A Study on Automatic Speech Segmentation Method Based on Human Perception Characteristics

By Keyi Zhou

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

September, 2013

A Study on Automatic Speech Segmentation Method Based on Human Perception Characteristics

By Keyi Zhou (1110204)

A thesis submitted to
School of Information Science,
Japan Advanced Institute of Science and Technology,
in partial fulfillment of the requirements
for the degree of
Master of Information Science
Graduate Program in Information Science

Written under the direction of
Professor Masato Akagi

and approved by
Professor Masato Akagi
Professor Jianwu Dang
Associate Professor Masashi Unoki

August, 2013 (Submitted)

Contents

1 Introduction **5**
- 1.1 Motivation . . . 5
- 1.2 Background . . . 5
- 1.3 Problem definition . . . 7
  - 1.3.1 Human perception characteristics . . . 7
  - 1.3.2 Principle of HMM-based segmentation . . . 8
  - 1.3.3 Drawback of HMM-based segmentation . . . 9
- 1.4 Purpose of this research . . . 10
- 1.5 Thesis structure . . . 11

2 An outline of the proposed method **13**
- 2.1 Conceptual idea for determining the phoneme boundaries . . . 13
- 2.2 An outline for the proposed method . . . 14

3 Databases **16**
- 3.1 TIMIT database . . . 16
- 3.2 ATR database . . . 16

4 Spectrum target prediction model **18**
- 4.1 Basic concept . . . 18
- 4.2 Preliminaries . . . 19
  - 4.2.1 Parameters for speech analysis . . . 19
  - 4.2.2 Differentiation for time derivative $\dot{f}_n$ of f_n . . . 20
- 4.3 Prediction method . . . 20
- 4.4 Simulated results . . . 22
- 4.5 Discussion and conclusion . . . 22

5 Experiments for segmentation **23**
- 5.1 Description for experiments . . . 23
- 5.2 Results of the experiments . . . 23
- 5.3 Discussion . . . 24
- 5.4 Conclusion . . . 26

6 Implementation for the proposed automatic segmentation method 27
6.1 Strategy for extracting the phoneme boundaries from all candidates 27
6.1.1 An outline for the strategy 27
6.1.2 Implementation of the strategy 28
6.2 Labelling methods 31
6.3 Experiments for labelling 31
6.3.1 Preprocessing for using TIMIT database 33
6.3.2 Description for experiments 33
6.3.3 Results of experiments 36
6.4 Conclusion 36

7 Objective Evaluation 37
7.1 Analysis for the HMM-based segmentation results 37
7.2 Evaluation method 38
7.3 Theoretical upper-limitation evaluation results 39
7.4 Experimental evaluation results for the proposed method 40
7.5 Discussion for evaluation results 41
7.6 Conclusion 42

8 Conclusion and future work 43

List of Figures

1.1 Concatenative speech synthesis based on the unit-selection technique . . . 6
1.2 Trajectory of spectra features (/aia/) . . . 8
1.3 The uaually used HMM for a phoneme . . . 9
1.4 A composite HMM for the continuous speech /aia/ . . . 9
1.5 HMM-based Segmentation for the speech /aia/ . . . 10
1.6 Ideal results by the proposed automatic speech segmentation method . . . 11

2.1 Conceptual idea for determining the phoneme boundaries . . . 14
2.2 Scheme of the proposed method . . . 15

4.1 Block diagram for solution of the spectrum target prediction model . . . 21
4.2 Simulated results for the spectrum target prediction model . . . 22

5.1 Results of the experiments for segmentation . . . 25

6.1 A sample labelling file for HMM-based segmentation . . . 28
6.2 A sample file to describe the mean errors for each boundary . . . 29
6.3 A sample labelling file for hand-made segmentation . . . 29
6.4 Flow-process diagram for obtaining the error list file . . . 30
6.5 Block diagram for the proposed method . . . 32
6.6 The block diagram for the experiments using TIMIT database . . . 35
6.7 An example labelling file obtained by the proposed method . . . 36

7.1 The distribution for the numbers of boundaries for the same error in HMM-based segmentation results . . . 38
7.2 Analysis for HMM-based segmentation results . . . 39
7.3 Ideal results for the proposed method . . . 40

List of Tables

2.1 Comparison for spectrum target prediction model and HMM 15

5.1 Parameters for obtaining the acoustic features (MFCC) 24

6.1 The phoneme set (61 phonemes) 34
6.2 Parameters for extracting MFCC features in HTK 34

7.1 Theoretical upper-limitation evaluation results 40
7.2 Evaluation results for the proposed method 41

Chapter 1

Introduction

1.1 Motivation

Speech segmentation is very helpful in many applications, such as speech summarization[1], video summarization[2], speech document indexing and retrieval[3]. Besides, it is a vitally important task in several audio processing applications like speaker diarization[4], speaker tracking[5], and automatic speech recognition (ASR)[6]. The requirements on the segmentation differ depending on the application.

There have been more and more applications with the requirement for highly accurate and reliable speech segmentation. Given a specific example, based on the unit-selection technique, concatenative speech synthesis has become an essential approach to text-to-speech (TTS) systems[7–10]. As seen in Fig. 1.1, using this technique, a sequence of pre-processed speech segments (in other words, "unit") are selected from a large speech corpus, and then concatenated sequentially. These segments or units present the given phonetic and prosodic descriptions when segmentation are carried out. Obviously, it is fair to say that the quality of synthesized speech by this technique is dependent on the quality of the speech corpus, to a large extent.

Traditionally, segmentation by professional human beings has been considered the most reliable and precise method to get the segments for a variety of TTS applications. However, it will be time-consuming and labor-intensive, especially when the required size of the speech database is huge. Therefore, an appropriate automatic method for segmentation is more feasible and practical.

1.2 Background

In the literature, some methods for automatic speech segmentation have been proposed based on Hidden Markov Model (HMM)[11]. In the HMM-based framework, each phone unit is modelled by a context-dependent or context-independent HMM. The model parameters are trained based on a collection of speech data with the corresponding transcripts and then the trained HMMs are used to align a speech signal along the associated transcripts by means of Viterbi decoding[12]. For segmentation, these methods divide

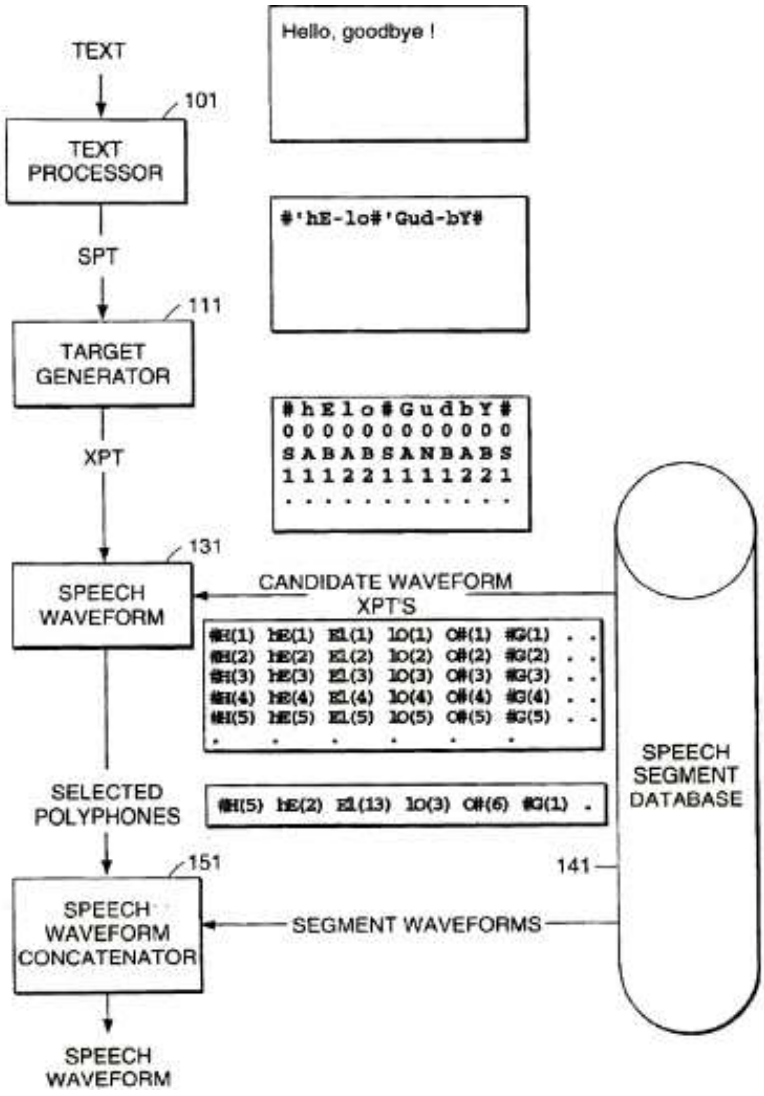

Figure 1.1: Concatenative speech synthesis based on the unit-selection technique.

continuous speech into segments (e.g. phonemes) based on the variety of pattern distance (e.g. likelihood in HMM). According to the reported segmentation accuracy[13,14], the results are good enough for some transitions, however, the HMMs do not perform well for other transitions and make similar error patterns depending on the transitions[11,15,16].

Although these methods are certified useful to train model parameters utilized for speech recognition and automatic speech segmentation can be implemented very efficiently using the BaumWelch algorithm and Viterbi algorithm[17], it is suspected for highly natural speech synthesis or some other application with the requirement for highly accurate and reliable speech segmentation, by using automatic segmented speech databases. The reason is that segmentation at points-in-time when output of the HMM changes to a following phoneme, according to Maximum Likelihood (ML) model, makes imprecise boundaries for the segments compared to professional manual segmentation. The detail will be given with a specific example in the next section.

Some methods have been proposed to solve the problem mentioned above, by refining the initial HMM-based segmentations. For instance, energy changes are used in different frequency bands for boundary correction[18], neural network are employed to refine phone boundaries[19,20], and support vector machine (SVM) classifiers are trained to differentiate boundaries from non-boundary positions[21]. In recent years, the method with several modifications has been proposed to an HMM-based system, including the use of energy-based features and distinctive phonetic features, and the use of observation-dependent state transition probabilities[22]. Besides, the use of phone boundary models is investigated for forced alignment within the HMM framework, treating phones and phone boundaries as independent HMMs and determining a boundary by the alignment of its own state with frames[23]. Some research has achieved great improvement for standard HMM-based segmentation[22,23]. However, considering the principle of the HMM-based segmentation, there has been some problems remained to be solved for further research. One of the problems is that HMM is a probability model without considering human perception characteristics[24–26]. Based on the human perception characteristics, it may be helpful to obtain more accurate boundaries for the segments.

1.3 Problem definition

1.3.1 Human perception characteristics

Analysis of continuous speech reveals that incomplete articulation causes transitional sound intervals. Given a specific example in Fig. 1.2, considering an utterance of the concatenate vowel /aia/, with the influence of co-articulation, the variation of spectra features from /a/ sound to /i/ sound or from /i/ sound to /a/ sound passes the area of /e/ sound. Such phenomenon makes difficulty when determining the boundaries between different phonemes.

However, for human beings, transitional sounds are not perceived, even in co-articulation intervals, when their physical characteristics approximate the characteristics of other sounds. For example, the transitional sound /e/ in the diphthong /ai/ will not be per-

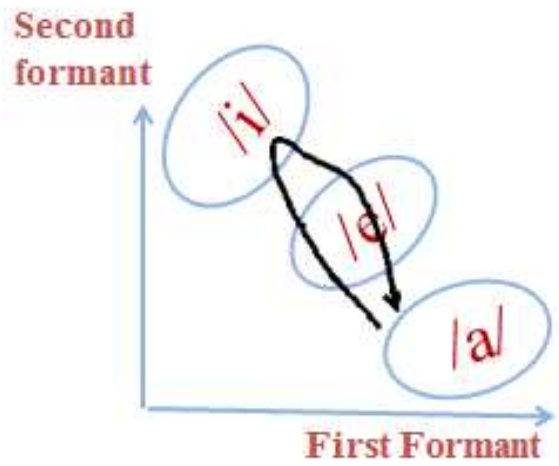

Figure 1.2: Trajectory of spectra features on the first (F1) and second formant (F2) space (/aia/).

ceived even though its physical characteristics will appear in the co-articulation interval between the /a/ sound and/i/ sound.

This phenomenon can be explained by a compensation mechanism that presumably exists in the speech perception mechanism[24–26]. For example, Lindblom and Studdert-Kennedy reported that there may be some overshoot or extrapolation in the processing of brief stimuli characterized by rapidly changing spectra[25]. Such mechanism helps human beings to perceive the accurate boundaries for continuous speech.

1.3.2 Principle of HMM-based segmentation

Considering the same example /aia/ as the last subsection, a brief segmentation procedure is described in this subsection. In simpler Markov models (like a Markov chain), the state is directly visible to the observer, and therefore the state transition probabilities are the only parameters. However, as shown in Fig. 1.3, in a HMM, the states are not directly visible, but outputs, dependent on the states, are visible. Moreover, for each HMM, there is a state transition to describe the connection between the HMM and a corresponding output. For each phoneme, a HMM with five states is usually used, including the non-emitting entry and exit states, as shown in Fig. 1.3. The parameters for each state in the HMM can be trained by Baum-Welch algorithm[17].

