Nagaoka University of Technology
Graduate School of Engineering
Department of Information Science and Control Engineering

A Study on Speech Classification Based on Deep Neural Network under Adverse Environments

Khomdet Phapatanaburi (14700984)

Submitted in part fullfillment of the requirements for the degree of
Doctor of Engineering of Nagaoka University of Technology, Japan
@Speech Information Processing Laboratory, July 2017

Abstract

Recently, speech classification plays an important role in real-world speech application such as mobile devices and voice control system of a car. However, additive noise and reverberation under adverse environments can degrade the performance of speech classification.

In this thesis, author focuses on improving speech classification considering additive noise and reverberation under adverse environments by using deep neural network (DNN). Two speech classification tasks: accent recognition and voice activity detection (VAD) are considered. The accent recognition task is investigated under reverberant environment while the VAD task will be examined under noisy environment.

For accent recognition task, DNN-based accent recognition is proposed for modeling the reverberation over multi-frame. Moreover, the combination of scores of DNN-based accent recognition and Gaussian Mixture Model (GMM)-based accent recognition is also proposed to utilize complementary characteristics of these two models. In reverberant environment, our proposed method outperformed conventional GMM-based accent recognition. The accent recognition rate was improved from 90.7 % with GMM to 93.0 % with DNN. The combination of GMM and DNN achieved recognition rate of 97.5 %, which outperformed than the individual GMM and DNN because the complementation of GMM and DNN.

For VAD task, a DNN-based joint phase- and magnitude -based feature (JPMF) enhancement (JPMF with DNN) and a noise-aware training (NAT)-DNN- based JPMF enhancement (JPMF with NAT-DNN) are proposed. Moreover, to improve the performance of the proposed feature enhancement, a combination of the scores of the proposed phase- and magnitude-based features is also applied. Specifically, Mel-frequency cepstral coefficients (MFCCs) and the mel-frequency delta phase (MFDP) are used as magnitude and phase features. The experimental results show that the proposed feature enhancement significantly outperforms the conventional magnitude-based feature enhancement. The proposed JPMF with NAT-DNN method achieves the best relative equal error rate (EER), compared with individual magnitude- and phase-based DNN speech enhancement. Moreover, the combined score of the enhanced MFCC and MFDP using JPMF with NAT-DNN further improves the VAD performance.

Furthermore, motivated by the great advantage of phase information, deep neural network (DNN) using magnitude and phase information called phase aware DNN is proposed to achieve better VAD performance. In noisy environment, record under low signal-to-noise ratios (SNRs), the results show that the proposed method significantly outperforms conventional method. Based on DNN classification, the EER was reduced from 23.70% of DNN using only magnitude information, to 19.92% of phase aware DNN. Moreover, the score combination of different phase aware DNN can further improves the VAD performance.

Acknowledgements

At first, I would like to express my gratitude to my supervisor Professor Longbiao WANG and Professor Masahiro IWAHASHI for giving me invaluable advices on the way to be a doctoral student. My supervisor taught and support me throughout my three years in doctoral study. I would like to thank them for helping me to complete my studies and research on time.

I would like to express my gratitude to Prof. Yasunori SUGITA, Prof. Takashi YUKAWA, Prof. Koichi YAMADA and Prof. Masato AKETAGAWA, to be my thesis examiners. Their insightful comments in the qualifying examination helped me to improve this dissertation.

I would also like to thank Japan government via Rajamangla University of Technology Thanyaburi for providing the Monkukagasho scholarship and giving a chance to have a doctoral study at Nagaoka University of Technology (NUT). I would also like to express my thankful to the International Student Affair department (Kokusaika) of NUT for taking care of me during my study. I wish to express my gratitude to Assistant Professor Warunee ARIYAWIRIYANAN and Assistant Professor Jakkree SRINONCHAT for advising me to study here.

I much thanks to my colleagues, Zhaofeng ZHANG, Yuta KAWAKAMI, Satoshi SHIOTA, Zeyan OO, Fairoza A.B. HAMZAH, Tin Moe AYE, and all other students in speech information processing laboratory and image information system laboratory and to all Thai students in NUT.

Last but not least, I would like to show my greatest appreciation to my parents, Pravit PATANABURI, Pradap PATANABURI, Thotsaphon PHAPATANABURI, and Kirana PHAPATANABURI for their constant love and support for my study in NUT.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

Recently, speech classification receives with attention from reserchers all over the world got by improving speech-based technologies [2, 3]. The purpose of speech classification is to take human speech signal into computation/some process and classify the speech into many categories such as accent recognition, voice activity detection (VAD) [4], spoofing speech detection [5], speaker recognition [6], phone classification [7], emotional speech classification [8] and etc [9, 10, 11]. Each categories have different target for analyzing spoken speech signal. For example, in accent recognition [8], it is used to detect foreign/regional accent of speaker from spoken utterance. The VAD [12] will select human speech boundary from background sound for speech utterance. The speaker recognition [13] is applied to identify speakers who speaking. The emotional speech recognition [14] is aimed to identify emotion of speakers in spoken speech such as sad, happy,

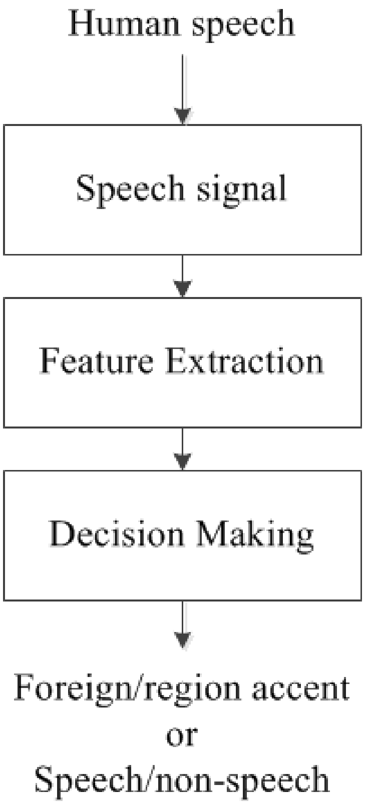


Figure 1.1: Structure of speech classification

and etc.The identification of foreign/regional accent of speaker and human speech boundary from background sound for speech utterance is focused in this study. Identifying accents of speaker and human speech boundary can provide significant benefits for speech application when improving automatic speech recognition [15, 16].

Figure 1.1 shows the most common speech classfication methods, which is based on machine-learning methods.

In this structure, a speech signal is first recorded from human by using the microphone. After that the speech s' feature is extracted and will undergo the the decision making procedure . Finally, the decision making is performed to identify the target task. In this dissertation, the targets are identified into the foreign/region accent of speaker or speech/non-speech boundary from spoken utterance are classified.

Traditional speech classification system [17] is known to work well when the speech signals are recorded using high quality system under clean environments. However, the existing speech

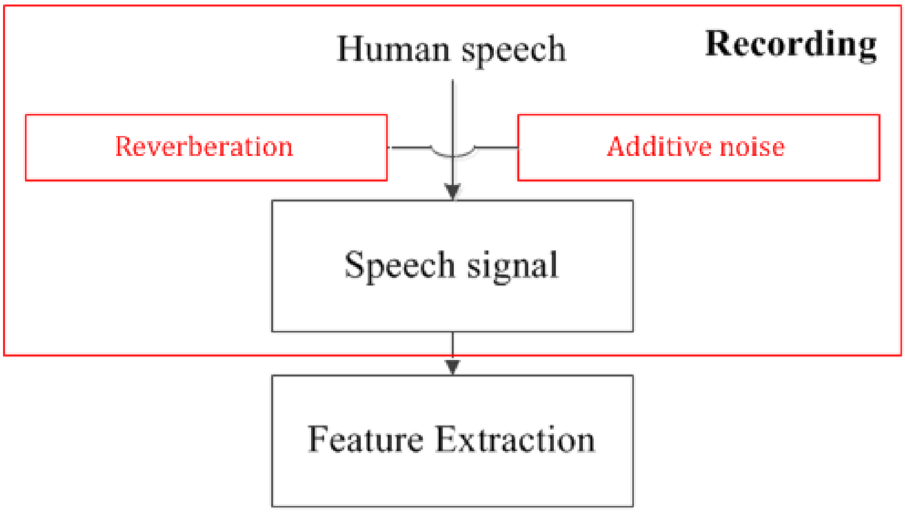


Figure 1.2: Reverberation and additive noise in speech classification system, where occurred when the speech signal is recorded by the microphone. The reverberation happened when a signal is reflected from surfaces of objects, such as the walls and furniture. The additive noise occurred when a speech signal is interfered by undesired sound such as machine interfaces, car, and air condition.

classification system can easily be applied in a clean environment compared to the adverse environment, which is not applicable in the real-world application. Additive noise and reverberation under adverse environments [18] degrade performance of speech classification. Therefore, it is important to consider speech classification under the adverse environments.

In order to obtain real-world-based speech classification applications, two types of adverse environment: reverberation and additive noise are considered. They occurred when the speech signal is recorded by the microphone as shown in Figure 1.2. The reverberation happened when a signal is reflected from surfaces of objects, such as the walls and furniture. The additive noise occurred when a speech signal is interfered by undesired sound such as machine interfaces, car, and air condition. In this dissertation, two types of speech classification is discussed which are the accent recognition task analyzed under the reverberation, and theVAD task affected by the additive noise.

To improve the performance of existing accent recognition and VAD task under adverse environments, the deep neural network (DNN), which has been proved to be efficient for speech system affected by the reverberation and additive noise, is applied. Two types of DNN are used in this study. The first one is classification based DNN where DNN can be applied as a speech classification to deal with reverberation and additive noise by using context dependent frames. The second one is regression based DNN which can be exploited to transform from noisy speech feature to clean speech feature for speech processing. The detail explanation on the two tasks is discussed in Section 1.2 and 1.3.

1.2 The introduction of classification-based DNN for distant talking accent recognition

Automatic accent recognition is a fundamental task in the speech processing community because the performance of practical speech application [19, 20, 21] is influenced by people s' accent. The accents can normally be classified into two kind accent : non-native accents caused by foreign speaker or non-native accents caused by regional speaker. In this dissertation, accent recognition is focused on native accent in speech utterance which has no accent gender information, no accent gender or information no transcripts.

In the field of accent recognition, various types [21, 22, 23] of accent models have been studied and proposed. The Gaussian mixture model (GMM) seems to be the most popular statistical model for accent recognition. This is because some accent-dependent spectral shapes can be represented by using Gaussian components, which is important in accent recognition process. For example, in [24], GMM is applied to automatically identify the regional Swiss French accents based on four different regions of Switzerland, where it is showed that the GMM works well in term of the accuracy rates of accent recognition.

Although conventional GMM-based accent recognition system [24, 25] is known to perform well when the close-talking microphone is used to recode speech signal, many speech applications [26, 27, 28, 29], in real world is under a distant-talking environment such as an room, office, etc. Consequently, accent recognition in a distant-talking environment need to be investigated to improve robust distant-talking speech recognition. Normally, the effectiveness of traditional system in speech processing is reduced when the mouth of speaker is far away from the microphone. This is due to reverberation and noise background. It is considered that the effectiveness of accent recognition in reverberant environment will be degraded too. However, there are very few studies which focus on accent recognition in reverberant environment. The conventional GMM-based method may not perform well in this situation because investigating reverberation over multi-frame cannot be modeled by GMM.

Nowadays, deep neural network (DNN) has been applied as speech classification (such as speech/speaker recognition) in various speech conditions. DNN improved discriminative training method which is the same as in [30], multi-layer perceptron (MLP). The merit of DNN is that it can be applied as a speech classification[31], based on non-linear transform to deal with reverberation and additive noise by using context dependent frames. Hence, many studies have showed that it is robust for speech recognition under distant-talking environment.

In this dissertation, author apply GMM as the baseline, and propose a classification-based DNN for distant-talking accent recognition in a reverberant environment. Moreover, to improve the accent recognition performance, we also propose the combination of likelihood of GMM and DNN.

1.3 The introduction of regression and classification-based DNN for robust noise VAD

VAD is a task to identify human speech from other sounds during speech utterance. It is very important in speech processing because the practical speech applications [32, 33, 34, 35] can be influenced by distinguishing speech and non-speech regions. In clean speech environment, methods based on energy in [36] [37] is widely applied. These methods work well since energy levels between speech and non-speech segments is quite different. However, the VAD based on energy cannot work well because of the problem of the corrupted speech signal from additive noise. Thus, energy-based VAD may not be suitable for speech applications under noise condition.

Nowadays, machine-learning-based VAD [38, 39] have been popular under noise condition. This is because they have a strong theoretical base that guarantees their performance. Although many types of machine-learning-based VAD have been proposed, support vector machine (SVM) and deep neural network (DNN) may be popular acoustic models for VAD detector. For SVM, [40] proposed SVM-based VAD as a single frame-based classifier. The results showed that SVM-based VAD outperformed conventional VAD such as short-term energy-based method, long-term spectral divergence method [41], and GMM-based VAD because the SVM captures speech-relevant information effectively. For DNN, a deep belief network-based VAD was proposed as multi-frame-based classifier. The experimental result showed that the DNN-based VAD outperformed traditional VAD (e.g., G.729B [42] and Sohn VAD [43]) because of its robustness to noise. Hence, theSVM and DNN-based VAD are considered as VAD a baseline in this dissertation. Although conventional SVM and VAD-based VAD system have proved to perform well, the conventional feature may not be effective under noisy conditions due to the corruption of the speech.

To improve conventional features, deep neural network (DNN) has attracted attention to produce effective feature in many speech processing tasks [44, 45, 46, 47, 48] . The merit of deep neural network (DNN) applied as a regression-based method is it can transform the noisy feature to a clean feature. For instance, DNN-based feature enhancement proposed in [44] was applied to VAD. Compared with traditional feature-based VAD, the VAD using a DNN-based enhanced feature improves performance because the nonlinear mapping function can predict clean features from corrupted features, hence the VAD is better in decision making process. However, the weakness of DNN-based feature enhancement is it is dependent to the training data. Thus, the performance of enhanced feature cannot perform well when the system is evaluated on previously unseen data.

To obtain effective feature enhancement, DNN-based feature enhancements using magnitude and phase information is considered to achieve better VAD performance. In this dissertation, a DNN-based joint phase- and magnitude -based feature (JPMF) enhancement (JPMF with DNN) and a noise-aware training (NAT)-DNN based JPMF enhancement (JPMF with NAT-DNN) for noise-robust voice activity detection (VAD) are proposed. Then, to further improve the performance of the proposed feature enhancement, a combination of the scores of the proposed phase and magnitude-based features is also applied.

In addition to proposed DNN based feature enhancement, this dissertation proposes a classification-based DNN using magnitude and phase information (that is, phase aware DNN where it can exploit full information in the original signal) to achieve better VAD performance compared to the conventional DNN based classification. Moreover. automatic score combination is applied to improve the performance of phase aware DNN.

1.4 The structure of this dissertation

The reminder of this dissertation is organized as following:

- In Chapter 2, the basic framework of accent recognition and voice activity detection (VAD) are introduced to explain the fundamental part for these two tasks.
- Chapter 3 investigates the distant-talking accent recognition with reverberation using conventional (that is Gaussian mixture model (GMM)) and proposed methods (that are DNN-based accent recognition and combination of GMM and DNN).
- In Chapter 4, the mechanism and experiment of several feature enhancements (including conventional DNN-based feature enhancements and proposed joint phase and magnitude enhancements) are considered for noise robust VAD.
- Section 5 also investigates conventional classification-based DNN and proposed classification-based DNN (that is phase aware DNN) for noise robust VAD.
- Section 6 summarizes the dissertation and future works are briefly described here.

Chapter 2

The basic framework of accent recognition and voice activity detection

In this Chapter, the basic framework of accent recognition and voice activity detection (VAD) task where both systems have similar speech processing are discussed here. Fig 2.1 shows the system components of both tasks. They can generally categorize into front-end and back-end processing phases. The feature extraction for the input of acoustic model is investigated in front-end processing, whereas the classification/modeling components are described in the back-end processing.

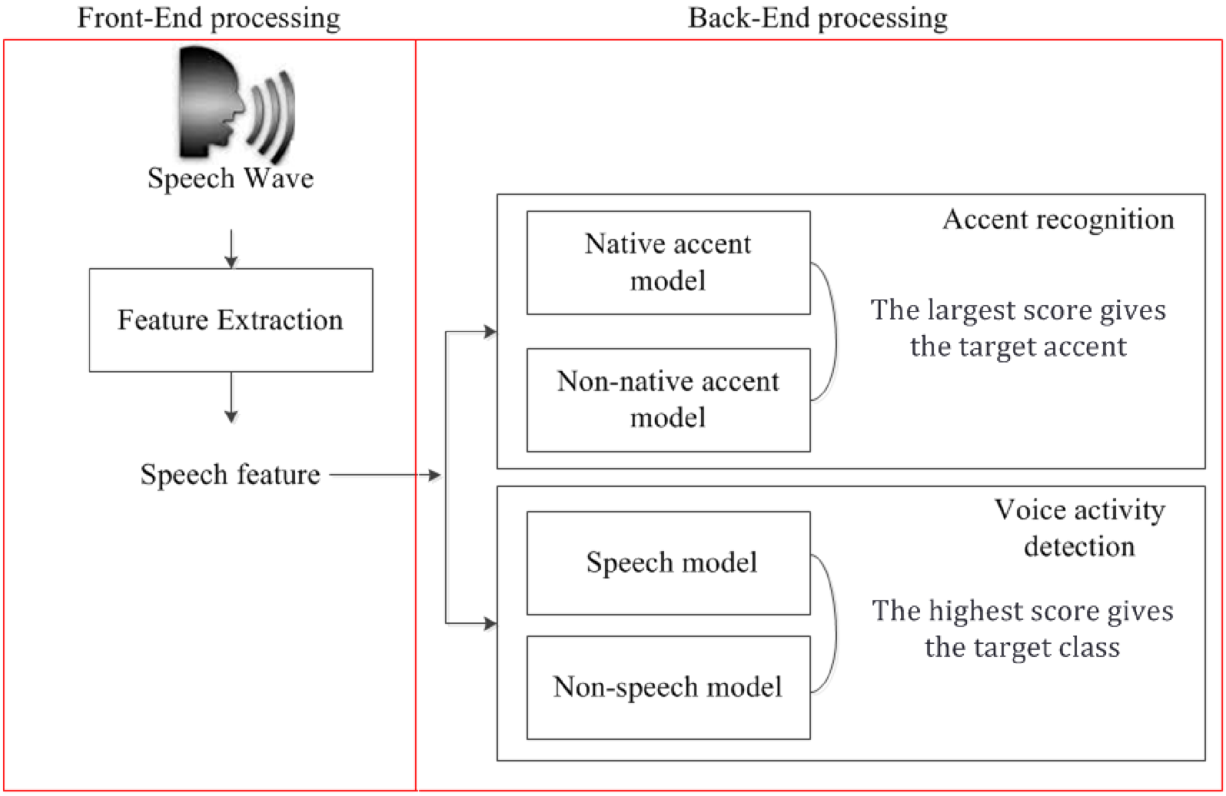


Figure 2.1: The flowchart of accent recognition and VAD systems categorized into including front-end and back-end processing part.

2.1 Front-end processing

In this Section, the conventional front-end processing for accent recognition and VAD that refer to a sequence of acoustic feature vectors, which is converted from the input of recorded speech waveform is introduced. The sequence of fixed vectors is called as feature extraction. There are many features being introduced for speech classification, such as mel-frequency cepstral coefficients (MFCC)[49], power-normalized cepstral coefficients (PNCC)[50], perceptual linear prediction (PLP) [51], mel-frequency Delta-phase cepstral coefficients (MFDP))[52], modified group delay cepstral coefficients (MGDCC)[53], baseband phase difference (BPD)[54], instantaneous frequency (IF)[55], etc. The choice of feature extraction depends according to the classification problem, computational capability, and the amount of speech data. Similarly, these features can be combined to exploit different specific characteristic of speech signal.

In accent recognition and VAD task, MFCC is the most common feature extraction. This is because it is easy to apply, yet still provide a good recognition rate. Therefore, the MFCC is used as the magnitude-based standard feature. The MFDP and MGDCC are also considered to utilize phase information in improving the performance of VAD. The detail of features will be described as follow.

2.1.1 Mel-frequency cepstral coefficients

In order to capture the MFCC representation from speech signal, the speech signal is segmented into overlapped frame of length (note that the 25ms frame of length is used in this dissertation), and then a high-pass filter is exploited to compensate the high-frequency component of input speech as follow:

$$y(n) = s(n) - \alpha s(n-1) \tag{2.1}$$

where α is defined as a chosen parameter between 0.9 to 1. and $y(n)$ is the pre-emphasized speech signal based on high-pass filter. Next, the multiplication of each frame of the pre-emphasized output speech with a window function in order to keep the important continuity of the first and the end point in the frame which is constructed as $y'_\imath(n)$ given by

$$y'_\imath(n) = y(n)w(n - \imath D) \tag{2.2}$$

where $w(n)$ is a causal window of length T (i.e., zero-valued outside the range $0 \leq n \leq T-1$), $\imath$ is the frame index, D is the number of samples between successive analysis frames (the step size, where $D \leq T$). For the causal window, this dissertation uses Hamming window since it provides low distortions in the spectral analysis of speech signals. The aim of this multiplication is to eliminate the distortions in finite-length effects.

