short utterance embeddings and tries to generate long utterance embeddings under the triplet PLDA objective which can be expressed as

$$L_{tripletPLDA} = \min(s(G(z), x_{target}) - s(G(z), x_{nontarget}) - \psi, 0). \tag{4.10}$$

4.2.3. Experimental Setup

In this work, the x-vector and PLDA model is trained with the voxceleb2 dataset that is described in Subsection 4.1.2. For the GAN training, utterances from voxceleb2, which have a duration longer than 10 seconds, are selected for training. Selected utterances are truncated to 10s randomly, and for short utterance examples, truncated utterances are further trimmed to 2 seconds. In the GAN training, semi-hard sampling [54] is applied over the batch of examples for selecting a negative example for the anchor in the triplet PLDA loss.

For the evaluation, voxceleb1 dataset that is described in Subsection 4.1.2 is used. All utterances are truncated to 2s in order to create a very short verification scenario.

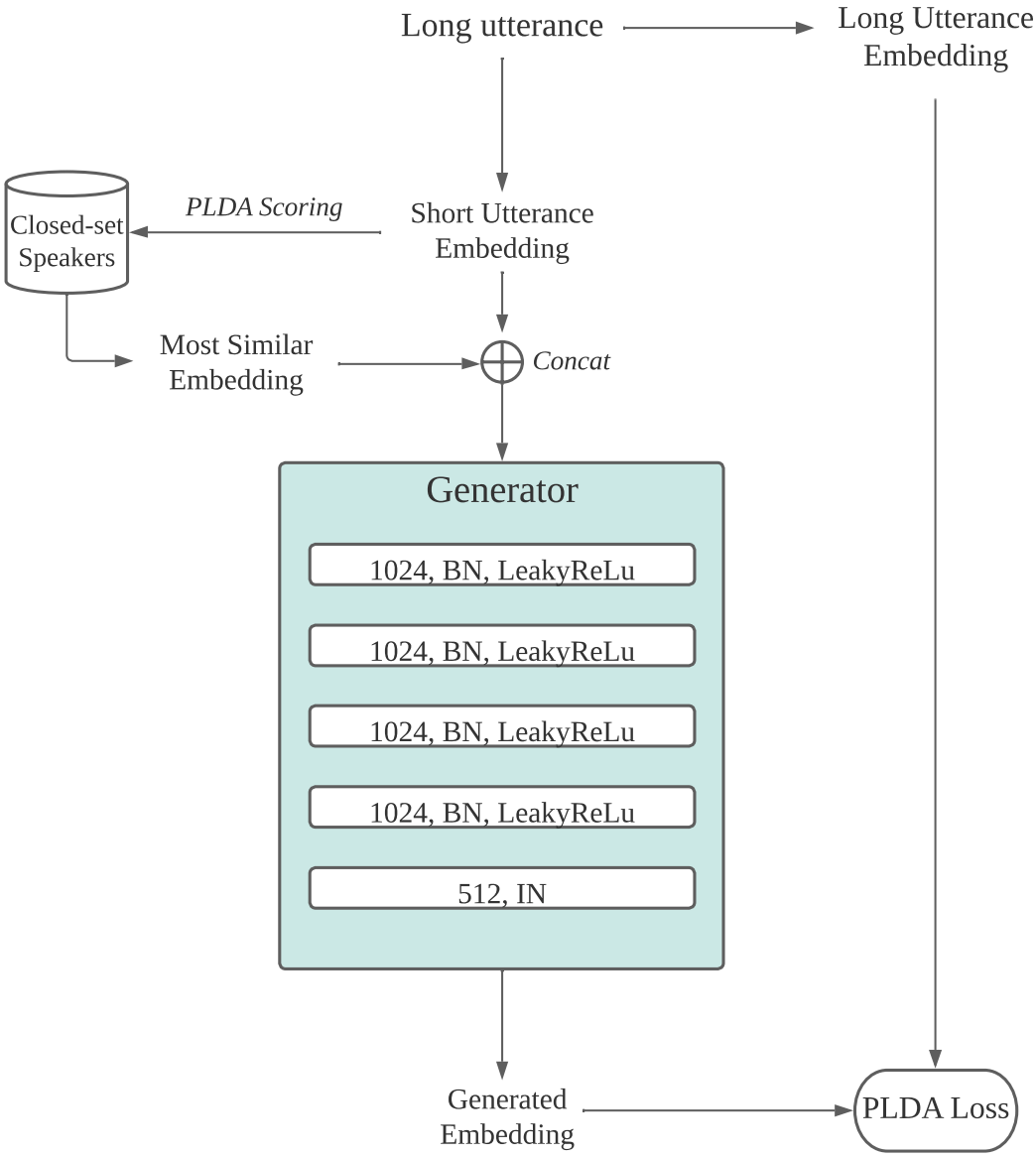

Figure 4.5. Proposed only generator training.

The discriminator is discarded in the test case for the GAN, and generated embeddings by using the embedding from the 2 seconds of utterance, and the auxiliary embedding from the closed-set data is used in the verification. The verification experiments are also performed with the only generator model and the only generator model without condition in order to see the effect of the adversarial training and auxiliary information separately. Equal error rate (EER) is measured as the performance metric.