When carrying out the phoneme segmentation (in other words, time-alignment) for continuous speech, trained HMMs will be connected together in sequence, according to the phoneme sequences of the speech signal. Each model in the sequence corresponds directly to the assumed underlying symbol. Considering the example /aia/, the HMMs for /a/ sound and /i/ sound will be used to construct a composite HMM, as shown in Fig 1.4. The non-emitting states "null" in the composite HMM indicate the transfer from one symbol to another symbol. The states "a_ 1", "a_ 2" and "a_ 3" indicate three trained

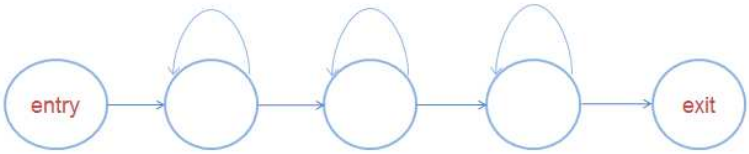

Figure 1.3: The uaually used HMM for a phoneme.

states in the HMM for the phoneme "a". Accordingly, the states "i_ 1", "i_ 2" and "i_ 3" indicate three trained states in the HMM for the phoneme "i". Then, according to the composite HMMs, given a obversation sequence in frame level obtained from the speech /aia/, the possibilities for different tokens will be calculated out. Then, a hidden sequence for the obversation sequence will be outputted by the means of Viterbi coding, according to Maximum Likelihood (ML) model[17]. As shown in Fig. 1.5, supposing that the hidden sequence is outputted corresponding to the obversation sequence in each frame at the maximal possibility, the tokens in hollow rectangles in different colors indicate different symbols in the phoneme sequence. Segmentation will be carried out at points-in-time when output of the HMM changes to a following phoneme, to get the boundaries (two red vertical lines in Fig. 1.5).

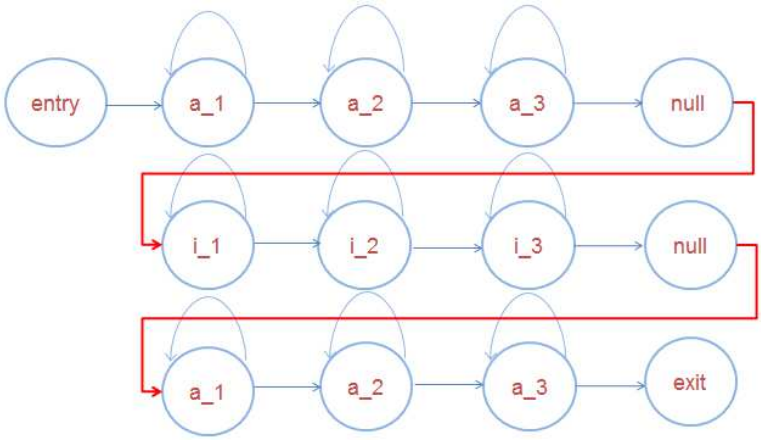

Figure 1.4: A composite HMM for the continuous peech /aia/. The non-emitting states "null" indicate the transfer from one symbol to another symbol. The states "a_ 1", "a_ 2" and "a_ 3" indicate three trained states in the HMM for the phoneme "a". Accordingly, the states "i_ 1", "i_ 2" and "i_ 3" indicate three trained states in the HMM for the phoneme "i".

1.3.3 Drawback of HMM-based segmentation

There has been brief description for human perception characteristics and HMM-based segmentation method in the last two subsections. The drawback of HMM-based segmentation will be stated in this subsection.

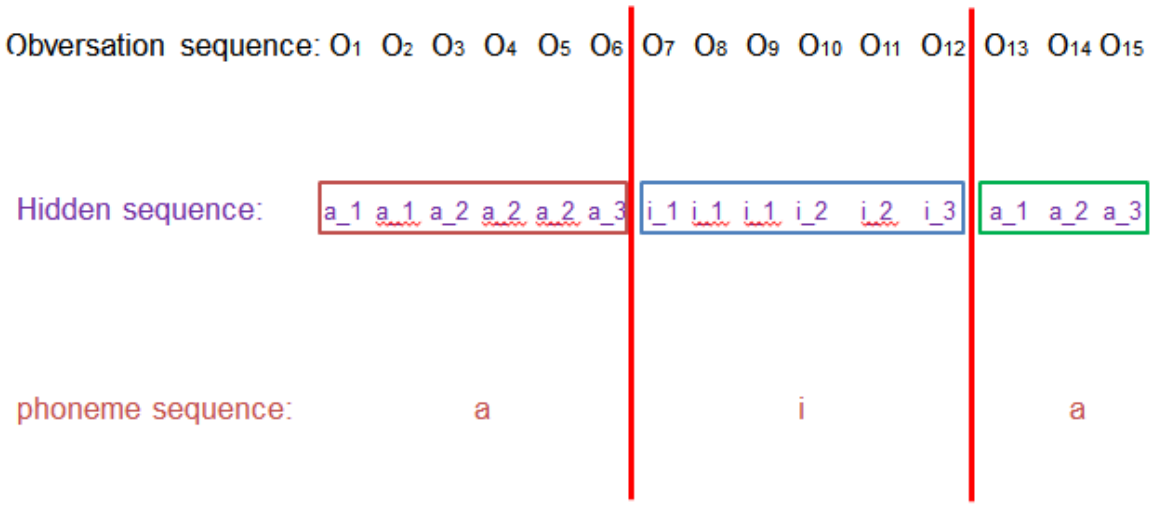

Figure 1.5: Segmentation at points-in-time when output of the HMM changes to a following phoneme. The hidden sequence is outputted corresponding to the obversation sequence in each frame by Viterbi agorithm. The tokens in hollow rectangles in different colors indicate different symbols in the phoneme sequence. The red vertical lines indicate the HMM-based boundaries.

When dividing the original signal /aia/ into frames, there are wide length for interval sound /e/. According to the first subsection in this section, because human beings do not perceive the transition sound /e/, the physically estimated critical points are identified as the points where the phonemes identified according to the minimum distances from the estimated target first changes from /e/ sound to /a/ sound or /i/ sound[27]. However, based on the variation of likelihood, the critical points may not be detected as the automatic boundaries between phonemes, which makes mismatching to human perception characteristics. Thus, without taking human perception in consideration, it is difficult for speech synthesis to use automatic segmented speech corpus based on HMM to maintain highly naturalness. Moreover, it is also suspect to supply automatic segmented speech corpus based on HMM for applications with the requirement for highly accurate and reliable speech segmentation.

1.4 Purpose of this research

According to the Subsection 1.3.3, the results by HMM-based automatic segmentation method are often found unsatisfactory to be directly applied to TTS or some other applications with the requirement for highly accurate and reliable speech segmentation. Thus, the boundaries should be adjusted to match human perception characteristics. According to the Subsection 1.3.1, a compensation mechanism presumably exists in the speech perception mechanism[24–26]. If this compensation mechanism can be modelled, it will be applicable to obtain the boundaries approximated to human perceptual results in speech

signal segmentation, to solve the problem that segmentation based on HMM makes mismatching to human perception characteristics.

The spectrum target prediction model[27–29] was constructed on the assumption that humans have a spectrum target prediction mechanism and use it to perceive predicted spectra. The model predicts each phoneme target by using dynamic spectral features which are important for phoneme perception.

Based on the spectrum target prediction model, the purpose of this research is to propose an automatic speech segmentation method. As shown in Fig. 1.6, by choosing segmentation points-in-time when the estimated phoneme target changes in human perception, appropriate segmentation and labelling are processed for highly natural speech synthesis or some other applications with the requirement for highly accurate and reliable speech segmentation using huge speech corpus.

To achieve the purpose, there are several sub-goals for this research. First of all, the candidates for the precise boundaries should be calculated out by the spectrum target prediction model. How to extract the boundaries from the candidates is the next sub-goals. Afterwards, based on the strategies for extracting the phoneme boundaries, the schemes for labelling methods should be constructed. The last, there should be evaluation for the proposed method compared to the HMM-based method.

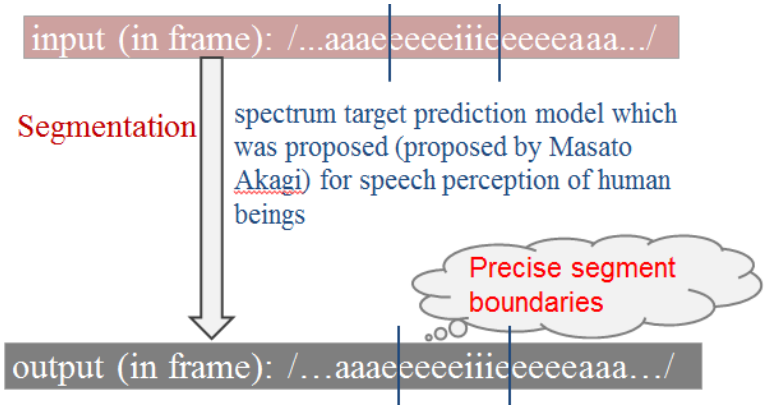

Figure 1.6: Ideal results by proposed automatic speech segmentation method. The blue lines indicates the boundaries in human perception.

1.5 Thesis structure

This thesis is structured as follows. In Chapter 2, an outline of the proposed method is introduced. The following chapter then introduce the databases used in this research. To achieve the first sub-goal presented in the last section, the outline of the spectrum target prediction model is introduced in Chapter 4, and simulated results are also included in this chapter. In the following Chapter, experiments for segmentation using ATR database[30] are presented to verify the effectiveness of the spectrum target prediction model. Based on the analysis results in Chapter 5, how to extract phoneme boundaries and implement

labelling are described in Chapter 6. The results of experiments for labelling, using TIMIT database[31], are also shown in this chapter. According to the experiments in the last chapter, evaluation for the proposed method is presented in Chapter 7. Meanwhile, based on the results of evaluation, the advantages and disadvantages for the proposed method are summarized in this chapter. In the last chapter, the conclusion for this thesis is presented. Moreover, according to the current work, future work is put forward to.

Chapter 2

An outline of the proposed method

This chapter will introduce an outline of the proposed method. Structured as following, the conceptual idea for determining the phoneme boundaries based on the spectrum target prediction model presented in Section 2.1, then the outline of the proposed method is constructed in Section 2.2.

2.1 Conceptual idea for determining the phoneme boundaries

Referring to spectrum target prediction model simulating the human speech perception mechanism[27–29], the model predicts the stable spectral target in each short-term interval and compensates for phonemic characteristics from co-articulation. When the trajectories of spectral features are approximated by a 2^{nd}-order critically damped system, the model can estimate target spectral features using short-term spectrum sequences without being given the onset positions of the spectral transition. If the prediction accuracy is high enough for segmentation points-in-time, it is possible to process appropriate segmentation and labelling.

A conceptual idea for determining the phoneme boundaries based on the spectrum target prediction model is shown in Fig. 2.1. Given the same example /aia/ in Section 1.3, the brown dotted line indicates the trajectory of spectral features in time domain. According to the trajectory of spectral features (physical characteristics) from /a/ to /i/, phoneme "a" (psychological characteristics) in human perception is estimated by the target prediction model. The black solid line indicates the sequence of the target. Then, segmentation is processed at the point-in-time when this estimated phoneme target changes. The result is, converting the gradual transition (the sequence of spectral features) into a step function (the sequence of the target) to get accurate segmentation boundaries for continuous speech.

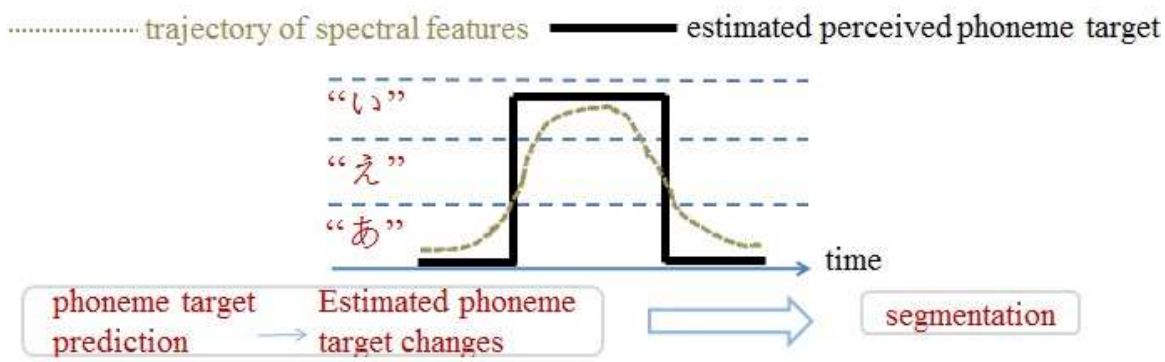

Figure 2.1: Conceptual idea for for determining the phoneme boundaries based on the spectrum target prediction model.

2.2 An outline for the proposed method

As shown in Fig. 2.1, to calculating out targets for each frame of the speech is needed to determine the phoneme boundaries, which makes the spectrum target prediction model is highly sensitive to the variation of targets. As the results, the number of the boundaries could not be determined, which will be discussed in detain in Chapter 5. How to solve this problem is the key point of the proposed method.

