Computational Linguistics and Chinese Language Processing
Vol. 10, No. 2, June 2005, pp. 145-166

Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpora for Concatenation-based TTS

Cheng-Yuan Lin*, Jyh-Shing Roger Jang* and Kuan-Ting Chen*

Abstract

Precise phone/syllable boundary labeling of the utterances in a speech corpus plays an important role in constructing a corpus-based TTS (text-to-speech) system. However, automatic labeling based on Viterbi forced alignment does not always produce satisfactory results. Moreover, a suitable labeling method for one language does not necessarily produce desirable results for another language. Hence in this paper, we propose a new procedure for refining the boundaries of utterances in a Mandarin speech corpus. This procedure employs different sets of acoustic features for four different phonetic categories. In addition, a new scheme is proposed to deal with the "periodic voiced + periodic voiced" case, which produced most of the segmentation errors in our experiment. Several experiments were conducted to demonstrate the feasibility of the proposed approach.

Keywords: speech assessment methods phonetic alphabet, speech corpus, sequential forward selection, k-nearest neighbor rule, leave-one-out, speaker-adapted model, context-dependent hidden Markov model (HMM).

1. INTRODUCTION

Corpus-based speech synthesis systems are becoming more and more popular due to the high degree of fluency achieved and the natural feel of the generated speech. However, such systems always require a significant amount of human effort in labeling the phonetic boundaries of the corresponding corpus [Van Erp *et al.* 1988] [Wang *et al.* 1999] [Cosi *et al.* 1991]. Therefore, a great deal of research on automatic phonetic labeling methods has been conducted over the past several years [Ljolje *et al.* 1993, 1994] [Demuynck *et al.* 2002]. In general, most of these methods involve the following two steps:

(1) rough phonetic segmentation by means of Viterbi forced alignment using HMM (hidden Markov models) or other statistical methods;

* Multimedia Information Retrieval Laboratory, Dept. of Computer Science, National TsingHua University, Hsing-Chu, Taiwan, Tel: +88635715131-3506
E-Mail: {gavins, jang, marco}@wayne.cs.nthu.edu.tw

[Received December 8, 2004; Revised April 6, 2005; Accepted April 6, 2005]

(2) high time-resolution analysis of the phonetic boundaries using boundary checking rules.

These HMM-based recognizers can be categorized in various ways. For example, some use context-dependent HMM, while others use context-independent HMM [Makashay *et al.* 2000]. Also, there are various types of HMM training methods, including speaker-dependent (SD), speaker-independent (SI), and speaker-adapted (SA) models. Although the HMM-based speech recognizer using MFCCs (mel-frequency cepstral coefficients) is well known for its excellent speech recognition, ability, its use of automatic phonetic segmentation and labeling does not always produce precise and satisfactory results necessary for the development of TTS. As a result, other acoustic features and refinement algorithms have been proposed in the literature to improve the phonetic labeling results obtained from HMM-based recognizers.

Several works have focused on automatic phonetic labeling, in the last few years. For example, in [Bonafonte *et al.* 1996], Bonafonte *et al.* took Gaussian probability density distribution as a similarity measure. In [van Santen *et al.* 1990], Jan P. H. van Santen *et al.* adopted broad-band and narrow-band edge detection. In [Torre Toledano *et al.* 1998], Toledano *et al.* tried to mimic human labeling using a set of fuzzy rules. In [Sethy *et al.* 2002], Sethy *at al.* employed adapted CDHMM (continuous density hidden Markov model) models [Lamel *et al.* 1993]. The main focus of all of these studies has been English speech, and they have seldom addressed the question of which phonetic class tends to be more error prone. Moreover, the methods proposed in the above papers may not perform equally well when dealing with another language. For example, most approaches for English utterance segmentation can be divided into two categories: rule-based [Torre Toledano *et al.* 1998] and statistics-based [Sethy *et al.* 2002] methods. For a rule-based approach, one needs to define a set of rules (crisp or fuzzy) for various phonetic transitions. For a statistics-based approach, one needs to collect a sample data set and label the set accordingly. Conceptually, the rule-based approaches for English corpora can be adapted for application to Chinese corpora. But in fact, it is hard to design such a system without the aid of human experts who have a thorough understanding of the similarities and differences between the phonetic sets of these two languages. It is our belief that the above two approaches should be used in a seamless, integrated manner. As a result, we have developed a hybrid approach, where most of the boundaries are identified via statistical pattern recognition (Sequential Forward Selection, K-Nearest Neighbor Rule and Leave-One-Out) [Whitney 1971] [Duda *et al.* 2001], while the most difficult cases (periodic voiced + periodic voiced) are handled using a rule-based approach.

Mandarin Chinese is a tonal language, and each character is associated with one or several syllables. A Chinese syllable is either composed of a CV (Consonant-Vowel or INITIAL-FINAL [Chou *et al.* 2002] [Lee 1997]) structure or a single V (Vowel) structure. Therefore, the primary effort in speech labeling focuses on precisely identifying the

boundaries of each syllable. Then the boundary between a consonant and a vowel within a syllable can be identified according to the type of a consonant. In most cases, the consonant is fricative, affricate, or plosive, and the consonant can easily be distinguished using several acoustic features other than MFCCs, such as zero-crossing rate or pitch, etc. If the consonant is periodic, as in the case of " ㄌ " ("l" in SAMPA (http://www.phon.ucl.ac.uk/home/sampa/home.htm), the acronym of 'Speech Assessment Methods Phonetic Alphabet'), then the consonant does not need to be segmented, and the whole syllable should be treated as a single unit for TTS, since further operations involving pitch or time scale modification should be performed on both the consonant and the vowel.

In [Chou *et al.* 1998, 2002], Chou *et al.* proposed an SD-based HMM model plus simple boundary correction rules for Mandarin Chinese. However, to construct this system is time consuming because of the iterative procedure used for forced alignment, the correction rules and, re-training. In addition, it becomes particularly inefficient if the speech corpus is updated incrementally and regularly, such as by adding one hour of speech data per week. Furthermore, the SD-based HMM model may not outperform the SA-based HMM if the size of the training data is moderate; for example, there is one hour of data for the same speaker.

In this paper, we propose an SA-based HMM recognizer that performs a forced alignment first and then employ a refinement procedure to modify the identified boundaries. The proposed refinement procedure uses several innovative acoustic features to refine boundaries for various phonetic categories. These approaches and experimental results obtained using them will be described in the following sections.