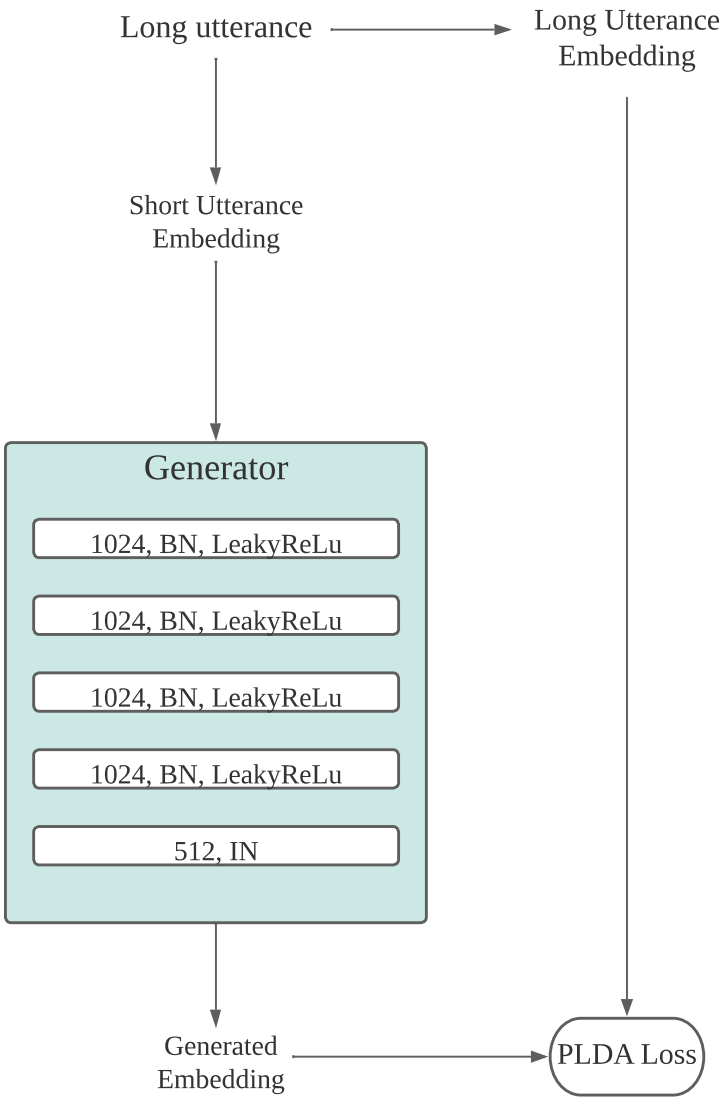

Figure 4.6. Proposed only generator training without auxiliary information.

The proposed method is applied to only enrollment embedding in the standard scenario. Additionally, a scenario in which the proposed method is applied to both enrollment and authentication embeddings are also experimented. Lastly, the fusion of original embedding from the short utterance and the generated embedding is applied to all scenarios.

4.2.4. Results

In Table 4.5 the verification EER's are compared. The baseline is standard verification with x-vector + PLDA, WGAN-GP is the proposed GAN model that tries to generate a more robust embedding from the enrollment x-vector embedding. Only G is the generator without the adversarial training, and Only G w/o condition is the generator without both adversarial training and auxiliary input. x-vector(2s) and x-vector(10s) correspond to embeddings extracted from 2 seconds and 10 seconds of audio, respectively. The result of verification with both generated embeddings at the enrollment and authentication sides is also presented.

Table 4.5. Verification results for generated enrollment embeddings.

			EER(%)	
Model	**Input**	**Condition**	**enroll gen**	**enroll-auth gen**
Baseline	-	-	*8.37*	
WGAN-GP	x-vector(2s)	x-vector(aux)	8.12	7.89
Only G	x-vector(2s)	x-vector(aux)	8.16	7.84
Only G(w/o cond)	x-vector(2s)	-	**8.05**	**7.67**

The results of the experiments where the generated embeddings are fused with the original embedding can be seen in Table 4.6. In the table, gen + orig refers to fusion of the generated and original embedding, while orig refer to the original embedding from the very short utterance.

The results in Table 4.5 suggest that the proposed GAN improves speaker verification performance slightly with generated embedding as the enrollment. However, the verification result of only generator training without the discriminator and adversarial loss suggests that there is no evident effect of the GAN training since the results are really similar. Also, training the generator without the conditional input, that is the auxiliary embedding from the closed-set data performed best, indicating that conditioning with the auxiliary information is not effective and even hurts the performance,

in this case.

Table 4.6. Verification results of fusion with original embeddings.

Model	**Enrollment**	**Authentication**	**EER**
Baseline	*orig*	*orig*	*8.37*
WGAN-GP	gen + orig	orig	7.64
WGAN-GP	gen + orig	gen + orig	7.25
Only G	gen + orig	orig	7.58
Only G	gen + orig	gen + orig	7.14
Only G(w/o cond)	gen + orig	orig	7.54
Only G(w/o cond)	gen + orig	gen + orig	**7.08**

The results in Table 4.6, show significant improvement over the baseline by fusing the generated embedding and original embedding in all cases. This would indicate that the generation process introduces complementary information to generated embeddings while losing some of the essential speaker information encoded in the original embedding. Compensating this loss by the fusion with the original embedding result in a robust speaker embedding that improves the verification performance noticeably. Only generator network without the auxiliary conditioning input is performed best overall.

In conclusion, we have shown that adversarial learning has no significant effect in our experiments. Conditioning the neural network with the auxiliary information also seemed ineffective in our case. It is suspected that the high-dimensionality of the conditional input makes the neural network task really complex. In addition, the auxiliary information that is the most-similar speaker model embedding to the original embedding; would not contain too much complementary information about the identity of the original speaker. Most importantly, it seems possible to generate the complementary information to some extent without the auxiliary information, as the only generator model without the condition performed really well.

5. EFFECT OF PHONETIC MISMATCH IN VERIFICATION

In this chapter, the analysis of the effect of phonetic mismatch on speaker verification with very short utterances is presented. As a compensation method, a back-end scoring module relying on deep neural networks is introduced.

The phonetic mismatch between enrollment and authentication utterance is usually seen as one of the main potential sources for the performance degradation in text-independent speaker verification with very short utterances. Especially in target trials, the dissimilarity of the features because of the difference in phonetic content makes the verification problem harder for a given verification method which has to produce high similarity scores for target verification trials.

In this work, an analysis of the correlation between the verification scores and phonetic mismatch is performed. A back-end scoring module for x-vector embeddings that takes the phonetic distribution of an utterance into consideration is experimented with to improve speaker verification with very short utterances.

5.1. Verification Score Analysis in Terms Of Phonetic Mismatch

In order to verify and analyze the effect of the phonetic mismatch between enrollment and authentication on the verification similarity score, a verification experiment was designed. In this experiment, the phonetic distribution of each utterance is defined by the occurence vector of phones in the utterance using phone labels. For each verification trial, the phonetic mismatch is measured as the distance between the phonetic distribution vectors of enrollment and authentication utterance.

