DISSERTATION

Voice Activity Detection Using Deep Neural Network

Graduate School of
Natural Science & Technology
Kanazawa University

Division of Electrical Engineering and Computer Science

Student ID Number: 1424042016
Name: Suci Dwijayanti
Chief Advisor: Masato MIYOSHI, Prof. Dr. Eng.
Date of Submission: December, 2017

Abstract

Voice activity detection (VAD) is utilized as a preprocessing for some speech applications to determine speech and non-speech periods in input signals. Many methods have been proposed to find acoustic features which are effective in distinguishing speech from non-speech periods especially in noisy signals. In this study, a deep neural network (DNN)-based VAD method is proposed to improve the performance of raw features, i.e., log power spectra (LPS), to detect such periods by utilizing dynamics that refer to the time-varying properties of speech signals. In the proposed method, the dynamics are highlighted by speech period candidates, which are calculated based on heuristic rules for simple patterns of the first and second derivatives of LPS that characterize the starting and ending points of utterances. The candidates, together with LPS, are input into the DNN to determine speech periods. Experiments are conducted to evaluate the proposed method by using speech signals smeared with five types of noise (white, babble, factory, car and pink) with signal to noise ratios (SNRs) of 10, 5, 0 and -5 dB. The experimental results show that the proposed method is excellent under all considered noise conditions. The addition of speech period candidates provides improvements to LPS. The results also show that the proposed method is superior to conventional methods for all noise conditions.

Contents

List of Figures

List of Tables

Acknowledgments

I would like to express my deepest gratitude to all the people who have contributed and supported me in making this thesis possible. I am indebted and would like to thank my advisor, Prof. Masato Miyoshi, for giving an excellent atmosphere for doing research in audio information processing laboratory, including his useful guidance and patience. I would also thank Assistant Prof. Takeshi Saitou for his support while doing this research. I would also like to thank Prof. Satoshi Yagitani, Prof. Yoshiya Kasahara, Prof. Yoshitaka Goto and Prof. Akihiro Hirano for their support and advisement while serving on my committee.

The most profound gratitude to my parents, Rahman HS and Suwasti who always give their best support, motivation and endless pray to me. I thank my sister, Agustin Darmayanti, and my brothers, Ridho Akbar and Norman Wibowo. Without their encouragement and love, I would not be able to accomplish it.

I am also very grateful to the audio information processing laboratory members who help me during my living in Kanazawa, especially to their patience with my limited Japanese ability. I would like to thank Yamamori san for sharing knowledge while doing this research. Also, thank you so much for all of Indonesia students group in Kanazawa, especially KU-DIKTI students batch 2014.

I also thank DIKTI for the scholarship and Universitas Sriwijaya for giving this great opportunity to pursue this degree.

Last and foremost, I thank The Almighty, Allah SWT, for guiding, blessing, and protecting me through this period. It would not be possible for me to accomplish this without the faith I have in the Almighty.

Kanazawa, December 2017

Chapter 1

Introduction

1.1 Research Background

Voice activity detection (VAD) is used as a preprocessing for various speech applications to determine speech and non-speech periods in their input. In digital cellular telecommunication systems, such as universal mobile telecommunication systems (UMTS) [1], VAD is employed to detect non-speech frames to reduce average bit rates [2] so that the highly efficient speech coding and low bit rate transmission may be achieved [3]. In speech enhancement, for example spectral subtraction, speech/non-speech detection is necessary to determine signal periods that contain only noise. This is useful for noise estimations which are used in the noise reduction process [4]. VAD is also utilized to identify speech periods in observed signals as it should be fed to a recognition engine, thus avoiding the processing of non-speech periods that do not convey information and may even affect the recognition process [5].

In prior studies, simple acoustic features have been used to detect speech periods, such as energy and zero crossing rates [6]. Various modifications of energy-based features, such as those described in [7] and [8], have been proposed to improve the VAD performance degraded as the signal-to-noise ratio (SNR) decreases. Other acoustic features have also been examined and investigated to improve such performance. For example, Ramirez *et al.* [9] introduced a feature called as long-term spectral divergence (LTSD) which measures divergency between speech and

noise spectra. Periodic to aperiodic components ratios were employed by [10]. Pek *et al.* [11] investigated an effective modulation frequency range for VAD and used indices of modulation spectra of speech data to distinguish speech and non-speech periods. Kinnunen and Rajad [12] introduced the likelihood ratio VAD method in which mel-frequency cepstral coefficients (MFCCs) of each input frame are utilized to obtain speech and non-speech models. Sohn *et al.* [13] proposed a statistical method based on a Gaussian model and used the mean of the likelihood ratios for individual frequency bands as the decision of speech or non-speech, assuming *a priori* known noise. Davis *et al.* [14] proposed a scheme that combines a low variance spectrum estimation and a statistical method for decision rules based on the SNR ratio measured. Although these methods perform well under stationary noise, their performances are degraded under non-stationary noise. Machine learning methods have also been investigated to improve the VAD performance particularly in noisy environment, for instance, support vector machine (SVM) methods [15–17] and neural networks [18, 19].

Because background noise is a challenging problem, selecting features that discriminate properties of speech and noise is important in designing VAD algorithms [20]. Combination of several features may improve VAD accuracy. Almajai and Milner [21] employed both audio and visual information. They use MFCCs together with their delta and delta-delta for audio features, and 2D discrete cosine transform (DCT) for visual features. Zhang *et al.* [22] attempted to optimize the capability of DNN-based VAD by combining several acoustic features: pitch, discrete Fourier transform (DFT), MFCCs, coefficients obtained from linear predictive coding (LPC), relative-spectral perceptual linear predictive analysis (RASTA-PLP) and amplitude modulation spectrograms (AMS) with their delta features, as the input to DNNs. Segbroeck *et al.* [23] exploited spectral shape, spectro-temporal modulation, harmonicity and long-term spectral variability to obtain a noise robust frontend for VAD.

Recently, DNNs have attracted increasing attention for VAD and have been found to be highly competitive with traditional VAD. The great flexibility, deep and generative training properties of DNNs are useful in speech processing [24] includ-

ing speech/non-speech detection. Espi *et al.* [25] utilized spectro-temporal features for a convolutional neural network (CNN) to detect non-speech acoustic signals. Zhang and Wang [26] explored the contextual information to determine speech and non-speech segments. Ryant *et al.* [27] utilized MFCCs as the input to a DNN to detect speech activity on Youtube. Mendelev *et al.* [28] proposed a DNN to improve the VAD performance when only small training data are available, using a max out activation function to improve the accuracy carried by a drop out model. However, choosing features as the input for a DNN is not a trivial problem. Research in automatic speech recognition has shown that raw features have potential to be used as the input of a DNN replacing "hand-crafted" features [29].

In this study, we first attempted to utilize only log power spectra (LPS) to detect speech periods in noisy signals using a DNN. In the preliminary experiment, we obtained two important findings. First, the performance of VAD using the log power spectra as the input of the DNN outperforms standard features, such as MFCCs and MFCCs combined with delta and delta-delta cepstra, for both clean and noisy signals. Second, the performance of DNN-based VAD using MFCCs improves when they are incorporated with their delta and delta-delta. These delta features are referred to as dynamics which are related to the time-varying of speech signals [30]. Thus, this result indicates that the dynamics may contribute to improve the VAD performance. Based on the second finding, we attempt to enhance the VAD performance based on the usage of LPS by introducing dynamics. The dynamics are expressed by speech period candidates which are derived from the first and second derivatives of the LPS.

Figure 1.1 shows the outline of the proposed method. Major speech characteristics are first highlighted by using a running spectral filter (RSF) [31]. Next, masks are composed by using the first and second derivatives of the LPS from the above-mentioned filtering process through heuristic rules. These masks, which consist of binary values, are then multiplied by the raw spectra expressed in decimal to obtain speech period candidates. Since not all subbands signals may contribute to the VAD decision, we consider obtaining the speech period candidates for individual subbands. These speech period candidates, together with the LPS, are input into a

DNN to determine speech periods.

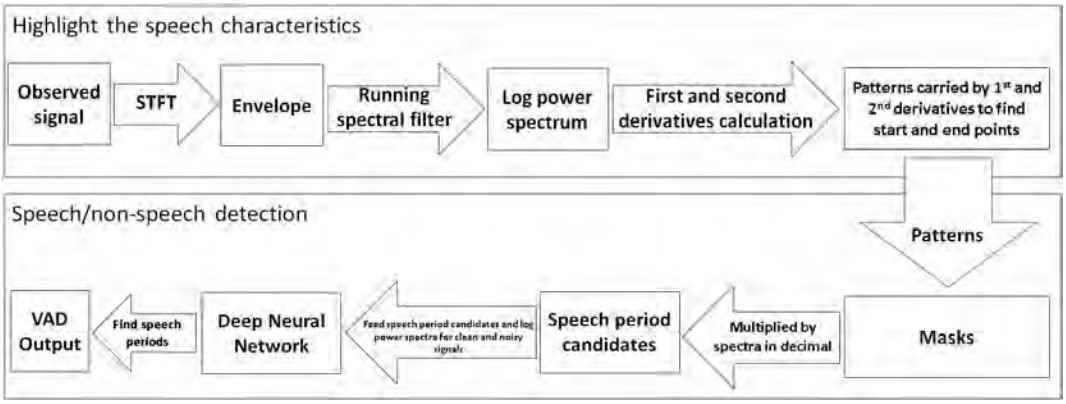


Figure 1.1. Outline of the proposed VAD method

1.2 Motivations and Objectives

This study is motivated by the belief that the first and second derivatives of log power spectra (LPS) may contribute to find starting and ending points of utterances. These derivatives are calculated just after highlighting speech characteristics in a modulation frequency range of 1 to 16 Hz. The starting and ending points are then characterized from simple patterns of these derivatives. The output of this process is referred to as speech period candidates which may highlight dynamics referring to time-varying properties of speech signals.

This study aims to enhance performance achieved by DNN-based VAD using LPS, by adding dynamics expressed by speech period candidates. These candidates, together with the LPS are fed to the DNN to get VAD output. As shown in the experimental results, the addition of speech period candidates is effective at improving the accuracy of the LPS for finding speech and non-speech periods, especially for low SNRs and non-stationary cases. The proposed method is superior to other methods such as [9], [12], [13] and [23]. We also observe the DNN utilizes the information carried by useful subbands which may correspond to speech fundamental frequencies (F0) or their neighbors.

1.3 Organization of Dissertation

This thesis is organized as follows. We describe a brief description of VAD including various features that have been employed to identify speech periods in Section 2. Section 3 describes the proposed method for detecting speech periods using LPS and speech period candidates. The experimental results and discussion of the results are provided in Section 4. Finally, Section 5 presents the conclusions of this study.

Chapter 2

Voice Activity Detection

2.1 Voice Activity Detection

Voice activity detection (VAD) is important for various speech applications as it discriminates the presence of speech periods from background noise in an audio signal [3]. Figure 2.1 shows an example of VAD. As shown in the figure, VAD has two classes; speech and non-speech periods. It has been utilized in different applications for a variety of purposes. In the following section, roles of VAD are described in speech enhancement, speech coding, and speech recognition.

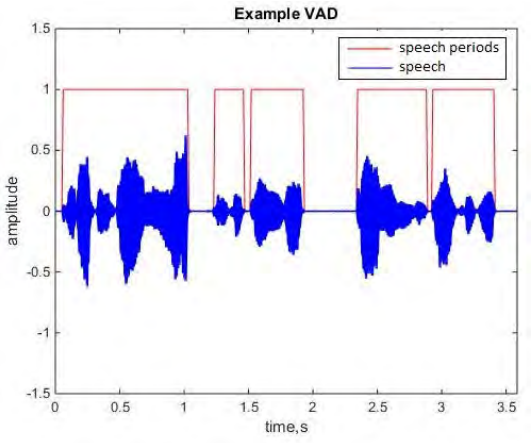


Figure 2.1. Voice activity detection (VAD)

2.1.1 Speech Enhancement

One major purpose of speech enhancement is to enhance the performance of speech communication by reducing background noise. In spectral subtraction, for example, VAD may be useful to find out if noise reference needs to be updated. Speech detection is essential to determine the periods that contain only noise. Such noise-only periods are essential to allow the noise reference to be updated, thereby suppressing noise more accurately [32]. Park *et al.* [33] proposed a VAD method using Teager energy to derive speech absence probability (SAP) to be implemented in speech enhancement algorithm to improve its performance, especially in the noisy environment.

2.1.2 Speech Coding

VAD output allows to deactivate transmission during the absence of speech in speech coding, and hence to reduce the amount of transmitted data while maintaining quality. In modern telecommunication systems, silence compression which is achieved by VAD in the discontinuous transmission (DTX) mode, may be utilized to reduce the average bit rates [9]. VAD has also been implemented in cellular networks as shown in [34]. Some standards in telecommunication have adopted VAD methods. For example, the International Telecommunication Union (ITU) [1] employs a VAD module in DTX mode to work with speech coding algorithm G.729. To determine speech periods, G.729 Annex B has recommended some features such as linear prediction (LP) spectra, full-band energy, low band (0-1 kHz) energy and zero crossing rates (ZCR). On the other hand, European Telecommunication Standard (ETSI) also uses VAD for the digital cellular telecommunication systems. In this standard, VAD is used in DTX for adaptive multi-rate (AMR) speech traffic channels.

2.1.3 Speech Recognition

Another important application of VAD is speech recognition. The accuracy of speech recognizers may degrade and indicate constraints to the operation in noisy

environment when training conditions do not match to testing conditions [3]. VAD is useful in such conditions since it can identify speech periods that should be fed to a recognition engine so that the error can be reduced. Moreover, ignoring non-speech periods can also save CPU power. Ramirez *et al.* [35] proposed a VAD method to get a robust speech recognition system by improving statistical test, namely a contextual likelihood ratio test (LRT).

2.2 Features for VAD

Feature extraction and decision scheme are general stages in a VAD algorithm to get VAD output. The first process plays a significant role because its purpose is to extract acoustic features, which represent discrimination between speech and noise, from the noisy speech signals [20]. Various features have been used in conventional VAD methods. In this section, the features used in VAD are divided into 5 groups:

2.2.1 Energy Based Features

Energy might be the earliest feature developed in the time domain for VAD. In [6], Rabiner *et al.* proposed to locate endpoints of an utterance using zero crossing rates and energy. The energy of a speech signal can be calculated as

$$E_j = \frac{1}{N} \sum_{i=(j-1)N+1}^{jN} x^2(i), \tag{2.1}$$

where E_j is the energy of the j-th frame, $x(i)$ is the *i*-th sample of the speech, and N is the number of samples in the considered frame. The zero crossing rate of the speech, Z_n, for a given frame of N samples can be calculated as

$$Z_n = \frac{1}{2N} \sum_{i=1}^{N} (|\text{sgn}(x_n(i)) - \text{sgn}(x_n(i-1))|). \tag{2.2}$$

Figure 2.2 shows a speech signal with its energy and zero crossing rate representation.

A speech production model suggests that speech energy is concentrated in the

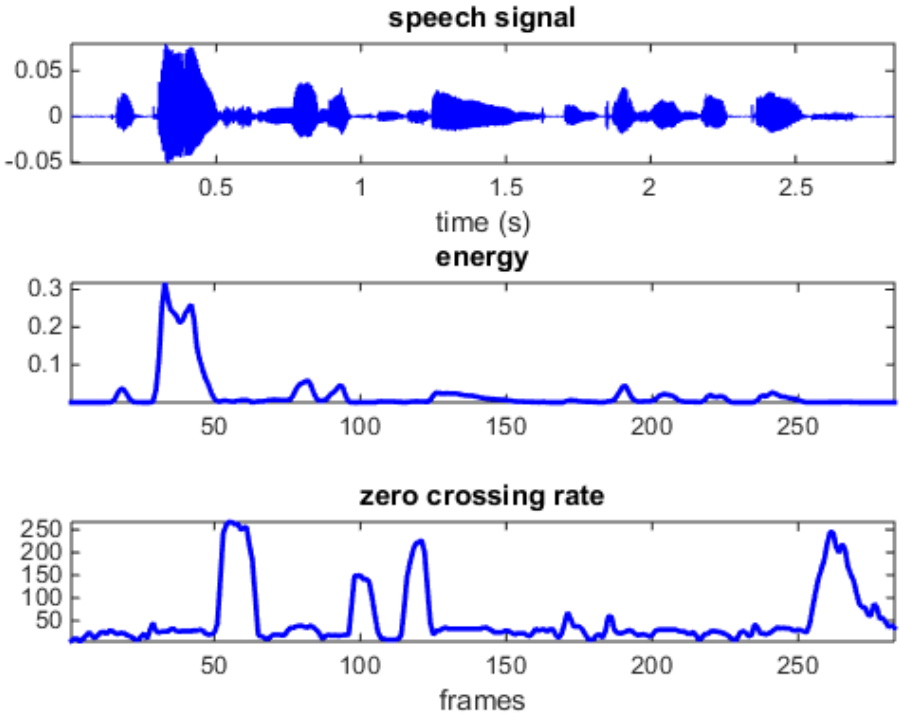


Figure 2.2. Energy and zero crossing features

frequency below 3 kHz. On the other hand, the noise energy is mostly located at higher frequencies. Hence, this technique is somewhat suitable for a clean condition, and its performance is degraded in low SNR. To overcome such problem, many studies have been proposed to modify the VAD method based on short time energy. Prasad *et al.* [8] proposed to use an adaptive linear energy-based detector to update a threshold value because a fixed threshold would be insufficient for varying acoustic environments of a speaker. Sakhnov *et al.* [7] utilized the energy of the speech based on root square energy (RMSE) instead of short time energy in combination with periodicity and the energy levels are later used to estimate the threshold.

These VAD methods based on energy features have been employed in some standards because of its low computation complexity, for example, ITU-T Recommendation G.729 Annex B [1].

2.2.2 Spectral Features

Frequency spectrum analysis that represents the frequency components of the signal over time has also been used to derive the features employed in VAD. A VAD algorithm proposed by Renevey *et al.* [36] utilizes entropy estimation measures of a time-frequency magnitude spectrum. The idea of this algorithm relies on a presumption that the signal spectrum shows to be more well-organized during the speech periods than the non-speech periods. The entropy can be measured as,

$$H(|\boldsymbol{Y}(\omega,t)|^2) = -\sum_{\omega=1}^{\Omega} P(|Y(\omega,t)|^2)\log(P(|Y(\omega,t)|^2)) \tag{2.3}$$

where $P(|Y(\omega,t)^2|) = \frac{|Y(\omega,t)|^2}{\sum_{\omega=1}^{\Omega}|Y(\omega,t)|^2}$ is the probability of the frequency band ω for the magnitude spectrum for frame t.

Ramirez *et al.* [9] also proposed a VAD method using a spectral feature. They assume that the information of spectral magnitude on time-varying signal is the most significant clue to detect speech periods in a noisy speech signal. In this method, they use long term spectral envelope (LTSE).Then, a long-term spectral distance (LTSD) between speech and noise is measured as a feature to derive a decision rule. The LTSD can be formulated as

$$LTSD(X_k) = 10\log_{10}(\frac{1}{L}\sum_{i=1}^{L}\frac{|\tilde{X}_{k,i}|^2}{|N_i|^2}) \tag{2.4}$$

$$\tilde{X}_{k,i} = \max\{X_{k-M,i},\ldots,X_{k,i},\ldots,X_{k+M,i}\}, \tag{2.5}$$

where M is the number of neighboring frames, and L is the number of fast Fourier transform (FFT) coefficients, N_i is the i-th coefficient of the average noise spectrum and $\tilde{X}_k = [\tilde{X}_{k,1},\ldots,\tilde{X}_{k,L}]^T$ is the estimated spectral envelope (LTSE) of the neighboring $(2M+1)$ frames.

