Robust speaker verification system with anti-spoofing detection and DNN feature enhancement modules

Du, Steven

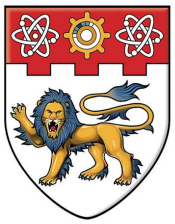
2015

Du, S. (2015). Robust speaker verification system with anti-spoofing detection and DNN feature enhancement modules. Master's thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/65396

https://doi.org/10.32657/10356/65396

Robust Speaker Verification System with Anti-spoofing Detection and DNN Feature Enhancement Modules



NANYANG TECHNOLOGICAL UNIVERSITY

A thesis submitted to
the School of Computer Engineering
of the Nanyang Technological University

by

STEVEN DU

in partial fulfillment of the requirement of the
Degree of Master of Engineering

August 11, 2015

Abstract

This thesis focuses on the robustness issues of speaker verification (SV) systems. Although current SV systems perform well under clean condition, their performance degrades dramatically under real-world uncontrolled environments. The reliability of current SV systems is also questionable under spoofing attacks. These pitfalls severely limit it's deployment in many applications. This thesis presents approaches to combat these two robustness issues, namely noise robustness and spoofing attacks.

To address the noise robustness issue, the use of deep neural networks (DNN) as a feature compensation method in the front-end module of the SV system is proposed. The motivation to use DNN is due to its success in various related speech fields, and its ability to model nonlinear relationships between high dimensional input and output. In this work, DNN is used to convert noisy input features into clean features. The proposed method is evaluated using the benchmarking speaker recognition evaluation (SRE) 2010 dataset provided by the National Institute of Standards and Technology(NIST). To focus on the effect of feature pre-processing, the SV system is trained using noise free speech and evaluated on noise corrupted speech. Results show that the proposed DNN feature compensation improves the equal error rate (EER) by 2%-25% under different unseen noise types for various SNR levels.

To address the spoofing attacks issue, the use of long temporal high dimensional speech features derived from both magnitude and phase spectra as input features to neural network (NN) classifiers is proposed. The long term temporal information is incorporated by concatenating 31 successive frames as input feature to the NN classifier. The classifier is then used to predict the posterior probability of the test speech being spoofing speech. Four speakers of CMU-ARCTIC database are selected for spoofing data generation and methods evaluation. Spoofing data is generated by four synthesis methods, namely: AHOcoder, STRAIGHT, JD-GMM with maximum likelihood parameter generation, and weighted correlation-based frequency warping (CFW). The results show that both long term information and detailed information maintained in high dimensional features improve the performance of synthetic speech detection significantly. The proposed method was extended and used to compete in the ASVspoof 2015 challenge and achieved best results in the closed set challenge among 16 teams worldwide.

Acknowledgments

It is a great pleasure for me to acknowledge the assistance and contributions of many individuals in making this dissertation a success.

First and foremost, I would like to thank my supervisor, Dr. Chng Eng Siong (NTU), for his assistance, ideas, and feedbacks during the process of doing this dissertation. Without his guidance and support, this dissertation could not have been completed on time.

Secondly, I would like to thank Dr. Li Haizhou (I^2R) and Dr. Lee Kong Aik (I^2R) for their invaluable guidance. Their constant encouragement helped me overcome the difficulties encountered in my research.

Thirdly, I want to thank my colleagues at the speech group in NTU for their generous help. I want to thank Dr. Xiao Xiong, Mr. Chong Tze Yuang, especially Mr. Tian Xiaohai for many fruitful discussions. I also want to express my sincere thanks to Osho Gupta for his support and suggestions.

Last but not the least, I wish to express my sincere gratitude to my family for their encouragement and moral support.

I also place on record, my sense of gratitude to one and all, who directly or indirectly, have lend their hand in this venture.

Contents

List of Figures

List of Tables

List of Abbreviations

ASR	Automatic Speech Recognition
DNN	Deep Neural Network
DET	Decision Error Trade-off
DTW	Dynamic Time Warping
EM	Expectation Maximization
EER	Equal Error Rate
FAR	False Alarm Rate
FRR	False Reject Rate
GMM	Gaussian Mixture Model
JFA	Joint Factor Analysis
iVector	Identity Vector
LDA	Linear Discriminant Analysis
LP	Linear Predictive
LPC	Linear Prediction Coefficient
minDCF	Minimum Decision Cost Function
MFCC	Mel Frequency Cesptral Coefficients
MAP	Maximum a posteriori
MLLR	Maximum Likelihood Linear Regression
NAP	Nuisance Attribute Projection
NIST	National Institute of Standards and Technology
NMF	Non-negative Matrix Factorization
PLDA	Probabilistic Linear Discriminant Analysis
PMC	Parallel Model Combination
RPCC	Residual Phase Cepstrum Coefficients
UBM	Universal Background Model
SNR	Signal to Noise Ratio
SRE	Speaker Recognition Evaluation
SV	Speaker Verification
SVM	Support Vector Machine
TTS	Text To Speech
VC	Voice Conversion
VTS	Vector Taylor Series
WCCN	Within Class Covariance Normalization

Chapter 1

Introduction

The speaker verification (SV) task uses speech signal to verify whether a claimed speaker is genuine or an impostor. Hence, its robust performance is an important consideration. This thesis examines two existing robustness issues that impact the performance of current SV systems.

The first robustness issue arises due to the high variability of speech signals. The variability can be categorized as intrinsic and extrinsic [3]. Intrinsic variability arises from the difference in language, speaking-style, health, and emotional state; while, extrinsic variability is due to distortions from channel and noisy conditions. Noisy condition is one of the major factors that have impacted SV performance. This is because the statistics of the noisy features in test conditions can be significantly different to the clean features used for model training and hence lead to poor classification performance. The thesis focuses on reducing the feature mismatch problem by mapping the noisy features to clean features to improve performance.

We consider the second robustness issue to be the vulnerability of SV system under spoofing attacks [4]. Spoofing is the purposeful generation of the target speaker's speech to fool a SV system. Spoofing can be performed by synthesizing a target speaker's

voice using techniques such as text-to-speech (TTS) [5] or voice conversion (VC) [6]. To detect and reject such input, this thesis proposes a NN classifier using high dimensional magnitude and phase spectra with long term information.

1.1 Motivation

In the past decade, SV system has improved significantly and achieved good performance under high signal-to-noise-ratio(SNR) conditions. Fig. 1.1(a) shows the improvement of equal error rate (EER) on the NIST SRE evaluation task [1] from the year 2001 to 2011. However, the performance degrades when there is a mismatch between training and testing condition; Fig. 1.1(b) shows the degradation of performance under additive car and babble noise based on the 2011's state of the art SV system. This clearly indicates the need for a more robust SV system for noisy conditions.

To improve the performance under noisy conditions, the existing methods can be grouped under feature or model domain approaches. For feature domain methods, noise robust features have been proposed, such as Mean Hilbert Envelop Coefficients (MHEC) [7], Residual Phase Cepstrum Coefficients (RPCC) [8], Teager Phase Cepstrum Coefficients (TPCC) [9], etc. in place of conventional MFCC features. For the model domain methods, model parameters trained from clean features are adapted to noisy conditions using techniques such as Vector Taylor Series (VTS) [10], Maximum Likelihood Linear Regression(MLLR) [11], Parallel Model Combination(PMC) [12, 13], etc. However, the performance of both feature and model domain methods depends significantly on the priori knowledge of noise type and SNR level and is not satisfactory under unseen noisy conditions [14].

To date, deep neural network (DNN) methods have been successfully applied in various tasks of speech processing [15–18]. They have achieved good performance for speech

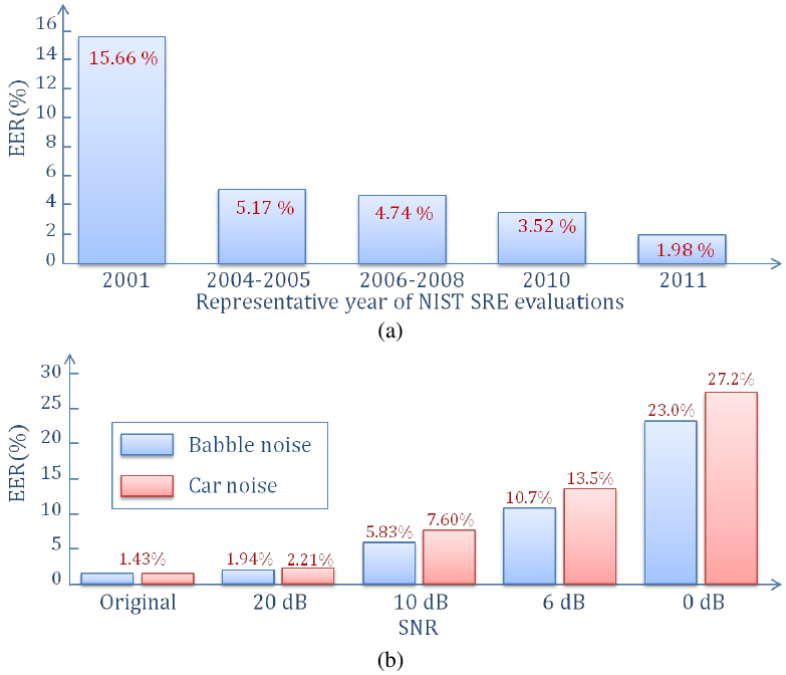


Figure 1.1: The figure is extracted from [1](a) Performance of representative state-of-the-art technologies from 2001 to 2011 on telephone data from NIST 2010 evaluation. (b) Performance degradation of the 2011 SV system as a function of SNR under babble and car noises.

processing tasks such as acoustic modeling [16, 19], speech enhancement [20] and de-reverberation [21]. In SV task, NNs and DNNs have been applied in front-end feature extraction [22–24], and back-end modeling [25–32]. DNN was shown to be able to model complicated nonlinear relationship between high dimensional input and output. Extending [23, 24], this thesis applies DNN to predict clean speech features from the noisy features in unseen conditions to improve the performance of SV systems.

Spoofing attack is also a serious challenge to SV systems. Advancements in speech synthesis techniques like TTS and VC have made it possible to generate high quality spoofing speech. This has made even the state-of-the-art SV systems vulnerable to such types of spoofing attacks [33, 34]. To detect spoofing attacks, several methods have been proposed [33, 35–38]. Most of these methods rely on GMM-based classifiers, which only allows the use of low dimensional features such as Mel-frequency cepstrum coefficients

(MFCC). This thesis proposes to use high dimensional long term magnitude and phase features to detect synthetic speech.

1.2 Contributions

Two contributions to improve the performance of current SV systems are examined.

The first contribution is proposed to use DNN as a feature compensation front-end [39] to predict the underlying clean cepstral features from noisy features. The DNN is trained using multi-conditioning training with different noise types and SNRs. The training is achieved by minimizing the mean square error (MSE) between the DNN's prediction and the clean cepstral features. Results are evaluated on the benchmarking SRE 2010 female core task, which shows that the proposed method can be used to reduce the equal error rate (EER) even for unseen noisy test conditions.

The second contribution is proposed to use long term high dimensional speech features derived from both magnitude and phase spectra to detect synthetic speech [40]. The long term temporal information is captured by concatenating features from a window of 31 frames to form the supervector. The high dimension supervector is used by a multilayer perceptron (MLP) neural network to detect spoofing attack. The proposed approach is evaluated on the CMU-ARCTIC database [41] with four algorithms to generate synthetic speech. Results show that the proposed method performs better than previous methods in terms of EER.

The first approach "DNN Feature Compensation For Noise Robust Speaker Verification" is published in [39] and the second approach "Detecting Synthetic Speech Using Long Term Magnitude And Phase Information" is published in [40].

1.3 Organization of Dissertation

The thesis is organized as follow:

Chapter 2 describes the overview of SV task and technologies, including the state-of-the-art SV systems, various features and models based techniques to improve robustness. This is followed by an overview of existing spoofing and anti-spoofing techniques.

Chapter 3 presents the proposed DNN feature compensation method. It first introduces the proposed method, then the baseline experimental setup, and finally the evaluation results.

Chapter 4 presents the proposed spoofing speech detection method. It first introduces the feature extraction module and neural network classifier of the proposed method. Next, it presents the experimental setup, including corpus and baseline method. Finally, the evaluation results and discussion are given.

Chapter 5 concludes the thesis with a summary of contributions and possible future directions.

Chapter 2

Literature Review

This chapter first presents a broad view of speaker verification(SV) system and the state-of-the-art methods. Next, we examine two existing challenges - noise robustness and spoofing attacks, faced by current SV systems. Finally, a summary of the recent proposed techniques to improve performance for these challenges is presented.

The chapter is organized as follow: sections 2.1 introduces the history of SV briefly, section 2.2 presents the general structure of speaker verification(SV) system, section 2.3 describes the state-of-the-art SV systems, section 2.4 presents the robustness issues of SV system. Finally, section 2.5 summarizes the chapter.

2.1 Overview of Speaker Verification Technologies

The Speaker Recognition (SR) task is to identify a person from his/her voice. This research is also called voice biometrics [42] and has been an area of active research for more than four decades. The initial use of SR can be traced back to 1960s when speech spectrograms were used to support legal proceedings [43], with human experts in the loop. The first Automatic Speaker Recognition (ASR) was developed by Pruzansky [44] in 1963, af-

ter which the first large scale ASR system was developed by Texas Instrument [45]. Since then, ASR has attracted much attention and has become a multi-disciplinary branch of biometrics that covers speaker verification (SV), speaker identification (SID), as well as other related researches such as speaker tracking, detection, and segmentation (diarization). More details of this subject can be found in [46].

SV and SID are the two major trends in voice biometrics [47, 48]. The purpose of SV is to verify the claimed identity of a given speech recording to a reference speaker. The SID's focus is to assign an identity to an unknown speech recording among a pool of known speakers. The SID faces additional problems when the test speech is not from a known speaker set, i.e., it requires an additional step to label unknown speaker. In other words, the SID system needs to update the known speakers set whenever a new speaker is added. This makes the SID system difficult to scale. Unlike SID, SV does not rely on known speakers dictionary. During verification, a SV system only checks the unknown speaker recording against the identity of claimed speaker [49, 50] and makes a decision. Thus, SV system has lower computation cost, and is more practical and can be generalized.

The SV task can be further classified into text-dependent and text-independent speaker verification. For text-dependent system, the test utterance must be the same as the phrase used in training. For text-independent system, there is no such constraint and hence is more flexible and challenging. This thesis will focus on the text-independent SV system.

2.2 General Structure of SV

SV system, like most of the problems in pattern recognition, could be divided into two parts: the front-end and the back-end module as shown in Fig. 2.1.

