A STUDY ON EFFICIENT ROBUST SPEECH RECOGNITION WITH STOCHASTIC DYNAMIC TIME WARPING

Sun Xihao

Graduate School of Information Science and Technology

Hokkaido University

Sapporo, Hokkaido, Japan

July 17, 2014

Abstract

In recent years, great progress has been made in automatic speech recognition (ASR) system. The hidden Markov model (HMM) and dynamic time warping (DTW) are the two main algorithms which have been widely applied to ASR system. Although, HMM technique achieves higher recognition accuracy in clear speech environment and noisy environment. It needs large-set of words and realizes the algorithm more complexly. Thus, more and more researchers have focused on DTW-based ASR system.

Dynamic time warping (DTW) is based on template matching,it can accomplish time alignment of reference and test speech features by dynamic programming. Conventional DTW is fast and less complexity, however its recognition accuracy is limited. Therefore, Conventional DTW has mostly been used for speech recognition in clear environment. Recently, a DTW with multireferences (mDTW) algorithm has also been developed to improve the recognition accuracy in comparison to the hidden Markov model (HMM) algorithm under noisy conditions. However the mDTW algorithm increases the calculation cost and requires more memory resources which reduce the system practicability.

It is possible to reconstruct the multireferences. The new method should be require less memory resources and reduce the calculation cost. Therefore, this study proposes a reconstruction method which add a training part to the DTW-based ASR system. The proposed reconstruction of references is aimed at making the DTW algorithm more effective. According to the DTW algorithm, the optimal warping path implies a minimum

error between any two given sequences. The algorithm that we have proposed will give us a way to build a new reference to replace the original two. This process will be done in three stages; First, for each reference word, speech utterances will be divided into two subsets. Second, for each pair of subsets, the optimal path will be computed and the new reference will replace the pair of subsets. Finally, the new references will be input to the DTW-based ASR system to get the recognition accuracy. The feasibility of the proposed technique was examined using computer simulations. The results demonstrated the effectiveness of the proposed technique. The simulation results show that our approach yields 96.94% accuracy compared with the 97.54% accuracy of mDTW in 20 dB white noise and 84.4% accuracy compared with 86.44% accuracy of mDTW in 10 dB white noise. Our approach yields 94.12% accuracy compared with 94.14% accuracy of mDTW in 20 dB babble noise and 80.82% accuracy compared with 81.64% accuracy of in 10 dB babble noise. Comparing our proposed technique to the mDTW, the calculation cost has been reduced 41.6%.

Acknowledgments

There are a lot of people I must recognize for their help. First, I would like to thank my supervisor, Prof. Yoshikazu Miyanaga, for his encouragement and support, for his faith in my work and for providing an excellent research environment. I also gratefully acknowledge Pro. Hiroshi Tsutsui from our laboratory. I always highly appreciated his guidance and his continuous support.

A very special thanks goes to Dr. Baiko Sai. The successful, learning great skills with him is the proof to me that the key to achieving the highest quality research in speech recognition area. The numerous discussion with him and his advice have been an invaluable contribution to my effort in further developing m understanding of signal processing algorithms and information theory.

My appreciation also goes to the Thesis Defense Committee members: Prof. Kunimasa Saitoh , Prof. Toshio Nojima and Prof. Yasutaka Ogawa from the Division of Media Network, Graduate School of Information Science and Technology (IST), Hokkaido University.

I am also extremely grateful to the present and past members of the Information Communication Network Laboratory (ICNL). They have been my companions for many years now and working with them convinced me that research flourished best in such an excellent team acting in concert to achieve common goals.

From a personal perspective, I would like to express my gratitude to my parents and

wife. They are the best parents I can imagine and I am very grateful to them for teaching me never to be satisfied with an achievement and to continue to strive for more. I also want to express my heartfelt gratitude to my wife for all her patience, love and support. The time with her gives me much of the strength for my work.

Contents

List of Figures

List of Acronyms

ASR	Automatic Speech Recognition
CMS	Cepstrum Mean Subtraction
CMVN	Cepstral Mean-variance Normalization
DC	Direct Component
DCT	Discrete Cosine Transformation
DFT	Discrete Fourier Transform
DP	Dynamic Programming
DRA	Dynamic Rang Adjustment
DTW	Dynamic Time Warping
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
HMM	Hidden Markov Model
IIR	Infinite Impulse Response
LPCC	Linear Predictive Coefficients
MFCC	Mel-Freqency Cepstrum Coefficients
PLP	Perceptual Linear Predictive
RASTA	RelAtve SpecTrAl
RSF	Running Spectrum Filtering
SNR	Signal-to-Noise Ratio

SS	Spectral Subtraction
VAD	Voice Activity Detection
ZCR	Zero-Crossing Rate

Chapter 1

Introduction

1.1 Background

Speech is the primary means of communication between people. For reasons ranging from technological curiosity about the mechanisms for mechanical realization of human speech capabilities to the desire to automate simple tasks inherently requiring human-machine interactions, research in automatic speech recognition (and speech synthesis) by machine has attracted a great deal of attention over the past five decades [50]. With the development of technology, machines can be competent for many works instead of human. Thus, it is our holy grail to make machines understanding human's speeches and able to communicate with human by speech recognition technology. The researched motivation of automatic speech recognition (ASR) is to transform human's tongue signals to texts or commands. It means machines can convert speeches of phonemes, words or sentences into messages, and then the messages are achieved some texts by message comprehension. In other word, the machine obtains the human's commands and makes an appropriate response by message comprehension. As for an interdisciplinary subject, the speech recognition is involved with computer, acoustics, phonetics, signal process-

ing, artificial intelligence, mathematics statistics, psychology, etc.

With the developments of computer, acoustic, signal processing and pattern recognition, Speech recognition has been made great strides recently. It is applied widely to many fields (i.e., industry, military, communication, medical, self-server, office automatic, etc.). In the industry field, speech recognition is applied to quality control and checking, acoustic control of numerically controlled lathe, etc [32, 56, 91]. In the military field, speech recognition is applied to flight vehicle control system, operational command and training of air traffic control [90]. In communication field, speech recognition is applied to use voice activation to make some calls and an interactive voice response system that offers automated employee benefit information on demand [72]. In medical field, speech recognition is applied to special utensils and other aids for disabled persons [11]. In the consumer electronics field, speech recognition is applied to produces of mobile terminal, car autonavigator, domestic robot, etc [10, 14, 17, 59, 94]. Moreover, the more applications include information inquiry, ticket reservation, audio retrieval, dictating machine, automatic translation, etc. As an more conveniently and efficiently man-machine interactive mode, the speech recognition is close to our everyday lives. It has many significant influence on our lifestyles.

Nowadays, speech recognition can obtain an very excellent performance in the ideal environment of laboratory. However, the performance of speech recognition drops rapidly in the noisy environment. The reasons are the variance of speech in the transmission and distort of speech in surrounding noisy environment. Furthermore, characteristics are difference from different speakers, i.e., age, spirit, sex, dialog, etc. This sound spectrum can be significant changed from different speakers. Even if the same speaker's sound characteristic can be difference under difference time or spirit. In addition, reference data applied to pattern matching may also not able to cover all the

human sound characteristics. All the factors above are considered as major obstacles for speech recognition when applying to actual practice. Hence, it is important to improve the performance of speech recognition in noisy environment.

1.2 Classification of speech recognition

(1) Classification according to total number of vocabulary

- Small vocabulary speech recognition: The total number of recognized word is usually between 1 and 100.
- Medium vocabulary speech recognition: The total number of recognized word is usually between 100 and 1000.
- Large vocabulary speech recognition: The total of number recognized word is usually more than 1000.

Because the total number of recognized speech is small, the feature difference of all words is large. Thus, recognition accuracy for ASR based on small vocabulary is high. On the contrary, for the large vocabulary, the recognition accuracy is low. Because the feature differences of all words are small. Moreover, in order to support faster and higher performance hardware requirement, recognition time should also be controlled. However, the classified circumscription is not changeless. The given circumscription is only a reference number, but the order of quantity is usually same.

(2) Classification according to recognized unit

- Word based speech recognition: The word is used as a recognized phonetic unit in speech recognition. All aforehand speeches of the words must be preprocessed to reference patterns or made as the training models. Hence, the word based

model only recognizes the word, which is exist in the reference patterns or training models. If a new word needs to be recognized, then its reference speeches must be added into the recognition model beforehand. With the increase of recognize word, the calculation cost and time are increase for recognition. However, the recognition accuracy is high.

- Phoneme based speech recognition: The syllable and phoneme is used as a recognized phonetic unit in speech recognition. Firstly, the speech is recognized as a sequence of phoneme. Then, all phonemes are combined to some words, phrases or sentences according to rules of syntax of spoken language. In theory, all phoneme phonetic units in spoken language are preprocessed to reference patterns or made the training models. Next, the phoneme based speech recognition system can recognize all words, phrases and sentences.

(3) Classification according to recognized object

- Isolated word speech recognition: The speech recognition system recognizes the speech into a word, or a set of speech segments by labeling or halting, which can be recognized into a set of word.

- Continues speech recognition: The speech recognition system recognizes natural and fluent continues speech into some words, phrases or sentences correctly. The continuous speech recognition system is most complex. However, it is ultimate object of speech recognition research.

(4) Classification according to difference speaker

- Speech recognition for specific speaker: The speech recognition system only can recognize specific speaker's speeches. The system is simple and recognition accuracy is high. However, it is necessary to obtain plenty of reference speeches of

the speaker beforehand.

- Speech recognition for unspecific speaker: A few people's standard speeches are used to make reference pattern or train the learning model. Moreover, the system can recognize all people's speeches, and has excellent versatility and wide application. However, such system is difficult to apply to practice with its low recognition accuracy.

(5) Others classification

- Speaker recognition system: The processing does not recognize the word or semantics of speech, but it can recognize the speaker who said the speech. Thus, the system can be used as identification. Some security access control systems apply the speaker recognition system to identify the visitors by their speeches, and give corresponding permissions.

1.3 Motivation

Dynamic time warping (DTW) is a popular automatic speech recognition (ASR) method based on template matching [37, 81]. DTW can accomplish time alignment of reference and test speech features by dynamic programming. Conventional DTW has fast search and low complexity, but it has poor speech recognition accuracy. Therefore, DTW has mostly been used for speech recognition in clean speech environments [6, 42, 51, 77, 99]. Recently, a DTW with multireferences (mDTW) algorithm has also been developed to improve the recognition accuracy under noisy conditions. However, the mDTW algorithm increases the calculation cost. Therefore, in this thesis, our motivation is to develop a DTW-based ASR system with training part to reduce the calculation cost. Unlike a conventional DTW or mDTW, we employ an appropriate reference utterances

to replace the original utterances. We attempt to improve the performance of the DTW-based speech recognition approach. First, we improve the short time energy algorithm. The new proposed approach is easily represent the smoothness properties between adjacent frames, substantially decreases the effect of pulse-noise. The endpoint detection accuracy is increased. Then, we propose the union of running spectrum filter (RSF), cepstrum mean substraction (CMS), and dynamic range adjustment (DRA) to reduce noise. The recognition accuracy is better than that of RSF, as well as calculation cost is lower than that of RSF. Last, we propose the DTW with training part is used to recognize. Compare with the mDTW, the recognition accuracy is almost same. However, the calculation cost have been reduced significantly.

1.4 Thesis Overview

Chapter 1 the background of automatic speech recognition (ASR) systems has been introduced. Current ASR recognizes ether the small set of words and phrases or the large vocabulary of speech sentences. For each task, a suitable ASR has been developed and improved recently. In this doctor thesis, dynamic time warping (DTW) has been explored and modified suitable for an efficient robust speech recognition system.

Chapter 2 introduces the basic technologies used into ASR. The speech features are extracted by speech analysis methods and they are used for speech recognition. Normally speech features are disturbed by various noises and thus its noise components should be reduced by using noise robust technologies. After that, noise robust speech features are estimated and used for speech recognition. As commonly used speech clustering technologies, DTW and hidden Markov model (HMM) have been already developed. In this chapter, the overview of these technologies have been explained.

Chapter 3 the importance of automatic voice activity detection (VAD) has been discussed. In particular, under noise circumstances, it has been quite difficult to design the automatic voice activity detection with a speech recognition system. The basic concept about VAD and its current techniques have been discussed in this chapter.

Chapter 4 introduces current noise reduction technologies used into speech processing. Among them, CMS, and RSF/DRA are explained in this chapter.

Chapter 5 introduces conventional DTW methods. Some DTW methods have been developed and applied into several real applications. However, they have somewhat weak against speaker independent mechanism and various noises. Some of issues in the conventional DTW have been discussed in this chapter.

Chapter 6 has proposed new techniques using DTW, VAD, CMS and RSF/DRA. It can realizes noise robust mechanism, robust automatic VAD and high speech recognition accuracy. In addition, the proposed method can reduce the total calculation cost drastically compared with other methods whose recognition accuracy is almost the same.

Chapter 7 summaries the above research and give a conclusion to highlight the research significance. Finally, we briefly describe some possible work for future research.

Chapter 2

Fundamentals of speech recognition

2.1 Situation of speech recognition

Fig. 2.1 shows a diagram of an ASR system that comprises modules for voice activity detection (VAD), feature extraction, noise reduction, and speech recognition [16, 36, 45, 46, 53, 62, 70] The unknown speech waveform is sampled, processed by these blocks, and compared with known waveforms to make a recognition decision. The blocks shown in this figure are discussed below and throughout the paper.

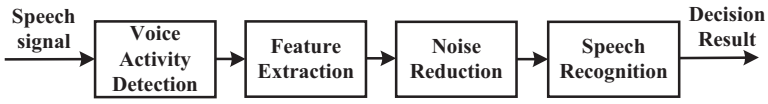


Figure 2.1. ASR system diagram

2.2 Feature extraction of speech signal

The feature vector is extracted form original speech signal at front-end processing of ASR system, which is easy to build model and recognize. The parameter of feature

vector is very important to improve the recognition accuracy. Usually, the differences of feature among speeches of same word should be as small as possible. On the contrary, that among differences of feature of different words should be as big as possible. Moreover, in order to reduce storage space, recognition cost and time, the number of dimension of feature vector should be as small as possible upon keeping the higher accuracy.

Since 1980s, the cepstrum parameter is widely to ASR. It includes linear predictive coefficients (LPCC) [4,5,38,49], Mel-frequency cepstrum coefficients (MFCC) [12] and perceptual linear predictive (PLP) [28,31]. The MFCC is most popular in ASR, because the MFCC better expresses the mechanism of human's ear. By analyzing the spectrum of speeches, we can obtain the better accuracy and robust. The Mel frequency better describes the nonlinear relation that human's ear feels the frequency of speech signal. The equation that linear frequency is converted to Mel frequency is

$$f_{mel} = 2595 \log_{10}(1 + \frac{f_{linear}}{700}) \tag{2.1}$$

where f_{mel} is Mel frequency and f_{linear} is real linear frequency. In Mel frequency domain, the perception of hearing is symmetrical for frequency. For different frequencies, the speech signal in corresponding critical-band can make the basilar membrane to vibrate. When the bandwidth of frequency is more than the critical-band, we can not perceive the signal. By Zwicker's research [104], the change of critical-band is same to that of Mel frequency. Under 1000 Hz, the Mel frequency is linear distribution, and it is logarithm distribution above 1000 Hz. This is also shown in Fig. 2.2. So a set of bandpass filters can be used to imitate hearing, thus, reducing the influence of noisy circumstance. According to the different critical-band, the frequency of speech signal is divided into a set of trilateral bandpass filters (Mel filter-banks). The weighted sums of all amplitudes of signals in the same critical-band is as the output of a trilateral bandpass

filter, and then a vector is obtained from all outputs by logarithm computation. Finally, the vector is transformed to MFCC parameter by discrete cosine transform (DCT).

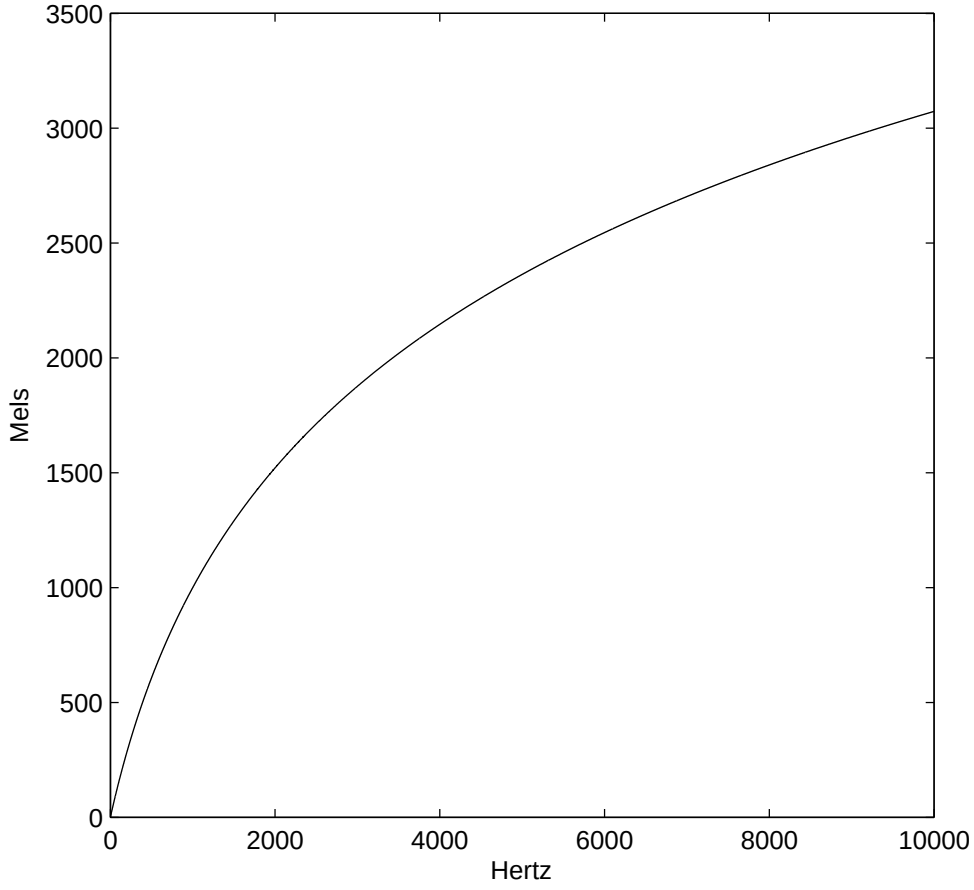


Figure 2.2. The relation between Mel frequency and linear frequency

Fig. 2.3 shows a block diagram of the MFCC processor for speech recognition. The basic steps in the processing include the following:

(1) Preemphasis

The digitized speech signal, $s(n)$, is through a first-order finite impulse response (FIR) filter, it is put into spectrally flatten signal and made less susceptible to finite precision effects later in the signal processing. The fixed first-order system is

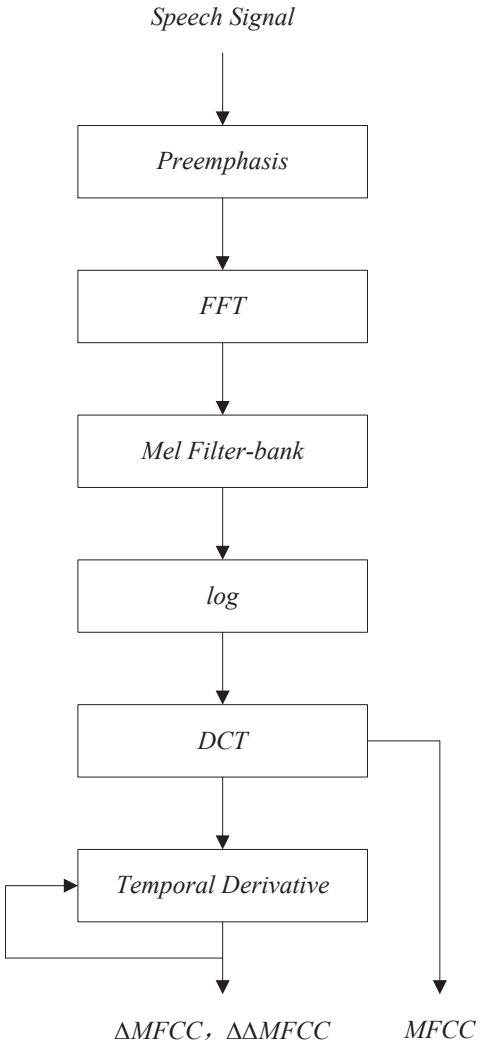


Figure 2.3. Block diagram of MFCC processor for speech recognition

$$H(z) = 1 - 0.97z^{-1} \tag{2.2}$$

In the case, the output of the preemphasis, $s'(n)$, is related to the input to the network, $s(n)$, by the difference equation

$$s'(n) = s(n) - 0.97s(n-1) \tag{2.3}$$

(2) Windowing

The next step in the processing is to window each individual frame. If we define the window as $w(n)$, $0 \leqslant n \leqslant N-1$, then the result of Hamming window, which has the form

$$w(n) = 0.54 - 0.46cos(\frac{2n\pi}{N-1}) \tag{2.4}$$

$$s_w(n) = s'(n)w(n) \tag{2.5}$$

$s_w(n)$ is the signal after windowing.

(3) Fast Fourier transform (FFT)

$s_w(n)$ is transformed to spectrum coefficient by FFT:

$$S(k) = \left| \sum_{n=0}^{N-1} s_w(n) e^{-j\frac{2\pi kn}{N}} \right|, \quad 0 \leqslant k \leqslant N-1 \tag{2.6}$$

(4) Mel fiter-banks

$S(k)$ is filtered with Mel filter-banks and the logarithm energy $X(m)$ is obtained.

$$X(m) = \ln \left(\sum_{m=0}^{N-1} S(k) H_m(k) \right), \quad 1 \leqslant m \leqslant M \tag{2.7}$$

where m is the number of filter, $H_m(k)$ is the weighted factor of the m^{th} filter in the frequency K and $X(m)$ is the output of m^{th} filter.

