Precise phone/syllable boundary labeling of utterances in a speech corpus plays an important role in constructing corpus-based TTS (text-to-speech) systems. 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 the design of a new procedure to refine the boundaries in a Mandarin speech corpus. This procedure employs different sets of acoustic features for four different phonetic categories. In addition, a new scheme is designed to deal with the case of “periodic voiced + periodic voiced” which produces most of the segmentation errors in our experiment. Several experiments are designed to demonstrate the feasibility of the proposed approach.
Corpus-based speech synthesis systems are becoming more and more popular due to the high degree of fluency and the natural feel of the generated speech. However, such systems always require a significant amount of human effort to label the phonetic boundaries of the corresponding corpus [3][16][25]. Therefore, a great deal of research regarding automatic phonetic labeling has been proposed over the past several years [1][2][20]. In general, most of these methods involve the following two steps:

1. Rough phonetic segmentation by Viterbi forced alignment using HMM (hidden Markov models) or other statistical methods.

2. High time-resolution analysis of the phonetic boundaries by boundary checking rules.

These HMM-based recognizers can be categorized in various dimensions. For example, some use context-dependent HMM and others use context-independent HMM [23]. 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 success in speech recognition, its usage for automatic phonetic segmentation and labeling does not always produce precise and satisfactory results for the development of TTS. As a result, other acoustic features and refinement algorithms are proposed in the literature to improve the phonetic labeling results from HMM-based recognizers.
With regard to automatic phonetic labeling, several articles have been devoted to this study in the last few years. For example, in [5], Bonafonte et al. took Gaussian probability density distribution as a similarity measure. In [18], Jan P. H. van Santen et al. adopted broad-band and narrow-band edge detection. In [13], Toledano et al. tried to mimic human labeling using a set of fuzzy rules. In [27], Sethy et al. employed adapted CDHMM (continuous density hidden Markov model) models [21]. The main focus of all of these studies is on English speech, and they seldom address the issue of which phonetic class tends to be more error prone. Moreover, the proposed methods in the above papers may not perform equally well when dealing with another language. For example, most approaches for English utterance segmentation can be classified into two categories: rule-based [13] and statistics-based [27]. For rule-based approaches, we need to define a set of rules (crisp or fuzzy) for various phonetic transitions. For statistics-based approaches, we need to collect a sample data set and label the set accordingly. Conceptually, the rule-based approaches for English corpora can be adapted for Chinese. But in fact, it is hard to design such a system unless we have human experts who have a thorough understanding of the similarities and differences between the phonetic sets of these two languages. In the present study, it is our belief that these two approaches should be used in a seamless integrated manner. As a result, we have chosen a hybrid approach, where most boundaries are identified via statistical pattern recognition (Sequential Forward Selection, K-Nearest Neighbor Rule and Leave-One-Out).
while the most difficult cases (periodic voiced + periodic voiced) are handled by 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 [14][22]) structure or a single V (Vowel) structure. Therefore, the primary work of speech labeling will focus on precisely identifying the boundaries of each syllable. Then the boundary between the consonant and the vowel within a syllable can be identified according to the type of the consonant. In most cases the consonant is fricative, affricate or plosive, and the consonant can easily be separated using several acoustic features other than MFCCs, such as zero-crossing rate or pitch, and others. If the consonant is periodic, like ‘ㄢ’ (‘l’ in SAMPA [10], SAMPA is the acronym of ‘Speech Assessment Methods Phonetic Alphabet’), or ‘ㄇ’ (‘m’ in SAMPA), then the consonant does not need to be segmented, and the whole syllable should be treated as a single unit for TTS, since the further operations of pitch-scale or time-scale modification should be performed on both the consonant and the vowel.

In [14][15], 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 its iterative procedure for forced alignment, correction rules and re-training. And, it becomes particularly inefficient if the speech corpus is updated incrementally and regularly,
such as the increment of 1-hour 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 of moderate size, for example, 1 hour for the same speaker.

In this paper we designed an SA-based HMM recognizer to perform a forced alignment as the first step and then employed a refinement procedure to modify the identified boundaries. The proposed refinement procedure adopts several innovative acoustic features to refine boundaries for different phonetic categories. These approaches and their related experimental results will be described in the following sections.

