A Hybrid Approach to Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpus

Cheng-Yuan Lin, Kuan-Ting Chen and J.-S. Roger Jang

Department of Computer Science
National Tsing Hua University, Taiwan
{gavins, marco, jang}@wayne.cs.nthu.edu.tw

Abstract

In this paper, we propose a hybrid approach to refine the phonetic boundaries in a Mandarin speech corpus. This approach employs different sets of acoustic features for different categories of phonetic transitions, except for the most difficult case of "periodic voiced + periodic voiced", which is therefore handled by a heuristic scheme. Several experiments are designed to demonstrate the feasibility of the proposed approach.

1. Introduction

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. Therefore, it is desirable to design an efficient method for automatic phonetic labeling, especially when the size of the speech corpus is very large. A great deal of research regarding automatic phonetic labeling has been proposed in the literature [6][7]. Most of these methods involve the following two steps:

1. Rough phonetic segmentation by forced alignment of Viterbi search using HMM (hidden Markov model) or other statistical methods.
2. High-resolution analysis and refinement of the phonetic boundaries by various boundary-checking rules.

Although forced alignment could be applied to identify the most probable boundaries, the segmental results based on such boundaries are not always accurate enough for TTS (text-to-speech) applications. As a result, numerous studies have proposed alternative approaches other than forced alignment. Drawbacks of some of the representative approaches are summarized as follows:

1. Wang et al. [8] proposed the use of MFCCs alone to refine the boundaries of all categories of phonetic transitions. This is too assertive. For example, if a boundary is of the case "silence + fricative", other simple features, such as energy, may well outperform MFCCs.
2. Toledano et al. [2] adopted multiple features based on manually tuned subject rules for each category of phonetic transition. This approach is labor intensive and not easily scaled up because different phonetic sets have different rules to be identified.
3. In particular, most methods (for instance, [6][7][10]) do not elaborate the issue of error analysis, as what categories of phonetic transitions tend to be more error-prone and how to deal with these transitions.

Therefore, in this paper, we have seven acoustic features as candidates and then employ several methods in statistical pattern recognition [9], including sequential forward selection, k-nearest neighbor rule and leave-one-out error estimation to identify the most suitable features for each transition of phonetic categories. This "divide and conquer" approach works well in general, except for certain transitions with strong coarticulation. Hence, we proposed a hybrid approach where most boundaries are identified via statistical pattern recognition while the most difficult cases (with strong coarticulation) are handled by a rule-based approach.

This paper is organized as follows: Section 2 explains the phonetic category transition of Mandarin Chinese. Section 3 introduces our use and design of statistical pattern recognition. Section 4 describes a new scheme to improve the most difficult cases. Finally, Section 5 gives the conclusions and possible future work.

2. The phonetic category transition of Mandarin Chinese

2.1. The four phonetic categories

There are 37 distinct phonetic alphabets in Mandarin Chinese. We divide them into four categories according to their acoustic characteristics. These four categories are fricative and affricate, unaspirated stop, aspirated stop, and periodic voiced, as listed in the following using SAMPA (Speech Assessment Methods Phonetic Alphabet) format [5].

- Fricative and affricate: (consonants only)
  (Fricative) SAMPA: f x 6 S s
  (Affricate) SAMPA: t6 t6_h TS TS_h ts ts_h
- Unaspirated stop: (consonants only)
  SAMPA: p t k
- Aspirated stop: (consonants only)
  SAMPA: p_h t_h k_h
- Periodic voiced: (consonants and vowels)
  (Consonants) SAMPA: m n l Z
  (Vowels) SAMPA: a o @ e ai ei au ou an @n aN @N 2 I U y

2.2. All phonetic category transitions in Mandarin

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.

Based on a similar analysis, we can construct Table 1 to list all possible transitions (from the left side to the right side of a boundary) for the beginning and ending boundaries of a syllable.

Table 1. Possible phonetic category transitions (denoted by O) of the beginning, and ending boundaries.

Left side	Right side	Beginning boundary	Ending boundary
S	F	O	X
S	U	O	X
S	A	O	X
S	P	O	X
P	F	O	O
P	U	O	O
P	A	O	O
P	P	O	O
P	S	X	O

PS. S: silence, F: fricative and affricate, A: aspirated stop, U: unaspirated stop, P: periodic voiced, O: possible transition, X: impossible transition.