(5) Discrete Fourier transform (DFT)

The MFCC coefficients $c(l)$ are obtained with DFT.

$$c(l) = \sqrt{\frac{2}{M}} \sum_{m=1}^{M} X(m) \cos \frac{\pi(2m+1)l}{2M}, \quad 0 \leqslant l \leqslant L-1 \tag{2.8}$$

where L is the total of dimension of MFCC vector.

(6) Temporal derivative

the first order difference $\triangle\, c(l)$ and second order difference $\triangle\triangle\, c(l)$ coefficients can be obtained in time by the functions:

$$\triangle\, c(l) = \frac{\sum_{\sigma=-\Phi}^{\Phi} \sigma c(l+\sigma)}{\sum_{\sigma=-\Psi}^{\Psi} \sigma^2} \tag{2.9}$$

$$\triangle\triangle\, c(l) = \frac{\sum_{\sigma=-\Psi}^{\Psi} \sigma c(l+\sigma)}{\sum_{\sigma=-\Psi}^{\Psi} \sigma^2} \tag{2.10}$$

where Φ and Ψ are the number of frame that is used to compute the difference at both front and back. The $c(l)$, $\triangle\, c(l)$ and $\triangle\triangle\, c(l)$ are spliced to MFCC feature vector.

2.3 Pattern comparison techniques

A key question of speech recognition is how speech patterns is compared to determine their similarity. According to the specifics of recognition system, pattern comparison can be done in a wide variety of ways [9, 64]. Usually the early speech recognition system uses the pattern comparison to identify. In the training, all template parameters are extracted from every speech unit by the feature vector sequences of training. In the recognition, the testing speech feature vector is compared with the all pattern parameters. The speech unit is the result, which similarity is highest. Because of the speech signals are random, the length of time that the some utterances for one word are pronounced by the same people are difference. Thus, the utterances must be flexed to same length of time before pattern comparison. Firstly, researchers align the speech parameters into time with linear flexing method. As all testing speech signals are flexed

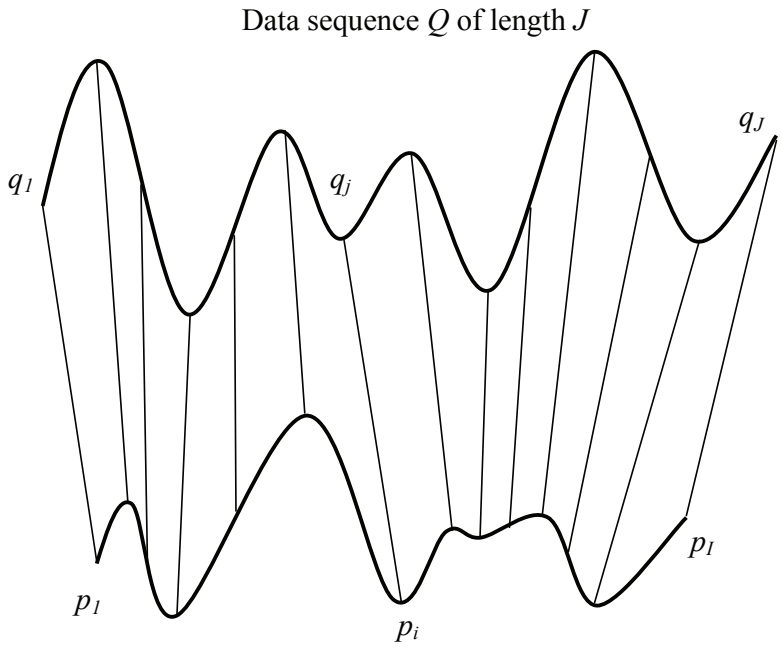
to length of the reference template. However, the utterance is nonlinear flexing. The consonants and the transition segments from consonant to vowel keep the fixed lengths and their changes are less. But the flexing of vowel segments are large. Thus, the linear flexing method can not be aligned so accurately and the result is unsatisfactory. Hence, the more advanced Pattern comparison techniques are proposed.

2.3.1 Dynamic time warping method

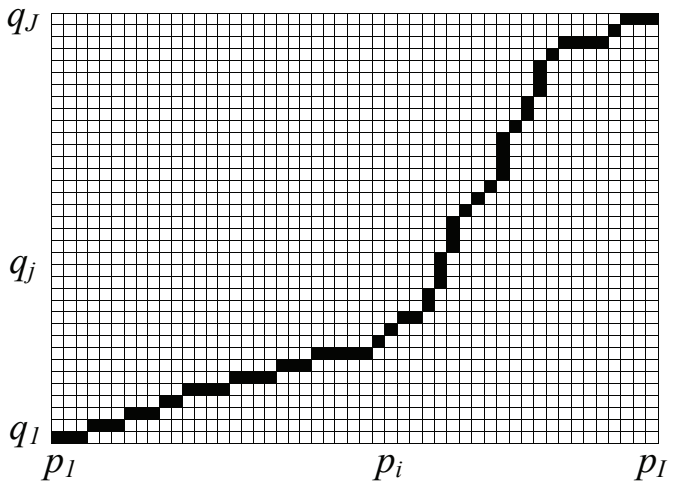
The dynamic programming (DP) can solve the problem of difference speaking velocity. Dynamic time warping (DTW) algorithm was proposed with DP by Sakoe [81], Vintsyuk [86], et al.. The DTW algorithm is nonlinear time alignment technology that combines the time alignment with distance computing technology. The DTW separates a problem of complex global optimization into some simple problems of local optimization. It calculates step by step and finds out the optimal matching path between the testing pattern and reference pattern. Fig. 2.4 shows the illustration of DTW algorithm. The DTW algorithm overcomes the problem that speaking speed is nonuniform and improves the performance of ASR system. The recognition accuracy of speech recognition for small vocabulary is very high by DTW. But the DTW algorithm is fit to that the recognition unit is word, prase or the whole sentence. For the large vocabulary, the DTW algorithm is difficult to apply, because the calculation cost is large [6, 42, 48, 66, 77].

2.3.2 Hidden Markov Model method

The hidden Markov model (HMM) [35, 40, 54, 67] is that the speech signal can be well characterized as a double parametric random processes. One is used to describe the statistical method of characterizing the spectral properties of the short-time nonstationary signal (or instantaneous character of signals), the other is used to describe the process



(a) The alignment of measurements by DTW for measuring the distance between two sequences



(b) DTW obtains a mapping between the sequences. The black squares denote the optimum warping path

Figure 2.4. Illustration of DTW

how a short-time stationary signal is made the transition to next short-time stationary signal, as well as dynamic character of the speech signal. Based on the double random processes, HMM approach can identify the short-time stationary speech signals of difference parameter. It also can follow the process of transition between these speech signals.

The human's process of speech also is a double stochastic processes. The speech signal is a observable sequence. It is the parameters sequences that the brain makes it to phonemes, words or sentences by the grammar and human's minds. Thus the parameters sequences is unobservable. Many experiments have shown the HMM approach can describe the processing of phonation of speech signal very accurately.

All parameters of the HMM are defined as follow.

(1) N is the number of states in the model. although the states are hidden, for many practical applications is often some physical significance attached to the sates or to sets of states of the model. The individual states are labeled as $\{1,2,\ldots,N\}$, q_i is the state at time t.

(2) M is the number of distinct observation symbols in the per state. The observation symbols are denoted as $V = \{v_1, v2, \ldots, v_M\}$. The observation sequence is denoted as $O = \{o_1, o_2, \ldots, o_T\}$. T is the size of observation sequence.

(3) The state-transition probability distribution $A = \{a_{ij}\}$ where

$$a_{ij} = P[q_{i+1} = j | q_i = i] \quad 1 \leqslant i \leqslant N, 1 \leqslant i \leqslant N \tag{2.11}$$

$$\sum_{j=1}^{N} a_{ij} = 1 \tag{2.12}$$

(4) the observation symbol probability distribution $B = \{b_j(k)\}$, in which

$$b_j(k) = P[o_t = v_k | q_i = j] \quad 1 \leqslant K \leqslant M, 1 \leqslant j \leqslant N \tag{2.13}$$

(5) The initial state distribution $\pi = \{\pi_i\}$ in which

$$\pi_i = P[q_1 = i] \quad 1 \leqslant i \leqslant N \tag{2.14}$$

An HMM can be described with specification of two model parameters N and M, specification of observation symbols, and the specification of the three sets of probability measures A, B, and π. For convenience, we use the compact notation

$$\lambda = (A, B, \pi) \tag{2.15}$$

With the time is changed, the states can be transferred each other, it is possible to the same states. Every observation sequence has corresponding state-transition probabilities for different states. Fig. 2.5 shows an HMM with four states $\{S_1, \cdots, S_4\}$. The state-transition is a_{ij} between all states. Each observation sequence is $\{o_1, o_2, \cdots, o_T\}$. The observation sequence is MFCC feature vector of speech signal.

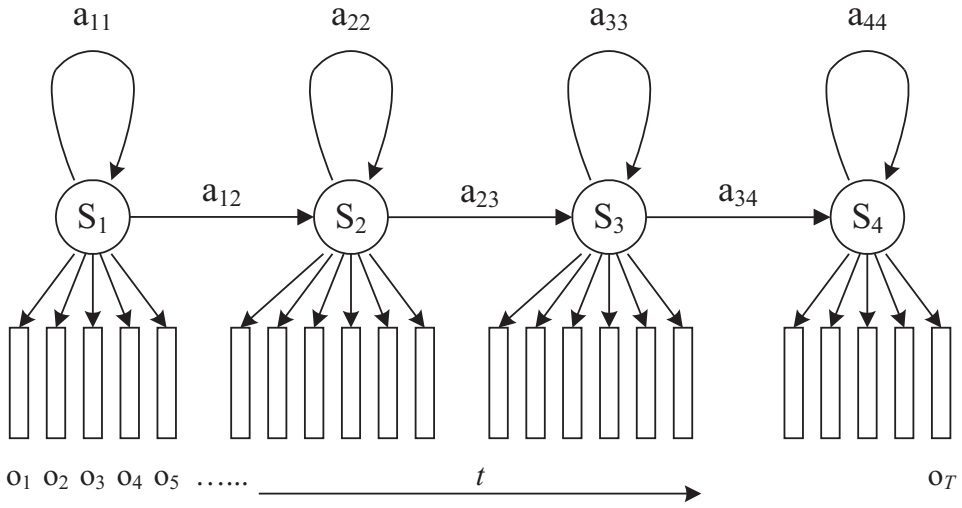


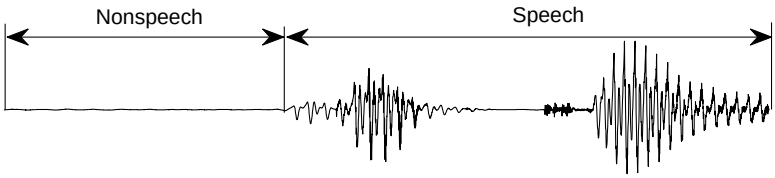
Figure 2.5. The relation between HMM chain and parameters of speech

2.4 Voice activity detection techniques

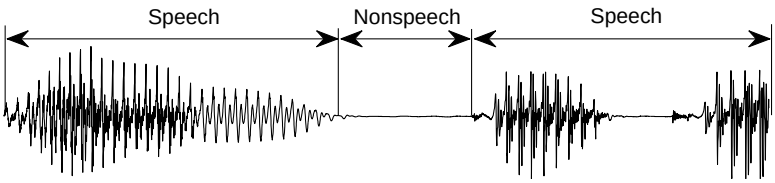
In the speech signal processing, the voice activity detection (VAD) technique [2, 34, 55, 92, 93] is important. The VAD can distinguish the speech segments and nonspeech segments from the input of digital speech signal, moreover, it can determine the start-point and end-point of speech signal accurately.

In the isolated word speech recognition system and continuous speech recognition system, the efficient VAD is important for improving the recognition accuracy and reducing the time of processing. In the noise reduction, the VAD is also important. For example, cepstrum mean subtraction (CMS) [24]. In order to compute the mean of energies of all speech frames, CMS must detect the endpoints of speech segment, in order to reduce the distortion of transmission channel and improve the robustness of recognition. Furthermore, when the silent segments are taken out beforehand, the estimation of energy of speech is more closer to real speech segments rather than the silent segments are influenced by the noise in silent or nonspeech segments. Moreover, it is good for creating the silent model and noise model that the nonspeech segments are taken out from the speech signal. Obviously it can decrease the collected digital data from the analog speech signal that the starting-point and end-point are detected accurately and the background noises segment without the speech. Thus, it can decrease the computation cost and processing time in speech processing systems. In the variable bit rate speech coding, the bit rate of silent segments can be reduced under the quality of received speech signal is kept the same. In order to decrease the transmitting power and economize the resources of channel, the mobile terminal usually uses the the variable bit rate speech coding. If the speech signals are nothing in the channel, it will reduce the bit rate. Whereas, it will raise the bit rate.

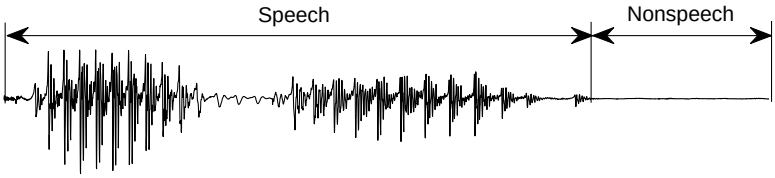
In the robust speech recognition, the intentions of VAD are the follows.



(a) Nonspeech segment at the start of a word



(b) Nonspeech segment at the median of a word



(c) Nonspeech segment at the end of a word

Figure 2.6. The nonspeech segments of a word

- Detecting the speech frame and background noise from the signal of speech frame. The VAD can affect the performance of ASR. If the speech segment is recognized to noise, then some important speech data are lost and the recognition accuracy is decreased. If the noise segment is recognized to speech, then calculation cost and the error probability of comparison with reference patterns will be raised, the recognition accuracy is also decreased.

- Dividing the sentence. For the continuous speech recognition system, the sentences are divided into the recognition unit (syllable, phoneme, word or phrase, e.g.) by VAD. For the man-machine interactive processing system, the system can respond to user by the every sentence. If the whole sentence is detected in error, the response of system may be mistake. If the system know the end of a sentence, then it do not respond the request.

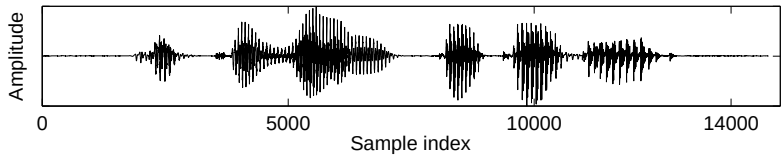
- Some speech recognition algorithms need estimate the spectrum characteristics of noise. The spectral subtraction (SS), e.g., the the spectrum characteristics of noise is estimated with the detected noise.

- Reducing the calculation cost. The calculation cost is important to low performance hardware, mobil device or embedded system. The VAD can take out the nonspeech segments and reduce the speech coding, then the ASR system can improve the recognition performance and time.

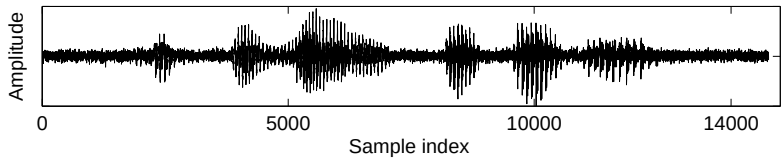
2.5 Noise reduction technique

In the early researches of speech recognition, the standard speech databases are recorded on the quiet circumstances. Thus, the better recognition accuracy can be gotten with

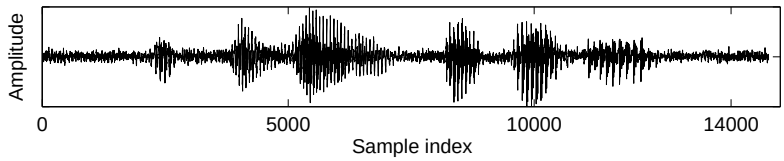
the recognition system that speeches are trained or created to reference models on the quiet circumstances. As the application of speech recognition system, the recognition environment is more complexity. Under real noise environment, the recognition performance is drastic lowering because the feature vectors are discrepant between the noisy speeches and the reference models, which are created under the quiet circumstances. [20, 44, 76]. Fig. 2.7 shows the waveforms of clean speech with white and babble noises. Fig. 2.8 shows the 3^{th} dimension feature vector of MFCC for the three waveforms. It shows the feature vectors of speech are so distorted by the noises.



(a) The waveform of clean speech



(b) The waveform of speech with 10 dB white noise



(c) The waveform of speech with 10 dB babble noise

Figure 2.7. The waveforms of speech with white and babble noises

The robust noisy speech recognition has been a research focus in the last twenty years, the researches proposed many ways and tried to improve the performance of ASR system. But any perfect solution has not been proposed for robust ASR system. The major influences are the follow.

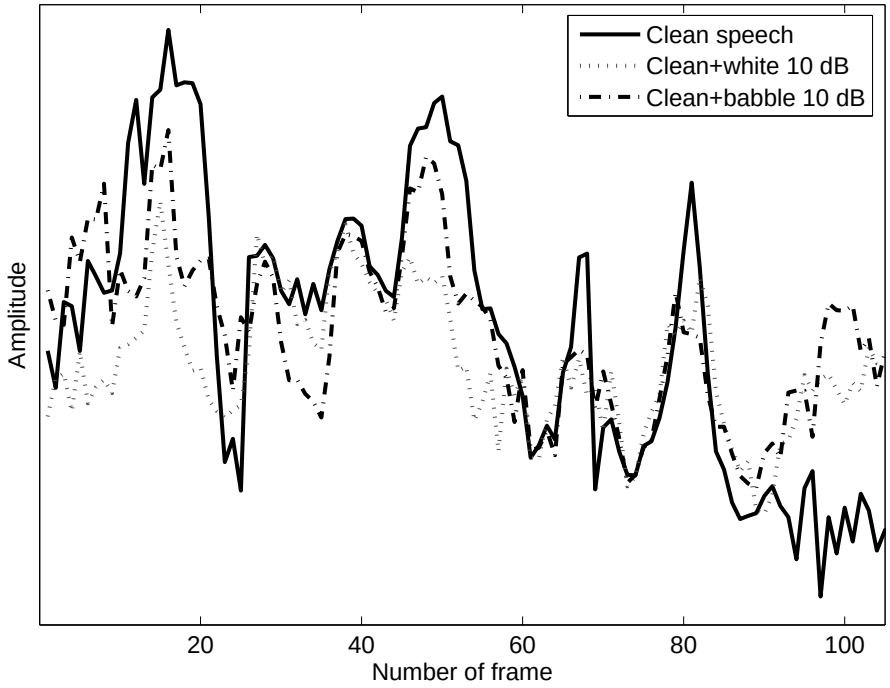


Figure 2.8. The 3^{th} dimension feature vector of MFCC of clean speech with 10 dB white and babble noises

- The influence of double articulation. The acoustic feature of speech signal is closely related with the pronounce. The acoustic features of speech signal may be made a great deal of different in different contexts, characterizes some language constructions. Moreover, two same utterances may express the different meanings.

- The influence of language complexity. The meaning of a sentence is closely related with the contexts and cultural background. Furthermore, the structure of sentence is variation in language grammar. But it is very difficult that the information of context are applied to ASR.

- The influence of variation of pronunciation for speaker himself. For the factors of

age, sentiment, health condition, speaking speed and so on, the acoustic features are different between utterances of same word.

- The influence of utterances for different speakers. The utterances between different speakers are big difference, because their vocal cords are difference.

- The influence of ambient environment. The speech signal can be distorted easily by the noise, reverberation, microphone, transmission channel and so on.

The noise reduction technique can reduce the noise and extract the real speech from the noisy speech. It tries to increase the acoustic feature of real speech signal possibly, in order to improve the recognition accuracy of ASR system.

Chapter 3

Voice Activity Detection

3.1 Introduce

The human's speech is discontinuous. Thus, the ASR system begins to work when speech is detected. Usually, only the VAD programming runs in order to reduce the calculation cost of ASR system, when speech signal is nothing. Furthermore, the endpoints of speech are accurately detected is important to improve the recognition accuracy of ASR system. Thus, VAD is a very important technique problem, especially in high ambient noise environments. The accurate endpoint detection of speech is a simple problem in the most benign circumstances. In practice, one or more problems usually make accurate VAD difficult in the noisy background (e.g., fans or machinery running). In nonstationary environments (e.g., the presence of door slams, irregular road noise, car horns) with speech interference (as from TV, radio). Other factors are that the distortion introduced by the transmission system when the speech is sent, (e.g., cross-talk, intermodulation distortion, and various types of tonal interference arise to various degrees in the communications channel) [41]. Many VAD methods have been proposed in speech recognition systems. VAD algorithm typically relies on the short-time energy

and zero-pass ratio [15,92]. The associated techniques use different features of syllables in the time-domain and are low computational complexity. In the chapter, we introduce these ways of VAD and propose modified VAD algorithm.

3.2 Short-time energy algorithm

Since the speech signal is a nonstationary processing, the way can not been used to process speech signal, which is used to process stationary signal. The produced processing of speech signal is closely-related with physical working of phonatory organ. This physical working is slower than vibrations of sounds. The speech signal in $10 \sim 30$ ms time can be as a quasi-steady signal (as short-time steady state), because the parameters of spectrum and physical characteristics are almost invariant [70,73]. Thus, a speech signal can be divided into many short frames and every frame is as a detecting unit. According to the energies of speech and nonspeech frame, short-time energy based VAD approach can identify endpoints of any speech signal, because the energy of speech frame is larger than that of nonspeech frame [34, 43, 68, 82, 92].