Ma and Nishihara proposed a long spectral flatness measure (LSFM) based VAD method that exploits the spectral flatness of a signal to differentiate speech from noise [37]. Gosh *et al.* also utilized spectral features by proposing a long-term signal variability (LTSV) which utilizes a very long window to compute averaged

spectrogram and entropies of each frequency bands [38].

2.2.3 Cepstral Features

Cepstra are particularly important in speech recognition. They provide an identifiable form of transcript for each particular utterance. Cepstral differences along time are keys to the success of speech recognition [39]. Mel-frequency cepstral coefficients (MFCCs) are commonly used as cepstral features for VAD, such as [15] and [12] which use MFCCs and their delta and double delta coefficients for the input features to SVM and GMM classification, respectively. Another VAD method which uses cepstra was proposed in [41]. Fukuda *et al.* [40] used the delta cepstra which express the dynamic changes of the cepstral frames in their proposed VAD. Delta cepstrum d_t at time t is estimated as

$$d_t = \sum_{k=1}^{K} k.(c_{t+k} - c_{t-k})/(2\sum_{k=1}^{K} k^2), \tag{2.6}$$

where c_t is defined as the cepstrum coefficient at time t.

2.2.4 Harmonicity Features

Multiple harmonics of fundamental frequency F_0 can be found in the voiced speech structure. In noisy condition, such harmonic structures of voiced speech are still preserved.

Autocorrelation function (ACF) can be utilized to capture the harmonic structure of the speech $x(n)$ as

$$\text{ACF}(\tau, \ell) = \frac{\sum_{n=\ell L}^{\ell L-N+1+\ell} x(n).x(n-\tau)}{\text{norm}(\tau, \ell)}, \tag{2.7}$$

where $\text{norm}(\tau, \ell) = \sqrt{\sum_{n=\ell L}^{\ell L-N+1+\tau} x^2(n).\sum_{n=\ell L-\tau}^{\ell L-N+1} x^2(n)}$. Some pitch-related features for VAD are usually derived from this ACF approach. For periodic signals, it is maximized for values of τ being integer multiples of the period. Features which indicate the maximum ACF peak [42] or the periodicity of the ACF [43] employ this property.

Ishizuka *et al.* [44] stated that a human auditory system has the capability to process the harmonicity (periodic) components, which may relate to vowels and voiced consonants and remains after eliminating the aperiodic components which deviate from predominant periodicity. In their work, they proposed an acoustic feature, namely, PAR (periodic and aperiodic ratio) which represents the power ratios between these two components in observed signals. Then, a decision scheme to obtain VAD decision is derived from a likelihood of the existence of target speech signals based on the relations to the PARs.

Basu [45] used three features for VAD, such as a non-initial maximum of normalized "noisy" autocorrelation, some autocorrelation peaks, and normalized spectral entropy, which are later used in HMM. The autocorrelation vector, $\mathbf{r}_k = [r_{k,1}, \ldots, r_{k,N}]^T$, of the $k-$th frame $\mathbf{x}_k$, and the estimated pitch, $\hat{f}_0$ are found as

$$r_{k,\tau} = \sum_{m=\tau}^{N} x_{k,m} x_{k,m-\tau} \tag{2.8}$$

$$\tau_{max} = \arg\max_{\tau} r_{k,\tau} \tag{2.9}$$

$$\hat{f}_0 = \frac{f_s}{\tau_{max}}, \tag{2.10}$$

where f_s represents the sampling frequency.

The autocorrelation approach is also developed in the spectral domain such as [46]. Krishnamachari *et al.* [46] investigated the behavior of spectral autocorrelation under co-channel a condition where the spectral autocorrelation ratio (SAR) parameter is defined as

$$SAR = 20 \log_{10} \frac{R(p_1)}{R(q_1)}, \tag{2.11}$$

where $R(p_1)$ is the local maximum of spectral correlation which occurs at lag p_1, and $R(q_1)$ is the second local maximum that is not related harmonically to the first peak or the local minimum between p_1 and $2p_1$. When the speech is silence or unvoiced, a peak which is not related to the peak harmonically will be lower in amplitude. That means SAR will be very high so that the frame can be used as speech. However, when speech and interferer were of comparable magnitude, the SAR ratio might

come near to zero, which indicate such frame is non-usable. The SAR will again be low when there is an occurrence of a false peak of comparable magnitude along with the harmonically related pulses. However, the physical interpretation is that such frames can not be used when a pure tone which is mixed with the speech signal is of comparable magnitude.

2.2.5 Modulation Features

A characteristic energy modulation, which peaks at about 2-5 Hz corresponding to the typical syllable rate of human speech [47–49], dominates a temporal structure of speech. This speech characteristic may be employed in VAD. Shadevsky *et al.* [50] exploited the characteristic in modulation spectra of speech by filtering noise components outside the modulation frequency domain of 1 to 16 Hz in their proposed VAD. Mesgarani *et al.* [51] proposed an auditory model which is based on multiscale spectro-temporal modulations to distinguish speech from noise. Modulation spectra is also used by [11] for detecting speech periods.

Batch *et al.* [52] considered to use amplitude modulation spectral (AMS) features to detect speech activity. SVM is then used as the classifier to decide speech or non-speech periods based on AMS. To obtain AMS, a signal is analyzed by using STFT and its output is computed by squared magnitude, then it is summed into rectangular, non-overlapping Bark bands and logarithmic amplitude compression. Another STFT is later applied to each spectral band to obtain modulation spectrum. In the end, an envelope extraction followed by a logarithmic compression is further applied.

2.3 VAD Decision

Feature classification of speech signals is the final stage of VAD algorithms to determine speech periods. It determines the class of features either as speech or non speech. The simplest way to classify a signal is by using a detection threshold which defines the features as speech or non speech. When scalar value $y_k \in \mathbb{R}$ is the feature extracted at frame $k-$th, $\mathbb{R}$ can be divided into two classes by a fixed

threshold η. It is assigned as speech when the frame has $y_k > \eta$, and $y_k < \eta$ is non-speech as shown as,

$$y_k \underset{non-speech}{\overset{speech}{\gtrless}} \eta. \tag{2.12}$$

More than one threshold is sometimes defined, such as [6] that computed three thresholds, namely, an upper energy threshold, a lower energy threshold, and a zero crossings rate threshold.

Since a fixed threshold is insufficient for varying acoustic environment, an adaptive threshold is used in many VAD algorithms, such as [8], [7] and [9]. Threshold needs to be updated and it is controlled by a learning factor α. If a new noise parameter is n_k at frame k-th, the threshold is updated as

$$\hat{\eta} = \alpha\eta + (1-\alpha)n_k. \tag{2.13}$$

An α with a smaller value is useful for a situation where the background noise constantly changes so the threshold can be more sensitive to such changes. Meanwhile, a larger α will be useful for the conditions where the background noise changes infrequently since it may make the threshold to be more modest to the changes [53].

A threshold method for a decision rule-set might be adequate when extracted features are highly discriminative and linearly separable because they are usually obtained heuristically. Once a feature space is non-linearly separable, a statistical method will be a better choice. Different from such threshold method, a statistical method needs a prior knowledge of speech. The distributions of features for speech and noise can be modeled explicitly by using probability density functions (PDFs) model of which parameters are estimated during a running time. The statistical approach is contrary to the heuristic approaches where the threshold is adopted to a likelihood ratio of the speech and noise model [20].

A well-known statistic approach is proposed by Sohn *et al.* [13]. They formulate two hypotheses to consider frames as

$$H_0 = \text{speech absent} : \mathbf{X} = \mathbf{N} \tag{2.14}$$

$$H_1 = \text{speech present} : \mathbf{X} = \mathbf{N} + \mathbf{S}, \tag{2.15}$$

where the bold letters represent L the discrete Fourier transform (DFT) coefficient vectors which are asymptotically independent Gaussian random variables. **S** is for speech with $k-$th elements S_k, **N** is for noise with $k-$th elements N_k, and **X** is for noisy speech with $k-$th elements X_k. The probability density functions conditions are calculated as,

$$p(\mathbf{X}|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \exp\{-\frac{|X_k|^2}{\lambda_N(k)}\} \tag{2.16}$$

$$p(\mathbf{X}|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi[\lambda_N(k) + \lambda_S(k)]} \exp\{-\frac{|X_k|^2}{\lambda_N(k) + \lambda_S(k)}\} \tag{2.17}$$

where the variance of N_k and S_k are represented as $\lambda_N(k)$ and $\lambda_S(k)$, respectively. The expression for the likelihood ratio for the kth frequency bin is as,

$$\Lambda_k \triangleq \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\{\frac{\gamma_k \xi_k}{1+\xi_k}\}, \tag{2.18}$$

where $\xi_k \triangleq \frac{\lambda_S(k)}{\lambda_N(k)}$ is a priori SNR and $\gamma_k \triangleq \frac{|X_k|}{\lambda_N(k)}$ is a posteriori SNR. A decision rule is then calculated as the mean of the likelihood ratios for the individual frequency bands as follows

$$\log \Lambda = \frac{1}{L} \sum_{k=0}^{L-1} \log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta. \tag{2.19}$$

The likelihood above is calculated for a single frame. Later, a hangover scheme is introduced to consider the dependency between successive frames. To this purpose, two states of HMM model speech and non-speech. For each frame n, the VAD score $\Gamma(n)$ is calculated as follows

$$\Gamma(n) = \frac{P(H_0)}{P(H_1)} \Big[\frac{a_{01} + a_{11}\Gamma(n-1)}{a_{00} + a_{10}\Gamma(n-1)}\Big] \Lambda(n) \underset{H_0}{\overset{H_1}{\gtrless}} \eta, \tag{2.20}$$

where $a_{ij} \triangleq P(q(n) = H_j | q(n-1) = H_i)$ is a state transition probability, $P(H_0)$ and $P(H_1)$ are prior probabilities, and $\Lambda(n)$ denotes the likelihood ratio at nth frame. Finally, the VAD score is contrasted to a decision threshold η to obtain speech/non-speech decisions.

Besides the above method, there are some other works which also employ statistical approach such as [14] and [54]. Chang *et al.* [54] combined multiple statistical models, namely, complex Laplacian, Gaussian model, and Gamma probability density functions to derive the speech and non-speech decision.

Besides statistical approach, a machine learning approach has recently got large interest to be used to obtain the VAD decision. Machine learning-based VADs are highly competitive to traditional VADs because of several factors. They can be integrated to speech recognition system, they guarantee to have good performance, and they can combine multiple features together [22]. Many methods have been proposed such as support vector machine (SVM) in [15–17], neural networks in [18, 19], and deep neural networks (DNNs) in [22, 26, 27, 55]. Recently, DNNs have been successfully applied in speech processing including VAD, given their capabilities. In fact, Mohammed *et al.* [24, 56] describe DNNs as a flexible model that does not require information on the specific data distribution. In addition, a DNN can include several nonlinear hidden layers, thus increasing its flexibility and discrimination capabilities. Furthermore, a generative pre-training allows a strong domain-dependent regularization of the networks weights.

Chapter 3

DNN-Based Method for VAD

A speech signal can be analyzed by using short-time Fourier transform (STFT) as follows:

$$X(m,k) = \sum_{n=-\infty}^{n=\infty} h(m-n)x(n)W_K^{kn}, \tag{3.1}$$

where a speech signal is represented by $x(n)$, an analysis window which is time reversed and shifted by m frames is represented by $h(n)$, k is a frequency variable, the number of frequency bins is represented by K, and $W_K = \exp^{-j(\frac{2\pi}{K})}$. The visualization of the spectrum in time is represented by spectrogram as shown in Fig. 3.1. $X(m,k)$ can be further written as,

$$X(m,k) = |X(m,k)|e^{j\angle X(m,k)}. \tag{3.2}$$

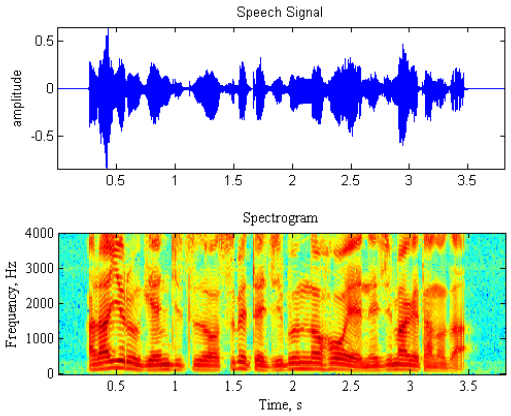


Figure 3.1. Spectrogram of speech

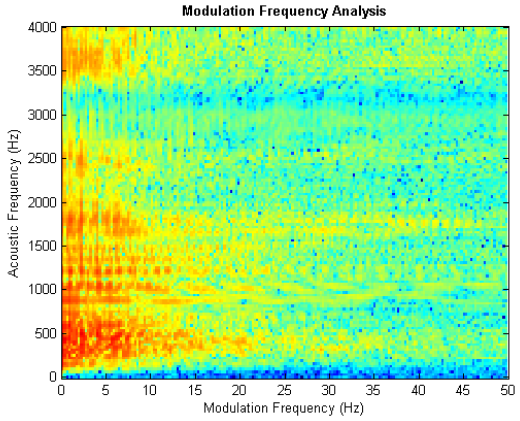


Figure 3.2. Modulation spectrum of speech

Magnitude operator, $|X(m,k)|$, can be defined as an envelope.These envelope spectra consist of short-term spectra arranged on the time axis to display the time variation of spectral characteristics. The representation of the short-term spectrum with respect to the time variation is called as a modulation spectrum. The modulation spectrum is achieved by transforming the spectral envelope using another STFT as follows,

$$X(l,k,i) = \sum_{m=-\infty}^{m=\infty} g(l-m)|X(m,k)|W_I^{im}, \tag{3.3}$$

where a modulation frequency analysis window is represented by $g(m)$, a corresponding hop size is expressed by l, k is an acoustic frequency and i is a modulation frequency. Figure 3.2 shows the modulation spectral energy $X(l,k,i)$ which is a magnitude of subband envelope spectra in the acoustic-modulation frequency plane.

Figure 3.2 shows that the speech signal energy is concentrated in a low modulation frequency range. Several investigations [47, 48, 50, 57, 58], have shown that the energy of clean speech signals is mostly concentrated within a modulation frequency range of 1 to 16 Hz. Hence, each subband envelope, $|X(m,k)|$, is filtered through an RSF to remove noise outside the modulation frequency range. The RSF output, $X_{rsf}(m,k)$, consists of both negative and positive values, and we follow [31] in replacing the negative values with zeros.

A subband log power spectrum of the RSF output, $E(m,k)$, is expressed as

$$E(m) = 10\log_{10}\sum_{k=1}^{k=N}(X_{rsf}(m,k))^2 \tag{3.4}$$

Hereafter, we call the log power spectrum of the RSF output as LPS-RSF. The first and second derivatives of the log power spectrum, $E(m,k)$, obtained through the above filtering, are calculated as follows:

$$\Delta_{-m}E(m) = E(m) - E(m-1) \tag{3.5}$$

$$\begin{aligned}\Delta_m^2 E(m) &= \Delta_{+m}\Delta_{-m}E(m) \\ &= [E(m+1) - E(m)] - [E(m) - E(m-1)].\end{aligned} \tag{3.6}$$

These derivatives are used to produce speech period candidates that highlight the dynamics in the LPS-RSF. These candidates are used together with the log power spectra, derived from Eq. (3.2), to detect speech periods in the DNN described later. The detailed block diagram of the proposed method is shown in Fig. 3.3.

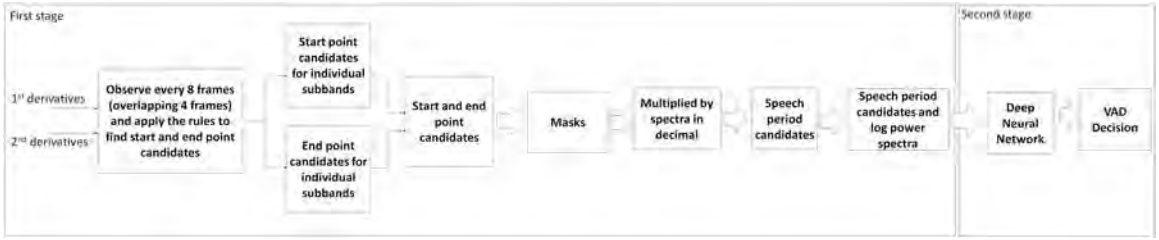
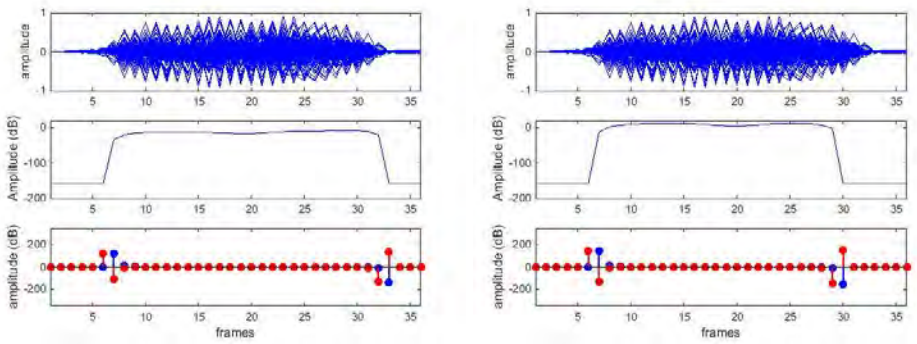


Figure 3.3. Block diagram of the proposed VAD method

3.1 Speech Period Candidates

In spoken language, an utterance is a continuous piece of speech that has a start and an end, and is separated from a successive utterance by a pause. Figure 3.4 shows the subband observations at 250 and 875 Hz of utterance /ha/ and the observations' first and second derivatives of the LPS-RSF obtained using Eqs. (3.5) and (3.6). The frame size used to obtain this representation is 20 ms. Since the sampling frequency is 8 kHz, each frame may consist of 160 samples, and a Hamming



(a) Subband at 250 Hz (b) Subband at 875 Hz

Figure 3.4. Subband observations of utterance /ha/, log power spectra, and their first and second derivatives. Blue and red lines represent first and second derivatives, respectively

window is used as the analysis window with a 10 ms frame shift.

As shown in Fig. 3.4, the starting and ending points of the utterance /ha/ may be identified from the patterns of the first and second derivatives. The starting and ending point candidates of utterance /ha/ in the subband at 250 Hz are located at frames 6 and 33, respectively. In contrast, in the subband at 875 Hz, the starting and ending point candidates are found to lie at frames 6 and 30, respectively. These observations indicate that not all subband signals may contribute to the VAD decision. Therefore, we calculate the first and second derivatives for the individual subbands to obtain the speech period candidates.

We will use Fig. 3.5 to explain the mechanism for identifying the starting (Fig. 3.5a) and ending (Fig. 3.5b) points. These two figures show the observation frames. To determine the starting and ending points, the speech signal is observed in segments of 8 frames with an overlap of 4 frames. The rules for identifying the starting and ending points are as follows:

(i) To identify a starting point, we consider 8 frames at once and check the former 4 frames, as shown in Fig. 3.5a. We observe these frames to find a frame that has the local maximum second derivative followed by a positive first derivative in the successive frame. When this pattern holds, the respective frame becomes a starting point candidate. This process continues for the successive eight frames with an overlap of four frames from the previous observation.