3. The statistical pattern recognition

3.1. Feature selection for each phonetic category transition

It is evident that not all features work equally well for each type of transition. Therefore we must systematically pick up the most suitable acoustic features for each transition. In addition, we collected a speech corpus called TRAIN3700 which contains about 3700 syllables of 30 long sentences (by one speaker). The beginning and ending phonetic boundaries of these 3700 syllables are manually labeled.

3.1.1. Defining candidate boundaries for training data

In order to find the most discriminative features, we must create a set of training data. This is achieved by adding several candidate boundaries, 5 ms apart, located within ± 50 ms of a true (manually labeled) boundary. A candidate boundary is labeled "correct" if it is within ± 10 ms of the true boundary. (According to Chou [3], manual labeling of two human experts can achieve about 90% consistency on 10 ms tolerance.) Therefore we chose to have 5 "correct" candidates (including the true one), all within ± 10 ms of the manually labeled one, for our experiments. On the other hand, we chose 8 "wrong" candidates located within [35, 50] ms and [-35,-50] ms of the true boundary. In other words, for each true boundary, we can create a set of 13 candidate boundaries, which have 5 labeled "correct" and 8 labeled "wrong" as their desired classification output as shown in Figure 1.

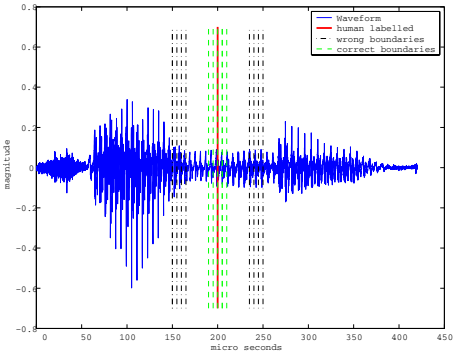

Figure 1. Training data of 5 correct boundaries and 8 wrong boundaries around the true boundary labeled by human. The content of this waveform is "將離" ("t6-I-aN, l-I" in SAMPA).

3.1.2. Feature definition

In this paper, we adopted seven acoustic features, including zero-crossing rate, log energy, entropy, bisector frequency [1], burst degree [1], pitch and MFCCs respectively. All features are scalars except for MFCCs which is a 24-dimensional vector. However, we did not take these features directly as our training features. In our implementation, for each candidate boundary, we evaluate the differences between all acoustic features of its left and right frames (The size of a frame is set to 20 ms). The "difference of acoustic features" is then used as the features for designing a classifier. Besides, these features also have to be normalized to the range [-1, 1] before training. Therefore, there are seven single-dimension features for representing each candidate boundary.

3.1.3. Feature selection by SFS, KNNR and LOO

In order to find the most influential features, we employed the method of sequential forward selection (SFS) proposed by Whitney [9] in the pattern recognition literature. The principle of SFS is to start with a single feature with the best classification rate. Then we try to identify a newly added feature to the already selected feature set that can most increase the classification rate. This greedy step is repeated until the desired number of features has been selected, or until there is no improvement in 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) as the performance criterion. The basic concept of 1-NNR is to assign the class of a given test vector based on the nearest data point in the training data. In order to have a better degree of robustness, we chose to have KNNR, in which the nearest k neighbors are selected around the test vector and the assigned class is determined by a voting mechanism.

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 average classification rate of these test vectors. We have performed a simple search to find the best value of K in KNNR is 9 in our experiment..

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.

3.1.4. *Classification rates for phonetic category transitions*

After the use of SKL (SFS, KNNR, and LOO), we can obtain the training data set of each phonetic category transition and its corresponding classification rates. In other words, there are totally 13 training data set (8 for beginning and 5 for ending) to be used for boundary refinement. Figure 2 shows the result of SFS for the "periodic voiced + aspirated stop" phonetic category at the beginning boundary, where the x-axis is the selected features and the y-axis is the LOO classification rates.

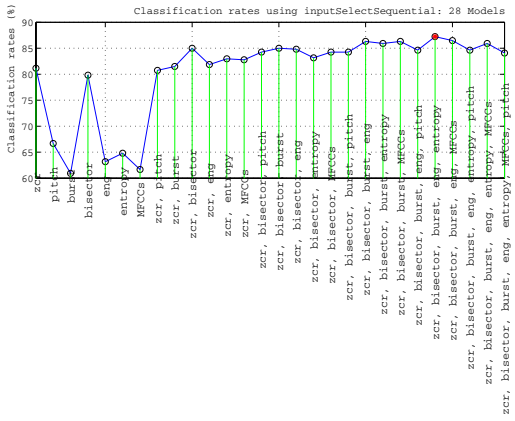

Figure 2. The LOO classification rates for different combinations of acoustic features for "periodic voiced + aspirated stop" transition at the beginning boundary.