The samples of a waveform of input speech signal is defined as $x(m)$, m is the sample index. The short-time square energy of speech signal $E_{sqr}(n)$ is defined as

$$E_{sqr}(n) = \sum_{m=-\infty}^{+\infty} [x(m)\omega(m-n)]^2 \tag{3.1}$$

The short-time average amplitude $E_{avg}(n)$ is defined as

$$E_{avg}(n) = \sum_{m=-\infty}^{+\infty} |x(m)|\omega(m-n) \tag{3.2}$$

The short-time logarithm energy $E_{log}(m)$ is defined as

$$E_{log}(n) = \sum_{m=-\infty}^{+\infty} \log[x(m)\omega(m-n)]^2 \tag{3.3}$$

The $\omega(n)$ is the a window function which is small width in samples, it represents the frame size n. Usually the rectangular, Hamming and Hanning window functions are used to speech signal processing. The rectangular window function is defined as

$$\omega(n) = \begin{cases} 1 & 0 \leqslant n \leqslant N-1 \\ 0 & other \end{cases} \tag{3.4}$$

The Hamming window function is defined as

$$\omega(n) = \begin{cases} 0.54 - 0.64\cos(\frac{2n\pi}{N-1}) & 0 \leqslant n \leqslant N-1 \\ 0 & other \end{cases} \tag{3.5}$$

The Hanning window function is defined as

$$\omega(n) = \begin{cases} 0.5(1 - \cos(\frac{2n\pi}{N-1})) & 0 \leqslant n \leqslant N-1 \\ 0 & other \end{cases} \tag{3.6}$$

The short-time square energy $E_{sqr}(n)$, logarithm energy $E_{log}(n)$, and average amplitude $E_{avg}(n)$ can embody the signal strength, but their characteristics are difference. To embody the dynamic range of amplitude, the E_{avg} is better than E_{sqr} and E_{log}. To embody the level difference between surd and sonant, the E_{avg} is worse than E_{sqr} and E_{log}. Hence we use the short-time square energy E_{sqr} and rectangular window function to detect endpoint in the chapter.

3.3 Zero-crossing rate algorithm

Sometime, aforesaid short-time energy algorithms are inaccurate for VAD. The human's pronunciation include the surd and sonant. The sonant is produced by the vibration of the vocal chords. The amplitude of sonant is high and periodicity is apparently. The surd is without vibration of the vocal chords, it is produced by the friction, impact or plosive that the suction of air into the mouth. Thus, the short-time energy is lower than that of

sonant. It can be identified into nonspeech easily by short-time energy method. Fig. 3.2 shows a waveform of word 'Sapporo'. The amplitude of surd segment is lower than that of sonant segment, and it is almost same to that of nonspeech segment. Hence, they are very difficult to identify with just our eyes. If the nonspeech and surd segments are zoomed, we found the waveform of surd segment goes up and down so quickly around zero level value, and the number of crossing zero level value for nonspeech segment is fewer. Fig. 3.5 shows the short-time energy and zero-crossing rate of frames of a speech signal. The the number of crossing zero level value can be used to distinguish the endpoint of speech signal. The method is described as zero-crossing rate (ZCR) [3, 8, 19, 52, 63, 101].

The zero-crossing rate is defined as

$$ZCR(n) = \frac{1}{2} \sum_{m=-\infty}^{+\infty} |sgn[x(m)] - sgn[x(m-1)]| \omega(m-n) \tag{3.7}$$

where the $sgn[\cdot]$ is symbol function, it is defined as

$$sgn[x] = \begin{cases} 1 & x \geqslant 0 \\ -1 & x < 0 \end{cases} \tag{3.8}$$

The $\omega(n)$ usually uses the rectangular window function.

3.4 Double thresholds algorithm based on short-time energy and zero-crossing rate

The double thresholds algorithm sets two thresholds for speech signal. The starting of speech signal is detected by the hither threshold, and then the other threshold is used to accurately detect the real starting point of speech signal. The algorithm is described as follow.

Firstly, the speech signal is divided into some frames. The time of each frame is about $20-30$ ms. There is $10-20$ ms overlaps for adjacent frames. We use the 23.2 ms (about 256 sample point) for a frame and the part of overlaps is 11.2 ms (about 128 sample point). The short-time energy of the i^{th} frame is defined as $E_i(N)$, the zero-crossing rate is defined as $ZCR_i(n)$. Initially, we do not known which frame is the nonspeech or speech, hence we assume that the inchoative short-time part is the nonspeech segment, there is only even distributed background noise in the part. The threshold of zero-crossing rate $IZCT$, lower threshold of short-time energy ITL and higher threshold of short-time energy ITU of the first N frames can be calculated. N is set as 5. The threshold of zero-crossing rate is defined as

$$IZCT = \min(IF, \overline{IZC} + 2\zeta_{IZC}) \tag{3.9}$$

where $\overline{IZC}$ is the average of ZCR of first five frames, ζ_{IZC} is the standard deviation of ZCR, IF is empirical value, usually $IF = 25$.

$$\overline{IZC} = \frac{1}{N}\sum_{i=1}^{N} ZCR_i(n) \tag{3.10}$$

$$\zeta_{IZC} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(ZCR_i(n) - \overline{IZC})^2} \tag{3.11}$$

The short-time energy $E_i(n)$ of the first N frames are calculated. The maximum among short-time energies of all frames is defined as IMX, and the minimum is defined as IMN.

$$\begin{aligned} I_1 &= 0.03 \times (IMX - IMN) + IMN \\ I_2 &= 4 \times IMN \end{aligned} \tag{3.12}$$

so ITL and ITU are defined as

$$ITL = \min(I_1, I_2) \tag{3.13}$$

$$ITU = 5 \times ITL \tag{3.14}$$

Then, we detect the $E_i(n)$ of each frame which is from No $N+1$ frame. If $E_i(n)$ of a frame is more than ITL, then the frame number is recorded as p_1, detecting continues. If $E_i(n)$ of a frame is lower than ITL, and all $E_i(n)$ are less than ITU, which frames are until current frame, then p_1 is updated to current frame number. Otherwise, the p_1^{th} frame is as the starting of speech signal.

Finally, we forward compare the ZCR of each frame from the p_1^{th} frame, If $ZCR(n)$ of continuous three frames are more than $IZCT$, then the p_1 is updated to first frame number of three frames. Otherwise, the starting of speech signal is still the p_1^{th} frame.

The method how to detect end of speech signal is same to above mentioned method. If we detect $E_i(n)$ of a frame is less than the ITL, then the frame number is set as p_2. If all $E_i(n)$ of the latter N frames are also less than ITL, then we detect all frames from the p_2^{th} frame by the ZCR, until the $ZCR_i(n) < IZCT$. The last frame which $ZCR_i(n) < IZCT$ is the end of speech signal.

Fig. 3.4 shows VAD method with double thresholds algorithm based on short-time energy and zero-crossing rate in clear environment. The solid line are the detected end-points with shot-time energy, the dashes lines are the detected end-points with ZCR.

3.5 Modified short-time energy for VAD

The detection accuracy with normally short-time energy and ZCR methods is not so ideal. By plenty of experiment results, the detection accuracy of normally methods is about 83%. In Eq. 3.9, the $\overline{IZC} + 2\zeta_{IZC}$ for first noisy segment may be zero or a very close approximation value to zero. Hence, the $IZCT$ may be zero or a very small value by Eq. 3.9. If the ZCR values of some latish noisy frames are bigger than that of first noisy frames, then it can lead to detect mistakenly. According to observe and statistics, if the short-time energy of unvoiced consonants is very low in the beginning segment of speech signal, then the ZCR values of continuous 10 frames are usually not more than the 20. It is only in chance cases, a few ZCR values of noisy frames are more than 20, as well as these noisy frames are discontinuous. Thus, the $IZCT$ is processed as

$$IZCT = \max(IZCT, 15) \tag{3.15}$$

For the ITL and ITU, the I_1 and I_2 are obtained by the IMX and IMN of the first N frames. Under the a relatively noise-free environment, the average of short-time energy of the first 10 frames is close to 0. However, in practice, there are many kinds of noise, the energies of these noises are difference in different cases. The differentials between IMX and IMN may be very big, ever it is several orders of magnitude. For the Eq. 3.12, ITL is closely related to IMN. If ITL value of a frame in first N frames is very small, then the other that of $N-1$ frames are meaningless. It is very easy to recognize the noise frame from noisy speech by the ITL in this case.

Hence, ITL and ITU should be dynamic. They are volatile with the average of energy of first N frames. The ITL and ITU are modified as

$$IMA = \frac{1}{N}\sum_{i=1}^{N} E_i(n) \tag{3.16}$$

$$ITL = \alpha_1 \cdot IMA \tag{3.17}$$

$$ITU = \alpha_2 \cdot IMA \tag{3.18}$$

where α_1 and α_2 are empirical value. By many results of experiment, the higher accuracy detection can be obtained when $\alpha_1 \approx 0.1$ and $\alpha_2 \approx 1.5$

Sometimes, pulse-noise strength can be significant in the nonspeech segment, because the background noise is uncertainty. These short-time energies of pulse-noise frames are very strong, leading to their short-time energies are more than ITU. In the case, the pulse-noise may be recognized to speech. In practical, human's utterance is continuous, the adjacent frames of speech are associated. It is impossible that the undulation of energy of adjacent frames is crazy. The case that the energy is instantaneous aggrandizement can only be noise.

This problem can render frame energy larger than ITU, and thus, it may be recognized as speech. To counter this, we propose smoothing the frame energy level during nonspeech segments using the following first-order recursive equation:

$$F_i(n) = \lambda F_{i-1}(n) + (1-\lambda)E_i(n), \tag{3.19}$$

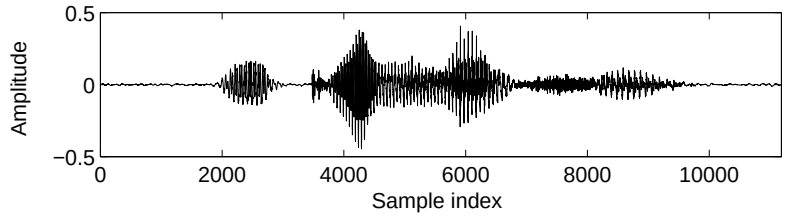
where $\lambda \in (0,1)$ is the forgetting factor. The initial condition $F_0(n) = 0$. Then, if $F_i(n) \geqslant ITU$, the update in (3.19) stops and frame i is classified as speech data. Conversely, if $F_i(n) < ITU$, updating continues.

The $F_i(n)$ is easily represent the smoothness properties between adjacent frames, substantially decreases the effect of pulse-noise.

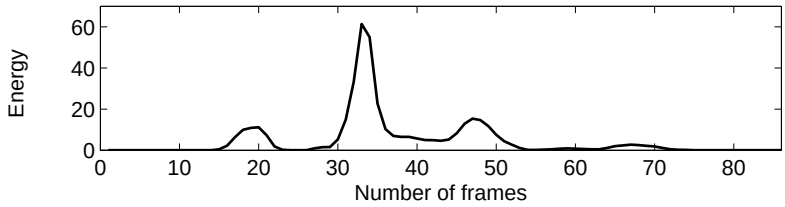
Fig. 3.6 shows the VAD with modified short-time energy. In Fig. 3.6(e), the dash dot lines indicate the detected endpoints with normalized short-time energy method,

the short dashes lines indicate the detected endpoints with modified short-time energy method, and the solid lines indicate the detected endpoints with ZCR after modified short-time energy. The modified short-time energy method can reduce the influence of noisy pulse. ZCR can detect the surd segment accurately.

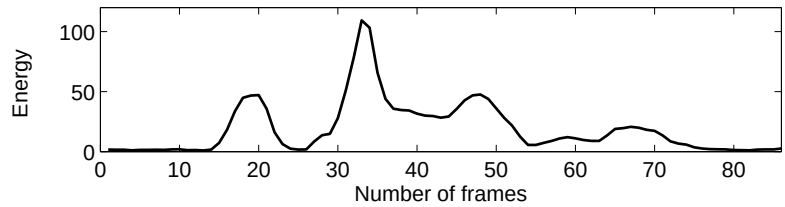
For the end-point detection of speech, the ZCR is not used to detect, and the ITL also is dropped, the only ITU is used to detected the end-point. The reason is that short-time energy of consonant is weak, and its ZCR is high. On the contrary, the short-time energy of vowel is high and its ZCR is low. There are many kinds of phoneme construction (e.g., C-V, V-C, C-V-C) in other language grammar (e.g., English). However, there are only two kinds of phoneme construction (C-V and V) in Japanese language. All of ending of Japanese phoneme are vowel. Hence, it is no benefit to detect the end-point with ZCR, even it is opposite effect. In the ending part of speech, there are usually some terminal sounds that the short-time energy is very weak. These terminal sounds are no benefit to improve the recognition accuracy of ASR system, thus it is fit that the only ITU is used to detected the end-point.



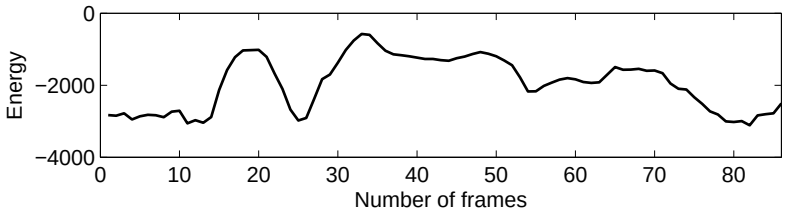
(a) The waveform of a speech signal



(b) Short-time square energy



(c) Short-time average energy



(d) Short-time logarithm energy

Figure 3.1. The short-time square energy, logarithm energy and average energy of a speech signal

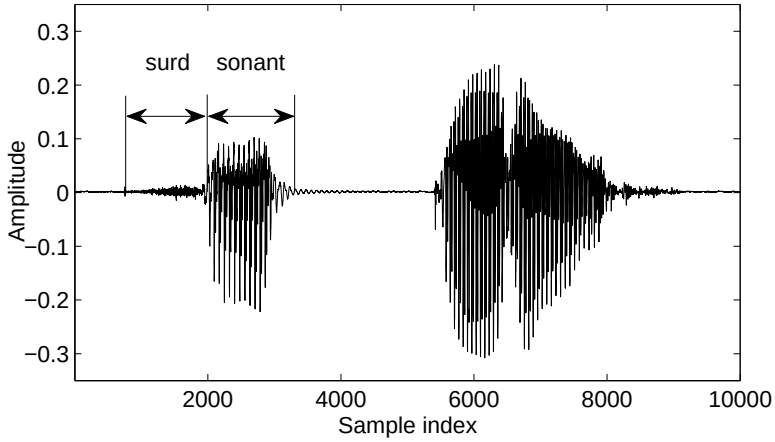
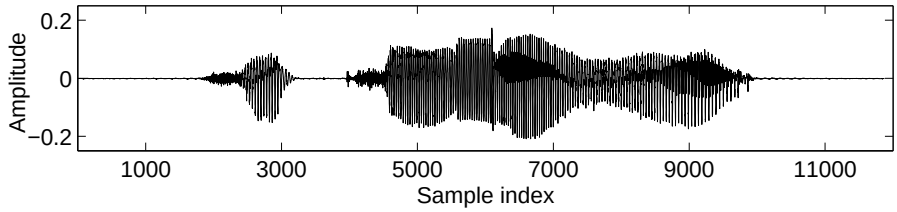
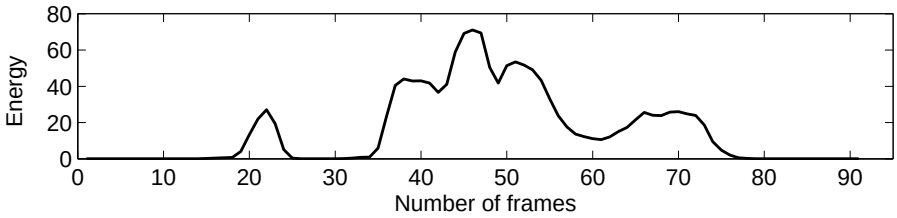


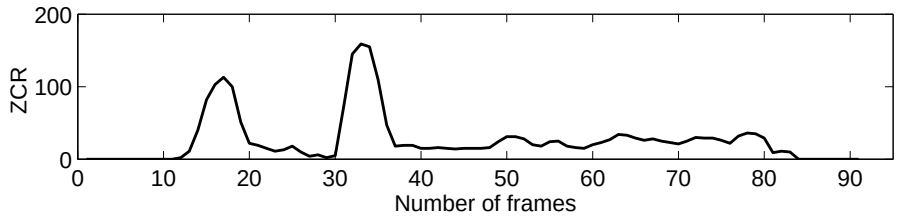
Figure 3.2. Waveform of word ‘Sapporo’



(a) The waveform of a speech signal

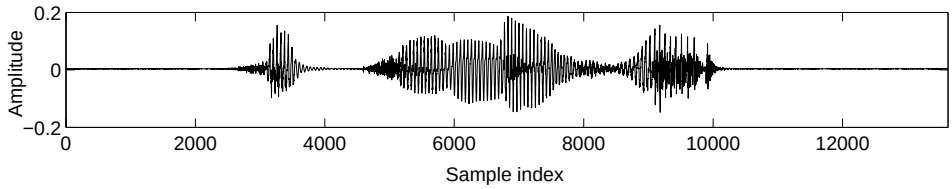


(b) Short-time energy

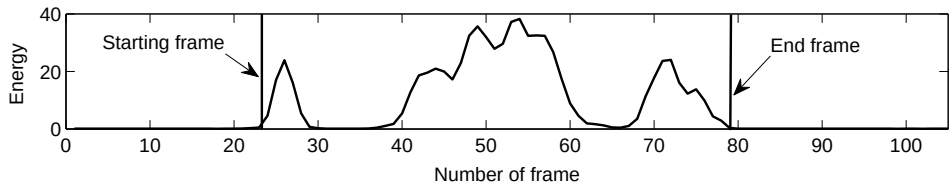


(c) Zero-crossing rate

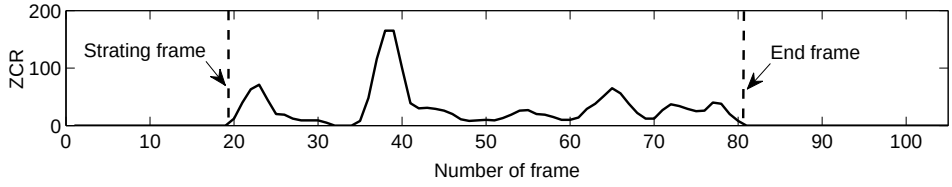
Figure 3.3. Short-time energy and zero-crossing rate of frames of a speech signal



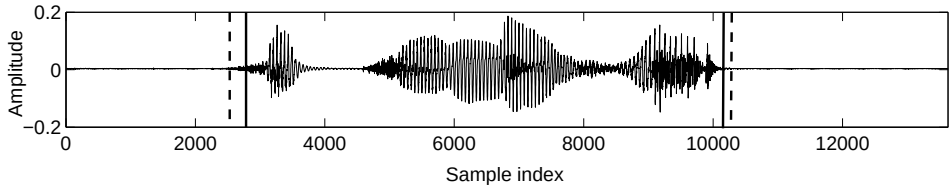
(a) The waveform of a clean speech signal



(b) VAD with short-time energy

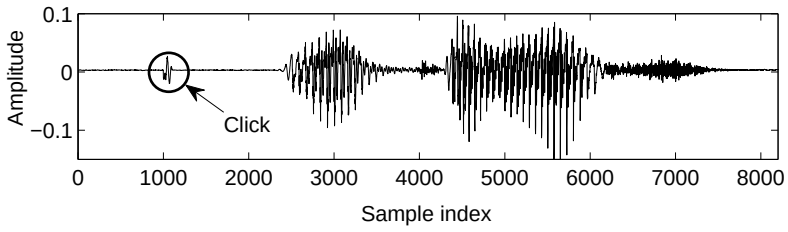


(c) VAD with zero-crossing rate

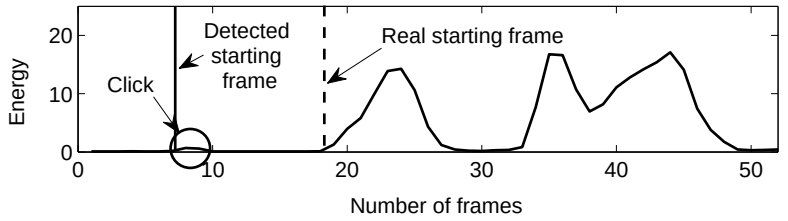


(d) The end-points of speech with double methods

Figure 3.4. VAD with double thresholds algorithm based on short-time energy and ZCR in clear environment

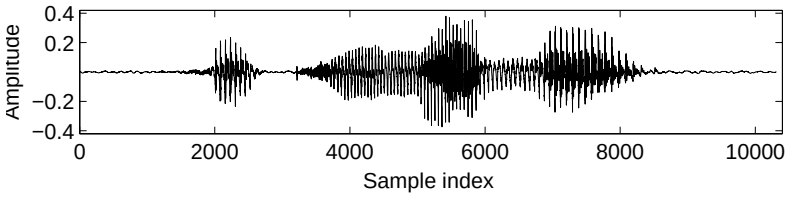


(a) The waveform of a speech signal

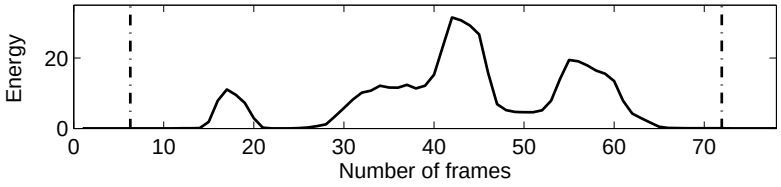


(b) Short-time energy

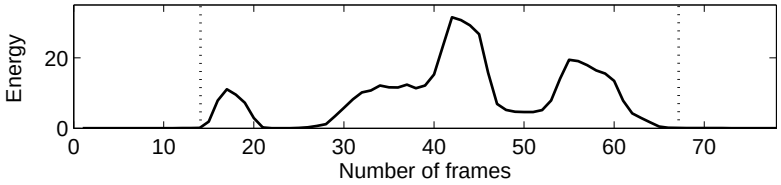
Figure 3.5. Short-time energy and zero-crossing rate of frames of a speech signal



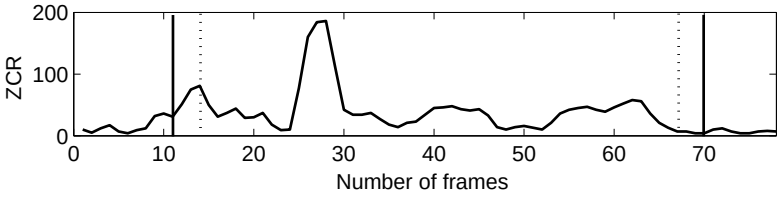
(a) The waveform of a clean speech signal



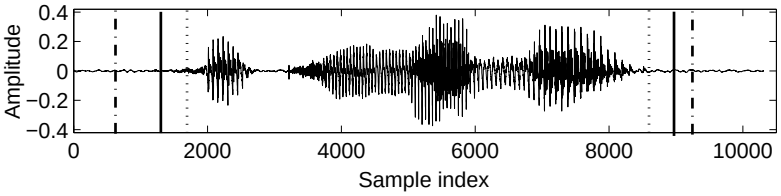
(b) VAD with normalized short-time energy method



(c) VAD with modified short-time energy method



(d) VAD with zero-crossing rate



(e) The endpoints of speech by VAD

Figure 3.6. VAD with double threshold algorithm based on modified short-time energy and zero-crossing rate in clear environment

Chapter 4

Noise Reduction

4.1 Influence of additivity and multiplicative noises

Usually, there are two kinds of noise by interrelation between single and noise. One is additivity noise, the other is multiplicative noise [57, 100, 102]. Assuming the speech signal is $s(t)$ and noise signal is $n(t)$. If the mixed superimposed waveform is $s(t) + n(t)$, then the noise is additivity noise. If the mixed superimposed waveform is $s(t) \otimes n(t)$, then the noise is multiplicative noise. The additivity noise and speech signal are independent with each other. It exists in all the time whether there are speech signal or not. We can only reduce the influence of additivity noise, but can not eliminate the additive noise completely. Thus, the additivity noise can effect the speech signal inevitably. The multiplicative noise is usually caused with the unfavorable channel. It exists with the presence of speech signal. If the speech signal disappear, then the multiplicative noise is also disappear.