The defined phonetic distribution vector of an utterance for a given phone set $\{p_1, p_2, .., p_C\}$ can be expressed as

$$g = [\frac{N_{p1}}{N}, \frac{N_{p2}}{N}, ..., \frac{N_p}{N}], \tag{5.1}$$

where C is the number of phones in the set, N_{pi} is the number of the occurrence of the specific phone p_i in the utterance, and N is the total number of phones in the utterance.

Additionally, the phonetic overlap is also measured as the number of matched, non-zero dimensions of the phonetic distribution vectors. The Pearson correlation coeffient is calculated using the verification scores and corresponding phonetic mismatch/overlap values of verification trials. As the verification method, the x-vector + PLDA model that is trained with the voxceleb2 dataset is used in the experiments.

5.1.1. Experimental Setup

In the experiments, two different datasets with phone labels are used. Firstly, TIMIT dataset [55] is used in the verification experiments. TIMIT dataset was originally designed for speech recognition task with utterances and their detailed transcriptions, including phone boundaries. The phone set $\{p_1, p_2, .., p_52\}$ consists of 52 distinct phones. The speaker labels are also present, and there are a total of 630 speakers in the dataset, and each speaker utters ten different, phonetically rich sentences among a set of 2400 different sentences. The official test set consists of utterances from 168 speakers. For the verification experiment, utterances from the official test set are split into 2 seconds segments in accordance with their detailed transcriptions. Verification trials are created from 2 seconds segments.

Secondly, a verification experiment is performed on LibriSpeech corpus [56]. Librispeech corpus is also designed originally for speech recognition, and speaker labels are also present. The phone set $\{p_1, p_2, .., p_4*\}$ consists of 40 distinct phones. The dataset consists of utterances collected from 1278 speakers who are prompted to read different passages from various books that are present in the LibriVox audiobooks

project. The test set contains utterances from 146 speakers. In this experiment, the same preprocessing is applied to the test set of Librispeech as test utterances are split into 2-second segments. The verification trials are created from the 2 seconds segments.

5.1.2. Results

The Pearson correlation values can be seen in Table 5.1 . Different correlation coefficients are calculated for target and non-target trials separately.

Table 5.1. Correlation values between verification scores and phonetic mismatch.

	Score vs Phonetic Distance		**Score vs Phonetic Overlap**	
Data	**target**	**non-target**	**target**	**non-target**
TIMIT(2s)	-0.22	-0.08	0.25	0.08
LibriSpeech(2s)	-0.249	0.027	0.235	-0.03

Results show that there is indeed a tendency for target verification scores to decrease with increasing phonetic distance. In parallel, target verification scores tend to increase with the increasing phonetic overlap. The correlation values are not very high would indicate that not only the phonetic mismatch is one of the many factors, but also crude analysis would not be sufficient to depict the effect accurately. A very small correlation factor is observed for non-target trials, as expected. The main factor for increasing non-target verification would be the actual resemblance of voices between two different speakers.

5.2. Back-end Scoring with Phonetic Distance Information

In order to compensate for the performance degradation that is related to phonetic mismatch, a neural network based back-end scoring module for x-vector embeddings is proposed. The scoring module would be aware of the phonetic mismatch by the introduction of the phonetic distribution information as the additional source of information. The main motivation is that it would be possible to condition the scoring network to

similar/dissimilar phonetic content along with the target/non-target embedding pairs. The network would make up for the target score decrease when a phonetic mismatch occurs.

5.2.1. System Overview

A feed-forward neural network is designed as the back-end scoring module. The network takes the concatenated enrollment and authentication embeddings and phonetic distribution vectors that are described in Section 5.1 . The network is trained with target/non-target trials under the binary classification objective, and the output target probability is used as the verification score. The block diagram of the proposed system can be seen in the Figure 5.1.

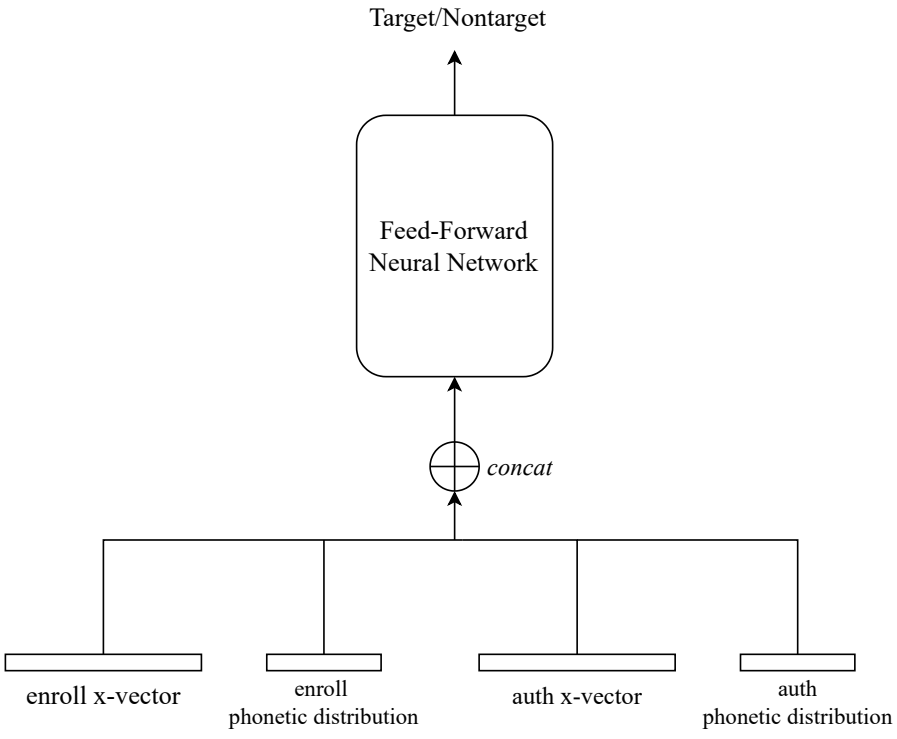

Figure 5.1. Proposed phonetic mismatch aware back-end scoring module.