This paper is organized as follows. Section 2 introduces our forced alignment procedure that uses an HMM recognizer to get initial estimations of all boundaries. Section 3 explains the refinement procedure specially designed for four phonetic categories and describes acoustic features are chosen by the SFS (Sequential Forward Selection) [Whitney 1971] algorithm. Section 4 describes the experiments conducted to demonstrate the performance of the proposed refinement procedure, and presents error analysis of irretrievable errors. Section 5 draws conclusions and discusses future work.

2. HMM BASED RECOGNIZER

2.1 From Orthographic Transcription to Phonetic Transcription

Forced alignment using the HMM-based recognizer relies on knowledge of the underlying phonetic transcription of a given utterance. In general, once the orthographic transcription and speech data are both available, we can employ forced alignment for automatic phonetic transcription. However, some commonly used Chinese characters have multiple syllables with different pronunciations, depending on the lexical contexts; For instance, the Chinese

character “重” (meaning “heavy”) is pronounced “ㄓㄨㄥˋ”(“TS-U-@N, 4th tone” in SAMPA) in “重要” (meaning “important”) and “ㄔㄨㄥˊ”(“TS_h-U-@N, 2nd tone” in SAMPA) in “重疊” (meaning “overlap”). As a result, word segmentation in the text sentence is necessary for correct phonetic transcription for the purpose of alignment. Commonly used approaches to word segmentation in Chinese NLP (natural language processing) include the forward or backward maximum word matching algorithm [Chen *et al.* 1992][Yeh *et al.* 1991], and the dynamic-programming-based statistic probability method [Sproat *et al.* 1990]. However, no word segmentation algorithm can guarantee perfect results for the following reasons:

(1) Word segmentation relies on a collection of Chinese words in the form of a dictionary, which cannot cover all existing words since new words are constantly being created.

(2) Even if the word dictionary were complete, some pronunciations could not be determined through dictionary lookup, especially for the case of Chinese poems. For instance, the first character of “朝辭白帝彩雲間” (meaning “leaving Baidi city in colored dawn”) is pronounced “ㄓㄠ” (“TS-au, 1st tone” in SAMPA, meaning “dawn”), not “ㄔㄠˊ” (“TS_h-au, 2nd tone” in SAMPA, meaning “to head for”). This error cannot be corrected through dictionary lookup since “朝” is a single-character word meaning “morning”.

(3) Conflicts in word segmentation can lead to different results. For instance “老掌櫃順手把錢揣在懷裡” (meaning “the old shopkeeper smoothly slipped the money into his pocket”) will be labeled as “老 掌櫃 順手 把 錢 揣 在 懷裡” (meaning “the old + shopkeeper + smoothly + slipped + the money + into + his pocket”) if forward maximum word matching is used. On the other hand, it will be labeled as “老 掌櫃 順 手把 錢 揣 在 懷裡” (meaning “The old + shopkeeper + smoothly + handle bar + the money + into + his pocket”) if the backward approach is adopted.

In order to avoid errors resulting from phonetic transcription, we perform the following two steps to achieve a better performance:

(1) We perform word segmentation using forward and backward maximum matching based on a word dictionary containing around 90,000 entries. We keep the phonetic transcriptions as candidates for use in the next step. (If the result is the same, then we have only a single phonetic transcription.)

(2) We expand the list of obtained phonetic transcription candidates by adding possible syllables for polyphonic characters that are not found in any of the words obtained through the above word segmentation process. We use these different phonetic transcription candidates to perform a forced alignment through Viterbi decoding. We accept the phonetic transcription that has the maximum log likelihood.

The above steps combine both word segmentation in NLP and forced alignment in speech recognition to achieve better phonetic transcription performance. When the TTS-455 speech corpus with about 6,000 Chinese syllables was used, the syllable error rate was 2.1% and 1.9% for forward and backward maximum matching, respectively. With the addition of step 2, the error rate was reduced to 1.0%, which represents a significant reduction of 50% in the error rate. Some of the error cases are shown in Table 1.

Table 1. Labeling errors when orthographic transcription was transformed to phonetic transcription.

Text sentences of speech corpus.	Human transcription	Machine transcription
春風秋月何時『了』	ㄌㄧㄠˇ ("l-I-au, 3rd tone")	ㄌㄜ˙ ("l-@, 5th tone")
他囊『括』七面金牌	ㄎㄨㄛˋ ("k_h-U-o, 4th tone")	ㄍㄨㄚ ("k-U-a,1st tone")
道『行』高深的老僧 掐指一算就知道對方的來意	ㄏㄤˊ ("x-aN, 2nd tone")	ㄒㄧㄥˊ ("6-I-@N, 2nd tone")

Note: Symbols in parentheses are described in SAMPA.

The last character of the first sentence is a typical single character having multiple pronunciations that cannot be identified through word dictionary lookup. Unfortunately, forced alignment cannot find the correct phonetic transcription, either, because the utterance itself is ambiguous and unclear. The second sentence demonstrates the inadequacy of the word dictionary since "括" in "囊括" (meaning "to obtain") is labeled "ㄍㄨㄚ" ("k-U-a, 1st tone"in SAMPA) in the dictionary, while it is also pronounced "ㄎㄨㄛˋ"("k_h-U-o, 4th tone" in SAMPA) colloquially. The error from the third sentence indicates the inadequacy of the word dictionary; the word "道行" (meaning "capability" or "achievement") should be in the word dictionary, but it is not.

2.2 Speech Corpus Introduction

Once a phonetic transcription is obtained, we can perform forced alignment by using a HMM recognizer. In this study, we used two Mandarin Chinese speech corpora:

(1) TTS-455 speech corpus: This corpus contains 455 sentences spoken by one speaker and covers about 6,000 syllables. It is mainly for TTS. The details are as follows:

I. time duration: 30 minutes (66MB of disk space);

II. sampling rate and bit rate: 20,000 Hz, 16bits;

III. base syllables: 408;

IV. tonal syllables: 1196.

More information on this corpus can be found in (http://speech.cs.nthu.edu.tw/gavins/Research/SpeechSynthesis/content_hsf455.txt).

(2) TCC-300 speech corpus (http://rocling.iis.sinica.edu.tw/ROCLING/MAT/Tcc_300brief.htm): It contains sentences spoken by 300 subjects from National Taiwan University, Chiao Tung University, and Cheng Kung University in Taiwan. The recorded texts were selected from the "Academia Sinica Balanced Corpus" (http://www.sinica.edu.tw/~tibe/2-words/modern-words).