As well known, continuous speech in English and in Japanese can be separately segmented and labelled by the Hidden Markov Model Toolkit (HTK)[17] and the Julious[32] based on HMM. Besides, there has be other software based on HMM for different languages. Thus, the labelling files obtained by the HMM-based automatic segmentation method can be used. As shown in Table 2.1, comparing the HMM-based method and the method based on the spectrum target prediction model, for the HMM-based segmentation and labelling, the number of boundaries can be determined, however, the boundaries are not accurate enough compared to hand-made boundaries based on human perception characteristics according to Section 1.3, on the opposite, there are accurate boundaries by the spectrum target prediction model, however, it is difficult to determine the number of boundaries. Thus, considering the advantages and disadvantages of the two methods, the proposed method is to combine the advantages of the two methods, by adjust the HMM-based boundaries to the boundaries obtained by the method based on the spectrum target prediction model. The scheme is shown in Fig. 2.2.

Table 2.1: Comparison for spectrum target prediction model and HMM.

Methods	Advantage	Disadvantage
HMM	the number of boundaries can be determined	the boundaries are not accurate enough
spectrum target prediction model	accurate boundaries are obtained	the number of boundaries can not be determined

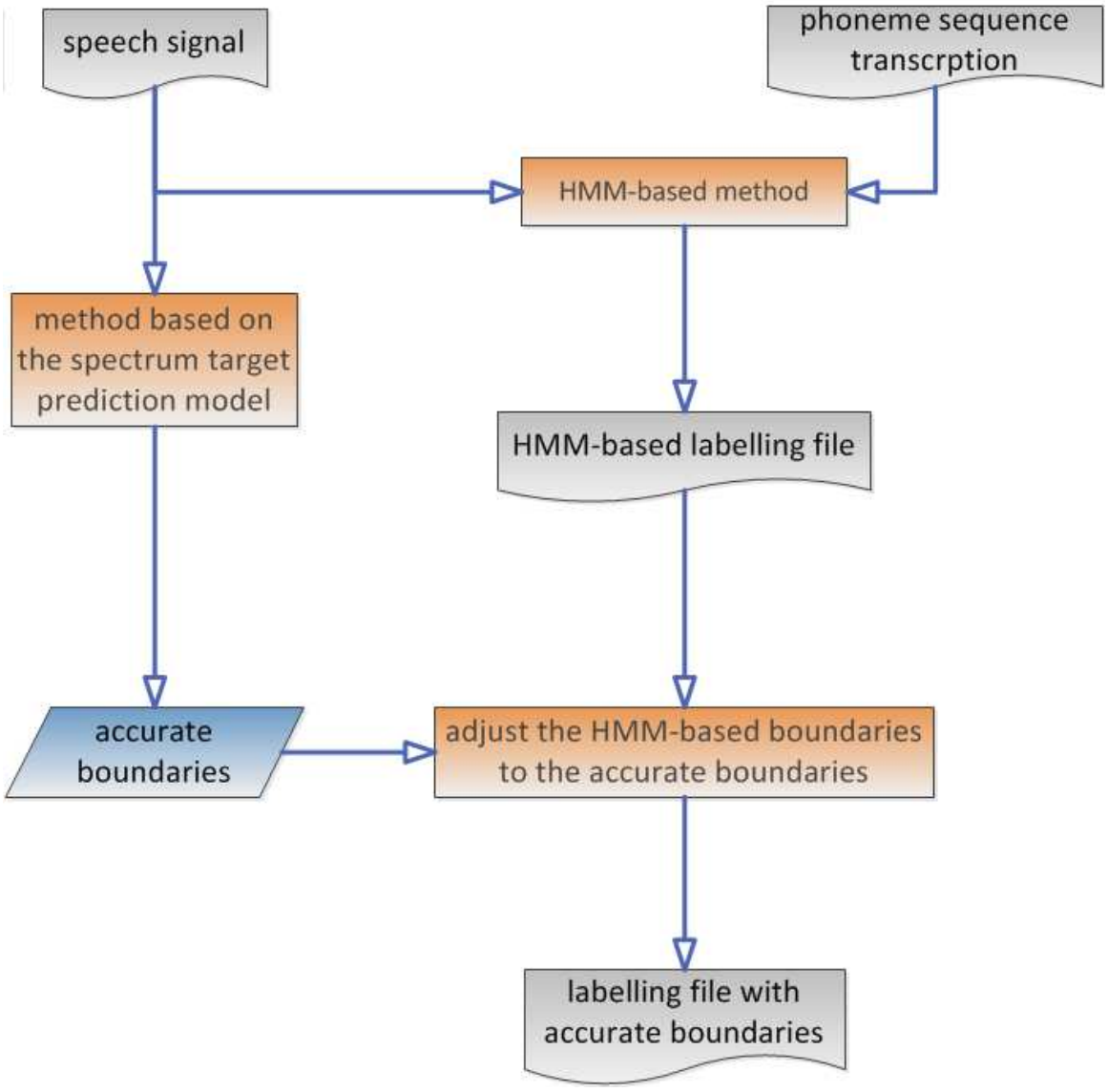

Figure 2.2: Scheme of the proposed method.

Chapter 3

Databases

Before the proposed method is presented in detail, there will be a brief introduction for the databased used for this research in this chapter. There are two databases to be selected: TIMIT database and ATR database. The reason to choose these two databases is that there are highly precise hand-made labelling files based on human perception corresponding to each utterances. For the purpose of this research is to obtain the accurate boundaries in human perception, these highly precise hand-made labelling files exactly meet the requirement of evaluation for the proposed method. The following two sections will introduce TIMIT database and ATR database in order.

3.1 TIMIT database

The TIMIT corpus of read speech is designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance. The speech was recorded at Texas Instruments, Inc (TI), transcribed at Massachusetts Institute of Technology (MIT).[31]

The TIMIT corpus transcriptions are made by audiologists and have been hand verified. In this research, the phonetic transcriptions will be treated as the most precise labelling results and used for comparison to automatic segmentation results. Test subset (including 1680 sentences) and training subset (including 4620 sentences), balanced for phonetic and dialectal coverage, are specified[31]. In this experiments, all test sentences will be used.

3.2 ATR database

Since 1986, Advanced Telecommunications Research Institute International (ATR) has taken the initiative in making a large-scale speech database that meets the needs of a variety of studies in speech research. The database basically consists of four types : a word speech database, a continuous speech database, a database for a large numbers of

speakers, and a database for speech synthesis. All speech data are digitized in 16 bits by 20 kHz sampling.[30]

The word database includes 5229 frequently used words selected from a Japanese word dictionary, 216 phonetically balanced words, 115 short sentences in a domain within which ATR is trying to develop an interpreting telephony system, and other supplementary words. The continuous speech database is a set of 503 phonetically balanced short sentences with no specific domain.[30]

Segmentation and transcription of the database were done by hand in order to keep the quality of the speech data as high as possible. By inspecting spectrograms, segmentation was done phoneme-by-phoneme and, for each vowel, transitional parts from/to adjacent consonants were also marked. In some cases, it is hard to define the boundaries between phonemes, especially between two consecutive vowels or between semi-vowel and vowel. In such cases, boundaries were defined as the center of the formant transitions between the two phonemes. However, no segmentation was made when two phonemes were fused into a single phoneme which we call "inseparable portion". Labellers were trained for a couple of months how to make segmentations and transcriptions by looking at the sound spectrograms.[30]

Chapter 4

Spectrum target prediction model

According to Section 1.4, the prediction target prediction model simulates the human speech perception mechanism, especially a compensating mechanism which presumably exists in the speech perception mechanism. Thus, this model will be adopted for automatic speech segmentation approximated by manual segmentation.

Although the details of the model have been presented in some references[27–29], this chapter will introduce an outline of the model firstly and then give simulated results for this model to indicate the effectiveness for this research. First, basic concepts and the algebraic derivations for the model are described in Section 4.1. Next, preliminaries for the model are explained in Section 4.2. Moreover, a method for short-term prediction is proposed in Section 4.3. In the last two section, simulated results are shown, then the discussion and conclusion are presented in order.

4.1 Basic concept

Considering a general 2^{nd}-order critical-damping model as follows:

$$\left(\Delta^2 - 2\lambda\Delta + \lambda^2\right) y_n = \lambda^2 b \tag{4.1}$$

In Equation (4.1), y_n denotes a vector obtained at the time index n. Δ is a differential operator in time. λ is a reciprocal time constant. The time position $n = 0$ is the onset position of the transition. And, b denotes a target to which y_n converges. Here, $\lambda < 0$ and $\lambda > 0$ indicate the past (or $n \leq 0$) and the future (or $n \geq 0$) respectively. The solution of Equation (4.1) is:

$$y_n = (a + cn)\, e^{\lambda n} + b \tag{4.2}$$

In Equation (4.2), a and c are constants determined by boundary conditions. Other parameters can be predicted by using Equation (4.3) and the following measure[33,34]:

$$e\left(n_0 or n_1, \lambda\right) = \sum_{n=n_0}^{n_1} \left|y_n^i - y_n\right|, n_0 < n_1 \tag{4.3}$$

where y_n^i is an unknown input sequence. For these methods, a long-term input sequence is essentially required, starting at the onset position of the transition $n_0 = 0$ when $\lambda < 0$ or $n_1 = 0$ when $\lambda > 0$. Thus, two values are needed for non-linear optimization. However, the purpose of the model is to estimate b only.

Divide Equation (4.1) such that:

$$(\Delta - \lambda)\{(\Delta - \lambda)\, y_n\} = \lambda^2 b \tag{4.4}$$

and assume that:

$$x_n = (\Delta - \lambda)\, y_n \tag{4.5}$$

Equation (4.1) may be represented more simply as:

$$(\Delta - \lambda)\, x_n = \lambda^2 b \tag{4.6}$$

and Equation (4.6) is a 1^{st}-order equation. The solution of Equation (4.6) is:

$$x_n = ce^{\lambda n} - \lambda b \tag{4.7}$$

At time $n = m$, assume that:

$$c_m = ce^{\lambda m} \tag{4.8}$$

The neighbourhood x_{m+t} of x_m is expressed by:

$$x_{m+t} = c_m e^{\lambda t} - \lambda b \tag{4.9}$$

The target b is the next estimated by minimizing the following error:

$$e(\lambda) = \sum_{t=n_0}^{n_1} \left|(\Delta - \lambda)\, y_{m+t}^i - x_{m+t}\right|^2 = \sum_{t=n_0}^{n_1} \left|x_{m+t}^i - x_{m+t}\right|^2 \tag{4.10}$$

where non-linear optimization under only λ is needed and it does not required the onset position of the transition in estimating the target b for the reason that x_{m+t} is an exponential function.

4.2 Preliminaries

4.2.1 Parameters for speech analysis

The speech waveform should be digitized to 16 bits at a sampling frequency of 8KHz in advance. Then, a 20 milliseconds Hamming window spaced every 1 millisecond is used to compute the values of acoustic features. To correspond with the purpose of this research, the Mel Frequency Cepstrum Coefficient (MFCC) is selected, which is extracted from speech signal based on the human auditory characteristics.[35]

4.2.2 Differentiation for time derivative $\dot{f}_n$ of f_n

A suitable measure of the time derivation $\dot{f}_n$ of f_n is needed to obtain stable values, as shown in Equations (4.5) and (4.6). Thus, the difference between adjacent spectra, f_n and f_{n+1}, is not used. Instead, $\dot{f}_n$ is calculated as the 1^{st}-order regression coefficient of f_n, as shown in Equation (4.11).

$$\dot{f}_n = \frac{\sum_{t=n-N/2}^{n+N/2} (t-n) * f_t}{\sum_{t=n-N/2}^{n+N/2} (t-n)^2} \tag{4.11}$$

where N is the length of the window used for differentiation. In this research, N is selected to be 30 milliseconds at first.

4.3 Prediction method

Assume that the model is used during an interval of length L centered at time m. In this research, L is selected to be 50 milliseconds originally. Actually, 50 milliseconds is the most suitable length of windows for estimation[27]. Given a value $\widehat{\lambda}_m$ for λ_m over the interval: $m - L/2 \leq n \leq m + L/2$, two sets of values can be calculated for the input spectrum x_n: $\widetilde{x}_n$, $\widehat{x}_n$. The former one could be calculated directly using Equation (4.12). The other one comes from fitting $\widetilde{x}_n$, in the least squares sense, as shown in Equation (4.13).

$$\widetilde{x}_n = -\widehat{\lambda}_m y_n + \dot{y}_n, m - \frac{L}{2} \leq n \leq m + \frac{L}{2} \tag{4.12}$$

$$\widehat{x}_n = a_m e^{\widehat{\lambda}_m (n-m)} - \widehat{x}_m b_m, m - \frac{L}{2} \leq n \leq m + \frac{L}{2} \tag{4.13}$$

And when calculating the second one, values for the constant vectors a_m and b_m (the target) are given. The next, the total squared error is calculated from Equation (4.14), according to Equation (4.10).

$$e\left(\widehat{\lambda}_m\right) = \sum_{n=m-L/2}^{m+L/2} |\widetilde{x}_n - \widehat{x}_n|^2 \tag{4.14}$$

It can be minimized by optimization $\widehat{\lambda}_m$ as a non-linear optimization problem under renewing parameter $\widehat{\lambda}_m$. The measurement is set to 0.0001 in this research. The optimization method is chosen to be "Golden Cut Algorithm", which is often used in the area of operations research (OR). The $\widehat{\lambda}_m$ obtained by this method is the reciprocal time constant λ_m, and the b_m is the target in the interval: $m - L/2 \leq n \leq m + L/2$.

The block diagram for estimation[27] is shown in Fig. 4.1.