(ii) To identify an ending point, we consider 8 frames at once and check the subse-

(a) Starting point (b) Ending point

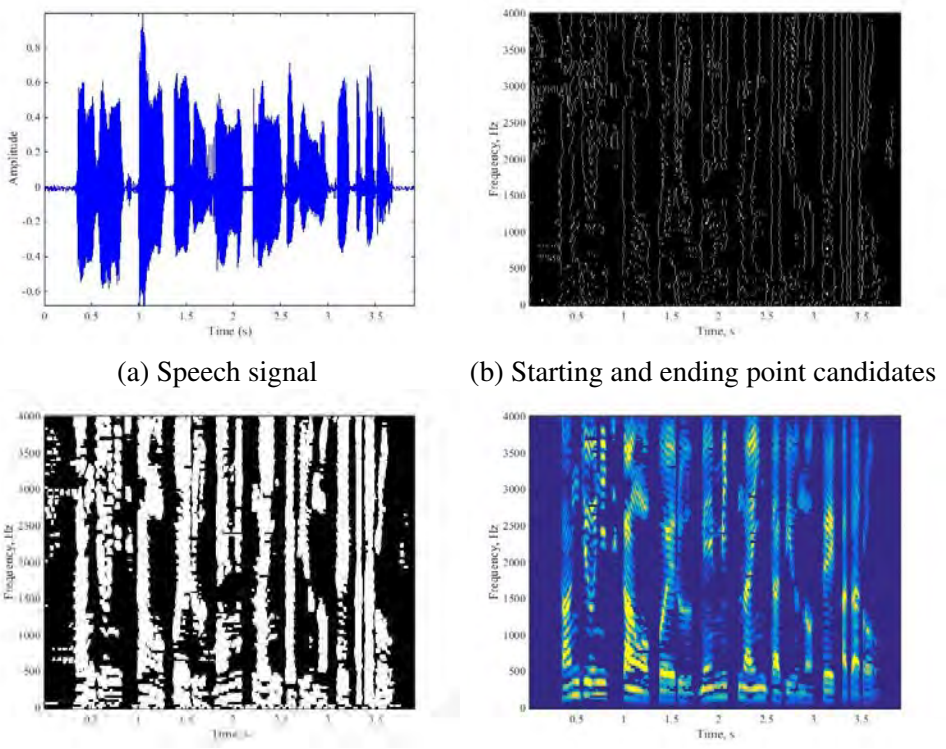
Figure 3.5. Method of identifying the starting and ending points

quent 4 frames, as shown in Fig. 3.5b. We observe these frames to find a frame that has the combination of a local minimum first derivative and a local maximum second derivative that is preceded by at least one negative first derivative. When this pattern holds, such frame becomes an ending point candidate. This process continues for the successive eight frames overlapped with four frames from the previous observation.

The above two processes continue until the last observation frames have been examined.

The starting and ending point candidates that are found based on rules (i) and (ii) are marked by the simple binary number of one. Figure 3.6b shows the starting and ending point candidates of the speech signal. We then simply add the binary ones between the starting and ending points to obtain the masks, as shown in Fig. 3.6c.

The masks, however, may cause misjudgments with respect to non-speech periods because such masks do not carry information coming from the amplitude of the observed signal. To minimize such misjudgement, we attempt to remove values of 'one' coming from the signal parts when the amplitudes are relatively small simply by multiplying the power spectra expressed in decimal by the masks. Hereafter, the output of this process is referred to as speech period candidates. The result of the process is shown in Fig. 3.6d. These speech period candidates, together with the log power spectra from Eq. (3.2) become input for the DNN.



(a) Speech signal

(b) Starting and ending point candidates

(c) Masks

(d) Speech period candidates

Figure 3.6. Representation of a speech signal (a), its starting and ending point candidates using rules (i) and (ii) (b), masks (c), and speech period candidates as a result of multiplying masks by power spectra expressed in decimal form (d)

3.2 Deep Neural Network for VAD

At the second stage, a deep neural network (DNN) is employed to get correct speech periods. A DNN has been shown to be effective to be used in various speech applications, including VAD, as shown in [27] and [55]. A DNN is a feed forward artificial neural network which has many hidden layers (often more than two) between its input and output, as shown in Fig. 3.7. For simplicity, if there is an $L+1$-layer DNN, layer 0 indicates the input layer and layer L is considered as the output layer. According to [59], in the first L layers, the activation vector $\mathbf{a}^{\ell}$ is obtained as follows:

$$\mathbf{a}^{\ell} = f(\mathbf{z}^{\ell}) = f(\mathbf{W}^{\ell}\mathbf{a}^{\ell-1} + \mathbf{b}^{\ell}), \text{for} \quad 0 < \ell < L, \tag{3.7}$$

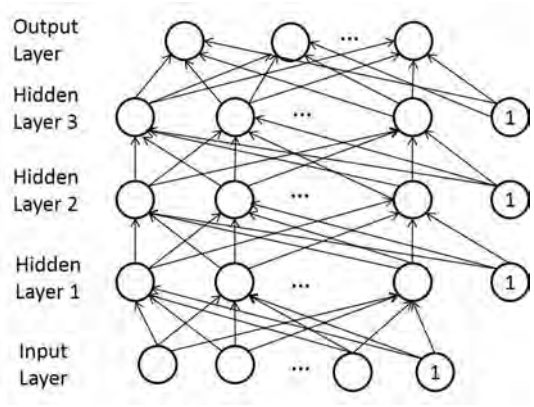


Figure 3.7. Illustration of DNN

where $\mathbf{z}^\ell = \mathbf{W}^\ell \mathbf{a}^{\ell-1} + \mathbf{b}^\ell \in \mathbb{R}^{N_\ell \times 1}$ is an excitation vector, $\mathbf{a}^\ell \in \mathbb{R}^{N_\ell \times 1}$ is an activation vector, $\mathbf{W}^\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ is a weight matrix, $\mathbf{b}^\ell \in \mathbb{R}^{N_\ell \times 1}$ is a bias vector, and $N_\ell \in \mathbb{R}$ are the number of neurons at layer ℓ. $\mathbf{a}^0 = \mathbf{o} \in \mathbb{R}^{N_0 \times 1}$ represents an observation or the input feature vector, where $N_0 = D$ is the dimension of feature, and $f(.) : \mathbb{R}^{N_\ell \times 1} \rightarrow \mathbb{R}^{N_\ell \times 1}$ is a transfer function (also known as activation function) applied to the excitation vector element-wise **z**. In most application, including VAD, the sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.8}$$

is used as the activation function.

For classification tasks such as VAD, each output neuron in the output layer indicates a class $i \in \{1, \ldots, C\}$, where $C = N_L$ represents the number of assigned classes. Here, VAD consists of two classes; speech and non-speech. To accommodate it, a softmax function is used to normalize the excitation and added in the final layer, as

$$\mathrm{a}_i^L = P_{\text{dnn}}(i|\mathbf{o}) = \text{softmax}_i(\mathbf{z}^L) = \frac{e^{z_i^L}}{\sum_{j=1}^{C} e^{z_J^L}} \tag{3.9}$$

where z_i^L is the *i*th element of the excitation vector $\mathbf{z}^L$. The value of the *i*-th output neuron a_i^L represents the probability $P_{dnn}(i|\mathbf{o})$ that the observation vector **o** belongs to class *i*.

The model parameters, $\{\mathbf{W,b}\} = \{\mathbf{W}^\ell, \mathbf{b}^\ell | 0 < \ell \leq L\}$ are calculated by following Eq. (3.7) to obtain the activation vectors. This is done layer by later from layer 1 to layer $L - 1$. In the final layer, Eq. (3.9) is used for the classification task to

obtain the DNN output. This process is called as forward computation and shown in Algorithm 1.

Algorithm 1 DNN Forward Computation

procedure FORWARDCOMPUTATION(($\mathbf{O}$)) ▷ Each column of $\mathbf{O}$ is an observation vector
  $\mathbf{A}^0 \leftarrow \mathbf{O}$
  for $\ell \leftarrow 1; \ell < L; \ell \leftarrow \ell + 1$ **do** ▷ L is the total number of layers
    $\mathbf{Z}^\ell \leftarrow \mathbf{W}^\ell \mathbf{A}^{\ell-1} + \mathbf{B}^\ell$ ▷ Each column of $\mathbf{B}^\ell$ is $\mathbf{b}^\ell$
    $\mathbf{A}^\ell \leftarrow f(\mathbf{Z}^\ell)$ ▷ $f(.)$ can be sigmoid, tanh, or other functions
  end for
  $\mathbf{Z}^L \leftarrow \mathbf{W}^L \mathbf{A}^{L-1} + \mathbf{B}^L$
  if regression **then** ▷ regression task
    $\mathbf{A}^L \leftarrow \mathbf{Z}^L$
  else ▷ classification task
    $\mathbf{A}^L \leftarrow$ softmax$(\mathbf{Z}^L)$ ▷ Apply softmax column wise
  end if
  return $\mathbf{A}^L$
end procedure

The unknown model parameters, **W** and **b**, can be estimated from training samples $\mathbb{S} = \{(\mathbf{o}^m, \mathbf{y}^m) | 0 \leq m < M\}$, where *M* is the number of training samples, $\mathbf{o}^m$ is the *m*-th observation vector and $\mathbf{y}^m$ is the corresponding desired output vector. A criterion for training and an algorithm for learning can determine this process. To minimize the expected loss (loss function), the model parameters should be trained as

$$J_{EL} = \mathbb{E}(J(\mathbf{W}, \mathbf{b}; \mathbf{o}, \mathbf{y})) = \int_{\mathbf{o}} J(\mathbf{W}, \mathbf{b}; \mathbf{o}, \mathbf{y}) p(\mathbf{o}) d(\mathbf{o}), \tag{3.10}$$

where $J(\mathbf{W,b}; \mathbf{o,y})$ represents the lost or cost function for the given model parameters $\{\mathbf{W, b}\}$, the observation **o**, and the corresponding output vector **y**, and $p(\mathbf{o})$ represents the probability density function of observation **o**. Typically, $p(\mathbf{o})$ is not known and needs to be estimated from the training set and $J(\mathbf{W,b}; \mathbf{o,y})$ is not well-defined for samples unseen in the training set. Thus, the DNN model parameters are usually trained to optimize the empirical criteria. In the classification taskk, the cross-entropy (CE) criterion,

$$J_{CE}(\mathbf{W}, \mathbf{b}; \mathbb{S}) = \frac{1}{M} \sum_{m=1}^{M} J_{CE}(\mathbf{W}, \mathbf{b}; \mathbf{o^m}, \mathbf{y^m}) \tag{3.11}$$

is often used, where

$$J_{CE}(\mathbf{W}, \mathbf{b}; \mathbf{o}, \mathbf{y}) = -\sum_{i=1}^{C} y_i \log \mathrm{a}_i^L, \tag{3.12}$$

$y_i = P_{emp}(i|\mathbf{o})$ is the empirical (observation in the training set) probability that the observation **o** belongs to class *i*, and $\mathrm{a}_i^L = P_{dnn}(i|\mathbf{o})$ is the same probability estimated from the DNN. Given the training criterion, the error backpropagation (BP) algorithm can learn the model parameters {**W,b**}. This algorithm is derived from the chain rule used for gradient computation as described in Algorithm 2.

Algorithm 2 Backpropagation Algorithm

procedure BACKPROPAGATION($(\mathbb{S} = \{(\mathbf{o}^m, \mathbf{y}^m)|0 \leq m < M\})$) ▷ $\mathbb{S}$ is the training set with *M* samples
  Randomly initialize $\{\mathbf{W}_0^\ell, \mathbf{b}_0^\ell\}, 0 < \ell \leq L$ ▷ L is the total number of layers
  while Stopping Criterion Not Met **do** ▷ Stop if reached max iterations or the training criterion improvement is small
    Randomly select a minibatch **O,Y** with M_b samples
    Call ForwardComputation(**O**)
    $\mathbf{E}_t^L \leftarrow \mathbf{A}_t^L - \mathbf{Y}$ ▷ Each column of $\mathbf{E}_t^L$ is $\mathbf{e}_t^L$
    $\mathbf{G}_t^L \leftarrow \mathbf{E}_t^L$
    for $\ell \leftarrow L; \ell > 0; \ell \leftarrow \ell - 1$ **do**
      $\nabla_{\mathbf{W}_t^\ell} \leftarrow \mathbf{G}_t^\ell(\mathbf{a}_t^{\ell-1})$
      $\nabla_{\mathbf{b}_t^\ell} \leftarrow \mathbf{G}_t^\ell$
      $\mathbf{W}_{t+1}^\ell \leftarrow \mathbf{W}_t^\ell - \frac{\epsilon}{M_b}\nabla\mathbf{W}_t^\ell$ ▷ Update **W**
      $\mathbf{b}_{t+1}^\ell \leftarrow \mathbf{b}_t^\ell - \frac{\epsilon}{M_b}\nabla\mathbf{b}_t^\ell$ ▷ Update **b**
      $\mathbf{E}_t^{\ell-1} \leftarrow (\mathbf{W}_t^\ell)^T\mathbf{G}_t^\ell$ ▷ Error backpropagation
      if $\ell > 1$ **then**
        $\mathbf{G}_t^{\ell-1} \leftarrow f'(\mathbf{Z}_t^{\ell-1}) \bullet \mathbf{E}_t^{\ell-1}$
      end if
    end for
  end while
  Return dnn $= \{\mathbf{W}^\ell, \mathbf{b}^\ell\}, 0 < \ell \leq L$
end procedure

The DNN is composed of multiple layers restricted Boltzmann machines (RBMs), where each of RBM is a stochastic generative neural network constructed by visible and hidden neurons, and there is no connection between visible-visible and hidden-hidden as shown in Fig. 3.8. Values of hidden neurons are usually binary followed by Bernoulli distribution. Meanwhile, real or binary values are usu-

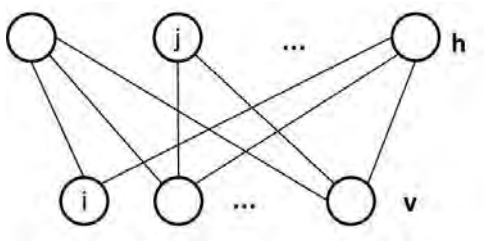


Figure 3.8. A representation of restricted Boltzmann machines

ally taken by the visible neurons. Energy is assigned by the RBM for both visible vector **v** and hidden vector **h** configuration. For the Bernoulli-Bernoulli RBM, the energy is

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{c}^T\mathbf{v} - \mathbf{b}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v}, \tag{3.13}$$

where $\mathbf{v} \in \{0,1\}^{N_v \times 1}$ and $\mathbf{h} \in \{0,1\}^{N_h \times 1}$. N_v is the numbers of visible neurons and N_h are the numbers of hidden neurons. $\mathbf{W} \in \mathbb{R}^{N_h \times N_v}$ is the weight matrix connecting visible and hidden neurons, and $\mathbf{c} \in \mathbb{R}^{N_v} \times 1$ and $\mathbf{b} \in \mathbb{R}^{N_h} \times 1$ are the visible and hidden layer bias vectors, respectively. For Gaussian-Bernoulli RBM, the energy is

$$E(\mathbf{v}, \mathbf{h}) = \frac{1}{2}(\mathbf{v} - \mathbf{c})^T(\mathbf{v} - \mathbf{c}) - \mathbf{b}^T\mathbf{h} - \mathbf{h}^T\mathbf{W}\mathbf{v} \tag{3.14}$$

for each configuration $(\mathbf{v}, \mathbf{h})$. Here, the visible neurons take real values, $\mathbf{v} \in \mathbb{R}^{N_v \times 1}$.

A probability associates to each configuration is defined as

$$P(\mathbf{v}, \mathbf{h}) = \frac{e^{-E(\mathbf{v},\mathbf{h})}}{Z}, \tag{3.15}$$

where $Z = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}$ is the normalization factor.The posterior probabilities $P(\mathbf{v}|\mathbf{h})$ and $P(\mathbf{h}|\mathbf{v})$ can be calculated as

$$P(\mathbf{h} = \mathbf{1}|\mathbf{v}) = \sigma(\mathbf{W}\mathbf{v} + \mathbf{b}), \tag{3.16}$$

$$P(\mathbf{v} = \mathbf{1}|\mathbf{h}) = \sigma(\mathbf{W}^\mathbf{T}\mathbf{h} + \mathbf{c}) \tag{3.17}$$

for binary values. For Gaussian visible neurons, $P(\mathbf{v}|\mathbf{h})$ is estimated as

$$P(\mathbf{v}|\mathbf{h}) = \mathcal{N}(\mathbf{v}; \mathbf{W}^\mathbf{T}\mathbf{h} + \mathbf{c}, \mathbf{I}), \tag{3.18}$$

where I is the appropriate identity covariance matrix. Meanwhile, the the conditional probability $P(\mathbf{h}{=}\mathbf{1}|\mathbf{v})$ is equivalent to the above equation. The RBM can be used to learn a probability distribution over its set of inputs.