From Figure 2, it is evident that the most discriminant features for the "periodic voiced + aspirated stop" category at the beginning boundary are zero-crossing rate, bisector frequency, log energy, entropy and burst degree, which are able to achieve a LOO classification rate of 87.2%. The classification rates for all phonetic category transitions are demonstrated in Table 2.

Table 2. Classification rates for all phonetic category transitions.

Beginning	Classification rate	Ending	Classification rate
S+F	95.1%	S+F	X
S+U	95.6%	S+U	X
S+A	92.3%	S+A	X
S+P	83.9%	S+P	X
P+F	94.8%	P+F	88.8%
P+U	89.7%	P+U	81.2%
P+A	87.2%	P+A	81.5%
P+P	63.8%	P+P	64.2%
P+S	X	P+S	83.2%

Here, S, F, U, A and P are the same with the ones in Table 1.

From Table 2, it can be observed that the classification rates of the "periodic voiced + periodic voiced" (later abbreviated as "P + P") transition are comparatively low, which are only about 63.8% for beginning boundaries and 64.2% for ending boundaries. We shall develop special method for this type of transition.

3.2. Boundary refinement using SKL

In this subsection, we explore the use of SKL for boundary refinement after HMM segmentation. In particular, we shall demonstrate the "P + P" transition still performs poorly and requires other special treatment.

We prepared two corpora for our experiment, TTS-455 and TCC-300 [4]. The TCC-300 corpus is used to train the HMM models for forced alignment, and the TTS-455 corpus, containing 455 sentences (about 6000 syllables by one speaker, all phonetic boundaries are manually labelled), is left for evaluating the performance of refined boundaries.

At first, we obtained an initial estimate of the beginning/ending boundaries for the TTS-455 corpus by the TCC-300-trained HMM-based recognizer (which uses context-dependent triphone models). Then for every initial boundary, we selected candidate boundaries as the test set, which are 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.

In order to observe the difference in performance between the non-"P + P" cases and the "P + P" cases, we gather classification rates of boundary refinement for all boundaries (beginning and ending) in the TTS-455 corpus, as shown in Figure 3.

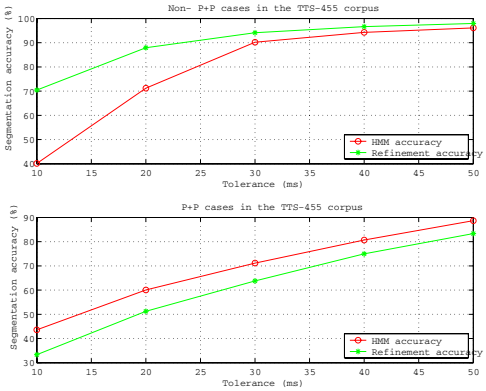

Figure 3. Segmentation accuracy of non-"P + P" and "P + P" cases.

From Figure 3, we can observe that "P + P" cases perform badly for boundary refinement, even worse than those of HMM segmentation. In contrast, non-"P + P" cases have consistently much better performance. This phenomenon is consistent with the classification rates of the SKL procedure mentioned in the previous subsection. As a result, it is necessary to design an alternative method for "P + P" category.

4. Method for "P + P" category

4.1. "Divide and conquer" method

According to our observation, not all cases in the "P + P" category perform poorly. In other words, combining all these transitions into a single category may be too coarse for further refinement. However, to divide the "P + P" category into all possible phonetic transitions is also impractical since the training data is limited. Hence, we used the TCC-300-trained HMM models to identify the boundaries of "P + P" category in the TRAIN3700 corpus. In fact, there are about 235 different phonetic transitions in this "P + P" category. We then divided them into two groups denoted by group B ("better", with segmentation accuracy >= 80% within 20ms) and group W ("worse", segmentation accuracy < 80% within 20ms). Group

B contains 72 phonetic transitions while group W contains 163. A heuristic approach is used to refine the boundaries in group W, as explained in the next subsection.

4.2. A heuristic approach for group W

We noticed that most human labeled boundaries in group W are located in a region with lower log energy. Thus we proposed a heuristic method to refine boundaries in group W using MFCCs and log energy, as follows:

1. For every initial boundary (by an HMM-based recognizer), we selected candidate boundaries, 2 ms apart, and within 80 ms at both sides of this boundary. In other words, there are 81 candidate boundaries in this ± 80 ms search region.
2. Calculate the log energy (being normalized to [0, 1]) in the search region and find its average value. Identify a new search region whose log energy is less than 90% of the average log energy.
3. For boundaries within the new search region, select the one with the maximum absolute difference between MFCCs of its left and right frames.