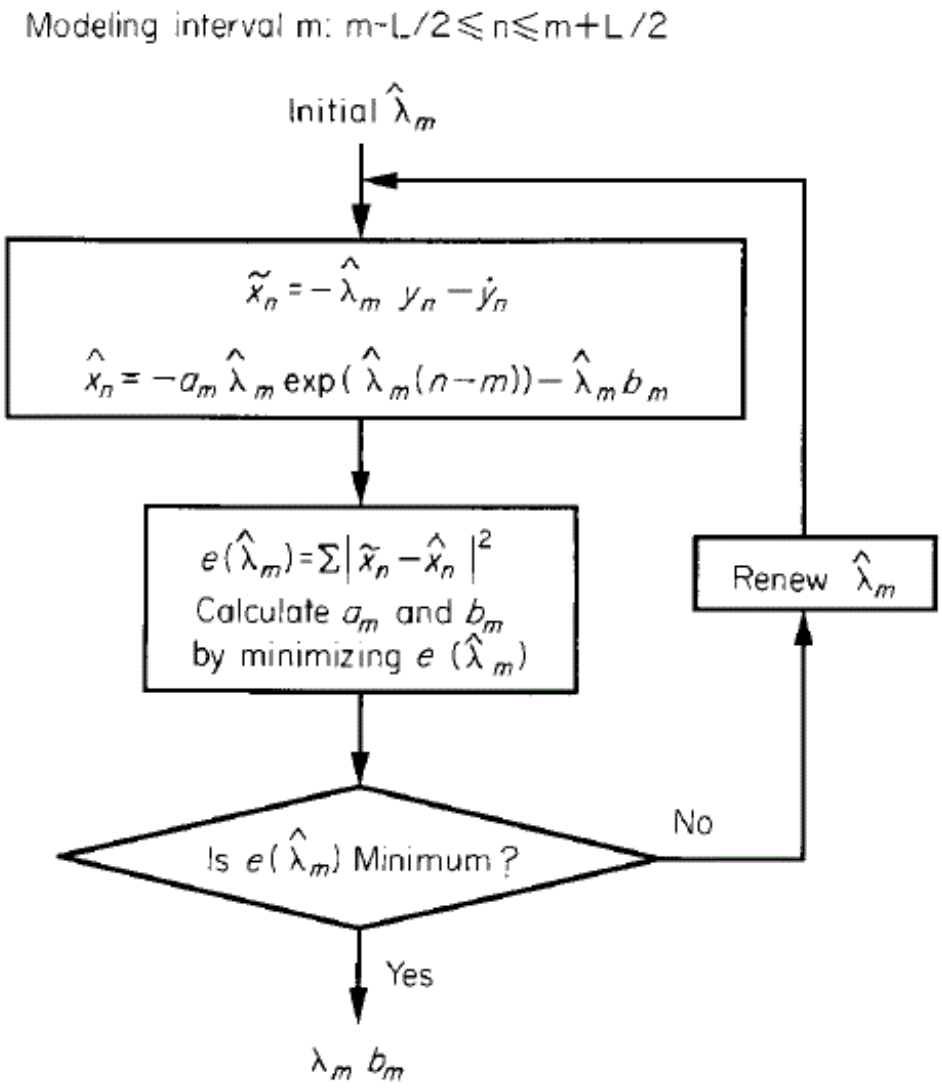

Figure 4.1: Block diagram for solution of the spectrum target prediction model.[27]

4.4 Simulated results

With the purpose of verifying the effectiveness of this model for the purpose of this research, simulated experiments has been carried out first and the results are shown in Fig. 4.2. the solid line indicates the 2nd-order critically damped function by Equation (4.2) as the input. The values are $a = -1$, $b = 1$, $c = 0.02$ and $\lambda = -0.02$, which means the input is: $y = (-1 + 0.02n)\, e^{-0.02\lambda} + 1$. The horizontal axis and the vertical axis indicate the time domain (the unit is set to 1 millisecond) and the variation of the amplitude with time passing. The dashed line indicates predicted target b.

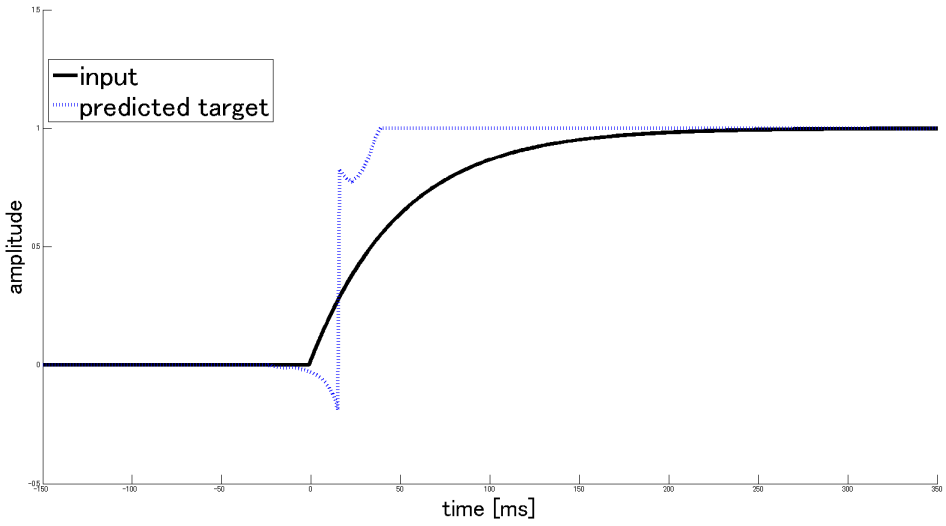

Figure 4.2: Simulated results for the spectrum target prediction model.

4.5 Discussion and conclusion

According to Fig. 4.2, the model nearly converts the gradual transition into a step function. The point-in-time where the input (solid line) of this model intersects with the output (dashed line) is the boundary, according to the conceptual idea for determine the phoneme boundaries in Section 2.1. Around the boundary, target changes from nearly 0 to nearly 1 rapidly while there are only small changes for the targets around other points-in-time. If the Euclidean distances of the adjacent targets are calculated, the result must be that the value of distance at the boundary is much bigger than the value of distances in other points-in-time, which means the position of boundary is a peak in the Euclidean distances of the adjacent targets. Thus, to select the peaks makes it possible to implement the conceptual idea for determine the boundaries in Section 2.1.

The simulated results indicate the effectiveness of the spectrum target prediction model for the purpose of this research. The experiments for segmentation using speech signal will be carried out in the next chapter.

Chapter 5

Experiments for segmentation

Based on the simulated results for the spectrum target prediction model in the last chapter, experiments for segmentation using the speech signal are carried out to achieve the first sub-goal (find the candidates for the precise boundaries) described in Section 1.4. This chapter is divided into four sections. In section 3.1, the description for experiments in brief is presented. Then, the results of the experiments are shown in section 3.2. In the last two sections, some discussion and conclusion for the results of experiments are presented, for next sub-goal of this research.

5.1 Description for experiments

Files for experiments include an sound file from ATR database and a corresponding manually labelling file used for comparison to experiment results. The used software for experiments is MATLAB. To be explained in advance, in this research, manual segmentation and labelling with listening tests will be thought as the most precise results, including the experiments described in the next chapter.

Steps for experiments are listed as below.

1. Re-sampling sound file at a sampling frequency of 8KHz (as mentioned in Subsection 4.2.1, The speech waveform should be digitized to 16 bits at a sampling frequency of 8KHz in advance to be provided to the spectrum target prediction model)
2. Setting parameters for obtaining the values of MFCCs, as shown in Table 5.1
3. Obtaining acoustic features (MFCC) from sound file by the programs similar with the executable file "HCopy" in HTK[17]
4. Calculating the differentiation for time derivative for each frame
5. Estimating targets for MFCCs by the spectrum target prediction model
6. Calculating the distances between the adjacent targets for discussion

5.2 Results of the experiments

Results of the experiments using MFCCs are shown in Fig. 5.2. In Fig. 5.2, the sub-graphs from the top to the bottom indicate the speech waveform (parts of the whole

Table 5.1: Parameters for obtaining the acoustic features (MFCC).

Parameter	Value
analysis frame duration	20ms
analysis frame shift	1ms
the preemphasis coefficient	0.97
the number of filterbank channels	20
the number of cepstral coefficient	12
the cepstral sine lifter parameter	22
lower frequency limit	300Hz
upper frequency limit	3700Hz

waveform), manual labelling results, the trajectory of the input MFCCs, the trajectory of the targets obtain by the spectrum target prediction model, the Euclidean distances of the adjacent targets. In all sub-graphs, the horizontal axis presents the time domain and the unit is second. The color in the third sub-graph and the fourth sub-graph presents different values of each points. The red solid lines indicate segmentation boundaries based on manual labelling results. The purpose to draw two hollow color circles is to compare the trajectory of MFCCs to the trajectory of targets, before one point-in-time for segmentation and after this point-in-time.

5.3 Discussion

Considering the purpose of this research that is to determine phoneme boundaries in human perception, based on the results of experiments, the advantages and disadvantages for the spectrum target prediction model are discussed as below.

Advantages of the spectrum target prediction model:

According to figures comparing original acoustic features (MFCCs) and estimated targets, when acoustic features change smoothly over time, as corresponding, estimated targets change rapidly, just as the variation of color in two hollow color circles. The same as the simulated results, the proposed method almost converts the gradual transitions into some step functions, which indicate the effectiveness of the spectrum target prediction model for the purpose of this research.

Then, comparing manual labelling and the peaks of Euclidean distances of the adjacent targets, around the hand-made segmentation boundaries in human perception(red lines), there are always peaks to be selected as the automatic segmentation boundaries, whose values are almost the biggest in a period. Besides, the positions of such peaks are similar as the hand-made boundaries. Based on the results, some strategies for extracting phoneme boundaries from all peaks (treated as candidates) can be constructed, to make it possible that the automatic speech segmentation results by the proposed method do not mismatch to human perception characteristics as HMM.

Disadvantages of the spectrum target prediction model:

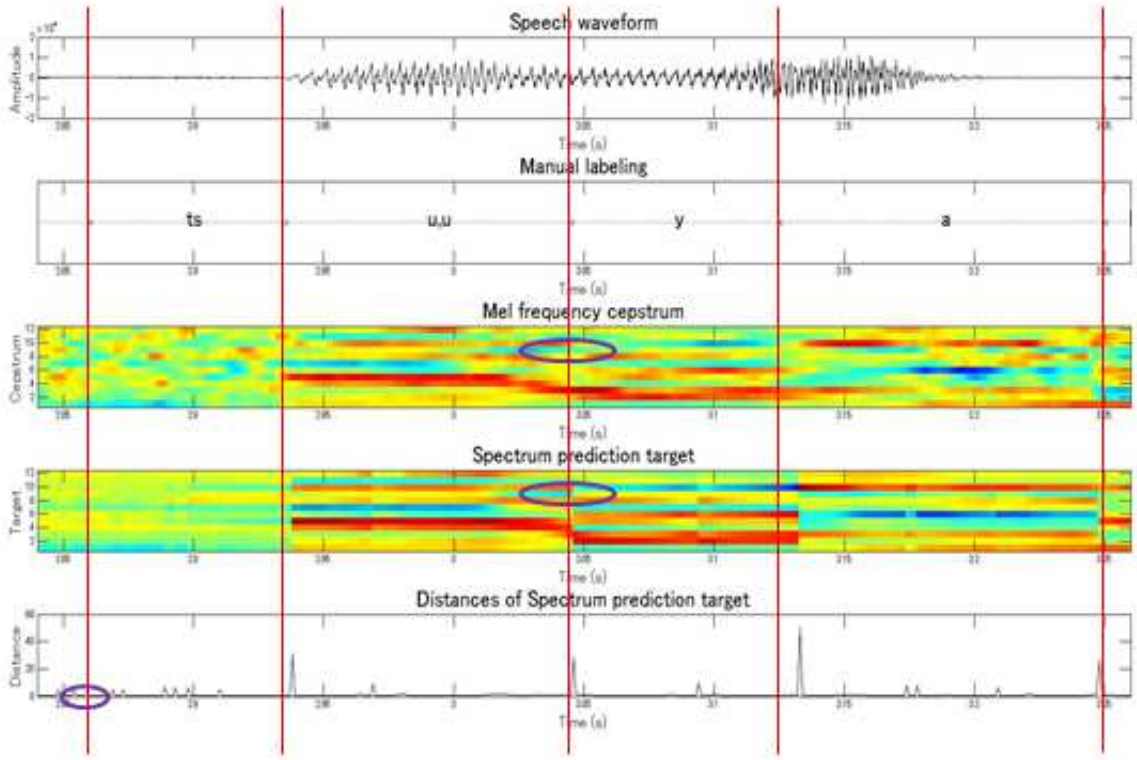

Figure 5.1: Results of the experiments for segmentation using MFCCs. The sub-graphs from the top to the bottom indicate the speech waveform (parts of the whole waveform), manual labelling results, the variety of MFCCs, the variety of the targets obtained by the spectrum target prediction model, the Euclidean distances of the adjacent targets. The red solid lines indicate segmentation boundaries based on manual labelling results.

The number of peaks is too much. Because the spectrum target prediction model is highly sensitive to the variation of targets, there are too much peaks whose values are small, making difficulty for determining the boundaries, as shown in the last sub-graph in Fig. 5.2. Thus, even though the numbers of boundaries are given when implementing the forced alignment for speech signal, it is hardly possible to extract the right peaks as boundaries, for there have been no rules to determine the boundaries using the framework of the spectrum target prediction model by now.