Stochastic gradient descent (SGD) is used to train an RBM with goal to minimize the negative log likelihood (NLL).

$$J_{NLL}(\mathbf{W}, \mathbf{c}, \mathbf{b}; \mathbf{v}) = -\log P(\mathbf{v}) = F(\mathbf{v}) + \log \sum_{v} e^{-F(v)}, \tag{3.19}$$

where $F(\mathbf{v})$ is the free energy which is defined as

$$F(\mathbf{v}) = -\log(\sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h})}). \tag{3.20}$$

Then, parameters are updated as

$$\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \epsilon \Delta \mathbf{W}_t, \tag{3.21}$$

$$\mathbf{c}_{t+1} \leftarrow \mathbf{c}_t - \epsilon \Delta \mathbf{c}_t, \tag{3.22}$$

$$\mathbf{b}_{t+1} \leftarrow \mathbf{b}_t - \epsilon \Delta \mathbf{b}_t, \tag{3.23}$$

where ϵ is the learning rate, and

$$\Delta \mathbf{W}_t = \rho \Delta \mathbf{W}_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{W}_t} J_{NLL}(\mathbf{W}, \mathbf{c}, \mathbf{b}; \mathbf{v}^m), \tag{3.24}$$

$$\Delta \mathbf{c}_t = \rho \Delta \mathbf{c}_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{c}_t} J_{NLL}(\mathbf{W}, \mathbf{c}, \mathbf{b}; \mathbf{v}^m), \tag{3.25}$$

$$\Delta \mathbf{b}_t = \rho \Delta \mathbf{b}_{t-1} + (1-\rho)\frac{1}{M_b}\sum_{m=1}^{M_b} \nabla_{\mathbf{b}_t} J_{NLL}(\mathbf{W}, \mathbf{c}, \mathbf{b}; \mathbf{v}^m), \tag{3.26}$$

where ρ is the momentum parameter, M_b is the minibatch size, and $\nabla_{\mathbf{W}_t} J_{NLL}(\mathbf{W, c, b}; \mathbf{v}^m)$ is the gradient of the NLL criterion on model parameter weight **W**, and $\nabla_{\mathbf{c}_t} J_{NLL}(\mathbf{W, c, b}; \mathbf{v}^m)$ and $\nabla_{\mathbf{b}_t} J_{NLL}(\mathbf{W, c, b}; \mathbf{v}^m)$ are for bias **W, c** and **b**, respectively. Derivatives of such NLL with respect to model parameters can be calculated as

$$\nabla_\theta J_{NLL}(\mathbf{W}, \mathbf{c}, \mathbf{b}; \mathbf{v}) = -\Big[\langle \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta} \rangle_{data} - \langle \frac{\partial E(\mathbf{v}, \mathbf{h})}{\partial \theta} \rangle_{model}\Big], \tag{3.27}$$

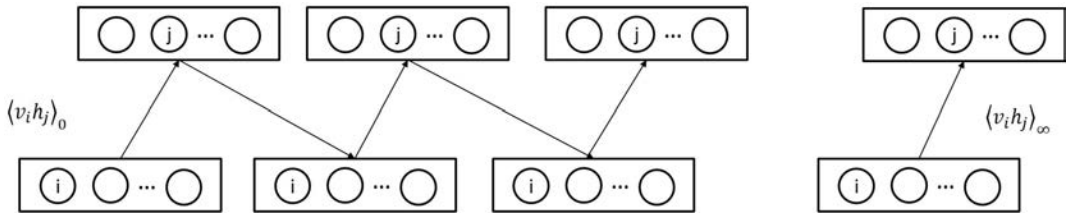


Figure 3.9. Illustration of the contrastive divergence algorithm

where model parameter is represented by θ, and $\langle x \rangle_{data}$ and $\langle x \rangle_{model}$ are the expectation of x estimated from the data and from the model, respectively. Weight update for the visible-hidden can be calculated as follows,

$$\nabla_{w_{ji}} J_{NLL}(\mathbf{W}, \mathbf{c}, \mathbf{b}; \mathbf{v}) = -[\langle v_i h_j \rangle_{data} - \langle v_i h_j \rangle_{model}]. \tag{3.28}$$

The term $\langle . \rangle_{model}$ takes exponential time to compute exactly when the hidden values are unknown, so the approximated learning algorithm, contrastive divergence (CD) [60] is employed. Illustration of the contrastive divergence algorithm is shown in Fig. 3.9. For the first step of Gibbs sample, initialization is implemented to a data sample. Then, a hidden sample is generated from the visible sample based on the posterior probability $P(\mathbf{v}|\mathbf{h})$. This process keeps continuing for many steps. When it runs for infinite steps, $\langle v_i h_j \rangle_{model}$ can be estimated from the samples as,

$$\langle v_i h_j \rangle_{model} \approx \langle v_i h_j \rangle_1 = v_i^1 \mathbb{E}_j(\mathbf{v}|\mathbf{h}^0) = v_i^1 P_j(\mathbf{v}|\mathbf{h}^0), \tag{3.29}$$

$$\langle v_i h_j \rangle_{data} \approx \langle v_i h_j \rangle_0 = v_i^0 \mathbb{E}_j(\mathbf{h}|\mathbf{v}^0) = v_i^0 P_j(\mathbf{h}|\mathbf{v}^0). \tag{3.30}$$

These update rules are also derived for models parameters **c** and **b**. The contrastive divergence algorithm is summarized in Algorithm 3.

The RBM and DNN have relations that suggest a training of a very deep generative model in layer-wise manner. In the training process of RBM, each data vector, **v** is used for the computation of a vector of expected hidden activation, **h**. Output of this process is used to generate new features for the next RBMs. Once training has stopped, the initial values of the weights of the hidden layers in DNN equals to the number of RBMs trained. The DNN can be further fine-tuned which is performed to the whole network [61].The DNN can be considered as a statistical graphical model

Algorithm 3 The contrastive divergence algorithm for training RBMs

procedure TRAINRBMWITHCD($(\mathbb{S} = \{\mathbf{o}^m | 0 \leq m < M\}, N)$) ▷ $\mathbb{S}$ is the training set with *M* samples, *N* is the CD steps

Randomly initialize $\{\mathbf{W_0}, \mathbf{c_0}, \mathbf{b_0}\}$

while Stopping Criterion Not Met **do** ▷ Stop if reached max iterations or the training criterion improvement is small

Randomly select a minibatch **O** with M_b samples

$\mathbf{V}^0 \leftarrow \mathbf{O}$ ▷ Positive phase

$\mathbf{H}^0 \leftarrow P(\mathbf{H}|\mathbf{V}^0)$ ▷ Applied column-wise

$\nabla_{\mathbf{W}} J \leftarrow \mathbf{H}^0(\mathbf{V}^0)^T$

$\nabla_{\mathbf{c}} J \leftarrow \text{sumrow}(\mathbf{V}^0)$ ▷ Sum along rows

$\nabla_{\mathbf{b}} J \leftarrow \text{sumrow}(\mathbf{H}^0)$

for $n \leftarrow 0; n < N; n \leftarrow n + 1$ **do** ▷ Negative phase

$\mathbf{H}^n \leftarrow \mathbb{I}(\mathbf{H}^n > rand(0, 1))$ ▷ Sampling, $\mathbb{I}(\bullet)$ is the indicator function

$\mathbf{V}^{n+1} \leftarrow P(\mathbf{V}|\mathbf{H}^n)$

$\mathbf{H}^{n+1} \leftarrow P(\mathbf{H}|\mathbf{V}^{n+1})$

end for

$\nabla_{\mathbf{W}} J \leftarrow \nabla_{\mathbf{W}} J - \mathbf{H}^N(\mathbf{V}^N)^T$ ▷ Subtract negative statistics

$\nabla_{\mathbf{c}} J \leftarrow \nabla_{\mathbf{c}} J - \text{sumrow}(\mathbf{V}^0)$

$\nabla_{\mathbf{b}} J \leftarrow \nabla_{\mathbf{b}} J - \text{sumrow}(\mathbf{H}^0)$

$\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t + \frac{\epsilon}{M_b} \delta \mathbf{W}_t$ ▷ Update **W**

$\mathbf{c}_{t+1} \leftarrow \mathbf{c}_t + \frac{\epsilon}{M_b} \delta \mathbf{c}_t$ ▷ Update **c**

$\mathbf{b}_{t+1} \leftarrow \mathbf{b}_t + \frac{\epsilon}{M_b} \delta \mathbf{b}_t$ ▷ Update **b**

end while

Return $rbm = \{\mathbf{W,c,b}\}$

end procedure

and the posterior probability can be modeled as Bernoulli distribution as follows

$$P(\mathbf{h}^{\ell}|\mathbf{v}^{\ell-\mathbf{1}}) = \sigma(\mathbf{W}^{\ell}\mathbf{v}^{\ell-\mathbf{1}} + \mathbf{b}^{\ell}), \tag{3.31}$$

and the output layer estimates the label $\mathbf{y}$ based on the input as a multinomial probability distribution as

$$P(\mathbf{y}|\mathbf{v}^{L-1}) = \text{softmax}(\mathbf{z}^{L}) = \text{softmax}(\mathbf{W}^{\ell}\mathbf{v}^{\ell-\mathbf{1}} + \mathbf{b}^{\ell}). \tag{3.32}$$

The exact modeling of $P(\mathbf{y}|\mathbf{o})$ needs integration of all possible hidden values $\mathbf{h}$ from the given input feature $\mathbf{o}$ and the label $\mathbf{y}$. This is in practice so the practical way to replace the marginalization with the mean field approximation by defining

$$\mathbf{v}^{\ell} = \mathbb{E}(\mathbf{h}^{\ell}|\mathbf{v}^{\ell-\mathbf{1}}) = P(\mathbf{h}^{\ell}|\mathbf{v}^{\ell-\mathbf{1}}) = \sigma(\mathbf{W}^{\ell}\mathbf{v}^{\ell-\mathbf{1}} + \mathbf{b}^{\ell}). \tag{3.33}$$

The above equation is equal to Eq. (3.7). Hence, the generative pre-training weights can be used to initialize the DNN. When the pre-training finishes, a randomly initialized softmax output layer is added and the backpropagation is used to fine-tune all the weights in the network discriminatively. Using generative pre-training to initialize DNN weights may potentially improve the performance of the DNN on the testing set [59].

3.2.1 Designed DNN for the Proposed VAD

In the proposed VAD using DNN, the DNN is constructed by five layer RBMs as shown in Fig. 3.10. In the input layer, $\mathbf{a}^0 = \mathbf{o} = \mathbf{P}(m,k) \in \mathbb{R}^{N_0 \times 1}$ denotes the input feature vector of the DNN, where $\mathbf{P}(m,k)$ denotes the log power spectra and the speech period candidates. $\mathbf{P}(m,k)$ is used as a learning instance and is mapped onto the correct speech periods that are identified during the training process. Following Eq. (3.7), the output of DNN through five nonlinear hidden layers as follows,

$$\begin{aligned}\mathbf{a}^{\ell} \;=\; & f(f(f(f(f((\mathbf{a}^{(0)}\mathbf{w}^{(0)} + \mathbf{b}^{(0)})\mathbf{w}^{(1)} + \mathbf{b}^{(1)})\mathbf{w}^{(2)} + \mathbf{b}^{(2)})\mathbf{w}^{(3)} + \mathbf{b}^{(3)}) \\ & +\mathbf{w}^{(4)} + \mathbf{b}^{(4)}) + \mathbf{w}^{(5)} + \mathbf{b}^{(5)}),\end{aligned} \tag{3.34}$$

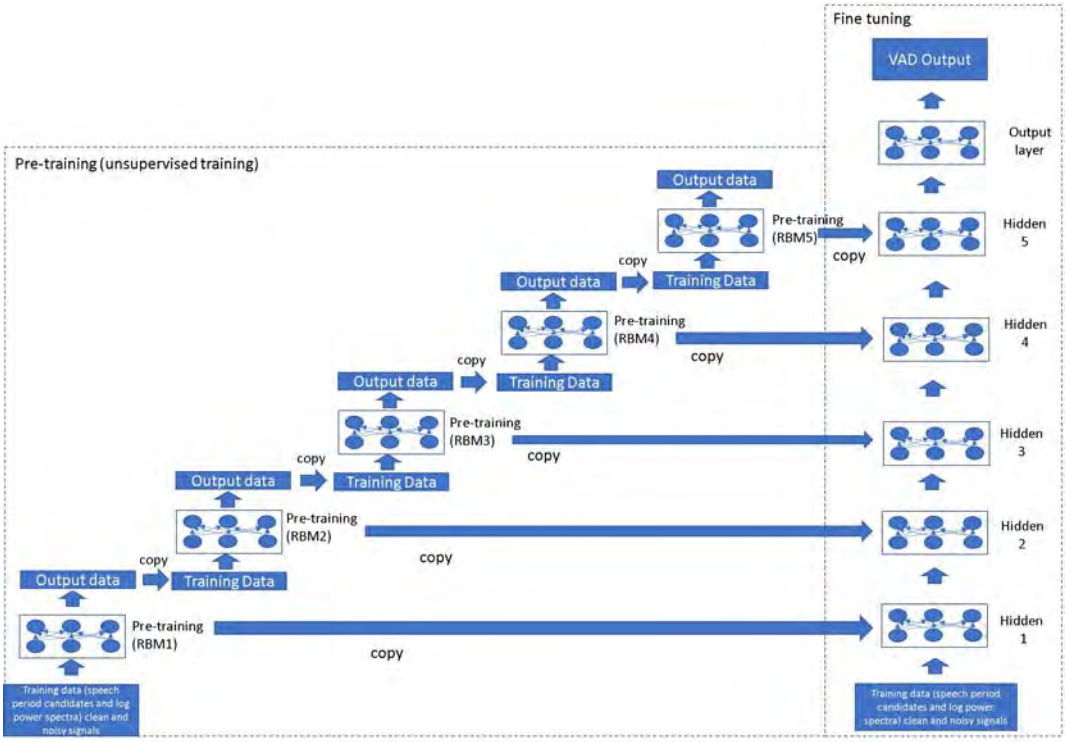


Figure 3.10. Configuration of proposed VAD based DNN

where $\mathbf{a}^{\ell}$ is the DNN output, $\mathbf{w}^{\ell}$ denotes weights and $\mathbf{b}^{\ell}$ denotes biases between the $\ell-$th layer and the $(\ell+1)$th layer. $f(.)$ denotes the activation function. Here, in this study, we use sigmoid, as described in Eq. (3.8) for the activation function followed by softmax as shown in Eq. (3.9) for the output layer. The training procedure of the DNN mentioned above includes two processes. First, a pre-training is performed as a greedy layer-wise unsupervised learning procedure. Contrastive divergence algorithms (see Algorithm 3) are used in the pre-training to approximate the negative log-likelihood gradient of the data with respect to RBM's parameters. The employed DNN in the proposed VAD is composed of five layers of RBMs which consist of visible and hidden neurons, v_j and h_j, respectively. Here, Bernoulli (visible)-Bernoulli (hidden) RBMs, i.e., $v_j \in \{0,1\}$ and $h_j \in \{0,1\}$, are used. Once the learning process has been completed for an RBM, the activity values of its hidden units can be used as *feature input* for training the next RBM [62]. After layer-by-layer pre-training, a backpropagation technique is applied throughout the entire net to fine-tune the weights to obtain optimal results [63] (see Algorithm 2).

Because the VAD output consists of two classes (i.e., speech and non-speech),

the DNN output for each frame, $\mathbf{y}(m)$, is a binary vector with elements determined as,

$$y(m) = \mathbf{a}^L = \begin{cases} 1, & \text{if speech is at frame } m, \text{ and} \\ 0, & \text{otherwise.} \end{cases} \tag{3.35}$$

The DNN outputs trains of 1s (ones) as the speech periods.

Chapter 4

Experiment and Discussion

4.1 Experimental Setup

In the experiments, we use 99 speech files from the ASJ Continuous Speech Corpus for Research vol. 2 [64]. These speech files are divided equally into 3 data set. Then, we create 3 groups, each with 66 files for training (combination of 2 data set to obtain a group is different to another group) and the rest of speech files are used for evaluation purposes. The objective of dividing the data is to evaluate the proposed method inside different data set. To obtain noisy signals, the clean speech files are mixed with 5 types of noise such as white, babble, factory, car and pink from NOISEX-92 [65]. Each noise signal is differently selected for the speech files as well as the SNRs of 10, 5, 0 and -5 dB. Consequently, 21 sets of data are used to produce 1386 speech files for training for each of group.

The input signals are sampled at 8 kHz for the experiments. The frame size is 20 ms, and the analysis window is a Hamming window with a 10 ms frame shift. After filtering process using the RSF, the LPS-RSF is calculated for an individual subband using Eq. (3.4). Then, the first and second derivatives for each individual subband are calculated using Eqs. (3.5) and (3.6). These derivatives are used to obtain the starting and ending point candidates in accordance with rules (i) and (ii). After the conversion of the starting and ending point candidates from the sparse representation to the mask, the masks are multiplied by power spectra, expressed in decimal form, to obtain speech period candidates. Then, the DNN is applied to

the speech period candidates in combination with the log power spectra from Eq. (3.2). This combination is fed to the DNN to obtain the VAD decision regarding the speech periods. Before feeding to the DNN, the input features are normalized to zero mean and unit variance in each dimension. The DNN configuration for the proposed VAD is shown in Fig. 3.10 (see Chapter 3). The DNN is composed by five RBMs which are stacked together. The parameters setting of DNN are shown in Table 4.1.

Table 4.1. Parameters setting of DNN

Number of layers	5 layers
Number of the hidden units in different layers	[200,200,200,200,100]
Learning rate	0.0001
Maximum epochs for pre-training and fine-tuning	200

Note that, after determining the speech and non-speech periods, we do not perform any post processing such as a *vad hangover* because such process is out of the range of this study.

4.2 Evaluation

To represent the proposed method performance, the receiving operation characteristic (ROC) curve, in which the true positive rate (TPR) is plotted against the false alarm rate (FPR), is considered. TPR and FPR are defined as follows,

$$TPR = \frac{TP}{TP + FN}, \text{ and} \tag{4.1}$$

$$FPR = \frac{FP}{FP + TN}, \tag{4.2}$$

where TP (true positive) represents the number of speech frames in the speech periods correctly detected as speech, TN (true negative) shows the number of non-speech frames in the speech periods correctly detected as non-speech, FN (false negative) shows the number of speech frames in the speech periods incorrectly detected as non-speech, and FP (false positive) represents the number of non-speech

frames in the speech periods incorrectly detected as speech. To obtain a quantitative ROC value, the area under the curves (AUCs) are calculated. These AUCs are the main metric for evaluation.

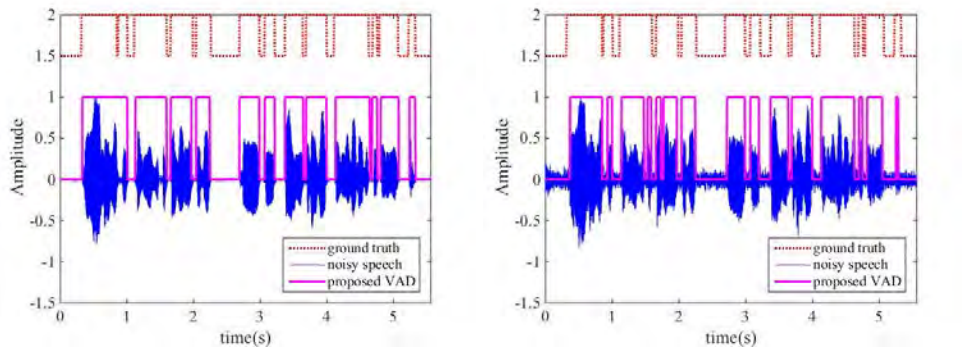
The proposed method is compared with the DNN-based VAD method which utilized the log power spectra, as the base-line of evaluation to evaluate its effectiveness. In the preliminary experiment, the log power spectra are first compared with features that are frequently used in speech processing, i.e., MFCC, delta features and delta-delta features [27,55]. These features were input into the DNN using the same configuration and parameters.

We also compare the proposed method to other well-known methods; VAD based on statistical method [13], Long-Term Spectral Divergence (LTSD) based VAD [9], the likelihood ratio based VAD from Mel-frequency cepstral coefficients (MFCCs) [12], and the combination of contextual, discriminative and spectral cues of human voice [23] to evaluate it.

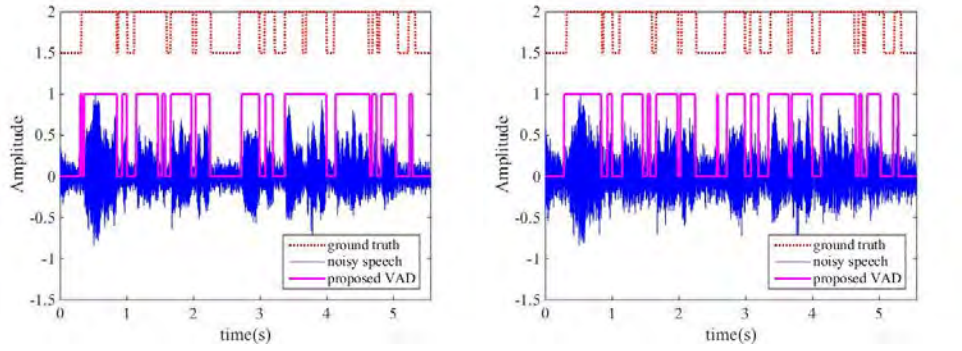
In addition, we also evaluate the useful subbands in the DNN-based VAD to observe which subbands have more valuable information to be used in the training process of the employed DNN.

