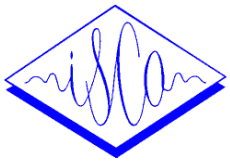

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

ROBUST SPEECH DETECTION WITH HETEROSCEDASTIC DISCRIMINANT ANALYSIS APPLIED TO THE TIME-FREQUENCY ENERGY

Ye Tian, Zuoying Wang, and Dajin Lu

Department of Electronics Engineering
Tsinghua University, Beijing
tianye@thsp.ee.tsinghua.edu.cn

ABSTRACT

In this paper, we propose a robust speech detection algorithm with Heteroscedastic Discriminant Analysis (HDA) applied to the Time-Frequency Energy (TFE). The TFE consists of the log energy in time domain, the log energy in the fixed band 250-3500 Hz, and the log Mel-scale frequency bands energy. The bottom-up algorithm with automatic threshold adjustment is used for accurate word boundary detection. Compared to the algorithms based on the energy in time domain [1], the ATF parameter [2], the energy and the LDA-MFCC parameter [3], the proposed algorithm shows better performance under different types of noise.

1. INTRODUCTION

A major cause of errors in automatic speech recognition system is the inaccurate detection of speech [1]-[4]. It is essential for robust speech recognition system that speech segments be reliably separated from noise segments. In this paper we address the problem of robust speech detection in quiet and in the presence of noise.

Parameters frequently used for speech detection such as the energy (in time domain), the zero-crossing rate, the pitch, and the linear prediction error energy show poor robustness to background noise. Even if complex decision strategies are used, they are not sufficient to get reliable speech detection in noisy environment [2]. For robust speech detection, the use of frequency energy is noticeable. Junqua *et al.* [4] proposed the Time-Frequency (TF) parameter, which is the sum of frequency energy in the fixed frequency band 250-3500 Hz and the time energy. Wu *et al.* [2] further propose an Adaptive Time-Frequency (ATF) parameter, which extends the TF parameter from single band to multi-band analysis. The ATF parameter can extract useful frequency information by adaptively choosing proper Mel-scale frequency bands which have the maximum word signal information. The frequency energy used in the ATF parameter is the sum of these useful bands energy. Experiments show the ATF parameter is better than the TF parameter for speech detection in noisy environment. However, the frequency bands energy are fusioned using binary weights in ATF, that is, each band is only classified as useful or useless. Since different types of noise have different frequency energy distribution, each frequency band should have different speech-noise discrimination ability. Thus different weight should be applied to different band, and optimum weights can further improve the performance.

Martin *et al.* [3] use Linear Discriminant Analysis (LDA) applied to the Mel-scale Frequency Cepstrum Coefficients (MFCC) for speech detection. MFCC are fusioned using a linear function calculated by LDA to get a parameter, referred as LDA-MFCC. The signal is classified as speech when both the energy and the LDA-MFCC parameter exceed corresponding threshold. The integration gives significant improvements for speech detection results especially in noisy environment. Nevertheless, Some problems should be further investigated.

1) MFCC is Discrete Cosine Transform (DCT) of the log frequency band energy. DCT compress the spectral information into low order coefficients. When LDA applied on MFCC to get a parameter for speech-noise discrimination, the DCT is harmful because it is not optimum for speech-noise discrimination and the reduction of dimensions may lose some useful information. Applying LDA directly on the log spectral band energy should be more reasonable for speech-noise discrimination. Moreover, MFCC has the energy normalization effect, that is, if the signal is linearly compressed or extended, MFCC keeps unchanged. This is useful for speech recognition, but it is not good for speech-noise discrimination. MFCC has little speech-noise discrimination ability when noise and speech have similar frequency bands energy distribution. According to above analysis, the frequency bands energy is better than MFCC when used for speech-noise discrimination.

2) In the speech detection algorithm, the energy and the LDA-MFCC parameter is used separately. Now that LDA has already provided a good framework to fusion all the features into one parameter to get maximum separation between speech and noise, the energy can be also integrated as one of the features for LDA.

3) LDA is known to be inappropriate for the case of classes with unequal sample covariance. There is no reason to assume that different types of noise have equal sample covariance with speech. In recent years, there has been an interest in generalizing LDA to Heteroscedastic Discriminant Analysis (HDA) [5] by removing the equal within-class covariance constraint.