After the multiplication, the speech signal is decomposed into its frequency component by fast discrete Fourier transform (DFT) as shown in the following equation.

$$Y_\imath(k) = \sum_{n=1}^{N} y'_\imath(n) e^{-jw_k n} \tag{2.3}$$

Here, $w_k = \frac{2\pi k}{L}$, where L is the number of analysis frequencies being considered in the DFT (with $L \geq T$). Normally, the magnitude part of DFT is used for MFCC feature computation. The power spectral $P_\imath(k)$ of $Y_\imath(k)$ is calculated by

$$P_\imath(k) = \frac{1}{N}|Y_\imath(k)|^2 \tag{2.4}$$

Then the power spectral $P_\imath(k)$ of each frequency is multiplied by Mel-scale to calculate mel-spaced filerbank, Y^{mel} between lower and upper frequency limits given in the designed frequency range (Hz). The Mel-scale is related to the pitch, or the perceived frequency of a speech signal. Combining the mel-scale makes the feature match more closely human's mechanism and reduce the number of coefficient as human can properly distinguish small changes on the pitch at low frequencies better than on the pitch at high frequencies,

The mel-spaced filerbank is calculated by applying a set of 20-40 (note that this dissertation use 24) triangular filter as shown in Figure 2.2

The next step is to apply discrete cosine transform (DFT) on the log of filterbank, $log(Y_a^{mel})$ to calculate $\Im$ mel-scale cepstral coefficients given by

$$F_m = \sum_{a=1}^{\Re}[log(Y_a^{mel})]cos[\frac{\pi m}{\Re}(a-0.5)], \qquad m = 1, 2, 3, ..., \Im \tag{2.5}$$

where $\Re$ is the number of frequency bin, $\Im$ is the total number of mel-scale cepstral coefficients.

By exploiting discrete cosine transform (DFT), the higher coefficients are cut due to more noise-like features. At the same time, the spectral shape of the signal is taken. The final result of F_m is called mel-frequency cepstral coefficients (MFCC) .

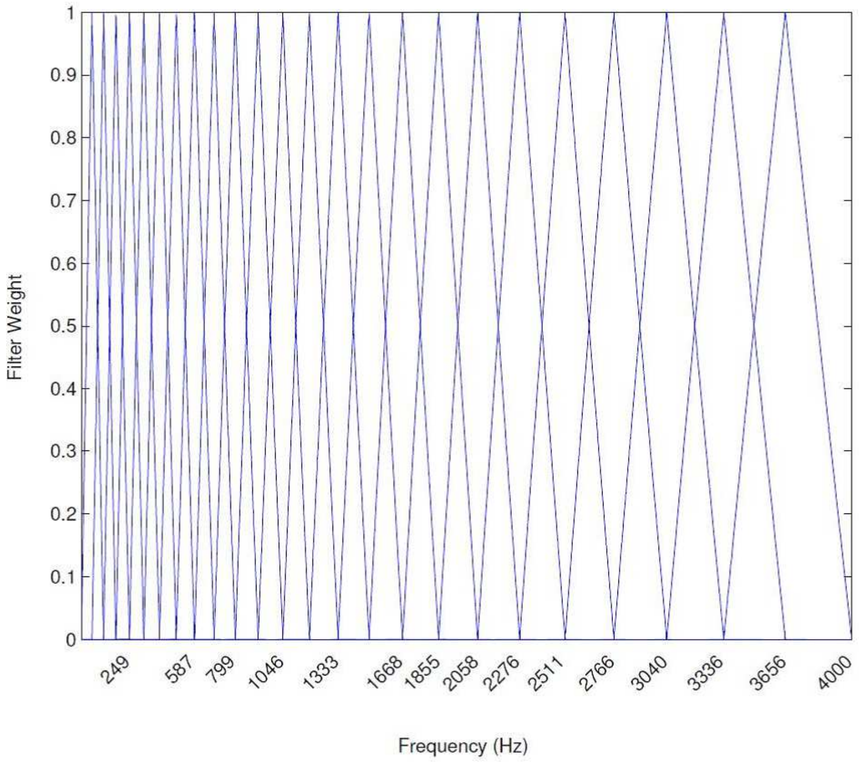


Figure 2.2: Shape of the Mel filter bank weights [1] based on a 24-filter system with an 4000 Hz sampling frequency.

2.1.2 Mel-frequency delta-phase

This Section presents the extraction of mel-frequency Delta-phase cepstral coefficients (MFDP) proposed by [52]. MFDP, which is derived from the phase spectrum has proven to be an effective feature for speech classification (including VAD and speaker recognition). It can be computed from the phase spectrum difference of the short-time discrete Fourier transform (STFT) between neighboring frames. The STFT of an input speech signal sequence is introduced in Eq. 2.3. Prior to calculating MFDP, the short-time delta phase spectrum $\triangle\phi_\imath(k)$ is designed to avoid some of the phase unwrapping problems and is given as the time-derivative of the short-time phase spectrum as follows,

$$\triangle\phi_\imath(k) = arg[Y_\imath(k)Y_{\imath-1}(k)^* e^{-jw_k D}] \tag{2.6}$$

where $(.)^*$ indicates the complex conjugate and D is the number of samples between successive analysis frames. The delta phase spectrum based on [52] is used to obtain the MFDP features by applying the mel filter bank to the absolute delta phase spectrum, followed by taking the logarithm of the filter bank energies and performing a discrete cosine transform to obtain the MFDP feature. Here, we use MFDP as phase information, using a rectangular window with a length of 25 ms instead of the common length of 256 ms because the longer length includes excess speech and non-speech segments in each window.

2.1.3 Modified group delay cepstral coefficients

Recently, modified group delay which is proved to be efficient in choosing the phase-based features has gained lots of interest from researchers in the field of speech recognition. This is because it can solve the phase wrapping problem where the phase information are stuffed into $(-\pi \leq \text{Ø} \leq \pi)$. It can be computed by improving the group delay (GD). The GD is defined as the negative derivative of the Fourier transform phase for frequency and can be computed directly as follow:

$$\tau_y(k) = -\left(\frac{d}{dk}\log\left(Y(k)\right)\right) \tag{2.7}$$

$$= \frac{Y_R\left(k\right)\vec{Y}_R\left(k\right) + Y_I\left(k\right)\vec{Y}_I\left(k\right)}{|Y(k)|^2} \tag{2.8}$$

where, $Y(k)$ is the Fourier transform of the signal $s(n)$, $\vec{Y}\left(k\right)$ denotes the Fourier transform of $ns(n)$, footnote R and I indicates the real and imaginary part of the complex, respectively. Although GD can give a good performance for speech classification, it cannot work well when $|Y(k)|^2$ is approximated to zero. In [53], the modified GD feature have been studied and report that the modification can give better performance than the original GD by changing $|Y(k)|^2$. The modified group delay function can be formed as

$$\tau_{mgd}(\omega) = \frac{Y_R(k)\vec{Y}_R(k) + Y_I(k)\vec{Y}_I(k)}{|S(k)|^{\partial\Phi}} \tag{2.9}$$

Here, $S(k)$ is cepstrally smoothed $Y(k)$. The range of ∂ and Φ are $(0 < \partial \leq 1.0)$. In this dissertation, τ_{mgd} (applying $\partial = 0.4$,$\Phi = 0.6$) is used to produce MGDCC feature for the input of robust noise VAD.

2.2 Back-end processing

In this Section, we address conventional back-end processing for accent recognition and VAD system that cover acoustic model to recognize/classify the target task.

2.2.1 Accent recognition

Accent recognition is a task that identify correct accents of speaker from their voice. Figure 2.3 shows a flowchart of an accent recognition. The process of system can be classified into 2 main parts which consist of the enrollment part and recognition part.

In the enrollment part, a accent model is trained by the feature vector of the target accent. In the recognition stage, a feature transformed from unknown accent is fed into the trained accent model. Then, the trained model is compared with tested feature to identify target accent.

In this study, Gaussian mixture model (GMM) is applied for conventional accent recognition compared to proposed method. The detail of accent recognition task is introduced as following.

GMM-based accent recognition

GMM has been successfully used in domains of accent speaker recognition/recognition. For example, [23] performed an automatic recognition of accents using GMM and reported that

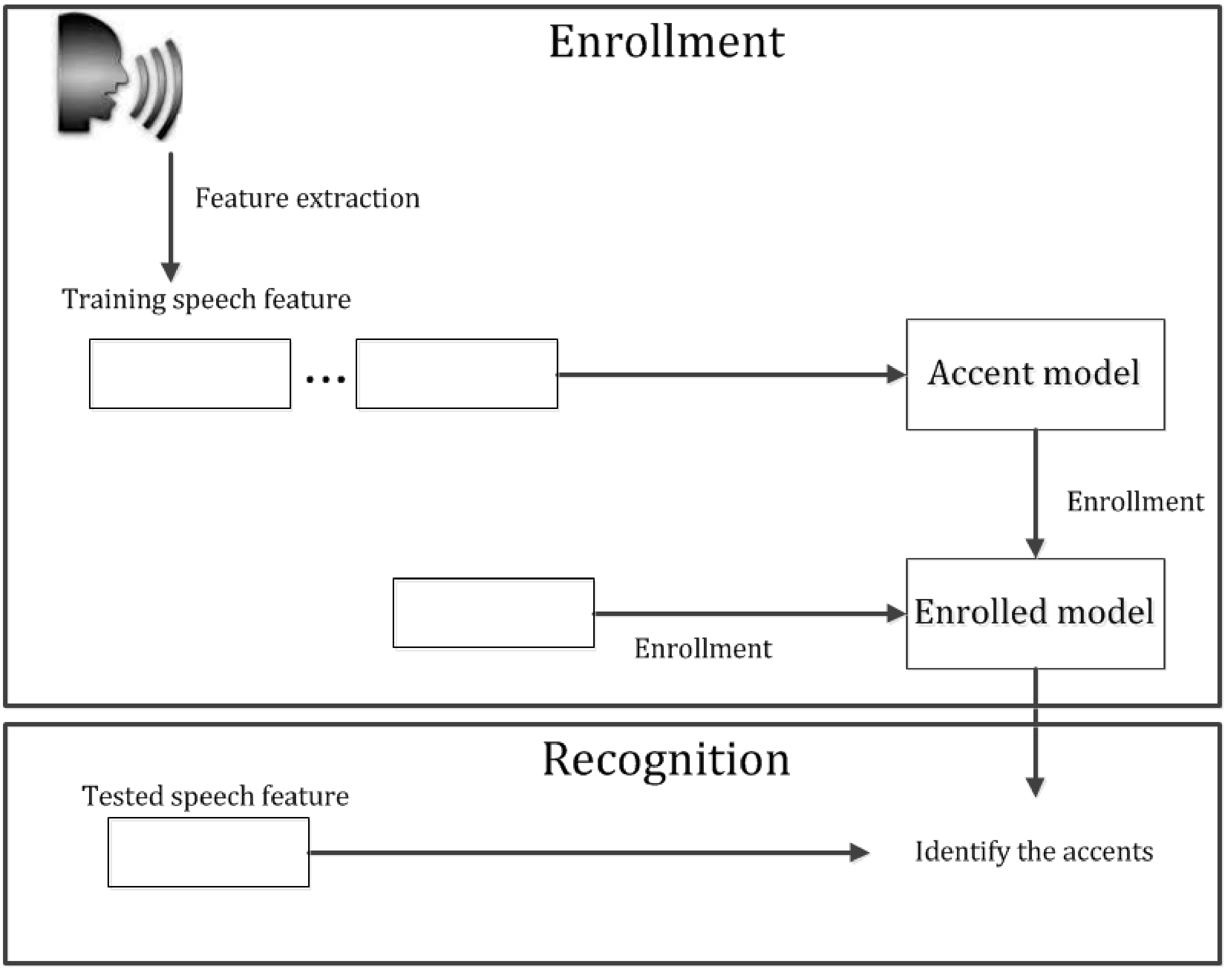


Figure 2.3: Diagram of accent recognition system for identifying native/non-native accent.

proper selection of number of component results in good recognition performance. From above reason, we use GMM to identify accents of speaker. We learn a set of 2 mixture models, one for native accent set and the second for non-native accent set. The native accent set is modeled using a GMM1 and the non-native accent set is modeled using a GMM2. The higher likelihood gives us the accent class to which the speaker belongs. The process step is briefly shown in Figure 2.4.

The brief description of GMM is summarized as follows. A GMM is a weighted sum of M component densities and is given by the form:

$$P(X|\lambda) = \sum_{i=1}^{M} c_i bi(x), \tag{2.10}$$

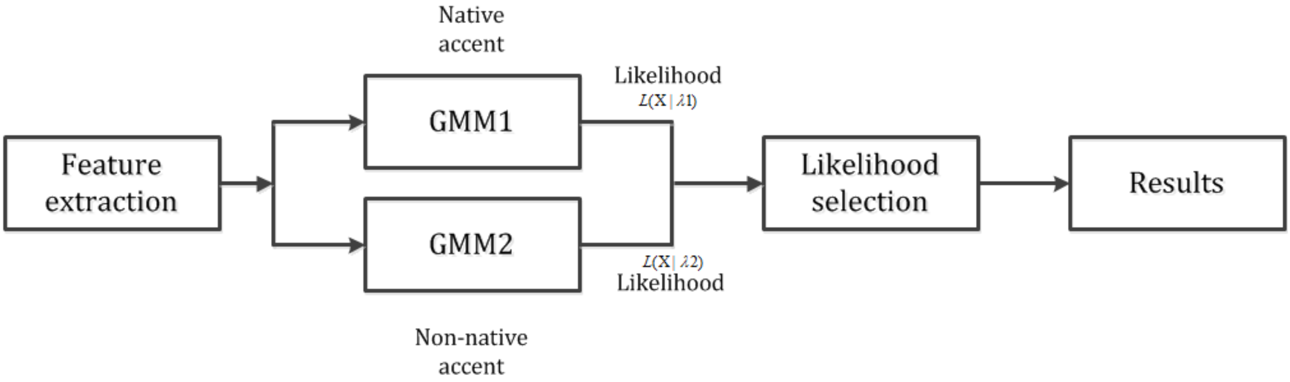


Figure 2.4: The diagram of GMM based accent recognition.

where w is a d-dimensional random vector, $b_i, i = 1, ..., M$, is the component densities and $c_i, i = 1, ..., M$, is the mixture weights. Each component density is a D-variate Gaussian function of the form:

$$b_i(x) = \frac{1}{(2\pi)^{d/2}|\sum_i|^{1/2}} exp\{\frac{1}{2}(x - \mu_i)^T \sum_i^{-1}(x - \mu)\}, \tag{2.11}$$

with mean vector μ_i and covariance matrix $\sum_i$. The mixture weights satisfy the constraint that:

$$\sum_{i=1}^{M} c_i = 1. \tag{2.12}$$

The complete Gaussian mixture model is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation:

$$\mu = \{c_i, \mu_i, \sum_i\}, i = 1, ..., M, \tag{2.13}$$

In our accent recognition system, each accent is represented by such a GMM and is referred to the model k. For a sequence of T test vectors $T = x_i, x_2, ..., x_T$ the standard approach is to calculate the GMM likelihood in the log domain as:

$$L(X|\lambda) = logp(X|\lambda) = \sum_{i=1}^{T} (x_i|\lambda), \tag{2.14}$$

The accent-specific GMM [56] parameters are estimated by using the expectation maximization (EM) algorithm [57] which used the training data uttered by the corresponding accent in the HTK toolkit [58] .

2.2.2 Voice activity detection

In this study, support vector machine (SVM) and deep neural network (DNN) are applied as conventional VAD classification. The detail is introduced as following subsection.

SVM-based VAD

SVM is one of of the popular methods for single condition based VAD since it can be applied as a two-class classification which is based on a separating hyperplane. A flowchart of the SVM based VAD system, which consists of the training phase and the testing part, is shown in Figure 2.5.

In training phase, the training set used to train the SVM consist of different vectors, which are the class labels. In this study, the first vectors labeled as 1 correspond to speech feature vectors and another vectors labeled as 0 correspond to non-speech feature vectors. The SVM decision function is formulated as follows:

$$f(\Upsilon) = \sum_{\jmath=1}^{H} \alpha_\jmath K(\mathcal{X}_\jmath, \Upsilon) + b \tag{2.15}$$

where $\mathcal{X}_\jmath$ are the support vectors, $\alpha_\jmath$ their weights and b is a constant bias. Υ is the unclassified tested vector, and $K(\mathcal{X}_\jmath, \Upsilon)$ is the kernel function. The support vectors are derived from the training feature sample through an optimization process, and therefore they are a subset of the

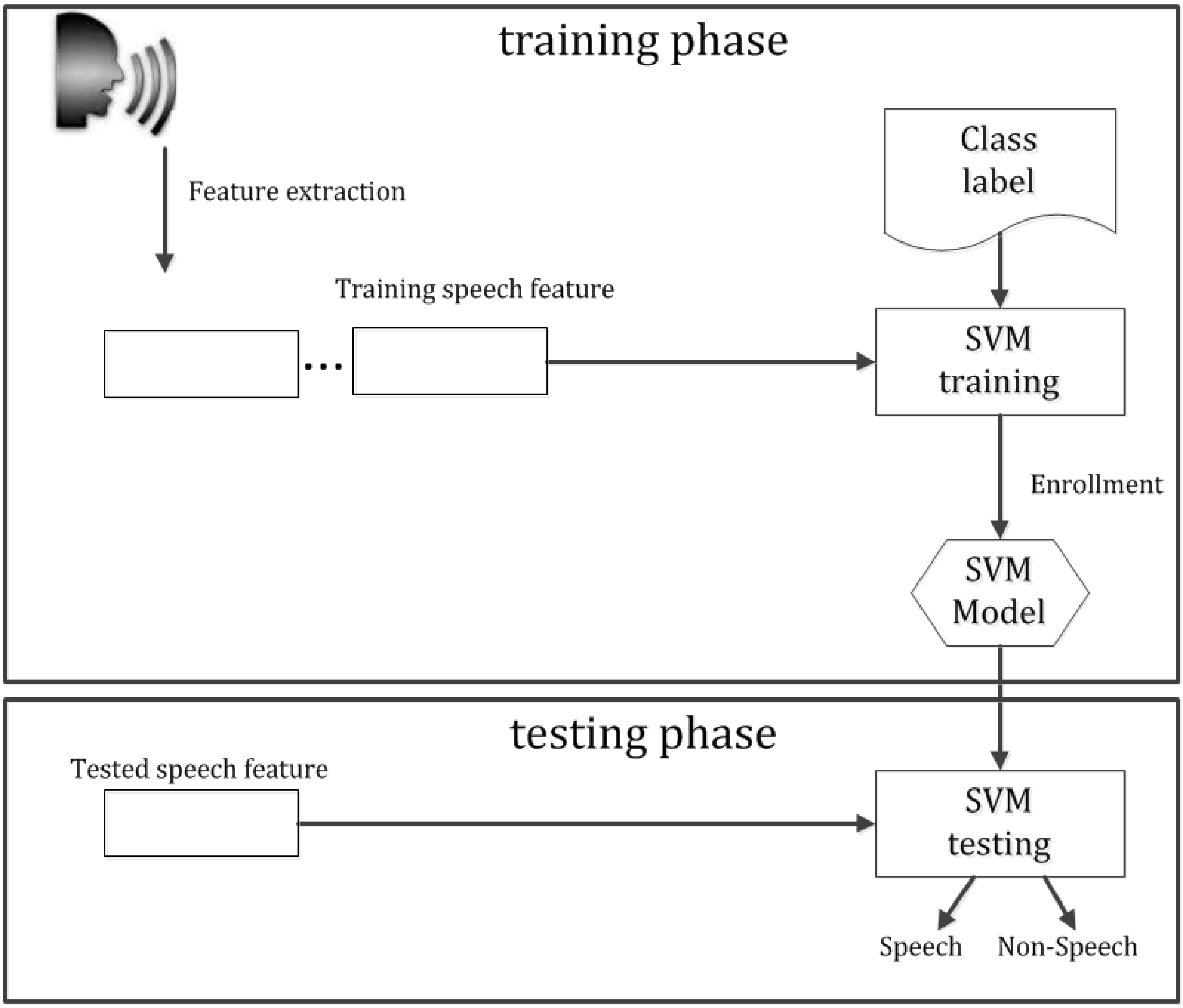


Figure 2.5: Diagram of SVM-based VAD system.

training feature sample. We use the publicly available LIBLINEAR tool [59] which considers linear kernels as following equation:

$$K_{linear}(\mathcal{X}, \Upsilon) =< \mathcal{X}, \Upsilon > \tag{2.16}$$

where $< . >$ denotes the inner product, For SVM, we use a grid-search using cross-validation which is based on [59] where various pairs of (C, γ) the values are tried and the one with the best cross-validation accuracy is picked. The regularization parameter is searched from the exponential grid $\{C = 2^{-5}, 2^{-3}, ..., 2^{15}, \gamma = 2^{15}, 2^{-13}, ...2^{3}\}$.

In testing phase, the SVM classifier trained SVM model obtained in the training phase will

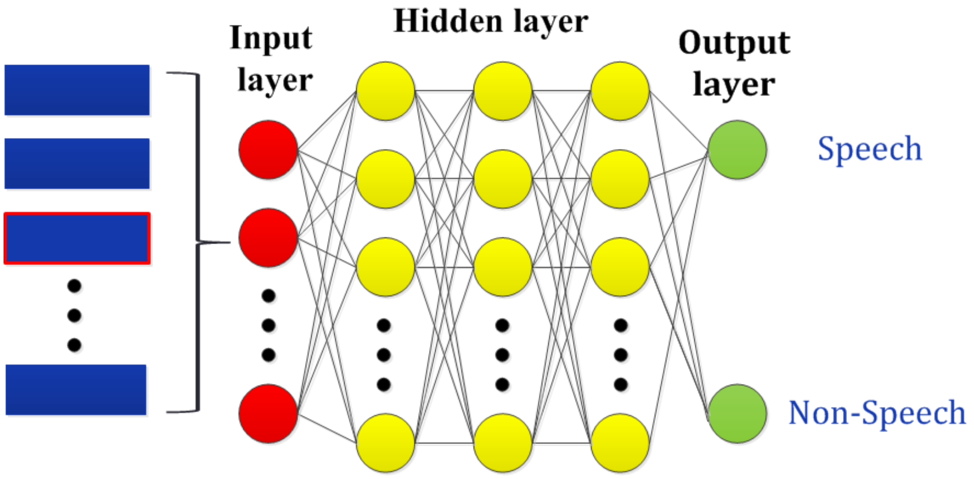


Figure 2.6: Diagram of DNN-based VAD system.

classify each frame of the tested speech feature vector to be speech and non-speech segment.