4.3 Results and Discussion

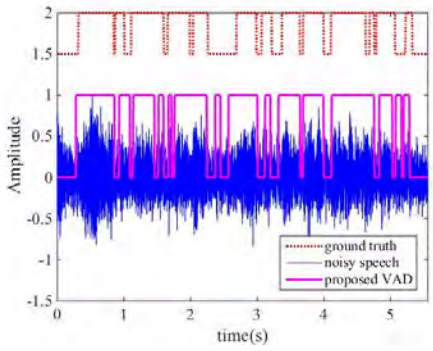
VAD decisions on noisy signals smeared with factory noise at various SNRs are shown in Fig. 4.1. In Fig. 4.1, the red dashed lines indicate the true speech and non-speech periods, whereas the solid magenta lines represent the generated VAD output.The output of the proposed VAD is reasonably close to the correct answers as shown in the figure. In the clean signal, the VAD decision of the proposed method is relatively the same as the correct label. When the signal is smeared by factory noise of 10 dB, there are only small numbers of misjudgments of speech and non-speech periods. Even for the case of -5 dB, the proposed method gives relatively correct judgment for the beginning and ending of the speech signals.



(a) Clean speech (b) Factory noise, SNR= 10 dB



(c) Factory noise, SNR= 5 dB (d) Factory noise, SNR= 0 dB



(e) Factory noise, SNR= -5 dB

Figure 4.1. Representative results of the proposed VAD method

4.3.1 Comparison between the proposed method and simple DNN-based VAD methods

In the preliminary experiment, we compared DNN-based VAD using log power spectra and "hand-crafted" features that are frequently used in speech processing, i.e., MFCC, delta features, and delta-delta features. The MFCCs were calculated using the same window length of 20 ms and the same shift of 10 ms. We consider 13 MFCCs combined with delta features and delta-delta features, resulting in 39

combined features.

Table 4.2. AUC (%) comparison between DNN-based VAD methods using log power spectra and MFCCs. The bold numbers represent the best results

Noise	SNR (dB)	AUC (%)- mean ± standard deviation		
		Log power spectra	MFCCs	MFCCs +Δ + ΔΔ
Clean		**98.72 ± 0.20**	98.18 ± 0.08	97.79 ± 0.41
White	10	**97.51 ± 0.49**	96.10 ± 0.59	96.91 ± 0.57
	5	**97.27 ± 0.48**	93.99 ± 0.85	95.11 ± 0.97
	0	**96.14 ± 0.76**	89.58 ± 1.46	90.85 ± 1.73
	-5	**93.88 ± 1.10**	81.43 ± 1.11	82.42 ± 1.53
Babble	10	**96.50 ± 0.55**	92.71 ± 0.92	93.51 ± 0.77
	5	**94.26 ± 0.66**	87.24 ± 1.03	87.73 ± 0.8
	0	**88.88 ± 0.74**	77.78 ± 0 .99	77.86 ± 0.82
	-5	**78.10 ± 1.10**	65.72 ± 1.40	65.59 ± 1.40
Factory	10	**96.80 ± 0.60**	95.16 ± 0.79	96.04 ± 0.74
	5	**95.14 ± 0.72**	91.60 ± 1.23	92.55 ± 1.06
	0	**91.17 ± 0.45**	84.19 ± 1.23	84.81 ± 1.13
	-5	**80.49 ± 1.54**	72.40 ± 1.14	72.70 ± 0.72
Car	10	**98.83 ± 0.15**	98.34 ± 0.23	98.26 ± 0.32
	5	**98.75 ± 0.16**	98.22 ± 0.34	98.23 ± 0.35
	0	**98.56 ± 0.16**	97.91 ± 0.44	98.08 ± 0.40
	-5	**98.06 ± 0.02**	97.27 ± 0.54	97.70 ± 0.46
Pink	10	**97.20 ± 0.66**	95.91 ± 0.81	96.64 ± 0.62
	5	**96.28 ± 0.79**	93.31 ± 1.00	94.28 ± 0.99
	0	**94.06 ± 0.95**	87.96 ± 1.54	88.91 ± 1.40
	-5	**88.01 ± 1.54**	78.30 ± 1.81	79.02 ± 1.22

As shown in Table 4.2, the DNN-based VAD performance that was achieved using the log power spectra is superior to that achieved using the MFCCs and that using MFCCs in combination with their delta and delta-delta cepstra. The log power spectra features capture more detailed information in the time-frequency domain. Consequently, these features represent a variety of important information that may be related to the speech characteristics. In contrast, the MFCCs may suffer from information loss, which may occur due to the dimension reduction caused by the

discrete cosine transform (DCT) compression. Inside the preliminary study, the DNN-based VAD performance that was achieved using the MFCCs is slightly improved when temporal derivatives, i.e., delta and delta-delta features, are considered in combination with the MFCCs. The enhancement achieved by using the MFCCs in combination with the delta and delta-delta cepstra implies that the dynamics expressed by delta and delta-delta cepstra, play a role to improve the VAD performance.

Table 4.3. AUC (%) comparison between the proposed method and DNN-based VAD methods using speech period candidates and log power spectra as the baseline. The bold numbers represent the best results

Noise	SNR (dB)	AUC (%)- mean ± standard deviation		
		Proposed	Log power spectra	Speech period candidates
Clean		**99.06 ± 0.13**	98.72 ± 0.20	98.10 ± 0.39
White	10	**97.91 ± 0.28**	97.51 ± 0.49	97.06 ± 0.54
	5	**97.44 ± 0.43**	97.27 ± 0.48	96.64 ± 0.46
	0	**96.59 ± 0.50**	96.14 ± 0.76	95.44 ± 0.57
	-5	**94.69 ± 0.66**	93.88 ± 1.10	93.40 ± 0.60
Babble	10	**96.84 ± 0.60**	96.50 ± 0.55	96.19 ± 0.68
	5	**95.19 ± 0.71**	94.26 ± 0.66	94.59 ± 0.92
	0	**91.30 ± 0.74**	88.88 ± 0.74	90.42 ± 0.53
	-5	**83.20 ± 0.87**	78.10 ± 1.10	81.85 ± 0.85
Factory	10	**97.25 ± 0.39**	96.80 ± 0.60	96.60 ± 0.56
	5	**95.96 ± 0.43**	95.14 ± 0.72	95.48 ± 0.77
	0	**93.18 ± 0.46**	91.17 ± 0.45	92.53 ± 0.67
	-5	**85.91 ± 0.29**	80.49 ± 1.54	84.57 ± 0.83
Car	10	**99.02 ± 0.11**	98.83 ± 0.15	97.60 ± 0.45
	5	**98.94 ± 0.11**	98.75 ± 0.16	97.37 ± 0.45
	0	**98.79 ± 0.09**	98.56 ± 0.16	97.02 ± 0.41
	-5	**98.40 ± 0.05**	98.06 ± 0.02	96.36 ± 0.32
Pink	10	**97.79 ± 0.39**	97.20 ± 0.66	96.86 ± 0.73
	5	**96.82 ± 0.59**	96.28 ± 0.79	95.98 ± 0.74
	0	**95.26 ± 0.70**	94.06 ± 0.95	94.26 ± 0.89
	-5	**91.56 ± 1.03**	88.01 ± 1.54	89.91 ± 1.20

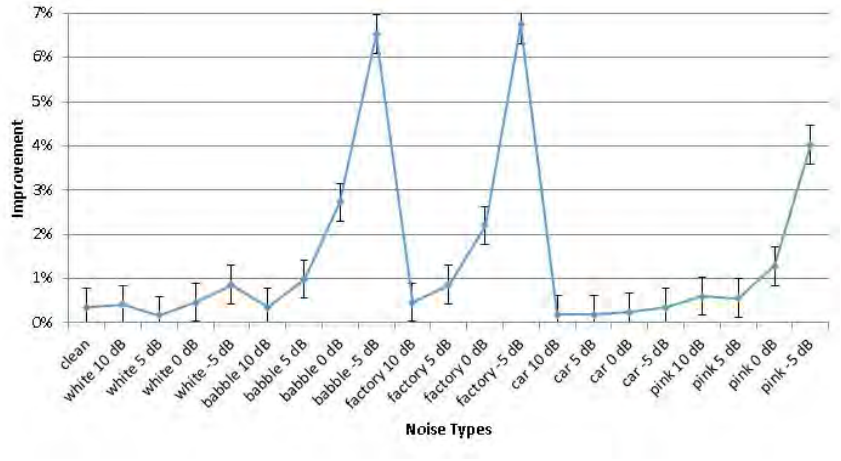


Figure 4.2. Improvement of VAD performance

The proposed method is used to improve the DNN-based VAD performance using the log power spectra alone by introducing speech period candidates. The results are shown in Table 4.3 that compares the average of AUCs achieved by DNN-based VAD methods using the log power spectra alone, the speech period candidates alone, and the proposed method. As shown in Table 4.3, the proposed method may improve the performance of DNN-based VAD method using the log power spectra alone for all cases. As shown in Fig. 4.2, a relatively good improvement is obtained in low SNR cases. The highest improvement occurs in low SNR signals at -5 dB, for example, the performance improves by 6.52% for babble noise.

The addition of speech period candidates to the log power spectra may add discriminant information to detect speech and non-speech periods. Figure 4.3 shows the DNN output from speech period candidates, log power spectra and the proposed method. As shown in the figure, in the beginning and ending of the speech signal which is smeared by babble noise at SNR 0 dB, the DNN-based VAD using log power spectra misclassifies more non-speech periods than the proposed method. The proposed method may identify non-speech periods better than the log power spectra. This indicates that the proposed method offers advantage over the usage of log power spectra as it may reduce misclassification caused by the use of log power spectra.

To evaluate the effect of introducing the speech period candidates, we measure the TPR (sensitivity) and the true negative rate (TNR or specificity). Sensitivity

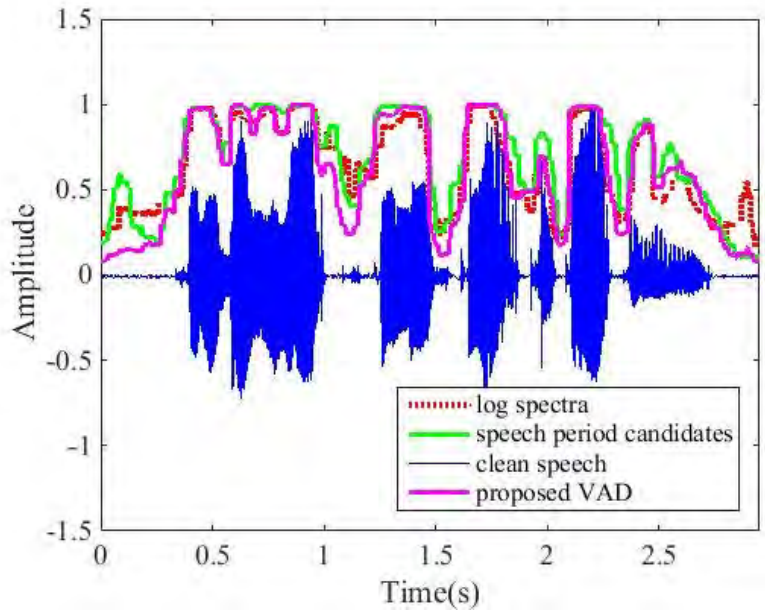
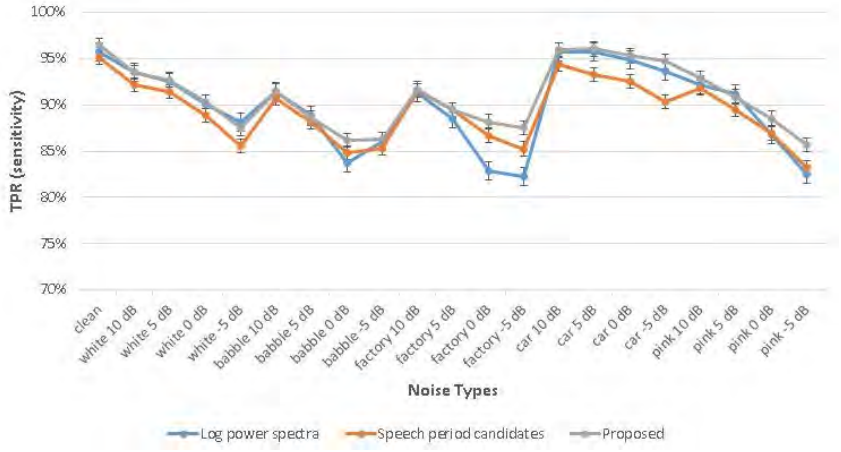


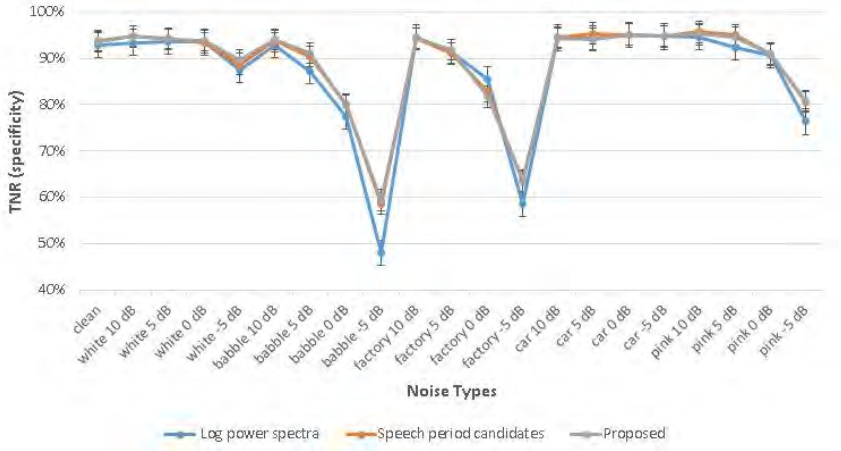
Figure 4.3. DNN-based VAD output in noisy speech (babble noise, SNR = 0 dB)

represents the percentage of correct frames detected as speech from all the speech frames and specificity represents the percentage of correct frames detected as non-speech from all the non-speech frames in the signal [66].

Figure 4.4 shows the mean TPR (sensitivity) and TNR (specificity). As shown in the figure, the proposed method has a high sensitivity and specificity for both the clean and noisy cases. Interestingly, the performance of the DNN-based VAD method using speech period candidates approaches and even outperforms the log power spectra in finding speech periods, as shown in Fig. 4.4a, particularly for low SNRs and non-stationary cases. This fact may imply that the addition of speech period candidates is useful to find speech periods for low SNRs and non-stationary cases. The specificity in speech period candidates is higher than the log power spectra as shown in Fig. 4.4b. This fact may imply that the speech period candidates may improve the log power spectra for finding non-speech periods. Thus, the speech period candidates do carry valuable information for judging speech and non-speech detection. They may reduce misclassification caused by the use of log power spectra. Hence, in the proposed method, the addition speech period candidates is effective at improving the performance of the log power spectra at finding speech and non-speech periods, especially for low SNRs and non-stationary cases.



(a) TPR (sensitivity)

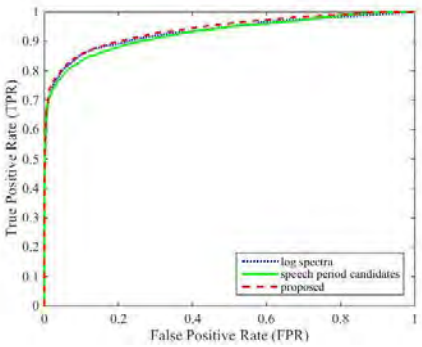


(b) TNR (specificity)

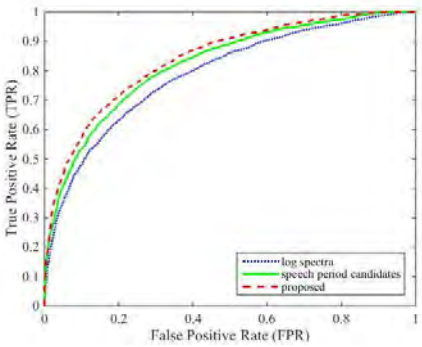
Figure 4.4. Sensitivity and specificity comparison between the proposed method, and the DNN-based VAD methods using speech period candidates and log power spectra

Figure 4.5 shows the ROC curves for the proposed method and the DNN-based VAD method using speech period candidates and log power spectra, respectively, at an SNR of -5 dB. As shown in the figure, the proposed method shows an advantage over the DNN-based VAD method using the log power spectra. The proposed method is effective for low SNR cases. In the cases of white and pink noise which represents the stationary noise, the working points of the proposed method are close to those of the DNN-based VAD method using the log power spectra. These methods achieve a high TPR and a low FPR. In the cases of babble and factory noise which represent non-stationary noise, the proposed method is less affected by the noise than the DNN-based VAD method using the log power spectra. The performance of the proposed method is superior to that of the DNN-based VAD

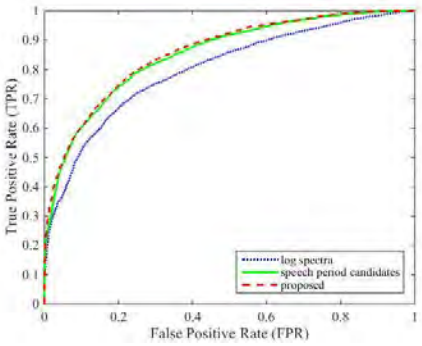
method using the log power spectra mainly due to introducing dynamics expressed by speech period candidates.



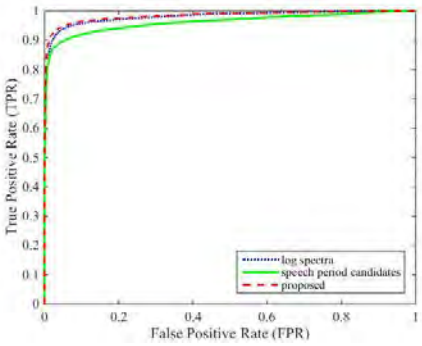
(a) White noise, SNR=-5 dB



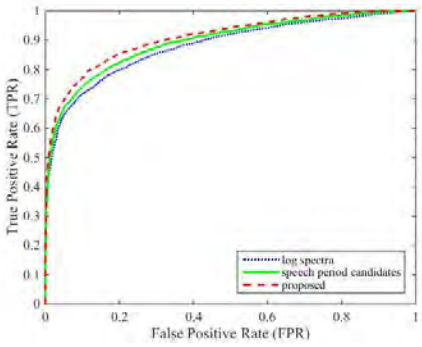
(b) Babble noise, SNR=-5 dB



(c) Factory noise, SNR=-5 dB



(d) Car noise, SNR=-5 dB



(e) Pink noise, SNR=-5 dB

Figure 4.5. ROC curves for the proposed method and for DNN-based VAD methods using log power spectra and speech period candidates, respectively

4.3.2 Evaluation of useful subbands

The employed DNN can learn characteristics carried by the input (log power spectra incorporating with speech period candidate). The input features may have strong local structure, sufficient variability, and model generatively to be used by

the DNN. When the input is fed to the DNN, the features carried by the input are transformed to be more variant and discriminative features. The hidden layers which are close to the input layer represents low level features capture local patterns that are sensitive to the change of input features. Meanwhile, the hidden layers which are close to the output layer represents high level features that are invariant and more abstract [59].

Here, in addition to the contribution of the speech period candidates, which may highlight dynamics, we attempt to find useful subbands for obtaining VAD decisions in the employed DNN. We evaluate which subbands have more valuable information than the others, by finding the similarity between the input (i.e., speech period candidates) and the VAD output. This similarity is evaluated by employing mutual information (MI), which aims to measure whether the inputs are dependent on the associated labels (VAD output). The mutual information between the discretized feature values, a, and the class labels, y, is evaluated according to the formula [67]

$$MI = \sum_{a \in A} \sum_{y \in Y} p(a, y) \log(\frac{p(a, y)}{p(a)p(y)}), \tag{4.3}$$

where $p(a, y)$ is a joint probability function of a and y, $p(a)$ is a marginal probability distribution functions of a and $p(y)$ is of y. Here, the feature values, a, are the input of the DNN (speech period candidates), and the class labels, y, are the VAD output. The larger the MI, the higher the dependency between the feature values, which represent speech period candidates for individual subbands, and the class labels (VAD output). Here, we rank the subbands according to their scores.