4) The introductions of multiple thresholds can get more accurate speech detection than single threshold. We use the bottom-up algorithm with automatic threshold adjustment [1], in which speech pulses are located and edited, to make final word

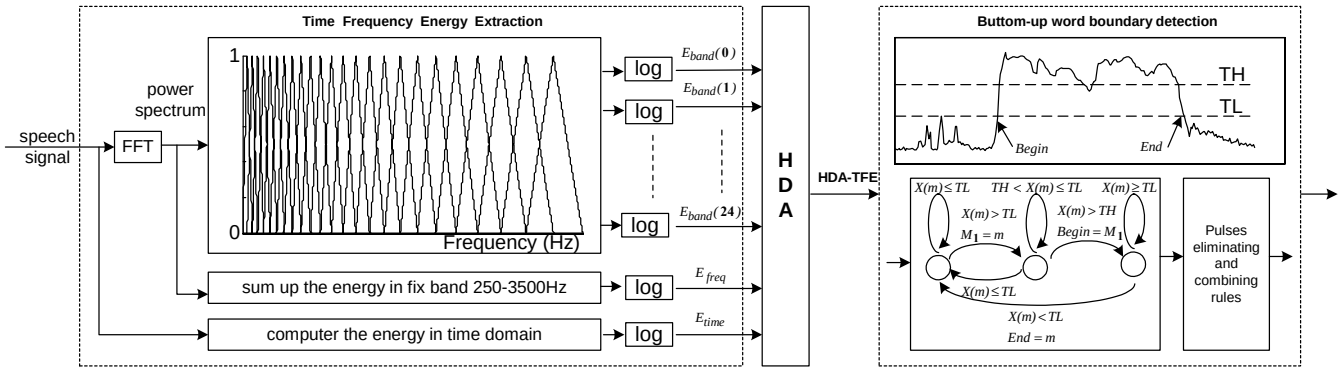

Fig.1. Flowchart of the proposed speech detection algorithm

boundary location.

To develop a robust speech detection algorithm and avoid the problems of the above approaches, this paper uses HDA applied to the Time-Frequency Energy (TFE), to get a new robust speech detection parameter, referred as HDA-TFE. The TFE consists of the log energy in time domain, the log energy in the fixed band 250-3500 Hz, and the log Mel-scale frequency bands energy. Based on the HDA-TFE parameter, the bottom-up algorithm with automatic threshold adjustment is proposed to get accurate word boundary detection. The proposed algorithm is compared to the algorithm based on the energy in time domain [1], the ATF parameter [2], the energy and the LDA-MFCC parameter [3] under different types of noise.

This paper is organized as follows. The speech detection algorithm is derived in Section 2 and the performance evaluation is proposed in Section 3. Finally, the conclusions are summarized in Section 4.

2. SPEECH DETECTION ALGORITHM

The proposed speech detection algorithm is illustrated in Fig.1. The algorithm consists of three main parts: time-frequency energy extraction, Heteroscedastic Discriminant Analysis, and bottom-up word boundary detection.

2.1 Time-Frequency Energy

The features used for the speech-noise classification contain both time and frequency energy.

Consider a given time-domain signal $x_{time}(m,n)$, representing the magnitude of the n th point of the m th frame. We calculate spectrum $x_{freq}(m,k)$, of this signal by K point Fast Fourier Transform (FFT).

The log energy in time domain is defined as

$$E_{time}(m) = \log\left(\frac{1}{N}\sum_{n=0}^{N} x_{time}^2(m,n)\right) \quad (1)$$

where N is the frame sample length.

The log energy in the fixed frequency band 250-3500 Hz is defined as

$$E_{freq}(m) = \log\left(\sum_{k=K1}^{K2} x_{freq}^2(m,k)\right) \quad (2)$$

where $K\mathbf{1}$ and $K\mathbf{2}$ is the start and the end index of the frequency band 250-3500 Hz ,respectively.