DNN-based VAD

Few years ago, DNN has been successfully applied for speech classification. This is due to powerful phone discrimination using context dependent classification, which is broadly reported in Automatic speech recognition (ASR). To consider multi-frame classification, DNN is used as VAD detector to differentiate the speech and non-speech boundary.

Figure 2.6 shows the structure of the conventional DNN-based VAD. On the left of the DNN as referred the figure above, a sequence of features, which is described in Section 2.1.1 is concatenated over several frames and then fed into the DNN. Next, the DNN is used to consider the concatenated features, x_{concat} (that is, context information) for distinguishing speech and non-speech parts. After the nonlinear transformation of the hidden layers, the speech and non-speech part based on probability scores is predicted by a linear output layer.

To train the DNN, the parameter of DNN is initialized by a pretrained restricted Boltzmann machines (RBM) [60] which firstly uses unsupervised training. After that stochastic gradient decent (SGD) is used to the cross entropy loss function as follow:

$$C_{ce} = -\sum_{q=1}^{\aleph}(r_q log(\hat{r}_q(xconcat)) + (1 - r_q)log(1 - \hat{r}_q(x_{concat}))). \tag{2.17}$$

Here, r and $\hat{r}$ denote the predicted and reference probabilities. $q \in \{n = 1, 2\}$ corresponds to class of speech or non-speech. Normally, this process step is often called fine-tuning which it is expected to correctly distinguish speech and non-speech after completely trained DNN.

Chapter 3

The classification-based DNN for distant-talking accent recognition

3.1 Motivation of this proposal

In previous Chapter 2, the framework of accent recognition was introduced. Traditional accent recognition has been proven to be effective when the speech signals are captured using a close-talking microphone. However, many real-world speechaided applications require to under a distant-talking environment [28, 29, 27, 61, 26, 62, 63], such as an office or the cabin of a car. Therefore, accent recognition in a distant-talking environment is a very important issue for robust distant-talking speech recognition with non-native accent. The performance of conventional system degrades dramatically as soon as the microphone is moved away from the mouth of the speaker. This is due to broad variety of effects such as background noise and reverberation. It is considered that the performance of accent recognition in a reverberant environment will be degraded dramatically too. However, very few studies focus on accent recognition performances in a reverberant environment. The conventional GMM-based approach may not achieve a good performance in such situation since the GMM cannot model the reverberation over multi-frame.

Recently, Deep Neural Networks (DNNs) have been proposed for audio classification (such as speech/speaker recognition) in various acoustic conditions. DNNs can be seen as improved multilayer perceptron (MLP), discriminative training method is the same [30]. The advantage of DNN is pretraining which can determine the significant initial value of the multi-layer networks. It employs an unsupervised pretraining method using restricted Boltzmann machine (RBM) have been proposed to train better initial values of deep networks [60]. A RBM is obtained by adding constraints that there is no connection of the same layer and connected to the symmetry to the Boltzmann machine. Its parameter is determined by the Greedy layer-wise training .This training method minimize the error between the input feature and the inverse transform feature of the input. After pretraining, DNNs are obtained by discriminative trained using BP algorithm from multilayer structure stacked RBMs. DNNs with pretraining have been shown better performance than the conventional MLP without pretraining on automatic speech recognition and large vocabulary business search task etc [60, 31]. Because DNN can train the non-linear transformation of context-dependent acoustic signal by using multi-frame speech signal as input, many studies have showed that it is robust for speech recognition under distant-talking environment.

In this Section, DNN based acoustic model is applied for accent recognition in a reverberant environment. Although DNN-based method can provide better discrimination on confusion regions of different classes than GMM but the performance of DNNs is more dependent on training data than that of GMM [60, 31]. In our previous study, fusion of GMM and Hidden Markov Model (HMM) was proposed, and improved the speaker recognition performance than the individual model in a distant-talking environment [64]. In this dissertation, the combination of likelihood of GMM and DNN to improve the accent recognition performance is also proposed.

3.2 Existing method

Gaussian mixture model (GMM)-based accent recognition using mel-frequency cepstral coefficients (MFCC) which is mentioned in Chapter 2 is used as the existing method. It will be compared to proposed method.

3.3 Proposed method

3.3.1 DNN-based accent recognition

DNN have received a lot of attentions again in recent year and become dominant in acoustic modeling in speech recognition and image classification [60, 30, 65, 31, 6]. As a consequence, this paper also considers DNN to identify speaker accent in a distant-talking environment. Although DNNs [60] can be seen as improved MLPs, discriminative training method is the same. The advantage of DNN is Restricted Boltzmann Machine (RBM) based pretraining which can determine the good initial value of the multi-layer networks.

Restricted Boltzmann machine

RBM has visible and hidden layer in which visible units that represent observations are connected to hidden units that learn to represent features using weighted connection. An RBM is restricted that there are no visible-visible or hidden-hidden connections. Different types of RBM are used in the case of binary or real-valued input. Bernoulli-Bernoulli RBMs used to convert binary stochastic variables to binary stochastic variables. Gaussian-Bernoulli RBMs is used to convert real-valued stochastic variables to binary stochastic variables. In a Bernoulli-Bernoulli RBMs, the weights on the connections and the biases of the individual units define a probability distribution over the joint states of the visible and hidden units via an energy function. The energy of a joint configuration is:

$$E(v,h|\theta) = -\sum_{i=1}^{v}\sum_{j=1}^{H} w_{ij}v_i h_i - \sum_{i=1}^{v} a_i v_i - \sum_{i=1}^{j=1} b_i v_H \tag{3.1}$$

where $\theta = (w, a, b)$ and w_{ij} represents the symmetric interaction term between visible unit and hidden unit while and are their bias term. v and H are the numbers of visible and hidden units.

$$p(v|\theta) = \frac{\sum_h exp(-E(v,h))}{\sum_v \sum_h exp(-E(v,h))} \tag{3.2}$$

The probability that an RBM assigns to a visible vector v is:

$$p(h_j = 1|v,\theta) = \sigma(b_i + \sum_{i=1}^{v} w_{ij}v_i) \tag{3.3}$$

Since there are no hidden-hidden connections, the conditional distribution is factorial and is given by:

$$p(v_i = 1|h,\theta) = \sigma(b_i + \sum_{i=1}^{v} w_{ij}h_i) \tag{3.4}$$

where $\theta(x) = (1 + exp(-x))^{-1}$. Similarly, since there are no visible-visible connections, the conditional distribution $p(h|v,\mu)$ is factorial and is given by:

$$E(v,h|\theta) = \sum_{i=1}^{v} \frac{(v_i - a_i)^2}{2} \sum_{i=1}^{v}\sum_{j=1}^{H} w_{ij}v_j h_j - \sum_{j=1}^{H} b_j h_j \tag{3.5}$$

In a Gaussian-Bernoulli RBMs, the energy of a joint configuration is:

$$p(v_i = 1|h,\theta) = N(v_i a_i + \sum_{j=1}^{H} w_{ij}h_i, 1) \tag{3.6}$$

The conditional distribution $p(h|v,\theta)$ is factorial and is given by:

where $N(\mu, V)$ is a Gaussian with mean μ and variance V .

Maximum likelihood estimation of RBM is to maximize the log likelihood for the parameter . Therefore, the weight update equation is given by:

$$\triangle w_{ij} =< v_i h_j >_d ata- < v_i h_j >_{model} \tag{3.7}$$

where $< . >_{data}$ data is the expectation that and are on together in the training set and is the same expectation calculated from the model. Because computation of is expensive, using contrastive divergence (CD) approximation for the compute gradient. It is possible to compute by once the Gibbs sampling.

DNN-based acoustic model

DNNs are obtained from completed discriminative training of Deep Belief Networks (DBNs), which is configured hierarchically by connecting the pretrained RBM. In order to obtain a pretrained RBM, we trains Gaussian-Gaussian RBM first, and later trains the Bernoulli-Bernoulli RBM. The top layer of DNNs uses a softmax layer. The softmax [65] operation is given by:

$$p(l|h) = \frac{exp(b_l + \sum_i h_i w_{ij})}{\sum_m exp(b_m + \sum_i h_i w_{im})} \tag{3.8}$$

where b_l is the bias of the label and w_{ij} is the weight from hidden unit in top layer to label.

After configure the DBNs using RBM, DBNs is discriminative trained by using BP algorithm to maximize the log probability of the class labels. For the accent recognition task, the training speech data of all accents and their corresponding accent labels are used to train a universal DNN whose output layer has N output nodes where N is the number of kinds of accents. In the testing procedure, the input audio is passed to the resulting DNN to compute the posterior probabilities of each frame, the label of the frame is determined as accent label which has the maximum posterior probability. After we get the label of all frames in an utterance, the label of the utterance is further voted from the label of all frames.

3.3.2 Combination of GMM and DNN

Although DNN-based method can provide better discrimination on confusion regions of different classes than GMM, and the influence of reverberation can be learned by using multiple frames of input reverberant speech, the performance of DNNs is more dependent on training data than that of GMM [30, 31]. GMM assumes that data obeys Gaussian distribution, so the models can be trained better than DNN when the training data is not sufficient. In this paper, we propose the combination of likelihood of GMM and DNN to improve the accent recognition performance. We expect that the combination of complementary characteristics of these two models will achieve better performance than the individual model. Combination of GMM and DNN-based model is proposed and evaluated. When a combination of two methods is used to identify speaker accent, the likelihoods of both models are linearly coupled to produce a new score L_{comb} given by

$$L_{comp}^{n} = (1-\alpha)L_{GMM}^{n} + \alpha L_{DNN}^{n}, n = 1, 2, ..., N, \tag{3.9}$$

where L_{GMM}^{n} and L_{DNN}^{n} are the likelihoods produces of the n-th GMM and DNN based reverberant speech models, respectively. N is the number of accents registered and α is weighing coefficient. An accent with the maximum likelihood is decided as the target accent.

3.4 Experiment

3.4.1 Experimental setup

Firstly, the accent recognition was evaluated using two accent corpuses. The first corpus used in our experiments is the TIMIT [66] corpus, which consists of eight kinds of American English

Table 3.1: Details of corpus of native and non-native accent English

No.	Corpus	Accent	#Spearker	#Senakers/Spearker
1	ERJ	Japanese English (non-native)	100 for training; 30 for test	10 for training; 10 for test
2	TIMIT	American English (native)	100 for training 30 for test	10 for training; 10 for test

Table 3.2: Detail of recording conditions for impulse responds measurement, where RT60 (second) is reverberation time in room, S is small, and L is large.

No.	Room	RT60(s)
	(a) CENSREC-4 data for training	
1	Japanese style room	0.40
2	Japanese style bath	0.60
3	elevator hall	0.75
	b) RWCP data for test	
4	echo room (cylinder)	0.38
5	Tatami-floored room(S)	0.47
6	Tatami-floored room(L)	0.60
7	conference room(L)	0.73
8	echo room(L)	1.30

accents as native accent. The second corpus is the ERJ corpus [67] , which consists of English accent which was spoken by Japanese student as non-native accent. It is briefly listed in Table 3.1.

To evaluate the performance of accent recognition system in artificial reverent speech for sake of convenience, eight impulse responses were selected from the Real World Computing Partnership (RWCP) sound scene database [68] and the CENSREC-4 database [69], which were convoluted with clean speech to create artificial reverberant speech. That is, TIMIT and ERJ corpus were used as clean speech. The training data are composed of 50 male and female speakers with 10 utterances taken from each and then the test data are also composed of 15 male and female speakers with 10 utterances.

Table 3.2 lists the impulse responses for the training and test sets. For the RWCP database, impulse responses were measured at several positions 2 m from the microphone. For the CENSREC-4 database, impulse responses were measured at several positions 0.5 m from the microphone.

Table 3.3: Detail of recording conditions for impulse responds measurement, where RT60 (second) is reverberation time in room, S is small, and L is large.

Model	**Setups**	
GMM	Number of diagonal covariance matrices	128
DNN	No.layer No. of hidden layer nodes	3x1024
	Batch size	128
	Bernoulli-Bernoulli RBM of learning rate	0.02
	Gaussian-Bernoulli RBM of,learning	0.002
	Weight decay	0.0002
	Iteration of Pre-training	50
	Iteration of Fine-tuning	100
	Context size	9

Table 3.4: Speech analysis conditions for accent recognition

Sampling frequency	16,000 Hz
Frame length	25 ms
Feature shift	10 ms
Feature space	25 (12MFCCs+12△MFCCs+△Power)

Table 3.3 gives the model parameters for accent identification. Both models used 25-dimensional mel-frequency cepstral coefficients (MFCCs) to be the feature as listed in Table 3.4.

For model training, we created the accent models as listed in following Table 3.5, which differ in that either clean speech data or multi-condition artificial reverberant speech data for training. The description of each model is explained below. Finally the scores of each respected model are combined to improve identification performance.

Clean GMM: A GMM-based model was trained with clean speech and was tested with artificial reverberant speech with five types of RWCP impulse responses.

MC-GMM: A GMM-based model (Multi-Condition training GMM) was trained using artificial reverberant speech with three types of CENSREC-4 impulse responses and was tested with reverberant speech as above baseline.

MC-DNN: A DNN-based accent model (Multi-Condition training DNN) was trained and tested using the same speech with MC-GMM.

Table 3.5: Speech analysis conditions for accent recognition

No.	Accent model	Training data	Test data
1	Clean GMM	Clean speech	
2	MC-GMM	Reverberant speech	Reverberant speech
3	MC-DNN	Reverberant speech	

Table 3.6: Speech analysis conditions for accent recognition

No.	Accentmodel	TIMIT	ERJ	Average
1	Clean GMM	99.9	0.00	49.9
2	MC-GMM	90.4	91.0	90.7

3.4.2 Experimental result

Result of GMM

The result of GMM model for accent recognition is shown in Table 3.6. In general accent recognition, there are two kinds of accents: accents from foreign or non-native speakers and accents from native speakers who speaker the dialects of the language. Table 3.6 shows the accent recognition rates of TIMIT (native speaker accent), ERJ (non-native speaker accent) and the average of TIMIT and ERJ. Here, the average results of reverberant speech with five reverberation times (RT60 from 0.38 s to 1.30 s) are used.

For clean speech model, we can notice that the result of reverberant speech is (49.9%) which is almost similar to the random result (50.0%). It is noticed that the accent recognition rate of clean test speech using clean GMM is 99.0%. This indicates that the performance of accent recognition is degraded significantly in the distant-talking environment. Performance of GMM model, which uses the multi-condition training data, was improved from 49.9% (clean speech model) to 90.7%.

Figure 3.1 shows the accent recognition rates using MC-GMM based method of reverberant speech of five recording conditions shown in Table 3.2. The performance of non-native speaker accent (ERJ) recognition is degraded with longer reverberation time. However, the trend of the result of native speaker accent (TIMIT) is reverse.

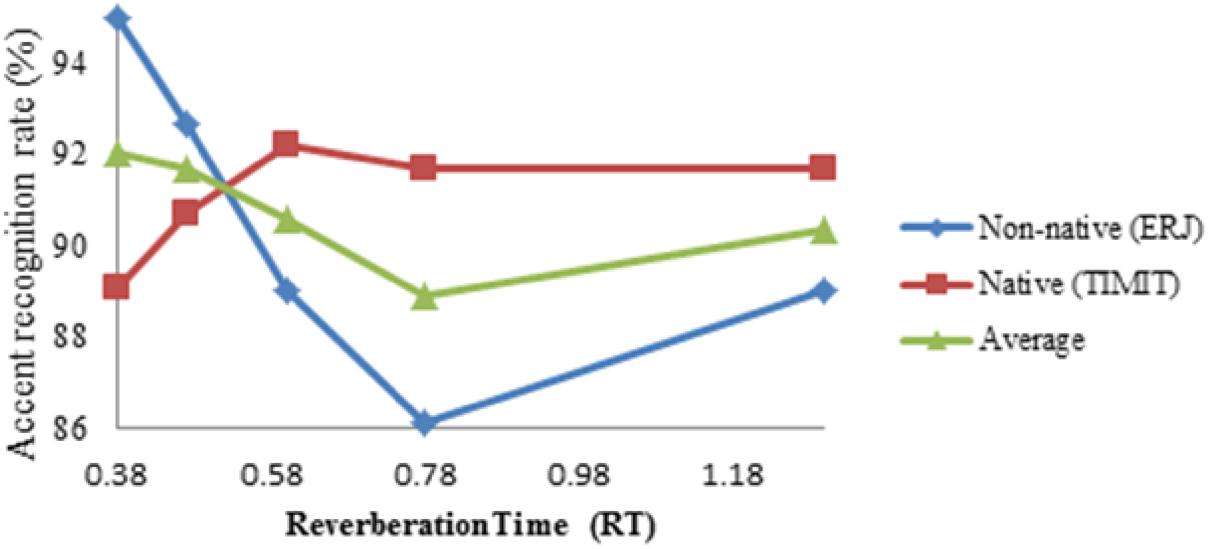


Figure 3.1: GMM based accent recognition rate of each reverberation time

Result of DNN

Prior to use DNN based model, different configurations of DNNs are investigated to obtain appropriate parameter for accent recognition. For this algorithm, the networks are initialized by the method proposed in [18]. After initialization, DNNs are trained using configuration as shown in following Table 3.3. The number of layers varies from 3 to 5 and the number of nodes in each hidden layer increases from 256 to 1024. The result of each layer and node is specified using Iteration of fine-tuning that increases from 10 to 100. The results of DNN with five and three hidden layers are shown in Figure 3.2 and 3.3, respectively. In Figure 3.2 and Figure 3.3, the average results of reverberant speech with five reverberation times (RT60 from 0.38 s to 1.30 s) of native speaker accent (TIMIT) and non-native speaker accent (ERJ) are used. From both figure, we can see that more iteration produces better accent recognition. It is found that the 1024 x 5 produce the best corresponding result for five hidden layer configuration. And 1024x3 produces the best corresponding result for three hidden layer configuration. To obtain optimal result, 1024x3 was chosen because it gives us the highest recognition rate as illustrated in Figure 3.3 with recognition rate of 93.0

Figure3.4 shows the accent recognition rates of MC-DNN based method of reverberant speech

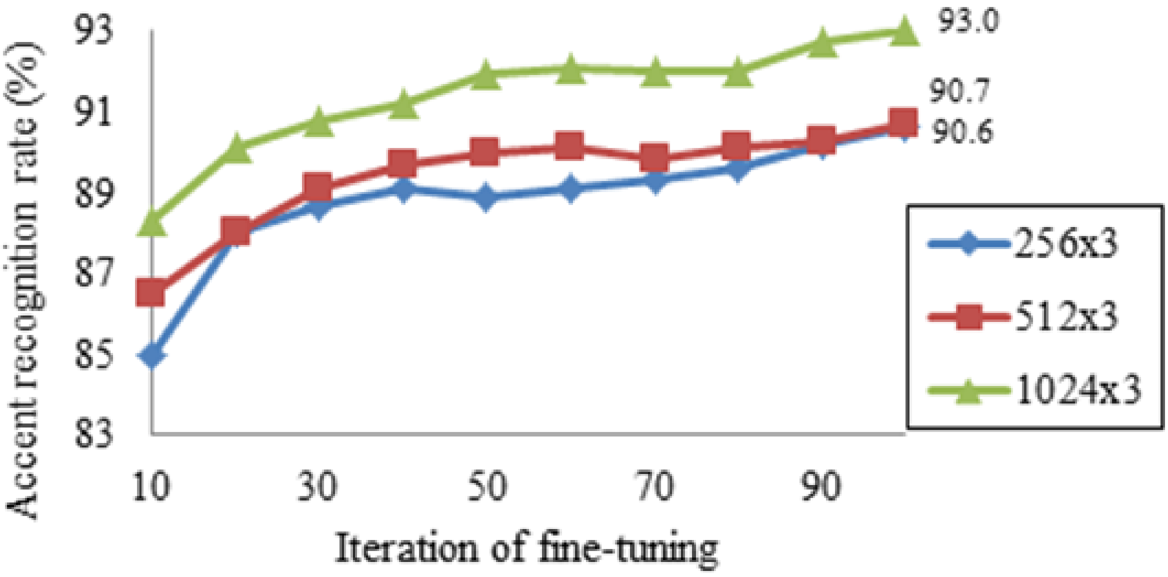


Figure 3.2: Results of DNN with three hidden layers.

of five recording conditions shown in Table 3.2. The trend of non-native speaker accent (ERJ) recognition performance and that of native speaker accent (TIMIT) recognition is reverse. Comparing with GMM-based method, the DNN-based method is robust to non-native speaker accent and not robust to native speaker accent. And Figure 3.5 summarizes the accent recognition rate of each reverberation time of GMM and DNN. Here, the average result of native speaker accent speech (TIMIT) and non-native speaker accent speech (ERJ) is used. DNN outperforms than GMM when the reverberation time is larger than 0.5 s. However, the results of GMM and DNN are similar when the reverberation time is 0.38 s.

Result by combination

In our previous study, the combination of GMM and HMM achieved better speaker recognition performance than individual model. In this section, the results of combination of GMM and DNN are presented.