The feed-forward neural network consists of 3 linear layers with dimensions of 1104,512,512, respectively. The leaky relu non-linearity is applied in the hidden layers, the sigmoid activation function is applied in the output layer for binary classification.

The phonetic distribution vectors are one-hot representations of present phones in the phone set $\{p_1, p_2, .., p_C\}$ for the given dataset. The network is trained in a supervised way under the objective of binary cross entropy (BCE) loss.

5.2.2. Experimental Setup

In the experiments, the x-vector model trained with the voxceleb2 dataset is used as the embedding extractor. As the training data, training subset of the TIMIT dataset that consists of utterances from around 600 speakers is used. The training utterances are split into 2 seconds segments in accordance with their phonetic transcripts, and verification trials are created from those segments for training. In order to see the effectiveness of the phonetic distribution vectors as additional input to the model, a standard model that takes only the concatenation enrollment and authentication embeddings is also trained with the same training data.

The same test set created from TIMIT corpus that is described in Subsection 5.1.1 is used in the evaluation as the standard evaluation scenario. In addition, a more challenging evaluation condition is created by sampling the top %20 target trials with the lowest phonetic overlap from the standard test set. Equal error rate(EER) is measured as the performance evaluation metric.

5.2.3. Results

The results of the verification experiments can be seen in Table 5.2. The results of the standard x-vector + PLDA framework is also included as the baseline. The standard denotes the standard evaluation scenario and hard denotes for the scenario with the sampled target trials that have lowest phonetic overlap. In the results, neural network based back-end performed very similarly to the standard i-vector + PLDA framework. Using phone vectors seems ineffective in our experiments, as the verification performance is slightly worse than in the case where only enrollment and authentication embeddings are used.

Table 5.2. Verification results for back-end scoring module with phonetic conditioning.

		EER(%)	
Model	**Input**	**Standard**	**Hard**
x-vector + PLDA	-	3.41	4.77
x-vector + FFNN	emb(enroll) + emb(auth)	**3.41**	**3.63**
x-vector + FFNN	emb(enroll) + emb(auth) + phone vectors	3.42	3.74

The main reason for this ineffectiveness would be the complexity of modeling the verification in relation to phonetic distributions for each phone in the set. Using discrete representations for phones that do not intrinsically model the acoustic properties of phones would make the task more challenging, as in our case. The most important result would be the performance of the proposed back-end module without the phonetic vectors in the hard test set. The performance of the standard x-vector+PLDA degrades noticeably in the hard test conditions compared to the standard set, but the proposed method is less affected by the more challenging conditions. This would be due to the embedding modeling capacity of the proposed method, as it is expected to model more complex distributions thanks to its non-linearity and supervised training in the target domain. It is expected that embedding from very short utterances distributed in a more complex way in the classification space and more complex modeling in the back-end scoring would be helpful to improve performance. A domain-specific modeling would also help as the embeddings from very short utterances to model the complex distribution of embeddings.

6. OTHER CONTRIBUTIONS

In this chapter, related contributions to speaker verification in general, are presented. The triplet PLDA loss proposed in Subsection 4.1.2, would be applicable as the standard objective for discriminative PLDA training and it is applied to Neural PLDA as the objective function. Additionally, a novel unsupervised domain adaptation method for PLDA is proposed. Although the unsupervised method is proposed for general conditions, it would be a potential remedy for performance degradation in verification with very short utterances.

6.1. Triplet Loss for Neural PLDA

The neural PLDA that is described in Subsection 3.2.3, is a discriminative, neural network based PLDA variant for speaker embeddings as a back-end scoring module. The authors of [3], proposed the soft detection cost function described in Subsection 3.2.3, as the main objective for training Neural PLDA.

The triplet PLDA loss proposed in Subsection 4.1.2, would be employed as the objective function in Neural PLDA training. triplet PLDA loss proved to be effective for ensuring speaker discriminability and would perform better than the soft detection cost(SDC) loss because of its focus on hard examples. The hard examples constitute the majority of errors, and the general focus of SDC would result in a sub-optimal point for the PLDA model.

6.1.1. Proposed Method

The triplet PLDA loss is described in Equation 4.10. In that work, the pre-computed PLDA parameters is used to obtain a similarity score, and only the generator parameters are updated in training while the PLDA parameters are fixed.

In this work, we propose to directly integrate triplet PLDA loss into Neural PLDA. The PLDA parameters are updated during training in order to have an optimal PLDA for verification. The mathematical expression of the triplet PLDA can be re-written as

$$L_{tripletPLDA} = \min(s(x_{anchor}, x_{positive}) - s(x_{anchor}, x_{negative}) - \psi, 0), \tag{6.1}$$

where $x_{anchor}, x_{positive}, x_{negative}$ are speaker embeddings from anchor, positive and negative examples, respectively. $s(x_{anchor}, x_{positive})$ is the PLDA score between the anchor and positive embedding pair, $s(x_{anchor}, x_{positive})$ is the PLDA score between anchor and negative embedding pair, and ψ it he safe-margin between positive and negative PLDA scores.

The proposed approach for sampling positive and negative examples for a given anchor is called semi-hard mining in [54]. Semi-hard mining samples positive examples randomly and samples the hardest negative under the constraint that the negative example would be farther to anchor than the positive example. If such a negative does not exist, the negative example with the maximum distance to the anchor is sampled.

Another approach for sampling positive and negative examples would be negative hard-mining. In negative hard-mining, the positives would be chosen randomly, while the negative example would be sampled as the closest negative to the anchor. The authors of [54] state that negative hard-mining results in poor initialization, but in this work, since the neural PLDA is initialized with GPLDA, it is less of a concern. Thus we have experimented with semi-hard mining and negative hard-mining separately in this work.

6.1.2. Experimental Setup

As the speaker embedding network, the x-vector model trained with voxceleb2 dataset is used. For training Neural PLDA, voxceleb2 dataset is also employed. Batches having the dimension of 1024 are formed from utterances of voxceleb2, and negative and positive example sampling is performed online. Each utterance in the batch is taken as the anchor; the positive examples are sampled randomly from the global training set while the negative examples are sampled within the batch for each anchor for the sake of computational speed.