The front-end is the feature processing unit which extracts speaker specific features from the speech recording. These features could be divided into low and high level features. Typical low level features include Linear Prediction Cepstral Coefficients (LPCC) [51], Perceptual Linear Prediction (PLP) [52] and Mel-Frequency Ceptral coefficients (MFCC) [53]. MFCC being the most commonly used feature. The high level features include word usage, pronunciation, phonotactics, prosody, etc [54]. The front-end module generally includes a voice activity detection (VAD) [55] unit to remove non-speech segments that does not contribute to discriminate speakers.

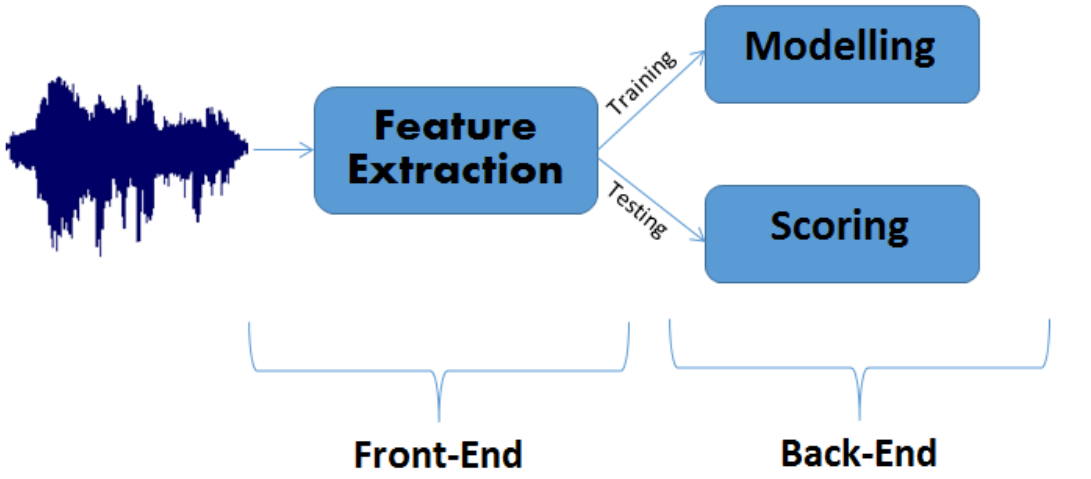


Figure 2.1: General structure of speaker verification system, showing front-end and back-end.

The role of the back-end module is to create models using the features extracted from the front-end. These models are used to verify whether a given utterance is authentic by generating a confidence score. If the confidence score is greater than a pre-determined threshold, the claimed identity is accepted. In recent years, many back-end modeling and scoring techniques have been studied; these include template models such as dynamic time wraping (DTW) [56] (for text-dependent SV), vector quantization (VQ) [57], and stochastic models such as GMM. The state-of-the-art modeling technique is the GMM

based iVector/PLDA system with MFCC as basic features [58]. The following section will discuss these back-ends techniques.

2.3 The State-of-the-art Speaker Verification Systems

2.3.1 Milestones

In recent years, the Gaussian Mixture Models (GMMs) [59, 60], adapted GMMs [61], GMM supervectors [62–64], JFA and iVectors have became the state-of-the-art approaches for SV task.

The GMM is one of the dominant approaches for speaker modeling in text-independent speaker recognition applications [65]. In 1990s, a GMM speaker model [59] was trained on twenty dimensional MFCC feature vectors extracted from enrolled utterances, the log likelihood between unknown utterances with respected enrolled speaker model was calculated as the confidence score. This scheme improved in 2000 [61] with the use of an Universal Background Model (UBM) for alternative speaker representation by Maximum A Posteriori (MAP) [66] adaptation. The UBM model is a GMM model trained on significantly large amount of speech data from many speakers. The MAP adaptation technique is then applied to generate the enrolled speaker model from the UBM model. The confidence score is the difference between the log likelihood of a test utterance from UBM and that from adapted speaker model.

Followed by the considerable success achieved by GMM-UBM, the speaker supervectors were then proposed. The aim was to convert a utterance with arbitrary duration to a fixed length speaker vector. This then allows the use of classifiers like supported vector machine [62–64] for the SV task. The supervector mentioned here is specific to the GMM supervector which is constructed by stacking the means of the adapted mixture compo-

nents. For example, an adapted GMM with 2048 components built on 60 dimensional feature vectors results into a 122880 (2048*60) dimensional supervector. A recent review of GMM techniques to supervector is presented in [42].

The use of a fixed length speaker vector enables the application of various factor analysis techniques such as the JFA [67] in SV task. The JFA approach decomposes a supervector s into speaker independent vector m, speaker dependent components V, channel dependent components U and residual components D, i.e.,

$$s = m + Vy + Ux + Dz$$

where y, x, z are assumed to have standard Gaussian distribution. In next section, the state-of-the-art method for SV, the iVector approach, is presented.

2.3.2 iVector Framework

The iVector framework [58] extends from the JFA technique. It was first proposed in [58], and further improved the SV performance. In general, iVector framework consists three parts: iVector extraction module, post-processing module and scoring module. The following sections present the details.

2.3.2.1 iVector Extraction

Contrasting to JFA technique, the iVector framework compresses the features extracted from a utterance into a much lower dimension by merging the speaker and channel dependent components into a new space known as total variability space. A couple of assumptions are made: (i) the speaker and channel components are statistically independent and (ii) these components have Gaussian distributions. A speaker and channel

dependent GMM supervector s can thus be represented as

$$s = m + Tw$$

where m denotes the speaker and channel independent background UBM supervector, T is the total variability matrix, which is used to represent the variation of primary directions across a large amount of training data. The coefficient w of this total variability is known as iVector. Given a SV system built on F dimensional MFCC features and UBM with C Gaussian components, the iVector extraction can be done as follow:

$$w = (I + T^T \Sigma^{-1} N T)^{-1} T^T \Sigma^{-1} A$$

where I is $F \times F$ identity matrix, N is a $CF \times CF$ diagonal matrix whose diagonal blocks are $N_c I$(c=1,2,....C), and the supervector A is generated by the concatenation of the centralized first-order Baum-Welch statistics. Σ is the covariance matrix of the residual variability not captured by T. The general process of computing the total-variability space is similar to that of JFA eigenvoice training except for one significant difference. In the JFA eigenvoice training, all the utterances of given speaker are considered to be the same person, while in total variability space training, they are considered to be different persons to capture the channel variation. An efficient method of computing the total-variability space is given in [58, 68].

The iVector's dimension is much lower than that of supervectors. This thus allows the use of various techniques that were not practical in high dimensional supervectors. The next section discusses some of these techniques.

2.3.2.2 Post-processing of iVector

Post-processing techniques attempt to reduce the effect of channel variability of the iVector, it is also known as channel compensation. Such techniques including nuisance attribute projection (NAP) [69], within class covariance normalization (WCCN) [70], and Linear Discriminant Analysis (LDA).

The NAP, introduced in [71, 72], is based on finding an appropriate projection matrix to remove the nuisance direction while keeping the relevant speaker features. This technique was initially designed to operate on supervector and later used on iVector [58]. It however only showed a little improvement when applied on iVectors.

The WCCN technique proposed in [73] is also used in the SVM system. It is based on linear separation of target speaker supervectors and background speaker supervectors using a one-versus-all decision. The WCCN is reported to provide an efficient channel compensation when combined with Cosine Distance on iVector based system [74].

LDA is a commonly used technique for dimension reduction and classification task [75, 76]. The motivation for using this technique for SV task is that LDA will treat each speaker as one class and then transforms the iVectors into an even lower dimensional space with new axes that attempts to minimize the intraclass variance and maximize the interclass variances. Then the cosine distance can be used as final scoring method on these transfered iVectors.

2.3.2.3 Scoring Techniques

Cosine Distance Based Scoring The most commonly used scoring technique is based on cosine distance because of its computational efficiency [58]. The cosine distance between two speaker iVectors t and e is defined as:

$$\cos(\mathbf{t}, \mathbf{e}) = \frac{< \mathbf{t}, \mathbf{e} >}{\|\mathbf{t}\|\|\mathbf{e}\|} = \frac{\sum_{i=1}^{n} \mathbf{t}_i \mathbf{e}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{t}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{e}_i)^2}} \tag{2.1}$$

The resulting value is between -1 and 1, where 1 means that the two speaker vectors are perfectly aligned and -1 denotes that they are inversely related. When the value is 0, it means that the two vectors are orthogonal and hence are not related.

Cosine distance can be used either alone or jointly with other methods. The best results in cosine distance based scoring are obtained when LDA is followed by WCCN [58].

SVM Scoring The SVM [77], in particular, is mainly used for binary classification and uses the criterion to maximize the gap between two opposing classes. In the case of speaker verification system, the two classes are the target speaker iVectors and the background speaker iVectors. It is found in [58], that the use of cosine distance as kernel function to measure the gap yields good results.

PLDA Scoring PLDA approach is a part of the current state of the art SV system [78]. It is a generative probabilistic model of iVector distributions for speaker verification scoring. Recall the assumption for iVector, i.e. the speaker and channel components are assumed to be independent and Gaussian distributed. In the PLDA approach, similar assumption (G-PLDA) [79] is applied. This approach was proposed by Prince for face recognition research and was recently applied to SV task. It's implementation is simple and fast due to it's closed-form solutions. The Heavy-Tailed PLDA (HT-PLDA) proposed by Kenny [80] in 2010 for speaker verification demonstrated superior performance but is computationally expensive. This result highlighted the non-Gaussian behavior of the speaker components in iVector framework. This motivated [78], in which iVector is transformed to reduce the non-Gaussian behavior [81] and then the G-PLDA is applied. His approach kept the simplicity of Gaussian model and yet achieved performance

equivalent to HT-PLDA. The PLDA score of two iVector $\mathbf{x}_s$ and $\mathbf{x}_t$ is defined as:

$$\begin{aligned} S_{\text{LR}}(\mathbf{x}_s, \mathbf{x}_t) &= \frac{P(\mathbf{x}_s, \mathbf{x}_t|\text{same speaker})}{P(\mathbf{x}_s, \mathbf{x}_t|\text{different speakers})} \\ &= \text{const} + \mathbf{x}_s^{\mathsf{T}}\mathbf{Q}\mathbf{x}_s + \mathbf{x}_t^{\mathsf{T}}\mathbf{Q}\mathbf{x}_t + 2\mathbf{x}_s^{\mathsf{T}}\mathbf{P}\mathbf{x}_t \end{aligned} \tag{2.2}$$

where

$$\begin{aligned} \mathbf{P} &= \mathbf{\Lambda}^{-1}\mathbf{\Gamma}(\mathbf{\Lambda} - \mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma})^{-1}; \quad \mathbf{\Lambda} = \mathbf{V}\mathbf{V}^{\mathsf{T}} + \mathbf{\Sigma} \\ \mathbf{Q} &= \mathbf{\Lambda}^{-1} - (\mathbf{\Lambda} - \mathbf{\Gamma}\mathbf{\Lambda}^{-1}\mathbf{\Gamma})^{-1}; \quad \mathbf{\Gamma} = \mathbf{V}\mathbf{V}^{\mathsf{T}}. \end{aligned} \tag{2.3}$$

and $\mathbf{V}$ is the factor loading matrix and $\mathbf{\Sigma}$ is the covariance of the PLDA model.

Just like the cosine distance scoring approach, PLDA can be used jointly with other approaches such as LDA. To date, the iVector/PLDA framework is the state of the art speaker verification technique.

2.4 Robustness Issues and Challenges of The Current Speaker Verification Systems

The robustness issues of SV are associated with the high variability of speech signal due to both intrinsic and extrinsic sources. The intrinsic sources of variability include speaking style, vocal effort, emotional and health state of speaker and linguistic factors [82] like speech content, utterance length, language, dialectal variations. The extrinsic sources are due to channel mismatch and environmental noise [14]. Currently, the channel mismatch issue has been somewhat addressed by the state-of-the-art approach of iVector/PLDA framework [58]. The environmental noise issue still remains a challenging problem.

In addition, the SV system faces new challenges from spoofing attacks. In this work,

the focus is to improve the robustness of the SV system under the environmental noise conditions as well as to detect spoofing attacks. The recent works on environmental noise and spoofing attacks will be discussed in following sections.

2.4.1 Environmental Noise

The environmental noise is a major factor that dramatically degrades the performance of the SV system [83–86]. The noise can severely affect the spectral features of speech. It is generally considered as additive, and its effect on speech features is hard to be quantified due to its randomness nature. This addition of noise results into arbitrary distortion of the speech features which causes the loss of discriminative speaker information. As speech features are the primary input of iVector computation [87], the resulting iVector becomes unreliable and fails to correctly represent the identity of the speaker.

To date, most of the researches have targeted the impact of noise with either feature or model domain approaches. The next section discusses some of these approaches.

2.4.1.1 Feature Domain Approaches

Feature Domain approaches could be divided into two groups: one group attempts to find the most robust speaker features, and the other group attempts to compensate the existing features using noise reduction algorithms [14]. Due to the low computational complexity and independence with back-end SV models, the second group is often used as a front-end processing unit.

Robust Feature Estimation There are mainly three classes of modifications to the MFCC feature for noise robust speaker verification.

The first class is based on linear prediction (LP). A comparative study of robust linear predictive analysis techniques for speaker identification is presented in [88]. It evaluated

various objective functions for LP coefficient estimation. Other studies [89–92] analyzed weighted LP, regularized variants of LP, and frequency domain LP. The evaluation given in [93] showed that for all the SNR levels, unweighted LP offered improvement over baseline MFCC, while the weighted LP could be improved depending on the noise type and SNR levels.

The second class depends on the use of different windows for feature extraction, i.e., in [94, 95], multi-taper spectrum estimations were presented and asymmetric G.729 speech coding windows were used in [96]. However, the evaluation in [93] showed that this class of methods was less effective than LP methods for all but clean test conditions.

The third class seeks for alternative spectrum estimation methods for speaker verification, such as wavelet transformation [97, 98], the denoised wave atom [99], using gammatone filter bank as a cochlear filter simulator [100] and Mean Hilbert Envelope Coefficients (MHEC) [7]. In [101], the instantaneous frequency has been shown to be robust against channel and speaking style. The instantaneous frequency is based on amplitude and frequency modulations (AM-FM) in speech signals. The study in [102] shows that AM-FM is more robust as compared to MFCC under babble noise conditions. In 2014, [9] indicated that physiologically motived features are more accurate than and complementary to traditional MFCC. These features include Residual Phase Cepstrum Coefficients (RPCC) [8], Glottal Flow Cepstrum Coefficients (GLFCC) [103], Teager Phase Cepstrum Coefficients (TPCC) [9]. Moreover, phase has been justified as an important feature for speech and audio processing [104–107]. Phase based modified group feature (MODGDF) [108] is shown to be robust to additive noise in [109].