We analyze the likelihood of both models as shown in Figure 3.6. It can be noticed that the likelihood of DNN with different accent provides clear boundary on non-native accent (ERJ) as noticed on the left of the figure. In contrast, the identification performance of native accent

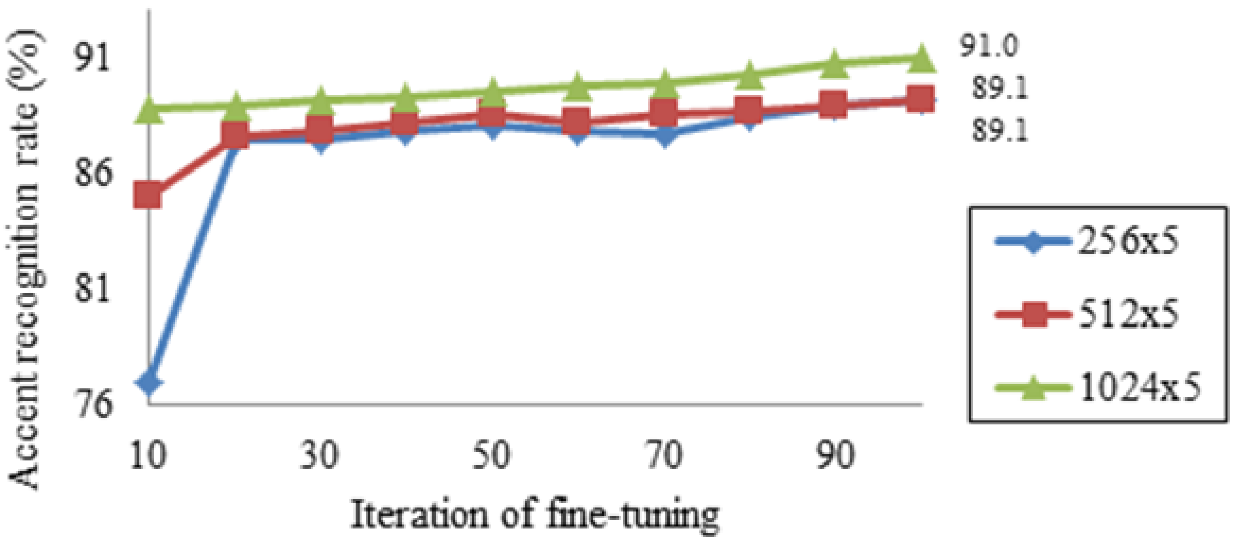


Figure 3.3: Results of DNN with five hidden layers.

(TIMIT) of DNN is worse than that of GMM as noticed on the right of the figure. We can notice that, the GMM and DNN based methods have complementary feature.

To obtain optimum performance, GMM and DNN based methods are combined using alpha parameter , which gives the weight to the DNN-based system and the GMM-based methods. Therefore Table 3.7 shows the detail of accent recognition rate of the combination method. It can be notice that, the best performance of 97.5% is achieved when weight coefficient equals to 0.6.It means that the likelihoods of GMM models are linearly coupled with DNN models to produce a new score at the rate of 6/4. This shows that the combination could improve the performance of both GMM and DNN respectively, which depend on parameter . The relative error reduction is 73.1% than the GMM-based method and 64.3% than the DNN-based method, respectively.

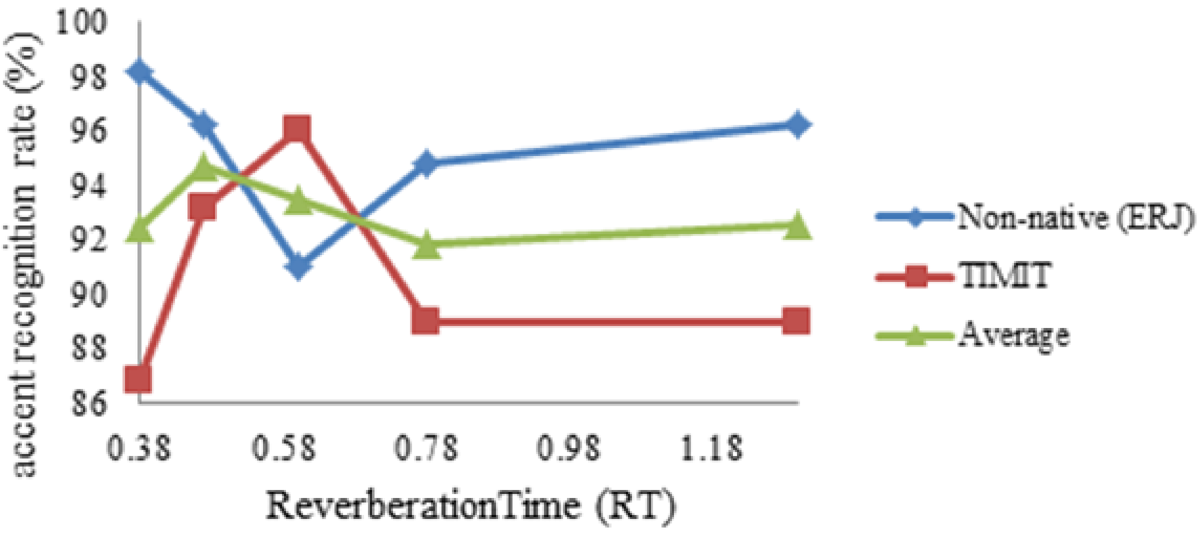


Figure 3.4: DNN based accent recognition rate of each reverberation time

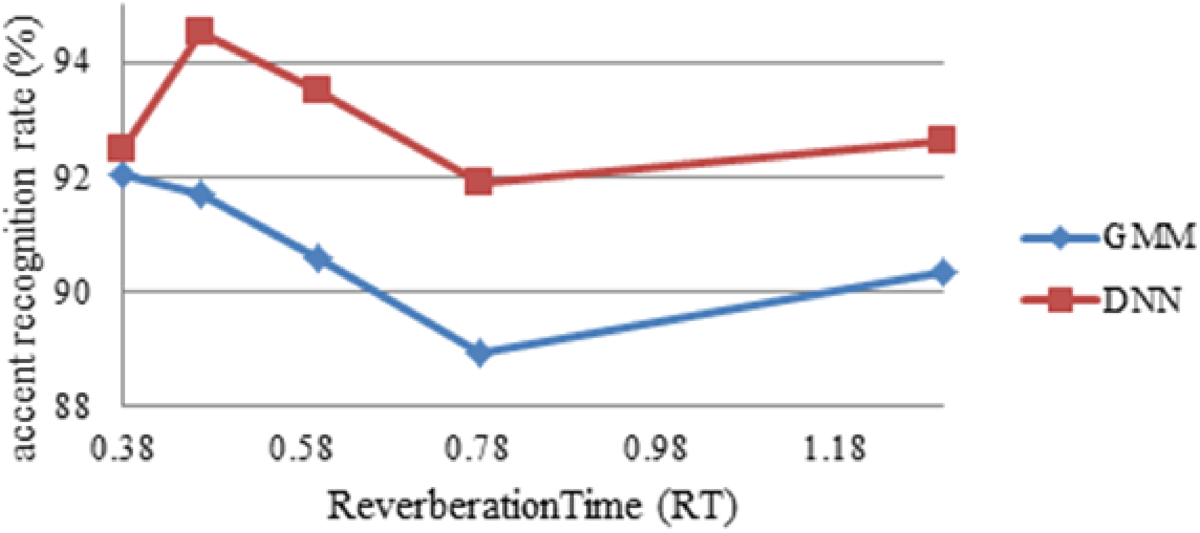
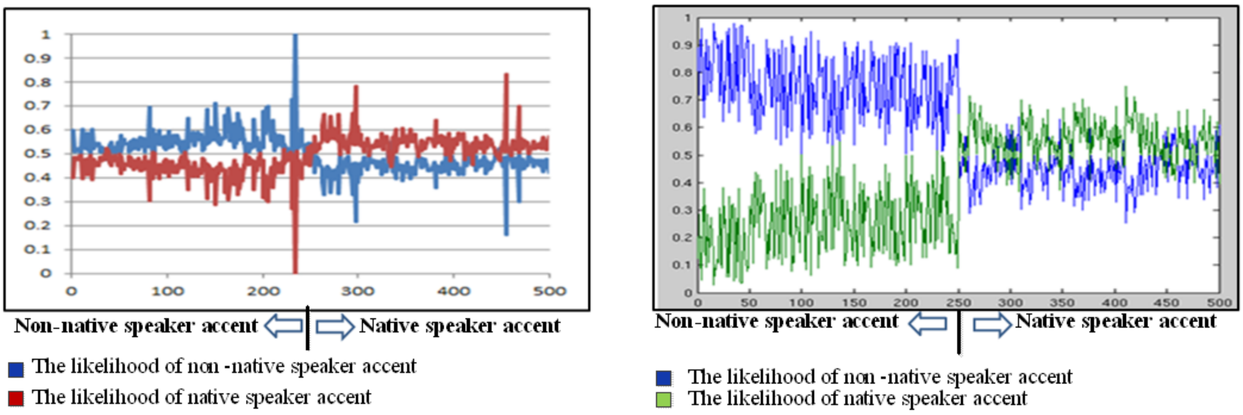


Figure 3.5: Accent recognition rate of each reverberation time of GMM and DNN



a) GMM-based model

b) DNN-based model

Figure 3.6: Accent recognition rate of each reverberation time of GMM and DNN

Table 3.7: The accent recognition rate of the combination method

Weight coefficient (α)	Recognition rate (%)		
	Non-native(ERJ)	Native(TIMIT)	Average
0 (GMM)	90.4	91.0	90.7
0.1	91.3	92.5	91.9
0.2	92.7	93.5	93.1
0.3	94.7	94.9	94.1
0.4	96.0	96.1	96.0
0.5	97.5	96.9	96.2
0.6	98.0	97.1	97.5
0.7	98.3	96.5	97.4
0.8	98.5	95.9	94.2
0.9	97.8	94.0	95.9
1 (DNN)	95.3	90.7	93.0

Chapter 4

The regression-based DNN for noise robust voice activity detection

4.1 Motivation of this proposal

In Chapter 2, the framework of voice activity detection (VAD) system has been discussed. the conventional VAD systems performs well in ideal conditional (e.g. in a clean condition). However, in a real-life application, the conventional system may not deal with adverse environment conditions when speech is corrupted by noise background (such as noise backgrounds in car, exhibition, restaurant, street, airport, station). This is because conventional feature requires a design with noise robustness.

To improve conventional features, feature enhancements have attracted attention in many speech processing tasks [44, 45, 46, 47, 48]. This is because of the better classifier performance obtained using the enhanced features. Deep neural networks (DNNs), which have been improved by the introduction of restricted Boltzmann machine (RBM)-based pretraining [70], have become popular for feature enhancement. [71] proposed a DNN-based feature enhancement (called a denoising autoencoder) that we here call DNN-based feature enhancement for noise-robust VAD. Compared with traditional feature-based VAD, the VAD using a DNN-based enhanced feature improves performance because the nonlinear mapping function can predict clean features from corrupted features, hence making the VAD decision better. [44] also applied DNN-based feature enhancement to speaker identification . Compared with the traditional method, the enhanced feature provided better speaker identification accuracy. However, the DNN-based feature enhancement has a weakness when evaluated on previously unseen data.

While DNN-based feature enhancement has been applied to many speech processing tasks, noise-aware training (NAT) [72] was introduced to solve the problem of poor DNN-based feature enhancement performance on unseen test data in the speech enhancement task. In [73], NAT-DNN-based feature enhancement obtained significantly better performance than conventional DNN-based feature enhancement because it could better predict clean features from corrupted features owing to the addition of noise information during DNN training. Although NAT-DNN-based feature enhancement can provide better performance than DNN-based feature enhancement, we observe that phase information is discarded during feature enhancement training because of the inflexibility of phase computation. Therefore, NAT-DNN-based feature enhancement may be improved by phase-aware training.

In this Chapter, two conventional DNN-based feature enhancements (DNN- and NAT-DNN-based feature enhancement) are used as baselines for enhancing individual magnitude- and phase-based features for noise-robust VAD. Although both DNN and NAT-DNN-based feature enhancement may achieve good noise-robust VAD performance, they have a weakness in that phase information is discarded during feature enhancement training. To overcome this weakness, this dissertation propose adding phase-aware training into DNN- and NAT-DNN-based feature enhancement, which we call DNN-based joint phase and magnitude feature (JPMF) enhancement (JPMF with DNN), and NAT-DNN-based JPMF enhancement (JPMF with NAT-DNN). Moreover, this dissertation apply a combined score of the magnitude- and phase-based features to improve the VAD performance.

4.2 Existing methods

This Section discuss two convention DNN based feature enhancement as existing methods where they will be compared to proposed methods for robust noise VAD. The detail are described as follows:

4.2.1 DNN-based feature enhancement

Neural networks (NNs) are universal mapping functions that can be used for both classification and regression problems [72]. Many researches have used NNs for feature enhancement for a quite some time. An NN with more than one hidden layer is usually called a deep NN (DNN). Recently, DNN has been improved because of the introduction of a pretraining algorithm [74] based on the RBM. This is why the deep structure of a DNN enables a much more efficient representation of many nonlinear transformations. In the past several years, DNNs have been utilized in many speech processing tasks such as acoustic modeling and feature enhancement. In this dissertation, we use DNN to enhance the MFCC magnitude feature. Our aim is to utilize the flexibility of a DNN to model the highly nonlinear and complicated mapping from a distorted MFCC to the underlying clean MFCC. Note that we also apply DNN to map distorted phase features, the mel-frequency delta phase (MFDP), to clean them, and the basic concepts of both methods are the same. Hence, we use a unified description for both features here.

The structure of the conventional DNN-based feature enhancement is shown in Figure 4.1 (a). On the left of the DNN in the figure, a sequence of feature vectors is generated from a noisy feature. To enhance the features, we use the MFCC as a magnitude feature and MFDP as a phase feature. To predict the clean feature of current feature (shown as gray in the figure), a sequence of features around the current vector is fed into the DNN. This allows the DNN to use context information to predict the clean feature vector. After the nonlinear transformation

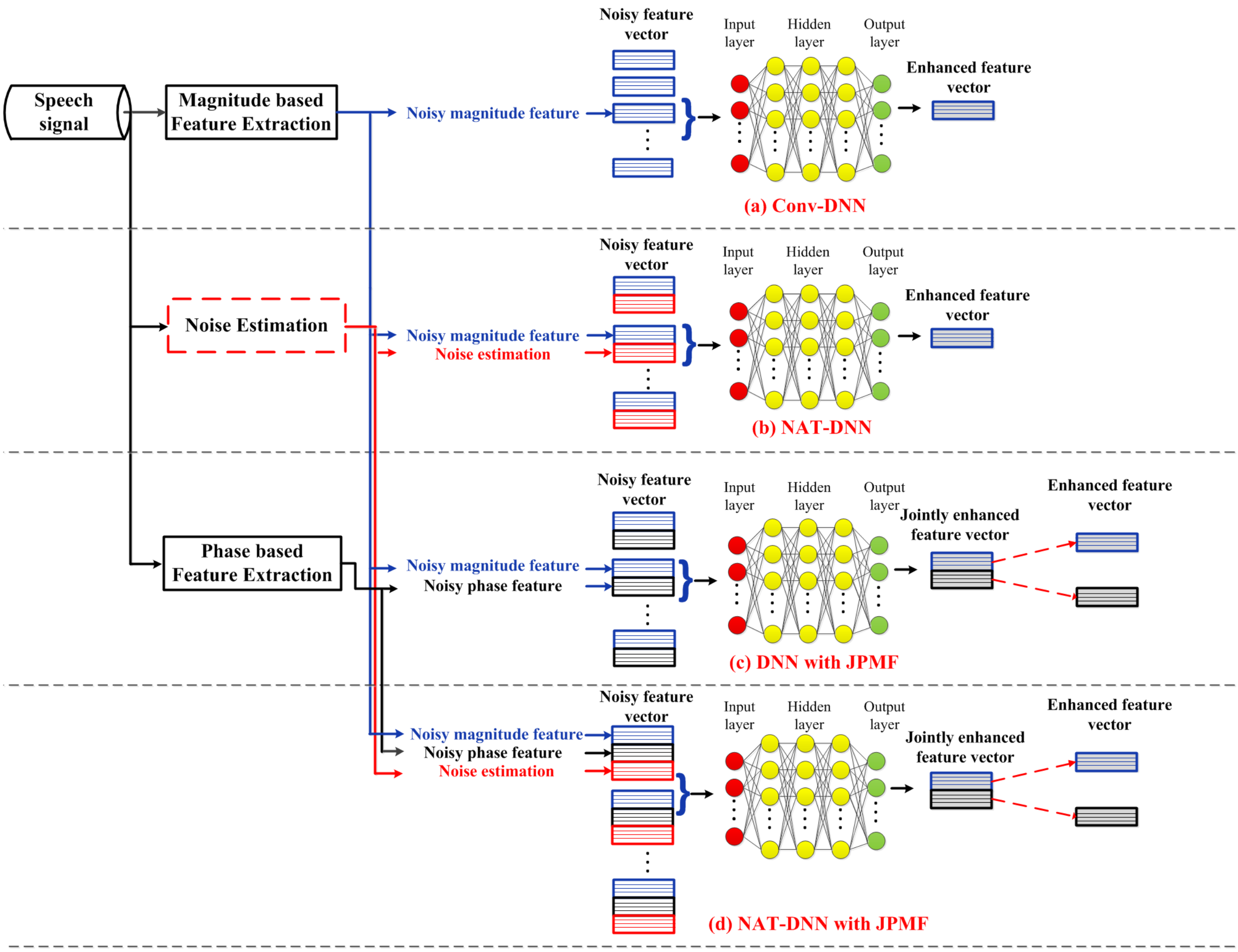


Figure 4.1: Structure of each feature enhancement type: (a) conventional DNN-based feature enhancement, (b) NAT-DNN-based feature enhancement, (c) JPMF with DNN, and (d) JPMF with NAT-DNN.

of the hidden layers, a linear output layer is used to predict the clean feature vector for the current frame.

To train the DNN for feature enhancement, parallel data comprising the clean and corrupted features of the same speech signal are required. The clean and corrupted feature vector sequences must be aligned accurately at frame level. We use clean and multi-condition data for training the DNN. The objective of training is to minimize mean square error (MSE) function between the output feature and the target features [75].

$$E_r = \frac{1}{N}\sum_{\eta=1}^{N} ||\hat{X}((Y_{\eta-\tau}^{\eta+\tau}, W, b) - X_\eta||_2^2. \tag{4.1}$$

Here, E_r is the MSE, $\hat{X}((Y_{\eta-\tau}^{\eta+\tau}), W, b)$ and X_η denote the estimated and reference normalized feature at sample index η, respectively, N represents the mini-batch size, $Y_{\eta-\tau}^{\eta+\tau}$ is the input feature, which is spliced at $\pm\tau$ context frames, W denotes the weight matrices, and b indicates the bias vector. The DNN parameters are then estimated iteratively by stochastic gradient decent [47] using the updated equation below.

$$\Delta(W_{i+1}, b_{i+1}) = -\lambda \frac{\partial E_r}{\partial(W_i, b_i)} - \kappa\lambda(W_i, b_i) + \omega\Delta(W_i, b_i). \tag{4.2}$$

Here, i denotes the number of update iterations, λ indicate the learning rate, κ is the weight decay coefficient, and ω is the momentum coefficient. This supervised training step is often called fine-tuning. Before the MSE step, the DNN is initialized by a pretrained RBM, which uses unsupervised training, and hence only a corrupted version of the speech is required. When the DNN is trained, it is expected to handle corrupted features well.

4.2.2 Noise aware training-DNN-based feature enhancement

In conventional DNN-based feature enhancement, the training is off-line because only a single magnitude feature is used for the regression function [72], as shown in Figure 4.1 (a). Although the mapping function can effectively deal with a previously seen noisy condition, an evaluation of mismatched noise types might affect the generalization capabilities of a non-speech segment owing to unseen noisy speech whose distortion characteristics are not similar to those of the training data. To solve this problem, NAT-DNN is applied to enable noise awareness. In this process, DNN is fed with a feature augmented with its corresponding estimation of the noise using a conventional minimum MSE (MMSE)-based noise estimation [76]. Hence, the DNN can use additional on-line noise information to improve the prediction of the clean feature. The input vector of the DNN with the noise estimate is constructed as follows:

$$\hat{Y}_{\eta-\tau}^{\eta+\tau} = [A_{\eta-\tau}, \hat{Z}_{\eta-\tau}, ..., A_\eta, \hat{Z}_\eta, ..., A_{\eta+\tau}, \hat{Z}_{\eta+\tau}] \tag{4.3}$$

where A_η represents the magnitude-based feature vector of the current noisy speech frame η, where the window size is $2 * \tau + 1$, and $\hat{Z}_\eta$ is the noise estimation based on MMSE [41]. In this procedure, the DNN is trained from the parallel data of the reference feature consisting of magnitude-based feature samples (MFCCs), like the traditional DNN and noise corrupted feature input vector $\hat{Y}_{\eta-\tau}^{\eta+\tau}$ of Eq. (3). The features are aligned at frame level. After training, the trained network can predict the underlying clean features when given noisy features, as shown in Figure 4.1 (b).

4.3 Proposed methods

In the methods described in the previous section, phase information is discarded during most of the feature enhancement training because of the inflexibility of phase computation. In this Section, integrating phase-aware training into two conventional DNN feature enhancements is proposed. By applying phase-aware training, the DNN-based feature enhancement is augmented with phase information during training, which is the proposed JPMF with the DNN method. Similarly, NAT-DNN-based feature enhancement can also be augmented with phase information during training, and this is the proposed JPMF with NAT-DNN method. The additional phase-aware training is expected to achieve better performance when compared with the conventional DNN feature enhancements described in Section 4.2.