In the time domain, we assume the interfered speech signal by additivity noise is $x(t)$

$$x(t) = s(t) + n(t) \tag{4.1}$$

The $x(t)$ is made Fourier transform, then the corresponding relation is follow in the frequency domain and power spectrum.

$$\begin{aligned} |X(t,i)|^2 &= |S(t,i)+N(t,i)|^2 \\ &= |S(t,i)|^2+|N(t,i)|^2+2|S(t,i)||N(t,i)|\cos(\theta(t,i)) \end{aligned} \tag{4.2}$$

Where $X(\cdot)$ is spectrum of the mixed superimposed signal, $S(\cdot)$ is spectrum of speech signal, $N(\cdot)$ is spectrum of additivity noise. t is frame index, i is the frequency components index of the t^{th} frame. $\theta(t,i)$ is the phase separation between speech signal and additivity noise on the i^{th} point. If the speech signal and additivity noise are assumed as independent distribution of zero-mean, then

$$|X(t,i)|^2 \approx |S(t,i)|^2+|N(t,i)|^2 \tag{4.3}$$

If we can extrapolate the $|N(t,i)|^2$, then the additivity noise can be removed in the frequency component $|S(t,i)|^2 = |X(t,i)|^2 - |N(t,i)|^2$, e.g., spectral subtraction (SS) method. These methods are based on that additivity noise is considered to approximately invariable. In fact, it is very difficult to extrapolate the power of additivity noise accurately. After subtracting the $|N(t,i)|^2$, a few additivity noise is still left. Furthermore, the distribution of additivity noise is variable, but the method is same.

Moreover, we can analyze the frequency spectral by the $|N(t,i)|^2$ in the all spectrum components. The frequency components which is most of $|N(t,i)|^2$ can be filtered with filter. The method can remove most of noise, but it is also very difficult to confirm the frequency of additivity noise. Some additivity noise is still left.

It is impossible that the multiplicative noises are removed with aforementioned two methods. Because of the multiplicative noise is appeared alongside of speech noise. In order to remove the multiplicative noise, the interfered speech noise must be processed. We assume the interfered speech noise by multiplicative noise is

$$x(t) = s(t) \otimes h(t) \tag{4.4}$$

The $x(t)$ is made fast Fourier transform (FFT), then the $x(t)$ is transformed as

$$X(t,i) = S(t,i) \cdot H(t,i) \tag{4.5}$$

Where $X(\cdot)$ is spectrum of the mixed superimposed signal, $S(\cdot)$ is spectrum of speech signal, $H(\cdot)$ is spectrum of multiplicative noise, t is frame index, i is the frequency components index of the t^{th} frame. Eq. (4.5) is made logarithms transformation on both sides.

$$\log|X(t,i)| = \log|S(t,i)| + \log|H(h,i)| \tag{4.6}$$

then, made cepstrum transformation on both sides.

$$X^{cep}(t,n) = S^{cep}(t,n) + H^{cep}(t,n) \tag{4.7}$$

Where $X^{cep}(\cdot)$ is cepstrum of the mixed superimposed signal, $S^{cep}(\cdot)$ is cepstrum of speech signal, $H^{cep}(\cdot)$ is cepstrum of additivity noise. n is the number of channel. Then, it is same as the additivity noise, we can extrapolate the $H^{cep}(t,n)$, and then the multiplicative noise can be removed in the frequency component $H^{cep}(t,n) = X^{cep}(t,n) - S^{cep}(t,n)$.

4.2 Running spectrum filtering algorithm

Running spectrum filtering (RSF) is a noise reduction method that exploits the difference of temporal variability between the spectra of speech and noise signals to remove the noise [18,22,24–27,39,47,60,87,103]. Thus, using RSF we have evaluated the different characteristics of speech and noise signals. In the modulation spectrum, we have found that the noise spectrum is concentrated in the direct component (DC). Most of the noise energy is distributed in the low-frequency band of the modulation spectrum. Fig. 4.1 shows the power spectrum of clean speech in the the modulation spectra. Fig. 4.2 shows

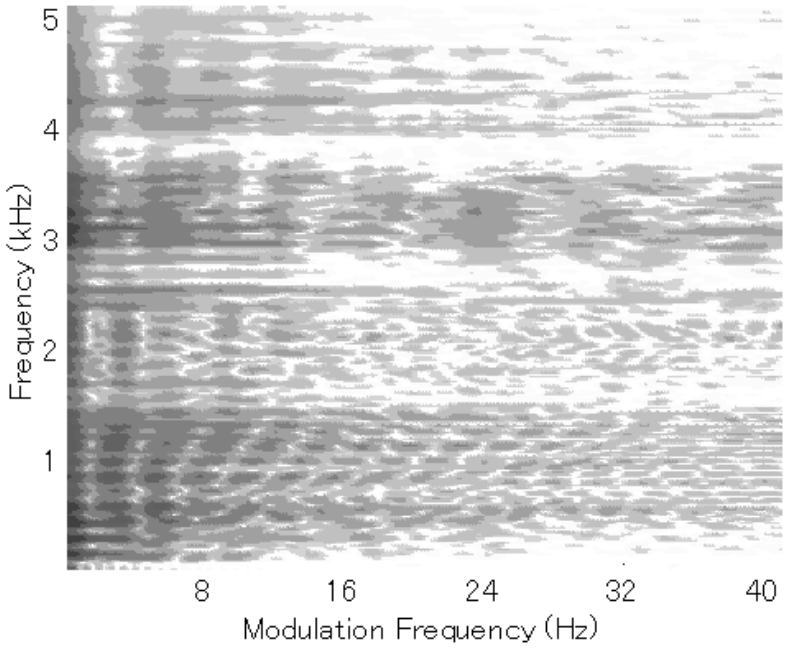


Figure 4.1. Power spectrum of clean speech in the modulation spectra

the power spectrum of 5 dB white noise in the the modulation spectra. Fig. 4.3 shows the power spectrum of mixed waveform with clean speech and 5 dB white noise in the the modulation spectra. The black shade means the strength of energy in three figures. The energy of clean speech becomes gradually weak with the modulation frequency raising in Fig. 4.1. Especially, where modulation frequency is about less than 16 Hz, the energy is particularly strong. This shows that the significant constituent of speech is in the band $[0, 16]$ Hz. Fig. 4.2 shows the noise is distributed on whole spectrum, but the energy of noise is stronger than another one in frequency band $[0, 1]$ Hz. The energy of speech is strengthened on the whole spectrum, since the noise is added in Fig. 4.3. The additivity noise on frequency band $[0, 1]$ Hz exerts such tremendous effect on speech signal.

Fig. 4.4 shows logarithm spectrum of clean speech in the modulation spectra. Fig. 4.5 shows logarithm spectrum of mixed waveform with clean speech and 5 dB white noise

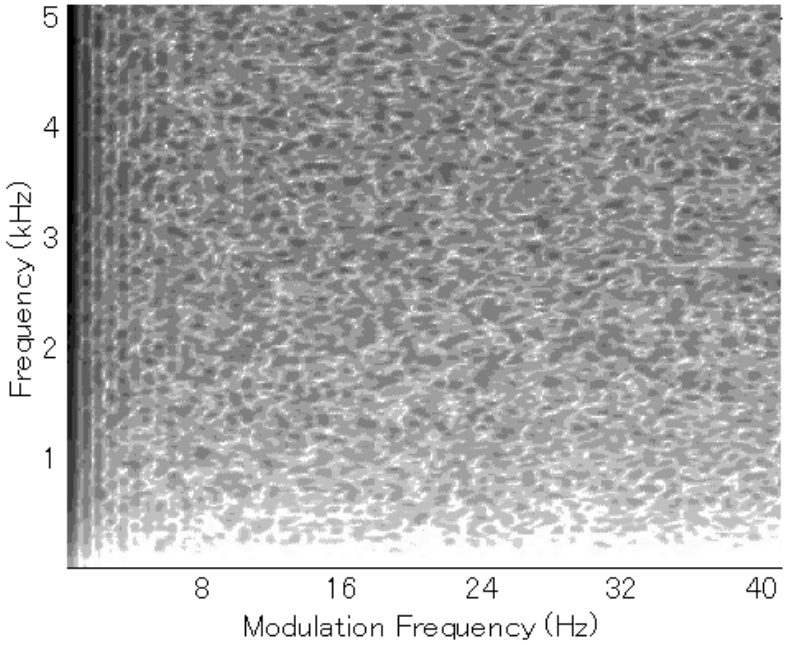


Figure 4.2. Power spectrum of 5 dB white noise in the modulation spectra

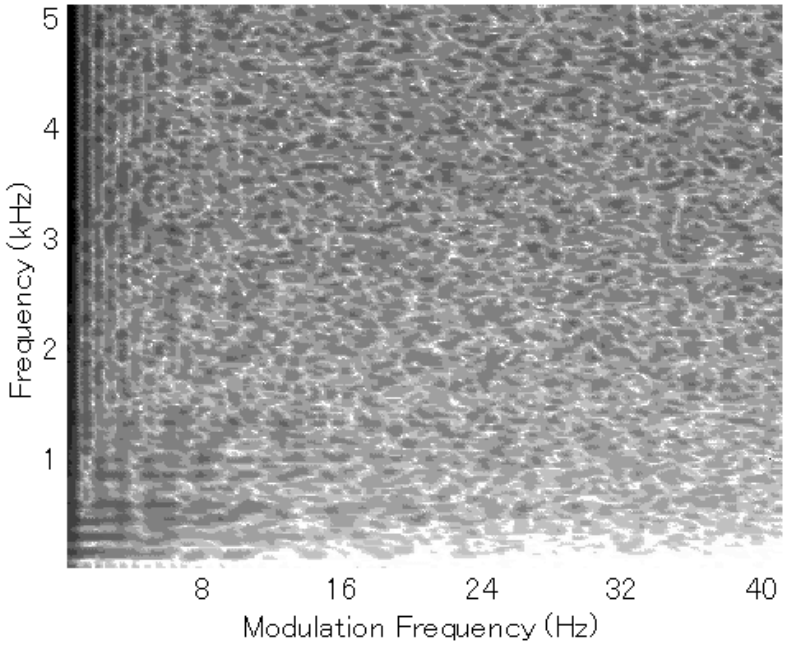


Figure 4.3. Power spectrum of mixed waveform with clean speech and 5 dB white noise in the modulation spectra

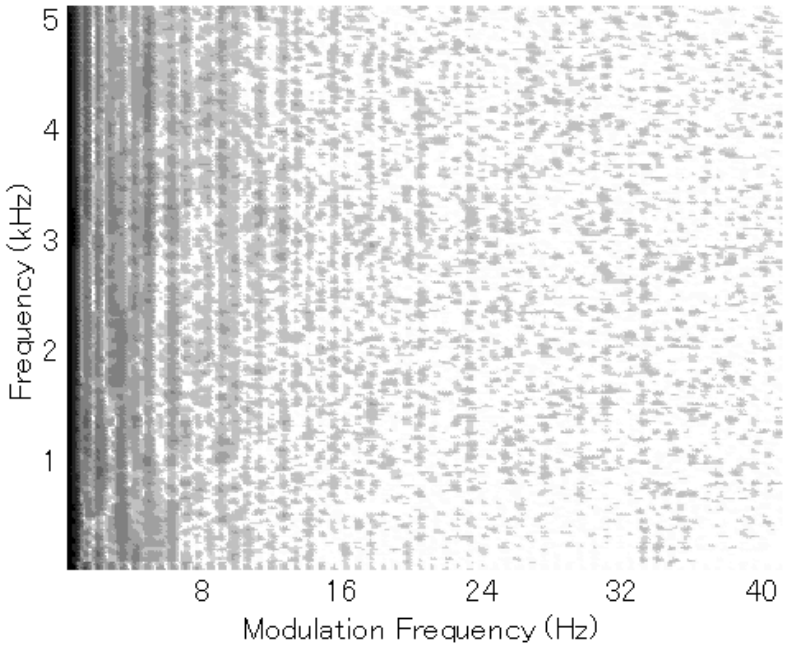


Figure 4.4. Logarithm spectrum of clean speech in the modulation spectra

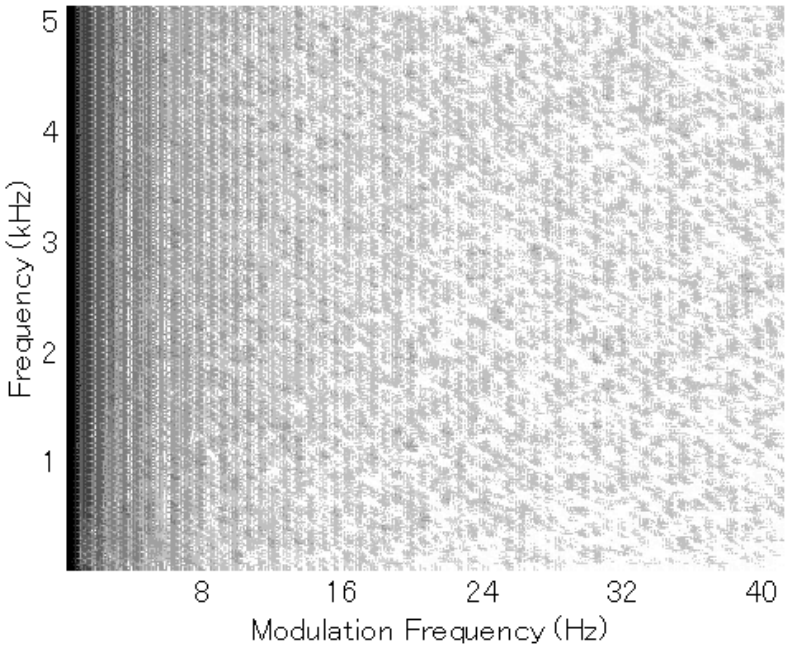


Figure 4.5. Logarithm spectrum of mixed waveform with clean speech and 5 dB white noise in the modulation spectra

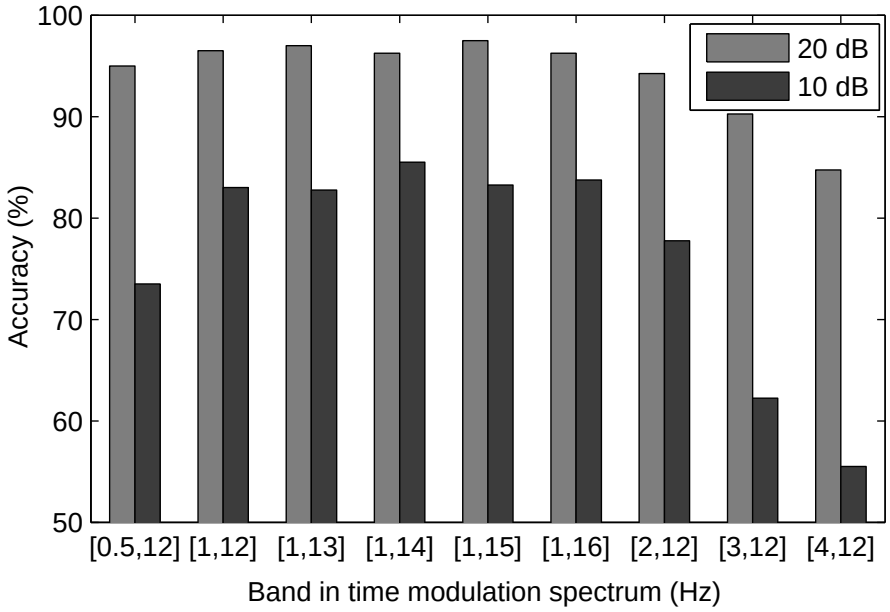


Figure 4.6. DTW recognition accuracy vs. band for RSF

in the modulation spectra. The distributions of energy are almost same to power spectra. Especially, where modulation frequency is about less than 16 Hz, the energy is particularly strong. To speech recognition, the important information of speech is about in the frequency band $[1,16]$ Hz. The multiplication noise exerts such tremendous influence on frequency band of close 0 Hz. Fig. 4.6 shows DTW recognition accuracies for different bands filtered by RSF. The $[1,16]$ Hz band is important for the speech spectrum. Recognition accuracy is much higher in band $[1,12]$ Hz vs than in band $[0.5,12]$ Hz. The figure shows that most of the noise is located in band $[0,1]$ Hz.

Thus, removing low-frequency components with a high-pass filter can reduce the noise. On the other hand, the speech spectrum covers a wider frequency range. There is a little low energy of noise in the high-frequency band. Therefore, we can use a band-pass filter to separate speech from noise. The overview of RSF processing is shown in Fig. 4.7. The additive noise is reduced in the power spectra and the multiplicative noise

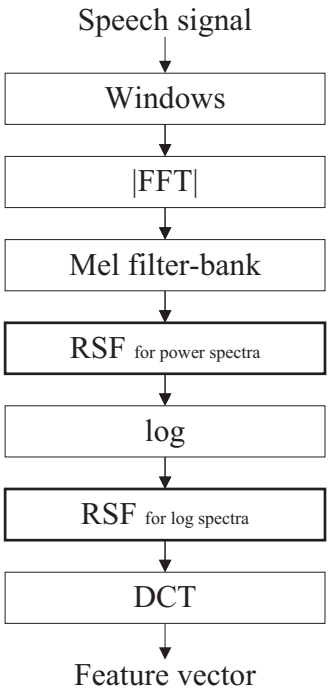


Figure 4.7. Overview of RSF

is reduced in the logarithm spectra by RSF.

RSF is similar to relative spectral (RASTA), which is proposed by Hermansky et al. [21, 29, 30, 33]. RASTA is that speech signal is filtered by a band-pass filter in each frequency channel, according to time tract of speech parameter. RASTA uses a band-pass filter with a sharp spectral zero at the zero frequency to cut-off slowly changing or steady-state factors in speech spectrum.

RASTA is usually used to logarithm or power spectra. It also can be applied to cepstrum or power spectra, which is transformed through expanding static nonlinear transformation. RASTA uses an infinite impulse response (IIR) filter [61, 69, 71]. Its transfer function is

$$H(z) = G \times \frac{z^{N-1} \sum_{n=0}^{N-1} \left(\frac{N-1}{2} - n \right) z^{-n}}{1 - \rho z^{-1}} \tag{4.8}$$

Usually, the $N = 5$, $G = 0.1$, and $\rho = 0.98$. Then,

$$H(z) = 0.1z^4 \times \frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{1 - 0.98z^{-1}} \tag{4.9}$$

The function of conventional IIR filter is

$$y(t) = \frac{\sum_{k=0}^{L} b_k z^{-n}}{1 + \sum_{k=1}^{M} a_k z^{-n}} x(t) \tag{4.10}$$

Where $x(t)$ is input signal, $y(t)$ is output signal, a_k and b_k are coefficients of filter. The IIR filter is also defined as L^{th}-order difference equation by Eq. (4.10)

$$y(t) = \sum_{k=0}^{M-1} a_k x(t-k) - \sum_{k=1}^{L} b_k y(t-k) \tag{4.11}$$

We known the output value is calculated with current input and last output values. Hence, the effect of steady background noise is still residue after many iterations [13]. In order to cut-off the effect of input signal, the RSF uses FIR filter instead of IIR filter [23]. The transfer function of FIR filter is

$$y(t) = \sum_{k=0}^{L} b_k z^{-n} x(t) \tag{4.12}$$

Where b_k is coefficients of filter. In order to get the sharp filter, the order of FIR filter must be very big. In our system, the order is usually 240. If the order is big, then the calculation cost is big. Hence, the calculation time is big. The higher order can affect the performance of ASR system. A high-performance FIR hardware with high order has been designed for solving the problem in [23, 96–98].

Fig. 4.8 shows the comparison of MFCC feature vector of 3^{th} channel between clean and noisy speech. Fig. 4.9 shows the comparison of MFCC feature vectors of 3^{th} channel between clean and noisy speech after RSF.