To conclude, feature estimation techniques could have a significant impact on speaker identification performance under noisy condition. However, those techniques still depend on prior knowledge of the noise type and SNR level [110].

Feature Compensation These methods try to alleviate the adverse effect of the noise corrupted features. Extensive research has resulted into some well-known methods such as RASTA filtering [111], cepstral mean normalization (CMN), relative autocorrelation spectrum (RAS) [112], feature warping [113], differential power spectrum (DPS) and power spectral subtraction (PSS) [114]. Significant performance improvement has been reported by use of these methods [114–118].

Speech enhancement techniques used in SV include wiener filtering, spectral equalization (SE) [119], minimum mean squared error (MMSE), log spectral amplitude estimator [120] and spectral subtraction (SS) [121]. [122] reported SS to be robust against real-world factory noise and aircraft noise. However it suffered from annoying residual musical noise effect [123] and required a prior knowledge of noise. In [117], it is highlighted that depending on the noise type and SNR levels, current speech enhancement front-ends might help or hurt SV performance.

Another approach for feature compensation is based on identifying and emphasizing/discarding features in specific regions, e.g., vowel like regions (VLRs) [124] emphasize specific features, while missing feature or missing data (MD) theory [125] discards the severely corrupted features in time-frequency (T-F) representation of noisy speech. In [126], the combination of missing feature theory and multi-condition training demonstrated improved performance over baseline system that was trained on clean data. The key task for missing feature approach is the estimation of a reliability mask. [127–132] used a binary mask to distinguish between the reliable and unreliable spetrographic features; [133, 134] focused on mask estimation inspired by computational auditory scene analysis (CASA); [135] proposed the use of matrix completion under missing data framework and [136] focused on short training and testing sessions.

There is another group of approaches based on matrix factorization such as PCA, LDA [137], and NMF[138] which have been shown to be effective for robust speech feature

extraction under noisy condition. Several studies have confirmed that the activation matrix of NMF can represent speaker identity and phonetic content in the speech [139, 140]. Moreover, Qiang [141] proposed Cortical Tensor Cepstral Coefficients (CTCC) feature extraction for speaker recognition based on constrained Non-negative Tensor Factorization (cNTF). They demonstrated that such feature led to higher recognition accuracy in noisy environment than MFCC, NMF, and PCA.

The most promising feature compensation approach is via deep learning. Deep learning has been shown to be very effective in many research areas such as image processing, speech processing and natural language processing. Current SV system has adopted deep learning techniques for both front-end and back-end [142, 143] units. In terms of feature compensation, neural network models have been used as mapping functions to transform emotion-specific features to emotion-independent features [144]. DNN has also been applied to estimate CASA mask in [145], while in [146], regularized Siamese deep network was used to extract speaker-specific information.

To summarize, the feature domain approach has the advantage of low computational complexity. However, it is considered to be a point estimate of clean speech features, and is thus unable to capture the uncertainty of observation [147].

2.4.1.2 Model Domain Approaches

The purpose of model domain approaches is mainly to adapt the existing speaker model parameters toward noise conditions. In [14], the author suggested that model domain approaches perform better than feature level compensation approaches. There are two major trends in model domain approaches: one approach uses data driven method to adapt trained model, another approach exploits a priori knowledge about the test environment.

In the first type of approaches, Maximum a Posteriori (MAP) [148] and Maximum

Likelihood Linear Regression (MLLR) [11] are the two most popular and traditional data-driven adaptation techniques. These were initially used in robust speech recognition, and then successfully applied to speaker verification task [61]. Recently, [149] proposed using Acoustic Factor Analyzer (AFA) to model acoustic features instead of GMM-UBM, and showed superior results [150].

In the second type of approaches, the Parallel Model Combination (PMC) was proposed in the late 1990s [12, 13]. Clean training model and a noise test GMM model were used in this work to estimate the parameters of a noise corruption function. Recently, [10] proposed the Vector Taylor Series (VTS) that used only mathematical derivation to represent the noise corruption process, and was much simpler than PMC. Following this, a noise robust iVector extractor using VTS was also proposed by Yun Lei [151] in 2013. In 2014, a simplified VTS was published [152], which was shown to be superior than JFA [58].

Comparing to feature domain approaches, model domain approaches are computationally expensive and often require substantial amount of training data.

2.4.2 Spoofing Attacks

The state-of-the-art technologies have raised the level of SV system's accuracy to enable real and massive deployments, but the reliability of SV based voice authentication system remains a concern when facing spoofing attacks. Spoofing is the purposeful generation of the target speaker's speech to fool a SV system. For this reason, a new direction of spoofing and anti-spoofing under SV research has recently appeared [153].

2.4.2.1 Spoofing Methods

Study in [153] classified spoofing attacks into two categories: direct and indirect attacks. Direct attacks are applied at the input signal of the SV system, while indirect attacks are performed within the SV system. Indirect attack requires system-level access to manipulate the internal operations of SV system such as feature extraction, models, score and the decision logic. As direct attacks does not require system access, they can be easily implemented and deployed. Thus, they pose the greatest threat to SV systems [154]. This thesis focuses on only the direct attack category. The most common direct attack techniques include: impersonation, replay attacks [155–157], speech synthesis via text to speech (TTS) [158–161] and voice conversion (VC) [162, 163]. These are discussed in the following paragraphs.

Impersonation It refers to a human impersonator training his voice to mimic the target speaker [164]. Zetterholm's analysis [165] showed that a professional impersonator could adjust his fundamental and formant frequencies of vowels to mimic the target speaker and fool a human listener but not a SV system. A recent study [166] used a professional impersonator to mimic five target speakers. Even though his voices sound convincing to a human listener, the impersonator failed to consistently spoof a SV system. The work by Lau [167] also showed limited success by impersonators to fool SV system. Therefore, human voice mimicking is not a generalized approach for spoofing SV system.

Replay Attack Another spoofing attack is replay attack [168] or cut-and-paste attack. It is easily implemented using a prerecorded voice of target user as input to SV system. A study in the late 1990s [169] highlighted that text-dependent SV system could fail to identify cut-and-paste speech as spoofing attacks. Similar results were also obtained in recent text-independent JFA [156, 170] and GMM-UBM based SV systems [171]. In 2015,

Wu et.al. [153] compared genuine to replayed speech and confirmed that the spectrogram and formant tracks between the two are almost indistinguishable.

Text-to-speech Text-to-speech [5] systems are used to generate speech signals for a given text. Recent TTS systems use technologies such as unit selection [172], statistical parametric [173], hybrid methods [174] and speaker adaptation techniques [175–177] to produce high quality and natural sounding speech. [178] employed a hybrid TTS approach to successfully attack a SV system obtaining a false alarm rate of 98%. [179] also mentioned that state-of-the-art TTS approach which models target speaker's voice is one of the most successful spoofing methods.

Voice Conversion Voice conversion technique morphs an input speech to sound like the target speaker. Hence, it can be used to attack SV system. The work in [180] evaluated the vulnerability of the classical GMM-UBM based SV system under VC attacks, and showed that VC technique can successfully morph the spectral shape of the attacker speech towards the genuine target speaker. They reported a drastic increase in the false acceptance rate of SV system from 8% to 100%. Similarly, Wu Et.al. [181] also showed that even the advanced speaker verification systems such as JFA and PLDA are vulnerable to VC attacks.

In summary, the attacks from TTS and VC represent a very genuine threat [182–184] to SV system.

2.4.2.2 Anti-spoofing

With the advancements in spoofing techniques, the research community has responded with various countermeasures to detect and defend such attacks [153]. However, the problem is far from being solved.

Spoofing attacks can be addressed by improving the robustness of the SV system or having a dedicated module to detect synthetic spoofing speech [160]. In [33, 35], various SV systems have been studied to examine their robustness against spoofing attacks, such as the JFA, GMM-UBM, etc. In [40, 185, 186] a dedicated module is used to detect whether the incoming speech is natural or synthetic. If a synthetic speech is detected, it will be immediately rejected, and hence can help to reduce FAR.

Several types of spoofing speech detection techniques have been studied. Most of these techniques are tailored to detect synthetic speech produced by specific synthesis algorithm. For example, the detection of synthetic speech generated by Hidden Markov Model (HMM) based TTS system was studied by [36–38, 187], and the detection of synthetic speech generated by VC techniques was studied by [33, 34, 181, 188, 189]. Most of these works rely heavily on GMM-based classifier using low dimensional features with limited temporal information [33, 35–38]. In [158, 190, 191], they demonstrated increased robustness to spoofing detection by using prosodic and phase features. These previous approaches motivate this thesis to examine high-dimensional features with long term temporal information to improve synthetic speech detection, which will be presented in chapter 4.

2.5 Summary

This chapter provides a brief review of speaker verification system, spoofing attack and anti-spoofing approaches. Firstly, the basic structure of SV system and the state-of-the-art iVector/PLDA framework are presented. Followed by challenges such as channel miss-match which are almost coped, while on the other hand noise robustness is still unanswered. Various attempts such as feature and model domain approaches towards noise robustness are discussed, most of methods relay on prior knowledge of noise data

to achieve best performance. Lastly, robustness of SV systems against spoofing attacks is discussed. It is still a concern for SV community.

The next chapter will introduce our baseline iVector/PLDA SV system built on clean data, including its dataset, evaluation metrics, detailed approach and results. After that, Chapter 4 will discuss more about spoofing and anti-spoofing and a new approach will be presented to tackle the spoofing attacks.

Chapter 3

DNN Feature Compensation For Noise Robust Speaker Verification

This chapter introduces the proposed DNN based feature compensation method. Firstly, we introduce the official benchmarking dataset used in this work. Then the baseline system is presented which is built on the well-established iVector/PLDA framework. Finally, the performance of SV system under both clean conditions and various noisy conditions are evaluated.

3.1 NIST Speaker Verification Evaluation Benchmarking

The goal of the NIST Speaker Recognition Evaluation (SRE) series is to contribute to the direction of research efforts and the calibration of technical capabilities of text independent speaker recognition. This section will briefly introduce the most commonly used speaker verification benchmarking databases.

3.1.1 Baseline Database

Many datasets are available for speaker verification benchmarking in literatures, such as Fisher Corpus [192], TIMIT [193], NTIMIT [194], KING, YOHO ,Switchboard I and II, Cellular Switch board and Tactical Speaker Identification from Linguistic Data Consortium (LDC) [195]. The European Language Resources Association (ELRA) provides over 140 corpus with large number of speakers and languages. The US National Institute of Standards and Technology (NIST) has also established a number of corpus. Based on these copus, they launched regular speaker recognition evaluations (SRE) since 1996, which provides a realistic and challenging benchmarking setup, attracting many researchers to make efforts for the competition.

The latest evaluation is SRE 2012. However, there are some drawbacks of this evaluation. Firstly, the rules of building speaker model are different from previous evaluations. In NIST's words, "SRE12 task conditions represent a significant departure from previous NIST SRE's (NIST 2012, p.1)". Secondly, different to previous SREs, SRE 2012 is an open-set speaker verification, which contains unknown speaker sets. Thirdly, some of the test segments include additive noise which is added in post processing.

The SRE2010 is a English-only text-independent close set speaker verification setup, with multiple training/testing conditions. The number of trials is large, e.g, there are 570176 trials for the core task setup broken down into 9 test conditions by match/mismatch training and testing conditions. The SRE2010 evaluation focuses on speaker detection in the context of conversational speech over multiple types of channels. The evaluation consists of one required task and eight optional tasks. The required task is called the core task and is the focus of this work.

In the core task, the training and testing data are from either i) conversational speech from a 5 minutes recording from telephone channel or room microphone channel, or ii)

microphone recorded conversational segment of three to fifteen minutes total duration involving the interviewee (target speaker) and an interviewer. Note that all the data has sampling rate of 8KHz. This mix of data further allows it to be divided into male and female speakers, and 9 conditions based on different mix of train and test data. It allows explicit examination of match and mis-match training and test conditions.

3.1.2 Evaluation Metrics

The evaluation of SRE is carried out in the following manner. Firstly, the system will generate scores for trials (refer section 2.2), the higher score indicates the higher confidence of the trail to be a target speaker trail. Secondly, the Equal Error Rate or minimum decision cost function will be used to evaluate the performance of system.

3.1.2.1 Equal Error Rate

Equal Error Rate (EER) is the value where the False Alarm Rate (FAR) is equal (or have the minimum distance) to the False Reject Rate (FRR). False alarm or false acceptance is the situation where the system incorrectly verifies or identifies a non-target speaker. False Alarm Rate states the ratio of the number of false alarm divided by the number of total target trials. On the other hand, False Reject means a false rejection of a true target speaker. The False Reject Rate is number of false rejections divided by the total non-target trials.

To compute EER, as illustrated in Fig. 3.1, the scoring algorithm will sweep over all possible scores and compute the FAR and FRR at each score by considering the score as decision threshold. Among the pairs of FAR and FRR the one with minimum distance is picked up as the EER candidate.

Decision Error Trade-off (DET) consists of FAR and FRR being its two axes. The

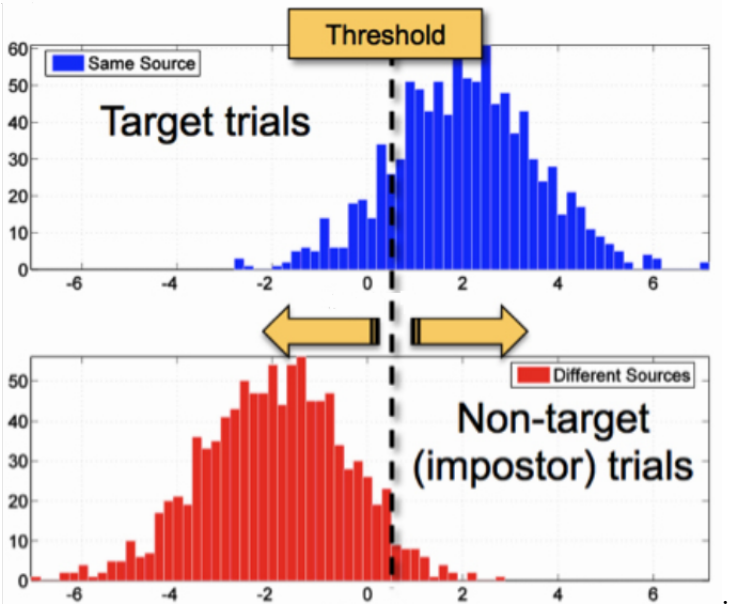


Figure 3.1: This figure is extracted from [2]. Compute EER by shifting threshold among the scores of target and non-target trials. In the histogram, the x axis indicates the range of score, the y axis is the occurrences of trials fall in that range.