Table 4.4 shows the top 4 subband ranks using MI. As shown in Tabel 4.4, the top 4 ranks show a similar tendency for clean and noisy signals. They mostly occur in frequency bins 6, 7, 8, and 9 (156.25 Hz, 187.5 Hz, 218.75 Hz, and 250 Hz). Such subband might play some roles in obtaining the VAD decision in the proposed method. To clarify this, we perform experiments in which the 4 top subband values in the proposed method are replaced with zeros, and the resulting VAD performance is shown in Fig. 4.6.

As shown in Fig. 4.6, at a high SNR, the performance of the proposed method is

Table 4.4. Subband (Hz) ranks using mutual information (MI)

		Subband (Hz) ranks using MI			
Noise	SNR (dB)	1	2	3	4
Clean		187.5	218.75	156.25	312.5
White	10	187.5	218.75	156.25	250
	5	187.5	218.75	156.25	250
	0	187.5	218.75	156.25	250
	-5	187.5	218.75	156.25	250
Babble	10	187.5	218.75	250	156.25
	5	187.5	218.75	250	156.25
	0	187.5	218.75	250	156.25
	-5	187.5	218.75	250	156.25
Factory	10	187.5	218.75	156.25	250
	5	187.5	218.75	156.25	250
	0	187.5	218.75	250	156.25
	-5	218.75	187.5	250	312.5
Car	10	187.5	218.75	156.25	250
	5	187.5	218.75	312.5	250
	0	187.5	218.75	250	312.5
	-5	218.75	187.5	250	312.5
Pink	10	187.5	218.75	156.25	250
	5	187.5	218.75	250	156.25
	0	187.5	218.75	250	156.25
	-5	187.5	218.75	250	156.25

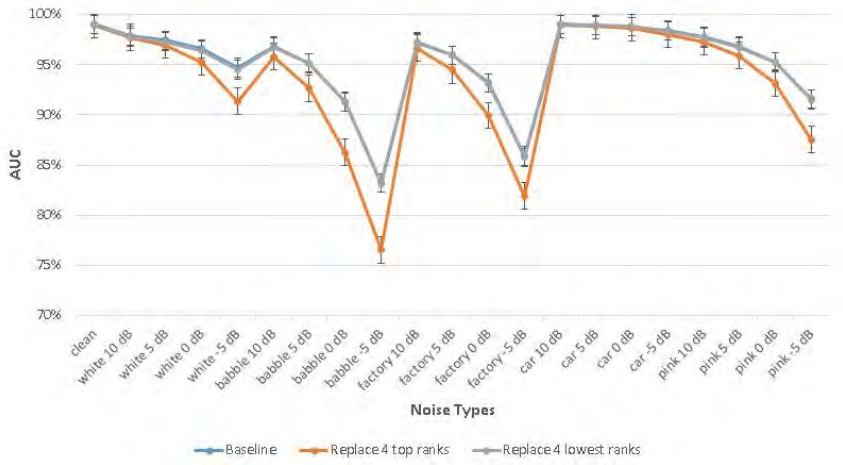


Figure 4.6. VAD performance of the proposed method after replacing the subband values of the top 4 ranks' and the lowest 4 ranks' with zeros

only slightly degraded when the frequency subbands of 156.25 Hz, 187.5 Hz, 218.75 Hz, and 250 Hz are replaced by zeros. In low SNR cases, the subbands are polluted

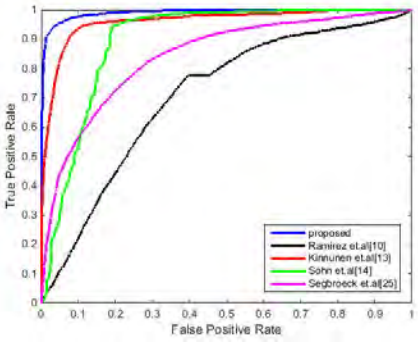
by noise. Consequently, the performance might be degraded, and this degradation worsens when these top 4 subbands are not utilized. In contrast, when the 4 lowest subbands are eliminated, the output accuracy can still be maintained. These results indicate that the top 4 subbands have a relatively important role in the decision-making process of the proposed method. We observe that the information carried by these subbands may correspond to the average F0 or its neighbors (the average F0 for the data is 179.87 Hz, the average F0 for a male is 149.09 Hz and 210.84 Hz for a female). Thus, in the proposed method, the DNN may utilize information coming from useful subbands which may correspond to the F0 or its neighbours.

4.3.3 Comparison between proposed method and other methods

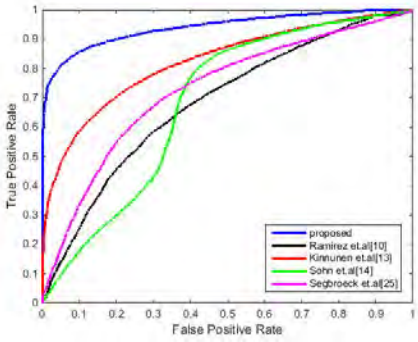
We also compare the proposed method with four other methods presented in [9], [12], [13] and [23]to evaluate its effectiveness. Illustration of VAD decisions for those four methods are shown in Appendix A.

Figure 4.7 shows the ROC comparison of the proposed and the other methods in clean and different noise signals with SNR -5 dB. Figure 4.7a shows that the proposed method outperforms the other methods in [9], [12], [13] and [23]. From Fig. 4.7b-f, we observe that the proposed method has the best performance among all methods for noisy signals of -5 dB. ROC curves show that when the background noise level increases, the proposed method is less affected than other methods. It enables working on the optimal point of ROC curve. On the contrary, the working points of other methods shift to the right when SNR declines in ROC space. The method proposed in [12] which use MFCCs yields clear advantages over other conventional methods in white, babble and car noise. In clean signal, the statistical method proposed by [13] has relatively good TPR which indicates high correct speech periods. But, it has more misjudgments because of high FPR. In babble noise -5 dB, the performance of [13], [9] and [23] are close to each other. On the other hand, the performance of [12] is close to [23] for factory noise -5 dB. The proposed VAD also shows better performance in SNRs of 10, 5 and 0 dB (see Appendix B).

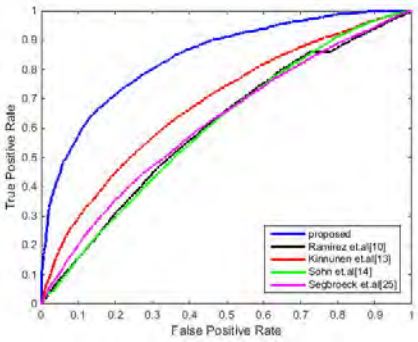
To get quantitative comparisons, we consider AUC. The results are summarized



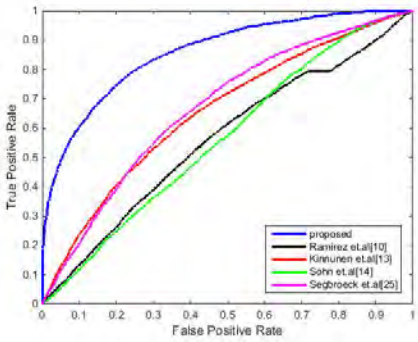
(a) Clean speech



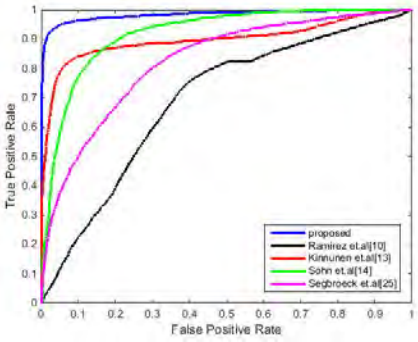
(b) White noise -5 dB



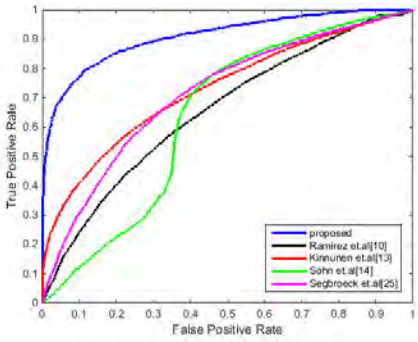
(c) Babble noise -5 dB



(d) Factory noise -5 dB



(e) Car noise -5 dB



(f) Pink noise -5 dB

Figure 4.7. ROC curve comparison between proposed method and other methods

in Table 4.5. As shown in the table, the proposed method is superior to other methods for both clean and noisy signals. The performance of the method in [13], which utilizes a statistical method, approaches the performance of the method in [12], which utilizes MFCCs and a GMM as the classifier. Their performance worsens for non-stationary noise. Alternatively, the method in [23] can give better performance for low SNRs and non-stationary noise such as factory and babble than the method in [9]. The proposed method is superior to that of the other methods due to the advantages of using a DNN and features to input, i.e., speech period candidates and log power spectra.

Table 4.5. AUC (%) comparison between the proposed method and other methods (Ramirez *et al.* [9], Kinnunen *et al.* [12], Sohn *et al.* [13], and Segbroeck *et al.* [23]). The bold numbers represent the best results

		AUC (%) - mean ± standard deviation				
Noise	SNR (dB)	Proposed	Ramirez	Kinnunen	Sohn	Segbroeck
Clean		**99.06 ± 0.13**	71.03 ± 1.02	95.65 ± 0.43	88.48 ± 1.45	84.57 ± 1.21
White	10	**97.91 ± 0.28**	74.09 ± 1.50	93.65 ± 1.05	93.32 ± 0.50	79.63 ± 0.24
	5	**97.44 ± 0.43**	73.57 ± 1.36	93.02 ± 1.31	87.84 ± 0.50	77.75 ± 0.34
	0	**96.59 ± 0.50**	72.33 ± 1.24	91.06 ± 1.59	77.34 ± 1.14	75.21 ± 0.67
	-5	**94.69 ± 0.66**	68.97 ± 1.07	83.85 ± 1.28	66.79 ± 1.85	71.89 ± 0.77
Babble	10	**96.84 ± 0.60**	68.84 ± 1.32	87.71 ± 0.91	87.56 ± 0.87	81.25 ± 0.35
	5	**95.19 ± 0.71**	67.28 ± 0.71	84.19 ± 0.80	79.97 ± 0.70	79.05 ± 0.69
	0	**91.30 ± 0.74**	63.62 ± 0.91	76.59 ± 0.99	70.05 ± 0.94	72.99 ± 1.13
	-5	**83.20 ± 0.87**	59.37 ± 1.01	66.73 ± 1.52	60.33 ± 0.93	62.71 ± 1.25
Factory	10	**97.25 ± 0.39**	70.35 ± 1.85	88.12 ± 1.73	88.04 ± 0.67	81.19 ± 0.96
	5	**95.96 ± 0.43**	67.54 ± 1.57	84.42 ± 1.67	79.55 ± 0.85	78.99 ± 1.06
	0	**93.18 ± 0.46**	62.78 ± 1.53	77.70 ± 1.07	67.15 ± 0.82	74.67 ± 0.87
	-5	**85.91 ± 0.29**	57.81 ± 1.71	66.38 ± 0.76	56.28 ± 0.56	67.23 ± 0.52
Car	10	**99.02 ± 0.11**	69.06 ± 1.60	94.62 ± 0.36	91.56 ± 1.29	84.46 ± 0.96
	5	**98.94 ± 0.11**	68.27 ± 1.32	93.64 ± 0.80	92.15 ± 0.83	84.38 ± 0.96
	0	**98.79 ± 0.09**	68.50 ± 1.02	92.42 ± 0.40	92.41 ± 0.34	84.08 ± 0.91
	-5	**98.40 ± 0.05**	68.77 ± 1.87	90.16 ± 0.17	91.81 ± 0.10	83.49 ± 1.06
Pink	10	**97.79 ± 0.39**	73.11 ± 1.67	90.37 ± 1.33	90.54 ± 0.40	81.00 ± 0.94
	5	**96.82 ± 0.59**	72.36 ± 1.63	88.87 ± 1.85	82.51 ± 1.39	78.96 ± 1.21
	0	**95.26 ± 0.70**	70.32 ± 1.53	84.13 ± 1.76	71.70 ± 1.70	76.10 ± 1.31
	-5	**91.56 ± 1.03**	65.69 ± 1.40	74.46 ± 1.36	62.81 ± 1.82	71.78 ± 1.04

Chapter 5

Conclusion and Future Work

5.1 Conclusion

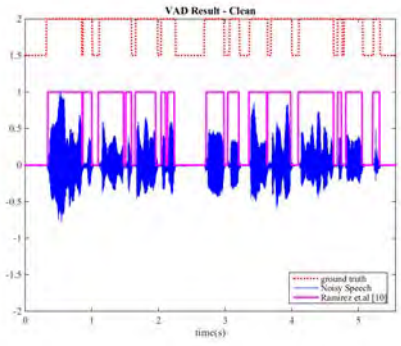
This study presents a DNN-based VAD method for improving the performance of VAD by introducing dynamics which may be highlighted by speech period candidates. These candidates are derived from the heuristic rules based on the first and second derivatives of the LPS-RSF. The speech period candidates are calculated for individual subbands and are then input into the DNN together with the log power spectra to generate the VAD decision. To evaluate the performance of the proposed method, we performed experiments using clean and noisy speech signals smeared with five types of noise, namely, white, babble, factory, car and pink, with SNRs of 10, 5, 0 and -5 dB. The experimental results indicate that the DNN-based VAD performance using the log power spectra is enhanced after utilizing the log power spectra together with the speech period candidates, particularly for noisy speech signals in low SNR and non-stationary cases. The addition of dynamics expressed by the speech period candidates provides positive information that contributes to the detection of speech periods. The information carried by individual subbands are also important for obtaining the VAD decision. In this study, we show that the employed DNN-based VAD utilizes subbands which may correspond to F0 or its neighbors.

5.2 Future work

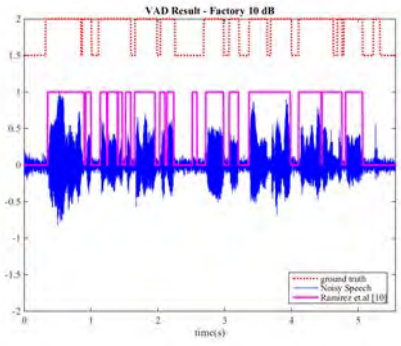
In this study, the DNN seems to learn the relations between the input (i.e., log power spectra and speech period candidates) and the correct speech periods in the training process. Through this study, we show that the DNN utilized information pieces coming from subbands which may correspond to F0 and its neighbors with relatively a good SNR. The VAD performance degrades when those subbands are eliminated. Further studies should be performed to clarify this and other factors that influence the behavior of the employed DNN. Moreover, the proposed method currently works in off-line which require access to the entire voice utterances. In the future, we intend to upgrade the proposed method to work in real time.

Appendix A

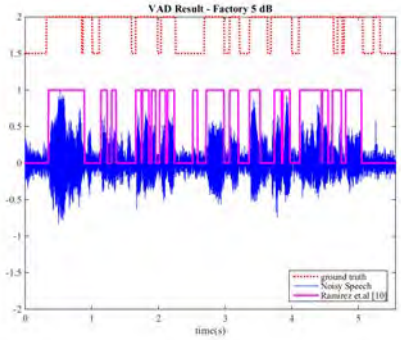
Illustration of Other Methods



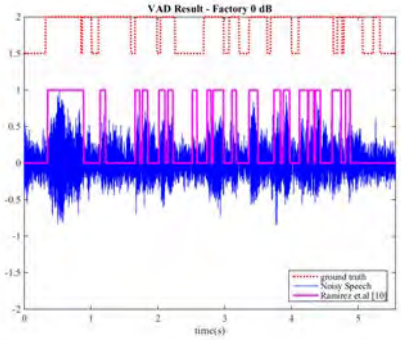
(a) Clean speech



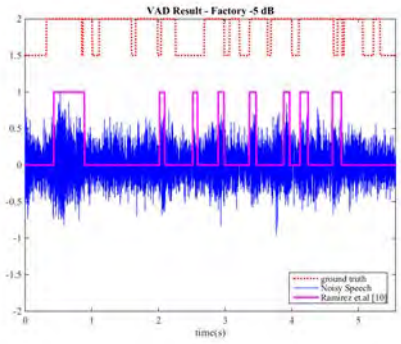
(b) Factory noise 10 dB



(c) Factory noise 5 dB

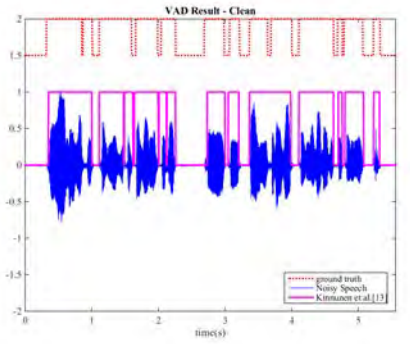


(d) Factory noise 0 dB

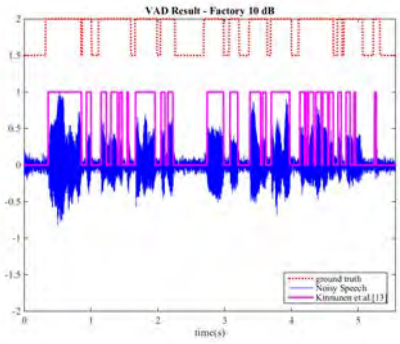


(e) Factory noise -5 dB

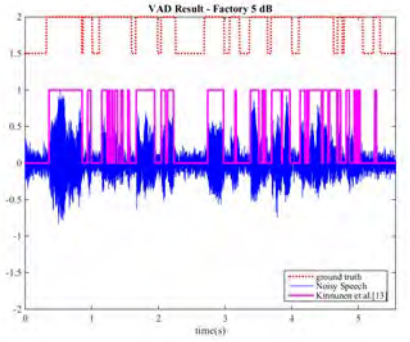
Figure A.1. Illustration results of VAD proposed by Ramirez *et al.* [9]



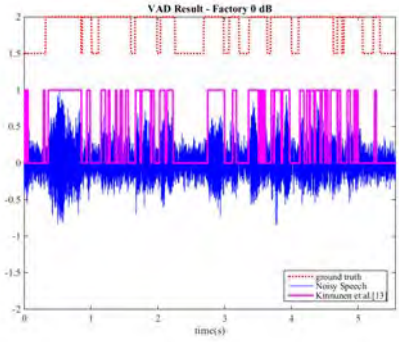
(a) Clean speech



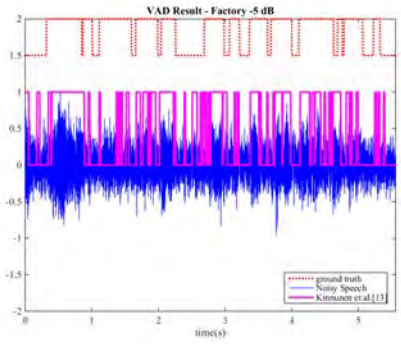
(b) Factory noise 10 dB



(c) Factory noise 5 dB

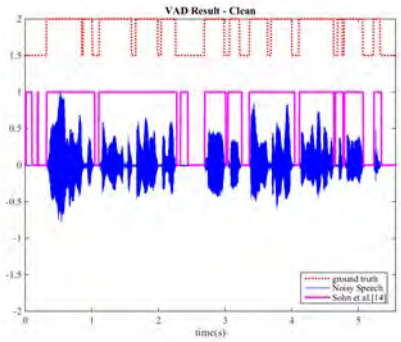


(d) Factory noise 0 dB

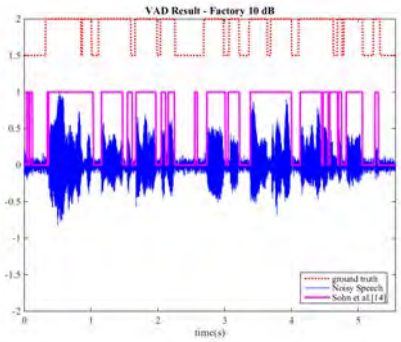


(e) Factory noise -5 dB

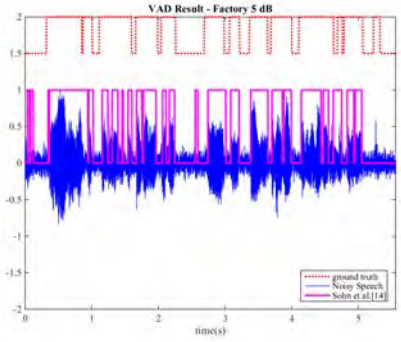
Figure A.2. Illustration results of VAD proposed by Kinnunen *et al.* [12]