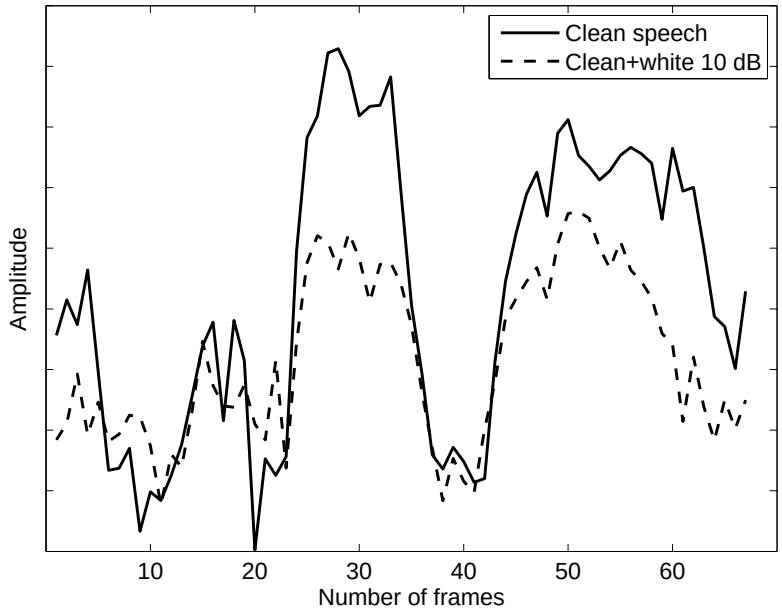


Figure 4.8. The comparison of MFCC feature vectors of 3^{th} channel between clean and noisy speech

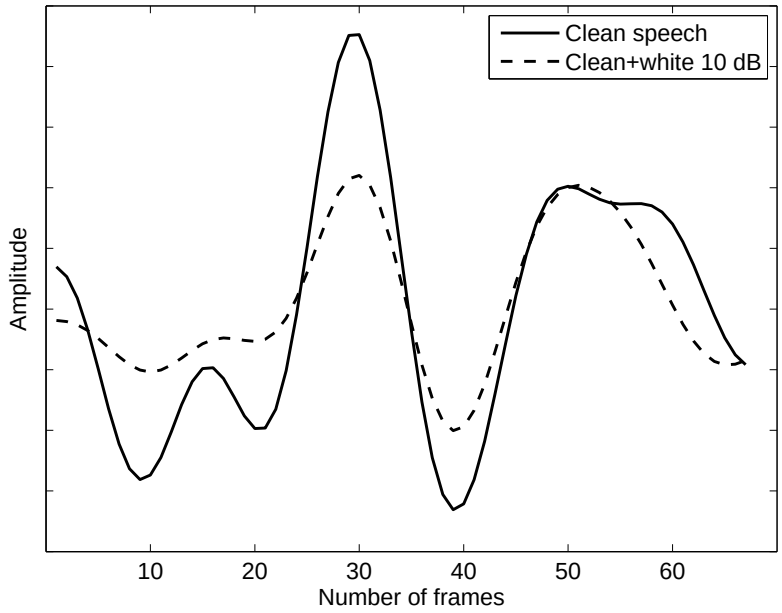


Figure 4.9. The comparison of MFCC feature vectors of 3^{th} channel between clean and noisy speech after RSF

4.3 CMS algorithm

CMS is a simple method of reducing noise [7, 58, 74, 75, 83]. White noise is uniformly distributed in a spectrum. After feature extraction, the MFCC feature vectors are obtained in the cepstral domain. In a long-time range, almost all speech features are changed with the progress of time. On the other hand, the time-invariant noise features in such a range are considered as almost constant. The subtraction of the time-invariant features from noisy speech features result in the reduction of noise components. We assume that a speech waveform is divided into h short frames. $f_i(t)$ is the t^{th} component of the i^{th} frame.

Noise reduction is then executed as Eq. (4.13).

$$f_i'(t) = f_i(t) - \frac{1}{h}\sum_{j=1}^{h} f_j(t) \tag{4.13}$$

Fig. 4.10 shows the comparison of MFCC feature vector of 3^{th} channel between clean and noisy speech after CMS.

4.4 Dynamic range adjustment algorithm

Usually, when white noise is added to a speech waveform, observing the speech waveform is more difficult than observing the clean speech. In addition, when RSF or CMS is applied for noise reduction, the signal amplitude is typically reduced.

The cepstral mean-variance normalization (CMVN) is proposed to adjust the amplitude [65, 84, 85]. The feature vector of each frame is normalized as follows

$$f_i'(t) = \frac{f_i(t) - \mu(t)}{\sigma(t)} \tag{4.14}$$

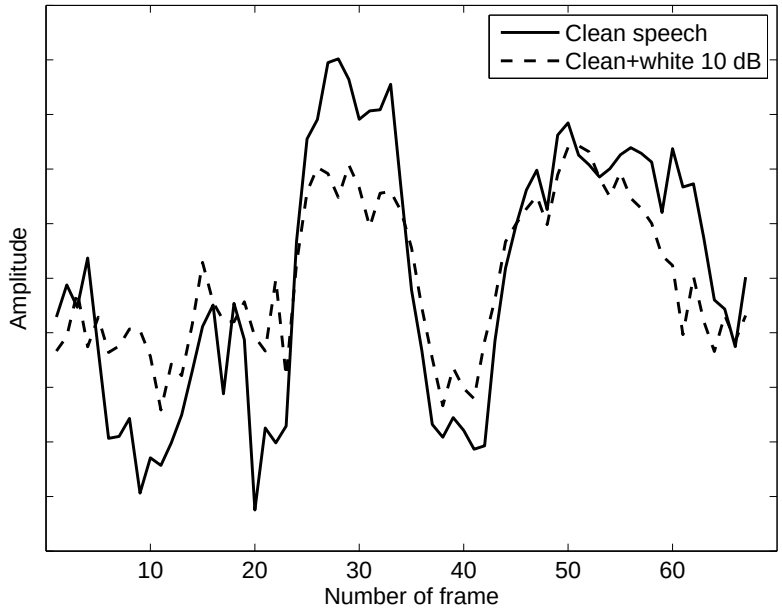


Figure 4.10. The comparison of MFCC feature vector of 3^{th} channel between clean and noisy speech after CMS

where $\mu(t)$ is mean of all frames in t^{th} component. It is calculated as follow

$$\mu(t) = \frac{1}{h}\sum_{i=1}^{h} f_i(t) \tag{4.15}$$

where $\sigma(t)$ is the standard deviation of all frames in t^{th} component. It is calculated as follow

$$\sigma(t) = \sqrt{\frac{1}{h}\sum_{i=1}^{h}(f_i(t) - \mu(t))^2} \tag{4.16}$$

Since clean speech is typically used as reference data, the amplitude difference between clean and RSF- or CMS-processed noisy speech deteriorates the recognition accuracy. The CMVN can normalize the waveform of each dimensional. But the lengths of voiceless segment of identical pronunciation in different time are different. Hence, the standard deviation are different. Then, the waveforms are made a great deal of difference for identical pronunciation in different time after CMVN processing. The shapes

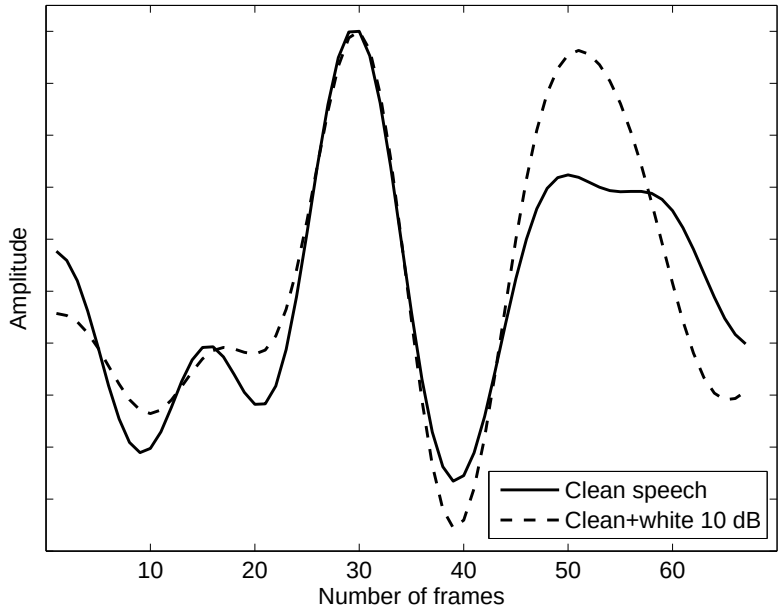


Figure 4.11. The comparison of MFCC feature vectors of 3^{th} channel between clean and noisy speech after RSF and DRA

of waveform are changed to original one.

Dynamic range adjustment (DRA) can be used to compensate for this difference using the following normalization [88, 89, 95].

$$f_i'(t) = \frac{f_i(t)}{arg \max\limits_{j=1,\cdots,h} |f_j(t)|} \tag{4.17}$$

DRA makes it possible to obtain similar cepstrum data for clean speech and noisy speech after CMS or RSF. However, The shapes of waveform are kept same to original one.

Fig.4.11 shows the comparison of MFCC feature vector of 3^{th} channel between clean and noisy speech after RSF and DRA. Fig.4.12 shows the comparison of MFCC feature vector of 3^{th} channel between clean and noisy speech after CMS and DRA.

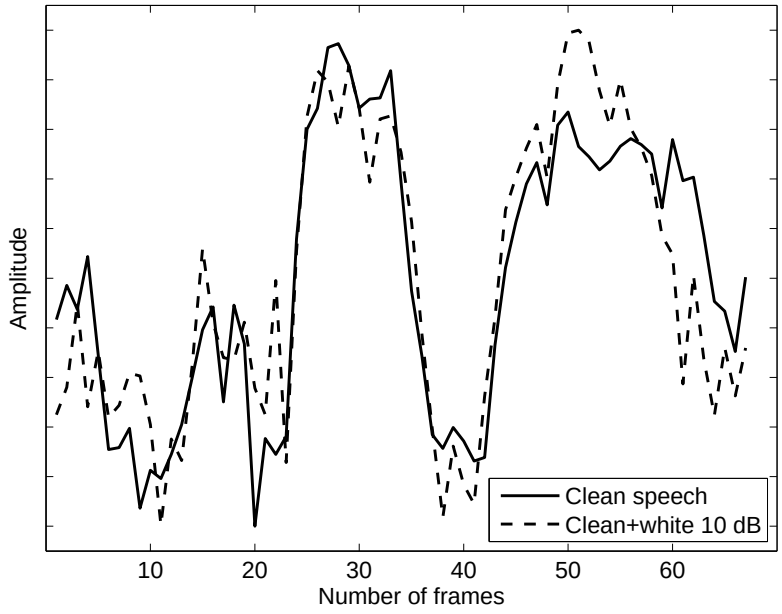


Figure 4.12. The comparison of MFCC feature vectors of 3^{th} channel between clean and noisy speech after CMS and DRA

4.5 Proposed noise reduction method

In real environment, the additivity and multiplicative noises are simultaneous. Hence, the mixed superimposed speech waveform is as follow in time domain [1].

$$x(t) = s(t) \otimes h(t) + n(t) \tag{4.18}$$

Where $x(t)$ is noisy speech signal, $s(t)$ is speech signal, $h(t)$ is multiplicative noise, and $n(t)$ is additivity noise. The Eq. (4.18) is Fourier transformed on both sides. In frequency and power spectrums, the equation is follow, which is effected by the additivity and multiplicative noises.

$$X(t,i) = S(t,i)H(t,i) + N(t,i) \tag{4.19}$$

$$
\begin{aligned}
|X(t,i)|^2 &= |S(t,i)H(t,i)+N(t,i)|^2 \\
&= |S(t,i)H(t,i)|^2+|N(t,i)|^2+2Re[S(t,i)H(t,i)N(t,i)] \\
&= |S(t,i)|^2|H(t,i)|^2+|N(t,i)|^2+2|S(t,i)||H(t,i)||N(t,i)|\cos(\theta(t,i))
\end{aligned} \tag{4.20}
$$

where $\theta(t,i)$ is the phase separation between speech signal and additivity noise on the i^{th} point. Because of the speech and noise can be supposed as mutually independent zero-mean distribution, the desired value of last item is zero in Eq. (4.20). Although instantaneous value of each frame is not zero in this item, the output value of each filter unit is equal to weighted sum of energies of all points when computing Mel-filter. Hence, Mel-energy of noisy speech signal is approximately equal to

$$
P_x(t,i) \approx P_s(t,i)P_h(t,i)+P_n(t,i) \tag{4.21}
$$

where $P_x(\cdot)$, $P_s(\cdot)$, $P_h(\cdot)$, and $P_n(\cdot)$ are Mel-energy of noisy speech, clean speech, additivity noise, and multiplicative noise.

In logarithm spectrum, we defined X^{log}, S^{log}, N^{log}, and H^{log} are as values of vector for noisy speech, clean speech, additivity noise, and multiplicative noise. So

$$
X^{log} = S^{log}+H^{log}+\log(I+e^{(N^{log}-S^{log}-H^{log})}) \tag{4.22}
$$

Similarly, we defined X^{cep}, S^{cep}, N^{cep}, and H^{cep} are as values of cepstrum feature vector for noisy speech, clean speech, additivity noise, and multiplicative noise in cestrum spectrum. So

$$
X^{cep} = S^{cep}+H^{cep}+D\log(I+e^{D^{-1}(N^{cep}-S^{cep}-H^{cep})}) \tag{4.23}
$$

where D is discrete cosine transformation (DCT) matrix.

According to Figs. 4.3 and 4.5, the most of energy of additivity noise distributes in lower modulation frequency on power spectrum, especially under 1 Hz. The most of energy of multiplication noise distributes under 1 Hz modulation frequency on logarithm

spectrum. But some energies of additivity and multiplication noises are also distributed in whole modulation frequency domain. RSF algorithm only can filter most of noise by band pass filter, but some noises are still remained. In Eq. (4.23), the H^{cep} can be almost removed by RSF, but the effect of $D\log(I+e^{D^{-1}(N^{cep}-S^{cep}-H^{cep})}$ is in the whole modulation frequency domain.

On the other hand, we known the calculation cost of RSF algorithm is high, since the high order (240) is used. In Fig. 4.7, the conventional RSF algorithm is used twice. One is in power spectrum, the other is in logarithm spectrum. Hence, the calculation time of ASR system with RSF is relatively high.

In order to improve the performance of ASR system, we remove the RSF for noise reduction in power spectra. After cepstrum computing, we use RSF with band-pass filter to reduce the noise. And then, CMS method is used to reduce the remanent noise in whole frequency domain. CMS is simpler than RSF. The calculation cost is far lower than that of RSF. The flowchart of this method is shown in Fig. 4.13.

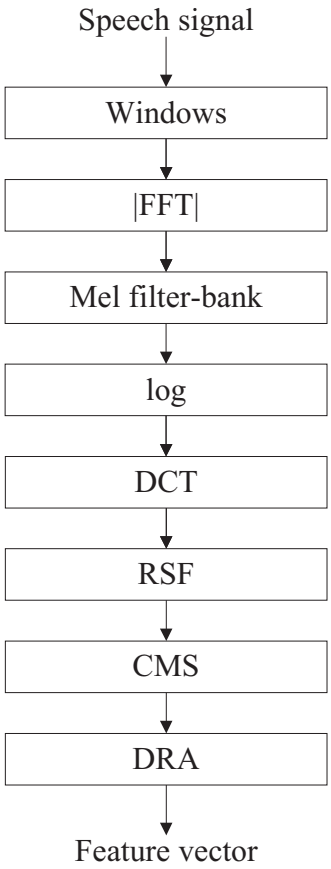


Figure 4.13. Overview of union of RSF CMS and DRA method

Chapter 5

Conventional Dynamic Time Warping Algorithm

5.1 Introduce

It is well known that speech signals can not be compared directly. Usually the lengths of time are difference, because human's speaking rate variations are difference and cause nonlinear fluctuation in a speech pattern time axis, even if the same utterances of same word also are difference in different times. Thus, the time-normalization or eliminating the fluctuation is necessary, and it has been one of central problems in speech recognition research. The DTW algorithm is based on dynamic programming (DP) algorithm and provides a solution to template matching for different lengths of pronunciation [37, 81]. It is a nonlinear warping technique where time series are stretched and compressed to match the reference speech. DTW aligns two sequences of feature vectors by warping the time axis iteratively until an optimal match between the two sequences is found. DTW is an appealing method because it does not require training.

5.2 Dynamic programming algorithm

DP-matching technique is a optimization algorithm [78–80]. It is a pattern matching algorithm for nonlinear time-normalization. DP technique can transform multistage decision process of ASR into many absolute single-stage decision processes, then it solves every absolute decision process one by one. DP technique can align two speech patterns that are time differences between them. DP warps the time axis of one and attains maximum coincident time-axis with other one. Then, the time-normalization distance (Euclidean distance) is calculated as similarity between them.

Usually, the speech signal can be expressed as a sequence of feature vector by feature extraction. We assume the sequence of feature vectors of test speech pattern is $P = [p_1, p_2, \cdots, p_i, \cdots, p_I]$. Where p_1 is the beginning frame, p_I is the end frame, I is the number of frames of test speech pattern. The sequence of feature vectors of reference speech is $Q = [q_1, q_2, \cdots, q_j, \cdots, q_J]$. Where q_1 is the beginning frame, q_J is the end frame, J is the number of frames of reference speech pattern. The P and Q must use the same kind of feature vector, length of frame, window function and vertical shift.

In order to calculate similarity between P and Q, the time-normalization distance $D(P,Q)$ is used to measure. The time-normalization distance is more small, the similarity is more high. The $D(P,Q)$ is total of distance of every pair of corresponding frames between two patterns. The frame time-normalization distance is defined as $d(p_i, q_j)$.

If number of frames of P and Q is same $I = J$, then the time-normalization distance $D(P,Q)$ can be calculated directly. It is as

$$D(P,Q) = \sum_{i=1}^{I} d(p_i, q_i) \tag{5.1}$$

Otherwise, number of frames of P and Q are aligned to same. This linear extension method can make it. If $I < J$, then P can be mapped into a sequence of J frames. The

the time-normalization distance $D(P,Q)$ can be calculated by Eq. (5.1). However, the method has never considered spoken times of each phoneme of speech are variable under different time or cases. Thus, the number of frame of a phoneme is variable. Therefore, the linear extension method is not accurate. The recognition may not be as good. Most of researchers use the DP algorithm in ASR field.

In order to describe the matching processing of DP algorithm, we consider a two-dimension rectangular coordinate system, where frame number of test pattern I is described as x-axis, where frame number of reference pattern J is described as y-axis. The intersection of frame number between them is considered as matching, the time differences can be depicted by a sequence of point $c = (i, j)$. The DP algorithm would find out a path, which passes some intersections of frame number. The all points in the path are corresponding frames which are used to calculated matching distance between two patterns. The path is not selected at random. Although pronunciation speed is variable, but the precedence order of frame in a speech pattern is is invariable. Hence, the selected path must be from upper dexter corner of rectangular coordinate system to the lower sinister corner. Fig. 5.1 shows a selected warping path by DP algorithm.

In order to describe the warping path, the time differences between them can be as a sequence of points:

$$C = [c_1, c_2, \cdots, c_l, \cdots, c_L] \tag{5.2}$$

where

$$c_l = (x(l), y(l)) \tag{5.3}$$

where $x(l)$ is frame number of test pattern in the path, $x(l) \in [1, \cdots, I]$, $y(l)$ is frame number of reference pattern in the path, $y(l) \in [1, \cdots, J]$. The sequence of points in path cab be considered to a function, which try to match Most similar frames from the time

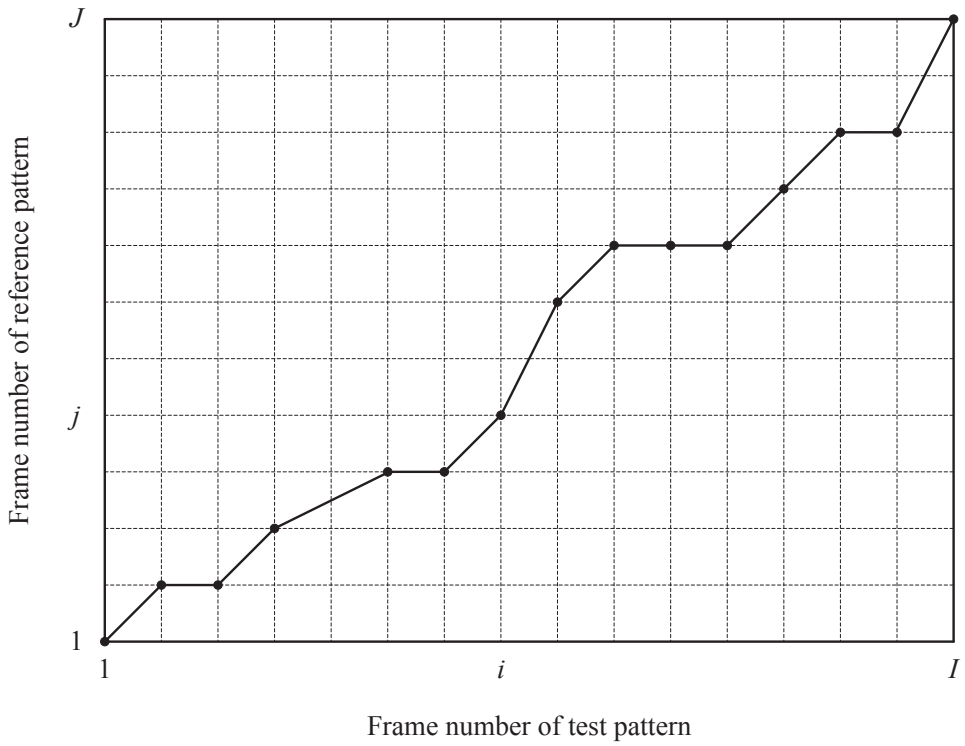


Figure 5.1. A warping path by DP algotihm

axis of pattern P onto that of pattern Q. The matching function is defined as

$$y(l) = \phi(x(l)) \tag{5.4}$$

As the measure of the similarity between two feature vectors of frame p_i and q_j, a Euclidean distance is defined as

$$d(c) = d(i,j) = \| p_i - q_j \| \tag{5.5}$$

Then, the distance of path is summation of all frame distances, and it is defined as

$$D(C) = \sum_{l=1}^{L} d(c_l) \tag{5.6}$$

Usually, the first point is $c_1 = (1,1)$, and the last point is $c_L = (I,J)$ in path. There are many pathes from point $(1,1)$ to (I,J). All of matching pathes must satisfy certain

restriction conditions η. It can obtain its minimum distance value when the matching function $\phi(\cdot)$ is determined to optimally warp the time difference between two patterns. The minimum distance value can be considered to be similarity between them. Hence, the matching distance between test pattern P and reference pattern Q is defined as

$$D(P,Q) = \min_{\substack{\phi(\cdot) \\ \phi(\cdot)\in\eta}} D(C) \tag{5.7}$$

Usually, the DP algorithm calculates all matching distances of frames and pathes (e.g. all intersections in Fig. 5.1). Then, selecting the path which is minimum distance value from (1,1) to (I,J) as the matching distance between two patterns. Hence, the calculation cost of DP algorithm is very large. It costs plenty of time to obtain the optimal path. In fact, some points are not used by the restriction conditions η, and the calculation cost and time can be reduced without these points. Hence, some modified DP algorithms are proposed and they are called DTW algorithm. Two major DTW algorithms have been used conventionally: the one proposed by Sakoe and Chiba in [81], and the one proposed by Itakula in [37]. These conventional DTW algorithms are showed in Fig. 5.2 and Fig. 5.6. The two DTW algorithms proposed different adjustment windows, warping function $\phi(\cdot)$ and restriction conditions η. The two DTW algorithms are more efficient than ordinary DP algorithm.