The official voxceleb1-e test set is used for verification evaluation. Equal error rate(EER) and minimum detection cost function (DCF) are measured as the evaluation metric. In the verification experiments, two scenarios are created in terms of utterance duration: full-full and 2s-2s.

In the full-full scenario, no truncation is applied to utterances of voxceleb1-e and voxceleb2, which have a median value of 7 seconds for the duration. In 2s-2s scenario, all utterances in the voxceleb1-e voxceleb2 data are truncated to 2 seconds in order to create very short duration conditions. In 2s-2s scenario, training is also performed with utterances of voxceleb2 data that are truncated to 2 seconds in order to create matched conditions between training and testing.

6.1.3. Results

Results for full-full scenario are presented in Table 6.1. GPLDA has also experimented with as a baseline in the verification experiments. Verification results of semi-hard mining and hard negative mining are presented separately.

Results firstly validate the findings in [3], as Neural PLDA with the soft detection cost outperformed standard Gaussian PLDA. Secondly, the triplet PLDA loss with semi-hard mining performed poorly as it produced the worst results. Semi-hard mining

is suspected of performing poorly because it never samples hard negative examples, and the PLDA model with semi-hard samples would be sup-optimal. The results of the triplet PLDA with hard negative sampling outperformed both baseline GPLDA and proposed soft detection cost indicating that focusing on the hard examples would result in a more optimal PLDA model.

Table 6.1. Verification results of different PLDA methods.

		voxceleb1-e (full-full)	
PLDA Method	**Loss**	**EER(%)**	**minDCF**
GPLDA	-	2.61	0.149
NPLDA	Soft Detection Cost	2.31	0.144
NPLDA	Triplet PLDA (Semi-Hard)	2.79	0.179
NPLDA	Triplet PLDA (Hard)	**2.11**	**0.127**

The results for 2s-2s scenario are presented in Table 6.2. In this scenario, triplet PLDA loss has experimented with only negative hard-mining.

Table 6.2. Verification results of different PLDA methods (2s-2s).

		voxceleb1-e (2s-2s)	
PLDA Method	**Loss**	**EER(%)**	**minDCF**
GPLDA	-	7.66	0.494
NPLDA	Soft Detection Cost	7.75	0.507
NPLDA	Triplet PLDA (Hard)	**7.49**	**0.458**

It is seen in the results that NPLDA with soft detection cost is performed poorly in very short duration conditions. NPLDA with the triplet loss outperformed both baseline methods slightly in EER and significantly in minDCF metric. This result would suggest that triplet loss focuses on compensating false alarm errors as the cost of a false alarm is higher in the minDCF metric. The hard mining for only negative examples would be the source of that limited improvement in equal error rate, and positive samples would be sampled in a more sophisticated manner in future work.

In conclusion, the proposed method would be applied to the very short scenario to improve verification performance.

6.2. Unsupervised Domain Adaptation of Neural PLDA

In this section, a novel unsupervised adaptation for PLDA for the target domain is proposed. PLDA parameters are estimated from labeled data, and a domain shift would degrade verification performance. In view of the high cost of obtaining a labeled dataset, we propose an unsupervised method for adapting PLDA to the target domain. Our proposed method relies on creating pseudo target/non-target labels directly for a discriminative PLDA method. The proposed method is applied to Neural PLDA, which is a discriminative variant of standard generative PLDA.

6.2.1. Proposed Method

The proposed adaptation method creates target examples for discriminative adaptation by sampling non-overlapping, random chunks from a single utterance. Non-target trials for adaptation are created by sampling random chunks from different utterances in the dataset. This kind of sampling approach would be expected to create valid target/non-target examples for Neural PLDA adaptation under the two strong assumptions about speech dataset:

- First Assumption: For target examples, two random chunks from a single utterance will most probably belong to the same speaker if there is most likely a single speaker in an utterance.
- Second Assumption: For non-target examples, random segments sampled from different utterances will most likely belong to different speakers if the number of speakers and utterances in an unlabeled dataset is reasonable.

The out-of-domain(OOD) Neural PLDA that is trained with labeled data from the different domain can be adapted to the target domain by training examples created by the proposed method from unlabeled target domain data. The block diagram of the proposed unsupervised adaptation can be seen in the Figure 6.2.

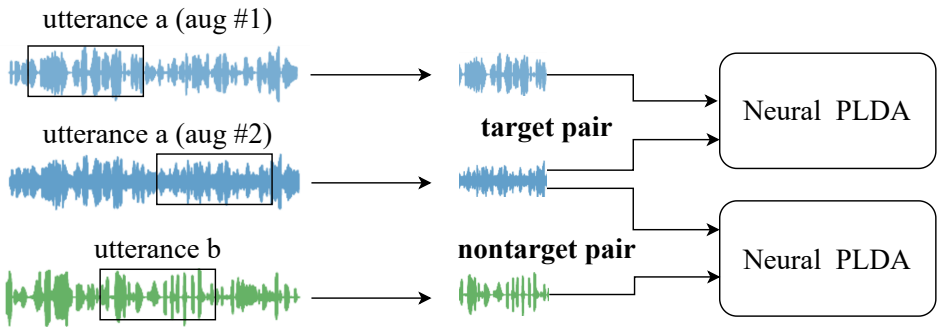

Figure 6.1. Proposed sampling approach.

By using the proposed method, two chunks will come from the same utterance for target trials. Thus, their nuisance factors will be similar. Thus, the speaker verification system's robustness against nuisance variability may be degraded. We used augmented versions of a single utterance for training in order to eliminate the insufficient modeling of nuisance variability. In addition to the original utterance, four different versions of an utterance are obtained: babble, music, non-speech noise, and reverberation added versions similar to the augmentation proposed in [1], in order to create pairs that have different nuisance conditions.

6.2.2. Experimental Setup

X-vector is used as the speaker embedding network in this work. X-vector model is trained with SRE04-10 data that is described in Subsection 4.1.2. This dataset which is determined as the labeled out-of-domain data mostly consists of English speech. As the target domain, Tunisian Arabic speech is determined in the experiments using a part of SRE18 evaluation data. The official evaluation set of SRE18 is used in the verification performance evaluation. The unlabeled part of SRE18 is used in unsupervised domain adaptation of the OOD Neural PLDA that is trained with SRE04-10 data. The duration of chunks in the adaptation process is fixed as 10 seconds. The diagram

of the proposed unsupervised adaptation can be seen in the Figure 6.2.