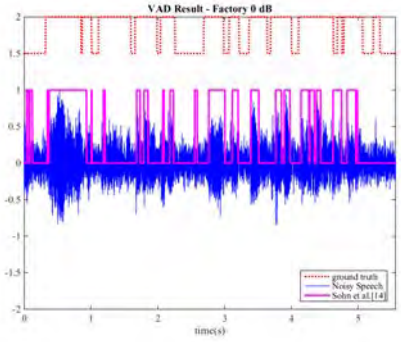
(a) Clean speech



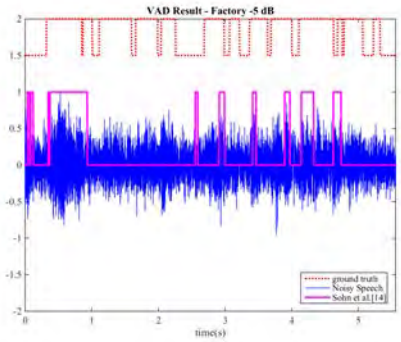
(b) Factory noise 10 dB



(c) Factory noise 5 dB

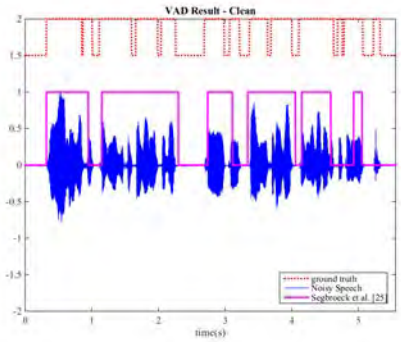


(d) Factory noise 0 dB

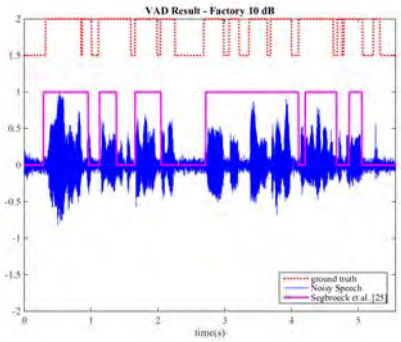


(e) Factory noise -5 dB

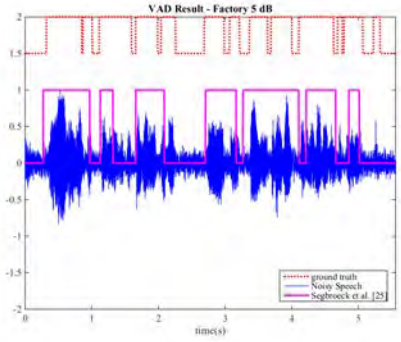
Figure A.3. Illustration results of VAD proposed by Sohn *et al.* [13]



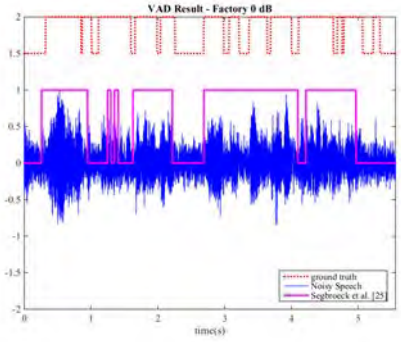
(a) Clean speech



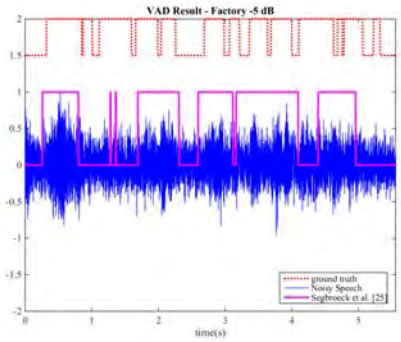
(b) Factory noise 10 dB



(c) Factory noise 5 dB



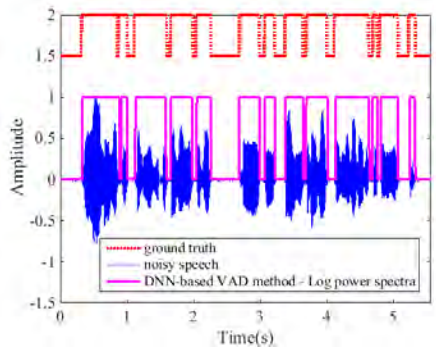
(d) Factory noise 0 dB



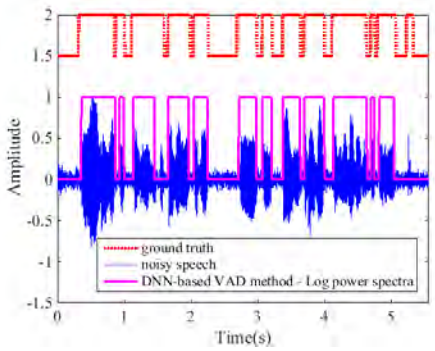
(e) Factory noise -5 dB

Figure A.4. Illustration results of VAD proposed by Segbroeck *et al.* [23]

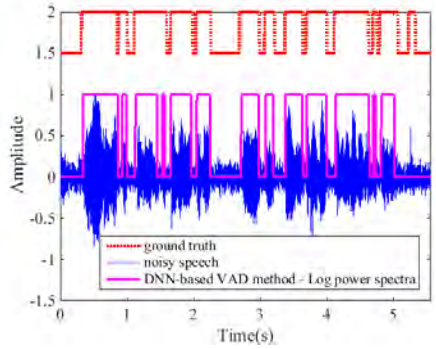
Illustration of DNN-based VAD Methods



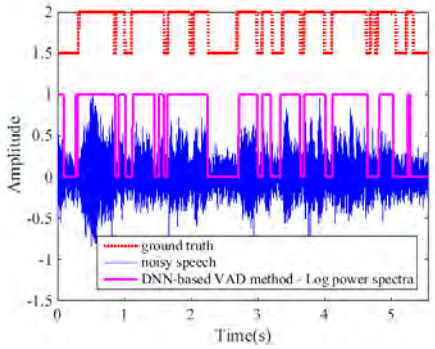
(a) Clean speech



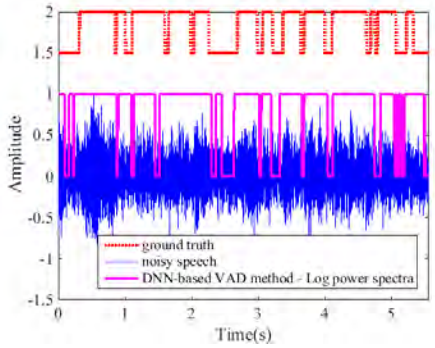
(b) Factory noise 10 dB



(c) Factory noise 5 dB

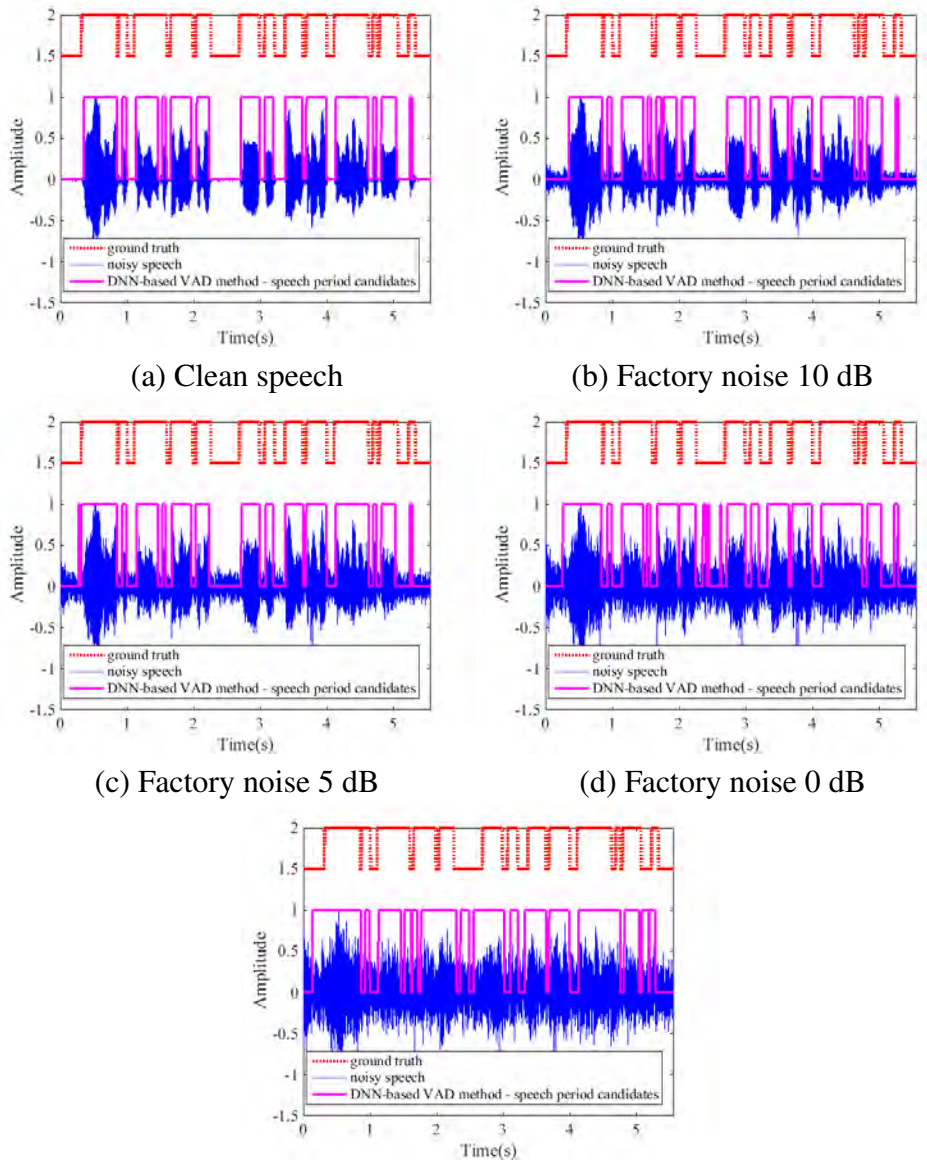


(d) Factory noise 0 dB



(e) Factory noise -5 dB

Figure A.5. Illustration results of DNN-based VAD method using log power spectra



(a) Clean speech

(b) Factory noise 10 dB

(c) Factory noise 5 dB

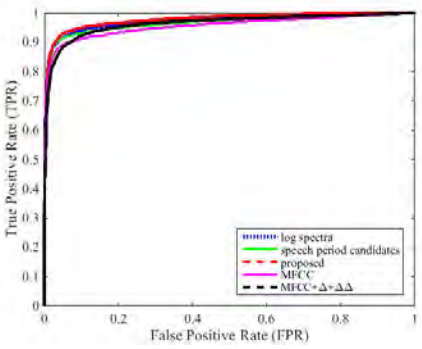
(d) Factory noise 0 dB

(e) Factory noise -5 dB

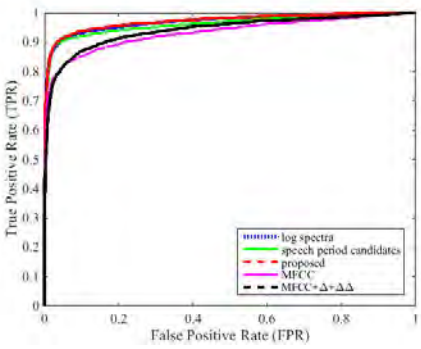
Figure A.6. Illustration results of DNN-based VAD method using speech period candidates

Appendix B

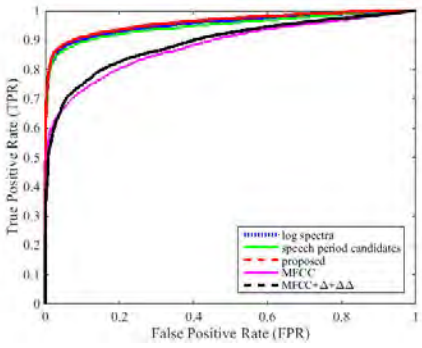
ROC Curve Comparison between Proposed Method and DNN-based VAD Methods



(a) White noise 10 dB

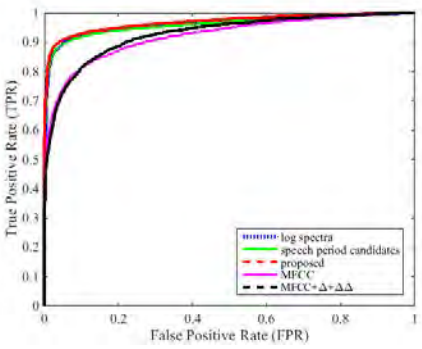


(b) White noise 5 dB

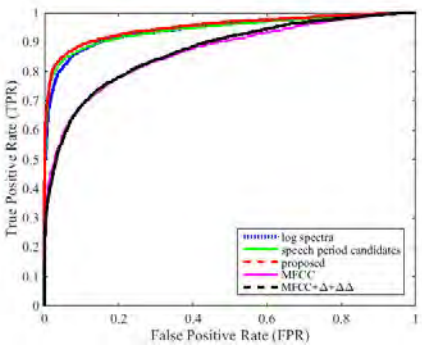


(c) White noise 0 dB

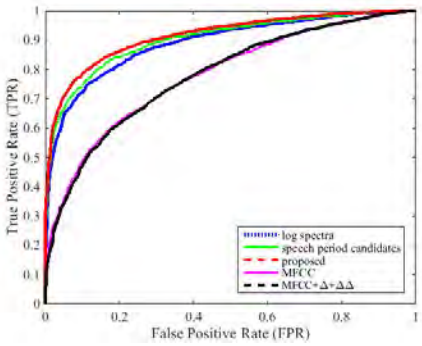
Figure B.1. ROC curve of noisy speech - white noise



(a) Babble noise 10 dB

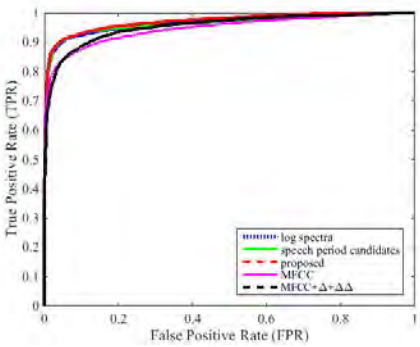


(b) Babble noise 5 dB

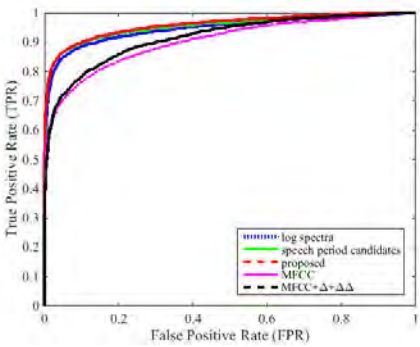


(c) Babble noise 0 dB

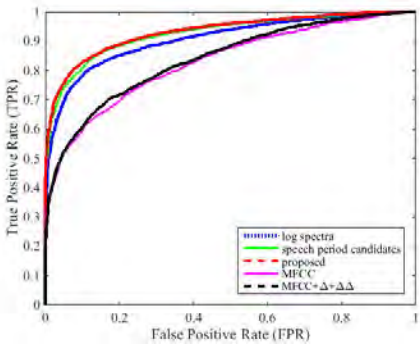
Figure B.2. ROC curve of noisy speech - babble noise



(a) Factory noise 10 dB

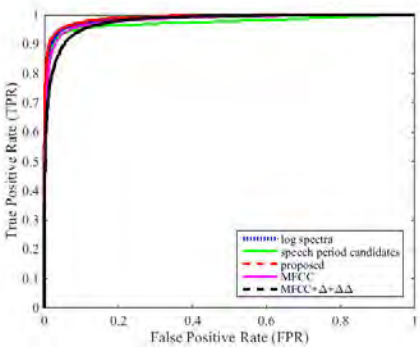


(b) Factory noise 5 dB

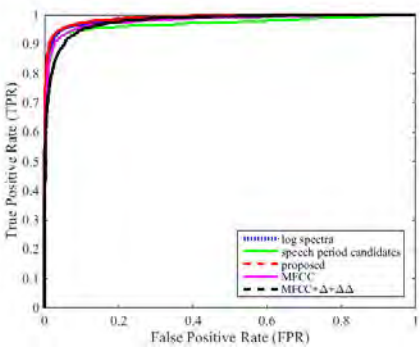


(c) Factory noise 0 dB

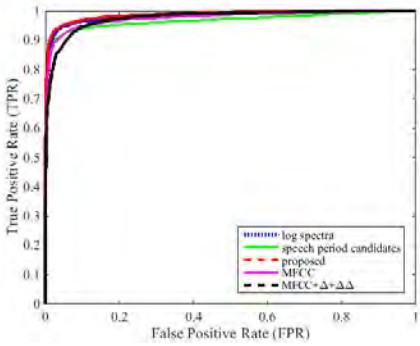
Figure B.3. ROC curve of noisy speech - factory noise



(a) Car noise 10 dB

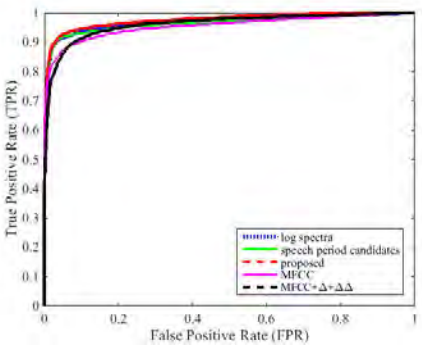


(b) Car noise 5 dB

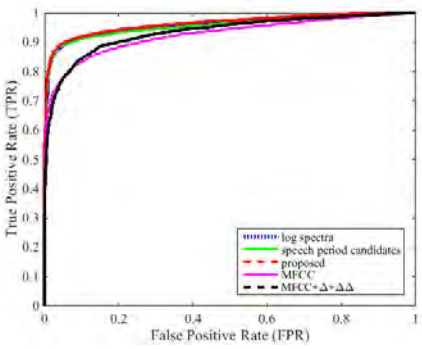


(c) Car noise 0 dB

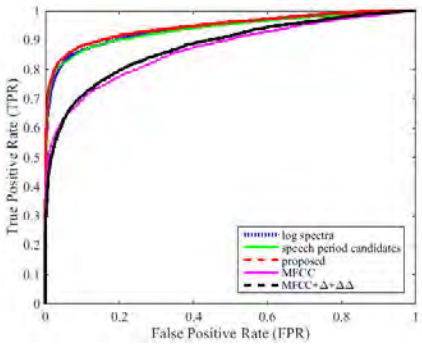
Figure B.4. ROC curve of noisy speech - car noise