5.3 Sakoe-Chiba proposed DTW algorithm

A nonnegative weighting coefficient w is intentionally introduced to measure flexible characteristic in the Eq. (5.6) by Sakoe and Chiba .Then the weighted summation of distances on the warping function $\phi(\cdot)$ is

$$\mathbb{D}(C) = \sum_{l=1}^{L} d(c_l)w(l) \tag{5.8}$$

where $w(l)$ is the weight coefficient of $c(l)$. So, the distance between test pattern P and reference pattern Q is defined as

$$D(P,Q) = \min_{\phi(\cdot)} \left(\frac{\sum_{l=1}^{L} d(c_l) w(l)}{\sum_{l=1}^{L} w(l)} \right) \tag{5.9}$$

Although the Eq. (5.9) is different with Eq. (5.6), but it is accordance with fundamental definition of time-normalized distance. The weight coefficient $w(l)$ is used to compensate the influence of every frame on the warping function $\phi(\cdot)$. Thus, the similarity between two patterns depends on the warping function and weight coefficient definition to every pair of frames.

The warping function $\phi(\cdot)$ must consider the characteristics of time sequence of speech signal, and voice versa. For example, the precedence order can not be changed after warped sequence, the distance between adjacent frames can not be so large and so on. Essential speech pattern time-axis structures are continuity, Monotonicity, limitation on the acoustic parameter transition speech in a speech. Hence, some restrictions conditions are very necessary to limit to match the frame by warping function $\phi(\cdot)$. These conditions can be realized as the follow and shown in Fig. 5.2.

1) Monotonic conditions:

$$\begin{aligned} x(l-1) &\leqslant x(l) \\ y(l-1) &\leqslant y(l) \end{aligned} \tag{5.10}$$

These monotonic conditions express the characteristics of time sequence of speech signal, and voice versa. The precedence order can not be changed after warped sequence.

2) Continuous conditions:

$$\begin{aligned} x(l) - x(l-1) &\leqslant 1 \\ y(l) - y(l-1) &\leqslant 1 \end{aligned} \tag{5.11}$$

The continuous conditions express how to choice the adjacent frame. By above two restrictions, we know the pervious adjacent point of c_l is one of $(x(l), y(l)-1)$, $(x(l)-1, y(l)-1)$ and $(x(l)-1, y(l))$.

$$c_{l-1} = \begin{cases} & (x(l), y(l)-1) \\ & (x(l)-1, y(l)-1) \\ or & (x(l)-1, y(l)) \end{cases} \tag{5.12}$$

3) Boundary conditions:

$$\begin{aligned} x(1) = 1, y(1) &= 1 \\ x(L) = I, y(L) &= J \end{aligned} \tag{5.13}$$

The boundary conditions define that the beginning point must be the point $c_1 = (1,1)$ and the end point is $c_L = (I,J)$ for all pathes. These also express the characteristics of time sequence of speech signal. The two endpoints of two patterns firstly must be aligned.

4) Adjustment window condition:

$$\left| x(l) - y(l) \right| \leqslant r \tag{5.14}$$

In fact all warping pathes from $(1,1)$ to (I,J) may not cross all points. Thus, the adjustment windows defines the computation area for warping function. These points out of adjustment windows are excluded from calculation. In other words, the pathes which cross the points out of adjustment windows may not be optimal path. However, the calculation cost of DTW algorithm can be reduced more much efficiently.

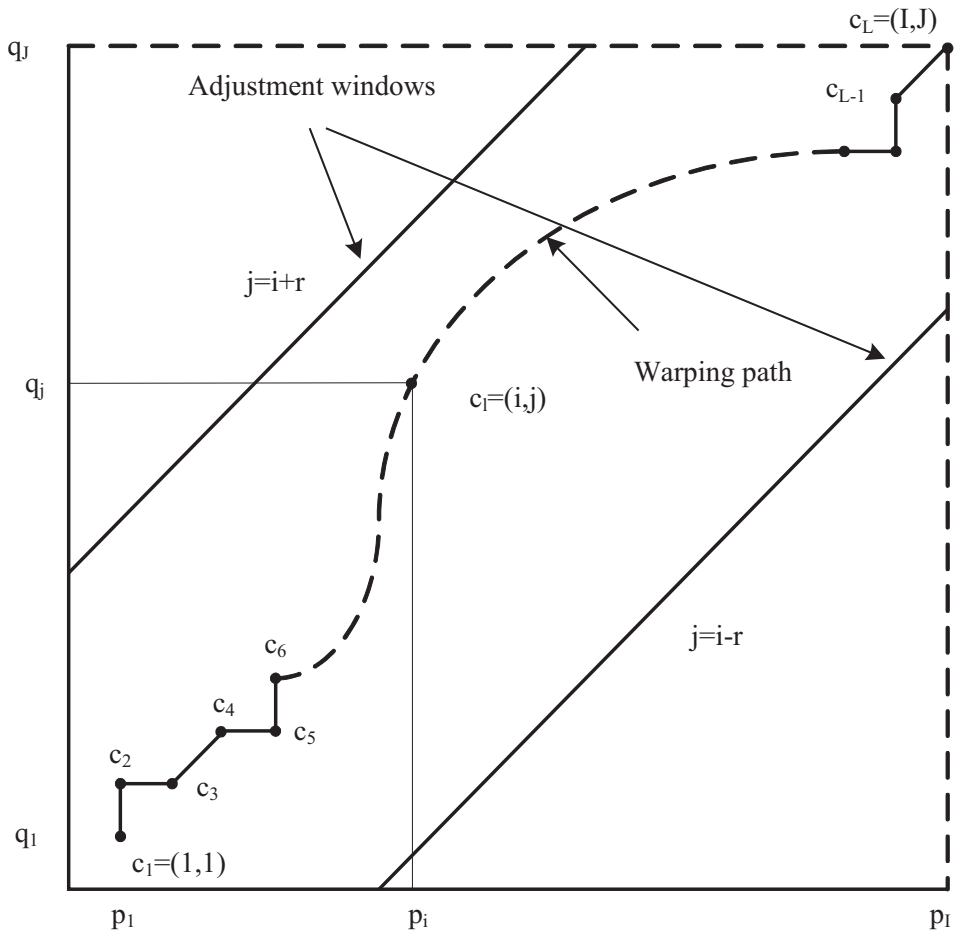


Figure 5.2. Warping function and adjustment window definition for Sakoe and Chiba's DTW algorithms

How to obtain the optimal path (i.e. $D(I,J)$) by above restrictions conditions? We consider it with a negative sequence recursive processing. We known only three pathes can pass the point (i,j) by Eq. (5.12). They are shown in Fig. 5.3. Then, these path distances $D(i,j)$ from $(1,1)$ to (i,j) are

$$\begin{aligned} D_1(i,j) &= d(i,j) + D(i-1,j) \\ D_2(i,j) &= d(i,j) + D(i-1,j-1) \\ D_3(i,j) &= d(i,j) + D(i,j-1) \end{aligned} \tag{5.15}$$

Thus, the optimal path from $1,1$ to i,j is the path whose distance is minimum among $D_1, D_2 and D_3$. And so on, the path distance of every point can be defined a recurrence formula as

$$D(i,j) = d(i,j) + \min \begin{pmatrix} D(i-1,j) \\ D(i-1,j-1) \\ D(i,j-1) \end{pmatrix} \tag{5.16}$$

We can obtain the optimal path of every point, which is from $(1,1)$ in the adjustment windows by Eq. (5.16). Until the last point (I,J), only one path is remained, then the path is matching optimal path between pattern P and pattern Q, and the path distance is similarity between them.

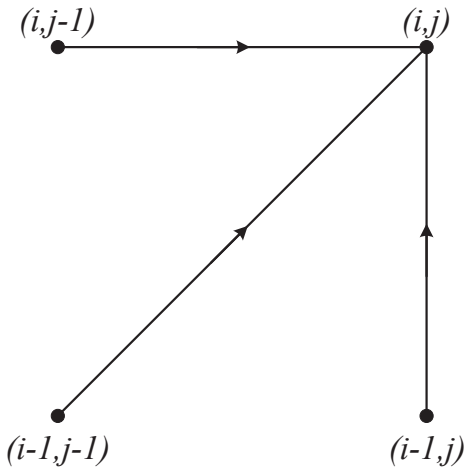


Figure 5.3. Continuous conditions for Sakoe and Chiba's DTW algorithms

In Eq. (5.9), the path distance is with weighting coefficient. Assuming the sum of all weighting coefficient is defined as

$$\mathbb{W}(C) = \sum_{l=1}^{L} w(l) \tag{5.17}$$

then, the time-normalization distance is

$$D(P,Q) = \min_{\phi(\cdot)} \left(\frac{\mathbb{D}(C)}{\mathbb{W}(C)} \right) \tag{5.18}$$

It is very important how to set the reasonable weighting coefficient. It can affect the performance of DTW algorithm. There are two typical weighting coefficients are defined and shown in Fig. 5.4. They are as follows.

1) Symmetric form

$$w(l) = (x(l) - x(l-1)) + (y(l) - y(l-1)) \tag{5.19}$$

then

$$\mathbb{W}(C) = I + J \tag{5.20}$$

2) Asymmetric form

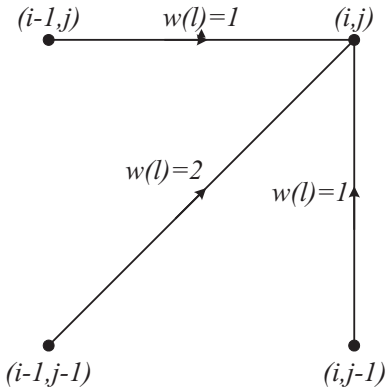
$$w(l) = \begin{cases} x(l) - x(l-1) \\ or \quad y(l) - y(l-1) \end{cases} \tag{5.21}$$

$$\mathbb{W}(C) = \begin{cases} I \\ or \quad J \end{cases} \tag{5.22}$$

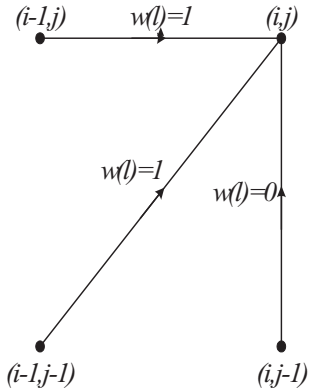
In the symmetric form, $D(I,J) = D(J,I)$. But $D(I,J) \neq D(J,I)$ in the asymmetric form. In the asymmetric form, the weighting coefficient $w(l)$ can reduce to 0 when the anterior point of (i,j) is the point $(i,j-1)$, and it is shown in Fig. 5.4(b). In this case, some feature vectors are possibly excluded the warping path, but the frame weighted distance $d(i,j) \cdot 0 = 0$. It is obvious that the zero weighting coefficient is unreasonable for the veritable path. We would discuss it in late part.

Hence, there are two kinds of optimal path distances under the symmetric and asymmetric forms.

DTW1: Symmetric Sakoe-Chiba's DTW



(a) Symmetric form



(b) Asymmetric form

Figure 5.4. Sakoe and Chiba proposed two weighting coefficients

The first point is $c(1) = (1,1)$, and its preorder point does not exist. The initial condition is

$$\mathbb{D}(c(1)) = d(c(1))w(1) \tag{5.23}$$

We assume the implicit point $c(0) = (0,0)$, then the weighting coefficient $w(1) = 1+1 = 2$ in symmetric form. So the weighted summation of distances of point $(1,1)$ is

$$\mathbb{D}(1,1) = 2d(1,1) \tag{5.24}$$

The optimal weighted warping path distance is:

$$\mathbb{D}(i,j) = \min \begin{pmatrix} \mathbb{D}(i-1,j) + d(i,j) \\ \mathbb{D}(i-1,j-1) + 2d(i,j) \\ \mathbb{D}(i,j-1) + d(i,j) \end{pmatrix} \tag{5.25}$$

The restricting condition is

$$j - r \leqslant i \leqslant j + r \tag{5.26}$$

The time-normalized distance of optimal warping path between two patterns is

$$D(I,J) = \frac{\mathbb{D}(I,J)}{\mathbb{W}(C)} \tag{5.27}$$

where $\mathbb{W}(C) = I + J$.

DTW2: Asymmetric Sakoe-Chiba's DTW

In a similar way, assuming the implicit point $c(0) = (0,0)$, then the weighting coefficient $w(1) = 1$ in asymmetric form. So the weighted summation of distances of point $(1,1)$ is

$$\mathbb{D}(1,1) = d(1,1) \tag{5.28}$$

The weighted optimal distance has been previously discussed. In order to avoid the influence of zero weighting coefficient, we define a new solution for it. Assuming there are k continuous points until the point (i,j) by the j-axis direction (e.g. Fig. 5.5). Then, the nethermost point is $(i, j-k+1)$. We assume its preorder point is (x_p, y_p).

We define the weighted summation of distances on the warping path is

$$\mathbb{D}(i,j) = \mathbb{D}(x_p,y_p) + \frac{1}{k} \sum_{y=j-k+1}^{j} d(i,y) \tag{5.29}$$

Hence, the optimal weighted warping path distance is

$$\mathbb{D}(i,j) = \min \begin{pmatrix} \mathbb{D}(i-1,j) + d(i,j) \\ \mathbb{D}(i-1,j-1) + d(i,j) \\ \mathbb{D}(x_p,y_p) + \frac{1}{k}\sum_{y=j-k+1}^{j} d(i,y) \end{pmatrix} \tag{5.30}$$

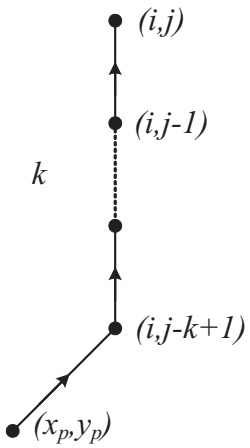


Figure 5.5. k continuous points until the point (i, j) by the j-axis direction

The restricting condition is

$$j - r \leqslant i \leqslant j + r \tag{5.31}$$

The time-normalized distance of optimal warping path between two patterns is

$$D(I,J) = \frac{\mathbb{D}(I,J)}{\mathbb{W}(C)} \tag{5.32}$$

where $\mathbb{W}(C) = I$ or J.

5.4 Itakutra proposed DTW algorithm

Another one is Itakura proposed DTW algorithm. It is different with the Sakoe and Chiba proposed that. The weighting coefficient never be considered. The restriction conditions can be realized as the follow and shown in Fig. 5.6.

1) Monotonic conditions:

$$\begin{aligned} x(l-1) &\leqslant x(l) \\ y(l-1) &\leqslant y(l) \end{aligned} \tag{5.33}$$

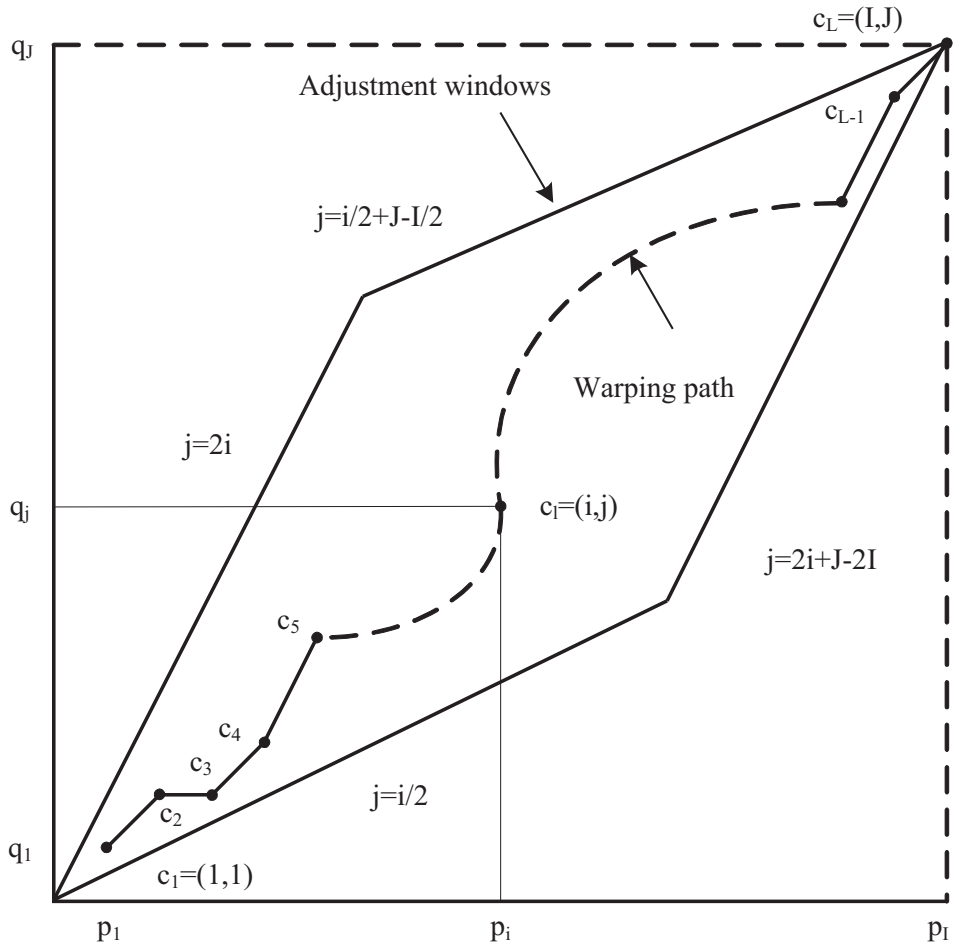


Figure 5.6. Warping function and adjustment window definition for Itakura's DTW algorithms

These monotonic conditions express the characteristics of time sequence of speech signal and voice versa. The precedence order can not be changed after warped sequence.

2) Continuous conditions:

$$x(l) - x(l-1) = 1 \tag{5.34}$$

$$y(l) - y(l-1) = \begin{cases} 0,1,2 & y(l-1) \neq y(l-2) \\ 1,2 & y(l-1) = y(l-2) \end{cases} \tag{5.35}$$

These continuous conditions express how to choice the adjacent frame. By these above two restrictions, we know the pervious adjacent point of c_l is one of $(x(l)-1,y(l))$,$(x(l)-1,y(l)-1)$ and $(x(l)-1,y(l)-2)$.

$$c_{l-1} = \begin{cases} & (x(l)-1,y(l)) \\ & (x(l)-1,y(l)-1) \\ or & (x(l)-1,y(l)-2) \end{cases} \tag{5.36}$$

3) Boundary conditions:

$$\begin{aligned} x(1) &= 1, y(1) = 1 \\ x(L) &= I, y(L) = J \end{aligned} \tag{5.37}$$

The boundary conditions define the beginning point must be point $c_1 = (1,1)$ and the end point is $c_L = (I,J)$ for all pathes. These also express the characteristics of time sequence of speech signal. The beginning and end points of two patterns must be aligned firstly.

4) Adjustment window condition:

$$\begin{aligned} y(l) &= 2x(l) \\ y(l) &= \tfrac{1}{2}x(l) \\ y(l) &= 2x(l) + J - 2I \\ y(l) &= \tfrac{1}{2}x(l) + J - \tfrac{1}{2}I \end{aligned} \tag{5.38}$$

By the continuous conditions Eq. 5.36, we known the slope of warping path is confined between 2 and $1/2$. Thus, the warping function is confined in the parallelogram area, which is constituted with the four straight lines in Eq. 5.38. Those points out of the area can not been calculated. In extreme cases, the i is added 1 then j is added 2,the last point is $J = 2I$ when the slope is 2. Conversely, i is added 2 then j is added 1, the last point is $J = \frac{1}{2}I$ when the slope is $1/2$. Hence, the Itakura proposed DTW algorithm can be realized when $\frac{1}{2}I \leqslant J \leqslant 2I$.

The Itakura proposed DTW algorithm is described as 'DTW3' in the thesis. The calculated processing of optimal path is follow.

DTW3: Itakura's DTW

The weighting coefficient is not used. The initial condition $c_1 = (1,1)$, $c_L = (I,J)$, $d(c_1) = d(1,1)$. Hence,

$$D(1,1) = d(1,1) \tag{5.39}$$

We also consider it with a negative sequence recursive processing. How to calculate the optimal path? We known only three pathes can pass the point (i,j) by Eq. (5.36). They are shown in Fig. 5.7. Then, these path distances $D(i,j)$ from $(1,1)$ to (i,j) are

$$D(i,j) = d(i,j) + \min \begin{pmatrix} D(i-1,j) \\ D(i-1,j-1) \\ D(i-1,j-2) \end{pmatrix} \tag{5.40}$$

Until the last point (I,J), Only one path is remained, then the path is matching optimal path between pattern P and pattern Q and its distance is similarity between them.