In order to perform a forced alignment on the TTS-455 speech corpus, we need to train an HMM-based recognizer. This recognizer will be described in Section 4.

3. DESIGN OF THE REFINEMENT PROCEDURE

A post-processing scheme must be used to refine the identified syllable boundaries. Specifically, since a forced alignment is based on MFCCs only, it makes sense to use other acoustic features to enhance precision. As mentioned in Section 1, using either a rule-based or a statistics-based approach alone is inadequate. Therefore, we combine these two methods to deal with a Mandarin Chinese speech corpus. First of all, we divide all Chinese phonemes into four categories. Then, we determine which set is suitable for which method (rule-based or statistics-based) by applying pattern recognition techniques. These steps will be described in detail in the following subsections.

3.1 Four Phonetic Categories

There are 37 distinct phonetic alphabets in Mandarin Chinese. This makes it difficult to develop a general method that can be used to refine labeling between all possible phonetic transitions. Hence, we divide all Chinese phonemes into four categories according to their acoustic characteristics. These four categories are fricative and affricate, unaspirated stop, aspirated stop, and periodic voiced [Lu 2002], as listed below in SAMPA format and in the MPA (Mandarin Phonetic Alphabet) format:

- Fricative and affricate: (consonants only)

  (Fricative)

  - SAMPA: f x 6 S s
  - MPA: ㄈ ㄏ ㄒ ㄕ ㄙ

  (Affricate)

  - SAMPA: t6 t6_h TS TS_h ts ts_h
  - MPA: ㄐ ㄑ ㄓ ㄔ ㄗ ㄘ

- Unaspirated stop: (consonants only)
  - SAMPA: p t k
  - MPA: ㄅ ㄉ ㄍ
- Aspirated stop: (consonants only)
  - SAMPA: p_h t_h k_h
  - MPA: ㄆ ㄊ ㄎ
- Periodic voiced:

  (Consonants)
  - SAMPA: m n l Z
  - MPA: ㄇ ㄋ ㄌ ㄖ

  (Vowels)
  - SAMPA: a o @ e ai ei au ou an @n aN @N 2 I U y
  - MPA: ㄚ ㄛ ㄜ ㄝ ㄞ ㄟ ㄠ ㄡ ㄢ ㄣ ㄤ ㄥ ㄦ ㄧ ㄨ ㄩ

Fricative and affricate are combined in a single category is mainly because of the similarity of the acoustic characteristics. In particular, for any given syllable with an affricate or fricative consonant, according to our observations, the duration ratio between the aperiodic and periodic parts is almost constant; in addition, there usually exists a high zero-crossing rate at the aperiodic part. As for the periodic voiced category, we include both consonants and vowels since they both contain stable harmonic or pitch structures.

3.2 Feature Definition

In order to refine the boundaries identified by the HMM-based recognizer, we need to employ several acoustic features other than MFCCs. Some of these acoustic features are commonly used in speech processing; they include the zero-crossing rate, log energy, pitch, and entropy [Shen *et al.* 1998]. In addition, we also adopt two new acoustic features, the bisector frequency and the burst degree, to help identify boundaries more precisely.

3.2.1 Bisector Frequency

The bisector frequency is defined in equations (1) and (2):

$$freqIndex = \arg\min_{1<k<N} \left| \sum_{f=1}^{k} A_f - \frac{\sum_{f=1}^{N} A_f}{2} \right|, \quad (1)$$

$$bisectorFreq = \frac{freqIndex}{N} \times sampleRate\,, \tag{2}$$

where A_f is the amplitude of the f^{th} frequency component and there are N distinct frequency components in the spectrum. The key characteristic of the bisector frequency is that its value is smaller for a voiced frame but larger for an unvoiced frame. Thus, we can use this feature to distinguish unvoiced from voiced patterns. Although the zero-crossing rate can also be used to detect unvoiced patterns, it is not sufficiently robust, especially when the mean amplitude of an unvoiced frame deviates from zero. For example, in Figure 1, the second unvoiced part of the waveform can be better detected by means of the bisector frequency than the zero-crossing rate.

In our implementation, we normalize the value of this feature to the range [0,1] according to equation (3):

$$bisectorfreq = \left(\frac{bisectorfreq - lowfreq}{highfreq - lowfreq} \right), \tag{3}$$

where the values of *highfreq* and *lowfreq* are empirically set to be $\frac{sampleRate}{2} \times 0.8$ and 100, respectively.

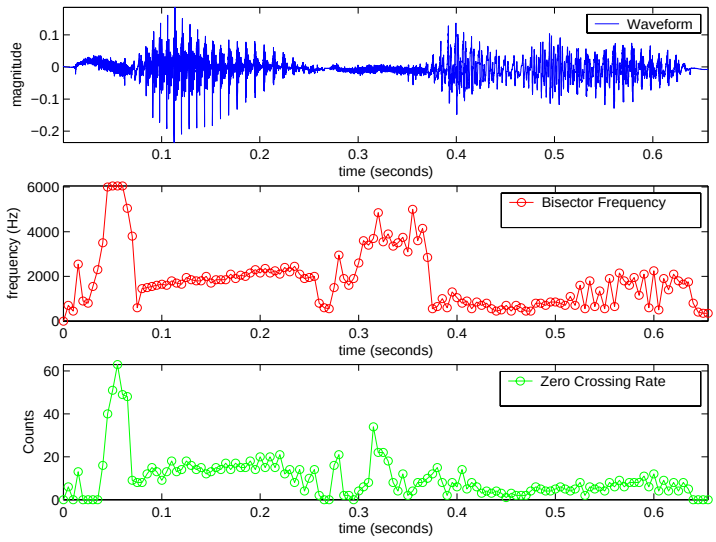

Figure 1. A comparison between the bisector frequency and the zero-crossing rate. The second unvoiced part of the waveform is better detected by means of the bisector frequency than the zero-crossing rate. The content of this waveform is "在視爲" ("ts-ai, S, U-ei" in SAMPA).