DET curve is widely used to represent speaker verification system performance. Fig. 3.2 shows the DET curves of one such system. On DET curve, the point where the FAR equals FRR is highlighted as the EER point.

3.1.2.2 Minimum Decision Cost Function

In the EER, the cost of false acceptance (alarm) and false reject (miss) are equally weighted. However, it may not be suitable for real applications. Consider a SV based banking system, where a false acceptance may cost the loss of millions dollars. To tackle this problem, the SRE proposed cost model for performance evaluation.

In SRE2010 the decision cost function (DCF) is defined as a weighted sum of miss and false alarm error probabilities:

$$C_{Det} = C_{Miss} * P_{Miss|Target} * P_{Target} + C_{FalseAlarm} * P_{FalseAlarm|Non-Target} * (1 - P_{Target})$$

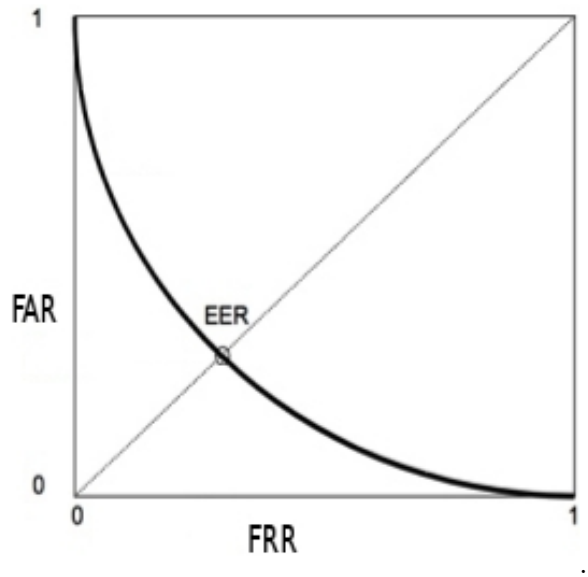


Figure 3.2: EER point on the DET cures, where the FAR equals FRR.

where C_{Miss},$C_{FalseAlarm}$ are the relative cost of errors and P_{Target} is the priori probability of the specified target speaker. The primary SRE2010 evaluation will use $C_{Miss} = 10$,$C_{FalseAlarm} = 1$,$P_{Target} = 0.01$. minDCF is the minimum value of DCF that a speaker verification system can achieve.

3.2 iVector/PLDA Baseline System

This section will introduce the baseline system. It is based on the well-established iVector/PLDA speaker verification framework.

3.2.1 Framework

Fig 3.3 shows the baseline iVector/PLDA framework used in this work. The framework is developed based on Kaldi [196] r3473 code base.

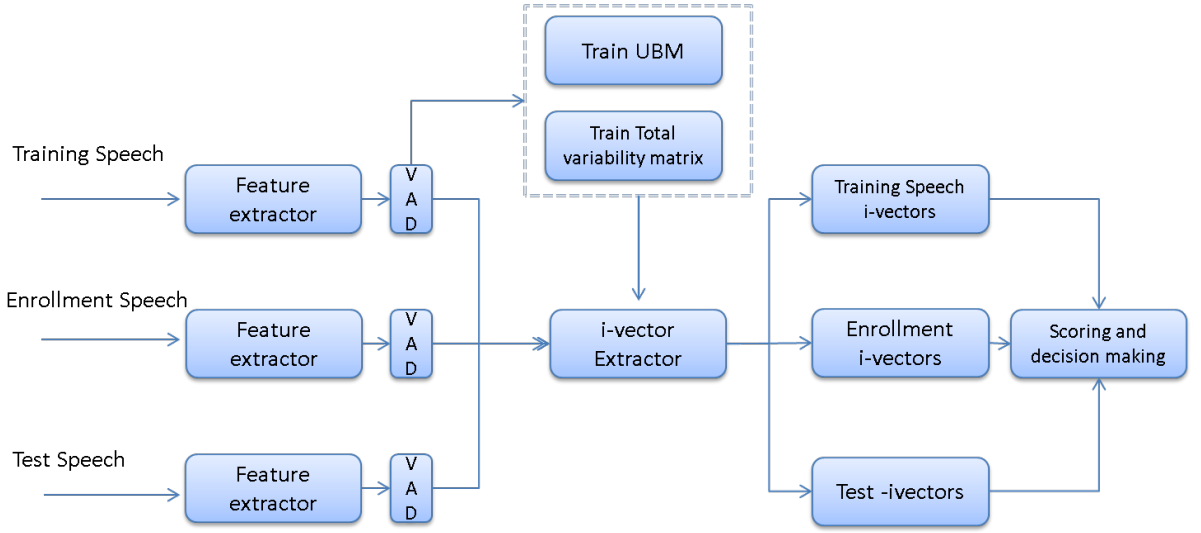


Figure 3.3: Kaldi SRE2010 framework.

3.2.1.1 Font-end and Back-end

The front-end mainly consists of feature extraction part, as shown in Fig 3.4. The raw speech signal used in this work is sampled in 8khz, 16bits depth, approximately 16k bytes per second. The MFCC features are used as the base features. The dimensionality of MFCC is 20. To include the temporal information, the delta and the double-delta of MFCC are included to form a 60-dimensional feature. An energy based voice activity detector (VAD) was used to remove the silence frames.

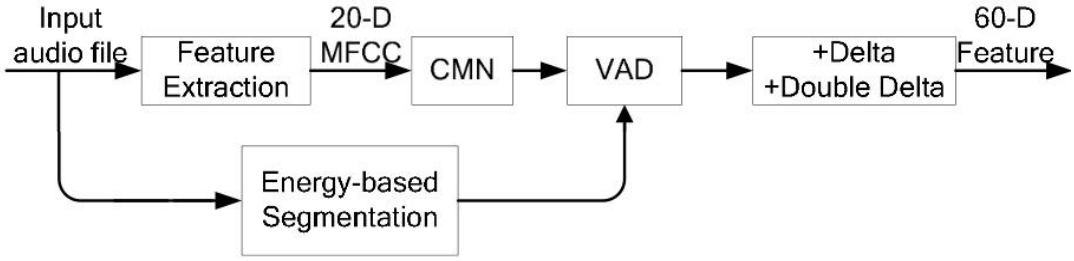


Figure 3.4: Front-end processing,raw audio input is processed resulting 60 dimensional feature vector for each 10ms frame.

The back-end includes UBM training, iVector Extractor training and scoring.

UBM Training The UBM is used to produce the supervector which is trained by hundreds of hours of speech data from Switchboard II Phase 2 and 3, Switchboard Cellular Part 1 and 2, and NIST SRE 2004, 2005, and 2006. The number of Gaussian mixtures of the UBM is 2048.

IVector Extractor Training For iVector extractor training, Total Variability Matrix is trained on the same data as the UBM training. The role of this iVector extractor is to infer the coefficients(iVector) of this Total Variability Matrix during run time. The iVector extractor converts a supervector of an utterance to a low-dimensional vector.

Scoring The Linear Discriminant Analysis transformation is applied followed by PLDA scoring to compute the log likelihood ratio between same and different speaker hypothesis. The PLDA transformation matrix is also trained on the same data as the UBM training.

3.2.1.2 System Training, Enrollment and Testing Stage

The three stages discussed above represent the major tasks of speaker verification system in time sequence.

During the system training stage, the UBM, Total Variability Matrix and iVector extractor are trained. During enrollment stage, the iVector of enrollment speaker will be extracted by iVector extractor. Each known speaker is represented by one iVector. During the testing stage, the iVectors of test utterance are extracted. A confidence score is given for each trial.

The performance of SV system under clean and noisy conditions will be investigated in the following subsections.

3.2.2 Test in Clean Condition

Clean condition refers to the scenario where both enrollment and test phases are carried out in a noise free environment, i.e. NIST SRE2010.

3.2.2.1 System Performance

Table 3.1 shows the performance of SV system under clean condition. The scores calculated by different methods, such as Cosine, LDA, PLDA, and LDA followed PLDA scoring, are reported. C1 to C9 indicate the nine core-core test conditions, each condition is a mixture of different channel and vocal efforts. Details of the conditions can be found in [197]. Among the four scoring approaches the combination of LDA and PLDA produced the best results in all the nine conditions. Since the best scoring method is the LDA followed by PLDA scoring, this is used as the baseline scoring method in this work.

Table 3.1: EER% score for four scoring methods under nine test conditions in noise free environment

Test Conditions	Scoring methods			
	Cosine	LDA	PLDA	LDA+PLDA
C1	9.89	4.92	3.87	3.26
C2	21.86	10.64	9.01	7.63
C3	11.31	5.65	3.52	3.51
C4	18.05	5.17	4.91	3.15
C5	7.89	3.94	3.10	2.31
C6	12.02	7.12	4.37	4.37
C7	17.22	10.00	7.92	5.00
C8	2.80	2.24	1.12	1.12
C9	7.54	1.74	1.94	1.11
Average	12.06	5.71	4.40	**3.50**

3.2.3 Test in Unknown Noisy Condition

This thesis assumes the unfavorable scenarios where both the noise type and SNR are unknown to the system. To simulate such scenarios, the iVector extractor and enrolled speaker iVector are kept unchanged, meanwhile four type of noises F16, Volvo, White and Babble noise from Noisex92 [198] are injected to test speech at SNR of 0dB, 10dB, and 20dB, resulting into 12 test datasets. The oracle VAD is applied before noise addition in order to evaluate the proposed system independent of the VAD quality, as suggested by [199]

The performance of the 12 noise corrupted SRE2010 core test female trials are shown in the Table 3.2. On the last column the baseline system trained in clean condition gives an average EER of 3.50%. The rest of columns show the performance of baseline system on the corrupted SRE 2010 trials.

It is observed that the EER degrades while SNR decreases. Among all of the noise types, Volvo noise obtain the least degradation in performance while SNR decreases, with only 1% absolute increase in EER at 0dB. On the other hand, large degradation of EER is found in white, F16, and babble noises at 0 dB. The results show that the state-of-the-art speaker verification system is not robust against these noise types.

3.3 DNN Feature Compensation

This chapter introduces the proposed DNN based feature compensation method for speaker verification system under unknown noise conditions.

Recently, deep neural networks (DNN) have been applied in various tasks of speech processing [15, 16]. It has been shown to offer better mapping capability over shallower network. In general, DNN has achieved good performance over many speech processing tasks, such as acoustic modeling [16, 19], speech enhancement [20] and de-

reverberation [21]. Moreover, DNN is capable to model complicated nonlinear relationships between high dimensional features. As shown by many studies, the DNN has achieved promising success in predicting clean speech from noisy and reverberant speech.

As inspired by these previous works, this chapter discusses an approach to improve the noise robustness of the SV system by using DNN. Specifically, DNN is used to improve the noise robustness capability of the speech features. The speech feature will then be used for i-Vector extraction, as shown in the Fig 3.5. The DNN is trained from parallel data of clean and noisy speech features, both types of feature are aligned at the frame level.

The proposed DNN feature compensation method, including the structure, training process and evaluation results, will be presented in following sections.

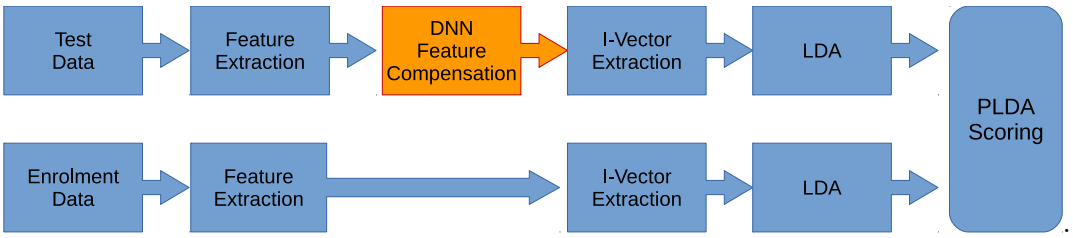


Figure 3.5: DNN feature compensation module in SV system during enrolment and test stage.

3.3.1 DNN Structure

The block diagram of the proposed DNN based feature compensation is shown in Fig. 3.6. The DNN is used to predict the corresponding clean feature set from given noisy feature vector sequences. $\mathbf{x(t)}$ denotes the raw 20 dimension feature, including 19-dimensional MFCC and 1-dimensional logEnergy, of the system at time t. To predict the the clean feature $\tilde{x}(t) \in R^{20}$, the DNN uses 11 frames, specifically, 5 past and 5 future along with the current frame. Thus, the input vector has a 11 frame context with the dimension of

220.

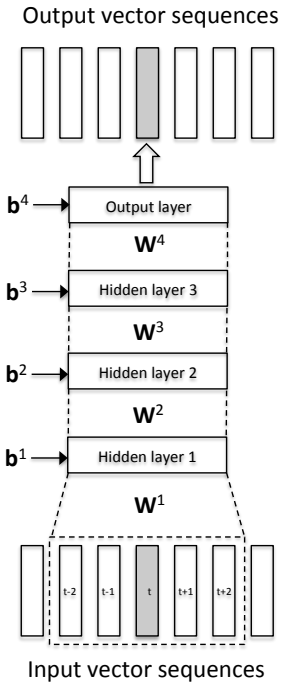


Figure 3.6: DNN structure for feature compensation

There are three hidden layers in the DNN with 2048 nodes per layer. Each layer has a linear transform and a nonlinear activation function. The input vector is affine transformed by $\mathbf{W}^1$ and $\mathbf{b}^1$ first, then goes through the activation function to form the output of hidden layer 1 which is then forwarded further to the subsequent layers till the output layer of DNN. A linear activation function in the output layer is used to formulate the regression task. The MSE is selected as the cost function for DNN training. The MSE between the predicted features and the target clean features are back propagated through all hidden layers. Specifically, the stochastic gradient descent algorithm is used to train the network parameters.

It is well known that the pre-training could result a better training performance as it could prevent training trapped in local minimum. Thus, the system was first pre-trained by using restricted Boltzmann machine (RBM) [200].

3.3.2 DNN Training Data Preparation

For DNN training, parallel features are generated from a subset of UBM training data (90 hours of speech was selected). The noisy data are created by adding noise from noise samples of Noisex92 [198]. Specifically, the buccaneer1, destroyerengine, destroyerops, factory1, hfchannel, leopard, machine-gun, pink noise are added to produce noisy signals with SNR uniformly selected from the range of 0-30 dB for each utterance. Hence a parallel set of 90 hours of noisy speech is created. The clean and the noisy speech are naturally aligned to each other at frame level. The test data was corrupted by noise types that are not observed during DNN training.