This paper is organized as follows. Section 2 introduces a forced alignment procedure using an HMM recognizer to get initial estimation of all boundaries. Section 3 explains the refinement procedure specially designed for four phonetic categories, and describes how to choose the acoustic features by SFS (Sequential Forward Selection) [4] algorithm. Section 4 describes the experiments for demonstrating the performance of the proposed refinement procedure, and makes an error analysis for those irretrievable errors. Section 5 draws the conclusions and discusses future work related to this paper.

2. HMM BASED RECOGNIZER

2.1 From Orthographic Transcription to Phonetic Transcription

Forced alignment using the HMM-based recognizer relies on the knowledge of the underlying
phonetic transcription of a given utterance. In general, once the orthographic transcription and speech data are both available, then we can employ forced alignment for automatic phonetic transcription. However, some commonly used Chinese characters have multiple syllables for different pronunciations, depending on the lexical contexts. For instance, the Chinese character “重” (meaning “heavy”) is pronounced as “ㄓㄨㄥˋ” (“TS-U-@N, 4th tone” in SAMPA) in “重要” (meaning “important”) and as “ㄔㄨㄥˊ” (“TS_h-U-@N, 2nd tone” in SAMPA) in “重疊” (meaning “overlap.”) As a result, word segmentation in the text sentence becomes very important for providing the correct phonetic transcription for alignment.

Commonly used approaches to word segmentation in Chinese NLP (natural language processing) include the forward or backward maximum word matching algorithm [12][30] or the dynamic-programming-based statistic probability method [28]. However, no word segmentation algorithm can guarantee perfect results due to the following facts:

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

(2) Even if the word dictionary is complete, some pronunciation cannot be determined through dictionary lookup, especially in Chinese poems. For instance, the first character of “朝辞白帝彩雲間” (meaning “leaving Baidi city in colored dawn”) is pronounced as “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 for “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”) using forward maximum word matching. On the other hand, it will be labeled as “老 掌櫃 順 手 把 錢 揄 在 懷裡” (meaning “The old + shopkeeper + smoothly + handle bar + the money + into + his pocket”) using the backward one.

In order to prevent any errors made as a result of phonetic transcription, we adopted the following two steps, combining word segmentation and forced alignment, to achieve a better performance:

(1) Perform word segmentation using forward and backward maximum matching based on a word dictionary of around 90,000 entries. Keep the phonetic transcriptions as candidates to be used in the next step. (If the result is the same, then we have only a single phonetic transcription.)

(2) 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 by the above word segmentation processes. Use these different phonetic transcription candidates to run
a forced alignment of the Viterbi decoding. 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 in order to achieve a better performance in phonetic transcription. By using the TTS-455 speech corpus [6] with about 6000 Chinese syllables, the syllable error rate is 2.1% and 1.9% for forward and backward maximum matching, respectively. With the addition of step 2 for alignment-based enhancement, this error rate is reduced to 1.0%, which is a significant reduction of 50% in error rate. Some of the error cases are shown in Table 1.

| Text sentences of speech corpus | Human transcription | Machine transcription |
|--------------------------------|---------------------|----------------------|
| 春風秋月何時了 |ㄌ一ㄠˇ (“l-I-au, 3rd tone”) |ㄉㄜ. (“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 word with multiple pronunciations that cannot be identified through word dictionary lookup. Unfortunately, the forced alignment cannot find the correct phonetic transcription either, because the utterance itself is ambiguous and unclear. The second sentence indicates the insufficiency 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 as “ㄎㄨㄛˇ” (“k_h-U-o, 4th tone” in SAMPA) colloquially. The error from the third sentence indicates the insufficiency 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 the phonetic transcription is obtained, we can perform a forced alignment by using a HMM recognizer. In this paper, we have two Mandarin Chinese speech corpora:

(1) TTS-455 speech corpus [6]: This corpus contains 455 sentences by one speaker and covers about 6000 syllables. It is mainly for TTS. Its detailed information is organized as following:

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

II. Sampling rate and bit rate: 20000 Hz, 16bits.

III. Base syllables: 408.

IV. Tonal syllables: 1196.

More information on this corpus can be found in [8].

(2) TCC-300 speech corpus [7]: It was recorded by 300 subjects from National Taiwan University, Chiao Tung University and Cheng Kung University in Taiwan. In addition, the recording texts were selected from “Academia Sinica Balanced Corpus” [11].
In order to perform a forced alignment on the TTS-455 speech corpus, we need to train an HMM-based recognizer. The details of this recognizer will be described in Section 4.

