Refining Segmental Boundaries for TTS Database Using Fine Contextual-Dependent Boundary Models

Lijuan Wang[1], *Yong Zhao*[2], *Min Chu*[2], *Jianlai Zhou*[2] *and Zhigang Cao*[1]

[1]Department of Electrics Engineering, Univ. of Tsinghua, China
[2]Microsoft Research Asia, Beijing, China
wlj01@mails.tsinghua.edu.cn

ABSTRACT

This paper proposed a post-refining method with fine contextual-dependent GMMs for the auto-segmentation task. A GMM trained with a super feature vector extracted from multiple evenly spaced frames near the boundary is suggested to describe the waveform evolution across a boundary. CART is used to cluster acoustically similar GMMs, so that the GMM for each leaf node is reliably trained by the limited manually labeled boundaries. An accuracy of 90% is thus achieved when only 250 manually labeled sentences are provided to train the refining models.

1. INTRODUCTION

Labeling the segmental boundary in speech waveforms is essential for a corpus-based concatenative TTS system. Manual labeling is reliable, yet, labor and time consuming. Thus, it is desirable to have an automatic approach for segmentation, especially when the speech corpus is very large. The most popular automatic segmentation method, termed *forced alignment*, is an HMM-based approach that has been widely used in the training stage of automatic speech recognition (ASR). For doing *forced alignment*, Viterbi algorithm is applied to find out the most probable boundaries for the known sequence of speech units. However, such boundaries are not necessary the best concatenation points for these units. Thus, post-refinement is often performed to search for the most suitable locations for all boundaries as illustrated in Figure1, in which a small amount of manually labeled boundaries have to be provided for learning the characteristics of the preferred boundary locations. Various refining techniques, such as Gaussian Mixture Model (GMM) [1], Hidden Markov Model (HMM) [1], Neural Networks (NN) [2] and MLPs [3][4], have been proposed to portray the boundary property. It has been found that, if boundaries were classified into groups by their phonemic context, such as Vowel, Nasals, Liquids etc, and a refining model was trained for each group, more precise auto-segmented boundaries were obtained [1][3]. However, such a classification is still coarse. The phonemic context within the same group may vary greatly. For example, 'i' and 'u', which are often clustered into the Vowel group, have quite different formant trajectories. Modeling them with the same refining model will loose precision. The ideal solution is to train an individual model for each pair of phone boundaries. However, there are normally not sufficient manually labeled boundaries for training so many individual models. Thus, this paper presents a CART based method that clusters segmental boundaries automatically according to their similarity in acoustic features. And an individual refining model is trained for each leaf node. With this method, it is convenient to adjust the number of models according to the amount of training data available.

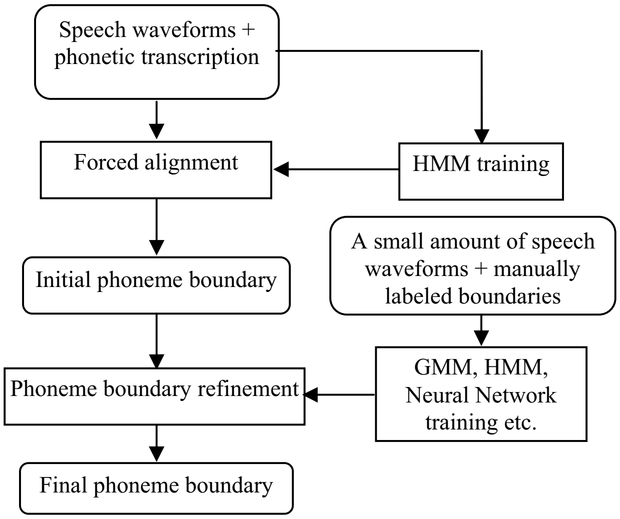
Speech waveforms + phonetic transcription
Forced alignment
HMM training
Initial phoneme boundary
A small amount of speech waveforms + manually labeled boundaries
Phoneme boundary refinement
GMM, HMM, Neural Network training etc.
Final phoneme boundary

Figure1: Block diagram of the two-step automatic segmentation

This paper is organized as follows: Section 2 elaborates the proposed refinement method. Section 3 presents the evaluation experiments and results. Finally, Section 4 gives conclusions and outlines for future work in this field.

2. THE FINE CONTEXTUAL-DEPENDENT BOUNDARY MODELS

2.1. The super feature vector for a boundary

0-7803-8484-9/04/$20.00 ©2004 IEEE

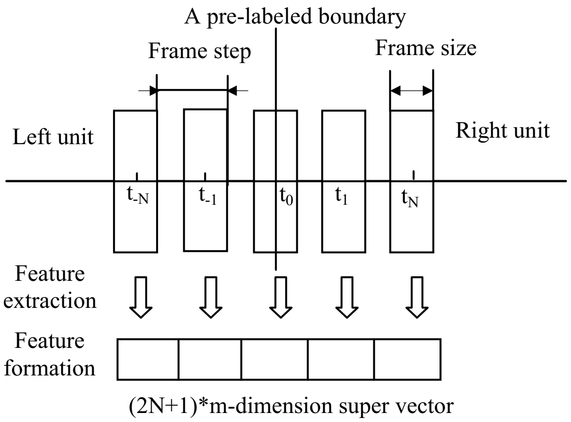

Figure2: An illustration of the extraction and formation of the super feature vector for a boundary

In previous studies [1], N GMMs were trained from the N frames of features nearby the pre-labeled boundaries and the weighted sum of the likelihoods from all GMMs is used in the boundary refinement. However, these weights are not easy to be optimized. Furthermore, the independent assumption of the N frames in such a method is not true. Therefore, in this paper, only one GMM is trained with a super feature vector as illustrated in figure2. First, 2N+1 frames of acoustic features, m-dimension each, are extracted from time t_{-N} to t_N, where t_0 is the pre-labeled boundary, t_{-N} to t_{-1} are N frame to the left of the boundary and t_1 to t_N are N frame to the right of the boundary. Then the 2N+1 acoustic features are put together to form a (2N+1)*m-dimension super vector for that boundary. Normally, the frame step is set to be larger than the frame size so that consecutive frames are not overlapped and the super feature vector contains more information about the boundary.