3.3.3 DNN Feature Compensated Result

The DNN feature compensated results are shown in the second half of Table. 3.2. It is observed that in all the 9 female core test conditions, the EER in most of the noisy cases is reduced. For average EER over all the 9 core conditions, the DNN based compensation reduce the EER by 19.99%, 7.01%, 25.93%, and 16.06% for the F16, Volvo, white, and babble noises, respectively at 0 dB. The improvement by DNN is small for Volvo noise and large for other noises. This is because the Volvo noise does not degrade the standard SV system's performance significantly. In addition, the improvement of DNN based feature compensation is higher at low SNR level and vice versa. In summary, the DNN achieves higher EER reduction at more distorted conditions.

3.3.4 MSE Measure of DNN Compensated Features

To understand why DNN feature compensation helps to improve the performance, the frame level MSE between the original clean features to both noisy and DNN enhanced features was calculated. Each feature frame is a 20 dimensional MFCC vector, the

Table 3.2: EER% of NIST SRE2010 female core test conditions(c1-c9) with uncompensated features versus compensated features. Speech features are corrupted by four types of noise at different SNR, then compensated by a DNN which is trained without assume prior knowledge of noise types and SNR levels

Test Cases	Noise Types												Clean
	F16			Volvo			White			Babble			
	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB	0 dB	10 dB	20 dB	
Uncompensated features													
c1	12.81	7.6	5.61	3.44	3.44	3.27	17.97	11.36	7.65	10.15	6.1	4.73	3.26
c2	22.35	16.51	12.87	7.88	7.37	7.22	28.27	21.86	16.77	18.28	13.43	10.24	7.63
c3	12.94	9.3	7.04	4.54	4.46	4.27	16.96	11.81	8.29	11.81	8.42	5.9	3.51
c4	15.08	9.03	6.31	3.33	3.16	3.07	20.6	13.33	9.55	10.38	6.57	4.5	3.15
c5	12.11	7.97	5.92	3.72	2.95	2.54	15.49	9.71	7.04	9.86	7.18	5.35	2.31
c6	20.77	16.94	12.04	8.2	7.65	6.24	26.23	19.13	13.12	18.04	15.3	11.65	4.37
c7	22.22	16.11	10.56	7.22	5.64	6.11	27.93	19.44	13.89	16.67	12.22	7.83	5
c8	6.15	2.99	1.69	1.15	1.12	0.71	8.39	3.95	2.24	4.73	2.23	1.12	1.12
c9	10.4	6.36	4.62	1.16	1.16	1.16	15.61	9.25	5.78	8.67	4.63	2.56	1.11
Average	**14.98**	**10.31**	**7.41**	**4.52**	**4.11**	**3.84**	**19.72**	**13.32**	**9.37**	**12.07**	**8.45**	**5.99**	**3.50**
DNN compensated features													
c1	10.41	7.82	6.45	4.04	3.96	3.91	14.03	10.62	8.77	9.20	6.28	4.56	4.04
c2	18.69	15.03	12.29	7.74	7.69	7.59	22.27	18.37	15.47	17.17	12.21	9.43	9.17
c3	11.68	8.04	6.16	4.02	3.64	3.91	12.19	8.79	7.41	8.79	6.16	4.90	5.40
c4	10.43	8.24	6.84	4.47	4.38	4.29	14.32	11.05	9.29	8.94	6.05	5.08	4.29
c5	9.86	7.32	5.95	3.10	2.54	2.25	11.35	8.17	5.92	7.94	5.63	4.51	3.38
c6	17.49	13.83	10.11	6.01	5.11	3.87	17.58	13.12	9.29	14.90	11.61	8.20	6.28
c7	14.46	10.56	8.33	5.56	5.56	5.12	19.45	13.44	10.00	12.22	9.45	7.22	5.00
c8	5.03	2.79	1.95	1.12	1.12	1.12	6.15	3.35	2.24	4.47	2.79	1.95	2.23
c9	9.83	6.36	4.05	1.74	1.73	1.74	14.11	9.83	6.94	7.51	3.47	2.89	1.73
Average	**11.99**	**8.89**	**6.90**	**4.20**	**3.97**	**3.76**	**14.60**	**10.75**	**8.37**	**10.13**	**7.07**	**5.42**	**4.61**
Rel. Impr. (%)	19.99	13.80	6.81	7.01	3.33	2.28	25.93	19.28	10.68	16.06	16.35	9.54	-

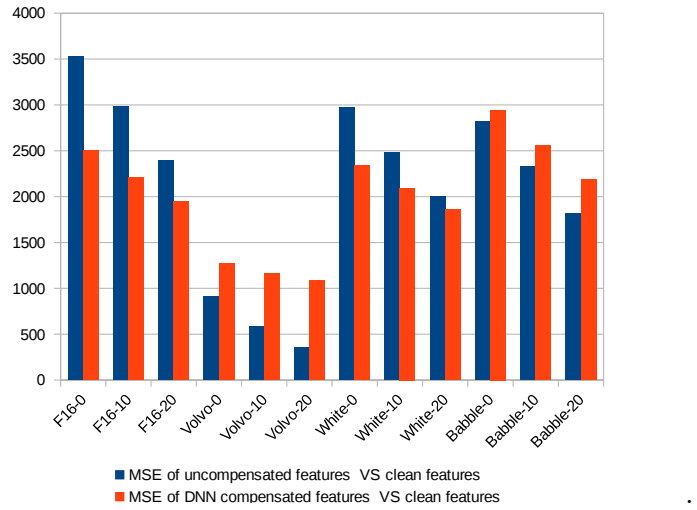


Figure 3.7: Frame level MSE measure (Y aixs) of uncompensated/compensated features under different noise conditions and SNRs (X aixs), F16-0 denotes noise type is F16 and SNR is 0 dB

squared error of each frame is accumulated and normalized by the number of frames for all the utterances of a given noise type. Fig. 3.7 shows the frame level MSE of the noise degraded (in blue) and DNN compensated (in red) features against clean features. DNN reduces the MSE of F16 and white noise, a large reduction is found in lower SNR. While MSE for the babble and Volvo noise increase for all level of SNR but yet the speaker verification performance is better on those noise types. This shows that the MSE measure of DNN compensated features does not correlate well with EER. This is possible as the SV using not only MFCC feature but also its dynamic feature like delta and delta delta as input. Even the MSE of raw feature becomes worse the dynamic feature may still preserve the important speaker specific feature. It could be investigated further on how to incorporate dynamic feature in DNN training/testing to improve the robustness of SV system.

3.4 Conclusion

This chapter described the proposed a DNN feature compensation method for noise robust speaker verification. The evaluation was conducted on the SRE 2010 benchmarking core test cases. Compare to the state-of-the-art speaker verification system based on iVector and PLDA, the SV system with DNN feature compensation obtained better performance. This indicates that DNN feature compensation is a possible way to improve the robustness of SV systems.

Chapter 4

Spoofing Speech Detection

Chapter 3 proposed an approach to address the noise robustness issues of SV system. This chapter examines another challenge faced by current SV systems, namely spoofing attack [170]. As discussed in chapter 2, spoofing attack in this work is defined as the purposeful generation of a target speaker's speech to fool the system. In this study, the scope of spoofing attacks is limited to synthetic speech. For example, an attacker may make use of existing text-to-speech [5] (TTS) or voice conversion [6] (VC) technologies to synthesize a target speaker's voice. To address this challenge, the thesis proposes to use a dedicated module based on long term magnitude and phase features as a front-end screening process to reject synthetic speech.

The chapter is organized as follows: section 4.1 presents the proposed spoofing speech detector module, section 4.2 discusses the experiments and evaluation criteria. Finally, section 4.3 concludes the chapter.

4.1 Spoofing Speech Detector Module

The internal blocks of dedicated module are shown in Fig. 4.1. It contains front-end features processing module and back-end decision making module. There are three distinct front-end modules to generate high dimensional features extracted from magnitude and phase spectrum. To capture long term temporal information, 31 successive frames are concatenated to form a super-vector. This super-vector is used as the input to back-end NN classifiers to predict the posterior probabilities of synthetic speech. If the posterior probability is greater than a predefined threshold, then the utterance will be rejected and not pass to the SV system.

The following sections describe the details of the front-end feature extractor and back-end module of the system.

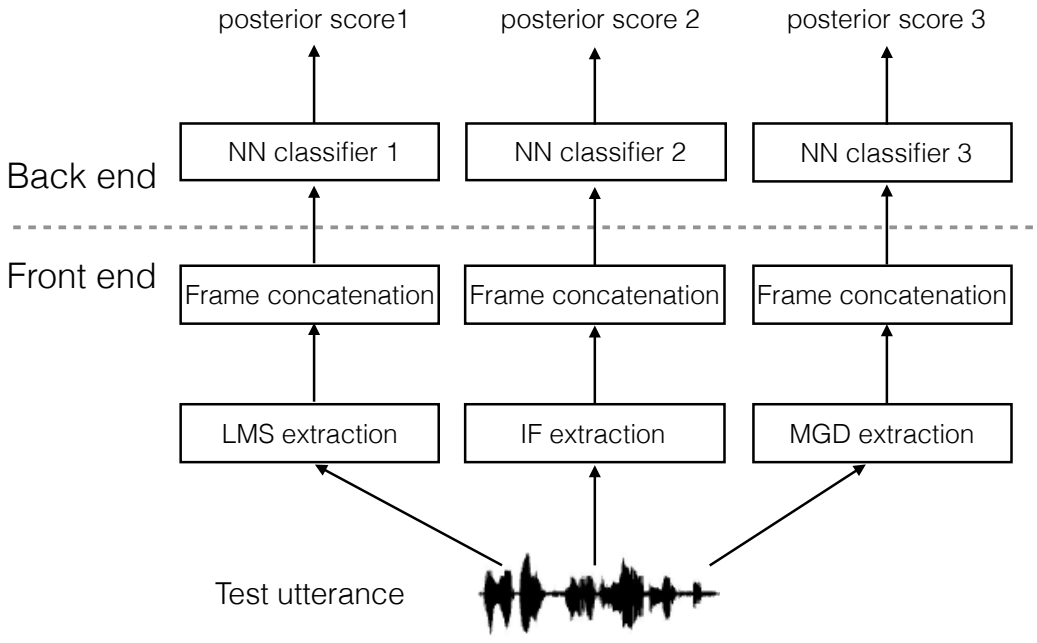


Figure 4.1: System architecture

4.1.1 Front-end Feature Extractors

An important part of spoofing speech detection is the selection of discriminative features that can precisely differentiate natural speech from synthetic speech. Most of the previous works use low dimensional features such as MFCC [33, 35–38]. This work considers three high dimensional features to detect spoofing speech, namely log magnitude spectrum (LMS), Instantaneous frequency derivative (IF), and Modified group delay (MGD).

4.1.1.1 Log Magnitude Spectrum (LMS)

The first feature used is the LMS. It is evaluated using short-time Fourier transform (STFT) on the speech signal. The speech is sampled at 16KHz and the STFT analysis window is 25ms with 15ms overlap. The Hamming window and DC offset is applied on each analysis frame. Given a speech signal at frame t, $\mathbf{x_t} \in R^{400}$, the complex spectrum is:

$$\mathrm{X}(t,\omega) = |\mathrm{X}(t,\omega)|e^{j\theta(t,\omega)}, \tag{4.1}$$

where, $|\mathrm{X}(t,\omega)|$ and $\theta(t,\omega)$ are the magnitude and phase spectrum at frame t and frequency ω, respectively. The FFT length is chosen to be 512. Hence, the LMS feature vectors is $\mathbf{X} = [log(|\mathrm{X}(t,0)|), \ldots, log(|\mathrm{X}(t,\pi)|)]^\top$

4.1.1.2 Instantaneous Frequency Derivative (IF)

The second feature used is the instantaneous frequency derivative (IF) [201]. It is the derivative of the phase spectrum $\theta(t,\omega)$ extracted in Equation 4.1,

$$\begin{aligned} IF(t,\omega) &= \frac{1}{2\pi}\frac{d\theta(t,\omega)}{dt} \\ &\approx \frac{1}{2\pi}(\theta(t,\omega) - \theta(t-1,\omega)). \end{aligned} \tag{4.2}$$

The range of $IF(t, \omega)$ is restricted to $[-\pi, \pi]$ by phase wraping. Unlike the original phase spectrum that hardly shows any patterns, there are clear patterns in the IF spectrum, making it possible to use them as features for synthetic speech detection.

4.1.1.3 Modified Group Delay (MGD)

The third feature used is the modified group delay (MGD) [201]. It is used to detect the non-linearity of the phase spectrum, which is generally absent in synthetic speech. Let $[x(1), ..., x(400)]$ represent the 400 samples in the analysis frame $\mathbf{x_t}$. The MGD feature at frame t is calculated as follows:

1) Compute the STFT of $x(n)$ and $nx(n)$ separately, denoted as $X(\omega)$ and $Y(\omega)$

2) Compute the smoothed spectrum of $|X(\omega)|$, denoted as $S(\omega)$

3) Compute the MGD as:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}} \tag{4.3}$$

4) Then, it is reshaped as:

$$\hat{\tau}(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|}|\tau(\omega)|^{\alpha} \tag{4.4}$$

Based on experimental results, the values of γ and α are set to 1.2 and 0.4 respectively.

4.1.2 Neural Network Based Synthetic Speech Classifier

These three features described above will be used to train three NN classifiers. The NN classifier is used as it has better capability to model high dimensional features as oppose to conventional GMM classifiers. The NN consists of 1 hidden layer with 2,048 sigmoid nodes. It is used to predict the posterior probability of the input super-vector

being synthetic speech. The posterior probabilities from a single utterance are averaged to produce a final posterior score.

4.2 Experimental Setup, Evaluation and Discussion

In the following sections, the spoofing attack corpus, baseline method and features, evaluation and discussion will be presented.

4.2.1 Spoofing Attack Corpus

CMU-ARCTIC database [41] is used to evaluated the proposed three features. This database contains 7 speakers, 4 with US English accent, 1 with Canadian English accent, 1 with Scottish English accent and 1 with Indian English accent. To guarantee the quality of converted speech, the 4 speakers with US English accent, (*bdl*, *rms*, *slt* and *clb*) are selected in this study. Each speaker has 1132 utterances which are sampled at 16 kHz, and the content of these utterances are same across all speakers.

There are four types of synthetic speech used in the study. Two of them are based on vocoder regeneration, and the other two are based on voice conversion techniques. The vocoder regenerated speech is considered as synthetic speech as most speech synthesis techniques use vocoders for speech analysis and synthesis.