We then multiply the spectrum $x_{freq}(m,k)$ by the weighting factors $f(i,k)$ on the Mel-scale frequency bank and sum the products for all k to get the log energy $E_{band}(m,i)$ of each frequency band i of the m th frame,

$$E_{band}(m,i) = \log\left(\sum_{k=0}^{K-1} x_{freq}^2(m,k) f(i,k)\right) \quad (3)$$

The time-frequency energy of the m th frame is defined as,

$$\mathbf{E}(m) = [E_{time}(m), E_{freq}(m), E_{band}(m,\mathbf{1}),...., E_{band}(m,\mathbf{24})] \quad (4)$$

The energy in time domain, according to Parseval Law, is equal to non-band-limited frequency energy. The energy in the fixed frequency band 250-3500Hz is selected because of its usefulness for detecting high energy regions that correspond essentially to the vowel portions of the speech signal. The Mel-scale bands energy give more detail description of the energy distribution in frequency domain. We use HDA to determine their weights for speech detection under different types of noise.

2.2 Heteroscedastic Discriminant Analysis

speech detection in noisy environment is also the problem of speech and noise classification. The goal of LDA and HDA is to find a linear transformation $\mathbf{A}$ which maximizes the class discrimination in the projected subspace. LDA assume that all the classes have equal sample covariance. HDA presented by Saon *et al*. [5] generalizes LDA by removing the equal within-class covariance constraint. Linear transformation $\mathbf{A}$ is obtained by maximize the objective function $H(\mathbf{A})$ according to the training data,

$$H(\mathbf{A}) = \prod_{j=1}^{J}\left(\frac{\left|\mathbf{A}\mathbf{B}\mathbf{A}^T\right|}{\left|\mathbf{A}\boldsymbol{\Sigma}_j\mathbf{A}^T\right|}\right)^{K_j} \quad (5)$$

where J is the total class number, K_j is the sample number belonging to class j, $\mathbf{B}$ is the between-class scatter, and $\mathbf{\Sigma}_j$ is the covariance of class j. More about HDA can refer [5].

We project the time-frequency energy to one-dimension subspace to make speech-noise discrimination. The HDA-TFE parameter of the m th frame is defined as,

$$HDA-TFE(m) = \mathbf{A}\mathbf{E}^T(m) \quad (6)$$

,where $\mathbf{A}$ is an $\mathbf{1}\times\mathbf{26}$ rectangular matrix obtained by HDA, and $\mathbf{E}(m)$ is the time-frequency energy of the m th frame. The HDA-TFE parameter will be used for speech detection.

2.3 Bottom-up word boundary detection

We take advantage of the bottom-up algorithm with automatic threshold adjustment [1], in which speech pulses are located and edited, to make final speech location. The state representation of the operation of the speech detector is shown in Fig. 1. Two different energy thresholds TH and TL are used in our detection.

$$TH = \mu_n + (\mu_s - \mu_n)/N_1 \quad (5)$$

$$TL = \mu_n + (\mu_s - \mu_n)/N_2 \quad (6)$$

where μ_n and μ_s are the mean value of the HDA-TFE parameter of noise and speech signal, respectively.

After getting all potential speech pulses within the recording interval, duration information is used to edit speech pulses. Pulses do not exceed a minimum length L_1 are eliminate. If proximity of two pulses less than a length L_2, they are combined as one pulse.

3. EXPERIMENTS

The proposed algorithm (SD-HDA-TFE) is compared to the algorithm based on the energy in time domain (SD-TE) [1], the ATF parameter (SD-ATF) [2], the energy and the LDA-MFCC parameter (SD-TE-LDA-MFCC) [3] under different types of noise. The prefix "SD" of each algorithm is referred as "Speech Detection".

The noise signals are taken from the noise database provided by the NATO Research Study Group on Speech Processing (RSG.10) NOISE-ROM-0 [6]. We take four typical types of noise for speech contamination in our experiments. They are white noise, vehicle interior noise, multi-talker babble noise, and destroyer engine room noise.

The speech data used are continuous Mandarin digit strings database which contains 10 speakers and 50 digit strings for each speaker. We select the first 10 strings of each speaker, totally 100 strings, as the train data. The other is used as the test data. To set up the noisy speech database, we added the prepared noisy signals to the recorded speech signals.