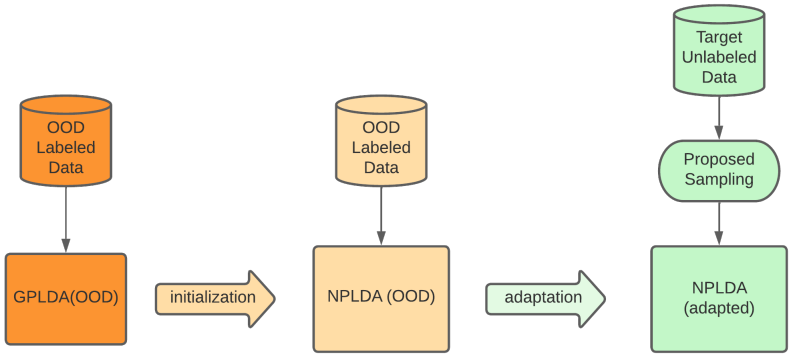

Figure 6.2. Proposed unsupervised adaptation of PLDA.

Along with the standard evaluation condition of SRE18, denoted as full-full, which consists of long telephone conversations, a more challenging test condition is created by truncating the original utterances to 10 seconds. This condition is denoted as 10s-10s. The out-of-domain GPLDA, out-of-domain Neural PLDA, and adapted out-of-domain Neural PLDA is compared in the test conditions. A statistical unsupervised adaptation approach that operates on GPLDA, which is included in kaldi toolkit [57] has also been experimented as a reference method.

6.2.3. Results

The results of speaker verification in the target domain are presented in Table 6.3. As a first step, discriminative OOD neural PLDA outperforms OOD GPLDA in almost all scenarios, which is consistent with the results found in [3]. Further, the proposed approach shows significant improvements when applied to OOD NPLDA. As compared to Kaldi's GPLDA adaptation algorithm, our proposed method performs better under a 10s-10s scenario but not under the full-full scenario. Since the proposed method cannot utilize full-duration utterances, it has been speculated to perform slightly worse in full-full scenarios. In contrast, Kaldi GPLDA adaptation is speculated to suffer from poor statistics when dealing with short utterances, whereas our proposed method has better performance. Besides improving the OOD PLDA on the target domain, comparison with a baseline adaptation approach shows the effectiveness of the proposed method

overall.

Table 6.3. Verification performance on SRE18 eval set.

	full-full		10s-10s	
Method	EER(%)	minDCF	EER(%)	minDCF
GPLDA (OOD)	9.09	0.433	16.23	0.705
NPLDA (OOD)	8.73	0.412	16.34	0.697
GPLDA adapt (kaldi)	**7.86**	**0.371**	16.43	0.685
NPLDA adapt (proposed sampling)	7.95	0.381	**14.54**	**0.659**

7. CONCLUSIONS

In this thesis, different methods are experimented with for improving the speaker verification performance in very short duration conditions. In addition to those methods, novel contributions for speaker verification in general conditions are also presented as they are also applicable to very short duration case. The proposed methods are applied to one of the most popular speaker verification methods, which is called x-vector speaker embeddings.

Firstly, methods that leverage the auxiliary information has experimented with in order to compensate for the insufficiency of speaker information in utterances with very short duration. The auxiliary information is defined as the most similar speaker model from large-scale, closed-set data. As the most basic approach, the auxiliary information is directly merged with the present speaker embedding from a very short utterance by weighted summation. The proposed basic approach improved the verification performance to some extent, indicating that there would be useful information in the auxiliary speaker model. A more sophisticated approach that utilizes generative adversarial networks has also been experimented with in order to generate a more robust speaker representation using the embedding from the very short utterance and the auxiliary information. In our experiments, both the adversarial training and the use of the auxiliary information seemed ineffective in improving the verification performance. However, we showed that it is possible to generate a more robust speaker representation from x-vector embedding from the very short utterance by a generative neural network.

Secondly, the effect of the difference between the phonetic content of very short utterances on the verification performance is analyzed. Analysis showed some correlation between the target verification score and the phonetic distribution distance. A neural network based back-end scoring module is experimented with to compensate for the decrease in verification score caused by the phonetic mismatch. The neural network

did not show significant gain in the standard test scenario. At the same time, the proposed back-end module proved its robustness to phonetic mismatch specifically as the performance was not affected by extreme mismatch conditions created by sampling a subset from the standard test data. The back-end scoring module is intended to be phonetic mismatch aware with the help of phonetic distribution vector as input. The task turned out to be too complex for the designed neural network, and the introduction of phonetic information to the back-end module did not improve the performance. In conclusion, the robustness of the proposed method to extreme conditions would be a key insight into the very short utterance duration problem in speaker verification. A more specialized scoring method for speaker embeddings has the potential to be very effective for very short utterances.

A novel loss function for PLDA is also presented in this thesis. The proposed triplet PLDA loss function is applied to Neural PLDA framework and shows superiority over standard loss functions designed for the Neural PLDA. The verification performance is improved on not only general duration conditions but also very short duration case.

Lastly, a novel unsupervised domain adaptation approach for PLDA is presented. The proposed adaptation method proved to be effective in compensating for the performance degradation that happens when a domain shift occurs between training and testing conditions. The proposed adaptation method would also be employed as a solution to speaker verification with very short utterances. The previous findings suggest the robustness of the specialized back-end scoring modules and unsupervised adaptation would be a feasible way to specialize the back-end scoring module with unlabeled data.

Future work for this thesis would focus on making speaker representations more robust to very short duration conditions. Designing a phonetic content invariant speaker representation network would be effective for improving the performance.

REFERENCES

1. Snyder, D., D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-Vectors: Robust DNN Embeddings for Speaker Recognition", *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, AB, pp. 5329–5333, 2018.

2. Ioffe, S., "Probabilistic linear discriminant analysis", *European Conference on Computer Vision*, Graz, Austria, pp. 531–542, 2006.