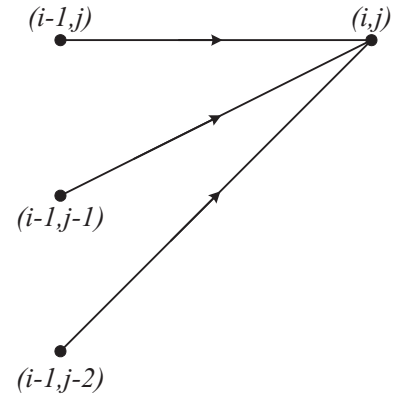


Figure 5.7. Continuous conditions for Itakura's DTW algorithm

5.5 DTW with multireferences

Conventional DTW is capable of fast search and low complexity, but it has poor speech recognition accuracy. In order to improve the recognition accuracy in noisy environments using DTW, a better way is to increase the number of utterances for the same word.

mDTW [15] has been developed. First, we assume there are M reference words, and each word has N speech utterances from difference speakers. The distance computed between the unknown speech waveform and the n^{th} utterance of the m^{th} reference word is denoted as d_{mn}, $1 \leqslant m \leqslant M$, $1 \leqslant n \leqslant N$. The distances computed between the unknown speech waveform and all utterances of the m^{th} reference word are collected in vector $\mathbf{d}_m = [d_{m1}\, d_{m2} \ldots d_{mn} \ldots d_{mN}]^{\mathrm{T}}$. Then, all distances between the unknown speech waveform and all reference utterances can be represented in matrix form as

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^{\mathrm{T}} \\ \mathbf{d}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{d}_M^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,N} \\ d_{2,1} & d_{2,2} & \ldots & d_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \ldots & d_{M,N} \end{bmatrix} \tag{5.41}$$

Sorting the distances for every reference word into ascending order yields $\mathbf{d}'_m$.

$$\mathbf{d}'_m = \begin{bmatrix} d'_{m,1} & d'_{m,2} & \ldots & d'_{m,N} \end{bmatrix} \tag{5.42}$$

That is, $d'_{m,1}$ and $d'_{m,N}$ are the minimum and maximum distances, respectively.

In contrast, in the mDTW approaches, the recognized word corresponds to

$$\underset{m=1:M}{\operatorname{argmin}}\ d'_{m,1}$$

Figure 5.8 shows the recognition accuracy of the mDTW algorithm for different numbers of reference utterances for each word. For this implementation, the reference

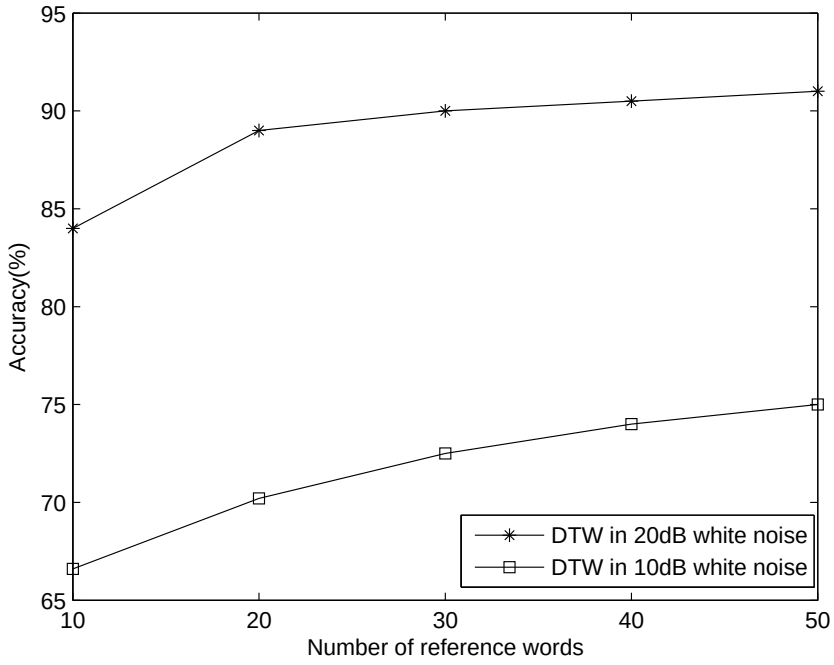


Figure 5.8. Recognition accuracy of mDTW

database consists of 100 isolated Japanese words, and every word has 10 to 50 wave-forms spoken by different persons, and the test words are 50 isolated Japanese words. Note that although accuracy continues to improve with a higher number of reference utterances for each word, calculation com-plexity also increases substantially because of the increasingly large reference database. In the following section, we present a way of finding an appropriate reference utterance to replace the increasing number of utterances, thus reducing the calculation cost while maintaining the high recognition accuracy.

Chapter 6

Reconstruct references DTW algorithm

As stated above, the more utterances we used for the same word, the more memory resources and computing time we need to pay. Therefore, the problem becomes how to find the best reference utterance to replace the large number of reference utterances. Actually, the DTW algorithm provides the optimal path for finding the best reference template. We give a detailed explanation in the following part.

6.1 One pair of vectors

For simplicity, first, we assume one pair of speech feature for the same word, $P = [p(1), p(2), \cdots, p(i), \cdots, p(I)]$ and $Q = [q(1), q(2), \cdots, q(j), \cdots, q(J)]$, as mentioned in Section 3. Then, by using the DTW algorithm, the optimal path between P and Q is defined as

$$C_{opt} = [c_1, c_2, \cdots, c_l, \cdots, c_L] \tag{6.1}$$

where c_l is a point on the $i-j$ plane, the coordinates of which are $(i(l), j(l))$, with the value $(p(i(l)), q(j(l)))$.

The optimal path C_{opt} is the one that minimizes the cumulative error path between P

and Q. In other words, the value of each optimal path point is the cloest value between P and Q. Therefore, let us consider defining a new vector C' to replace P and Q on the basis of the optimal path.

First, we consider the optimal path to represent a function that approximately realizes mapping from the axis of speech feature P onto that of speech feature Q. The slope of every two points in this function is calculated by

$$S = \frac{i(l+1) - i(l)}{j(l+1) - j(l)} \tag{6.2}$$

where S is the slope of two points. Actually, there are only three kinds of slope, as represented in Fig. 6.1.

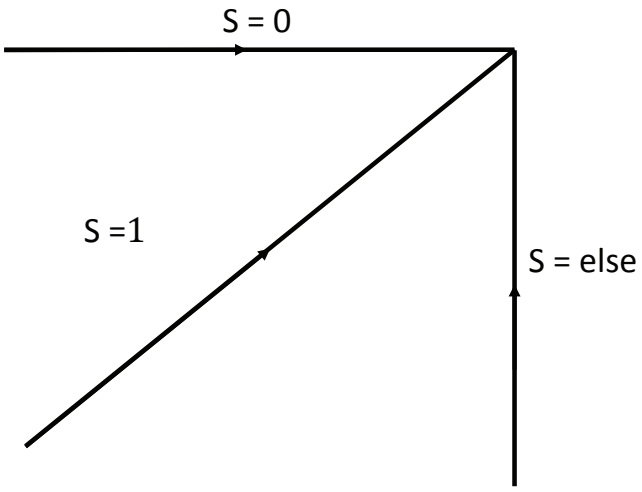


Figure 6.1. Types of slope

Then, every two points other than the starting point and end point will be merged into a new point with the value expressed as

$$c'(l) = \begin{cases} \frac{p(i(l))+p(i(l+1))+q(j(l))}{3} & \text{if S = 0} \\ \frac{p(i(l))+q(j(l))}{2} & \text{if S = 1} \\ \frac{p(i(l))+q(j(l))+q(j(l+1))}{3} & \text{else} \end{cases} \tag{6.3}$$

The starting point and end point remain the original values.

Finally, we define the new vector C' as a set of centroids calculated using Eq. (19).

$$C' = [c'_1, c'_2, \cdots, c'_N] \tag{6.4}$$

Figure 6.2 shows the details of the merging rule.

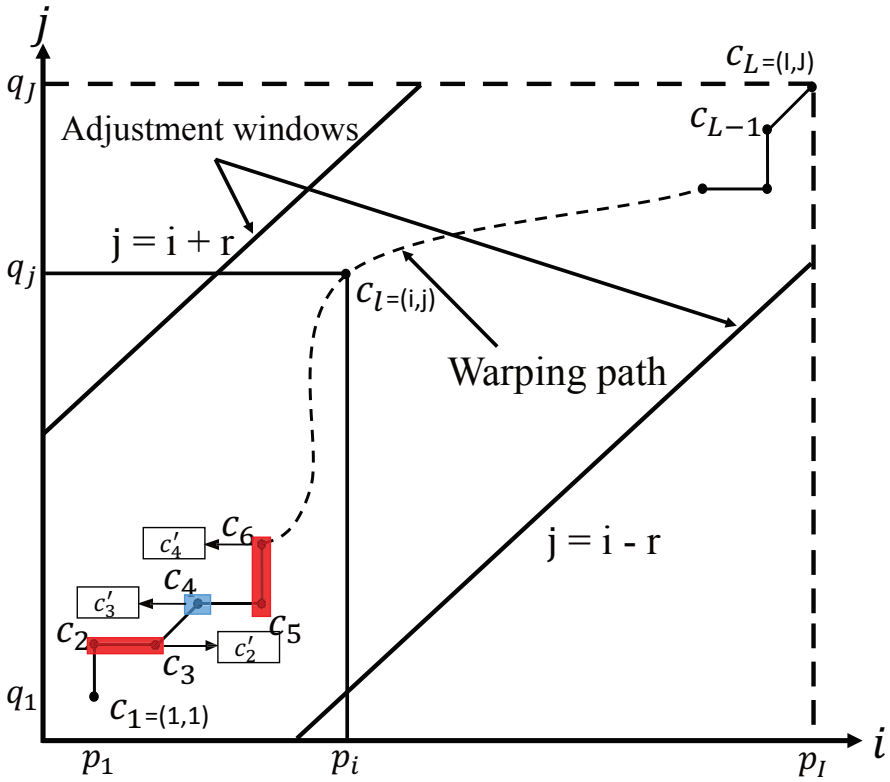


Figure 6.2. Merging rule

6.2 Pairs of vectors

On the basis of the above explanation, we solve the condition in which there are pairs of vectors. The proposed method proceeds in the following steps.

1. We assume there are *M* reference words, where each word has *N* speech utterances

from difference speakers. For each reference word, N speech utterances will be divided into two subsets.

2. For each pair of subsets, the optimal path will be computed. According to Eq. (19), the new vector will replace the pair of subsets. The number of speech utterances will be reduced to $N' = N/2$.

3. If we repeat step 2, the number of speech utterances will be further reduced. In other words, if we repeat step 2 t times (we call it training t times), then the number of speech utterances will be reduced to $\frac{1}{2^t}N$.

4. The distances computed between the unknown speech waveform and all utterances of reference word M are collected in a matrix as

$$\mathbf{D} = \begin{bmatrix} \mathbf{d}_1^{\mathrm{T}} \\ \mathbf{d}_2^{\mathrm{T}} \\ \vdots \\ \mathbf{d}_M^{\mathrm{T}} \end{bmatrix} = \begin{bmatrix} d_{1,1} & d_{1,2} & \dots & d_{1,N'} \\ d_{2,1} & d_{2,2} & \dots & d_{2,N'} \\ \vdots & \vdots & \ddots & \vdots \\ d_{M,1} & d_{M,2} & \dots & d_{M,N'} \end{bmatrix} \tag{6.5}$$

5. As in the mDTW algorithm, sort the distances for each reference word. The recognized word corresponds to

$$\underset{m=1:M}{\operatorname{argmin}}\ d'_{m,1}$$

6. Finally, in the recognition part, the recognition accuracy will be calculated.

Figure 6.3 shows the basic algorithm of three kinds of DTW. The conventional DTW uses the reference speech compared with the test speech. Its algorithm is simple and fast, but the recognition accuracy is low. mDTW uses N reference speeches compared with the test speech. Although the algorithm increases the robustness of the reference speeches and the recognition accuracy is very high, the computation cost is significantly increased. The proposed method not only reduces the computation cost but maintains a high recognition accuracy (Case of training once).

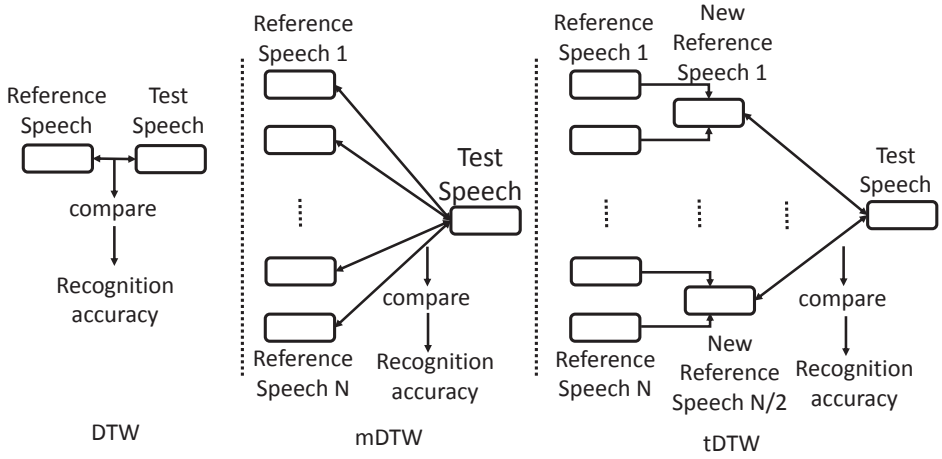


Figure 6.3. Basic algorithms of different DTW methods

6.3 Evaluation measure and results

Conventional recognition systems consist of ordinary feature extraction based on MFCC. The entire recognition system is implemented using MATLAB. The reference database consists of 100 isolated Japanese words, and each word has 100 waveforms spoken by 50 persons. The test words are 50 isolated Japanese words, and each word has 100 waveforms spoken by another 50 persons. MFCC feature vectors are extracted. These vectors comprise 36 dimensions: 12 cepstral coefficients ($s_i(k), i = 1,2,\ldots,12$, k : time index), 12 delta cepstral coefficients ($\Delta s_i(k) = s_i(k) - s_i(k-1)$), and 12 delta-delta cepstral coefficients ($\Delta^2 s_i(k) = \Delta s_i(k) - \Delta s_i(k-1)$). Other conditions are described in Table 6.1.

In this study, we have two main goals. One is to reduce the calculation cost. In the following part, we will show the calculation costs of mDTW and tDTW.

To obtain the calculation cost of mDTW, we must evaluate the following cost:

$$C_{T,i}^{ID}(\mathbb{A}) = MNC_D(H_i, \mathbb{A}) + C_R(\mathbb{A}) \tag{6.6}$$

where $C_{T,i}^{ID}$ is the total calculation cost of mDTW, C_D is the calculation cost of DTW, and

Table 6.1. Experimental settings and parameters

Recognition task	Isolated 100 words
Speech data	100 Japanese region names
Sampling	11.025 kHz, 16 bits
Window length	23.2 ms (256 samples)
Frame length	11.6 ms (128 samples)
Band of bandpass filter	1-15 Hz
Feature vector	36-dimensional MFCC
Noise type	white noise and babble noise

C_R is the calculation cost of noise reduction. M is the total number of target words, and N is the total number of speeches for each speech word (in the experiment, M is 100 and N is 100). We define $\mathbb{A}$ as a feature vector of speech and H_i as the i^{th} reference feature vector.

In the case of tDTW-based ASR, the total calculation cost is

$$C_{T,i}^{TD}(\mathbb{A}) = C_T(H_i,\mathbb{A}) + \frac{1}{2t}MNC_D(H_i,\mathbb{A}) + C_R(\mathbb{A}) \tag{6.7}$$

where $C_{T,i}^{TD}$ is the total calculation cost of tDTW, C_T is the calculation cost of the training part, and t is number of training repetitions. We assume training of only once, then, $C_T(H_i,\mathbb{A})$ can be expressed as

$$C_T(H_i,\mathbb{A}) = \frac{1}{2}MNC_D(H_i,\mathbb{A}) \tag{6.8}$$

Since $\frac{1}{2}MNC_D(H_i,\mathbb{A})$ is M times $\frac{1}{2}NC_D(H_i,\mathbb{A})$, in other words, $C_T(H_i,\mathbb{A}) \ll MNC_D(H_i,\mathbb{A})$, then $C_{T,i}^{TD}(\mathbb{A}) \approx \frac{1}{2}C_{T,i}^{ID}(\mathbb{A})$. Apparently, the calculation cost of mDTW has been reduced;

after training only once, the calculation cost has been reduced by almost 50%.

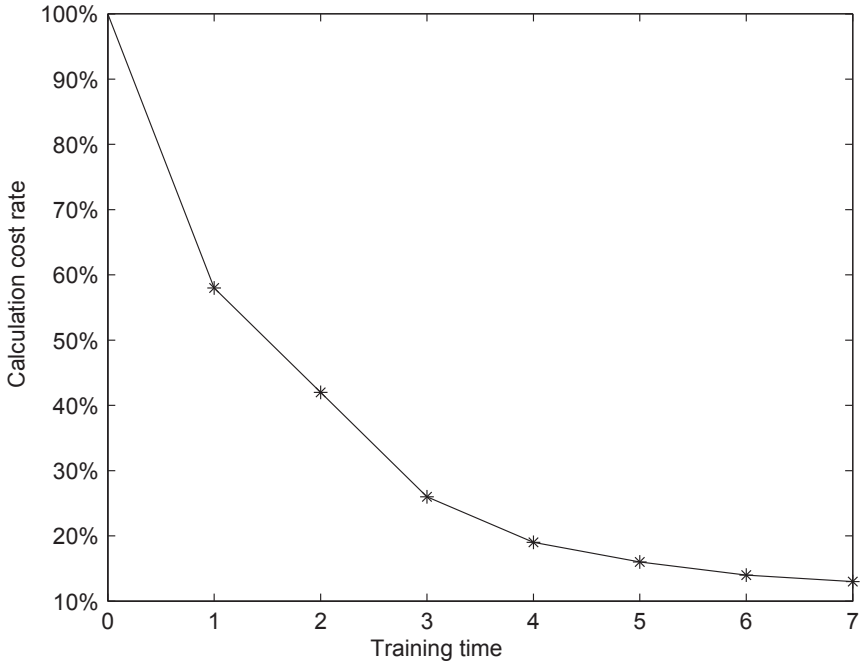


Figure 6.4. Computing time of proposed DTW

Figure 6.4 shows the practical calculation cost rate. We used the Epson Pro7500 computer with the Core(TM) i7-3820 CPU @ 3.6 GHz. Note that zero training time represents the mDTW calculation cost, and all the calculation cost rate were compared with the mDTW calculation cost. Apparently, after training once, computing time has been reduced 41.6%. On the other hand, when the numbers of reference words becomes half, the computing time is significantly reduced.

Our other goal is to maintain a high recognition accuracy. Figure 6.5 shows the recognition accuracy of the two DTW algorithms with 10 dB and 20 dB white and babble noise. Our approach yields 96.94% accuracy compared with the 97.54% accuracy of mDTW in 20 dB white noise and 84.4% accuracy compared with 86.44% accuracy of mDTW in 10 dB white noise. Our approach yields 94.12% accuracy compared with

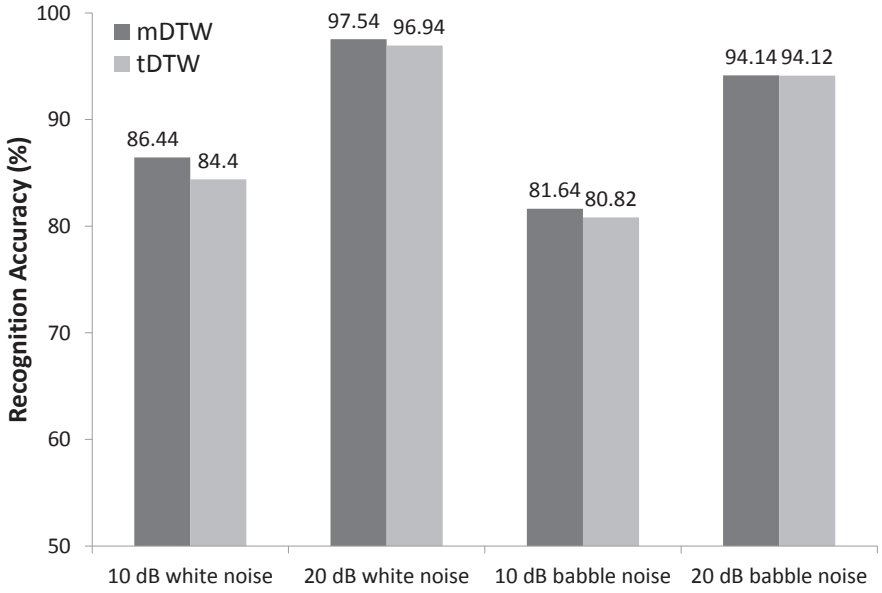


Figure 6.5. Recognition accuracy of tDTW algorithms with 10 dB and 20 dB white and babble noise

94.14% accuracy of mDTW in 20 dB babble noise and 80.82% accuracy compared with 81.64% accuracy of in 10 dB babble noise (case of training once).

Furthermore, Fig. 6.6 shows the tDTW recognition accuracy when the reference utterances have been trained more than once in 10 dB and 20 dB white and babble noise.

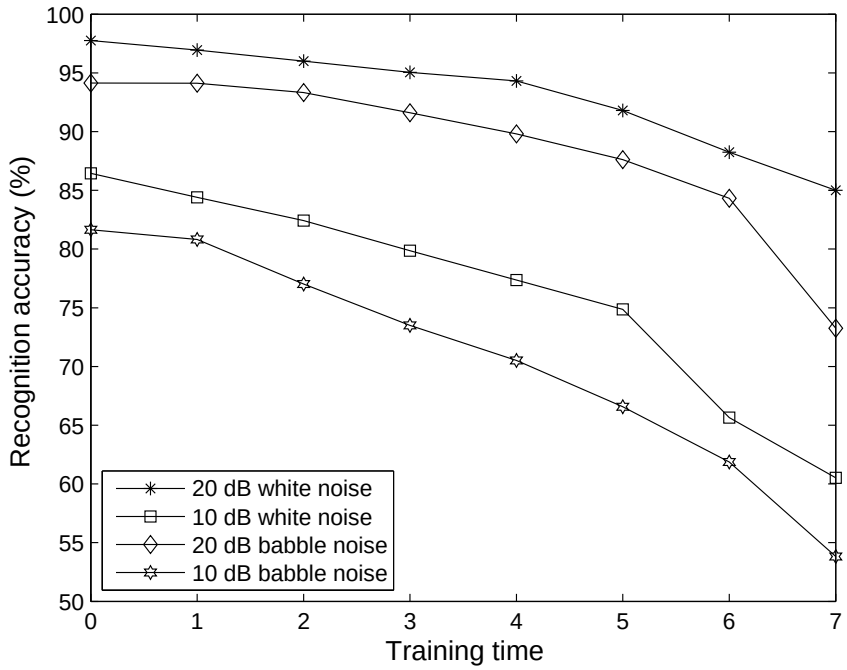


Figure 6.6. Recognition accuracy of proposed DTW

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this paper we provide an account of the results of this work. In this work we have proposed a new robust ASR technique that exploits VAD, noise-reduction, and DTW-based processing. We have found that the calculation cost of mDTW has been reduced 41.6% and recognition accuracy of the proposed method is similar to that of the mDTW.