5.4 Conclusion

Based on the discussion in last section, the candidates to be selected as phoneme boundaries have been calculated out, which indicates that the first sub-goal for this research presented in Section 1.4 has been achieved. How to extract the phoneme boundaries from all candidates is the next sub-goal, which will be described in the next chapter.

Chapter 6

Implementation for the proposed automatic segmentation method

According to the discussion for the results of experiments for segmentation in Section 5.3, although the similar boundaries as the hand-made results can be obtained by the spectrum target prediction model, how to extract these boundaries from all candidates (peaks in the sub-graph of Fig. 5.2) is a problem.

In this chapter, the strategy for extracting the phoneme boundaries from all candidates will be described, to achieve the second sub-goal presented in Section 1.4. The labelling method after obtaining boundaries in human perception is also presented in this chapter, to achieve the third sub-goal presented in Section 1.4. The two parts presented in the first two sections implement the scheme of the proposed automatic segmentation method presented in Section 2.2. Then, results of the experiments using TIMIT database are shown in Section 6.3. The last section will give a conclusion for this chapter.

6.1 Strategy for extracting the phoneme boundaries from all candidates

6.1.1 An outline for the strategy

From the discussion in Section 3.3, there is an idea for the strategy, to treat the labelling files obtained by other automatic segmentation methods as references just as the manual labelling file used in the experiments presented in Chapter 3. There must be no manual labelling files as references when automatic speech segmentation is carried out for real speech by the proposed method. However, continuous speech in English and in Japanese can be separately segmented and labelled by HTK and Julious based on HMM. Besides, there has been other software based on HMM for different languages. Thus, the labelling files obtained by HMM-based automatic segmentation methods can be used. Comparing HMM-based methods and the proposed method, for the HMM-based segmentation and labelling, the number of boundaries can be determined, however, the boundaries are not accurate enough compared to hand-made boundaries, on the opposite, there are always

candidates close to the hand-made boundaries by the spectrum target prediction model, however, there have been no suitable rules for choosing these accurate automatic segmentation boundaries from a number of candidates directly. Thus, to adjust the HMM-based boundaries to the accurate boundaries obtained by the spectrum target prediction model is a possible strategy for extracting the phoneme boundaries from all candidates.

In addition, according to the discussion for the advantage of the spectrum target prediction model in Section 5.3, there are always peaks near the manual boundaries, as shown in Fig. 5.2. Thus, if the estimation for the boundaries in human perception is feasible using the existing labelling files by the HMM-based method, it is foreseeable to obtain more precise boundaries compared to hand-made ones.

6.1.2 Implementation of the strategy

To implement this strategy, the labelling files obtained by the HMM-based method are needed. Fig. 6.1 gives an example for the labelling files, which is an revised output for time-alignment by HTK. There are three columns of data for each row in this file. The first number is the start time of a phoneme and the second number is the end time of this phoneme. The unit for two numbers is 1 millisecond for the proposed method. The third data is the phoneme.

In original labelling files by HTK, the unit for the time is 100 nanosecond, which should be divided by 10000 to be converted to 1 millisecond. Besides, in each row, there is a additional data, which is the same as the third data in the sample labelling file. It is used to identify the word sequence[17] and not used in this research.

```
0 580 h#
580 700 sh
700 790 iy
790 900 hv
900 1000 ae
1000 1040 dcl
1040 1070 d
1070 1100 y
1100 1150 er
```

Figure 6.1: A sample labelling file for HMM-based segmentation.

To estimate the boundaries in human perception, one more file is needed, as shown Fig 6.2. In this file, there are five columns of data in each row. The former three numbers are the mean errors, maximum errors and minimum errors for a boundary separately. The rest data records the phoneme before the boundary and the phoneme after the boundary separately. It can also be obtained statistically by calculating the errors between all automatic segmentation boundaries by the HMM-based method and the corresponding manual boundaries. The revised manual labelling file corresponding the HMM-based labelling file from TIMIT database in Fig. 6.1 is shown in Fig. 6.3. The similar as the labelling file in Fig. 6.1, in each row, the first number is the start time of a phoneme, the

second number is the end time of this phoneme and the third data is the phoneme in the same order as the HMM-based labelling file.

The similar as the original labelling files by HTK, in the original manual labelling files in TIMIT database, the time in the phonetic transcriptions are recorded at the sampling points, which should be divided by 16 to convert the unit to 1 millisecond.

```
mean_error max_error min_error former  later
                                phoneme phoneme
-15.247596 22.500000 -43.125000 h# sh
1.735697 13.750000 -12.500000 sh iy
-6.421779 45.562500 -49.750000 iy hv
13.031250 111.250000 -55.187500 hv ae
-2.618398 18.687500 -16.937500 ae dcl
-6.227541 48.687500 -94.625000 dcl d
-2.884868 17.687500 -41.875000 d y
9.115132 41.250000 -46.000000 y er
-1.001488 39.500000 -22.500000 er dcl
4.526316 24.625000 -5.625000 d aa
13.439660 139.187500 -87.875000 aa r
0.999662 47.937500 -24.125000 r kcl
-5.391537 85.562500 -73.125000 kcl k
-2.311368 33.500000 -40.000000 k s
2.781250 23.750000 -7.500000 s uw
-5.687500 3.562500 -7.750000 uw dx
-1.192969 22.500000 -163.250000 dx ih
-8.637242 20.687500 -82.500000 ih ng
3.272629 52.500000 -30.062500 ng gcl
-5.358675 66.875000 -71.875000 gcl g
6.334086 38.000000 -12.500000 g r
-9.194943 86.875000 -57.062500 r iy
-1.435516 32.562500 -12.500000 iy s
```

Figure 6.2: A list to describe the mean errors for each boundary.

```
0 602.5 h#
602.5 702.5 sh
702.5 798.9375 iy
798.9375 879.875 hv
879.875 1009.8125 ae
1009.8125 1055 dcl
1055 1068.9375 d
1068.9375 1099.1875 y
1099.1875 1172.5 er
```

Figure 6.3: A sample labelling file for hand-made segmentation in human perception.

Utilizing the HMM-based labelling files and the corresponding manual labelling files, there are some steps to obtain the errors list file in Fig. 6.2. As shown in Fig. 6.4, firstly, all errors are calculated between HMM-based boundaries and the corresponding boundaries in human perception, while recording the phonemes before the boundaries and the phonemes after the boundaries. Then, the numbers and the total errors for the same boundaries (the phonemes before the boundaries and the phonemes after the boundaries

are the same) can be calculated statistically based on the results in the first step. Finally, the mean errors for each boundary can be calculated out, meanwhile, the maximum errors and the minimum errors can be obtained..

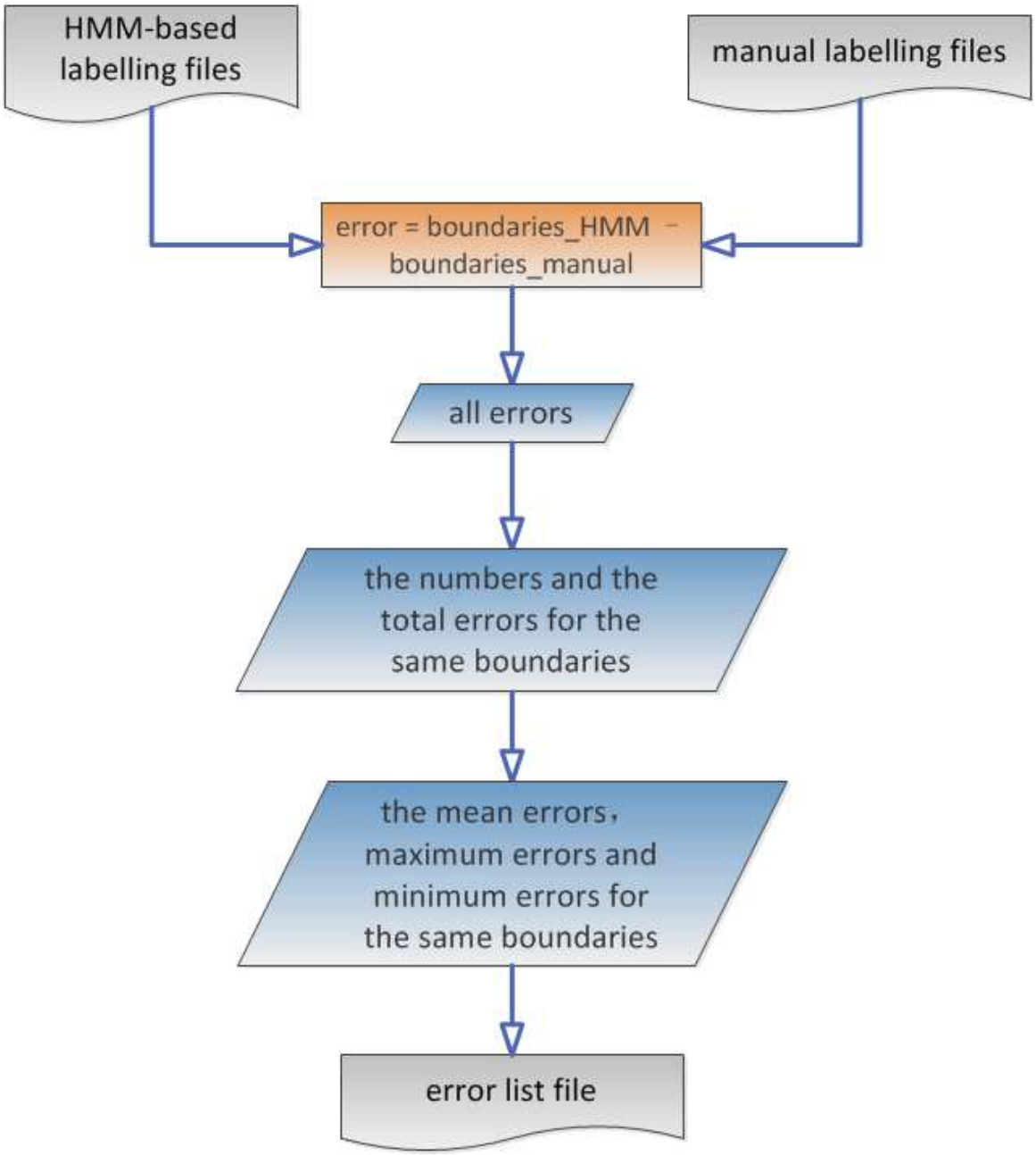

Figure 6.4: Flow-process diagram for obtaining the error list file.

Based on the list file for errors, a block diagram implementing this strategy for each speech signal is shown in Fig. 6.5. The speech signal and the HMM-based labelling file are the input, and the labelling file by the proposed method is the output. When a HMM-based boundary is provided to the proposed method, firstly the mean error should

be found for this boundary by checking in the errors list file according to the former phoneme and the later phoneme, and then the absolute value of the mean error will be judged if it is less than 20 milliseconds. If the answer is "Yes", a estimated manual boundary is determined by moving the HMM-based boundaries according to the found error and the nearest peak will be selected to be the boundary by the proposed method. Otherwise, a range $[hmm\_boundary - max\_error, hmm\_boundary - min\_error]$ for estimating manual boundary will be used for selecting the peak with biggest value as the automatic segmentation boundary, where $hmm\_boundary$ is the HMM-based boundary, meanwhile, max_error and min_error are the maximum error and the minimum error for this boundary compared to the manual boundary checked from the errors list file.

The reason compared to 20 milliseconds is that the accuracy of automatic segmentation is generally measured in terms of what percentage of the automatically labelled boundaries are within a given time threshold (tolerance) of the manually labelled boundaries. 20 milliseconds has been most widely used as a tolerance for measuring phone segmentation quality[22]. If the absolute error between automatic segmentation boundaries and manual segmentation boundaries is less than 20 milliseconds, the automatic segmentation boundaries will be treated as the suitable boundaries.

6.2 Labelling methods

In this section, based on the strategies for extracting the phoneme boundaries from all candidates presented in the last section, how to implement labelling (the third sub-goal presented in Section 1.4) will be constructed.

First of all, three points can be concluded from the labelling file in Fig. 6.1 or Fig. 6.3 as following.

(1) the start time of the first phoneme (actually, the value is 0) and the end time of the last phoneme are the start time and end time of the speech separately.

(2) The end time except the last one is the boundary.

(3) the end time of a phoneme is also the start time of the next phoneme.

Thus, after the phoneme boundaries are obtained using the speech signal and the corresponding HMM-based labelling file by the strategy described in Section 6.1, firstly the start time of the first phoneme and the end time of the last phoneme are copied from the HMM-based labelling file, then the rest of first row and second row of numbers are the obtained boundaries ordered by time. The last row of data are a copy from the phoneme sequence file without time-alignment in the same order as the HMM-based labelling file. The procedure is the adjustment for HMM-based boundaries.