3. Ramoji, S., P. Krishnan and S. Ganapathy, "NPLDA: A Deep Neural PLDA Model for Speaker Verification", arXiv:2002.03562[eess.AS], 2020.

4. Mandasari, M. I., M. McLaren and D. A. van Leeuwen, "Evaluation of I-vector Speaker Recognition Systems for Forensic Application", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2011*, Florence, Italy, pp. 21–24, 2011.

5. Sarkar, A. K., D. Matrouf, P. M. Bousquet and J.-F. Bonastre, "Study of the Effect of I-vector Modeling on Short and Mismatch Utterance Duration for Speaker Verification", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2012*, Portland, OR, pp. 2662–2665.

6. Sordo Martinez, P. L., B. Fauve, A. Larcher and J. S. D. Mason, "Speaker Verification Performance with Constrained Durations", *2nd International Workshop on Biometrics and Forensics*, Valletta, Malta, pp. 1–6, 2014.

7. Poddar, A., M. Sahidullah and G. Saha, "Performance Comparison of Speaker Recognition Systems in Presence of Duration Variability", *Annual IEEE India Conference (INDICON)*, New Delhi, India, pp. 1–6, 2015.

8. Kanagasundaram, A., S. Sridharan, G. Sriram, S. Prachi and C. Fookes, "A Study of X-vector Based Speaker Recognition on Short Utterances", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2019*, Graz, Austria, pp. 2943–2947, 2019.

9. Poddar, A., M. Sahidullah and G. Saha, "Speaker Verification with Short Utterances: A Review of Challenges, Trends and Opportunities", *IET Biometrics*, Vol. 7, No. 2, pp. 91–101, 2018.

10. Viñals, I., A. Ortega, A. Miguel and E. Lleida, "An Analysis of the Short Utterance Problem for Speaker Characterization", *Applied Sciences (Switzerland)*, Vol. 9, No. 18, pp. 1–19, 2019.

11. Kanagasundaram, A., D. Dean, S. Sridharan, J. Gonzalez-Dominguez, J. Gonzalez Rodriguez and D. Ramos, "Improving Short Utterance I-vector Speaker Verification using Utterance Variance Modelling and Compensation Techniques", *Speech Communication*, Vol. 59, pp. 69–82, 2014.

12. Das, R. K. and S. R. M. Prasanna, "Speaker Verification from Short Utterance Perspective: A Review", *IETE Technical Review*, Vol. 35, No. 6, pp. 599–617, 2018.

13. Zhang, W.-Q., Y. Deng, L. He and J. Liu, "Variant Time-Frequency Cepstral Features for Speaker Recognition", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2010*, Makuhari, Chiba, Japan, pp. 2122–2125, 2010.

14. Li, Z.-Y., W.-Q. Zhang and J. Liu, "Multi-resolution Time Frequency Feature and Complementary Combination for Short Utterance Speaker Recognition", *Multimedia Tools and Applications*, Vol. 74, No. 3, pp. 937–953, 2015.

15. Todisco, M., H. Delgado and N. W. D. Evans, "Articulation Rate Filtering

of CQCC Features for Automatic Speaker Verification", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2016*, San Francisco, CA, pp. 3628–3632, 2016.

16. Zhang, C. and T. F. Zheng, "A Fishervoice Based Feature Fusion Method for Short Utterance Speaker Recognition", *2013 IEEE China Summit and International Conference on Signal and Information Processing*, Beijing, China, pp. 165–169, 2013.

17. Krishnamoorthy, P., H. S. Jayanna and S. R. M. Prasanna, "Speaker Recognition under Limited Data Condition by Noise Addition", *Expert Systems with Applications*, Vol. 38, No. 10, pp. 13487–13490, 2011.

18. Mahadeva Prasanna, S. R., C. S. Gupta and B. Yegnanarayana, "Extraction of Speaker-specific Excitation Information from Linear Prediction Residual of Speech", *Speech Communication*, Vol. 48, No. 10, pp. 1243–1261, 2006.

19. Chan, W. N., N. Zheng and T. Lee, "Discrimination Power of Vocal Source and Vocal Tract Related Features for Speaker Segmentation", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 15, No. 6, pp. 1884–1892, 2007.

20. Das, R. K., S. Abhiram, S. R. M. Prasanna and A. G. Ramakrishnan, "Combining Source and System Information for Limited Data Speaker Verification", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2014*, Singapore, pp. 1836–1840, 2014.

21. Das, R. K. and S. R. Mahadeva Prasanna, "Exploring Different Attributes of Source Information for Speaker Verification with Limited Test Data", *The Journal of the Acoustical Society of America*, Vol. 140, No. 1, p. 184, 2016.

22. Park, S. J., G. Yeung, J. Kreiman, P. A. Keating and A. Alwan, "Using Voice Quality Features to Improve Short-Utterance, Text-Independent Speaker Verification

Systems", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2017*, Stockholm, Sweden, pp. 1522–1526, 2017.

23. Kenny, P., "Bayesian Speaker Verification with Heavy-Tailed Priors", *The Speaker and Language Recognition Workshop*, Brno, Czech Republic, 2010.

24. Kanagasundaram, A., R. Vogt, D. Dean and S. Sridharan, "PLDA Based Speaker Recognition on Short Utterances", *Odyssey 2012 - Speaker and Language Recognition Workshop*, Singapore, pp. 28–33, 2012.

25. Kanagasundaram, A., D. Dean, S. Sridharan, H. Ghaemmaghami and C. Fookes, "A Study on the Effects of Using Short Utterance Length Development Data in the Design of GPLDA Speaker Verification Systems", *International Journal of Speech Technology*, Vol. 20, No. 2, pp. 247—-259, 2017.

26. Hautamäki, V., Y.-C. Cheng, P. Rajan and C.-H. Lee, "Minimax I-vector Extractor for Short Duration Speaker Verification", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2013*, Lyon, France, pp. 3708–3712, 2013.

27. Mandasari, M. I., R. Saeidi, M. McLaren and D. A. van Leeuwen, "Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 21, pp. 2425–2438, 2013.