Chapter 3 the importance of automatic voice activity detection (VAD) has been discussed. In particular, under noise circumstances, it has been quite difficult to design the automatic voice activity detection with a speech recognition system. The basic concept about VAD and its current techniques have been discussed in this chapter.

Chapter 4 introduces current noise reduction technologies used into speech processing. Among them, RASTA, CMS, and RSF/DRA are explained in this chapter.

Chapter 6 has proposed new techniques using DTW, VAD, CMA and RSF/DRA. It can realizes noise robust mechanism, robust automatic VAD and high speech recognition accuracy. In addition, the proposed method can reduce the total calculation cost drastically compared with other methods whose recognition accuracy is almost the same.

7.2 Future work

Although we proposed method has improved the performance of ASR system with DTW algorithm, the recognition accuracy is not so high in low SNR. The real environment is complicated and volatile, we must try to improve the recognition accuracy of ASR system in order to practical application.

The VAD algorithm need to be modified and thus improved accuracy of endpoint detection. Although the modified VAD method with short-time energy and ZCR minimizes this effect of noise pulse, our method is limited to detect endpoint in low SNR. Hence, we must try to research and explore new technology in order to detect endpoint accurately in low SNR.

Since the number of reference words for the same word decrease, the recognition accuracy also reduced. our future work will attempt to find the best compromises between accuracy and complexity.

Bibliography

[1] A. Acero and R. Stern, "Environmental robustness in automatic speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, no. 2, Apr. 1990, pp. 849–852.

[2] S. Al-Haddad, S. Samad, A. Hussain, K. Ishak, and H. Mirvaziri, "Decision fusion for isolated Malay digit recognition using dynamic time warping (DTW) and hidden Markov model (HMM)," in *Proc. The* 5*th Student Conference on Research and Development*, Dec. 2007, pp. 1–6.

[3] B. Atal and L. Rabiner, "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 3, pp. 201–212, Jun. 1976.

[4] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, Jun. 1974.

[5] B. Atal, "Automatic recognition of speakers from their voices," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 460–475, Apr. 1976.

[6] A. Berstein and I. Shallom, “Noise processing DTW algorithms for speech recognition systems,” in *Proc. The* 17*th Convention of Electrical and Electronics Engineers in Israel*, Mar. 1991, pp. 293–296.

[7] Z.-H. Chen, Y.-F. Liao, and Y.-T. Juang, “Eigen-prosody analysis for robust speaker recognition under mismatch handset environment,” *Electronics Letters*, vol. 40, no. 19, pp. 1233–1235, Sep. 2004.

[8] D. G. Childers, M. Hahn, and J. N. Larar, “Silent and voiced/unvoiced/mixed excitation (four-way) classification of speech,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1771–1774, Nov. 1989.

[9] W. Chou and B.-H. Juang, *Pattern recognition in speech and language processing*. CRC Press, 2003.

[10] I. Chung and I.-J. Chung, “Memory efficient and fast speech recognition system for lowresource mobile devices,” *IEEE Transactions on Consumer Electronics*, vol. 52, no. 3, pp. 792–796, Aug. 2006.

[11] A. Cohen and D. Graupe, “Speech recognition and control system for the severely disabled,” *Journal of Biomedical Engineering*, vol. 2, no. 2, pp. 97–107, Apr. 1980.

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

[13] J. de Veth and L. Boves, “Channel normalization techniques for automatic speech recognition over the telephone,” *Speech Communication*, vol. 25, no. 1, pp. 149–164, 1998.

[14] P. Ding, L. He, X. Yan, R. Zhao, and J. Hao, “Robust mandarin speech recognition in car environments for embedded navigation system,” *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 584–590, May 2008.

[15] N. Erdol, C. Castelluccia, and A. Zilouchian, “Recovery of missing speech packets using the short-time energy and zero-crossing measurements,” *IEEE Transactions on Speech and Audio Processing*, vol. 1, no. 3, pp. 295–303, Jul. 1993.

[16] P. Foster and T. Schalk, *Speech Recognition The Complete Practical Reference Guide*, 1st ed. CMP Books, 1993.

[17] M. Fried-Oken, “Voice recognition device as a computer interface for motor and speech impaired people,” *Journal of Biomedical Engineering*, vol. 66, no. 10, pp. 678–681, Oct. 1985.

[18] K. Fujioka and Y. Miyanaga, “A new noise reduction method of speech signal with running spectrum filtering,” in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems*, Nov. 2004, pp. 173–176.

[19] C. Gan and R. Donaldson, “Adaptive silence deletion for speech storage and voice mail applications,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 36, no. 6, pp. 924–927, Jun. 1988.

[20] Y. Gong, “Speech recognition in noisy environments: A survey,” *Speech Communication*, vol. 16, no. 3, pp. 261–291, Apr. 1995.

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

[22] N. Hayasaka and Y. Miyanaga, “Spectrum filtering with FRM for robust speech recognition,” in *Proc. IEEE International Symposium on Circuits and Systems*, Nov. 2006, pp. 3285–3288.

[23] N. Hayasaka, S. Yoshizawa, N. Wada, Y. Miyanaga, and N. Hataoka, “A study of robust speech recognition system and its LSI design,” *The Society of Instrument and Control Engineers*, vol. 41, no. 5, pp. 473–480, May 2005.

[24] N. Hayasaka, K. Khankhavivone, Y. Miyanaga, and K. Songwatana, “New robust speech recognition by using nonlinear running spectrum filter,” in *Proc. International Symposium on Communications and Information Technologies*, Oct. 2006, pp. 133–136.

[25] N. Hayasaka and Y. Miyanaga, “A study of robust speech recognitoin sing FRM filter,” in *Proc. IEEE Region 10 Conference (TENCON)*, Nov. 2004, pp. 80–83.

[26] N. Hayasaka, Y. Miyanaga, and N. Wada, “Running spectrum filtering in speech recognition,” in *Proc. International Conference on Soft Computing and Intelligent Systems (SCIS)*, Oct. 2002, pp. 154–157.

[27] N. Hayasaka, N. Wada, S. Yoshizawa, and Y. Miyanaga, “A robust speech recognition system using FRM running spectrum filtering,” in *Proc. International Symposium on Control, Communications and Signal Processing (ISCCSP)*, Mar. 2004, pp. 401–404.

[28] H. Hermansky and J. Cox, L.A., “Perceptual linear predictive (PLP) analysis-resynthesis technique,” in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1991, pp. 37–38.

[29] H. Hermansky and N. Morgan, “RASTA processing of speech,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, Oct. 1994.

[30] H. Hermansky, N. Morgan, and H.-G. Hirsch, “Recognition of speech in additive and convolutional noise based on RASTA spectral processing,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, Apr. 1993, pp. 83–86.

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

[32] M. Herscher and R. Cox, “Voice programming of numerically controlled machines,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, May 1977, pp. 452–455.

[33] M. Holmberg, D. Gelbart, and W. Hemmert, “Automatic speech recognition with an adaptation model motivated by auditory processing,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 43–49, Jan. 2006.

[34] G. Hongbin, P. Weiyi, H. Chunru, and Z. Yongqiang, “A speech endpoint detection based on dynamically updated threshold of box-counting dimension,” in *Proc. International Forum on Information Technology and Applications*, vol. 2, May. 2009, pp. 397–401.

[35] X. D. Huang, Y. Ariki, and M. A. Jack, *Hidden Markov Models for Speech Recognition*, 1st ed. Edinburgh University Press, 1990.

[36] X. Huang, A. Acero, A. Acero, and H.-W. Hon, *Spoken language processing: a guide to theory, algorithm, and system development*. Upper Saddle River, NJ: Prentice Hall, 2001.

[37] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 23, no. 1, pp. 67–72, Feb. 1975.

[38] ——, "Line spectrum representation of linear predictive coefficients of speech signals," *Journal of the Acoustical Society of America*, vol. 57, no. S1, p. S35, Mar. 2000.

[39] Y. Jie and W. Zhenli, "Noise robust speech recognition by combining speech enhancement in the wavelet domain and lin-log RASTA," in *Proc. International Colloquium on Computing, Communication, Control, and Management*, vol. 2, Aug. 2009, pp. 415–418.

[40] B. H. Juang, "On the hidden markov model and dynamic time warping for speech recognition—-a unified view," *AT&T Technical journal*, vol. 63, no. 7, pp. 1213–1243, 1984.

[41] B. H. Juang, L. R. Rabiner, and J. G. Wilpon, "On the use of bandpass liftering in speech recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 7, pp. 947–954, Jul. 1987.

[42] S. B. Junior, R. C. Guido, S.-H. Chen, L. S. Vieira, and F. L. Sanchez, "Improved dynamic time warping based on the discrete wavelet transform," in *Proc. The* 9*th IEEE International Symposium on Multimedia Workshops*, Dec. 2007, pp. 256–263.

[43] J. C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 3, pp. 406–412, Jul. 1994.

[44] J.-C. Junqua and J. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*, 1st ed. Kluwer Academic, 1995.

[45] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. Prentice Hall Inc., 1996.

[46] ——, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2009.

[47] K. Khankhavivone, N. Hayasaka, Y. Miyanaga, and K. Songwatana, "A low cost running spectrum filter for speech recognition using modified frequency response masking technique," *Journal of Signal Processing*, vol. 11, no. 2, pp. 227–236, May 2007.

[48] C. Kim and K. deok Seo, "Robust DTW-based recognition algorithm for hand-held consumer devices," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 2, pp. 699–709, May 2005.

[49] H. K. Kim, S. H. Choi, and H. S. Lee, "On approximating line spectral frequencies to LPC cepstral coefficients," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 2, pp. 195–199, Mar. 2000.

[50] D. H. Klatt and K. N. Stevens, "On the automatic recognition of continuous speech: Implications from a spectrogram-reading experiment," in *Proc. IEEE Trans. Audio Electroacoust.*, vol. AU-16, June 1973, pp. 210–217.

[51] S. Kwong, C. Chau, and W. Halang, "Genetic algorithm for optimizing the non-linear time alignment of automatic speech recognition systems," *IEEE Transactions on Industrial Electronics*, vol. 43, no. 5, pp. 559–566, Oct. 1996.

[52] Y.-K. Lau and C.-K. Chan, "Speech recognition based on zero crossing rate and energy," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 1, pp. 320–323, Feb. 1985.

[53] C. H. Lee, F. K. Soong, and K. K. Paliwal, *Automatic Speech and Speaker Recognition*. Kluwer Academic Publisher, 1996.

[54] S. E. LeVinson, L. R. Rabiner, and M. M. Sondhi, "An introduction to the application of the theory of probabilistic functions of a markov process to automatic speech recognition," *Bell System Technical journal*, vol. 62, no. 4, pp. 1035–1074, 1983.

[55] Z. Lu, B. Liu, and L. Shen, "Speech endpoint detection in strong noisy environment based on the Hilbert-Huang transform," in *Proc. International Conference on Mechatronics and Automation*, Aug. 2009, pp. 4322–4326.

[56] T. Martin, "Practical applications of voice input to machines," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 487–501, Apr. 1976.

[57] J. Ming, "Noise compensation for speech recognition with arbitrary additive noise," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 833–844, May 2006.

[58] D. Naik, "Pole-filtered cepstral mean subtraction," in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, no. 1, May 1995, pp. 157–160.

[59] Y. Oh, J. Yoon, J. Park, M. Kim, and H. Kim, "A name recognition based call-and-come service for home robots," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 2, pp. 247–253, May 2008.

[60] N. Ohtsuki, Q. Zhu, and Y. Miyanaga, "The effect of the musical noise suppression in speech noise reduction using RSF," in *Proc. International Symposium on Communications and Information Technologies*, vol. 2, Oct. 2004, pp. 663–667.

[61] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing*, 2nd ed. Prentice-Hall Inc., 1998.

[62] K. K. Paliwal, *Automatic speech and speaker recognition: advanced topics*, 1st ed. Springer, 1996.

[63] C. Panagiotakis and G. Tziritas, "A speech/music discriminator based on RMS and zero-crossings," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 155–166, Feb. 2005.

[64] R. Pieraccini, "Pattern compression in isolated word recognition," *Signal Processing*, vol. 7, no. 1, pp. 1–15, Sep. 1984.

[65] P. Pujol, D. Macho, and C. Nadeu, "On real-time mean-and-variance normalization of speech recognition features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, May 2006, pp. I773–I776.

[66] L. Rabiner and S. Levinson, "Isolated and connected word recognition–theory and selected applications," *IEEE Transactions on Communications*, vol. 29, no. 5, pp. 621–659, May 1981.

[67] L. R. Rabiner, "A tutorial on hidden markov models and sclected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[68] L. R. Rabiner and M. R. Sambur, “An algorithm for determining the endpoints for isolated utterances,” *The Bell System Technical Journal*, vol. 54, no. 2, pp. 297–315, 1975.

[69] L. R. Rabiner and R. W. Schafer, “Introduction to digital speech processing,” *Foundations and Trends in Signal Processing*, vol. 1, no. 1-2, pp. 1–194, 2007.

[70] L. Rabiner and B.-H. Juang, *Fundamentals of speech recognition*, 1st ed. Upper Saddle River, New Jersey, USA: Prentice Hall PTR, 1993.

[71] L. R. Rabiner and R. W. Schafer, *Theory and Application of Digital Speech Processing*. Prentice-Hall Inc., 2009.

[72] L. Rabiner, “Applications of voice processing to telecommunications,” *Proceedings of the IEEE*, vol. 82, no. 2, pp. 199–228, Feb. 1994.

[73] L. Rabiner and R.W.Schafer, *Digital Processing of Speech Signals*, 1st ed. Upper Saddle River, Prentice HallUSA: Rainbow-Bridge Book Company PTR, 1978.

[74] M. Rahim, B.-H. Juang, W. Chou, and E. Buhrke, “Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification,” *Journal of the Acoustical Society of America*, vol. 55, no. 6, pp. 1304–1312, 1974.

[75] ——, “Signal conditioning techniques for robust speech recognition,” *IEEE Signal Processing Letters*, vol. 3, no. 4, pp. 107–109, Apr. 1996.

[76] B. Raj, V. Parikh, and R. Stern, “The effects of background music on speech recognition accuracy,” in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, no. 2, Apr 1997, pp. 851–854.

[77] M. Rashwan and M. Fahmy, “New technique for speaker-independent isolated-word recognition,” *IEEE Proceedings-F on Communications, Radar and Signal Processing*, vol. 135, no. 3, pp. 251–257, Jun. 1988.

[78] H. Sakoe, “Two-level DP-matching–A dynamic programming-based pattern matching algorithm for connected word recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 27, no. 6, pp. 588–595, 1979.

[79] H. Sakoe and S. Chiba, “A similarity evaluation of speech patterns by dynamic programming,” *Dig. 1970 Nat. Meeting, Inst. Electron. Comm. Eng.*, p. 136, Jul. 1970.

[80] ——, “Comparative study of DP-pattern matching techniques for speech recognition,” *Tech. Group Meeting Speech, Acoust. Soc.*, pp. S73–22, Dec. 1973.

[81] ——, “Dynamic programming algorithm optimization for spoken word recognition,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[82] S. Seneff, “Real-time harmonic pitch detector,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 26, no. 4, pp. 358–365, Aug. 1978.

[83] N. Shabtai, Y. Zigel, and B. Rafaely, “The effect of GMM order and CMS on speaker recognition with reverberant speech,” in *Proc. Hands-Free Speech Communication and Microphone Arrays*, May 2008, pp. 144–147.

[84] O. Viikki and K. Laurila, “Noise robust HMM-based speech recognition using segmental cepstral feature vector normalization,” in *Proc. Robust Speech Recognition for Unkown Communication Channels Pont-à-Mousson*, Apr. 1997, pp. 107–110.

[85] ——, “Cepstral domain segmental feature vector normalization for noise robust speech recognition,” *Speech Communication*, vol. 25, pp. 133–147, Aug. 1998.

[86] T. K. Vintsyuk, “Speech discrimination by dynamic programming,” *Cybernetics and systems analysis*, vol. 4, no. 1, pp. 81–88, 1968.

[87] N. Wada, N. Hayasaka, S. Yoshizawa, and Y. Miyanaga, “Robust speech recognition with feature extraction using combined method of RSF and DRA,” in *Proc. International Symposium on Communications and Information Technologies*, vol. 2, Oct. 2004, pp. 1001–1004.

[88] N. Wada, N. Hayasaka, N. Hataoka, and Y. Miyanaga, “Noise robust speech detection/recognition system including RSF/DRA and MFCC,” in *Proc. International Symposium on Communications and Information Technologies (ISCIT)*, vol. 1, Sep. 2003, pp. 455–458.

[89] N. Wada, S. Yoshizawa, N. Hayasaka, and YoshikazuMiyanaga, “Robust speech feature extraction using RSF/DRA and burst noise skipping,” *Transactionson Electrical Engineering, Electronics, and Communications (ECTI-EEC)*, vol. 3, no. 2, pp. 100–107, Aug. 2005.

[90] C. Weinstein, “Opportunities for advanced speech processing in military computer-based systems,” *Proceedings of the IEEE*, vol. 79, no. 11, pp. 1626–1641, Nov. 1991.

[91] G. White, “Speech recognition: A tutorial overview,” *Computer*, vol. 9, no. 5, pp. 40–53, May 1976.

[92] G. Xu, B. Tong, and X. He, “Robust endpoint detection in Mandarin based on MFCC and short-time correlation coefficient,” in *Proc. International Conference*

on Intelligent Computation Technology and Automation, vol. 2, Oct. 2009, pp. 336–339.

[93] K. Yamamoto, F. Jabloun, K. Reinhard, and A. Kawamura, "Robust endpoint detection for speech recognition based on discriminative feature extraction," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, May. 2006, pp. 805–808.

[94] I.-C. Yoo and D. Yook, "Automatic sound recognition for the hearing impaired," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 2029–2036, Nov. 2008.

[95] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, "Noise robust speech recognition focusing on time variation and dynamic range of speech feature parameters," in *Proc. International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, Dec. 2003, pp. 484–487.

[96] S. Yoshizawa, N. Hayasaka, N. Wada, and Y. Miyanaga, "VLSI architecture for robust speech recognition systems and its implementation in verification platform," *Journal of Robotics and Mechatronics*, vol. 17, no. 4, pp. 447–455, Aug. 2005.

[97] S. Yoshizawa, A. Kageyama, N. Hayasaka, and Y. Miyanaga, "Development of a dedicated hardware system for noise robust speech recognition using RSF/DRA technique," in *Proc. International Symposium on Communications and Information Technologies (ISCIT)*, vol. 1, Sep. 2003, pp. 463–466.

[98] S. Yoshizawa, N. Wada, N. Hayasaka, and Y. Miyanaga, "Scalable architecture for word HMM-based speech recognition and VLSI implementation in complete

system," *IEEE Transactions on Circuit and Systems-I*, vol. 53, no. 1, pp. 70–77, Jan. 2006.

[99] K. Yu, J. Mason, and J. Oglesby, "Speaker recognition using hidden Markov models, dynamic time warping and vector quantisation," *IEE Proceedings-F on Vision, Image and Signal Processing*, vol. 142, no. 5, pp. 313–318, Oct. 1995.

[100] K.-H. Yuo and H.-C. Wang, "Robust features derived from temporal trajectory filtering for speech recognition under the corruption of additive and convolutional noises," in *Proc. The 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 1, May 1998, pp. 577–580.

[101] T. Zhang and C.-C. Kuo, "Audio content analysis for online audiovisual data segmentation and classification," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 441–457, May 2001.

[102] Y. Zhao, "Frequency-domain maximum likelihood estimation for automatic speech recognition in additive and convolutive noises," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 3, pp. 255–266, May 2000.

[103] Q. Zhu, N. Ohtsuki, Y. Miyanaga, and N. Yoshida, "Robust speech analysis in noisy environment using running spectrum filtering," in *Proc. International Symposium on Communications and Information Technologies*, vol. 2, Oct. 2004, pp. 995–1000.

[104] E. Zwicker, "Masking and psychological excitation as consequences of the ear's frequency analysis," in *Proc. Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and GF Smoorenburg (Sijthoff, Leiden), 1970, pp. 376–394.

List of Publications

Peer-reviewed Journal

[1] X. Sun, Y. Miyanaga and B. Sai, “Dynamic Time Warping for Speech Recognition with Training Part to Reduce the Computation”, *Journal of signal processing (JSP)*, Vol.18, No.2, pp.89-96, March 2014.

Peer-reviewed International Conference

[1] X. Sun, Y. Miyanaga and S. Yoshizawa, “A speech recognition system with microphone array”, *International Symposium on Multimedia and Communication Technology (ISMAC)*, Manila, Philippines, September 8-9, pp.165-168, 2010.

[2] X. Sun and Y. Miyanaga, “Reduce the Computing Time Using Dynamic Time Warping (DTW) for Speech Recognition”, *International Electrical Engineering Congress (IEECON)*, Chiang Mai, Thailand, March 13-15, 2013.

[3] X. Sun and Y. Miyanaga, “Dynamic time warping for speech recognition with training part to reduce the computation”, *International Symposium on Signals, Circuits and Systems. (ISSCS 2013)*, Iasi, Romania, July 11-12, 2013, DOI 10.1109/ISSCS.2013.6651195 pp 1-4.

Technical Report

[1] X. Sun and Y. Miyanaga, “Efficiency improvement in dynamic time warping algorithms for isolated word recognition”, *Smart Info-Media System (SIS)*, Kagojima, Japanese, June 13-14, vol. 113, no. 78, SIS2013-13, pp. 65-68, 2013.