The performances of the four algorithms are evaluated in terms of the Receiver Operating Characteristics (ROCs). We compare automatic speech segment detection to manual segmentation. At the test SNR, the LDA and HDA linear functions are trained on noisy speech train data. Fig. 2 shows the ROCs of the four speech detection algorithms under 0dB noisy environment and clean environment. The X-axis is the percentage of the false-rejected speech frame relative to the total

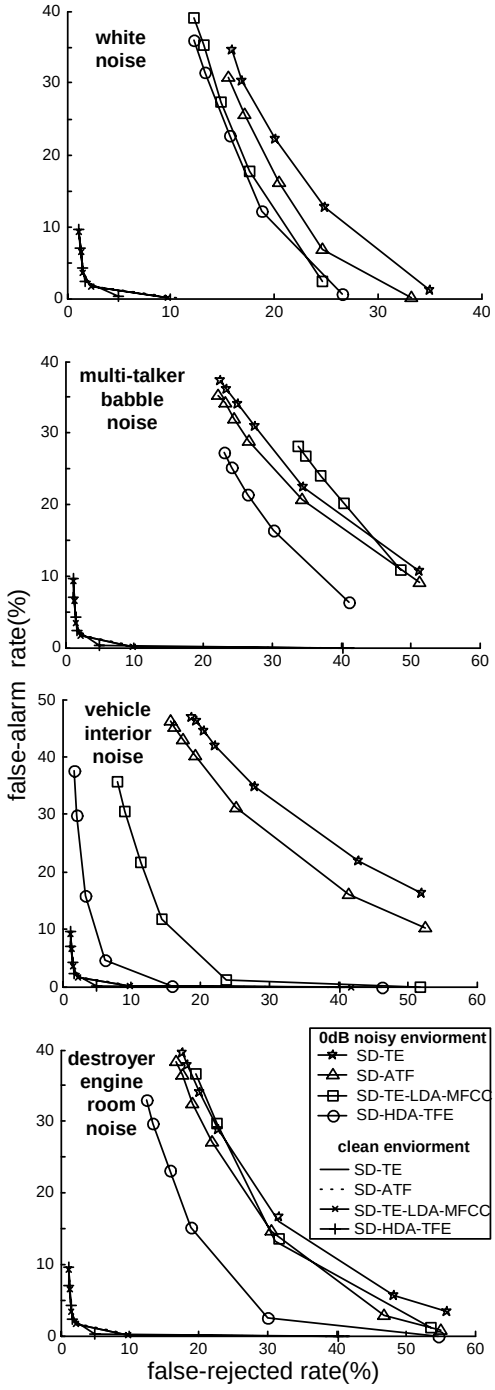

Fig.2. the ROCs of the four speech detection algorithms under clean environment and 0dB noisy environment
(The curves of the four detection algorithms under clean environment are in the left-bottom of each figure and overlap each other)

SD-TE: Speech detection based on the energy in time domain
SD-ATF: Speech detection based on the ATF parameter
SD-TE-LDA-MFCC: Speech detection based on the energy and the LDA-MFCC parameter
SD-HDA-TFE: Speech detection proposed in this paper

speech frame, and the Y-axis is the percentage of the false-alarmed noise frame relative to the total noise frame.

From the figure, we can see that all the four algorithms show almost the same speech detection performance under clean environment. But their performances are quite different under noisy environment. The SD-ATF algorithm has better performance than the SD-TE algorithm under all the types of noise because of adaptively choosing proper frequency bands energy to enhance the energy in time domain. This improvement shows the efficiency of the frequency bands energy for speech detection in noisy environment.

However, the SD-TE-LDA-MFCC algorithm shows unstable detection performance under different types of noise. Because the frequency bands energy are fusioned with optimum weights calculated by LDA, the SD-TE-LDA-MFCC algorithm shows better performance than the SD-ATF algorithm that uses binary weights under white noise and vehicle interior noise. Under white noise false-rejected speech frame relative to the total speech frame, and speech bands are contaminated by noise, thus the improvement is limited. The energy of vehicle interior noise is mostly focused on 0-1KHz, which is separated from speech frequency bands. Under this type of noise the speech bands is almost not contaminated by noise, thus the detection performance is significantly improved. Nevertheless, the performance of the SD-TE-LDA-MFCC algorithm is not satisfied under multi-talker babble noise and destroyer engine room noise. The reason is that the energy distribution of these two kinds of noise is similar to that of speech, and MFCC has no speech-noise discrimination ability because of the energy normalization effect. The most useful feature for speech-noise discrimination under this environment should be the frequency energy and the energy in time domain which are not normalized.