3.2.2 Burst Degree

It is difficult to recognize a burst pattern in speech using the zero-crossing rate and/or pitch. This is a stable pitch structure does not exist, and the zero-crossing rate is relatively low. To deal with this situation, we adopt a new feature called the burst degree, which is a weighted average between the log energy and the reciprocal average distance between the local maxima, as shown in equation (4):

$$burst\deg ree = \frac{\left(W_1 \times \frac{1}{avg(local\max Interval)} + W_2 \times \log Engergy \right)}{\left(W_1 + W_2 \right)}, \quad (4)$$

where W_1 and W_2 are two weighting factors with values of 4 and 1, respectively. The expression *avg*(*local* max *Interval*) is the average distance between the positions of neighboring local maxima of sample points. For instance, suppose that there are 4 local maxima located at positions 12, 52, 92 and 130 in a frame. Then, the intervals are 40, 40 and 38 and *avg*(*local* max *Interval*) is (40+40+38)/3. Figure 2 shows the result of the burst degree.

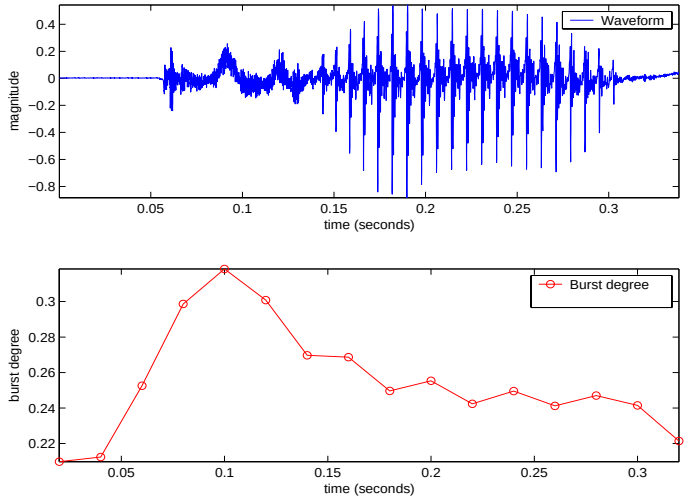

Figure 2. A speech waveform and its burst degree. The content of this waveform is "咖" ("k_h-a" in SAMPA).

3.3 Feature Selection Based on Phonetic Categories

In Section 3.1, we divided all phonemes into four phonetic categories. In this section, we divided boundaries into groups according to the transitions between phonetic categories. For instance, the boundaries of a given syllable with an aspirated stop consonant can be analyzed as follows:

(1) Beginning boundary: “silence + aspirated stop” or “vowel + aspirated stop”.

(2) Ending boundary: “vowel + silence” or “vowel + X”, where X is the consonant of the next syllable, which can be fricative and affricate, aspirated stop, unaspirated stop, or periodic voiced.

(3) INITIAL/FINAL boundary: “aspirated stop + vowel”. (The INITIAL/FINAL boundary is the boundary between the consonant and the vowel within a syllable. In our experiments, we did not try to find these kinds of boundaries since they were not the focus of this study. However, we still discuss all three kinds of boundaries for the sake of completeness.)

Based on similar analysis, we constructed Table 2 which lists all possible transitions from the left side to the right side for beginning, ending, and INITIAL/FINAL boundaries.

Table 2. All possible category transitions of beginning, ending, and Initial/Final boundaries.

Left side	Right side	Beginning boundary	Ending boundary	Initial/Final boundary
Silence	Fricative and affricate	O	X	X
Silence	Aspirated stop	O	X	X
Silence	Unaspirated stop	O	X	X
Silence	Periodic voiced	O	X	X
Fricative and affricate	Periodic voiced	X	X	O
Aspirated stop	Periodic voiced	X	X	O
Unaspirated stop	Periodic voiced	X	X	O
Periodic voiced	Silence	X	O	X
Periodic voiced	Fricative and affricate	O	O	X
Periodic voiced	Aspirated stop	O	O	X
Periodic voiced	Unaspirated stop	O	O	X
Periodic voiced	Periodic voiced	O	O	O

O: possible transition; X: impossible transition.

It is evident that not all features work equally well for each phonetic group. Therefore we must design an efficient method to distinguish the most outstanding among all possible features. In our experiment, we collected a speech corpus that contained about 2,100 syllables from 20 long sentences from speech lasting a total of 10 minutes. This corpus was used for feature selection and was fully independent of our speech corpus mentioned in Section 2.2. The syllables covered every Mandarin Chinese phoneme. The beginnings and endings of the phonetic boundaries of these 2,100 syllables were manually labeled. In the following we describe the steps we performed to find the best combination of features for each of these phonetic category transitions.

(1) In order to find the most discriminative features, we had to create a set of training data. This was done by adding several candidate boundaries, 10 ms apart, located within ±80 ms of a true (manually labeled) boundary. A candidate boundary was labeled "correct" if it was within ±20 ms of the true boundary. (According to [Chou *et al.* 2002], manual labeling by two human experts can achieve about 90% consistency with 10 ms tolerance and 100% with 20 ms tolerance.) Therefore, we chose to use 5 correct candidates, all within 20 ms of the manually labeled one, in our experiments. If we had chosen only one, then the number of "correct" data might have been too small, leading to an unbalanced sample data set. In other words, for each true boundary, we created a set of 17 candidate boundaries (including the true one), with 5 labeled "correct" and 12 labeled "wrong" as the desired classification output as shown in Figure 3.

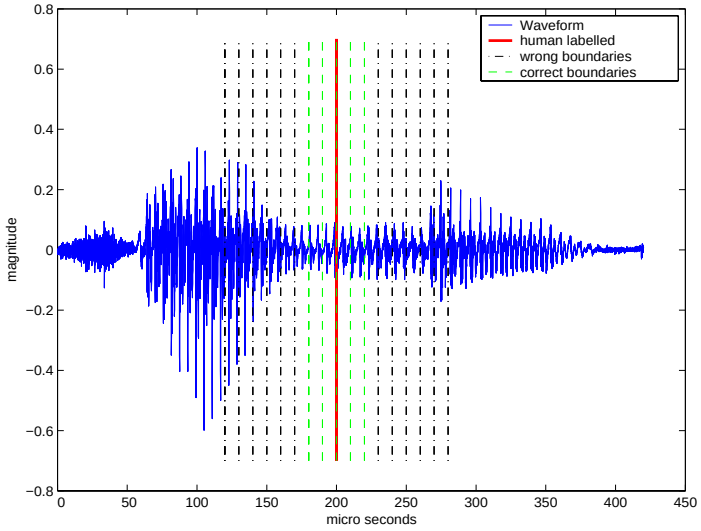

Figure 3. Training data of 5 correct boundaries and 12 wrong boundaries around the true boundary labeled by humans. The content of this waveform was "將離" ("t6-I-aN, l-I" in SAMPA).