6.3 Experiments for labelling

Experiments are carried out using TIMIT database to verify the effectiveness of the proposed method. Considering the language of TIMIT database is English, HTK is used to obtain rough automatic labelling files for the strategies presented in Section 6.1. Then,

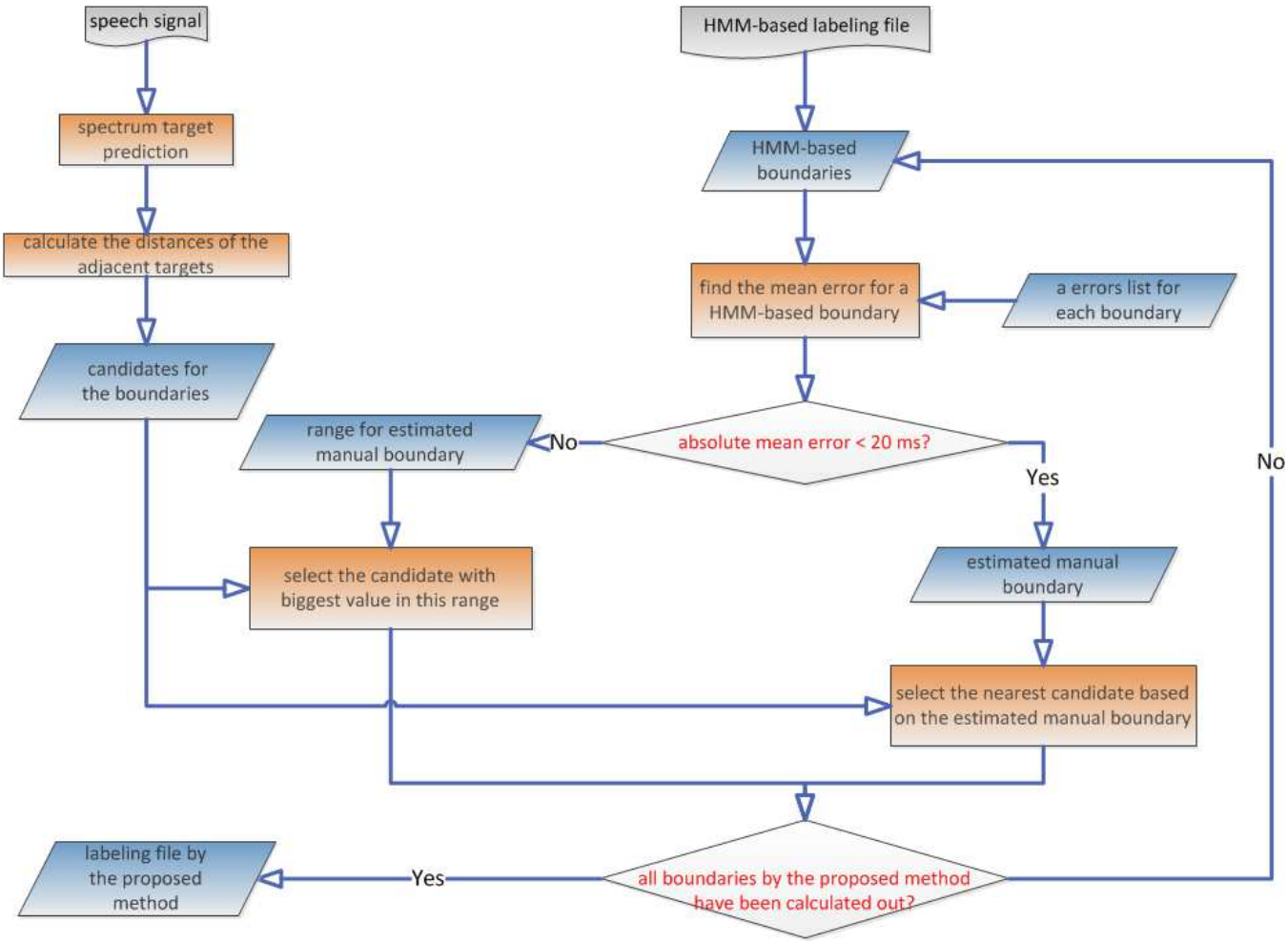

Figure 6.5: The block diagram for the proposed method. When a HMM-based boundary is provided to the proposed method, firstly the mean error should be found for this boundary by checking in the errors list file according to the former phoneme and the later phoneme, and then the absolute value of the mean error will be judged if it is less than 20 milliseconds. If the answer is "Yes", a estimated manual boundary is determined by moving the HMM-based boundaries according to the found error and the nearest peak will be selected to be the boundary by the proposed method. Otherwise, a range for estimating manual boundary will be used for selecting the peak with biggest value as the automatic segmentation boundary.

the error list file is obtained following the method shown in Fig. 6.4. Finally, the proposed method is utilized to get automatic labelling files with accurate boundaries. Particularly mentioned, when using TIMIT database, some preprocessing should be implemented.

6.3.1 Preprocessing for using TIMIT database

When using TIMIT database, some preprocessing should be implemented.

1. The TIMIT speech waveform files are SPHERE-headed[31] and can not be used in HTK and directly. Thus, firstly the waveform files should be SPHERE-striped to obtain the raw matrices and rewritten by the function "wavwrite" in MATLAB.

2. As mentioned in Subsection 4.2.1, The speech waveform should be digitized to 16 bits at a sampling frequency of 8KHz in advance to be provided to the spectrum target prediction model. Thus, the SPHRER-striped raw matrices should be re-sampled to 8KHz firstly when supplied for spectrum target prediction.

3. As presented in Subsection 6.1.2, the start time and the end time for each phoneme in the phonetic transcriptions are recorded at the sampling points. They are unable to be used in HTK, because the unit of time is 100 nanoseconds in HTK[21]. Besides, they are not suitable for comparison to the results by the proposed method, because the unit of time is 1 millisecond in the results by the proposed method. Thus, there should be a small change for the original phonetic transcriptions.

Particularly mentioned, in some research especially speech recognition, in order to improve the results of automatic recognition, the number of phonemes in original TIMIT phone set(61 phonemes) will be reduced by some rules[17,36]. However, in this experiments, all phonemes will be used for evaluation for the proposed method in the next chapter.

6.3.2 Description for experiments

Considering that the procedure to obtain the candidates for the automatic boundaries by the spectrum target prediction model is the same as what is described in Section 5.1, a detained description for obtaining the HMM-based labelling files will be presented in this subsection.

The whole training partition of TIMIT database (4620 utterances) were used for training and the whole test partition (1680 utterances) were used for testing. The 61 phonemes in original TIMIT phone set are listed in Table 6.1, classified by some rules[23,36].

Based on the classification in Table 6.1, for each phoneme, monophone HMM and GMM (Gaussian Mixture Model) acoustic models, with the standard 39 MFCC features extracted following the parameters shown in Table 6.2, were trained using the HTK[17]. For the parameters, compared to Table 5.1, "TARGETRATE" and "WINDOWSIZE" correspond to "analysis frame shift" and "analysis frame duration", separately. The units for this two parameters is 100 nanoseconds. The values are set to 5 milliseconds and 20 milliseconds widely used for forced alignment in HTK[37,38]. The number of states in the HMM models and the number of Gaussian mixtures were optimized for best alignment performance with 20 milliseconds tolerance presented in Subsection 6.1.2. Stops, stop closures, the vowel /ax-h/ ("devoiced schwa"), nasals, and liquids (/l/, /r/) are 1-state HMMs; the"true"

Table 6.1: The phoneme set (61 phonemes).

Type	Phoneme
Pauses and stop closures	pau, pcl, bcl, tcl, dcl, kcl, gcl, h#, eqi, q
Vowels	aa, ae, ah, ao, aw, ax, ax-h, axr, ay, eh er, ey, ih, ix, iy, ow, oy, uh, uw, ux
Glides	l, r, w, y, hh, hv, el
Nasals	m, n, ng, nx, em, en, eng
Plosives	b, d, g, p, t, k, dx, jh, ch
Fricatives	s, z, sh, zh, f, v, th, dh

diphthongs (/ay/, /aw/, /oy/) are 5-state HMMs; and the other phonemes are 3-state HMMs. Eight Gaussian mixtures were used.[23]

Table 6.2: Parameters for extracting MFCC features in HTK.

TARGETKIND	MFCC_ 0_ D_ A
TARGETRATE	50000.0
WINDOWSIZE	200000.0
USEHAMMING	T
PREEMCOEF	0.97
NUMCHANS	20
CEPLIFTER	22
NUMCEPS	12

Particularly mentioned, in forced alignment, unlike in automatic speech recognition, monophone HMMs are more commonly used than triphone HMMs, because monophone HMMs outperform triphone HMMs for medium tolerances (15-30 ms different from manual segmentation)[11,23,39]. Besides, the use of precise phonetic segmentation for training HMMs can improve the performance of forced alignment[12,23]. Thus, in this research, the acoustic observations of individual phonemes by extracting frames within the phone boundaries from observations (features) of utterances are used as input for training HMMs similar as the isolated unit training, instead of entire utterances with transcribed phone sequences[17].

Following the steps in HTK Book[17], time-alignment for the train subset and the test subset can be implement to obtain automatic labelling files. Then, the unit of the start time and the end time for each phoneme should be changed from 100 nanoseconds to 1 millisecond, which has been presented in Subsection 6.1.2. The next, following the method shown in Fig. 6.4, a errors list file for all boundaries can be obtained for the proposed method, utilizing the time-alignment results for the train subset and corresponding manual labelling files. Based on the results of spectrum target prediction, the HMM-based labelling files for the test subset and the errors list file, the automatic labelling files by the proposed method can be obtained, according to the block diagram in Fig. 6.5.

The description in this Subsection can be concluded in Fig 6.6.

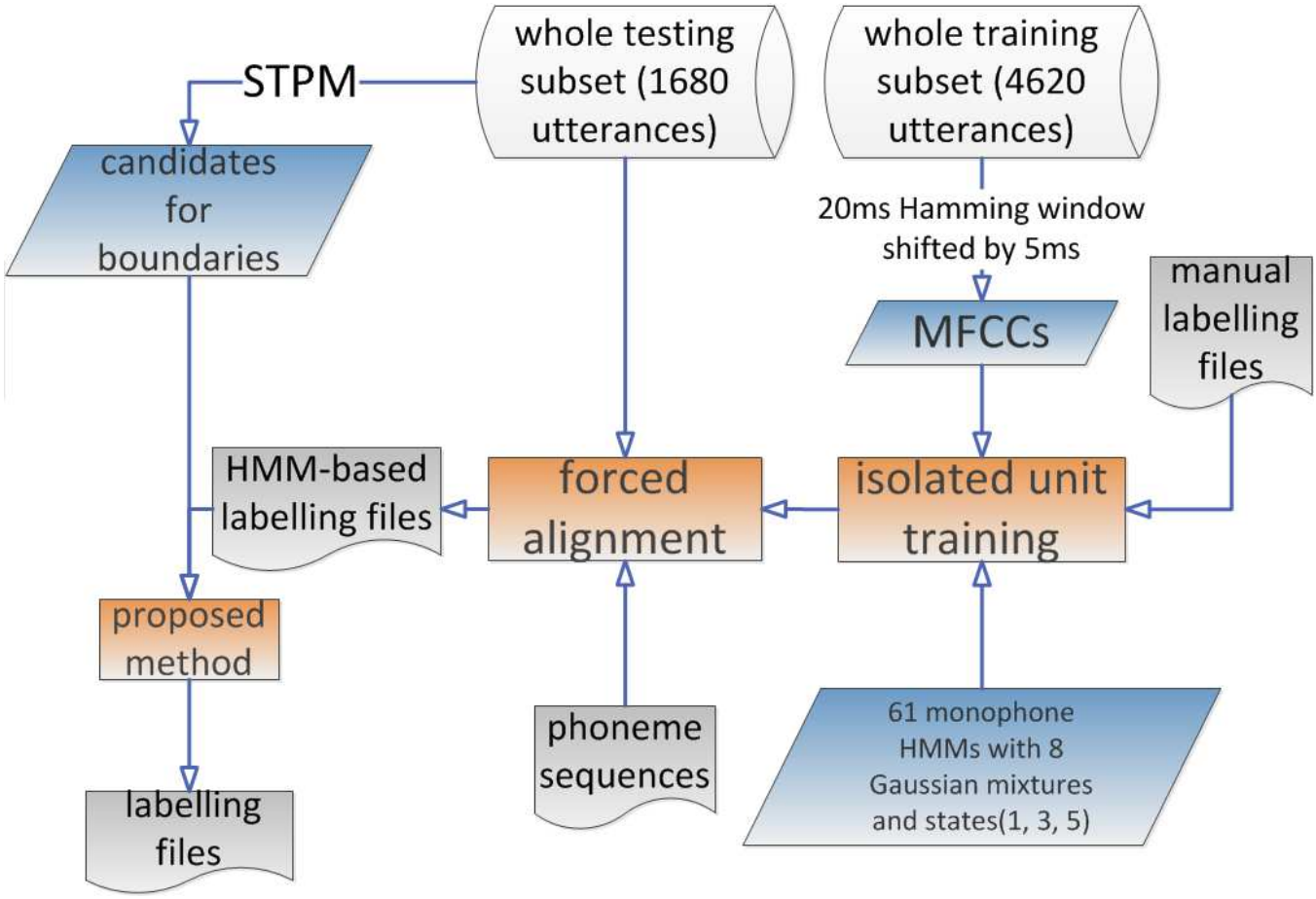

Figure 6.6: The block diagram for the experiments using TIMIT database. Firstly, the protos of 61 monophone HMMs are constructed by the rules presented in this Subsection and the MFCCs are extracted from utterances by a 20 millisecond Hamming window shifted by 5 millisecond, as the preparation. Then, utilizing the whole training subset and corresponding manual labelling files, isolated unit training for the protos is carried out. The next, the HMM-based labelling files will be obtained through forced alignment, utilizing the whole testing subset and corresponding phoneme sequences. Meanwhile, the whole training subset should be provided for forced alignment, to obtain a errors list file for the proposed method following the method shown in Fig. 6.4. The procedure above is processed in HTK. Also, the testing utterances are used to obtain candidates for the accurate phonemes boundaries by the means of spectrum target prediction model(STPM), similar to the experiments presented in Chapter 5. The last, the labelling files with more accurate phonemes boundaries are outputed following the block diagram in Fig. 6.5.