4.3.1 Joint phase and magnitude feature enhancement (JPMF with DNN)

The DNN-based feature enhancement in Section 4.2.1 is only trained with magnitude information, hence phase information is missing during training. A complex spectrum includes the magnitude spectrum and phase spectrum. In [77, 78], the authors proposed a monaural speech separation in the complex domain. The experimental results show that complex traditional ratio masking outperforms ratio masking in the magnitude domain. That is, the results indicate that jointly enhancing the real and imaginary components can be better than enhancing the magnitude and phase independently. Joint enhancement of the magnitude and phase can improve speech quality, and is expected to enhance features well for VAD. To improve DNN-based feature enhancement, we propose JPMF with the DNN, which uses both the magnitude and phase information in one NN, which is expected to provide more accurate prediction than the DNN-based feature enhancement. In the training process, the input vector of the DNN is augmented using phase information as follows:

$$\bar{Y}_{\eta-\tau}^{\eta+\tau} = [P_{\eta-\tau}, A_{\eta-\tau}, ..., P_{\eta}, A_{\eta}, ..., P_{\eta+\tau}, A_{\eta+\tau}] \tag{4.4}$$

where P_η represents the phase-based feature vector of the current noisy speech frame η, where the window size is $2 * \tau + 1$. The DNN is trained on parallel data consisting of the reference feature with a clean JPMF and the input vector of the corrupted feature $\bar{Y}_{\eta-\tau}^{\eta+\tau}$. After training, the enhanced phase and magnitude features were derived separately from the jointly enhanced phase and magnitude information predicted by the trained network. Figure 4.1 (c) briefly shows the concept of JPMF enhancement.

4.3.2 Noise aware training-DNN-based joint phase and magnitude feature enhancement (JPMF with NAT-DNN)

The NAT-DNN-based feature enhancement described in Section 4.2.2 was introduced to solve the problem of DNN-based feature enhancement when the testing and training data do not match [73, 79]. However, phase information is discarded during the NAT-DNN-based feature enhancement training because of the complexity of the phase computation. Therefore, we introduce phase-aware training into traditional NAT-DNN-based feature enhancement. This is the proposed JPMF with the NAT-DNN method that uses magnitude information, phase information, and noise estimation in the DNN training. This method is expected to achieve more accurate prediction than the NAT-DNN-based feature enhancement. In the training process, the input vector of the NAT-DNN is augmented using phase information as follows:

$$\check{Y}_{\eta-\tau}^{\eta+\tau} = [P_{\eta-\tau}, A_{\eta-\tau}, \hat{Z}_{\eta-\tau}, ..., P_{\eta}, A_{\eta}, \hat{Z}_{\eta}, ..., P_{\eta+\tau}, A_{\eta+\tau}, \hat{Z}_{\eta+\tau}] \tag{4.5}$$

Figure 4.1 (d) briefly shows the concept of the JPMF. The DNN is trained from the parallel data of the reference feature with the clean feature of Eq. 5, $\check{Y}_{\eta-\tau}^{\eta+\tau}$, and the input vector of the corrupted JPMF feature. After training, the enhanced phase and magnitude features are derived separately from the jointly enhanced features of the phase and magnitude information predicted by the trained network.

4.3.3 Score combination

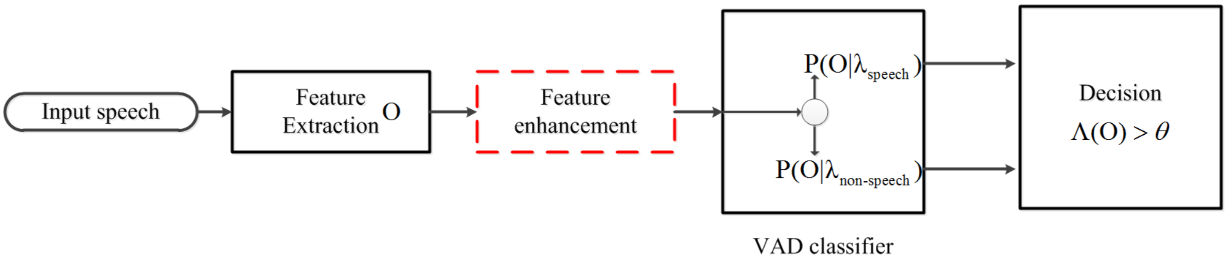


Figure 4.2: Flowchart of the VAD system

In the previous Section, both magnitude and phase feature are exploited in our method. Therefore, we propose a combination of the phase and magnitude feature scores to take advantage of the different benefits of these features. This technique is briefly described in this Section. A flowchart of the VAD system is shown in Figure 4.2. The SVM or DNN is used as a typical two-class classification for VAD. The decision about whether a given segment is speech or non-speech is based on the difference in probability that the segment is speech or non-speech.

$$\wedge(O) = p(O|\lambda_{speech}) - p(O|\lambda_{non-speech}), \tag{4.6}$$

where O is the feature vector of the input speech and λ_{speech} and $\lambda_{non-speech}$ are the models (SVM or DNN) of the speech and non-speech segments, respectively. Here, MFCC and MFDP are both usedas feature vector. To combine the scores, the probabilities obtained from the different features are combined by the following equation.

$$p(O_{comp}|\lambda_j) = \alpha p(O_{MFCC}|\lambda_j) + (1-\alpha)p(O_{MFDP}|\lambda_j). \tag{4.7}$$

$$p(O_{comp}|\lambda_j) = \alpha p(O_{MFCC}|\lambda_j) + (1-\alpha)p(O_{MFDP}|\lambda_j). \tag{4.8}$$

Where $p(O_{MFCC}|\lambda_j)$ and $p(O_{MFDP}|\lambda_j)$ are the probabilities based on MFCC and MFDP. Here, $j \in \{1, 2\}$ corresponds to class of speech or non-speech. In addition, α denotes the weighting coefficient.

4.4 Phase-based feature

In this work, we use two feature extraction methods to utilize both magnitude-based MFCC and phase-based MFDP where the details are described in section 2.1.1 and 2.1.3.

4.5 Experiment

4.5.1 Experimental setup

Our experiments were conducted on the CENSREC-1-C database [80]. The speech data were sampled at 16 kHz and finally downsampled to 8 kHz. The details of recording condition, utterances, and speaking style are the same as in CENSREC-1(AURORA-2J). To create the noisy speech, the samples were corrupted by two noise sets A and B, as shown in Table 4.1. Each noise set includes four different noise environments with SNRs from -5 dB to 20 dB in increments of 5 dB.

Figure 4.3 shows an overview a VAD system using the different kinds of feature enhancements introduced in Sections 4.2 and 4.3. The VAD system is fed with either MFCC or MFDP features. In Figure 4.4, the MFCC and MFDP features are separately enhanced with the same type of

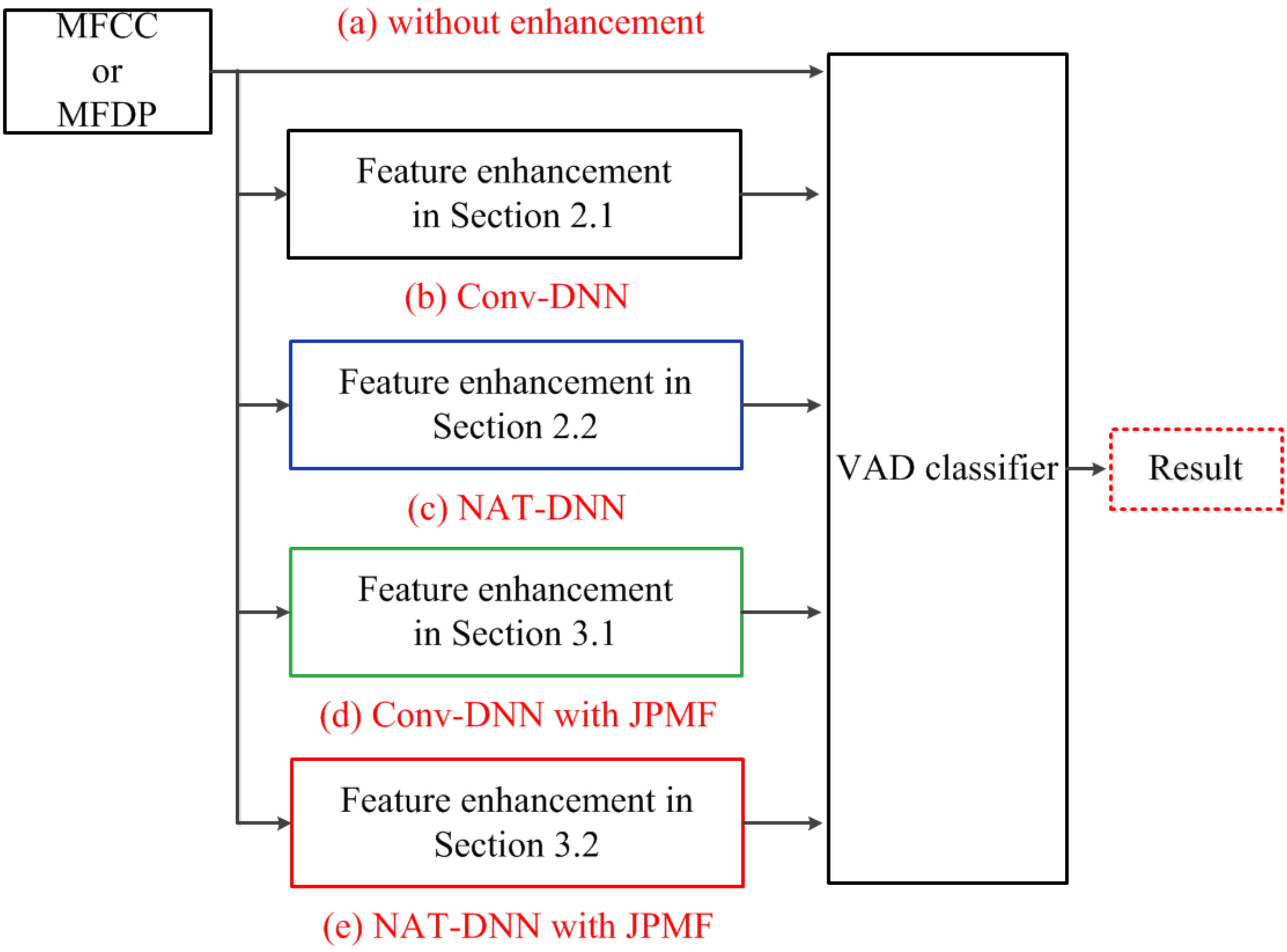

MFCC
or
MFDP
(a) without enhancement
Feature enhancement in Section 2.1
(b) Conv-DNN
Feature enhancement in Section 2.2
(c) NAT-DNN
Feature enhancement in Section 3.1
(d) Conv-DNN with JPMF
Feature enhancement in Section 3.2
(e) NAT-DNN with JPMF
VAD classifier
Result

Figure 4.3: Overview of the VAD system based on different feature enhancement approaches: (a) raw feature-based approach, (b) conventional DNN-based feature enhancement, (c) NAT-DNN-based feature enhancement, (d) JPMF with DNN, and (e) JPMF with NAT-DNN.

enhancement method, as shown in Figure 4.3 then their scores are obtained from the VAD classifier, which are then combines as described in Section 4.3.3. For the VAD detector, there are two classifier methods including SVM and DNN. Both classifier methods were trained using set A of the CENSREC-1-C database (artificially noisy speech segments with 19.39 hours were derived from subway, babble, car, and exhibition noise) and tested with set B of the CENSREC-1-C database (artificially noisy speech segments with 19.39 hours were derived from restaurant, street, airport, and station noise). For testing, the equal error rate (EER) is used as a measure of VAD performance. The result was evaluated using a broad range of SNR levels that were divided into three groups: high SNR, medium SNR, and low SNR. High SNR speech files were only corrupted by a small amount of noise and have SNR values of 20 dB and 15 dB. Medium SNR files have SNR values of 10 dB and 5 dB. To test the worst case scenario, the low SNR

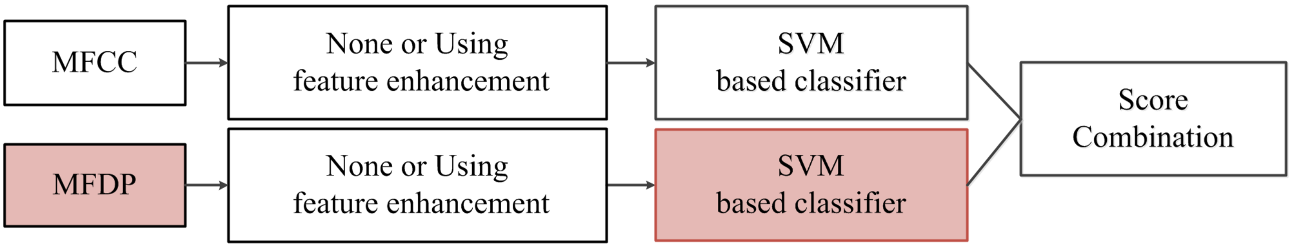

MFCC
MFDP
None or Using feature enhancement
None or Using feature enhancement
SVM based classifier
SVM based classifier
Score Combination

Figure 4.4: Overview of the VAD system with score combination.

Table 4.1: Noise environment in the CENSREC-1-C database.

Set	Additive Noises
A	Subway, Babble, Car, Exhibition
B	Restaurant, Street, Airport, Station

group was evaluated with SNR values of 0 dB and −5 dB.

To train the DNN, the multi-condition speech data of set A of the CENSREC-1-C database was used. Both MFCC and MFDP under the analysis conditions shown in Table 4.2 were tested. The input features consisted of seven spliced frames. A sigmoid type hidden layer was used for all layers except the input layer, where a linear hidden unit was used. To train the model for the feature enhancement approach we performed unsupervised RBM pretraining before supervised fine-tuning. To speed up the training, we performed RBM pretraining first. The Kaldi toolkit [81] was used for the pretraining task. The layers were trained in a layer-wise greedy fashion to maximize the likelihood over the training sample. The pretraining only requires a corrupted version of the utterance. For the back propagation to train the DNN, parallel data consisting of clean and distorted versions of the same utterance were used. The objective of this training is to minimize the MSE between the features. A stochastic gradient decent algorithm was used to improve the MSE error function. In the fine-tuning stage, the learning rate was 0.01, the

Table 4.2: Analysis conditions for MFCC and MFDP.

	MFCC	MFDP
Frame length	25 ms	
Frame shift	10 ms	
FFT size	512 point	
Dimensions	39 (13 MFCCs, 13 $\triangle$s, and 13$\triangle\triangle$ s)	39 (13 MFDPs, 13 $\triangle$s, and 13$\triangle\triangle$ s)

weight decay coefficient was 0.5, and the momentum was 0.5.

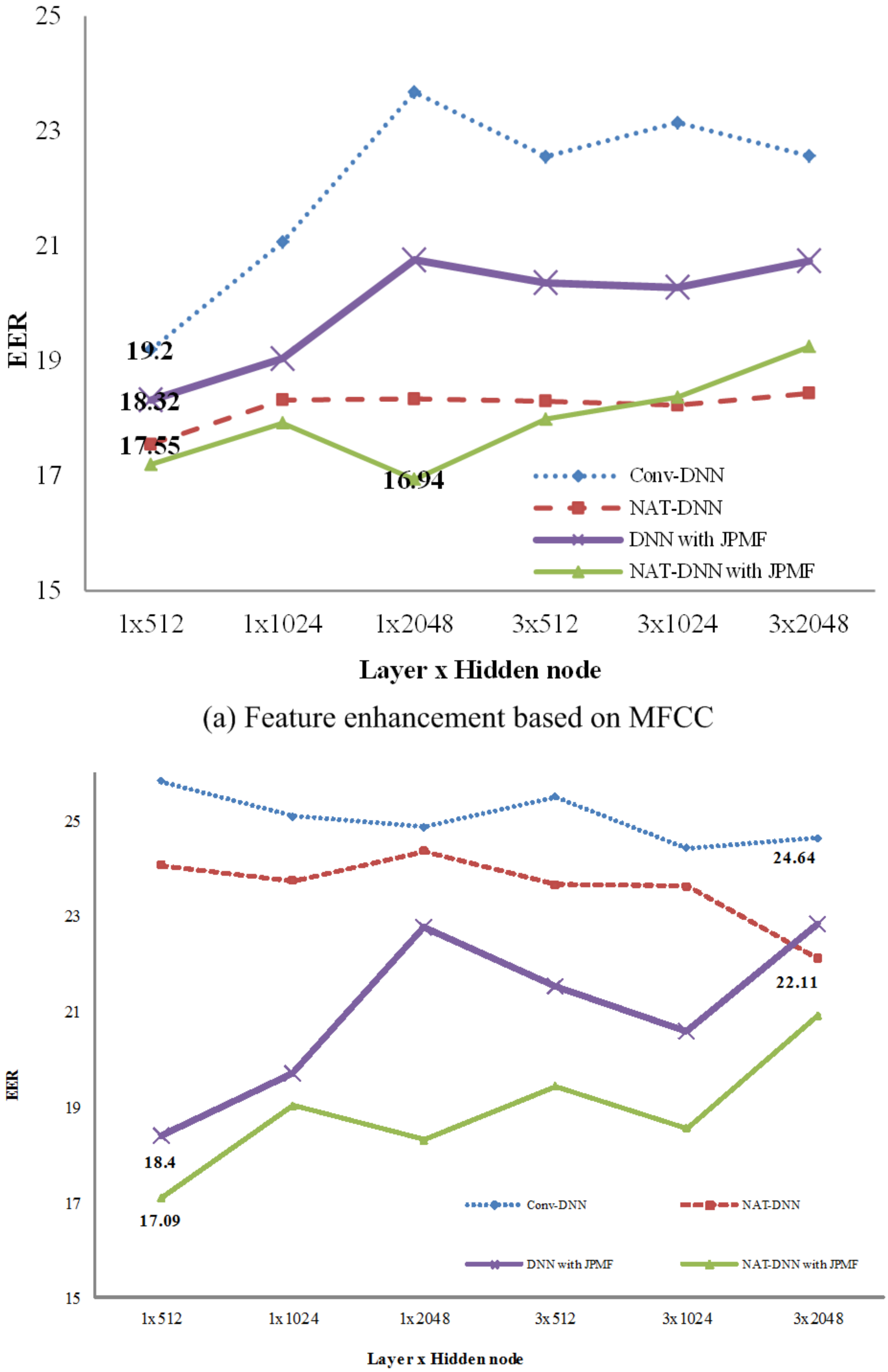
4.5.2 Experimental result

To evaluate the proposed JPMF enhancements for noise-robust VAD, this Section presents the results compared with two conventional DNN-based feature enhancements.

Results for the SVM classifier

In this subsection, SVM was used as a VAD to consider the performance of a single frame-based classifier which was based on [40]. To rapidly optimize the support vectors, we used the publicly available LIBLINEAR tool [82], which considers linear kernels for our experiment.

First, the results of VAD using unenhanced features (raw feature-based VAD) shown in Figure 4.3 (a) are first compared. The EERs are shown in Table 4.3 for MFCC and Table 4.4 for MFDP (rows highlighted by gray). Next, the VAD using different DNN of feature enhancement configurations (shown in Figure 4.3 (b)–(e)) were computed to achieve better performance. The DNNs of the enhanced features are trained using the parameters described in the subsection. The number of layers was one or three, and the number of nodes in each hidden layer was varied from 512 to 2,048. The results of each layer and node were specified using 30 fine-tuning iterations. The EER results of the DNN for each feature enhancement method are shown in Fig. 4.5 and show performance for MFCC (Fig. 4.5 (a)), and MFDP (Figure 4.5 (b)). The results show that feature enhancement using a DNN with three layers did not perform well according to our expectations when unseen noise was used in the VAD evaluation. This might be because training data are not sufficient. However, when the data were limited to seen noise, feature-enhanced VAD with all DNN configurations achieved better performance for both MFCC and MFDP than the VAD based on raw features (Tables 4.3 and 4.4). We selected the configuration with the best results for later experiments from Figure 4.5.



(a) Feature enhancement based on MFCC

(b) Feature enhancement based on MFDP

Figure 4.5: EERs of each enhancement system shown in Fig. 4.3 (b)(e): (a) using MFCC and (b) using MFDP.

Table 4.3: EER (%) of the SVM classifier based on MFCC: None is the result of the system in Figure 4.3(a), Conv-DNN is the result of the system in Figure 4.3(b), NAT-DNN is result of the system in Figure 4.3(c), DNN with JPMF is the result of the system in Figure 4.3(d), and NAT-DNN with JPMF is the result of the system in Fig. 4.3(e). Avg is short for average.

	Restaurant			Street			Airport			Station			**Avg**
Enhanced **method**	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	16.47	22.86	35.03	13.31	21.25	33.72	16.34	22.86	34.23	11.26	19.73	32.71	23.31
Conv-DNN	8.55	18.59	34.41	9.41	17.67	31.67	10.00	17.39	32.49	5.75	13.56	30.93	19.20
NAT-DNN	**5.19**	15.79	33.01	**5.48**	18.21	33.34	**7.39**	16.16	31.20	3.32	10.66	30.80	17.55
DNN with JPMF	7.21	18.13	34.20	7.41	**16.00**	**30.78**	9.53	17.23	32.20	4.39	12.46	30.31	18.32
NAT-DNN with JPMF	5.62	**15.26**	**31.95**	5.64	17.29	33.32	7.78	**14.78**	**29.64**	3.25	**9.48**	**29.31**	**16.94**

Table 4.4: EER (%) of the SVM classifier based on MFDP: Avg is short for average.