(2) For each candidate boundary, we evaluated the differences between all the acoustic features of its left and right frames. The size of each frame was 20 ms, and the "difference of acoustic features" was then used a feature for designing a classifier.

(3) In order to find the most influential acoustic features, we employed the method of sequential forward selection (SFS) [Whitney 1971] in the literature on pattern recognition. The idea behind SFS is to start with a single feature having the best classification rate. Then, we can keep the already selected features and try to identify a newly added feature that can increase

the classification rate the most. For instance, if features 2 and 5 are the currently selected features, then we will try to find another feature that, when combined with the selected features, can produce the best classification rate. This greedy step is repeated until the desired number of features has been selected or until there is no further improvement in the classification rate. In order to use SFS, we need to select a classifier together with its performance evaluation scheme. Here, we used KNNR (K-Nearest Neighbor Rule) as the classifier and LOO (Leave-One-Out) [Duda *et al.* 2001] as the performance criterion. The basic idea behind 1-NNR is to assign the class of a given test vector as the data point in the training data that is nearest to the given vector. In order to achieve better robustness, we can choose KNNR, where the K nearest neighbors are selected around the test vector and the assigned class is determined by means of a voting mechanism among these K points. Then, we performed a simple search to find the best value of K in KNNR is 9 in our experiment. To evaluate the performance of KNNR, we apply LOO, where a vector is selected as the test vector and all the other data as the training data. This process is repeated until each data point has served as the test vector. The final classification rate is the overall classification rate of these test vectors. KNNR with LOO is the most straightforward approach due to its simplicity, although other classifiers or performance criteria could also be used, too.

(4) We applied the procedure described above to two parts of each syllable, that is, the beginning and ending boundaries.

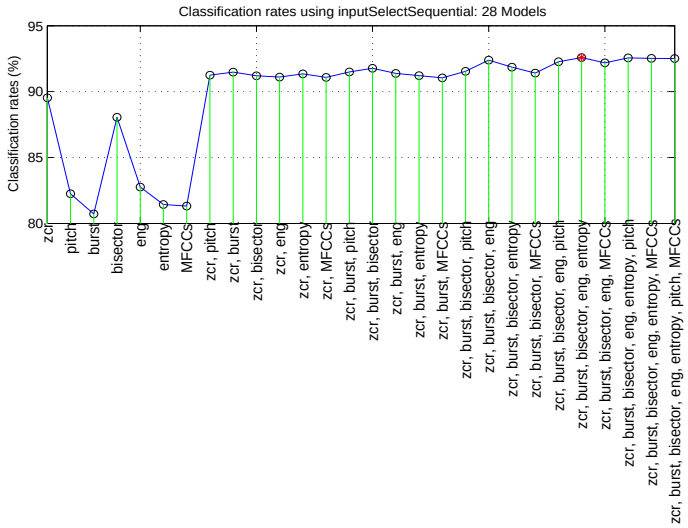

Figure 4. The LOO classification rates for different combinations of features for "silence + fricative and affricate" phonetic category at the beginning boundary.

Figure 4 shows the SFS results for the "silence + fricative and affricate" phonetic category at the beginning boundary, where the x-axis is the selected features and the y-axis is the LOO classification rates. From Figure 4, it is evident that the most distinguishing features for the "silence + fricative and affricate" category at the beginning boundary are the zero-crossing rate, bisector frequency, log energy, entropy, and burst degree, with which a LOO classification rate of 92.6% could be achieved. By following this same procedure, we could identify the most distinguishing features and their corresponding LOO classification rates, as shown in Table 3.1 and Table 3.2.

Table 3.1 Classification rates of the beginning boundaries of syllables for four phonetic categories.

Phonetic category transitions		Classification rate	Selected features
Left side	Right side		
Silence	Fricative and affricate	92.6%	Zero-crossing rate, bisector frequency, log energy, entropy, and burst degree
Silence	Aspirated stop	89.0%	Zero-crossing rate, log energy, bisector frequency, and burst degree
Silence	Unaspirated stop	92.1%	Entropy, log energy, burst degree and bisector frequency, and MFCCs
Silence	Periodic voiced	89.1%	Log energy, pitch, and burst degree
Periodic voiced	Fricative and affricate	92.7%	Bisector frequency, log energy, zero-crossing rate, entropy, and burst degree
Periodic voiced	Aspirated stop	87.6%	Zero-crossing rate and bisector frequency
Periodic voiced	Unaspirated stop	89.2%	Zero-crossing rate, log energy, entropy, and bisector frequency
Periodic voiced	Periodic voiced	71.8%	Bisector frequency, log energy, zero-crossing rate, entropy, MFCCs, and burst degree

Table 3.2 Classification rates of the ending boundaries of syllables for four phonetic categories.

Phonetic category transitions		Classification rate	Selected features
Left side	Right side		
Periodic voiced	Silence	87.4%	Log energy, burst degree, entropy, and bisector frequency
Periodic voiced	Fricative and affricate	89.6%	Zero-crossing rate, bisector frequency, pitch, log energy, burst degree, and entropy
Periodic voiced	Aspirated stop	89.9%	Zero-crossing rate, bisector frequency, pitch, log energy, burst degree, and entropy.
Periodic voiced	Unaspirated stop	86.4%.	Pitch and log energy
Periodic voiced	Periodic voiced	70.7%	Zero-crossing rate, bisector frequency, pitch, log energy, MFCCs, and entropy

The classification rates of "periodic voiced + periodic voiced" were only 71.8% at the beginning boundaries and 70.7% at the ending boundaries, respectively, which are comparatively low. This is mainly due to inseparable co-articulation. Later in this paper, we shall propose and detail other heuristic rules that can be applied to enhance the performance.

3.4 Further Improvement for "Periodic Voiced + Periodic Voiced" Cases

In our implementation, we first obtained an initial estimate of the beginning/ending boundaries based on the TCC-300 trained HMM with adaptation performed by means of a TTS-455 corpus. For every initial boundary, we selected candidate boundaries that were 2 ms apart and within 40 ms at both sides of this boundary. In other words, there were 41 candidate boundaries. The final boundary was determined by KNNR, where K was equal to 9, and the training data set is the one used for SFS and LOO mentioned above. The adopted features were those selected by the SFS as mentioned above.