3. DESIGN OF THE REFINEMENT PROCEDURE

Consequently, a post-processing scheme must be developed to refine the identified syllable boundaries. Specifically, since the forced alignment is based on MFCCs only; it makes sense to use other acoustic features to enhance the precision. As mentioned in Section 1, the use of either a rule-based or a statistics-based approach alone leaves much room for improvement. Therefore, we combined these two methods to deal with the Mandarin Chinese speech corpus.

First of all, we divide all Chinese phonemes into four categories. Afterwards, we determine which set is suitable for which methods (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 to refine the 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 [17], as listed below in SAMPA format [10] 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
  
  - (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

The reason for grouping fricative and affricate categories into a single one is mainly due to the similarity of 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, and in addition there usually exists a high zero-crossing rate at the aperiodic part. For the periodic voiced category, we included both consonants and vowels since they all contain a stable harmonic or pitch structure.
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, including zero-crossing rate, log energy, pitch, and entropy [19]. In addition, we also introduced two new acoustic features, bisector frequency and burst degree, to help identify the boundaries more precisely.

3.2.1 Bisector Frequency

The definition of bisector frequency is shown in equations (1) and (2):

\[
freqIndex = \arg \min_{1 \leq k < N} \left| \sum_{f=1}^{k} A_f - \frac{\sum_{f=k+1}^{N} A_f}{2} \right|
\]

\[
\text{bisectorFreq} = \frac{freqIndex}{N} \times \text{sampleRate}.
\]

where \( A_f \) is the amplitude of the \( f \)th frequency component and there are \( N \) distinct frequency components in the spectrum. We can use this feature to distinguish the unvoiced from the voiced patterns. Although the zero-crossing rate can also be used for detecting unvoiced patterns in most cases, it is not robust enough, 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 the bisector frequency than by the zero-crossing rate.
In our implementation, we normalized the value of this feature to the range [0,1] according to equation (3):

$$\text{bisorctorfreq} = \left( \frac{\text{bisorctorfreq} - \text{lowfreq}}{\text{highfreq} - \text{lowfreq}} \right),$$

(3)

where the value of \(\text{highfreq}\) and \(\text{lowfreq}\) are empirically set to be \(\frac{\text{sampleRate}}{2} \times 0.8\) and 100, respectively.

### 3.2.2 Burst Degree

It is difficult to recognize a burst pattern in speech using zero-crossing rate and/or pitch. This is because there does not exist a stable pitch structure, and the zero-crossing rate is relatively speaking not that high. To deal with this situation, we proposed a new feature called 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):
\[
\text{burst degree} = \frac{1}{W_1 \times \text{avg}(\text{local max Interval}) + W_2 \times \log \text{Energy}}.
\]

(4)

where \( W_1 \) and \( W_2 \) are two weighting factors of values 4 and 1, respectively. The \( \text{avg}(\text{local max Interval}) \) is the average distance between the positions of neighboring local maxima of sample points. For instance, suppose that there are 4 local maxima of sample points located at positions 12, 52, 92, 130 in certain frame. Then, the intervals are 40, 40, 38 and the \( \text{avg}(\text{local max Interval}) \) is \((40+40+38)/3\). Figure 2 shows a typical result of 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 have divided all phonemes into four phonetic categories. In this section, we divided the boundaries into groups according to the transition 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 this paper, our experiments do not try to find those kinds of boundaries since it is not the focus of this paper. However, we will still discuss all three kinds of boundaries for the sake of completeness.)

Based on a similar analysis, we can construct Table 2 to list all possible transitions from the left side to the right side for the beginning, ending, and INITIAL/FINAL boundaries.

Table 2. All Possible category transitions of the 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 contains about 2100 syllables of 20 long sentences, with a total of 10 minutes [9]. This corpus is used for feature selection and it is fully independent of our speech corpus mentioned in Section 2.2. These syllables cover every Mandarin Chinese phoneme. The beginning and ending of the phonetic boundaries for these 2100 syllables are manually labeled. The following we describe the steps for finding the best combination of features for each of these phonetic category transitions.

(1) In order to find the most discriminative features, we must create a set of training data. This is achieved by adding several candidate boundaries, 10 ms apart, located within ±80 ms of a true (manually labeled) boundary. A candidate boundary is labeled “correct” if it is within ±20 ms of the true boundary. (According to Chou [14], manual labeling of two human experts can achieve about 90% consistency on 10ms tolerance, and 100% on 20ms.) Therefore we chose to have 5 correct candidates, all within 20 ms of the manually labeled one, for our experiments. If we had chosen only one, 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 can create a set of 17 candidate boundaries (including the true one), which have 5 labeled “correct” and 12 labeled “wrong” as their
desired classification output as shown in Figure 3.