1. **AHOcoder-syn**: The AHOcoder [202] is used to generate the synthetic speech from the natural speech without feature transformation.

2. **STRAIGHT-syn**: The STRAIGHT [203] is used to generate the synthetic speech from the natural speech without feature transformation.

3. **GMM-VC**: The JD-GMM with maximum likelihood parameter generation method

[204] using Mel-Cepstral Coefficients (MCC) as feature with a model of 64 Gaussian mixtures is used to convert natural speech to impostor's speech.

4. **CFW-VC**: The weighted correlation-based frequency warping [205] with GMM-based residual compensation on voiced frames are used for voice conversion.

For the two voice conversion methods, the STRAIGHT is used to extract 513-dimensional spectrum and $\log F_0$. The 25-dimensional Mel-Cepstral Coefficients (MCCs) and 15-dimensional linear spectrum frequencies (LSFs) are then extracted from the spectrum.

The above four methods are used to generate spoofing speech for classifiers training and testing. Two types of synthetic speech (STRAIGHT-syn and GMM-VC) and natural speech from speaker *bdl* and *clb* are used for classifiers training. All four types of synthetic speech and natural speech from the rest of the two speakers, *slt* and *rms*, are used for testing.

4.2.2 Baseline Method

In order to highlight the advantages of using high dimensional features for synthetic speech detection, two low dimensional features and GMM-based classifier are selected for comparison. As introduced in 2.4.2.2, most of previous works rely heavily on GMM-based classifier using low dimensional features with limited temporal information [33, 35–38]

4.2.2.1 Baseline Feature 1: Mel-frequency Cepstrum Coefficient (MFCC)

MFCC [206] has been widely used for ASR/SV and synthetic speech detection [33–35]. It is a compact representation of the LMS. To extract MFCC of a speech signal, the Mel-frequency scaled filter banks are first computed over the magnitude spectrum to generate 23 filter bank coefficients. Then the logarithm function is applied on these coefficients, followed by discrete cosine transform to reduce the dimensionality to 13. To include the

temporal information, the delta and the double-delta of MFCC are included to form a 39-dimensional feature.

4.2.2.2 Baseline Feature 2: Modified Group Delay Cepstral Coefficient (MGD-Cep)

Second baseline feature is MGD-Cep, which have been successfully applied for synthetic speech detection [162]. MGD-Cep is extracted in the same manner as the MFCC feature, as discussed in previous subsection, except that the MGD feature $\hat{\tau}(\omega)$ (see Equation 4.4) is used instead of the LMS.

4.2.2.3 Baseline Classifier

GMM [162] is used as baseline classifier for MFCC feature. Given $\lambda_{synthetic}$ and $\lambda_{natural}$ the GMM models for the synthetic and the natural speech respectively, the log likelihood ratio of the two classes is used for the detection:

$$\mathbf{\Gamma}(O) = \log p(O|\lambda_{synthetic}) - \log p(O|\lambda_{natural}), \tag{4.5}$$

where, O is the observation feature. The number of Gaussian components is set to 1024.

4.2.3 Evaluation and Discussion

To validate the performance of the proposed approaches, the effects of input window size and different systems will be examined in following subsections. Two evaluation metrics, the EER value and DET curve as described in Section 3.1.2, are used. The EER is obtained by selecting an operating point which gives the equal miss rate and false alarm rate. The DET curve, on the other hand, illustrates the miss rate against the false alarm rate of all operating points.

4.2.3.1 The Effects of Input Window Size

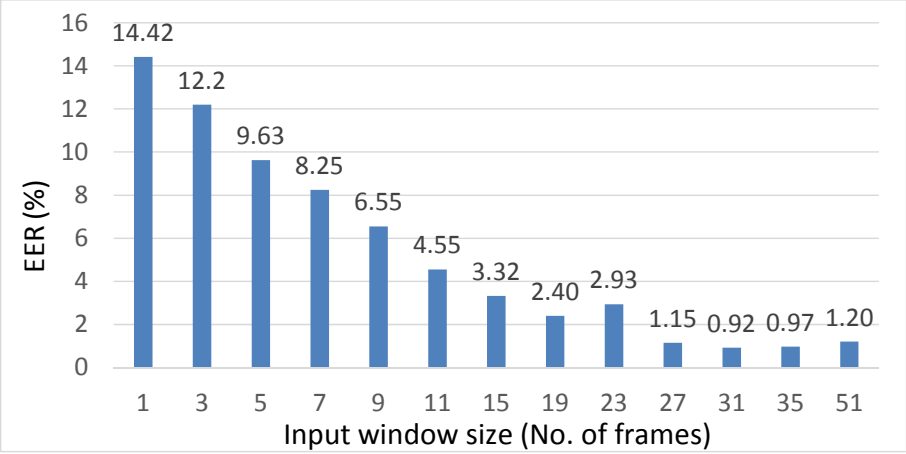


Figure 4.2: Equal error rate of LMS as a function of the input window size.

To examine the impact of the long term temporal information to the synthetic speech detection system, the input window size is increased from 1 to 51 frames. The NN is used for the classification while the LMS is used as the feature. The EER results of different window sizes are presented in Fig. 4.2.

As observed, the performance improves with respect to the window size, i.e. EER decreases from 14.42% to 0.92% as window size increases from 1 to 31 frames. The performance saturates after 31 frames. Longer window size, e.g. 51, degrades the performance slightly which may due to over-fitting. These observations confirm the effectiveness of long term temporal information in synthetic speech detection.

4.2.3.2 Comparison of Different Systems

High-dimensional and low-dimensional features based synthetic speech detection systems are compared and the EERs are shown in Table 4.1. The first four rows correspond to the systems (1-4) using low-dimensional features; following three rows depict the EERs of the systems (5-7) based on high-dimensional features; while the last four rows depict

Table 4.1: Equal error rate (EER, %) of detection performance of different systems, including 4 low-dimensional feature based system, 3 high-dimensional feature based system and the fusion of systems. Note that, the number after feature name indicates the window size of input feature.

No.	Systems	Natural against individual synthetic data				Natural against all synthetic
		STRAIGHT-syn	GMM-VC	AHOcoder-syn	CFW-VC	
Low dimensional features						
1	MFCC(1)-GMM	2.65	0.35	51.52	4.33	21.34
2	MFCC(1)-NN	17.51	0.29	51.57	33.28	30.85
3	MFCC(31)-NN	4.44	2.74	50.77	12.15	22.41
4	MGD-Cep(31)-NN	22.99	1.21	33.68	7.93	19.48
High dimensional features						
5	Magnitude(31)-NN	**0.09**	**0.00**	0.22	2.14	0.92
6	MGD(31)-NN	9.14	0.31	5.17	5.04	5.78
7	IF(31)-NN	0.77	0.04	**0.13**	**0.80**	0.54
Fusion						
	Fusion(5+6)	0.09	0.00	0.00	0.53	0.23
	Fusion(5+7)	0.04	0.00	0.04	0.35	0.13
	Fusion(6+7)	1.02	0.04	0.15	0.71	0.68
8	Fusion(5+6+7)	**0.04**	**0.00**	**0.00**	**0.18**	**0.09**

the EERs obtained by system fusion. The results include both EERs of natural speech against each type of synthetic speech and the EERs of natural speech against all four types of synthetic speech.

For systems 1-4, both performance of GMM and NN classifier are generally not satisfactory. While, the systems using high dimensional features, namely LMS (system 5), MGD (system 6), and IF (system 7), all perform significantly better than the systems using low dimensional features. This is due to the low dimensional feature is not capable in capturing as much detailed information as high dimensional feature. Comparing system 4 and 6, the only difference is the MGD-Cep used in system 4 is a smoothed and dimension reduced version of MGD used in system 6. However, the EER of system 6 (5.78%) is much lower than that of system 4 (19.48%). This is due to that compare to MGD-Cep, the MGD contains more detailed information of the input speech. Similarly,

by comparing system 5 and 3, the main difference is that the system 5 uses full detailed information of the magnitude spectrum while system 3 only uses the formant shape which is represented by the MFCC features. Again, the system with detailed information performs much better. Hence, it concludes that the detailed magnitude or phase information of speech signal is the key for achieving good performance in synthetic speech detection.

Also, this work examines the robustness of the features in anticipating unseen spoofed speech during testing. Noticed that the EERs in the first two columns are obtained from the experiment which synthetic speech is included in the training data; While, the EERs in the following two columns are obtained from the experiment which synthetic speech is unseen during the training. For low dimensional feature systems, the performance degrades significantly when synthetic speech is not included for the training. However, for high dimensional feature systems, only marginal degradation is observed. Thanks to the high dimensional feature which has embraced much more detailed discriminative features. A comparable performance can be achieved even the system is trained without any prior knowledge of the attacking speech.

The fusion of systems 5-7 is shown in the last 4 rows of Table 4.1. The fusions are obtained by averaging the scores of individual systems. The results show that although system 6 performs much worse than system 5 and 7, it provides complementary information and the best fusion results are obtained by fusing all systems together.

The detection error trade-off (DET) curves of our eight systems are presented in Fig. 4.3. The MGD-Cep feature with long term information reduces the EER from 21.34% of MFCC(1)-GMM to 19.48%; by using high-dimensional IF feature with long term information, the EER could be reduced to 0.54% and the combination of three high-dimensional features with long term information further reduces the EER to 0.09%.

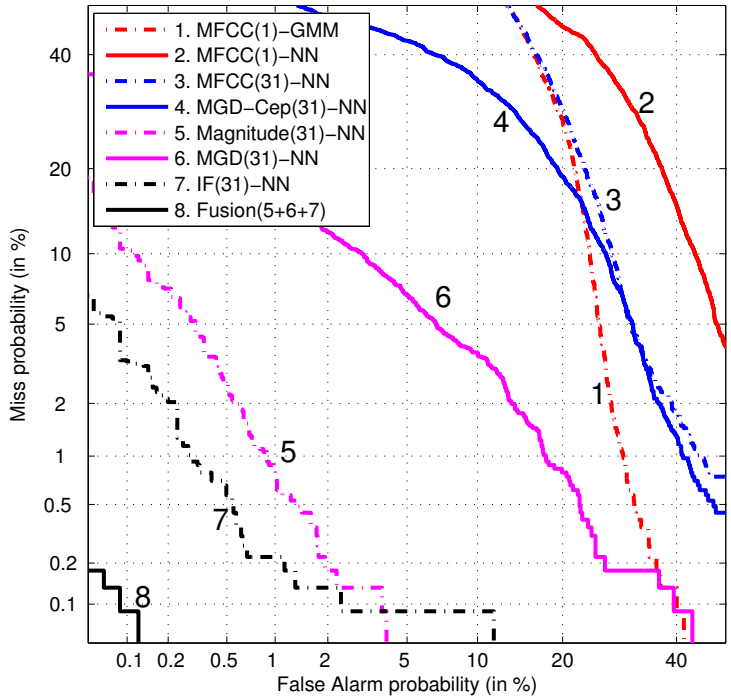


Figure 4.3: DET curve of synthetic detection performance of different systems.

4.3 Conclusion

This chapter conducted two studies of spoofing speech detection, including the use of long term temporal information and high dimensional feature vectors that contain detailed magnitude and phase information of the input speech signal. Results show that both long term temporal information and detailed speech information are vital for accurate detection of spoofing speech and thus, protecting the SV system from spoofing attack. In addition, the LMS feature is shown to be complementary to the phase derived features, such as MGD and IF.

Currently, the proposed system achieved higher performance in detecting synthetic speech using microphone recorded speech in clean environment. However, many real applications may involve noisy environments, which introduce distortion to natural speech.

Thus, the robust detection of synthetic speech in noisy conditions can be studied in future.

The proposed method has been used to compete in the ASVspoof 2015 challenge and has achieved the best results in the closed set condition among 16 other teams worldwide [186].

Chapter 5

Conclusions and Future Work

This dissertation has focused on two major robustness issues of SV system, namely noise robustness and spoofing attacks. Current speaker verification systems do not perform well under unseen noise conditions as well as are highly vulnerable to malicious spoofing attacks. This thesis has proposed approaches to address both of these issues. Section 5.1 first summaries the proposed methods and their evaluation. Finally, Section 5.2 discusses some of the future directions that can be explored.

5.1 Contributions

5.1.1 DNN Based Feature Compensation

In Chapter 3, the issue of current SV systems requiring prior knowledge of noise type and condition is examined. This thesis proposed a DNN based feature compensation approach [39] to transform the noisy input features into clean features for noise robust capability. The experimental results evaluated using the SRE 2010 showed that the proposed method generalizes well under various unseen noise conditions and reduced relative EER by 2-26%.

5.1.2 High Dimensional and Long Term Features for Spoofing Speech Detection

This thesis has also proposed the use of high dimensional features and long term temporal information as input features to detect spoofing speech. The high dimensional features are log magnitude spectrum, high-dimensional instantaneous frequency (IF), and modified group delay (MGD). The NN classifier trained with these features has shown an EER improvement from 19.48-30.85% to 0.54-5.78% over the baseline system, evaluate on the CMU-ARCTIC database [41]. This method has been used to compete in the ASVspoof 2015 challenge and has achieved the best results in the closed set condition among 16 other teams worldwide [186].

5.2 Future Work

This thesis can be further extended in many aspects. e.g, The proposed DNN based feature compensation method (Chapter 3) can be evaluated by using other types of robust speaker-specific features. Also, the proposed spoofing speech detection system (Chapter 4) should be examined under noisy conditions for many real world applications.

In addition, deep learning approach can be used to learn the speaker-specific features as well as more discriminative features to improve the performance of both SV system and spoofing detection system.

List of Publications

(i) Steven Du, Xiong Xiao, Eng Siong Chng, "DNN Feature Compensation For Noise Robust Speaker Verification", in proceedings *China Summit and International Conference on Signal and Information Processing (ChinaSIP 2015)*, 2015, ChengDu, China.

(ii) Xiaohai Tian , Steven Du , Xiong Xiao , Haihua Xu , Eng Siong Chng, and Haizhou Li, "Detecting Synthetic Speech Using Long Term Magnitude And Phase Information", in proceedings *China Summit and International Conference on Signal and Information Processing (ChinaSIP 2015)*, 2015, ChengDu, China.

(iii) Xiong Xiao, Xiaohai Tian, Steven Du , Haihua Xu , Eng Siong Chng, Haizhou Li, "Spoofing Speech Detection Using High Dimensional Magnitude and Phase Features: the NTU Approach for ASVspoof 2015 Challenge", in proceedings *INTERSPEECH 2015, 16th Annual Conference of International Speech Communication Association (Interspeech 2015)*, 2015, Dresden, Germany

References

[1] D. Garcia-Romero, *Robust Speaker Recognition based on Latent Variable Models*. PhD thesis, University of Maryland at College Park, 2012.