	Restaurant			Street			Airport			Station			**Avg**
Enhanced **method**	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	22.61	35.22	44.22	16.3	25.25	39.53	19.74	29.89	42.34	16.95	27.40	40.60	30.00
Conv-DNN	14.49	23.54	38.44	12.64	21.77	36.82	14.22	22.05	36.13	15.24	23.16	34.54	24.42
NAT-DNN	12.95	21.56	35.91	11.28	21.58	35.46	11.94	19.05	31.89	12.81	19.05	31.89	22.11
DNN with JPMF	**6.20**	14.47	31.55	10.58	**19.05**	**32.64**	8.07	14.87	30.3	6.57	14.12	32.34	18.4
NAT-DNN with JPMF	5.33	**13.47**	**31.62**	**7.49**	19.79	32.86	**6.12**	**13.61**	**29.06**	**4.25**	**10.97**	**30.49**	**17.09**

Table 4.3 shows the equal error rates (EERs) of SVM classifier using a magnitude-based feature (MFCC). By applying the feature enhancement with noise-robust VAD, the EERs were improved from the raw MFCC-based VAD in [83]. Similarly, the EERs were also improved from the raw MFCC-based VAD by applying the NAT-DNN-based feature enhancement. This is because of the enhanced MFCC using DNN. Moreover, the results show that the proposed feature enhancement, which uses both magnitude and phase information in one NN, provides better performance than conventional DNN-based feature enhancements. Using JPMF with DNN, the EERs were better than those of the DNN-based feature enhancement. In same way, the EERs of JPMF with the NAT-DNN were better than those of the NAT-DNN-based feature enhancement. This is because these methods better predict features, as the training input contains phase information to make DNN more efficient.

Table 4.4 shows the EERs of SVM classifier using the phase-based feature (MFDP). The abbreviations of the methods are the same as those in Table 4.3.

By applying the DNN-based feature enhancement and the NAT-DNN-based feature enhancement, the EERs were improved compared with those of the raw MFDP. Therefore, we can confirm that the DNN enhancement was also effective for phase features. Moreover, applying JPMF with DNN and JPMF with NAT-DNN have the same tendency as for the magnitude feature. This is because the DNN can use both magnitude and phase information for the enhancement, and hence more accurate clean features can be estimated.

Result of DNN classifier

In this section, the DNN was used as VAD detector to consider the effect of a multi-frame-based classifier. SignalGraph [84] was used to train the DNN. The DNN has one layer containing 512 neurons. The input feature for the DNN contains nine frames, and cross entropy was used. The learning rate started from 0.1 and was changed to 1 for the second epoch. It then decayed by a factor of 0.5 each time the cross entropy on a cross validation set between two consecutive epochs increased. The features used were the same as those used by the SVM classifier.

Before using feature enhancements, we investigated the DNN configurations with 1, 2 and 3 layers, each using raw MFCC as the input [85]. The results are shown in Table 4.5. DNN-VAD-based on more hidden layers did not perform according to our expectations. It might not be effective to do this if we simply consider VAD as a binary-class classification problem with the noisy speech and the background noise as the two classes. Therefore, we selected the DNN-VAD-based on one hidden layer with feature enhancements.

Table 4.5: EER (%) of DNN classifiers with different numbers of layers.

	Restaurant			Street			Airport			Station			Avg
# Number layer	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
1	3.57	11.14	32.50	3.40	9.67	27.35	5.54	11.48	26.60	2.88	8.21	27.75	**14.17**
2	3.78	11.30	32.58	3.44	9.75	28.89	5.28	11.11	26.13	2.88	8.21	27.66	14.25
3	3.94	11.33	32.63	3.39	9.84	28.95	5.41	11.35	26.42	2.88	8.02	27.82	14.33

Table 4.6: EER (%) of the DNN classifier based on MFCC: Ave is short for average.

	Restaurant			Street			Airport			Station			Ave
Enhanced method	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	3.57	11.14	32.5	3.4	**9.67**	27.35	5.54	11.48	**26.6**	2.88	8.21	**27.75**	14.17
Conv-DNN	3.95	11.79	31.77	4.03	8.98	24.87	5.88	12.37	28.27	3.57	9.43	28.82	14.48
NAT-DNN	**2.69**	8.92	30.64	2.84	11.61	28.76	**4.08**	11.96	28.46	**2.49**	8.39	30.5	14.28
DNN with JPMF	3.51	10.75	30.63	3.19	8.99	**24.86**	5.75	12.49	29.11	3.24	9.38	28.69	14.21
NAT-DNN with JPMF	2.78	**8.91**	**28.52**	**2.71**	11.25	28.37	4.49	**11.24**	26.93	2.51	**7.62**	28.59	**13.66**

Table 4.7: EER (%) of the DNN classifier based on MFDP: Avg is short for average.

	Restaurant			Street			Airport			Station			Avg
Enhanced method	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	14.20	23.9	39.08	12.17	22.11	37.69	12.90	21.58	36.28	13.92	22.49	35.24	24.29
Conv-DNN	14.11	23.51	38.34	11.45	20.7	35.89	13.30	21.28	35.49	13.95	21.60	34.49	23.67
NAT-DNN	11.96	20.44	36.78	10.25	21.55	35.22	11.44	18.67	33.89	12.81	20.07	37.18	22.52
DNN with JPMF	5.37	14.22	32.61	7.37	**16.10**	**30.64**	7.88	14.86	30.38	5.38	13.55	31.82	17.51
NAT-DNN with JPMF	**3.58**	**10.85**	**26.42**	**4.25**	16.73	31.25	**5.18**	**12.69**	**29.55**	**3.28**	**9.33**	**31.79**	**15.41**

Table 4.6 shows the EERs of the DNN classifier using MFCC. When applying the most feature enhancement, the EER results did not perform to our expectations. This is because multi-frame-based classification requires significantly enhanced features. However, JPMF with NAT-DNN could provide better performance than the other features because of its significantly enhanced feature.

Table 4.7 shows the EERs of the DNN classifier using MFDP. When feature enhancement is applied, the results have the same tendencies as the SVM classifier. This is because the DNNs significantly contribute to the performance of the phase-based feature.

Result of score combination

This Section reports the results of combining the MFCC and MFDP scores. Table 4.8 shows the results of score combination based on the SVM, and Table 4.9 shows the results of score combination based on the DNN. We can see that the VAD based on the combined score outperformed the systems using individual feature. This is because it takes advantage of the combination of different decisions of the systems.

Table 4.8: EER (%) of SVM classifier based score combination of MFCC and MFDP.

	Restaurant			Street			Airport			Station			**Avg**
Enhanced **method**	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	12.14	19.99	33.33	8.68	15.62	32.12	11.86	18.51	31.15	7.53	13.50	30.22	19.55
Conv-DNN	7.06	15.82	33.06	5.69	15.24	30.75	8.57	15.34	30.52	5.08	12.32	30.31	17.48
NAT-DNN	4.53	14.59	33.09	**4.65**	14.06	33.28	7.00	14.93	30.78	3.29	10.49	30.84	16.79
DNN with JPMF	5.9	14.23	31.59	7.19	**14.04**	**30.51**	8.04	14.73	30.11	4.28	12.1	30.41	16.93
NAT-DNN with JPMF	**4.41**	**13.01**	**31.57**	6.11	14.58	30.98	**5.97**	**13.47**	**29.17**	**2.95**	**9.16**	**29.82**	**15.93**

Table 4.9: EER (%) of DNN classifier based score combination of MFCC and MFDP.

	Restaurant			Street			Airport			Station			Avg
Enhanced method	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	High SNR	Medium SNR	Low SNR	All SNR
None	3.43	9.93	31.1	3.11	8.55	25.97	5.51	11.31	**25.77**	2.85	8.07	**27.68**	13.61
Conv-DNN	3.76	10.81	30.86	3.99	8.13	26.50	5.91	12.28	27.12	3.47	9.17	28.16	14.18
NAT-DNN	**2.62**	8.92	29.49	2.84	11.61	28.76	3.97	11.84	27.86	2.45	7.80	30.00	14.01
DNN with JPMF	3.42	10.44	30.13	3.28	**8.07**	**24.04**	5.75	12.29	29.00	3.18	9.18	28.66	13.95
NAT-DNN with JPMF	2.77	**8.87**	**28.41**	**2.7**	11.25	26.95	**4.52**	**11.2**	26.81	**2.43**	**7.44**	28.42	**13.48**

Analytic illustration of noise suppression

To better visualize the enhanced magnitude and phase-based feature described in the previous Sections, this Section displays the spectrogram of the enhanced MFCC and MFDP features and their scores.

Figure 4.6 shows MFCC feature spectrograms of an utterance example corrupted by stationary noise at SNR = 0. The spectrograms of noisy MFCC, clean MFCC, and MFCCs enhanced by the conventional DNN, NAT-DNN, JPMF with DNN (the first proposed method), and JPMF with NAT-DNN (the second proposed method) are shown in Figures 4.6 (b)–(g). Comparing Figure 4.6 (b) with Figures. 4.6 (d)–(f), the spectrograms using feature enhancement provide better speech/non-speech segment boundaries than those of the raw feature. This is because of DNN enhancement. To observe how the proposed method works, the spectrogram of Figure. 4.6 (d) can be compared with that of Figure 4.6 (f). It is clear that the enhanced MFCC using JPMF with the DNN provides a better boundary between the speech segment than that using the conventional DNN. The same tendency can be found for NAT-DNN. This is because of the phase-aware training. Hence, we confirmed that introducing JPMF in both conventional DNN and NAT-DNN training improves noisy MFCC features.

Figure 4.7 shows the MFDP spectrograms of an utterance example corrupted by stationary noise at SNR = 0. The spectrograms of noisy MFDP, clean MFDP, and MFDP enhanced by the conventional DNN, NAT-DNN, JPMF with the DNN (the first proposed method), and JPMF with NAT-DNN (the second proposed method) are shown in Figures 4.7 (b)–(g). Comparing Figure 4.7 (b) with Figures. 4.7 (d)–(f), spectrograms using feature enhancement provide better speech/non-speech segment boundaries than those using the raw feature. Again, this is because of DNN enhancement. To observe the proposed method, the spectrograms of Figure 4.7 (d) with Figure 4.7 (f) can be compared. Clearly, the enhanced MFCC using JPMF with DNN provide better boundaries between speech segments than those using conventional DNN and the same tendency can be found for NAT-DNN. This is due to introducing the magnitude information during DNN training.

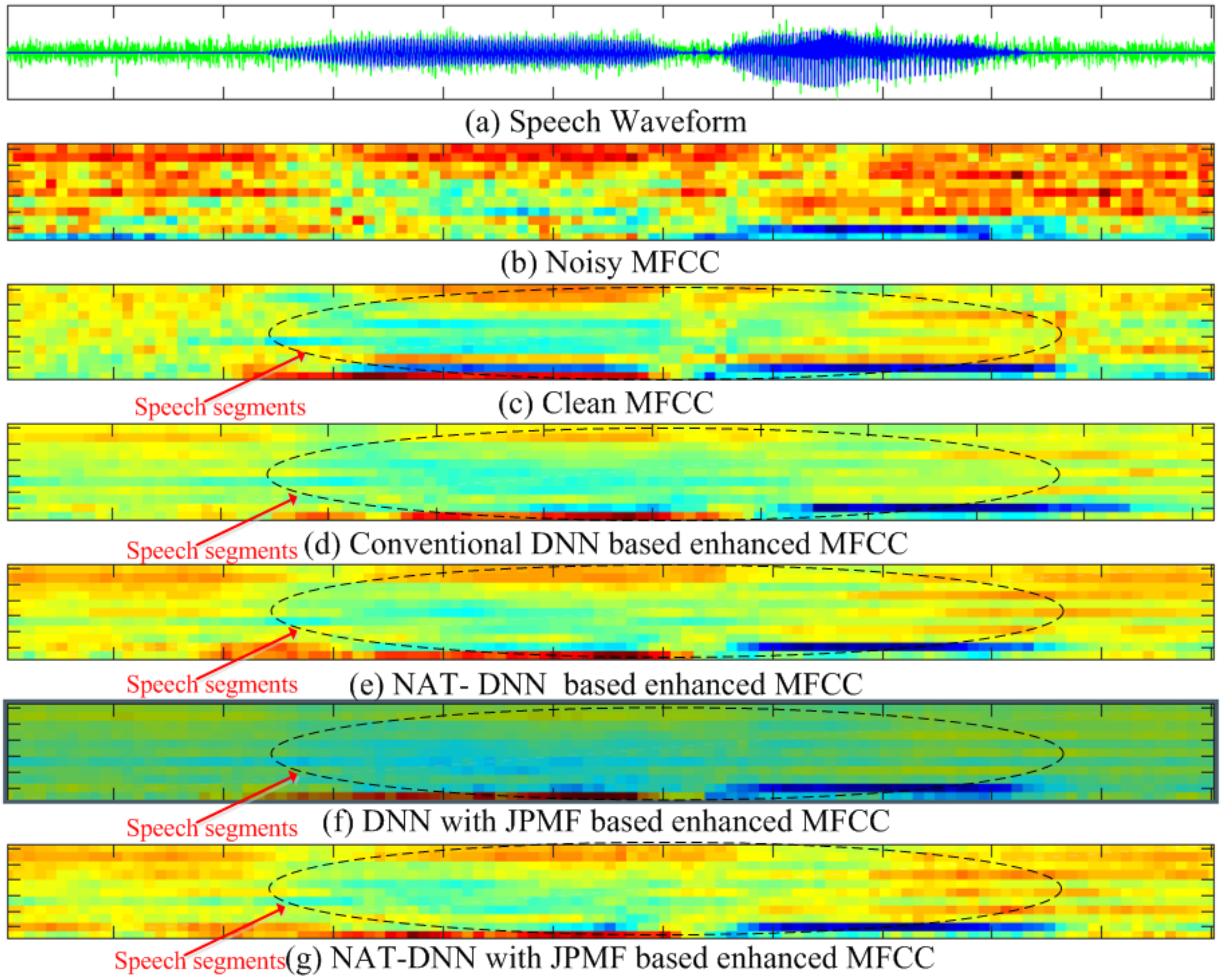


Figure 4.6: Spectrograms and VAD score of each enhancement method based on the MFCC feature: (a) the speech waveform, where the blue line is clean speech and the green line is stationary noise at SNR = 0, (b) noisy MFCC, (c) clean MFCC, (d) DNN-based enhanced MFCC, (e) NAT-DNN-based enhanced MFCC, (f) enhanced MFCC using JPMF with DNN, and (g) enhanced MFCC using JPMF with NAT-DNN.

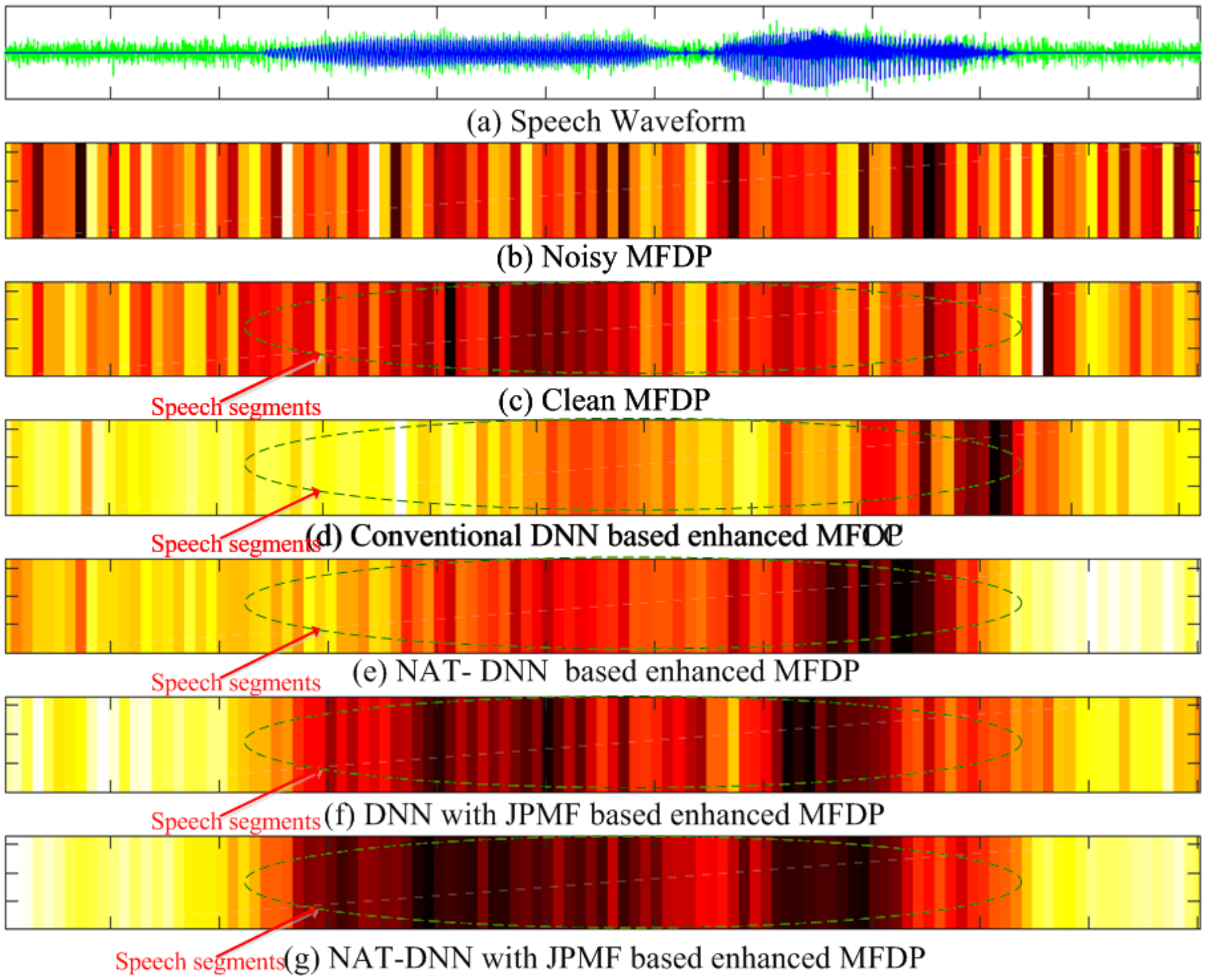


Figure 4.7: Spectrograms and VAD scores of each enhancement method based on the MFDP feature: (a) the speech waveform, where the blue line is clean speech and the green line is stationary noise at SNR = 0, (b) noisy MFDP, (c) clean MFDP, (d) DNN-based enhanced MFDP, (e) NAT-DNN-based enhanced MFDP, (f) enhanced MFDP using JPMF with DNN, and (g) enhanced MFDP using JPMF with NAT-DNN .

Chapter 5

The classification-based DNN for noise robust voice activity detection

5.1 Motivation of this proposal

In previous Chapter 4, several feature enhancement based on deep neural network (DNN) are investigated for voice activity detection (VAD). Especially, DNN-based joint phase- and magnitude -based feature (JPMF) enhancement (JPMF with DNN) and a noise-aware training (NAT)-DNN based JPMF enhancement (JPMF with NAT-DNN) are proposed for noise-robust voice activity detection (VAD). However, classification-based DNN using magnitude and phase information has been discarded. In this Chapter, we focus on a deep neural network (DNN) using magnitude and phase information (that is, phase aware DNN) to obtain better VAD performance than DNN using magnitude information.

To classify speech and non speech segment, many features have been considered for machine-learning based VAD. In [86], perceptual linear predictive coefficients (PLP) was proposed to distinguish speech and non-speech segment from speech sample. In [49], Mel-frequency cepstral coefficient (MFCC) was also proposed to distinguish speech and non-speech segment. In [87], concatenating feature from PLP, MFCC, pitch, discrete Fourier transform (DFT), and amplitude modulation spectrograms (AMS) was applied to distinguish speech and non-speech segment. These features have been proven to be powerful for VAD. However, they may have weakness due to missing phase information, which is the half of information present in original signal. DNN-based classifier achieves the best performance for VAD, but almost DNN-based VAD only as magnitude information. The phase information has been proven to be important for many speech processing tasks [88, 89, 90]. Therefore, this study proposes a phase aware DNN which jointly uses magnitude and phase information for noise robust VAD to determine speech segment or non speech segment.

The joint use of magnitude and phase feature was motivated to receive a set of practical features for robust noise VAD under low SNRs, without loss of phase information or magnitude information. To generate jointly magnitude and phase feature, we use both magnitude and phase features in our experiment. In magnitude based features, MFCC and power-normalized cepstral coefficients (PNCC) [50], that provide better result than MFCC for noise robust speech recognition, are used. For phase based features, instantaneous frequency derivative (IF) [55] derived from the derivative of the phase along time axis, baseband phase difference (BPD) [54] derived from difference of baseband short time Fourier transform, and modified group delay cepstral coefficient (MGDCC) [91] derived from negative derivative of the Fourier transform phase are applied. These features will be variously concatenated in serial. They are expected to achieve a better performance than single magnitude based feature.

5.2 Existing method

support vector machine (SVM) and deep neural network (DNN)-based VAD using MFCC introduced in Chapter 2 are used as existing method.

5.3 Proposed method

5.3.1 Phase aware DNN

The flowchart of the VAD system is shown in Figure 5.1. In this study, a deep neural network (DNN) is used as speech/non-speech detector. The DNN output refers to the posterior probabilities of two classes, including probability of speech segment, ϱ_{speech} and non-speech segment, $\varrho_{non-speech}$. The decision about whether speech is speech or non-speech segment is based on the posterior probability differences:

$$\wedge(I) = P(\varrho_{speech}|I) - P(\varrho_{non-speech}|I) \tag{5.1}$$

where I is the feature vector of input speech. Power-Normalized cepstral coefficients (PNCC),instantaneous frequency derivative (IF), baseband phase difference (BPD) and modified group delay (MGDCC) described in Section 5.4 are used. For conventional DNN-based VAD, only magnitude based feature is used as the feature vector of input speech as follows,

$$I = F_{mag} \tag{5.2}$$

where F_{mag} only uses magnitude based feature. In this paper, MFCC or/and PNCC is/are used. The structure of magnitude based DNN is shown in Figure 5.2(a). In [85] , conventional DNN-VAD was successfully applied to magnitude feature (MFCC). However, MFCC only contains the magnitude information of the speech, therefore, the speech/non-speech classifier might be incomplete.