(2) For each candidate boundary, we evaluate the differences between all acoustic features of its left and right frames. The size of each frame is 20 ms, and the “difference of acoustic features” is then used as the features for designing a classifier.

(3) In order to find the most influential acoustic features, we employed the method of sequential forward selection (SFS) proposed by Whitney [4] in the pattern recognition literature. The principle of SFS is to start with a single feature with the best classification rate. Then we keep the already selected features and try to identify a newly added feature that can most increase the classification rate. For instance, if feature 2 and 5 are the currently selected features. Then we shall try to find another features, that, when combined with the selected features, can achieve the best classification rate. This greedy step is repeated until the desired number of features has been selected, or until there is no improvement of the classification rate. In order to use SFS we need to select a classifier together with its performance evaluation scheme. Here we use KNNR (K-Nearest Neighbor Rule) as the classifier and LOO (Leave-One-Out) [26] as the performance criterion. The basic concept of 1-NNR is to assign the class of a given test vector as the one of the data point in the training data that is nearest to the given vector. In order to have a better degree of robustness, we can choose to have KNNR, in which the K nearest neighbors are selected around the test vector and the assigned class is determined by a
voting mechanism among these K points. For evaluating the performance of KNNR, we applied LOO, in which 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 selection due to its simplicity in concept and computation, although other classifiers or performance criteria could be used too “(We also performed a simple search to find the best value of K in KNNR is 9 in our experiment.).”

(4) We applied the procedure proposed in the previous steps to two parts of each syllable, that is, beginning and ending boundaries.

Figure 3. Training data of 5 correct boundaries and 12 wrong boundaries around the true boundary labeled by human. The content of this waveform is “將離” (“t6-I-aN, l-I” in SAMPA).
Figure 4 demonstrates the result of SFS 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 zero-crossing rate, bisector frequency, log energy, entropy and burst degree, able to achieve a LOO classification rate of 92.6%. By following this same procedure, we can identify the most distinguishing features and their corresponding LOO classification rates, as shown in Table 3.1 and Table 3.2.

| 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 | Left side | Right side | Classification rate | Selected features |
|-------------------------------|-----------|------------|---------------------|-------------------|
| 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” are only 71.8% at the beginning boundaries and 70.7% at the ending boundaries, respectively, which is comparatively low. This is mainly due to inseparable co-articulation. We shall propose and detail other heuristic rules to enhance the performance later in this paper.
3.4 Further Improvement of “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 by a TTS-455 corpus. For every initial boundary, we selected candidate boundaries, 2 ms apart, and within 40 ms at both sides of this boundary. In other words, there are 41 candidate boundaries. The final boundary is determined by KNNR, where K is equal to 9, and the training data set is the one used for SFS and LOO mentioned previously. The adopted features are those selected by the SFS as mentioned earlier.

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

Figure 5. The three common errors in “periodic voiced + periodic voiced” cases. The content of 1st waveform is “幼兒” (“I-ou, 2” in SAMPA), the content of 2nd waveform is “將離” (“t6-I-an, l-I” in
SAMPA) and the content of 3rd waveform is “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 the typical cases for 幼兒 (“I-ou” and “2” in SAMPA), 將離 (“t6-I-aN” and “l-I” in SAMPA), 蟲類 (“TS_h-U-@N” and “l-ei” in SAMPA). Refining the boundary of this category is more complicated, and there is little related work reported in the literature. In this paper, we propose a new scheme to deal with this category using MFCCs and log energy, as follows:

(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) Calculate the average log energy in the search region. Set the new search region to be the one whose log energy is less than the log energy threshold, which is defined as 0.9 times
the average log energy empirically.

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

Figure 7. The refined result based on MFCCs and log energy for the “periodic voiced + periodic voiced” case.
Figure 8. Typical result of refined boundary of the “periodic voiced + periodic voiced” case. The content of this waveform is “將離” (“t6-I-aN, l-I” in SAMPA).

Figure 7 demonstrates a typical result 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 are illustrated 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 for verifying the segmentation result. First, we employed different types of model training for constructing an HMM recognizer for forced alignment, as follows:

(1) 1<sup>st</sup> model: Speaker-independent (SI) model by the TCC-300 corpus.

(2) 2<sup>nd</sup> model: Speaker-dependent (SD) model by the TTS-455 corpus, with uniform segmentation.