[2] J. Gonzalez-Rodriguez, "Evaluating automatic speaker recognition systems: an overview of the NIST speaker recognition evaluations (1996-2014)," *Loquens*, vol. 1, no. 1, p. e007, 2014.

[3] B. T. Meyer and B. Kollmeier, "Robustness of spectro-temporal features against intrinsic and extrinsic variations in automatic speech recognition," *Speech Communication*, vol. 53, no. 5, pp. 753–767, 2011.

[4] F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 3068–3072, IEEE, 2013.

[5] J. Allen, M. S. Hunnicutt, D. H. Klatt, R. C. Armstrong, and D. B. Pisoni, *From text to speech: The MITalk system*. Cambridge University Press, 1987.

[6] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, "Voice conversion through vector quantization," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 655–658, IEEE, 1988.

[7] S. O. Sadjadi, T. Hasan, and J. H. Hansen, "Mean hilbert envelope coefficients (MHEC) for robust speaker recognition.," in *INTERSPEECH*, 2012.

[8] J. Wang and M. T. Johnson, "Residual phase cepstrum coefficients with application to cross-lingual speaker verification," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[9] J. Wang and M. T. Johnson, "Physiologically-motivated feature extraction for speaker identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 1690–1694, IEEE, 2014.

[10] P. J. Moreno, B. Raj, and R. M. Stern, "A vector taylor series approach for environment-independent speech recognition," in *Acoustics, Speech, and Signal*

Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on, vol. 2, pp. 733–736, IEEE, 1996.

[11] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171–185, 1995.

[12] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 245–257, 1994.

[13] T. Matsui, T. Kanno, and S. Furui, "Speaker recognition using HMM composition in noisy environments," *Computer Speech & Language*, vol. 10, no. 2, pp. 107–116, 1996.

[14] K. S. Rao and S. Sarkar, *Robust Speaker Recognition in Noisy Environments*. Springer, 2014.

[15] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6645–6649, IEEE, 2013.

[16] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[17] L. Deng and D. Yu, "Deep learning: Methods and applications," Tech. Rep. MSR-TR-2014-21, May 2014.

[18] D. Yu and L. Deng, *Automatic Speech Recognition - A Deep Learning Approach*. Springer, October 2014.

[19] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7398–7402, IEEE, 2013.

[20] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *Signal Processing Letters, IEEE*, vol. 21, no. 1, pp. 65–68, 2014.

[21] X. Xiao, S. Zhao, D. H. H. Nguyen, X. Zhong, D. L. Jones, E. S. Chng, and H. Li, "The ntu-adsc systems for reverberation challenge 2014,"

[22] Y. Konig, L. Heck, M. Weintraub, K. Sonmez, *et al.*, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition," in *Proc. RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications*, pp. 72–75, 1998.

[23] H. Lee, P. Pham, Y. Largman, and A. Y. Ng, "Unsupervised feature learning for audio classification using convolutional deep belief networks," in *Advances in neural information processing systems*, pp. 1096–1104, 2009.

[24] K. Chen and A. Salman, "Extracting speaker-specific information with a regularized siamese deep network," in *Advances in Neural Information Processing Systems*, pp. 298–306, 2011.

[25] S. Authenticity, "Speaker recognition using neural networks and conventional classifiers," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. 2, no. 1 PART II, 1994.

[26] Y. Bennani and P. Gallinari, "Neural networks for discrimination and modelization of speakers," *Speech Communication*, vol. 17, no. 1, pp. 159–175, 1995.

[27] M. Senoussaoui, N. Dehak, P. Kenny, R. Dehak, and P. Dumouchel, "First attempt of boltzmann machines for speaker verification," in *Proceedings of the Odyssey: The Speaker and Language Recognition Workshop*, 2012.

[28] V. Vasilakakis, S. Cumani, and P. Laface, "Speaker recognition by means of deep belief networks," *Proc. Biometric Technologies in Forensic Science*, 2013.

[29] T. Yamada, L. Wang, and A. Kai, "Improvement of distant-talking speaker identification using bottleneck features of dnn.," in *INTERSPEECH*, pp. 3661–3664, 2013.

[30] O. Ghahabi and J. Hernando, "i-vector modeling with deep belief networks for multi-session speaker recognition," *network*, vol. 20, p. 13, 2014.

[31] Y. Lei, L. Ferrer, M. McLaren, and N. Scheffer, "A deep neural network speaker verification system targeting microphone speech," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[32] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4052–4056, IEEE, 2014.

[33] T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4401–4404, IEEE, 2012.

[34] Z. Wu, A. Larcher, K.-A. Lee, E. Chng, T. Kinnunen, and H. Li, "Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints.," in *INTERSPEECH*, pp. 950–954, 2013.

[35] J.-F. Bonastre, D. Matrouf, and C. Fredouille, "Artificial impostor voice transformation effects on false acceptance rates.," in *INTERSPEECH*, pp. 2053–2056, 2007.

[36] T. Masuko, T. Hitotsumatsu, K. Tokuda, and T. Kobayashi, "On the security of HMM-based speaker verification systems against imposture using synthetic speech.," in *Eurospeech*, 1999.

[37] T. Masuko, K. Tokuda, and T. Kobayashi, "Imposture using synthetic speech against speaker verification based on spectrum and pitch.," in *INTERSPEECH*, pp. 302–305, Citeseer, 2000.

[38] T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system.," in *INTERSPEECH*, pp. 759–762, 2001.

[39] S. Du, X. Xiong, and E. S. Chng, "DNN feature compensation for noise robust speaker verification," IEEE, 2015.

[40] X. Tian, S. Du, X. Xiao, X. Haihua, E. S. Chng, and H. Li, "Detecting synthetic speech using long term magnitude and phase information," IEEE, 2015.

[41] J. Kominek and A. W. Black, "The CMU arctic speech databases," in *Fifth ISCA Workshop on Speech Synthesis*, 2004.

[42] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[43] R. H. Bolt, F. S. Cooper, E. E. David Jr, P. B. Denes, J. M. Pickett, and K. N. Stevens, "Speaker identification by speech spectrograms: some further observations," *The Journal of the Acoustical Society of America*, vol. 54, no. 2, pp. 531–534, 1973.

[44] S. Pruzansky, "Pattern-matching procedure for automatic talker recognition," *The Journal of the Acoustical Society of America*, vol. 35, no. 3, pp. 354–358, 1963.

[45] S. Furui, "40 years of progress in automatic speaker recognition," in *Advances in Biometrics*, pp. 1050–1059, Springer, 2009.

[46] H. Beigi, *Fundamentals of speaker recognition*. Springer Science & Business Media, 2011.

[47] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP journal on applied signal processing*, vol. 2004, pp. 430–451, 2004.

[48] J. P. Campbell Jr, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[49] S. Furui, "An overview of speaker recognition technology," in *Automatic speech and speaker recognition*, pp. 31–56, Springer, 1996.

[50] S. Furui, "Recent advances in speaker recognition," in *Audio-and Video-based Biometric Person Authentication*, pp. 235–252, Springer, 1997.

[51] S. Furui, "Cepstral analysis technique for automatic speaker verification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 2, pp. 254–272, 1981.

[52] H. Hermansky, B. Hanson, and H. Wakita, "Perceptually based linear predictive analysis of speech," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, vol. 10, pp. 509–512, IEEE, 1985.

[53] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, 1980.

[54] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, *et al.*, "The supersid project: Exploiting high-level information for high-accuracy speaker recognition," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 4, pp. IV–784, IEEE, 2003.

[55] M.-W. Mak and H.-B. Yu, "Robust voice activity detection for interview speech in nist speaker recognition evaluation," *Proc. APSIPA ASC 2010*, 2010.

[56] S. Furui, "Cepstral analysis technique for automatic speaker verification," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 29, no. 2, pp. 254–272, 1981.

[57] L. R. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, vol. 14. PTR Prentice Hall Englewood Cliffs, 1993.

[58] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[59] R. C. Rose and D. A. Reynolds, "Text independent speaker identification using automatic acoustic segmentation," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 293–296, IEEE, 1990.

[60] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 1, pp. 72–83, 1995.

[61] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[62] W. M. Campbell, "Generalized linear discriminant sequence kernels for speaker recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–161, IEEE, 2002.

[63] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, IEEE, 2006.

[64] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[65] U. Ben Simon, I. Lapidot, and H. Guterman, "Comparison between normalizations for svm-gmm supervectors speaker verification," in *Electrical and Electronics Engineers in Israel (IEEEI), 2010 IEEE 26th Convention of*, pp. 000621–000625, IEEE, 2010.

[66] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and audio processing, ieee transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[67] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435–1447, 2007.

[68] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of interspeaker variability in speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 5, pp. 980–988, 2008.

[69] A. Solomonoff, W. M. Campbell, and C. Quillen, "Nuisance attribute projection," *Speech Communication*, 2007.

[70] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition.," in *Interspeech*, 2006.

[71] W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "Svm based speaker verification using a GMM supervector kernel and NAP variability compensation," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, IEEE, 2006.

[72] R. J. Vogt, S. Kajarekar, and S. Sridharan, "Discriminant NAP for SVM speaker recognition," 2008.

[73] A. O. Hatch, S. S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition.," in *Interspeech*, 2006.

[74] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques.," in *Odyssey*, p. 15, 2010.

[75] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Institute for Signal and information Processing*, 1998.

[76] B. Scholkopft and K.-R. Mullert, "Fisher discriminant analysis with kernels," in *Proceedings of the 1999 IEEE Signal Processing Society Workshop Neural Networks for Signal Processing IX, Madison, WI, USA*, pp. 23–25, 1999.

[77] N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification.," in *Interspeech*, vol. 9, pp. 1559–1562, 2009.

[78] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems.," in *Interspeech*, pp. 249–252, 2011.

[79] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pp. 1–8, IEEE, 2007.

[80] P. Kenny, "Bayesian speaker verification with heavy-tailed priors.," in *Odyssey*, p. 14, 2010.

[81] S. Lyu and E. P. Simoncelli, "Nonlinear extraction of independent components of natural images using radial gaussianization," *Neural computation*, vol. 21, no. 6, pp. 1485–1519, 2009.

[82] M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, *et al.*, "Automatic speech recognition and speech variability: A review," *Speech Communication*, vol. 49, no. 10, pp. 763–786, 2007.

[83] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "The effect of noise on modern automatic speaker recognition systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4249–4252, IEEE, 2012.

[84] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6788–6791, IEEE, 2013.

[85] D. Garcia-Romero, X. Zhou, and C. Y. Espy-Wilson, "Multicondition training of gaussian PLDA models in i-vector space for noise and reverberation robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4257–4260, IEEE, 2012.

[86] G. Liu, Y. Lei, and J. H. Hansen, "Robust feature front-end for speaker identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4233–4236, IEEE, 2012.

[87] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[88] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A comparative study of robust linear predictive analysis methods with applications to speaker identification," *Speech and Audio Processing, IEEE Transactions on*, vol. 3, no. 2, pp. 117–125, 1995.

[89] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions.," in *INTERSPEECH*, pp. 1477–1480, 2010.

[90] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *Signal Processing Letters, IEEE*, vol. 17, no. 6, pp. 599–602, 2010.

[91] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertas, "Regularization of all-pole models for speaker verification under additive noise," in *Odyssey, The Speaker and Language Recognition Workshop*, 2012.

[92] S. Ganapathy, S. Thomas, and H. Hermansky, "Feature extraction using 2-d autoregressive models for speaker recognition," *ISCA Speaker Odyssey*, 2012.

[93] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, J. Sandberg, and M. Hansson-Sandsten, "Comparing spectrum estimators in speaker verification under additive noise degradation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4769–4772, IEEE, 2012.

[94] T. Kinnunen, R. Saeidi, J. Sandberg, and M. Hansson-Sandsten, "What else is new than the hamming window robust MFCCs for speaker recognition via multitapering.," in *INTERSPEECH*, pp. 2734–2737, 2010.

[95] T. Kinnunen, R. Saeidi, F. Sedlák, K. A. Lee, J. Sandberg, M. Hansson-Sandsten, and H. Li, "Low-variance multitaper MFCC features: A case study in robust speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 7, pp. 1990–2001, 2012.

[96] M. J. Alam, P. Kenny, and D. O Shaughnessy, "On the use of asymmetric-shaped tapers for speaker verification using i-vectors," in *Proc. Odyssey*, vol. 2012, 2012.

[97] B. Jawerth and W. Sweldens, "An overview of wavelet based multiresolution analyses," *SIAM review*, vol. 36, no. 3, pp. 377–412, 1994.

[98] C.-T. Hsieh, E. Lai, and Y.-C. Wang, "Robust speaker identification system based on wavelet transform and gaussian mixture model," *J. Inf. Sci. Eng.*, vol. 19, no. 2, pp. 267–282, 2003.

[99] M. Alhanjouri, M. A. Lubbad, and M. Z. Alkurdi, "Robust speaker identification using denoised wave atom and GMM," *International Journal of Computer Applications*, vol. 67, no. 5, pp. 17–23, 2013.

[100] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–277, IEEE, 2007.

[101] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 16, no. 6, pp. 1097–1111, 2008.

[102] M. S. Deshpande and R. S. Holambe, "Am-fm based robust speaker identification in babble noise," *environments*, vol. 6, no. 10, p. 19, 2011.

[103] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds, "Modeling of the glottal flow derivative waveform with application to speaker identification," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 5, pp. 569–586, 1999.

[104] B. Bozkurt and L. Couvreur, "On the use of phase information for speech recognition," in *Proc. EUSIPCO*, pp. 2–5, 2005.

[105] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," *Speech communication*, vol. 49, no. 3, pp. 159–176, 2007.

[106] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[107] E. Loweimi, S. M. Ahadi, and T. Drugman, "A new phase-based feature representation for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7155–7159, IEEE, 2013.

[108] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde, "Significance of the modified group delay feature in speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 190–202, 2007.

[109] P. Rajan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase based features for speaker recognition," in *proceedings of INTERSPEECH*, no. LIDIAP-CONF-2009-046, 2009.

[110] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, J. Sandberg, and M. Hansson-Sandsten, "Comparing spectrum estimators in speaker verification under additive noise degradation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4769–4772, IEEE, 2012.

[111] D. Hardt and K. Fellbaum, "Spectral subtraction and RASTA-filtering in text-dependent HMM-based speaker verification," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, pp. 867–870, IEEE, 1997.