However, for "periodic voiced + periodic voiced," the performance was not good enough due to co-articulation. Hence, we devised a special scheme for this category. Specifically, we adopted only two features to determine the boundary. This approach is based on the observation that most boundaries labeled by humans are located in a region with lower log energy.

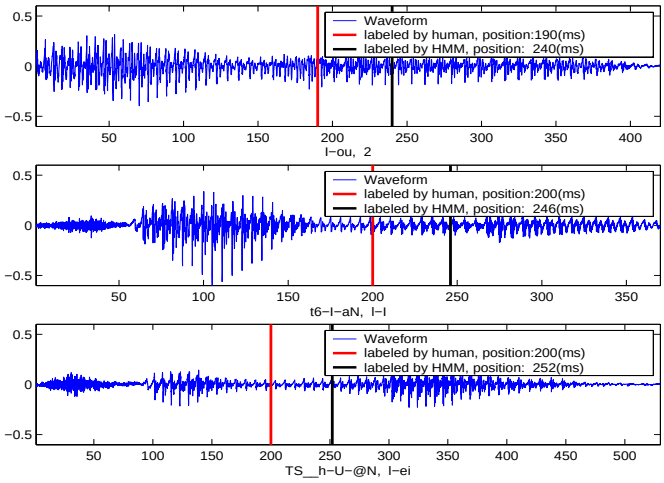

Figure 5. The three common errors in "periodic voiced + periodic voiced" cases. The content of the 1st waveform was "幼兒" ("I-ou, 2" in SAMPA), the content of the 2nd waveform was "將離" ("t6-I-aN, l-I" in SAMPA), and the content of the 3rd waveform was "蟲類" ("TS_h-U-@N, l-ei " in SAMPA).

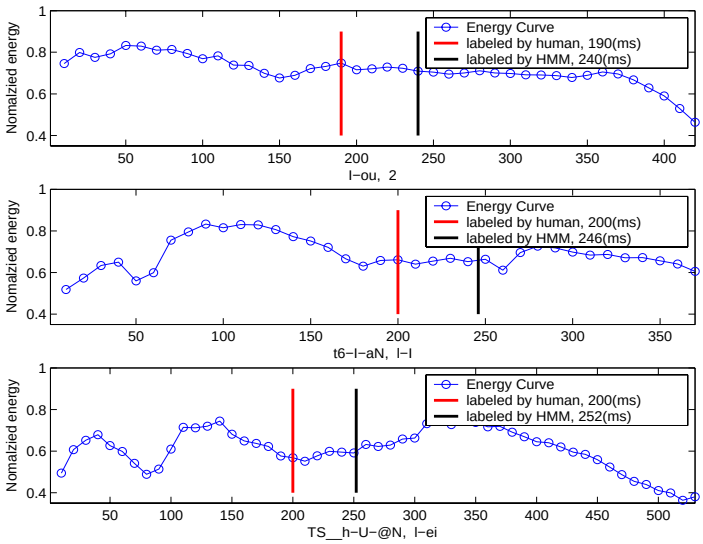

Figure 6. The corresponding log energy profiles for three "periodic voiced + periodic voiced" cases.

Figure 5 and Figure 6 show typical cases for 幼兒 ("I-ou" and "2" in SAMPA), 將離 ("t6-I-aN" and "l-I" in SAMPA), and 蟲類 ("TS_h-U-@N" and "l-ei" in SAMPA). Refining the boundary of this category is more complicated, and little related research has been reported in the literature. In this paper, we propose a new scheme to deal with this category using MFCCs and log energy, as described below:

(1) The search region is increased from ±40 ms to ±80 ms since large deviations over 50 ms are common in the "periodic voiced + periodic voiced" category. The number of candidate boundaries is increased from 41 to 81.

(2) We calculate the average log energy in the search region. We then set the new search region to be the one whose log energy is less than the log energy threshold, which is empirically defined as 0.9 times the average log energy.

(3) Among the boundaries within the new search region, we select the one with the maximum distance between the MFCCs of its left and right frames.

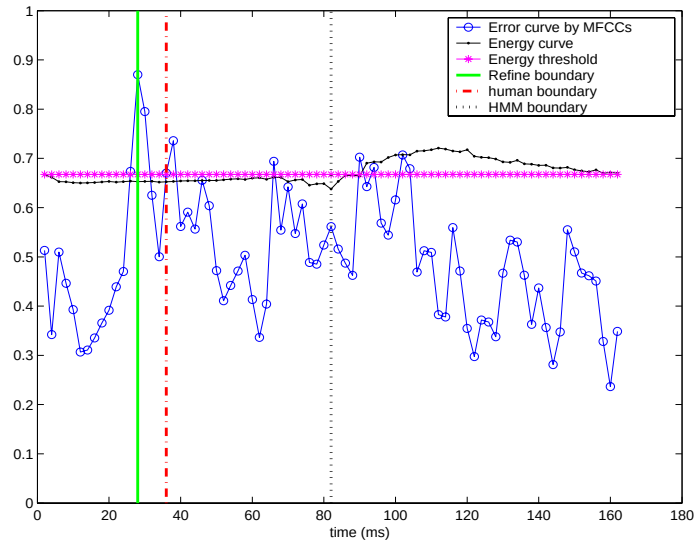

Figure 7. The refined results obtained based on MFCCs and log energy for the "periodic voiced + periodic voiced" case.

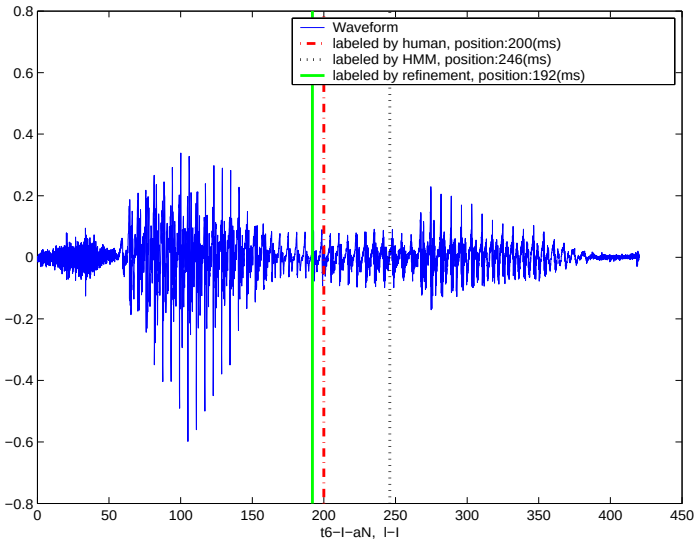

Figure 8. Typical results for the refined boundary of the "periodic voiced + periodic voiced" case. The content of this waveform was "將離" ("t6-I-aN, l-I" in SAMPA).