Figure 4 shows a typical result.

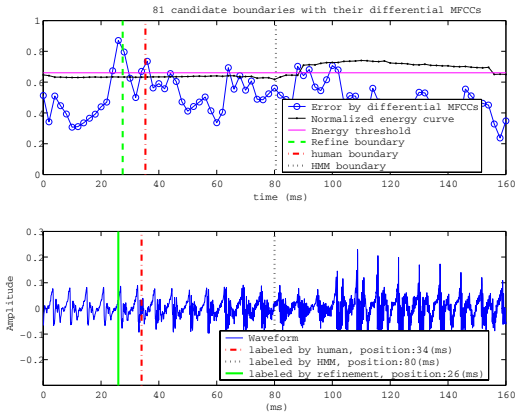

Figure 4. A typical result of the proposed heuristic method for group W.

4.3. Performance of boundary refinement using SKL together with the heuristic rule

In order to verify the feasibility of the proposed method, we repeated the same experiment mentioned in Section 2.4, but only focused on the "P + P" category. Here, we used the heuristic rule to carry out the boundary refinement for group W of "P + P".

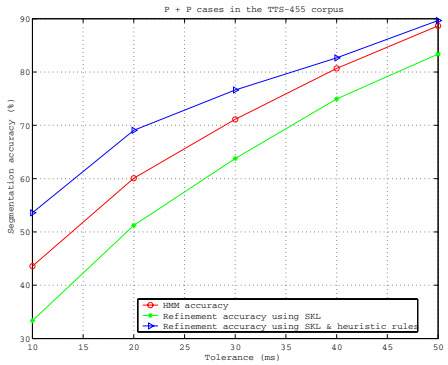

Figure 5. The comparison of boundary refinement between using SKL method and using SKL together with heuristic rules.

For group B of "P + P" transition, we applied the same SKL procedure. Figure 5 shows the experimental result on the test data of "P + P" category in the TTS-455 corpus. It is obvious that the performance is significantly improved.

5. Conclusions and future work

This paper proposed a hybrid approach to boundary refinement of automatic phonetic labeling. The most suitable acoustic feature sets for each phonetic category is found by the SKL procedure. Moreover, we have divided the poorly performed "P + P" category into group B and W, and proposed a heuristic method for improving the performance of group W. The experimental results demonstrate the improved performance.

Though the performance of group W is improved, but admittedly it is still not good enough for fully automatic labeling of corpus-based TTS applications. Further observations lead to future research directions, as follows:

1. Error cases in group W usually have a very strong co-articulation between neighboring syllables, such "第一" ("t-I, I" in SAMPA). Consistent segmentation of this type is also difficult for human. As a result, we need to take other factors into consideration, such as text-dependent time duration of each phone. For TTS applications, we may want to keep the transitions in group W as a single synthesis unit to avoid possible segmentation mistakes.
2. The size of TRAIN3700 data is not large enough. An on-going project is to use a larger corpus that can support more training data for better analysis.

6. References

[1] Cheng-Yuan Lin, Jyh-Shing Roger Jang, Kuan-Ting Chen, "Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpus for Concatenation-based TTS", International Journal of Computational Linguistics and Chinese Language Processing, 2005.

[2] D. Torre Toledano et al. "Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-correction Rules", Proc. Third ESCA/COCOSDA Workshop on SPEECH SYNTHESIS, 1998.

[3] 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.

[4] http://rocling.iis.sinica.edu.tw/ROCLING/MAT/Tcc_300brief.htm

[5] http://www.phon.ucl.ac.uk/home/sampa/home.htm

[6] Jan P. H. van Santen, J., Sproat, R. "High-accuracy automatic segmentation", Proceedings of European Conference on Speech Communication and Technology, 1990.

[7] 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.

[8] LiJuan Wang et al. "Refining Segmental Boundaries for TTS database Using Fine Contextual-Dependent Boundary Models", ICASSP 2004.

[9] Richard O. Duda, Peter E. Hart, David G. Stork, "Pattern classification, 2nd edition", New York, Wiley, 2001.

[10] Sethy, A. Narayanan, S, "Refined Speech Segmentation for Concatenative Speech Synthesis", ICSLP, 2002, pp. 149-152.