(3) 3<sup>rd</sup> model: Speaker-dependent (SD) model by the TTS-455 corpus, with initial segmentation by the model trained by the TCC-300 corpus.

(4) 4<sup>th</sup> model: Speaker-independent (SI) model by the TCC-300 corpus first, and then adapted by the TTS-455 corpus using the MLLR method [29].
Each acoustic model of these four types was constructed based on context-dependent tri-phones. The MLLR method used in the 4th model employed the regression class tree to estimate a set of linear transformation 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 employed a binary regression tree with thirty-two base classes to our adapted data. In order to speed up the adaptation process and save storage capacity, we used the diagonal transform matrix instead of the full transform matrix. 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 used uniform segmentation while the 3rd model used the segmentation derived by the recognizer trained from 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 of different modeling methods for forced alignment. The acoustic model for the HMM forced alignment is based on context-dependent triphone modeling. From Table 4, it is evident that 4th model achieves the best performance, which confirms our earlier conjecture.

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

| Model | Errors | <=10ms | <=20ms | <=30ms | >50ms |
|-------|--------|--------|--------|--------|-------|

(Complete the table with actual data)
| Model | <=10 ms | <=20 ms | <=30 ms | >50 ms |
|-------|---------|---------|---------|--------|
| 1<sup>st</sup> model | 49.47% | 70.58% | 84.24% | 4.90% |
| 2<sup>nd</sup> model | 45.02% | 69.83% | 81.96% | 7.64% |
| 3<sup>rd</sup> model | 43.49% | 65.55% | 79.60% | 8.49% |
| 4<sup>th</sup> model | 46.09% | 72.07% | 87.40% | 4.20% |

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

We have chosen the 4<sup>th</sup> 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 4<sup>th</sup> model. It is most likely that the system can be further improved. The following experiment is based on the initial boundaries identified by 4<sup>th</sup> model.

In our experiment, we divided the segmentation task into groups of phonetic category, as mentioned previously in Section 3.4. Table 5.1 exhibits the results for each phonetic category transition at the beginning boundary and Table 5.2 demonstrates the results for each phonetic category transition at the ending boundary. Table 6 shows the overall segmentation rate for comparison between the HMM-based recognizer (4<sup>th</sup> model) and our refinement procedure.

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

| Phonetic category Transitions | <=10 ms | <=20 ms | <=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 |
Table 5.2 Segmentation rates of the HMM recognizer and the refinement procedure for all phonetic category at the ending boundaries of syllables.

| Phonetic category Transitions | <=10 ms H | <=10 ms R | <=20 ms H | <=20 ms R | <=30 ms H | <=30 ms R | >50 ms H | >50 ms R |
|-------------------------------|----------|----------|-----------|----------|-----------|-----------|---------|---------|
| Left side                    | Right side | Silence | 56.0 | 58.7 | 76.2 | 78.6 | 85.0 | 86.8 | 5.7 | 5.2 |
| Periodic voiced              | Silence | 58.3 | 75.0 | 88.2 | 92.8 | 97.2 | 97.3 | 0.5 | 0.3 |
| Periodic voiced + Periodic voiced | Fricative and affricate | 47.5 | 57.8 | 80.7 | 84.1 | 94.4 | 94.5 | 1.4 | 1.5 |
| Periodic voiced + Periodic voiced | Aspirated Stop | 57.0 | 73.1 | 91.5 | 91.6 | 98.1 | 97.8 | 0.1 | 0.3 |
| Periodic voiced + Periodic voiced | Unaspirated Stop | 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 rate of 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 of each phonetic category transition is satisfactory except for the category of “periodic voiced + periodic voiced”. It may
seem that our refinement method performed poorly for this category. We have carried out another experiment to use the statistical method (just like the one used in other phonetic categories) on this “periodic voiced + periodic voiced” category; the average segmentation rate of \( \leq 30\text{ms} \) in this “periodic voiced + periodic voiced” category is lower than 60%. This clearly indicates that our refinement method (rule-based in this case) is definitely better. All in all, this category is still a difficult task for automatic segmentation since there is usually a very strong co-articulation between the 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 an improvement of 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 refinement procedure is 94.2%, which is desirable for general TTS systems. However, admittedly, there is still some room left for future improvement, as follows:

1. The size of our TTS-455 corpus is not large enough. A larger corpus will lead to a better adapted acoustic model, which will minimize the segmentation errors larger than 50 ms.