6.3.3 Results of experiments

An example labelling file is shown in Fig. 6.7, corresponding the HMM-based labelling file in Fig. 6.1. The different numbers in the second column for the same row in two files indicate the different boundaries obtained by two methods. These numbers will be used for evaluation.

```
0 598 h#
598 703 sh
703 804 iy
804 883 hv
883 1025 ae
1025 1063 dcl
1063 1080 d
1080 1093 y
1093 1169 er
```

Figure 6.7: An example labelling file obtained by the proposed method corresponding the HMM-based labelling file in Fig. 6.1.

6.4 Conclusion

The results of experiments indicate the effectiveness of the proposed method. By now, the second and third sub-goals have been achieved. For the last sub-goal, an appropriate objective evaluation method should be considered to verify that the proposed method is better than the HMM-based method for automatic speech segmentation.

Chapter 7

Objective Evaluation

In this chapter, an objective evaluation method is designed according to analysis for the results by the HMM-based segmentation method. Then, the objective evaluation will be carried out for the segmentation results by the proposed method, compared to the results by the HMM-based segmentation method.

This chapter is structured as follows. In Section 7.1, the results of statistical errors based on the rough automatic labelling files obtained by the HMM-based segmentation method is shown. The analysis for the results is also included in this section. Based on Section 7.1, the method for objective evaluation is constructed in Section 7.2. In the following two section, the theoretical upper-limitation evaluation results and experimental evaluation results for the proposed method are presented separately. According to the evaluation results for the proposed method, the discussion and conclusion are described in order.

7.1 Analysis for the HMM-based segmentation results

During the experiments for labelling for the proposed method, HMM-based segmentation results have been obtained. There are totally 62465 boundaries for 1680 sentences. According to the HMM-based labelling files and the corresponding manual labelling files, all errors can be calculated using equation "$error = boundary\_HMM - boundary\_manual$", where $boundary\_HMM$ is the end time of the phoneme in the HMM-based labelling files and $boundary\_manual$ is the corresponding end time in the manual labelling files. Then, the numbers of boundaries for the same error can be calculated statistically based on the results in the last step. Then, the distribution for the numbers of boundaries for 2414 different errors can be obtained, as shown in Fig. 7.1. The horizontal axis indicates the different errors (unit is set to 1 millisecond) and the vertical axis indicates the numbers of the boundaries for the same error.

As is well-known, the mean value and the variance value are used to analyse Gaussian distribution. However, the distribution for errors is not Gaussian distribution. Thus, other values should be considered instead of the mean value and the variance value. In

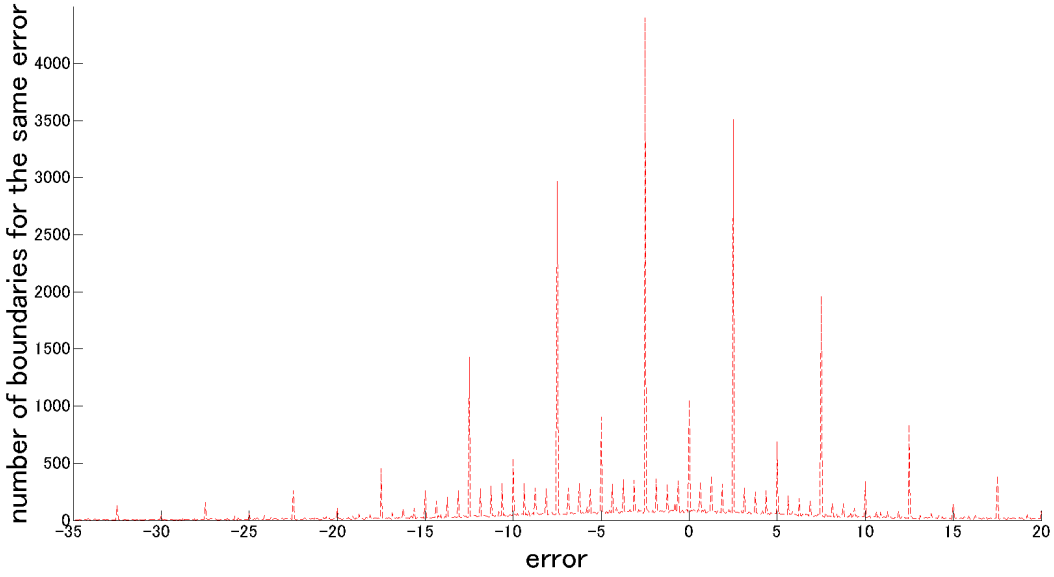

Figure 7.1: The distribution for the numbers of boundaries for the same error in HMM-based segmentation results

this research, the quartiles[40] are selected, which are the three points that divide the data set into four equal groups in descriptive statistics. The three points are called the first quartile (lower quartile), the median (second quartile), the third quartile (upper quartile) ordered by value from small to big. The quartile deviation[41], which is half the difference between the first and third quartile, measure the spread of a distribution usually used in non-linear situations and non-normal distributions, for it is not influenced by extremely high or extremely low scores.

7.2 Evaluation method

According to the analysis procedure for HMM-based segmentation results in last section, there are three values of errors to be thought highly of: the median, the first quartile, the third quartile. The median can be treated as a value of expectation. Obviously, it is the best situation that the median equals to 0, which indicates that the automatic segmentation boundaries make no errors compared to the hand-made boundaries in expectation. Between first quartile and third quartile there is a range making sure 50% of the automatic segmentation boundaries make small errors. Thus, the length of this range indicates the concentration level of the automatic segmentation boundaries which make small errors. The quartile deviation presented in the last section can be used for the measurement.

In addition, as presented in subsection 6.1.2, the accuracy of automatic segmentation is generally measured in terms of what percentage of the automatically labelled boundaries

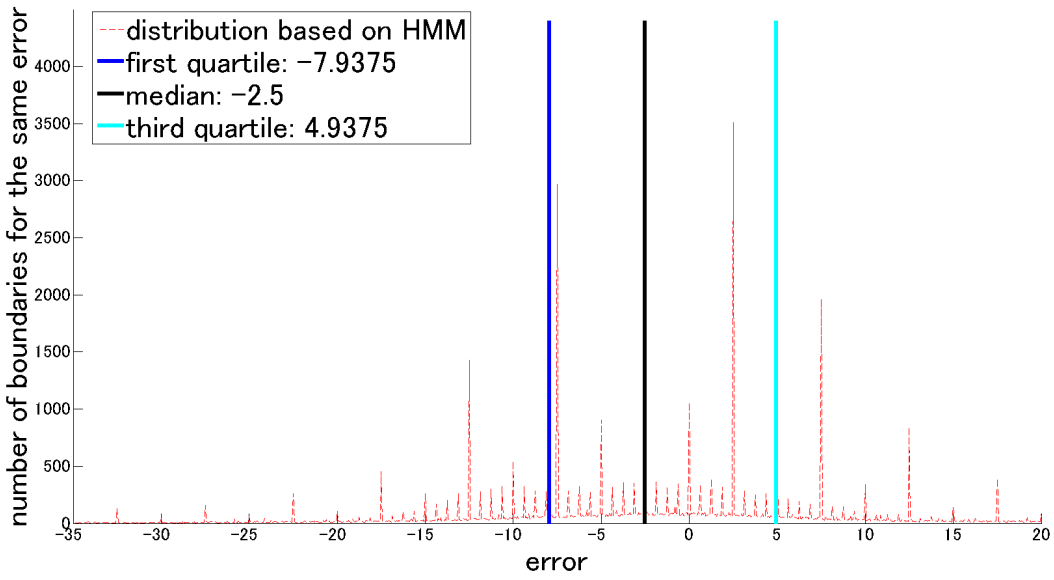

Figure 7.2: Analysis for HMM-based segmentation results based on the distribution for the numbers of boundaries for the same error

are within a given time threshold (tolerance) of the manually labelled boundaries. 20 milliseconds has been most widely used as a tolerance for measuring phone segmentation quality.

The results by the proposed method is compared to the standard HMM-based segmentation results and the distribution for the numbers of boundaries for the same error between manual results and HMM-based segmentation results has been known. Based on the discussion in the last two paragraph, for three evaluation criterion (the median, the quartile deviation, the percentage of the automatically labelled boundaries for 20 milliseconds tolerance), the proposed method should do better than HMM-based segmentation method for the three criterion. The ideal evaluation results is shown in Fig. 7.3, which indicates the proposed method improves the accuracy of segmentation boundaries compared to the standard HMM-based segmentation results for all three evaluation criterion: as well as that the median is moved to the position close to 0, the quartile deviation is smaller and the percentage of the automatically labelled boundaries for 20 milliseconds tolerance is bigger than ones in HMM-based segmentation results.

7.3 Theoretical upper-limitation evaluation results

According to the evaluation method presented in Section 7.2, the theoretical upper-limitation evaluation results are shown in Table 7.1. Assuming that manual labelling files in TIMIT database could be used as references for determining the boundaries from all candidates obtained by spectrum target prediction model, as shown in Fig. 5.2, the

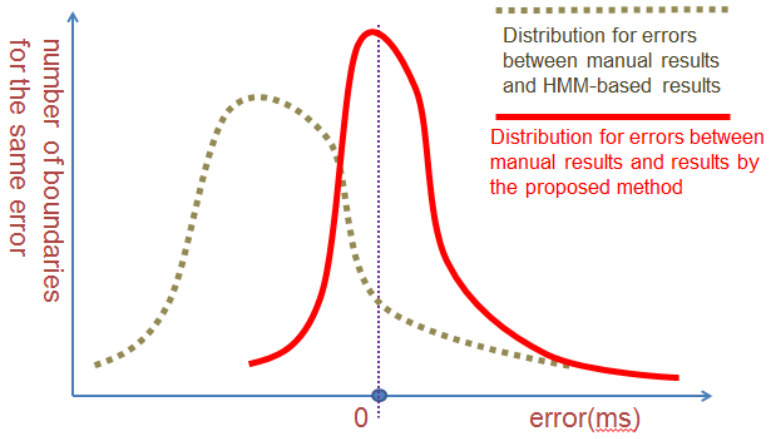

Figure 7.3: Ideal results for the proposed method

nearest peaks from the manual labelling files are always selected as the boundaries, to verify whether the highly accurate automatic boundaries can be obtained. In Table 7.1, the percentages of the automatically labelled boundaries for 5 milliseconds tolerance and 10 milliseconds tolerance are also reported and the explanation for the items in the first column has been presented in the Section 7.2. The unit of the other numbers is 1 millisecond.

Table 7.1: Theoretical upper-limitation evaluation results.

Median	-0.2188
First quartile	-1.5313
Third quartile	1.4062
Quartile deviation	1.46875
5(ms) tolerance	94.24%
10(ms) tolerance	99.92%
20(ms) tolerance	100%

7.4 Experimental evaluation results for the proposed method

Before evaluation for the proposed method, forced alignment boundaries obtained by HMM-based method are adjusted by a simple correction procedures. The HMM-based boundaries are shifted to new positions based on the mean errors in the errors list file, to achieve better results for evaluation.

The analysis results are summarized in Table 7.2. Different with theoretical upper-limitation evaluation results, the percentages of the automatically labelled boundaries for 10 milliseconds tolerance and 50 milliseconds tolerance are reported, because of the

actual accuracy in the experiments. The unit of the other numbers is 1 millisecond. The explanation for the items in the first column has been presented in the Section 7.2.

Table 7.2: Evaluation results for the proposed method.

	HMM-based method	The proposed method
Median	-0.2336	-0.0025
First quartile	-5.6480	-3.3725
Third quartile	5.4708	3.2875
Quartile deviation	5.5594	3.3300
10(ms) tolerance	71.58%	76.18%
20(ms) tolerance	90.02%	92.07%
50(ms) tolerance	98.53%	99.05%

7.5 Discussion for evaluation results

1. Compared to the analysis results for HMM-based segmentation boundaries, according to the three evaluation criterion described in Section 7.2, the proposed method performs automatic speech segmentation better than HMM-based segmentation method. The median has moved to the position closer to 0, the quartile deviation has reduced about 40%, and the the percentages of the automatically labelled boundaries for 20 milliseconds tolerance has increased about 2%.

2. Although the proposed method improves the accuracy of segmentation boundaries compared to the standard HMM-based segmentation results, there are some drawbacks for the proposed method. The file recording the mean error for each boundary is the key factor to improve the accuracy of boundaries, however, it is also the key drawback for the proposed method. The file can not be commonly used because there will be different phone sets for different databases to make the file variable. If there are different files to be needed for different databases, two problem come out.

(1) As described in Subsection 6.1.2, the file is made using all automatic segmentation boundaries by some methods and corresponding hand-made boundaries. If a database is used for research, it can be believable that there are manual labelling files for corresponding speech, just as TIMIT corpus. However, the file is made statistically. Thus, the scale of the databases should be big enough to make sure that the file can be used for the proposed method.