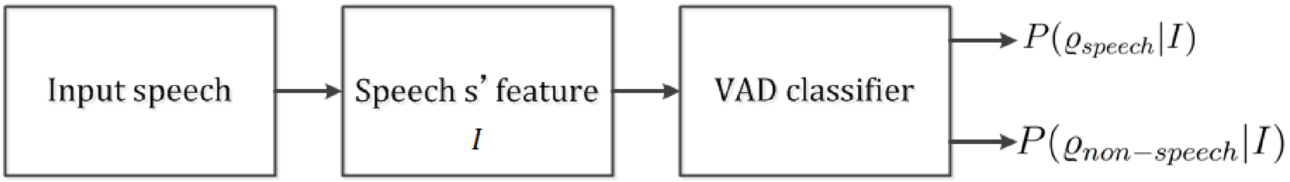


Figure 5.1: Flowchart of VAD system.

Recently, the phase information has been proven to be important for many speech processing tasks [88, 92]. In previous work[90], phase feature (MGDCC) augmented with corresponding MFCC could improve the performance of the DNN training as a regression task. This can be seen in improvement in performance in simultaneous enhancement of amplitude and phase feature. With this in mind, we expect that phase information can also improve performance of the DNN training as speech/non-speech classifier. Therefore, this study proposes a phase aware DNN which jointly uses magnitude and phase information for noise robust VAD to determine speech segment or non speech segment. The structure of phase aware DNN is shown in Figure 5.2(b). The feature vector of input speech, covering magnitude and phase information, is used as follows,

$$I = [F_{mag}, F_{phase}] \tag{5.3}$$

where F_{phase} is phase based features. MGDCC, IF, BPD, dual phase feature (augmenting MGDCC and IF, MGDCC and BPD, or of IF and BPD), or triple phase features (augmenting MGDCC, IF, BPD) is used to be augmented with magnitude based feature.

5.4 Magnitude and phase based feature

Five features are used in this work. They include two magnitude-based features, namely Mel-frequency cepstral coefficients (MFCC) and Power-Normalized cepstral coefficients (PNCC); and three phase-based features, namely instantaneous frequency derivative (IF), baseband phase difference (BPD), and modified group delay (MGDCC).

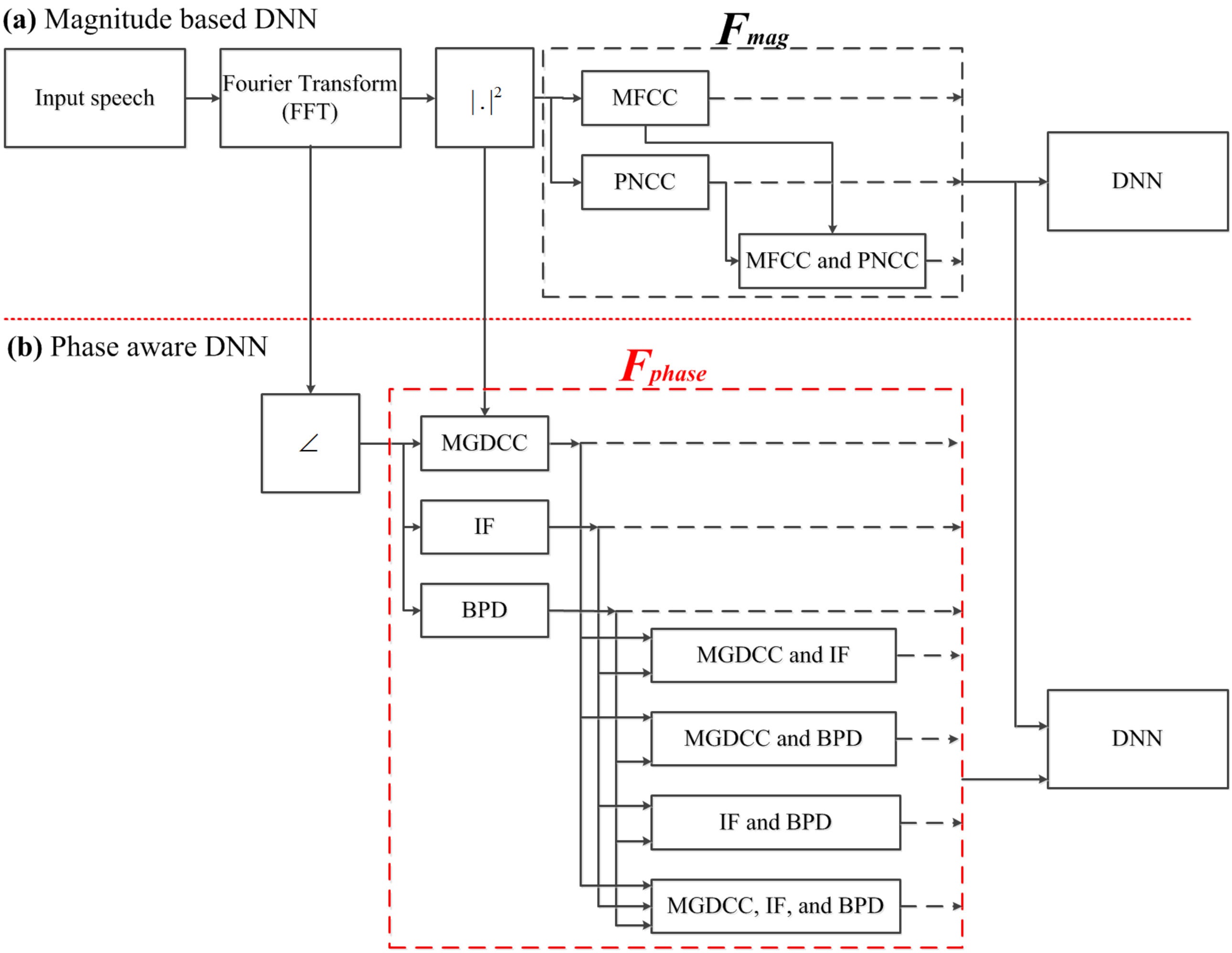


Figure 5.2: A block diagram of phase aware DNN.

- **MFCC** [49] is the most popular feature for speech processing including voice activity detection. We used MFCC as an amplitude feature for the VAD input.

- **PNCC** is another feature based on magnitude information. It has been developed to enhance the robustness of speech recognition systems under noisy condition. The major innovations of this feature when comparing with MFCC are the use of a power-law nonlinearity that replaces the traditional log nonlinearity used in MFCC coefficients, a noise-suppression algorithm based on asymmetric filtering that suppress background excitation, and a module that accomplishes temporal masking, which in detail is found in [50]. So, this method may be a useful feature for VAD when speech is corrupted by noise.

- **IF** instantaneous frequency [55] is a phase feature designed to provide clearer phase pattern than original phase spectrum that hardly displays any patterns. The computational algorithm

Table 5.1: Analysis conditions for MFCC, PNCC, IF, BPD and MGDCC.

	Dimension	**Frame length**	**Frame shift**	**FFT size**
MFCC	39 (13 MFCCs, 13 $\triangle$s, and 13$\triangle\triangle$s)	25 ms	10 ms	256 point
PNCC	39 (13 PNCCs, 13 $\triangle$s, and 13$\triangle\triangle$s)			
IF	127			
BPD	127			
MGDCC	36 (12 MGDCCs, 12 $\triangle$s, and 12$\triangle\triangle$s)			

is based on derivative of the phase along time axis. Therefore, it can capture the temporal information of phase. Unlike the original phase spectrum that has the problem called phase wrapping, there are better patterns in the IF spectrum, making it possible to be used as a feature.

- **BPD** baseband phase difference [54] is a phase feature extracted from baseband STFT which is different from IF processing using the phase difference between two successive segments, which can also yield significant phase information.

- **MGDCC** modified group delay [91] is a representation of filter phase response, which is defined as the negative derivative of the Fourier transform phase. It is designed to obtain better phase performance than group delay. Two factors, α and γ, are used for control the dynamic range of the modified group delay. Here, we perform the same setting as recommended in [53].

5.5 Experiment

Our experiments were conducted on CENSREC-1-C database [93]. The speech data is sampled at 16 kHz, and finally downsampled to 8 kHz. The details of recording condition, utterances, and speaking style are the same in CENSREC-1(AURORA-2J) [94]. As for training data, 104 clean speech data (52 males and 52 females) per one noise environment, which constructed by concatenating several utterances spoken by one speaker (the number of utterances in concatenated speech data is nine or ten), were used to create artificial noisy speech. Each artificial noisy speech is obtained by corrupting each speech data with one of 4 noise types (Subway, Babble, Car, Exhibition) at one of six noise levels of SNR, i.e., 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, and -5 dB. For the test data, 104 speech data per one noise types were corrupted by three unseen noise type, namely, Restaurant, Airport, and Station at three SNR levels: 5 dB, 0 dB, and -5 dB.

5.5.1 Baseline setup

We applied support vector machine (SVM) -based VAD using the concept based on [49]. The SVM is a binary classifier, which models the decision boundary between the two classes as a separating hyperplane. To rapidly optimize the support vectors, we used the publicly available LIBLINEAR tool [59] which considers linear kernels. For SVM, a grid-search on C and γ using cross-validation is used. Various pairs of (C, γ) values are tried and the one with the best cross-validation accuracy is picked. The regularization parameter is searched from the exponential grid $\{C = 2^{-5}, 2^{-3}, ..., 2^{15}, \gamma = 2^{15}, 2^{-13}, ...2^{3}\}$.

SignalGraph [84] was used to train the DNN. The DNN has one, two, three layers each of which contains 512 neurons. The N-layers of DNN are denoted as DNN-LN. The input feature for DNN contains 9 frames, Cross entropy (CE) was used for a learning criterion. The learning rate was started from 0.1 and changed to 1 for the second epoch, then it was decayed by a factor of 0.5 when the cross entropy on a crossvalidation set between two consecutive epochs increases. 39-dimensional MFCC (plus deltas and double deltas) was used for training SVM and DNN.. The equal error rate (EER) is used as measure of VAD result.

Table 5.2 shows the results of baseline SVM and DNN-VAD. The highlighted contents are the best performance on the averaged noise scenario.

From Table 5.2, DNN-based VADs outperformed SVM-VAD because DNNs have high feature representation ability to distinguish speech and non-speech accurately. However, DNN-VAD based on more one hidden layer did not perform according to our expectation. It seem that increasing the number of hidden layer for conventional DNN could to yield consistent performance gain for unseen noise types which the same tendency can be found in [87, 95, 96]. This might be due to the different characteristics of unseen noises or the weird architecture of DNN with only two-dimensional output which could not handle the diversified training data well.

5.5.2 Experimental result of individual feature

In this subsection, we used DNN-VAD using 1 layer, which is the best result from previous subsection, to investigate five types of the feature described in Section 5.4.

Table 5.3 shows result of each types of extraction feature. The spectrograms of each feature is shown in Figure 5.3, where (a) waveform,(b) clean MFCC, (c) clean PNCC, (d) clean IF, (e) clean BPD, and (f) clean MGDCC. The figure on the left is corresponding feature on the right under station noise at 5 dB. Comparing (a-f), the spectrogram of PNCC shows the distribution of values with greater variance than all other features. This can be seen in speech segments (highlighted by black dashed lines) and non-speech segments (highlighted by red dashed lines).

From Table 5.3, we can observe that PNCC outperformed all other feature due to strong of asymmetric noise suppression (ANS) and temporal masking in PNCC processing leading to clear spectrogram compared to other features. In this result, phase based feature worked worse than magnitude based feature in both MFCC and PNCC, but it had some ability of speech and non-speech as observed in Figure 5.3. Therefore, it might be useful to use phase information to augment magnitude based feature.

Table 5.2: Performance (EER%) comparison of SVM and DNN classifier using MFCC

Classifier	**Restaurant**			**Airport**			**Station**			**Ave. of all noise types**			**Ave.**
	5	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**ALL**
SVM	27.36	30.35	38.46	23.95	29.69	34.86	20.84	26.12	34.57	24.05	28.72	35.96	29.58
DNN-L1	14.69	26.69	38.31	14.12	21.15	32.04	10.84	21.84	33.65	13.22	23.23	34.67	**23.70**
DNN L2	14.05	24.04	35.81	16.09	23.89	33.92	12.38	22.57	34.96	14.17	23.50	34.90	24.19
DNN-L3	14.12	23.91	35.81	16.09	24.15	33.95	12.24	22.19	34.81	14.15	23.42	34.86	24.14

Table 5.3: Performance (EER%) comparison of DNN-VAD using individual feature

	Restaurant			**Airport**			**Station**			**Ave. of all noise types**			**Ave.**
	5	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**ALL**
MFCC	14.69	26.69	38.31	14.12	21.15	32.04	10.84	21.84	33.65	13.22	23.23	34.67	23.70
PNCC	14.35	24.53	36.88	13.51	20.14	31.23	9.07	17.13	27.04	12.31	20.60	31.72	**21.54**
IF	24.53	31.65	39.06	24.04	30.17	38.47	22.60	30.60	39.17	23.72	30.81	38.90	31.14
BPD	19.78	27.41	36.13	22.72	30.37	38.66	23.28	31.32	40.60	21.93	29.70	38.46	30.03
MGDCC	24.84	32.62	41.34	21.45	26.45	34.93	16.57	24.37	33.83	20.95	27.81	36.70	28.49

5.5.3 Experimental result of proposed method

According to [96], deep neural network based voice activity detection (VAD) has been proved to be powerful in fusing the advantages of multiple features, especially based on magnitude information. However, fusing the advantages of joint magnitude and phase features information has not been well investigated. In this subsection, we take multiple features contains magnitude and phase information into our experiment. Table 5.4 shows the result of each multiple features.

From Table 5.4, it can be seen that multi information based feature provided better performance than individual feature for robust noise VAD. This is due to combining the advantage of multiple features. In feature derived from two s, augmenting PNCC with MGDCC outperformed all of other multiple feature because of exploiting the advantage of robust noise magnitude and phase information. Moreover, the result showed that the feature augmented by dual magnitude features with single feature or augmented by dual phase features with single magnitude features gave better performance by feature augmented by dual feature. This is due to more clear information. However, for multiple feature derived four features, it could not perform our expected result. This might be because additional IF or BPD make feature complicated.

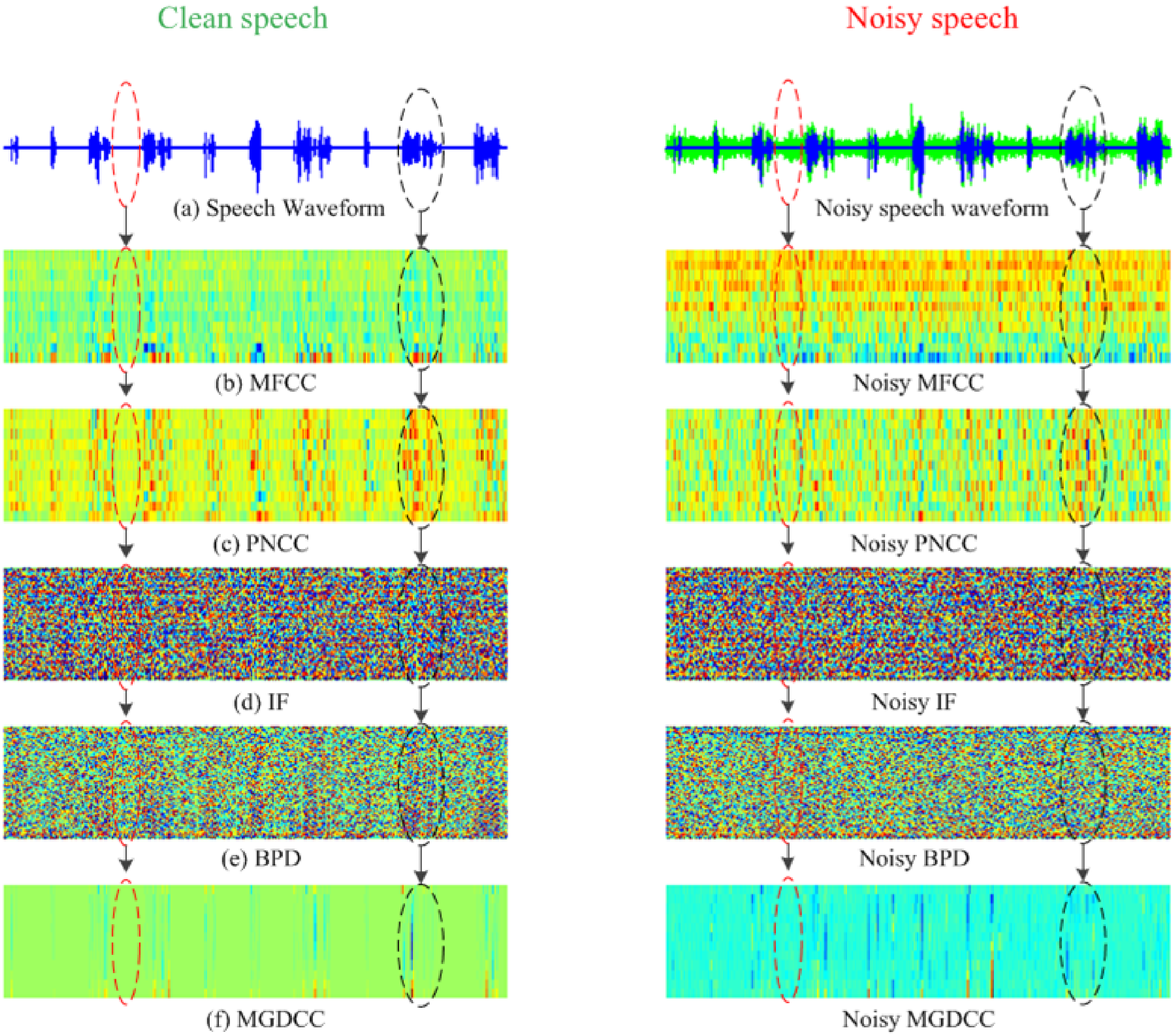


Figure 5.3: Spectrograms of five types of feature.

Table 5.4: Performance (EER%) comparison of DNN-VAD using different multiple features

	Restaurant			**Airport**			**Station**			**Ave. of all noise types**			**Ave.**
	5	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**ALL**
MFCC-PNCC	13.18	24.08	35.69	12.97	19.31	29.38	8.40	16.93	26.70	11.51	20.10	30.59	20.74
MFCC-IF	14.93	26.08	37.15	13.88	20.65	31.18	11.17	21.35	32.49	13.33	22.69	33.61	23.21
MFCC-BPD	15.59	26.31	37.66	13.78	20.31	30.56	10.98	20.66	32.41	13.45	22.43	33.54	23.14
MFCC-MGDCC	14.18	26.18	37.91	13.16	19.53	30.03	10.45	20.15	30.92	12.60	21.95	32.95	22.50
PNCC-IF	18.99	28.34	38.01	18.27	25.99	35.00	12.55	21.64	31.92	16.60	25.32	34.98	25.63
PNCC-BPD	13.47	23.77	35.79	13.12	20.18	31.06	8.87	17.00	27.37	11.82	20.32	31.41	21.18
PNCC-MGDCC	13.18	23.94	35.85	13.21	19.37	30.53	8.26	16.03	25.77	11.55	19.78	30.72	**20.68**
IF-BPD	21.17	28.66	37.12	24.24	31.72	40.40	23.36	31.97	40.81	22.93	30.78	39.44	31.05
IF-MGDCC	21.17	28.67	37.12	24.24	31.72	40.40	23.36	31.97	40.81	22.93	30.78	39.44	31.05
BPD-MGDCC	14.74	25.60	37.81	14.96	22.83	36.07	15.53	25.53	38.72	15.08	24.66	37.53	25.75
PNCC-MGDCC-MFCC	12.72	23.07	34.88	12.89	18.51	28.65	7.73	15.95	24.88	11.11	19.18	29.47	**19.92**
PNCC-MGDCC-IF	12.82	22.45	35.21	13.05	19.24	30.36	8.37	16.23	26.14	11.41	19.31	30.57	20.43
PNCC-MGDCC-BPD	12.92	22.55	35.25	13.09	19.07	30.29	8.33	16.27	26.13	11.45	19.30	30.55	20.43
MFCC-PNCC-MGDCC-IF	12.94	23.00	34.40	12.86	18.73	29.08	8.13	16.44	26.21	11.31	19.39	29.9	20.20
MFCC-PNCC-MGDCC-BPD	13.46	23.88	35.4	12.68	18.53	29.37	8.83	17.50	26.80	11.65	19.97	30.52	20.72
PNCC-MGDCC-BPD-IF	13.05	22.54	34.46	13.69	20.04	30.41	8.89	16.64	26.86	11.88	19.74	30.58	20.73

Table 5.5: Performance (EER%) of score combination

	Restaurant			**Airport**			**Station**			**Ave. of all noise types**			**Ave.**
	5	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**5**	**0**	**-5**	**ALL**
PNCC-MGDCC-MFCC + PNCC-MGDCC-IF	12.15	21.70	34.12	12.74	18.12	28.45	7.62	15.66	24.43	10.84	18.50	29.00	**19.44**
PNCC-MGDCC-MFCC + PNCC-MGDCC-BPD	12.23	21.78	34.36	12.75	18.29	28.72	7.52	15.60	24.28	10.83	18.55	29.12	19.50

5.5.4 Score combination

In this subsection, score combination is proposed to exploit the complementary characteristics of these two feature sets. The score ratio (that is differences of speech and non-speech segments) from different kind of joint magnitude and phase features, based on the combination of three features which obtained the best result from the previous Section, are combined linearly by following equation.

$$\begin{aligned} \wedge_{comb}(I) &= (1-\beta) \wedge (I_{dmsp}) + (\beta) \wedge (I_{dpsm}), \\ \beta &= \frac{\wedge(I_{dpsm})}{\wedge(I_{dpsm)} + \wedge(I_{dmsp})}. \end{aligned} \tag{5.4}$$

where β is a weighing coefficient, $\wedge(I_{dmsp})$ and $\wedge(I_{dpsm})$ denote the score ratio of joint dual magnitude and single phase feature, and of joint dual phase and single magnitude feature, respectively. Table 5.5 lists the result of score combination.