[112] K.-H. Yuo, T.-H. Hwang, and H.-C. Wang, "Combination of autocorrelation-based features and projection measure technique for speaker identification," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 565–574, 2005.

[113] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," 2001.

[114] J. Ortega-García and J. González-Rodríguez, "Overview of speech enhancement techniques for automatic speaker recognition," in *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on*, vol. 2, pp. 929–932, IEEE, 1996.

[115] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1, pp. 121–124, IEEE, 1998.

[116] A. Moreno-Daniel, J. A. Nolazco-Flores, T. Wada, and B. Juang, "Acoustic model enhancement: An adaptation technique for speaker verification under noisy environments," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–289, IEEE, 2007.

[117] S. O. Sadjadi and J. H. Hansen, "Assessment of single-channel speech enhancement techniques for speaker identification under mismatched conditions.," in *INTERSPEECH*, pp. 2138–2141, 2010.

[118] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *Signal Processing Letters, IEEE*, vol. 17, no. 6, pp. 599–602, 2010.

[119] T. G. Stockham Jr, T. M. Cannon, and R. B. Ingebretsen, "Blind deconvolution through digital signal processing," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 678–692, 1975.

[120] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 33, no. 2, pp. 443–445, 1985.

[121] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 27, no. 2, pp. 113–120, 1979.

[122] T. Ganchev, I. Potamitis, N. Fakotakis, and G. Kokkinakis, "Text-independent speaker verification for real fast-varying noisy environments," *International Journal of Speech Technology*, vol. 7, no. 4, pp. 281–292, 2004.

[123] E. Verteletskaya and B. Simak, "Enhanced spectral subtraction method for noise reduction with minimal speech distortion," 2010.

[124] G. Pradhan, B. Haris, S. Prasanna, and R. Sinha, "Speaker verification in sensor and acoustic environment mismatch conditions," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 381–392, 2012.

[125] A. Drygajlo and M. El-Maliki, "Speaker verification in noisy environments with combined spectral subtraction and missing feature theory," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 1, pp. 121–124, IEEE, 1998.

[126] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 5, pp. 1711–1723, 2007.

[127] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4, pp. IV–277, IEEE, 2007.

[128] D. Pullella, M. Kuhne, and R. Togneri, "Robust speaker identification using combined feature selection and missing data recognition," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4833–4836, IEEE, 2008.

[129] T. May, S. van de Par, and A. Kohlrausch, "Noise-robust speaker recognition combining missing data techniques and universal background modeling," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 108–121, 2012.

[130] M. T. Padilla, T. F. Quatieri, and D. A. Reynolds, "Missing feature theory with soft spectral subtraction for speaker verification.," in *Interspeech*, 2006.

[131] D. Ribas, J. A. Villalba, E. Lleida, and J. R. Calvo, "Speaker verification in noisy environment using missing feature approach," in *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pp. 220–227, Springer, 2010.

[132] Y. Shao and D. Wang, "Robust speaker recognition using binary time-frequency masks," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, IEEE, 2006.

[133] Y. Shao and D. Wang, "Robust speaker identification using auditory features and computational auditory scene analysis," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 1589–1592, IEEE, 2008.

[134] X. Zhao, Y. Shao, and D. Wang, "CASA-based robust speaker identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 5, pp. 1608–1616, 2012.

[135] C. Tzagkarakis and A. Mouchtaris, "Robust speaker identification using matrix completion under a missing data imputation framework," in *Proc. Workshop on Sig. Proc. with Adaptive Sparse Structured Representations (SPARS-13), Lausanne, Switzerland*, 2013.

[136] C. Tzagkarakis and A. Mouchtaris, "Reconstruction of missing features based on a low-rank assumption for robust speaker identification," in *Information, Intelligence, Systems and Applications, IISA 2014, The 5th International Conference on*, pp. 432–437, IEEE, 2014.

[137] Y. Mami and D. Charlet, "Speaker identification by anchor models with PCA/LDA post-processing," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1, pp. I–180, IEEE, 2003.

[138] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors.," in *ICASSP*, pp. 4029–4032, Citeseer, 2008.

[139] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2067–2080, 2011.

[140] A. Hurmalainen, J. F. Gemmeke, and T. Virtanen, "Modelling non-stationary noise with spectral factorisation in automatic speech recognition," *Computer Speech & Language*, vol. 27, no. 3, pp. 763–779, 2013.

[141] Q. Wu, L.-Q. Zhang, and G.-C. Shi, "Robust feature extraction for speaker recognition based on constrained nonnegative tensor factorization," *Journal of computer science and technology*, vol. 25, no. 4, pp. 783–792, 2010.

[142] J. Wang, D. Wang, T. F. Zheng, and F. Bie, "DNN-based discriminative scoring for speaker recognition based on i-vector,"

[143] E. Variani, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4052–4056, IEEE, 2014.

[144] S. R. Krothapalli, J. Yadav, S. Sarkar, S. G. Koolagudi, and A. K. Vuppala, "Neural network based feature transformation for emotion independent speaker identification," *International Journal of Speech Technology*, vol. 15, no. 3, pp. 335–349, 2012.

[145] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 4, pp. 836–845, 2014.

[146] K. Chen and A. Salman, "Extracting speaker-specific information with a regularized siamese deep network," in *Advances in Neural Information Processing Systems*, pp. 298–306, 2011.

[147] T. T. Kristjansson and B. J. Frey, "Accounting for uncertainity in observations: a new paradigm for robust automatic speech recognition," in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1, pp. I–61, IEEE, 2002.

[148] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and audio processing, ieee transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[149] T. Hasan and J. H. Hansen, "Acoustic factor analysis for robust speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 842–853, 2013.

[150] T. Hasan and J. H. Hansen, "Maximum likelihood acoustic factor analysis models for robust speaker verification in noise," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 2, pp. 381–391, 2014.

[151] Y. Lei, L. Burget, and N. Scheffer, "A noise robust i-vector extractor using vector taylor series for speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 6788–6791, IEEE, 2013.

[152] Y. Lei, M. McLaren, L. Ferrer, and N. Scheffer, "Simplified vts-based i-vector extraction in noise-robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4037–4041, IEEE, 2014.

[153] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, and H. Li, "Spoofing and countermeasures for speaker verification: a survey," *Speech Communication*, vol. 66, pp. 130–153, 2015.

[154] M. Faundez-Zanuy, M. Hagmüller, and G. Kubin, "Speaker verification security improvement by means of speech watermarking," *Speech communication*, vol. 48, no. 12, pp. 1608–1619, 2006.

[155] J. Villalba and E. Lleida, "Preventing replay attacks on speaker verification systems," in *Security Technology (ICCST), 2011 IEEE International Carnahan Conference on*, pp. 1–8, IEEE, 2011.

[156] J. Villalba and E. Lleida, "Detecting replay attacks from far-field recordings on speaker verification systems," in *Biometrics and ID Management*, pp. 274–285, Springer, 2011.

[157] F. Alegre, A. Janicki, and N. Evans, "Re-assessing the threat of replay spoofing attacks against automatic speaker verification," in *Biometrics Special Interest Group (BIOSIG), 2014 International Conference of the*, pp. 1–6, IEEE, 2014.

[158] P. L. De Leon, I. Hernaez, I. Saratxaga, M. Pucher, and J. Yamagishi, "Detection of synthetic speech for the problem of imposture," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pp. 4844–4847, IEEE, 2011.

[159] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 8, pp. 2280–2290, 2012.

[160] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7234–7238, IEEE, 2013.

[161] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "A crossvocoder study of speaker independent synthetic speech detection using phase information," in *Proc. Interspeech*, pp. 1663–1667, 2014.

[162] Z. Wu, C. E. Siong, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition.," in *INTERSPEECH*, 2012.

[163] F. Alegre, R. Vipperla, A. Amehraye, and N. Evans, "A new speaker verification spoofing countermeasure based on local binary patterns," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon: France (2013)*, p. 5p, 2013.

[164] M. Blomberg, D. Elenius, and E. Zetterholm, "Speaker verification scores and acoustic analysis of a professional impersonator," in *Proc. FONETIK*, pp. 84–87, 2004.

[165] E. Zetterholm, *Voice imitation: a phonetic study of perceptual illusions and acoustic success*, vol. 44. Lund University, 2003.

[166] R. G. Hautamäki, T. Kinnunen, V. Hautamäki, and A.-M. Laukkanen, "Comparison of human listeners and speaker verification systems using voice mimicry data," *TARGET*, vol. 4000, p. 5000, 2014.

[167] Y. W. Lau, M. Wagner, and D. Tran, "Vulnerability of speaker verification to voice mimicking," in *Intelligent Multimedia, Video and Speech Processing, 2004. Proceedings of 2004 International Symposium on*, pp. 145–148, IEEE, 2004.

[168] Z. Wu, S. Gao, E. S. Cling, and H. Li, "A study on replay attack and anti-spoofing for text-dependent speaker verification," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pp. 1–5, IEEE, 2014.

[169] J. Lindberg, M. Blomberg, *et al.*, "Vulnerability in speaker verification-a study of technical impostor techniques.," in *Eurospeech*, vol. 99, pp. 1211–1214, 1999.

[170] J. Villalba and E. Lleida, "Speaker verification performance degradation against spoofing and tampering attacks," in *FALA workshop*, pp. 131–134, 2010.

[171] Z.-F. Wang, G. Wei, and Q.-H. He, "Channel pattern noise based playback attack detection algorithm for speaker recognition," in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*, vol. 4, pp. 1708–1713, IEEE, 2011.

[172] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1, pp. 373–376, IEEE, 1996.

[173] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[174] Y. Qian, F. K. Soong, and Z.-J. Yan, "A unified trajectory tiling approach to high quality speech rendering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 2, pp. 280–290, 2013.

[175] E. Allefs, "Speech synthesis apparatus with personalized speech segments," Sept. 12 2003. US Patent App. 10/529,976.

[176] J.-c. Junqua, F. Perronnin, R. Kuhn, and P. Nguyen, "Voice personalization of speech synthesizer," Nov. 29 2005. US Patent 6,970,820.

[177] N. Kibre, S. Pearson, B. Hanson, and J.-c. Junqua, "Text selection and recording by feedback and adaptation for development of personalized text-to-speech systems," Sept. 14 2004. US Patent 6,792,407.

[178] V. Shchemelinin and K. Simonchik, "Examining vulnerability of voice verification systems to spoofing attacks by means of a TTS system," in *Speech and Computer*, pp. 132–137, Springer, 2013.

[179] K. Simonchik and V. Shchemelinin, "STC spoofing database for text-dependent speaker recognition evaluation," in *Spoken Language Technologies for Under-Resourced Languages*, 2014.

[180] D. Matrouf, J.-F. Bonastre, and C. Fredouille, "Effect of speech transformation on impostor acceptance," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, vol. 1, pp. I–I, IEEE, 2006.

[181] Z. Wu and H. Li, "Voice conversion and spoofing attack on speaker verification systems," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, pp. 1–9, IEEE, 2013.

[182] B. L. Pellom and J. H. Hansen, "An experimental study of speaker verification sensitivity to computer voice-altered imposters," in *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, vol. 2, pp. 837–840, IEEE, 1999.

[183] P. Perrot, G. Aversano, R. Blouet, M. Charbit, and G. Chollet, "Voice forgery using ALISP: Indexation in a client memory.," in *ICASSP (1)*, pp. 17–20, 2005.

[184] D. Matrouf, J.-F. Bonastre, and J.-P. Costa, "Effect of impostor speech transformation on automatic speaker recognition," *Biometrics on the Internet*, p. 37, 2005.

[185] Z. Wu, T. Kinnunen, N. Evans, and J. Yamagishi, "ASVspoof 2015: Automatic speaker verification spoofing and countermeasures challenge evaluation plan," *Training*, vol. 10, no. 15, p. 3750, 2014.

[186] X. Xiong, T. Xiaohai, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: the NTU approach for ASVspoof 2015 challenge," IEEE, 2015.

[187] P. L. De Leon, M. Pucher, and J. Yamagishi, "Evaluation of the vulnerability of speaker verification to synthetic speech," 2010.

[188] Q. Jin, A. R. Toth, A. W. Black, and T. Schultz, "Is voice transformation a threat to speaker identification?," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 4845–4848, IEEE, 2008.

[189] Q. Jin, A. R. Toth, T. Schultz, and A. W. Black, "Voice convergin: Speaker de-identification by voice transformation," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pp. 3909–3912, IEEE, 2009.

[190] A. Ogihara, U. Hitoshi, and A. Shiozaki, "Discrimination method of synthetic speech using pitch frequency against synthetic speech falsification," *IEICE transactions on fundamentals of electronics, communications and computer sciences*, vol. 88, no. 1, pp. 280–286, 2005.

[191] P. L. De Leon, B. Stewart, and J. Yamagishi, "Synthetic speech discrimination using pitch pattern statistics derived from image analysis.," in *INTERSPEECH*, 2012.

[192] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: a resource for the next generations of speech-to-text.," in *LREC*, vol. 4, pp. 69–71, 2004.

[193] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon Technical Report N*, vol. 93, p. 27403, 1993.

[194] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: A phonetically balanced, continuous speech, telephone bandwidth speech database," in *Acoustics, Speech, and Signal Processing, 1990. ICASSP-90., 1990 International Conference on*, pp. 109–112, IEEE, 1990.

[195] "Linguistic data consortium (LDC), http://www.ldc.upenn.edu,"

[196] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwarz, *et al.*, "The kaldi speech recognition toolkit," 2011.

[197] A. F. Martin and C. S. Greenberg, "The NIST 2010 speaker recognition evaluation," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[198] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech communication*, vol. 12, no. 3, pp. 247–251, 1993.

[199] C. Yu, G. Liu, S. Hahm, and J. H. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4017–4021, IEEE, 2014.

[200] G. Hinton, "A practical guide to training restricted boltzmann machines," *Momentum*, vol. 9, no. 1, p. 926, 2010.

[201] L. D. Alsteris and K. K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, vol. 17, no. 3, pp. 578–616, 2007.

[202] D. Erro, I. Sainz, E. Navas, and I. Hernaez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 2, pp. 184–194, 2014.

[203] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigné, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3, pp. 187–207, 1999.

[204] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2222–2235, 2007.

[205] X. Tian, Z. Wu, S. Lee, and E. S. Chng, "Correlation-based frequency warping for voice conversion," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, pp. 211–215, IEEE, 2014.

[206] S. B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," in *Readings in speech recognition*, pp. 65–74, Morgan Kaufmann Publishers Inc., 1990.