(a) Pink noise 10 dB



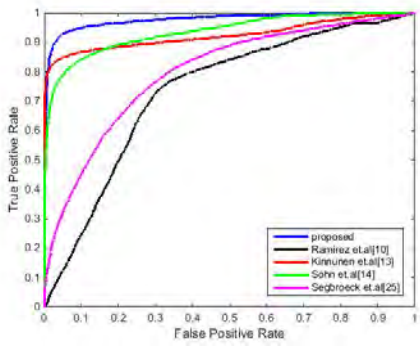
(b) Pink noise 5 dB



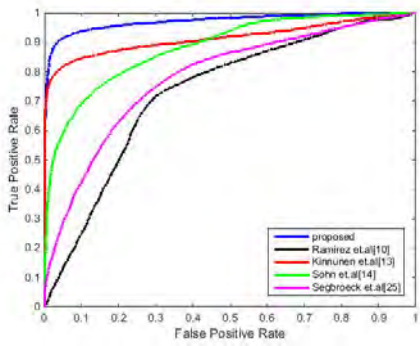
(c) Pink noise 0 dB

Figure B.5. ROC curve of noisy speech - pink noise

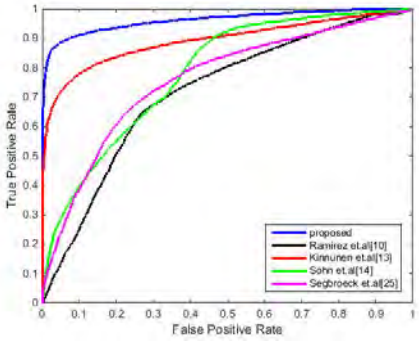
ROC Curve Comparison between Proposed Method and Other Methods



(a) White noise 10 dB

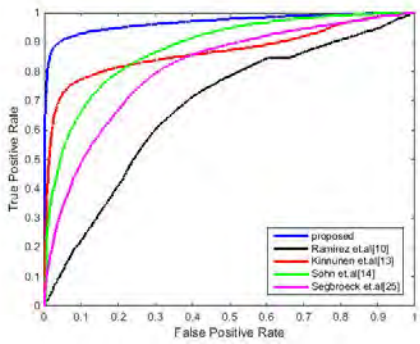


(b) White noise 5 dB

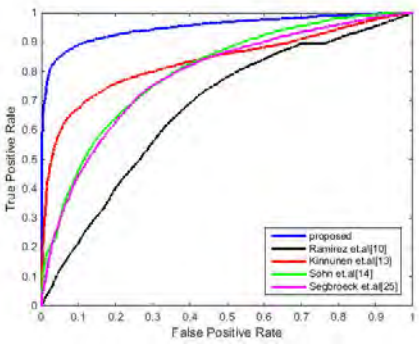


(c) White noise 0 dB

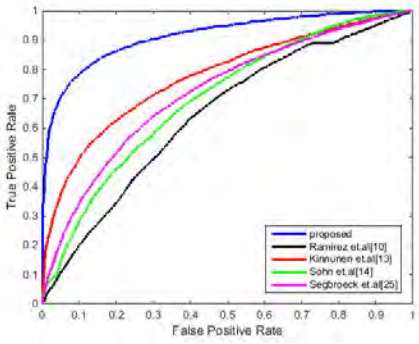
Figure B.6. ROC curve of noisy speech - white noise



(a) Babble noise 10 dB

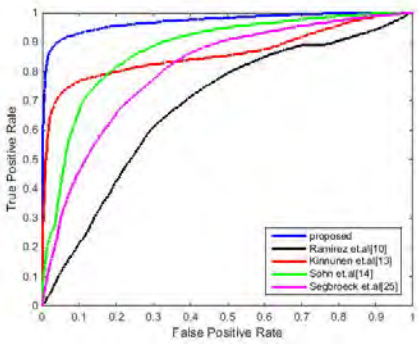


(b) Babble noise 5 dB

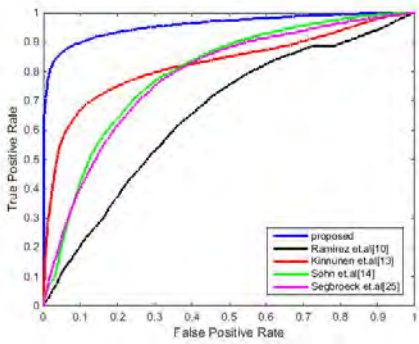


(c) Babble noise 0 dB

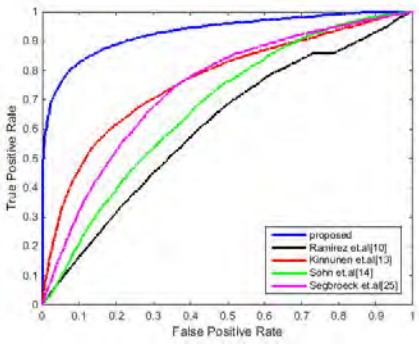
Figure B.7. ROC curve of noisy speech - babble noise



(a) Factory noise 10 dB

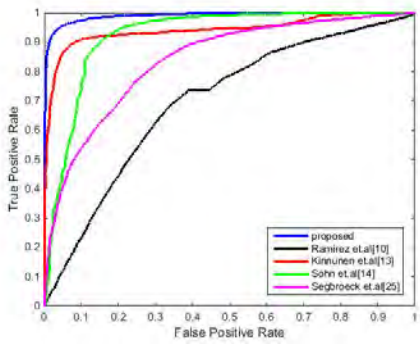


(b) Factory noise 5 dB

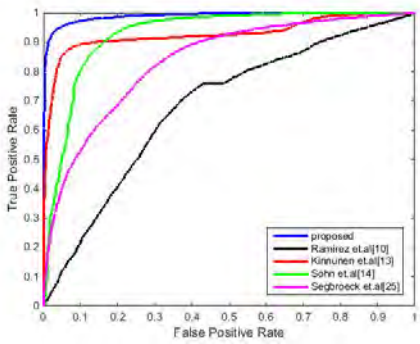


(c) Factory noise 0 dB

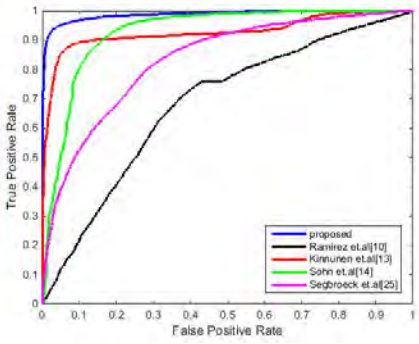
Figure B.8. ROC curve of noisy speech - factory noise



(a) Car noise 10 dB

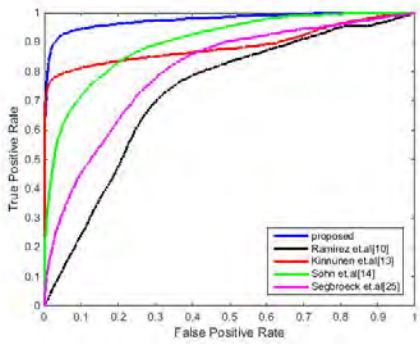


(b) Car noise 5 dB

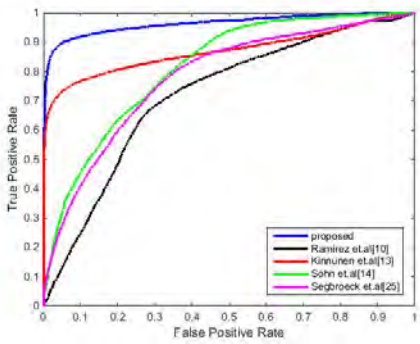


(c) Car noise 0 dB

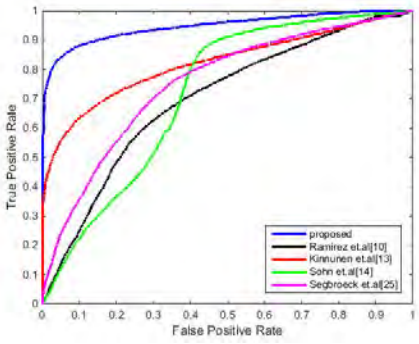
Figure B.9. ROC curve of noisy speech - car noise



(a) Pink noise 10 dB



(b) Pink noise 5 dB



(c) Pink noise 0 dB

Figure B.10. ROC curve of noisy speech - pink noise

Publications

Journal paper

1) Suci Dwijayanti, Kei Yamamori and Masato Miyoshi, Enhancement of speech dynamics for voice activity detection using DNN, *EURASIP Journal on Audio, Speech and Music Processing*, No. 1,Vol. 2018, (ISSN: 1687-4722, DOI: https://doi.org/10.1186/s13636-018-0135-7).

2) Suci Dwijayanti and Masato Miyoshi, Evaluation of Features for Voice Activity Detection Using a Deep Neural Network, *Journal of Theoretical and Applied Information Technology*, No. 9, Vol. 96. (ISSN: 1992-8645).

Conference paper

1) S. Dwijayanti and M. Miyoshi: Voice activity detection in noisy environment using dynamic changes of speech in a modulation frequency range of 1 to 16 Hz, The Journal of the Acoustical Society of America 140.4 pp. 3115-3115, 2016.

2) S. Dwijayanti and M. Miyoshi: A novel method to detect rising and setting tones, The Journal of the Acoustical Society of America 140.4 pp. 3404-3404, 2016.

Bibliography

[1] 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.

[2] F. Beritelli, S. Casale, and G. Ruggeri, "Performance evaluation and comparison of itu-t/etsi voice activity detectors," in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, vol. 3, pp. 1425–1428, IEEE, 2001.

[3] J. Ramirez, J. M. Górriz, and J. C. Segura, *Voice activity detection. fundamentals and speech recognition system robustness*. INTECH Open Access Publisher NewYork, 2007.

[4] R. Le Bouquin-Jeannès and G. Faucon, "Study of a voice activity detector and its influence on a noise reduction system," *Speech communication*, vol. 16, no. 3, pp. 245–254, 1995.

[5] K. Sreekumar, K. K. George, K. Arunraj, and C. S. Kumar, "Spectral matching based voice activity detector for improved speaker recognition," in *Power Signals Control and Computations (EPSCICON), 2014 International Conference on*, pp. 1–4, IEEE, 2014.

[6] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *Bell Labs Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[7] K. Sakhnov, E. Verteletskaya, and B. Simak, "Approach for energy-based voice detector with adaptive scaling factor," *IAENG International Journal of Computer Science*, vol. 36, no. 4, p. 394, 2009.

[8] R. V. Prasad, A. Sangwan, H. Jamadagni, M. Chiranth, R. Sah, and V. Gaurav, "Comparison of voice activity detection algorithms for voip," in *Computers and Communications, 2002. Proceedings. ISCC 2002. Seventh International Symposium on*, pp. 530–535, IEEE, 2002.

[9] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3, pp. 271–287, 2004.

[10] K. Ishizuka, T. Nakatani, M. Fujimoto, and N. Miyazaki, "Noise robust voice activity detection based on periodic to aperiodic component ratio," *Speech Communication*, vol. 52, no. 1, pp. 41–60, 2010.

[11] K. Pek, T. Arai, and N. Kanedera, "Voice activity detection in noise using modulation spectrum of speech: Investigation of speech frequency and modulation frequency ranges," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 33–44, 2012.

[12] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data.," in *ICASSP*, pp. 7229–7233, Citeseer, 2013.

[13] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[14] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 2, pp. 412–424, 2006.

[15] T. Kinnunen, E. Chernenko, M. Tuononen, P. Fränti, and H. Li, "Voice activity detection using mfcc features and support vector machine," in *Int. Conf. on*

Speech and Computer (SPECOM07), Moscow, Russia, vol. 2, pp. 556–561, 2007.

[16] Q.-H. Jo, J.-H. Chang, J. Shin, and N. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing*, vol. 3, no. 3, pp. 205–210, 2009.

[17] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Signal Processing, 2002 6th International Conference on*, vol. 2, pp. 1124–1127, IEEE, 2002.

[18] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 7378–7382, IEEE, 2013.

[19] T. V. Pham, C. T. Tang, and M. Stadtschnitzer, "Using artificial neural network for robust voice activity detection under adverse conditions," in *Computing and Communication Technologies, 2009. RIVF'09. International Conference on*, pp. 1–8, IEEE, 2009.

[20] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, p. 91, 2015.

[21] I. Almajai and B. Milner, "Using audio-visual features for robust voice activity detection in clean and noisy speech," in *Signal Processing Conference, 2008 16th European*, pp. 1–5, IEEE, 2008.

[22] 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.

[23] M. Van Segbroeck, A. Tsiartas, and S. Narayanan, "A robust frontend for vad: exploiting contextual, discriminative and spectral cues of human voice.," in *INTERSPEECH*, pp. 704–708, 2013.

[24] A.-R. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, pp. 4273–4276, IEEE, 2012.

[25] M. Espi, M. Fujimoto, K. Kinoshita, and T. Nakatani, "Exploiting spectro-temporal locality in deep learning based acoustic event detection," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 26, 2015.

[26] X.-L. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016.

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

[28] V. S. Mendelev, T. N. Prisyach, and A. A. Prudnikov, "Robust voice activity detection with deep maxout neural networks," *Modern Applied Science*, vol. 9, no. 8, p. 153, 2015.

[29] L. Deng, J. Li, J.-T. Huang, K. Yao, D. Yu, F. Seide, M. Seltzer, G. Zweig, X. He, J. Williams, *et al.*, "Recent advances in deep learning for speech research at microsoft," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 8604–8608, IEEE, 2013.

[30] L. Deng, "Dynamic speech models: theory, algorithms, and applications," *Synthesis Lectures on Speech and Audio Processing*, vol. 2, no. 1, pp. 1–118, 2006.

[31] K. Fujioka, N. Hayasaka, Y. Miyanaga, and N. Yoshida, "Noise reduction of speech signals by running spectrum filtering," *Systems and Computers in Japan*, vol. 37, no. 14, pp. 52–61, 2006.

[32] E. Verteletskaya and K. Sakhnov, "Voice activity detection for speech enhancement applications," *Acta Polytechnica*, vol. 50, no. 4, 2010.

[33] Y. S. Park and S. M. Lee, "Speech enhancement through voice activity detection using speech absence probability based on teager energy," *Journal of Central South University*, vol. 20, no. 2, pp. 424–432, 2013.

[34] K. Srinivasan and A. Gersho, "Voice activity detection for cellular networks," in *Speech Coding for Telecommunications, 1993. Proceedings., IEEE Workshop on*, pp. 85–86, IEEE, 1993.

[35] J. Ramírez, J. C. Segura, J. M. Górriz, and L. García, "Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2177–2189, 2007.

[36] P. Renevey and A. Drygajlo, "Entropy based voice activity detection in very noisy conditions," *threshold*, vol. 5, no. 5.5, p. 6, 2001.

[37] Y. Ma and A. Nishihara, "Efficient voice activity detection algorithm using long-term spectral flatness measure," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, no. 1, p. 87, 2013.

[38] P. K. Ghosh, A. Tsiartas, and S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[39] J. Haigh and J. Mason, "Robust voice activity detection using cepstral features," in *TENCON'93. Proceedings. Computer, Communication, Control and Power Engineering. 1993 IEEE Region 10 Conference on*, vol. 3, pp. 321–324, IEEE, 1993.

[40] T. Fukuda, O. Ichikawa, and M. Nishimura, "Phone-duration-dependent long-term dynamic features for a stochastic model-based voice activity detection.," in *INTERSPEECH*, pp. 1293–1296, 2008.

[41] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recog-

nition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2026–2038, 2011.

[42] T. Kristjansson, S. Deligne, and P. Olsen, "Voicing features for robust speech detection," *Entropy*, vol. 2, no. 2.5, p. 3, 2005.

[43] H. Ghaemmaghami, B. J. Baker, R. J. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," *Proceedings of Interspeech 2010*, 2010.

[44] K. Ishizuka and T. Nakatani, "Study of noise robust voice activity detection based on periodic component to aperiodic component ratio.," in *SAPA@ INTERSPEECH*, pp. 65–70, 2006.

[45] S. Basu, "A linked-hmm model for robust voicing and speech detection," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, vol. 1, pp. I–I, IEEE, 2003.

[46] K. R. Krishnamachari, R. E. Yantorno, D. S. Benincasa, and S. J. Wenndt, "Spectral autocorrelation ratio as a usability measure of speech segments under co-channel conditions," in *IEEE International Symposium Intelligent. Sig. Process. and Comm. Sys*, Citeseer, 2000.

[47] L. Atlas and S. A. Shamma, "Joint acoustic and modulation frequency," *EURASIP Journal on Advances in Signal Processing*, vol. 2003, no. 7, p. 310290, 2003.

[48] N. Kanedera, T. Arai, H. Hermansky, and M. Pavel, "On the relative importance of various components of the modulation spectrum for automatic speech recognition," *Speech Communication*, vol. 28, no. 1, pp. 43–55, 1999.

[49] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, pp. 1331–1334, IEEE, 1997.

[50] A. Shadevsky and A. Petrovsky, "Bio-inspired voice activity detector based on the human speech properties in the modulation domain," in *Information Processing and Security Systems*, pp. 43–54, Springer, 2005.

[51] N. Mesgarani, M. Slaney, and S. A. Shamma, "Discrimination of speech from nonspeech based on multiscale spectro-temporal modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 920–930, 2006.

[52] J.-H. Bach, B. Kollmeier, and J. Anemüller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pp. 41–44, IEEE, 2010.

[53] P. C. Khoa, *Noise robust voice activity detection*. PhD thesis, Citeseer, 2012.

[54] 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.

[55] F. Bie, Z. Zhang, D. Wang, and T. F. Zheng, "Dnn-based voice activity detection for speaker recognition," tech. rep., Tech. Rep, 2015.

[56] 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.

[57] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[58] T. Arai, M. Pavel, H. Hermansky, and C. Avendano, "Syllable intelligibility for temporally filtered lpc cepstral trajectories," *The Journal of the Acoustical Society of America*, vol. 105, no. 5, pp. 2783–2791, 1999.

[59] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*. Springer, 2014.

[60] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning.," in *AISTATS*, vol. 10, pp. 33–40, Citeseer, 2005.

[61] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *Journal of Machine Learning Research*, vol. 10, no. Jan, pp. 1–40, 2009.

[62] G. E. Hinton, "Learning multiple layers of representation," *Trends in cognitive sciences*, vol. 11, no. 10, pp. 428–434, 2007.

[63] M. A. Keyvanrad and M. M. Homayounpour, "A brief survey on deep belief networks and introducing a new object oriented toolbox (deebnet)," *arXiv preprint arXiv:1408.3264*, 2014.

[64] T. Kobayashi, "ASJ continuous speech corpus for research," *Acoustic Society of Japan (ASJ) Trans*, vol. 48, no. 12, pp. 888–893, 1992.

[65] The Rice University, "Noisex-92 database." http://spib.linse.ufsc.br/noise.html. Accessed: 2017-02-22.

[66] M. Myllymäki and T. Virtanen, "Voice activity detection in the presence of breathing noise using neural network and hidden markov model," in *Signal Processing Conference, 2008 16th European*, pp. 1–5, IEEE, 2008.

[67] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," *Computer Speech & Language*, vol. 29, no. 1, pp. 145–171, 2015.