Figure 7 shows typical results obtained by using the above refinement method. The refined boundary for the original waveform is shown in Figure 8. The experimental results and error analysis will be discussed in the next section.

4. EXPERIMENT RESULTS AND ERROR ANALYSIS

4.1 The Different Acoustic Models of HMM-based Recognizers

To evaluate the performance of our proposed system, we used the TTS-455 corpus to verify the segmentation results. First, we employed different types of model training to construct an HMM recognizer for forced alignment, as described below:

(1) 1st model: Speaker-independent (SI) model constructed by using the TCC-300 corpus.

(2) 2nd model: Speaker-dependent (SD) model constructed by using the TTS-455 corpus, with uniform segmentation.

(3) 3rd model: Speaker-dependent (SD) model constructed by using the TTS-455 corpus, with initial segmentation performed by the model trained using the TCC-300 corpus.

(4) 4th model: Speaker-independent (SI) model constructed by using the TCC-300 corpus first and then adapted by using the TTS-455 corpus.

Each of these four types of acoustic models was constructed based on context-dependent tri-phones. The MLLR method [Huang *et al.* 2001] used to construct the 4th model employs the regression class tree to estimate a set of linear transformations for the mean vectors and covariance matrices of a Gaussian mixture HMM system. The tree was constructed using a centroid-splitting algorithm based on the Euclidean distance measure. We applied a binary regression tree with thirty-two base classes to our adapted data. In order to speed up the adaptation process and preserve storage capacity, we used the diagonal transform matrix instead of the full transform matrix [Odell *et al.* 1995]. Hence, the 4th model can be regarded as a speaker-adapted (SA) model.

The difference between the 2nd and 3rd models lies in the initial segmentation for training. The 2nd model uses uniform segmentation, while the 3rd model uses the segmentation derived by the recognizer trained using the TCC-300 corpus. Both of them can be viewed as SD models derived from the TTS-455 corpus.

4.2 The Performance of Different Acoustic Modes for Labeling the TTS-455 Corpus

Table 4 summarizes the results obtained with different modeling methods. The acoustic model for HMM forced alignment is based on context-dependent triphone modeling. From Table 4, it is evident that the 4th model achieved the best performance.

Table 4. Segmentation results w.r.t. model training (including beginning and ending boundaries).

Model / Errors	<=10ms	<=20ms	<=30ms	>50ms
1st model	49.47%	70.58%	84.24%	4.90%
2nd model	45.02%	69.83%	81.96%	7.64%
3rd model	43.49%	65.55%	79.60%	8.49%
4th model	46.09%	72.07%	87.40%	4.20%

4.3 A Comparison of the Segmentation Rate between Forced Alignment and Our Refinement Procedure

We have chosen the 4th model as our primary speech recognizer. However, its performance in segmentation is still not good enough for TTS application. The segmentation rate within 20 ms is only 72% when using the 4th model. It is probable that the system can be further improved. The following experiment was based on the initial boundaries identified by the 4th model. In our experiment, we divided the segmentation task according to groups of phonetic categories, as mentioned previously in Section 3.4. Table 5.1 shows the results for each phonetic category transition at the beginning boundary and Table 5.2 shows the results for each phonetic category transition at the ending boundary. Table 6 compares the overall segmentation rates obtained with the 4th model recognizer and our refinement procedure.

Table 5.1 Segmentation rates obtained with the HMM recognizer and the refinement procedure for all phonetic categories at the beginning boundaries of syllables.

Phonetic category Transitions		<=10 ms		<=20 m		<=30 ms		>50 ms	
Left side	Right side	H	R	H	R	H	R	H	R
Silence	Fricative and affricate	23.1	77.3	59.1	91.1	91.8	96.1	1.9	1.7
Silence	Aspirated Stop	13.7	81.9	54.5	94.3	93.3	98.7	0.3	0
Silence	Unaspirated stop	13.2	89.5	53.0	98.2	92.3	99.6	0.2	0
Silence	Periodic voiced	8.8	70.1	46.8	86.7	86.9	92.1	3.4	2.5
Periodic voiced	Fricative and affricate	59.4	84.7	83.6	94.7	95.3	97.8	0.7	0.7
Periodic voiced	Aspirated Stop	27.0	81.5	61.7	94.6	93.1	96.5	1.2	1.2
Periodic voiced	Unaspirated stop	30.0	85.2	67.0	95.8	91.4	98.4	1.3	0.5
Periodic voiced	Periodic voiced	45.0	66.3	60.0	75.3	71.9	79.2	10.9	6.7

Note: H: HMM results; R: Refined results; unit: %.

Table 5.2 Segmentation rates obtained with the HMM recognizer and the refinement procedure for all phonetic categories at the ending boundaries of syllables.

Phonetic category Transitions		<=10 ms		<=20 m		<=30 ms		>50 ms	
Left side	Right side	H	R	H	R	H	R	H	R
Periodic voiced	Silence	56.0	58.7	76.2	78.6	85.0	86.8	5.7	5.2
Periodic voiced	Fricative and affricate	58.3	75.0	88.2	92.8	97.2	97.3	0.5	0.3
Periodic voiced	Aspirated Stop	47.5	57.8	80.7	84.1	94.4	94.5	1.4	1.5
Periodic voiced	Unaspirated stop	57.0	73.1	91.5	91.6	98.1	97.8	0.1	0.3
Periodic voiced	Periodic voiced	42.9	63.5	60.8	72.8	70.9	79.6	11.5	8.4

Note: H: HMM results; R: Refined results; unit: %.

Table 6. The overall segmentation rates obtained with this system. (including beginning and ending boundaries).

	<=10ms	<=20ms	<=30ms	>50ms
HMM-based forced alignment	46.1%	72.1%	87.4%	4.2%
The proposed refinement method	69.1%	87.7%	94.2%	3.5%

4.4 Results and Discussions

From Table 5.1 and Table 5.2, we can observe that the performance for each phonetic category transition is satisfactory except for the category “periodic voiced + periodic voiced.” It may seem that our refinement method performed poorly for this category. We have carried out another experiment in which we applied the statistical method (just like the one applied to other phonetic categories) to this “periodic voiced + periodic voiced” category. The average segmentation rate of <=30ms for this “periodic voiced + periodic voiced” category was 60% lower. This clearly indicates that our refinement method (rule-based in this case) is definitely better. All in all, this category still poses a difficulty for automatic segmentation since there is usually very strong co-articulation between two neighboring syllables, such “第一” (meaning “number one”), “蘇武” (an ancient Chinese person’s name), and so on.