Under all types of noise, the SD-HDA-TFE algorithm shows the best performance of the four algorithms, and the improvement is quite significant. To analyze the improvement visually, the four detection of an utterance, "8 0 1 1", under vehicle interior noise at 0dB is plotted in Fig.3. From the figure, we can observe both the energy in time domain and the ATF parameter show little discrimination ability between noise and speech signal. It can be also seen that the ATF parameter makes the speech signal more obvious than the energy in time domain. The LDA-MFCC parameter shows better speech-noise discrimination, but the SD-TE-LDA-MFCC algorithm can not give satisfying speech detection results because the energy exceeds the energy-threshold is not well satisfied for speech segmentation. Compared to them, the HDA-TFE parameter has better speech-noise discrimination ability. Moreover, the HDA-TFE parameter is more consistent for noise and speech segments than the LDA-MFCC parameter. The improvement is due to the removing of the assumption of the equal covariance of noise and speech signals, and also, by non-normalized bands energy. Combined with bottom-up algorithm with automatic threshold adjustment, speech detection is correctly given even at 0dB.

4. CONCLUSIONS

Reliable speech detection is important to subsequent processing such as speech recognition and speech enhancement. In this paper, we use HDA applied to the time-frequency energy to make speech detection. Experiments show the proposed algorithm is more beneficial than other detection algorithms under different types of noise, especially when the energy of noise and speech are focused on different frequency bands.

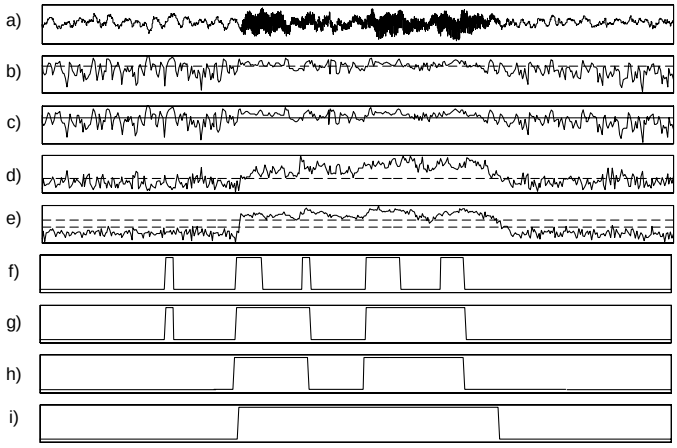

Fig.3. Detection of utterance "8 0 1 1" under vehicle interior noise at 0dB

(a) waveform (b) the energy in time domain (c) the ATF parameter (d) the LDA-MFCC parameter (e) the HDA-TFE parameter

(f),(g),(h),(I) detected speech pulses given by the *SD-TE* algorithm, the *SD-ATF* algorithm, the *SD-TE-LDA-MFCC* algorithm, the *SD-HDA-TFE* algorithm, respectively.

For real-time detection, HDA linear mapping function can be online trained to obtain an environment adaptation. The background noise can be estimated by the first several frames of the recording, or by two microphones so that the noise signal is recorded separately.

5. REFERENCE

[1] L. F. Lamel, L. R. Rabiner, A. E. Rosenberg, and J. G. Wilson, "An improved endpoint detector for isolated word recognition," *IEEE Trans. Acoustic, Speech and Signal Processing*, v29, pp. 777–785, Aug. 1981.

[2] G. D. Wu and C. T. Lin, "Speech detection with mel-Scale frequency bank in noisy environment". *IEEE Trans. Speech and Audio Processing*, v8, pp. 541-554, Sep 2000.

[3] A. Martin, D. Charlet, and L. Mauuary, "Robust speech/non-speech detection using LDA applied to MFCC", *Proceedings of ICASSP'2001*, v1, pp. 237-240, 2001.

[4] J. C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for speech detection in the presence of noise", *IEEE Trans. Speech and Audio Processing*, v2, pp. 406–412, July 1994.

[5] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen, "Maximum likelihood discriminant feature spaces", *Proceedings of ICASSP'2000*, v2, pp. 1129-1132, 2000.

[6] Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, v12, pp. 247–251, 1993.