From Table 5.5, It can be observed that the score combination of joint magnitude and phased feature outperformed individual phase aware DNN-based VAD system. This is because of combination of complementary characteristics of different features.

Chapter 6

Conclusion

6.1 Summary

This thesis proposed deep neural network (DNN) as a classifier and feature enhancements for speech classification under adverse environments.

In Chapter 3, DNN based classification was applied for distant talking accent recognition. Then, the combination of Gaussian Miture model (GMM) and DNN-based acoustic model was proposed. In reverberant environment, the accent recognition rate was increased from 90.7 % with GMM to 93.0 % with DNN due to modeling reverberation over multi frame. The proposed combination approach (that is, combining GMM and DNN) provided better accent recognition rate than the individual GMM and DNN. This was because the GMM and DNN based classification have complementary feature. Although DNN-based approach could give better discrimination on confusion boundaries of different classes than GMM, the performance of DNNs considerably depended on training data than that of GMM. Therefore, the combination of scores of GMM and DNN outperformed than the conventional approaches.

In Chapter 4, several feature enhancement approaches for robust noise voice activity detection (VAD) were proposed. DNN-based feature enhancement was applied as conventional method at first. Then, a DNN-based joint phase and magnitude feature (JPMF) enhancement called JPMF with DNN and a NAT-DNN-based JPMF enhancement called JPMF with NAT-DNN were proposed to obtain better result than conventional DNN-based feature enhancement. Moreover, to improve performance of feature enhancement, a combination of the scores of the phase- and magnitude-based features was also applied. The experimental results showed that the proposed feature enhancement produced better performance than the conventional DNN-based feature enhancement in both magnitude and phase feature. The proposed JPMF with NAT-DNN method yielded the best relative equal error rate (EER), compared with individual magnitude- or phase-based DNN speech enhancement. Moreover, the combined score of the enhanced magnitude- and phase-based feature using JPMF with NAT-DNN further improved the VAD performance.

In Chapter 5, to improve performance of conventional DNN based classification (that is, DNN using only magnitude information), a deep neural network (DNN) using magnitude and phase information (that is, phase aware DNN) are proposed to exploit full information in the original signal. Mel-frequency cepstral coefficient (MFCC), power-normalized cepstral coefficients (PNCC), instantaneous frequency derivative (IF), baseband phase difference (BPD) and modified group delay cepstral coefficient (MGDCC) were exploited as magnitude and phase information. The results showed that the phase aware DNN significantly outperforms the DNN using only magnitude information. For DNN-based classifier, the equal error rate (EER) was improved from 23.70% of MFCC, to 20.43% of joint dual magnitude and single phase features (augmenting PNCC, MGDCC and IF), to 19.92% of joint dual phase and single magnitude feature features (augmenting PNCC, MGDCC and BPD). By combining joint dual magnitude and single phase features with joint dual phase and single magnitude features, the EER was improved to 19.44%.

Table 6.1: Merit of each proposed method

	Proposed methods	**Merits**
Proposal 1 (Chapter 3)	DNN-based classification	Improve the performance of accent recognition under reverberant environment by modeling reverberation over multiple frames.
Proposal 2 (Chapter 3)	Combination of GMM and DNN	Improve the performance of accent recognition under reverberant environment by combining the complementary of GMM and DNN.
Proposal 3 (Chapter 4)	JPMF with DNN	Improve the performance of conventional DNN based feature enhancement under noise environment by DNN using augmented information (JPMF).
Proposal 4 (Chapter 4)	JPMF with NAT-DNN	Improve the performance of conventional DNN based feature enhancement under noise environment by NAT-DNN using augmented information (JPMF).
Proposal 5 (Chapter 5)	Phase aware DNN	Improve the performance of VAD under noise environment by phase aware DNN approach.

The Table 6.1 concluded this thesis.

6.2 Future work

This section introduce future work for distant-talking accent recognition and robust noise VAD.

In Chapter 3, optimal weight coefficient should be considered to automatically combine GMM and DNN for distant-talking accent recognition. Moreover, dereverberation methods and more training speech data may be extended to further improve the performance of accent recogntion rate.

In Chapter 4, the DNN based methods can be improved by feeding more information to the DNN training in order to increase the DNN-based feature enhancement quality for robust noise VAD. Therefore, other efficient noise estimator can be added for improving the DNN training. Moreover, other magnitude-based and phase-based spectral features can also be applied for improving the DNN and proposed score combination.

In Chapter 5, we will try to use more training speech data to improve proposed method (that is, phase aware DNN). Moreover, LDA based dimensional reduction can be considered for noise robust VAD.

List of abbreviations

ANS	asymmetric noise suppression
ASR	automatic speech recognition
BPD	baseband phase difference
CD	contrastive divergence
DBN	deep belief network
DCT	discrete cosine transform
DFT	discrete Fourier transform
DNN	deep neural network
DSP	digital signal processing
EER	equal error rate
EM	expectation maximization
GMM	Gaussian mixture model
HMM	hidden Markov model
IF	instantaneous frequency
JPMF	joint phase and magnitude feature
LPCC	linear predictive cepstral coefficients

MC	multi-condition
MFCC	mel-frequency cepstral coefficients
MFDP	mel-frequency Delta-phase
MGDCC	modified group delay cepstral coefficients
ML	maximum likelihood
MLP	multilayer perceptron
MMSE	minimum mean square error
MSE	mean square error
NAT	noise-aware training
NAT-DNN	noise aware training based deep neural network
PLP	perceptual linear prediction
PNCC	power-normalized cepstral coefficients
RBM	restricted Boltzmann machines
RT	reverberation time
SNR	signal-to-noise ratio
STFT	short-time discrete Fourier transform
SVM	support vector machine
VAD	voice activity detection

Bibliography

[1] Z. Zhang, *A study on speech signal processing for noise robust speaker and speech recognition*, Ph.D. thesis, Nagaoka university of technology, 2017.

[2] M. Benzeghiba, Renato De M., 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.

[3] C. Lee, F.K Soong, and K. Paliwal, *Automatic speech and speaker recognition: Advanced topics*, vol. 355, Springer Science & Business Media, 2012.

[4] I. Hwang and J.H. Chang, "Voice activity detection based on statistical model employing deep neural network," in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2014 Tenth International Conference on*. IEEE, 2014, pp. 582–585.

[5] L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, and Y. Kawakami, "Spoofing speech detection using modified relative phase information," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

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

[7] A. Gunawardana, M. Mahajan, A. Acero, and J.C. Platt, "Hidden conditional random fields for phone classification.," in *Interspeech*. Citeseer, 2005, pp. 1117–1120.

[8] D. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

[9] Y. Zeng, Z. Wu, T. Falk, and W. Chan, "Robust gmm based gender classification using pitch and rasta-plp parameters of speech," in *Machine Learning and Cybernetics, 2006 International Conference on.* IEEE, 2006, pp. 3376–3379.

[10] G.C. Van Orden, J.C. Johnston, and B.L Hale, "Word identification in reading proceeds from spelling to sound to meaning.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 14, no. 3, pp. 371, 1988.

[11] D. Jurafsky, "Speech and language processing: An introduction to natural language processing," *Computational linguistics, and speech recognition*, 2000.

[12] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[13] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position dependent cepstral mean normalization.," in *INTERSPEECH*, 2005, pp. 1977–1980.

[14] V.B. Kobayashi and V.B. Calag, "Detection of affective states from speech signals using ensembles of classifiers," 2013.

[15] F. Biadsy, *Automatic dialect and accent recognition and its application to speech recognition*, Ph.D. thesis, Columbia University, 2011.

[16] T. Petsatodis, *Far-Field Voice Activity Detection and Its Applications in Adverse Acoustic Environments*, Ph.D. thesis, Videnbasen for Aalborg UniversitetVBN, Aalborg UniversitetAalborg University, Det Teknisk-Naturvidenskabelige FakultetThe Faculty of Engineering and Science, 2012.

[17] A. Temko, C. Nadeu, D. Macho, R. Malkin, C. Zieger, and M. Omologo, "Acoustic event detection and classification," *Computers in the Human Interaction Loop*, pp. 61–73, 2009.

[18] D. Zhao, Y.and Wang, I. Merks, and T. Zhang, “Dnn-based enhancement of noisy and reverberant speech,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 6525–6529.

[19] L.M Arslan and J.HL Hansen, “Language accent classification in american english,” *Speech Communication*, vol. 18, no. 4, pp. 353–367, 1996.

[20] M.Y. Tsai and L.S. Lee, “Pronunciation variation analysis based on acoustic and phonemic distance measures with application examples on mandarin chinese,” in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 117–122.

[21] G. Choueiter, G. Zweig, and P. Nguyen, “An empirical study of automatic accent classification,” in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 4265–4268.

[22] I. Fohr, D.and Illina, “Text-independent foreign accent classification using statistical methods,” in *Signal Processing and Communications, 2007. ICSPC 2007. IEEE International Conference on*. IEEE, 2007, pp. 812–815.

[23] T. Chen, C. Huang, E. Chang, and J. Wang, “Automatic accent identification using gaussian mixture models,” in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 343–346.

[24] A. Lazaridis, E. Khoury, J. Goldman, M. Avanzi, Sébastien Marcel, and P.N Garner, “Swiss french regional accent identification,” in *Proceedings of odyssey*, 2014.

[25] S. Deshpande, S. Chikkerur, and V. Govindaraju, “Accent classification in speech,” in *Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on*. IEEE, 2005, pp. 139–143.

[26] M. Wu and D. Wang, “A two-stage algorithm for one-microphone reverberant speech enhancement,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 774–784, 2006.

[27] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speech recognition by combining multiple microphone-array processing with position-dependent cmn," *EURASIP Journal on Applied Signal Processing*, vol. 2006, pp. 204–204, 2006.

[28] H.G. Hirsch and H. Finster, "A new approach for the adaptation of hmms to reverberation and background noise," *Speech Communication*, vol. 50, no. 3, pp. 244–263, 2008.

[29] A. Sehr, R. Maas, and W. Kellermann, "Reverberation model-based decoding in the logmelspec domain for robust distant-talking speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1676–1691, 2010.

[30] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[31] A-r. Mohamed, G.E Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[32] O. Tanrikulu, "Residual echo signal in critically sampled subband acoustic echo cancellers based on IIR and FIR filter banks," *Signal Processing, IEEE Transactions on*, pp. 901–912, 1997.

[33] D.K. Freeman and G. Cosier, "The voice activity detector for the Pan-European digital cellular mobile telephone service," *Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on,*, pp. 369–372, 1989.

[34] D. Malah, R.V. Cox, and A.J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *In Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, pp. 789–792, 1999.

[35] D. Enqing, Z. Heming, and L. YongLi, "Low bit and variable rate speech coding using local cosine transform," *TENCON'02. Proceedings. 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering*, pp. 423–426, 2002.

[36] J.C. Junqua, B. Reaves, and B. Mak, “A study of endpoint detection algorithms in adverse conditions: incidence on a DTW and HMM recognizer,” *Second European Conference on Speech Communication and Technology*, 1991.

[37] R. Tucker, “Voice activity detection using a periodicity measure,” *Communications, Speech and Vision, IEE Proceedings I*, pp. 377–380, 1992.

[38] J.H. Chang, N.S. Kim, and S.K. Mitra, “Voice activity detection based on multiple statistical models,” *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.

[39] J. Wu and X.L. Zhang, “Efficient multiple kernel support vector machine based voice activity detection,” *Signal Processing Letters, IEEE,*, pp. 466–469, 2011.

[40] T. Kinnunen and E. Chernenko, “Voice activity detection using MFCC features and support vector machine,” *Conf. on Speech and Computer (SPECOM07), Moscow, Russia*, pp. 556–561, 2007.

[41] R. Tong, B. Ma, K.A Lee, and C. You, “The IIR NIST 2006 Speaker Recognition System: Fusion of Acoustic and Tokenization Features,” *In presentation in 5th Int. Symp. on Chinese Spoken Language Processing, ISCSLP*, 2006.

[42] A. Benyassine, E. Shlomot, H.Y Su, D. Massaloux, C. Lamblin, and J.P. Petit, “Itu-t recommendation g. 729 annex b: a silence compression scheme for use with g. 729 optimized for v. 70 digital simultaneous voice and data applications,” *IEEE Communications Magazine*, vol. 35, no. 9, pp. 64–73, 1997.

[43] Y. Ying, D.and Yan and F.K. Dang, J.and Soong, “Voice activity detection based on an unsupervised learning framework,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 8, pp. 2624–2633, 2011.

[44] Y. Ueda, L. Wang, A. Kai, and B. Ren, “Environment-dependent denoising autoencoder for distant-talking speech recognition,” *EURASIP Journal on Advances in Signal Processing*, vol. 2015:92, pp. 1–11, 2015.

[45] L. Wang, B. Ren, and Y. Ueda, “Denoising autoencoder and environment adaptation for distant-talking speech recognition with asynchronous speech recording,” *In Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*, pp. 1–5, 2014.

[46] B. Xia and C. Bao, “Speech enhancement with weighted denoising auto-encoder.,” *In INTERSPEECH*, pp. 3444–3448, 2013.

[47] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, “Speech enhancement based on deep denoising autoencoder.,” *In INTERSPEECH*, pp. 436–440, 2013.

[48] B. Ren, L. Wang, L. Lu, and A. Ueda, Y.and Kai, “Combination of bottleneck feature extraction and dereverberation for distant-talking speech recognition,” *Multimedia Tools and Applications*, vol. 75, no. 9, pp. 5093–5108, 2016.

[49] T. Kinnunen et al., “Voice activity detection using mfcc features and support vector machine,” in *Proc.SPECOM07*, 2007, vol. 2, pp. 556–561.

[50] C. Kim and R.M. Stern, “Power-normalized cepstral coefficients (pncc) for robust speech recognition,” in *Proc.ICASSP*. IEEE, 2012, pp. 4101–4104.

[51] H. Hermansky, “Perceptual linear predictive (plp) analysis of speech,” *the Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[52] I. McCowan and D. Dean, “The delta-phase spectrum with application to voice activity detection and speaker recognition,” *Audio, Speech, and Language Processing, IEEE Transactions on*, pp. 2026–2038, 2011.

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

[54] M. Krawczyk and T. Gerkmann, “Stft phase reconstruction in voiced speech for an improved single-channel speech enhancement,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1931–1940, 2014.

[55] L.D. Alsteris and K.K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," *Digital Signal Processing*, pp. 578–616, 2007.

[56] D.A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," *Speech communication*, vol. 17, no. 1, pp. 91–108, 1995.

[57] A.P. Dempster, Nan M. Laird, and Donald B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[58] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., "The htk book," *Cambridge university engineering department*, vol. 3, pp. 175, 2002.

[59] R.E. Fan et al., "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

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

[61] L. Wang, K. Odani, and A. Kai, "Dereverberation and denoising based on generalized spectral subtraction by multi-channel lms algorithm using a small-scale microphone array," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, pp. 12, 2012.

[62] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, T. Maas, R.and Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[63] Z. Zhang, L. Wang, and A. Kai, "Distant-talking speaker identification by generalized spectral subtraction-based dereverberation and its efficient computation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2014, no. 1, pp. 1–12, 2014.

[64] L. Wang, N. Kitaoka, and S. Nakagawa, "Robust distant speaker recognition based on position-dependent cmn by combining speaker-specific gmm with speaker-adapted hmm," *Speech communication*, vol. 49, no. 6, pp. 501–513, 2007.

[65] A. Krizhevsky, I. Sutskever, and G.E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[66] Linguistic Data Consortium et al., "Timit acoustic-phonetic continuous speech corpus," *URL http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp*, 1993.

[67] N. Minematsu, K. Okabe, K. Ogaki, and K. Hirose, "Measurement of objective intelligibility of japanese accented english using erj (english read by japanese) database.," in *INTERSPEECH*, 2011, pp. 1481–1484.

[68] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," 2000.

[69] M. Nakayama, T. Nishiura, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, et al., "Censrec-4: development of evaluation framework for distant-talking speech recognition under reverberant environments.," in *INTERSPEECH*, 2008, pp. 968–971.

[70] G.E. Hinton and R.R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, pp. 504–507, 2006.

[71] X.L. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," *In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 853–857, 2013.

[72] Y. Xu, J. Du, L.R. Dai, and C.H. Lee, "A regression approach to speech enhancement based on deep neural networks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, pp. 7–19, 2015.

[73] Y. Xu, J. Du, L.R. Dai, and C.H. Lee, “Dynamic noise aware training for speech enhancement based on deep neural networks.,” *In INTERSPEECH*, pp. 2670–2674, 2014.

[74] G.E. Hinton, S. Osindero, and Y.W. Teh, “A fast learning algorithm for deep belief nets,” *Neural computation*, pp. 1527–1554, 2006.

[75] X. Xiao, S Zhao., and D.H.H Nguyen, “The NTU-ADSC systems for reverberation challenge 2014,” *In Proc. REVERB challenge workshop*, 2014.

[76] R.C. Hendriks, “MMSE based noise PSD tracking with low complexity,” *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 4266–4269, 2010.

[77] D.S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for joint enhancement of magnitude and phase,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5220–5224.

[78] D.S. Williamson, Y. Wang, and D. Wang, “Complex ratio masking for monaural speech separation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[79] 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.

[80] N. Kitaoka, T. Yamada, and S. Tsuge, “CENSREC-1-C: An evaluation framework for voice activity detection under noisy environments,” *Acoustical Science and Technology*, pp. 363–371, 2009.

[81] D. Povey and A. Ghoshal, “The Kaldi speech recognition toolkit,” *In IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society.*, 2011.

[82] R.E. Fan, K.W. Chang, and C.J. Hsieh, “LIBLINEAR: A library for large linear classification,” *The Journal of Machine Learning Research*, pp. 1871–1874, 2008.

[83] Y.X. Zou, W.Q. Zheng, W. Shi, and H. Liu, "Improved voice activity detection based on support vector machine with high separable speech feature vectors," *In Digital Signal Processing (DSP), 2014 19th International Conference on*, pp. 763–767, 2014.

[84] X. Xiao, "Signalgraph: a deep learning toolkit for signal processing, 2016. [online]," *Available: https://github.com/singaxiong/SignalGraph.*

[85] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on youtube using deep neural networks.," in *INTERSPEECH*, 2013, pp. 728–731.

[86] F. Eyben et al., "Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies," in *Proc. ICASSP*. IEEE, 2013, pp. 483–487.

[87] X.L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2013.

[88] S. Nakagawa, L. Wang, and S. Ohtsuka, "Speaker identification and verification by combining mfcc and phase information," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 4, pp. 1085–1095, 2012.

[89] P. Mowlaee, R. Saeidi, and Y. Stylianou, "Advances in phase-aware signal processing in speech communication," *Speech Communication*, vol. 81, pp. 1–29, 2016.

[90] Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, and M. Iwahashi, "DNN-Based Amplitude and Phase Feature Enhancement for Noise Robust Speaker Identification," *Interspeech 2016*, pp. 2204–2208, 2016.

[91] B. Yegnanarayana and H.A. Murthy, "Significance of group delay functions in spectrum estimation," *IEEE Transactions on signal processing*, pp. 2281–2289, 1992.

[92] L. Wang, Y. Yoshida, Y. Kawakami, and S. Nakagawa, "Relative phase information for detecting human speech and spoofed speech," in *Proc. Interspeech*, 2015, pp. 2092–2096.

[93] N. Kitaoka et al., "Censrec-1-c: An evaluation framework for voice activity detection under noisy environments," *Acoustical Science and Technology*, vol. 30, no. 5, pp. 363–371, 2009.

[94] S. Nakamura et al., "Aurora-2j: An evaluation framework for japanese noisy speech recognition," *IEICE transactions on information and systems*, vol. 88, no. 3, pp. 535–544, 2005.

[95] Q. Wang et al., "A universal vad based on jointly trained deep neural networks," in *Proc. Interspeech*, 2015, pp. 2282–2286.

[96] X. Zhang and J. Wu, "Denoising deep neural networks based voice activity detection," in *Proc.ICASSP*. IEEE, 2013, pp. 853–857.

Appendix A

Publication List

Publication for journal

[1] K. Phapatanaburi, L. Wang, R. Sakagami, Z. Zhang, X. Li and M. Iwahashi, ”Distant-talking accent recognition by combining GMM and DNN, ” Multimedia Tools and Applications, Vol. 75, No. 9, pp. 5109-5124, May 2016. DOI 10.1007/ s11042-015- 2935-4

[2] K. Phapatanaburi, L. Wang, Z. Oo, W. Li, S. Nakagawa and M. Iwahashi, ”Noise Robust Voice Activity Detection using Joint Phase and Magnitude based Feature Enhancement,” Journal of Ambient Intelligence and Humanized Computing, Issue 11, pp.1-15, April 2017. DOI 10.1007/ s12652-017- 0482-8

Publication for conference

[3] K. Phapatanaburi, L. Wang, Z. Oo, W. Li, S. Nakagawa and M. Iwahashi, ”Noise Robust Voice Activity Detection by Combining Magnitude and Phase-Based Spectral Features with Denoising Autoencoder, Proc. ASJ 2016 spring Meeting, pp. 33-36,2016.

[4] L. Wang, K. Phapatanaburi, Z. Oo, S. Nakagawa, M. Iwahashi and J. Dang, ”Phase Aware Deep Neural Network for Noise Robust Voice Activity Detection,” Proc. of IEEE International

Conference on Multimedia Expo (ICME), July, 2017, (Accepted).