28. Nautsch, A., R. Saeidi, C. Rathgeb and C. Busch, "Robustness of Quality-based Score Calibration of Speaker Recognition Systems with respect to low-SNR and short-duration conditions", *Proc. The Speaker and Language Recognition Workshop (Odyssey 2016)*, Bilbao, Spain, pp. 358–365, 2016.

29. Poddar, A., Sahidullah and G. Saha, "Quality Measures for Speaker Verification

with Short Utterances", *Digital Signal Processing*, Vol. 88, pp. 66–79, 2019.

30. Rao, H., K. Phatak and E. Khoury, "Improving Speaker Recognition with Quality Indicators", *2021 IEEE Spoken Language Technology Workshop (SLT)*, Shenzhen, China, pp. 338–343, 2021.

31. Yang, I.-H., H.-S. Heo, S.-H. Yoon and H.-J. Yu, "Applying Compensation Techniques on I-vectors Extracted from Short-test Utterances for Speaker Verification using Deep Neural Network", *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, pp. 5490–5494, 2017.

32. Zhang, J., N. Inoue and K. Shinoda, "I-vector Transformation Using Conditional Generative Adversarial Networks for Short Utterance Speaker Verification", arXiv:1804.00290 [eess.AS], 2018.

33. Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Networks", arXiv:1406.2661 [stat.ML], 2014.

34. Liu, K. and H. Zhou, "Text-Independent Speaker Verification with Adversarial Learning on Short Utterances", *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 6569–6573, 2020.

35. Villalba, J., N. Brümmer and N. Dehak, "Tied Variational Autoencoder Backends for I-Vector Speaker Recognition", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2017*, Stockholm, Sweden, pp. 1004–1008, 2017.

36. Prasanna, S. R. M. and G. Pradhan, "Significance of Vowel-Like Regions for Speaker Verification Under Degraded Conditions", *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, No. 8, pp. 2552–2565, 2011.

37. Larcher, A., P.-M. Bousquet, K. A. Lee, D. Matrouf, H. Li and J.-F. Bonastre, "I-vectors in the Context of Phonetically-constrained Short Utterances for Speaker Verification", *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Kyoto, Japan, pp. 4773–4776, 2012.

38. Li, L., D. Wang, C. Zhang and T. F. Zheng, "Improving Short Utterance Speaker Recognition by Modeling Speech Unit Classes", *IEEE/ACM Transactions on Audio Speech and Language Processing*, Vol. 24, No. 6, pp. 1129–1139, 2016.

39. Chen, X. and C. Bao, "Phoneme-Unit-Specific Time-Delay Neural Network for Speaker Verification", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 29, pp. 1243–1255, 2021.

40. Zhou, T., Y. Zhao, J. Li, Y. Gong and J. Wu, "CNN with Phonetic Attention for Text-Independent Speaker Verification", *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Singapore, 2019.

41. Liu, Y., L. He, J. Liu and M. T. Johnson, "Introducing Phonetic Information to Speaker Embedding for Speaker Verification", *EURASIP Journal on Audio, Speech, and Music Processing*, Vol. 2019, No. 1, pp. 1–17, 2019.

42. Wang, S., J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu and J. Černocký, "On the Usage of Phonetic Information for Text-independent Speaker Embedding Extraction", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2019*, Graz, Austria, pp. 1148–1152, 2019.

43. Tawara, N., A. Ogawa, T. Iwata, M. Delcroix and T. Ogawa, "Frame-Level Phoneme-Invariant Speaker Embedding for Text-Independent Speaker Recognition on Extremely Short Utterances", *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, pp. 6799–6803, 2020.

44. Chung, J. S., J. Huh, S. Mun, M. Lee, H.-S. Heo, S. Choe, C. Ham, S.-Y. Jung, B.-J. Lee and I. Han, "In Defence of Metric Learning for Speaker Recognition", arXiv:2003.11982 [eess.AS], 2020.

45. Desplanques, B., J. Thienpondt and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification", arXiv:2005.07143 [eess.AS], 2020.

46. Snyder, D., J. Villalba, N. Chen, D. Povey, G. Sell, N. Dehak and S. Khudanpur, "The JHU Speaker Recognition System for the VOiCES 2019 Challenge", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2019*, Graz, Austria, pp. 2468–2472, 2019.

47. Ko, T., V. Peddinti, D. Povey, M. L. Seltzer and S. Khudanpur, "A Study on Data Augmentation of Reverberant Speech for Robust Speech Recognition", *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, pp. 5220–5224, 2017.

48. Kuttruff, H., *Room Acoustics (6th ed.)*, CRC Press., Boca Raton, FL, 2016.

49. Snyder, D., G. Chen and D. Povey, "MUSAN: A Music, Speech, and Noise Corpus", arXiv:1510.08484 [cs.SD], 2015.

50. Chung, J. S., A. Nagrani and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2018*, Hyderabad, India, pp. 1086–1090, 2018.

51. Nagrani, A., J. S. Chung and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset", *Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH) 2017*, Stockholm, Sweden, pp. 2616–2620, 2017.

52. Arjovsky, M., S. Chintala and L. Bottou, "Wasserstein GAN", arXiv:1701.07875 [stat.ML], 2017.

53. Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin and A. C. Courville, "Improved Training of Wasserstein GANs", *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*, Long Beach, CA, p. 5769–5779, 2017.

54. Schroff, F., D. Kalenichenko and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering", *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, pp. 815–823, 2015.

55. Garofolo, J., L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren and V. Zue, "TIMIT Acoustic-phonetic Continuous Speech Corpus", *Linguistic Data Consortium*, 1992.

56. Panayotov, V., G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR Corpus Based on Public Domain Audio Books", *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, Queensland, Australia, pp. 5206–5210, 2015.

57. Povey, D., A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. K. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, J. Silovský, G. Stemmer and K. Veselý, "The Kaldi Speech Recognition Toolkit", *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Waikoloa, HI, 2011.

58. Ang, A.-S., W. Tang and S. Ross, *Probability Concepts in Engineering*, John Wiley & Sons, Hoboken, New Jersey, 2007.

59. Ren, Z., Z. Chen and S. Xu, "Triplet Based Embedding Distance and Similarity Learning for Text-independent Speaker Verification", *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*,

Lanzhou, China, pp. 558–562, 2019.