2. Acoustic features other than MFCCs are potential candidates for leading to a better segmentation rate. 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 automatic phonetic labeling and segmentation for corpora to be used in TTS become a critical issue. In this paper, we proposed a specific refinement procedure suitable for Mandarin Chinese. We divided all Chinese phonemes into four categories and employed the SFS algorithm to select the best features for each phonetic category. However, the proposed method does not work well in the case of “periodic voiced + periodic voiced”. Hence, we 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.

For future work, we will focus on finding new features for improving the segmentation rate for 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 try other methods for feature extraction, such as linear discriminant analysis and principal component analysis.

6. Reference

[1] A. Ljolje and M. D. Riley. "Automatic segmentation of speech for TTS", Proceedings of European Conference on Speech Communication and Technology, 1993, pp. 1445-1448.
[2] A. Ljolje, J. Hirschberg and J.P.H van Santen, “Automatic Speech Segmentation for Concatenative Inventory Selection”, Progress in Speech Synthesis, Springer, 1997.
[3] A. Van Erp and L. Boves, “Manual segmentation and labelling of speech,” Proceedings of Speech, 1988, pp. 1131-1138
[4] A. Whitney, "A direct method of nonparametric measurement selection", IEEE Transactions on Computers, vol. 20, 1971, pp.1100-1103.
[5] Antonio Bonafonte, Albino Nogueiras and Antonio Rodriguez Garrido, “Explicit segmentation of speech using Gaussian models”, Proceedings of International Conference on Spoken Language Processing, 1996, pp. 1269-1272.
[6] Available: http://214lab.ee.ntut.edu.tw/ Dr.Hwang supports this speech corpus.
[7] Available: http://rocling.iis.sinica.edu.tw/ROCLING/MAT/Tcc_300brief.htm
Available: http://speech.cs.nthu.edu.tw/gavins/Research/SpeechSynthesis/content_hsf455.txt
Available: http://speech.cs.nthu.edu.tw/gavins/Research/SpeechSynthesis/training2100.txt
Available: http://www.phon.ucl.ac.uk/home/sampa/home.htm
Available: http://www.sinica.edu.tw/~tibe/2-words/modern-words
Chen, K. J., and Liu, S. H. “Word identification for mandarin Chinese sentences”, 
Proceedings of the Fifteenth International Conference on Computational Linguistics, 1992, pp. 101-107.

D. Torre Toledano, M. A. Rodriguez Crespo, J. G. EscaladaSardina, “Trying to Mimic Human 
Segmentation of Speech Using HMM and Fuzzy Logic Post-correction Rules”, Proc. Third 
ESCA/COCOSDA Workshop on SPEECH SYNTHESIS, 1998.

Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee, “A Set of Corpus-based Text-to-speech 
Synthesis Technologies for Mandarin Chinese”, IEEE Transactions on Speech and Audio 
Processing, Vol.10, No.7, 2002, pp.481-494,

Fu-chiang Chou, Chiu-yu Tseng and Lin-shan Lee, “Automatic Segmental and Prosodic 
Labeling of Mandarin Speech”, Proceedings of International Conference on Spoken Language Processing, 1998, pp. 1263-1266,

H. C. Wang, 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, vol. 22, no. 5, 1999, pp. 529-534.

Hung-mao Lu, “An implementation and Analysis of Mandarin Speech Synthesis Technologies”, Department of Communication Engineering, National Chiao Tung University, master thesis, 2002.

Jan P. H. van Santen, J., Sproat, R. “High-accuracy automatic segmentation”, Proceedings of European Conference on Speech Communication and Technology, 1999, pp. 2809–2812

Jialin Shen, Jeihwei Hung, Linshan Lee, “Robust entropy-based endpoint detection for 
speech recognition in noisy environments”, Proceedings of International Conference on 
Spoken Language Processing, 1998.

Kris Demuynck and Tom Laureys. “A Comparison of Different Approaches to Automatic 
Speech Segmentation”, Proceedings of International Conference on Text, Speech and 
Dialogue, 2002, pp. 277–284.

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

Lin-shan Lee, ”Voice Dictation of Mandarin Chinese”, IEEE Signal Processing Magazine, 
Vol.14, No.4, 1997, pp.63-101.

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

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

P. Cosi, 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.

Richard O. Duda, Peter E. Hart, David G. Stork, “Pattern classification, 2nd edition”, New 
York, Wiley, 2001.

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

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

Xuedong Huang, Alex Acero, H. W. Hon, “Spoken language processing”, New Jersey, 
Prentice Hall, 2001.

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