From Table 6, it is evident that our refinement approach leads to improvement in the overall segmentation rate. The segmentation rate within 20 ms is significantly increased by about 15.6%; and the segmentation rate within 30 ms after the refinement procedure is performed is 94.2%, which is acceptable for general TTS systems. Admittedly, however, there is still some room left for future improvement, as described in the following:

(1) The size of our TTS-455 corpus is not large enough. A larger corpus will result in a better adapted model, which will reduce the segmentation errors that are larger than 50 ms.

(2) Acoustic features other than MFCCs can potentially be used to obtain better segmentation rates. We are now in the process of identifying other more discriminative acoustic features for this purpose.

5. CONCLUSIONS

Correct phonetic labeling is very important for concatenation-based speech synthesis. Consequently, the application of automatic phonetic labeling and segmentation for corpora to be used in TTS has become a critical issue. In this paper, we have proposed a specific refinement procedure suitable for Mandarin Chinese. We divide all Chinese phonemes into four categories and employ the SFS algorithm to select the best features for each phonetic category. However, the proposed method does not work well in the "periodic voiced + periodic voiced" case. Hence, we have proposed an additional scheme to deal specifically with this case, using log energy and MFCCs. Several experiments have demonstrated the feasibility of the proposed approach.

In future work, we will focus on finding new features to improve the segmentation rate in the "periodic voiced + periodic voiced" case. We will also apply other classifiers, such as SVM (support vector machine), to further improve the classification results. Finally, we will apply other methods for feature extraction, such as linear discriminant analysis and principal component analysis.

Reference

Bonafonte, A., A. Nogueiras and A. Rodriguez-Garrido, "Explicit segmentation of speech using Gaussian models," *Proceedings of International Conference on Spoken Language Processing*, 1996, pp. 1269-1272.

Chen, K. J. and S. H. Liu, "Word identification for mandarin Chinese sentences," *Proceedings of the Fifteenth International Conference on Computational Linguistics*, 1992, pp. 101-107.

Chou, F.-C., C.-Y. Tseng and L.-S. Lee, "Automatic Segmental and Prosodic Labeling of Mandarin Speech," *Proceedings of International Conference on Spoken Language Processing*, 1998, pp. 1263-1266.

Chou, F.-C., C.-Y. Tseng and L.-S. Lee, "A Set of Corpus-based Text-to-speech Synthesis Technologies for Mandarin Chinese," *IEEE Transactions on Speech and Audio Processing*, 10(7), 2002, pp.481-494.

Cosi, P., D. Falavigna and M. Omologo, "A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies," *Proceedings of European Conference on Speech Communication and Technology*, 1991, pp. 693-696.

Demuynck, K. and T. Laureys, "A Comparison of Different Approaches to Automatic Speech Segmentation," *Proceedings of International Conference on Text, Speech and Dialogue*, 2002, pp. 277--284.

Duda, R. D., P. E. Hart and D. G. Stork, *Pattern Classification, 2nd ed.*, Wiley, New York, 2001.

Huang, X., A. Acero and H. W. Hon, *Spoken language processing*, Prentice Hall, New Jersey, 2001.

Lamel, L. F. and J. L. Gauvain, "High Performance Speaker-Independent Phone Recognition Using CDHMM," *Proceedings of European Conference on Speech Communication and Technology*, 1993, pp. 121-124.

Lee, L.-S., "Voice Dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, 10(4), 1997, pp.63-101.

Ljolje, A. and M. D. Riley, "Automatic segmentation of speech for TTS," *Proceedings of European Conference on Speech Communication and Technology*, 1993, pp. 1445-1448.

Ljolje, A., J. Hirschberg and J. P. H. van Santen, "Automatic Speech Segmentation for Concatenative Inventory Selection," *Proceedings of ESCA/IEEE Workshop on speech synthesis*, 1994, pp. 93-96.

Lu, H.-M., "An implementation and Analysis of Mandarin Speech Synthesis Technologies," MD thesis, National Chiao Tung University at Taiwan, 2002.

Makashay, M. J., C. W. Wightman, A. K. Syrdal and A. Conkie, "Perceptual evaluation of automatic segmentation in text-to-speech synthesis," *Proceedings of International Conference on Spoken Language Processing*, 2000, pp. 431-434.

Odell, J., D. Ollason, P. Woodland, S. Young and J. Jansen, *The HTK Book for HTK V2.0*, Cambridge University Press, Cambridge UK, 1995.

Sethy, A. and S. Narayanan, "Refined Speech Segmentation for Concatenative Speech Synthesis," *Proceedings of International Conference on Spoken Language Processing*, 2002, pp. 149-152.

Shen, J.-L., J.-W. Hung and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," *Proceedings of International Conference on Spoken Language Processing*, 1998.

Sproat, R. and C. Shih, "A statistical method for finding word boundaries in Chinese text," *Computer Processing of Chinese and Oriental Languages*, 1990, pp.336-351.

Torre Toledano, D., M. A. Rodrguez Crespo and J. G. EscaladaSardina, "Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-correction Rules," *Proceedings of Third ESCA/COCOSDA Workshop on speech synthesis*, 1998, pp. 207-212.

Van Erp, A. and L. Boves, "Manual segmentation and labelling of speech," *Proceedings of Speech*, 1988, pp. 1131-1138.

van Santen, J. P. H. and R. Sproat, “High-accuracy automatic segmentation,” *Proceedings of European Conference on Speech Communication and Technology*, 1990, pp. 2809–2812.

Wang, H. C., R. L. Chiou, S. K. Chuang and Y. F. Huang, “A phonetic labeling method for MAT database processing,” *Journal of the Chinese Institute of Engineers*, 22(5), 1999, pp. 529-534.

Whitney, A., "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, 20(9), 1971, pp.1100-1103.

Yeh, C. L. and H. J. Lee, “Rule-based word identification for Mandarin Chinese sentences - A unification approach,” *Computer Processing of Chinese and Oriental Languages*, 1991, pp. 97-118.