(2) Actually, the automatic speech segmentation method is designed for time-alignment using speech signal and a corresponding phoneme sequence file, instead of manual segmentation method. In such situation, there must be no hand-made boundaries for comparison to obtain such file.

In conclusion, a reliable commonly-used file listing the errors for all possible boundaries is essential when using this proposed method. However, considering different languages and dialects for same language, to obtain such file is a hard task.

3. In the experiments, only the standard HMM-based segmentation results are used for comparison. In fact, there have been some methods for adjusting the standard HMM-based segmentation boundaries[18–23], as presented in Section 1.2. Thus, the proposed method should be also compared to these methods. Besides, the proposed method should be implemented using more databases in English or other language.

4. Considering the motivation of this research, there should be a subjective evaluation method for the proposed method. For example, speech synthesis is carried out using the segments obtained by the proposed method, and then the naturalness of synthesized speech will be checked through listening tests.

5. According to the theoretical upper-limitation evaluation results, the percentage of the automatically labelled boundaries within 20 milliseconds threshold (tolerance) of the manually labelled boundaries has achieved 100%. The percentages of the boundaries with 10 milliseconds tolerance and 5 milliseconds tolerance are also extreme high. Besides, the quartile deviation is only 1.46875 milliseconds. All the evidences could be used for verifying that it is possible to obtain automatic segmentation boundaries similar as manual boundaries through spectrum target prediction model. However, by the proposed method in this research, the evaluation results reveal that the most suitable candidates are not always selected as phonemes boundaries. By optimizing the proposed method, there is great potential for this research.

7.6 Conclusion

According to the discussion for the evaluation results, the proposed method improves the accuracy of segmentation boundaries compared to the standard HMM-based segmentation results, which indicates the achievement of the last sub-goal presented in Section 1.4 and the achievement of the purpose of the proposed method. However, there have still be plenty of work for further research, such as the subjective evaluation and the comparison to other automatic segmentation methods. Especially the errors list file, used to estimate the boundaries in human perception, is the key factor to improve the accurate of phoneme boundaries compared to the HMM-based method. Thus, to prove the possibility to obtain a reliable commonly-used errors list file for all possible boundaries is a essential and hard task for extending the proposed method to all databases. Besides, compared to the theoretical upper-limitation evaluation results, how to optimizing the proposed method is also the future work to achieve higher performance in the domain of automatic speech segmentation.

Chapter 8

Conclusion and future work

In this thesis, a method based on human perception characteristics is proposed to obtain more precise phoneme boundaries close to hand-made boundaries for automatic speech segmentation, in order to supply automatic segmented speech corpus for highly natural speech synthesis or some other applications with the requirement for highly accurate and reliable speech segmentation. Utilizing the spectrum target prediction model, some candidates for extracting the phoneme boundaries can be calculated out. The results of experiments for segmentation indicates the effectiveness of the spectrum target prediction model for the purpose of this research. Based on the results above, the strategy for extracting the phoneme boundaries have constructed, meanwhile, the scheme for labelling method is established. Furthermore, evaluation is carried out based on some relevant experiments for the strategy. The experiments results indicated that the proposed method improves the accuracy of segmentation boundaries compared to the standard HMM-based segmentation results, by refining HMM-based segmentation boundaries. Above all, the propose of this research can be achieved along with the implementation for all sub-goals.

However, considering the proposed method may be provided for different databases, a reliable commonly-used file listing the errors for all possible boundaries should be obtained by some means in the future work. Another important future work is the subjective evaluation for the proposed method. In addition, more experiments and comparison to other automatic speech segmentation methods should be carried out for the revise of the proposed method. Besides, compared to the theoretical upper-limitation evaluation results, how to optimizing the proposed method is also the future work to achieve higher performance in the domain of automatic speech segmentation.

Acknowledgements

With the completion of this thesis, my master course at School of Information Science in JAIST will be close to the end. I would like to express my sincere gratitude to all those who have helped me for my study and research during the past year.

My deepest gratitude goes first and foremost to my supervisor, Professor Masato Akagi, who gave me the chance to come to JAIST and led me to enter into the world of speech processing. Plenty of instructive advice, useful suggestions, insightful criticism and professional guidance provided by him make me try my best to explore my potential in this research. During the preparation of this thesis, he has generously spent much time reading through each draft and provided me with inspiring advice. Without his consistent and illuminating instruction, the completion of this thesis would not be possible.

Secondly, I would like to give my heartfelt gratitude to Professor Jianwu Dang, Associate Professor Masashi Unoki and Assistant Professor Ryota Miyauchi, for their invaluable comments and suggestions on my research. Any progress that I have made is the result of their help, especially when I met bottleneck during my study and research.

I am also greatly indebted to all the teachers who once offered me valuable courses and advice during my study at School of Computer Software in Tianjin University in China. Two teacher to be specially mentioned, Associate Professor Yikui Zhang, has been providing me invaluable help on my study and life, as well as Lecturer Hongcui Wang, inspired my interest on speech processing.

Special thanks should go to my friends, Yi Kong, Yeteng An, and one of Yeteng An's senior students (sincere apology for that I have not known his name by now). They kindly gave me a hand when I was in trouble with utilizing HTK and Unix server during my research. Their profound knowledge and selfless devotion has assist me through the hard periods.

Last but not the least, my thanks would go to my beloved family for their loving considerations and great confidence in me all through these years. My presents have always been helping me out of difficulties and supporting without a word of complaint, which has always been the biggest motive power for my study.

Once again, I would like to give my sincere admiring to my supervisor, Professor Masato Akagi. His rigorous, conscientious and earnest attitude to the research has been impressing me since I came to JAIST, which would be invaluable wealth for my future work, study and life.

Reference

[1] K. Zechner. *Summarization of Spoken Language Challenges, Methods, and Prospects*, in Technology Expert eZine, Jan. 2002, Issue 6

[2] Y.-F. Ma, L. Lu, H.-J. Zhang and M. J. Li. *Attention Model for Video Summarization*, in 10th ACM International Conference on Multimedia, 2002

[3] K. Koumpis and S. Renals. *role of Prosody in a Voicemail Summarization System*, in Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding

[4] Speaker diarisation. *http://en.wikipedia.org/wiki/Speaker_ diarisation*

[5] Robert B. Dunn, Douglas A. Reynolds, and Thomas F. Quatieri. *Approaches to Speaker Detection and Tracking in Conversational Speech*, in Conversational Speech, Digital Signal Processing 10, 2000, pp. 93-112

[6] Speech recognition. *http://en.wikipedia.org/wiki/Speech_ recognition*

[7] A. J. Hunt and A. W. Black. *Unit selection in a concatenative speech synthesis system using a large speech database*, in Proc. ICASSP, Atlanta, GA, 1996, pp. 373-376

[8] B.Möbius. *The Bell Labs German text-to-speech system*, in Comput. Speech Lang., 1999, vol. 13, pp. 319-358

[9] A. K. Syrdal, C. W.Wightman, A. Conkie, Y. Stylianou, M. Beutnagel, J. Schroeter, V. Strom, K.-S. Lee, and M. J. Makashay. *Corpus-based techniques in the AT& T NextGen synthesis system*, in Proc. ICSLP, Beijing, China, Oct. 2000, vol. 3, pp. 410-415

[10] G. Coorman, J. Fackrell, P. Rutten, and B. Van Coile. *Segment selection in the L& H Realspeak laboratory TTS system*, in Proc. ICSLP, Beijing, China, Oct. 2000, vol. 2, pp. 395-398

[11] D. T. Toledano, L. A. H. Gömez, and L. V. Grande. *Automatic phonetic segmentation*, in IEEE Trans. Speech Audio Process., Nov. 2003, vol. 11, No. 6, pp. 617-625

[12] S. S. Park, and N. S. Kim. *On Using Multiple Models for Automatic Speech Segmentation*, in IEEE Trans. Audio, Speech, Lang. Process., Nov. 2007, vol. 15, No. 8, pp. 2202-2212

[13] H. Kawai and T. Toda. *An evaluation of automatic phone segmentation for concatenative speech synthesis*, in Proc. ICASSP, Montreal, QC, Canada, 2004, vol. I, pp. 677-680

[14] S. Nefti and O. Boëffard. *Acoustical and topological experiments for an HMM-based speech segmentation system*, in Proc. Eurospeech, Aalborg, Denmark, 2001, pp. 1711-1714

[15] J. Matouseˇk, D. Tihelka, and J. Psutka. *Automatic segmentation for Czech concatenative speech synthesis using statistical approach with boundary-specific correction*, in Proc. Eurospeech, Geneva, Switzerland, 2003, pp. 301-304

[16] J. Adell, A. Bonafonte, J. A. Gömez, and M. Castro. *Comparative study of automatic phone segmentation methods for TTS*, in Proc. ICASSP, Philadelphia, PA, 2005, vol. I, pp. 309-312

[17] S. Young, G. Evermann, M. Gales, T. Hain, X. Liu, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK Book (for HTK Version 3.4)*, Cambridge, U.K.: Cambrige Univ., March, 2009

[18] Y. -J. Kim and A. Conkie. *Automatic segmentation combining an HMM-based approach and spectral boundary correction*, in Proc. ICSLP, 2002, pp. 145-148

[19] D. T. Toledano. *Neural network boundary refining for automatic speech segmentation*, in Proc. ICASSP, 2000, pp. 3438-3441

[20] K. S. Lee. *MLP-based phone boundary refining for a TTS database*, in IEEE Trans. Audio, Speech, Lang. Process., May 2006, vol. 14, no. 3, pp. 981-989

[21] H. -Y. Lo and H-M. Wang. *Phonetic boundary refinement using support vector machine*, in Proceedings of ICASSP, 2007, pp. 933-936

[22] J. P. Hosom. *Speaker-independent phoneme alignment using transition-dependent states*, in Speech Communication, 2009, vol. 51, pp. 352-368

[23] J. Yuan, N. Ryant, M. Liberman, A. Stolcke, V. Mitra, and W. Wang. *Automatic Phonetic Segmentation using Boundary Models*, in Proceedings Interspeech, 2013

[24] P. T. Brady, A. S. House, and K. N. Steven. *Perception of Sounds Characterized by a Rapidly Changing Resonant Frequency*, in J. Acoust. Soc. of American, 1961, vol. 33, pp. 1357-1362

[25] B. E. F. Lindblom, and M. Studdert-Kennedy. *On the Role of Formant Transition in Vowel Recognition*, in J. Acoust. Soc. of American, 1967, vol. 42, pp. 830-843

[26] H. Kuwabara. *An Approach to Normalization of Coarticulation Effects for Vowels in Connected Speech*, in J. Acoust. Soc. of American, 1985, vol. 77, pp. 686-694

[27] M. Akagi. *Evaluation of a spectrum target prediction model in speech perception*, in J. Acoust. Soc. of American, 1990, vol. 87, pp. 858-865

[28] M. Akagi, and Y. Tohkura. *Spectrum target prediction model and its application to speech recognition*, in Computer Speech and Language, 1990, vol. 4, pp. 325-344

[29] T. Aritsuka, M. Akagi and S. Katagiri. *Speech recognition using spectrum target prediction model as a front-end processor*, in Speech Group Tech. Report, IEICEJ, SP91-36

[30] A. Kurematsu, K. Takeda, Y. Sagisaka, S. Katagiri, H Kuwabara, K. Shikano. *ATR Japanese speech database as a tool of speech recognition and synthesis*, in Speech Communication, August, 1990, Vol. 9, Issue 4, pp. 357-363

[31] TIMIT. *http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1*

[32] Julius. *http://julius.sourceforge.jp/*

[33] H. Fujisaki, M. Yoshida, Y. Sato, and Y. Tanabe. *A Model of the Coarticulatory Process in the Formant Frequency Domain and its Application to Recognition of Connected Vowel*, in Tech. Rep. Acoust. Soc. Jpn, S73-03

[34] S. Itahashi, and S. Yokoyama. *Description and Segmentation of Formant Trajectory with Second Order Linear System Model*, in Electrotech. Lab. Rev., 1976, vol. 40, No. 6, pp. 530-541

[35] MFCC. *http://en.wikipedia.org/wiki/Mel-frequency_ cepstrum*

[36] C. Lopes, and F.Perdig$\widetilde{a}$o. *Phone Recognition on the TIMIT Database*, in Speech Technologies, 2011, pp. 285-302

[37] R. Yan, Y. Zu, Y. Zhu. *AUTOMATIC SPEECH SEGMENTATION COMBINING AN HMM-BASED APPROACH AND RECURRENCE TREND ANALYSIS*, in Proc. ICASSP, 2006, vol. I, pp. 797-800

[38] J. -W. Kuo, H. -Y. Lo, and H. -M. Wang. *Improved HMM/SVM Methods for Automatic Phoneme Segmentation*, In Proceedings Interspeech, 2007, pp. 2057-2060

[39] J. Dines, S. Sridharan and M. Moody. *AUTOMATIC SPEECH SEGMENTATION WITH HMM*, In Proceedings of the 9th Australian International Conference on Speech Science & Technology, Melbourne, Dec. 2002, pp. 544-549

[40] R. J. Hyndman, and Y. Fan. *Sample quantiles in statistical packages*, in American Statistician, Nov., 1996, vol. 50, No. 4, pp. 361-365

[41] Quartile derivation. *http://wiki.answers.com/Q/What_ is_ quartile_ deviation*