2.2. Clustering segmental boundaries by CART

The evolution of the speech waveform across a segmental boundary is determined by the property of units on the left and right sides of the boundary. Thus, a boundary can be represented by a pseudo-triphone in the formation of X-B-Y, where B represents a boundary, X represents the phoneme to the left of the boundary, and Y represents the phoneme to the right of it. Theoretically, there are N_X*N_Y possible such pseudo-triphones, where N_X is the number of X in categories and N_Y is that of Y. N_X and N_X are not necessarily the same.

For modeling each type of boundaries precisely, training a *GMM* for each pseudo-triphone is desired. However, since there are normally limited manually labeled data available for training, it is not realistic to train a reliable model for each pseudo-triphone. Therefore, Classification and Regression Tree (CART) is used to cluster similar pseudo-triphones into the same category. Those unseen pseudo-triphones can be mapped to a suitable leaf node as well. Then a GMM is trained for each leaf node and it is used for refining the boundaries of the types belong to that leaf node.

Since the segmental boundaries are treated as a pseudo-triphone, the model clustering procedure is the same as what is done in training acoustic models for speech units. In fact, the same question set can be used as well.

2.3. Boundary refinement

Once the training is completed, automatic refinement of the corpus can start. An approach similar to that used in [1] is adopted in our studies. For a specific boundary to be refined, the optimal boundary is assumed to be in the vicinity of the initial boundary, i.e. a more suitable boundary is to be searched in a certain range around the initial one. Normally, a small frame step is used in the refining stage in order to get precise locations of boundaries. Acoustic features are first extracted for frames in the search range. Next, a leaf node on the CART is found by querying to it corresponding pseudo-triphone. Then, the likelihood for each frame in the search range is calculated using the pre-trained GMM for that leaf node. The frame that has the maximum likelihood is regarded as the optimal boundary. Obviously, the smaller the frame step is, the more precise the optimal boundary will be, however, at the cost of more calculations.

3. EXPERIMENTS AND RESULTS

3.1. Speech Corpus

The Microsoft Mandarin TTS speech corpus, which contains manually verified syllable boundaries, is used to evaluate the performance of the proposed method. The whole corpus, containing about 12,000 sentences are used to train the HMM models for forced alignment. However, only a small part of the manually labeled boundaries (1000-50,000 boundaries) are used to train the refining models. These models are tested on 10,000 boundaries out of the training set.

3.2. Method for performance evaluation

Although HMM models are trained for units smaller than syllables for performing forced alignment, only syllable boundaries are kept in this study since syllable is normally the base unit in Chinese concatenative speech synthesis. The automatically labeled boundaries are compared with the manually labeled boundaries. All auto-boundaries that have distances to the manual-boundaries smaller than a pre-defined tolerance threshold, such as 10ms or 20ms, are counted as correct labels. Then the percentage of correct labels is said to be the accuracy for the given threshold. For TTS applications, a 10 to 20ms threshold is normally used.

3.3. The baseline performance

A speaker-dependent large-vocabulary continuous speech recognition (LVCSR) system is first trained with the HTK toolkits [6] on the entire speech corpus, in which tone-independent triphone models are created. Then the system is used to perform forced alignment over the corpus. To make sure that the models are trained well enough, they are used to do decoding on the 500 utterances out of the training set. The pretty low base syllable error rate 7.34% shows that these models are trained well. However, the accuracy for the aligned syllable boundaries is only 73.67% for the 20ms threshold. Thus, we believe that there is not much room for increase the boundary accuracy by improving the HMM models. Therefore, several experiments on post-refinement are carried out. The accuracy of boundaries obtained by forced alignment is used as the baseline for evaluation the validity of the refining model.

3.4. Configuration of the refining model

During the training phase, 5 frames of acoustic features are extracted around the manually labeled boundary as illustrated in Figure2. The frame size is 20ms and the frame step is 30ms. For each frame, a 39-dimension vector, composed of 12 MFCC + energy, 13 first order deviations and 13 secondary deviations, is calculated. Thus, a super feature vector of (5*39) dimensions is formed. Both the GMM training and CART based clustering are carried out using HTK toolkits [6]. Each individual GMM is represented by a 1-state HMM.

3.5. Experiment 1: accuracy vs. the number of Gaussian components

This experiment is designed to investigate how many Gaussian components should be used in the refining model. So, the size of training set is set to be 20,000 instances and the stop criteria for growing the CART is that each leaf node contains at least 20 instances. As a result, 154 leaf nodes are obtained. Then, 1 to 8 components are trained for each leaf node. The testing results for these models are listed in Table1. It is interesting to see that the accuracy drops when the number of Gaussian components increases.

The results show that single Gaussian component is good enough to model the distribution of the super feature vector across a class of boundaries at the leaf node of the CART, i.e. boundaries that are clustered into the same leaf node have very simple distributions. In fact, increasing the number of mixtures does hurt the accuracy of the refinement to some extent. The reason for this may be that, when the number of instances on some leaf nodes is small, the parameters of multiple Gaussian mixtures cannot be estimated reliably.

Table 1: refinement accuracy vs. the number of Gaussian components

Tolerance (ms)	Base-line	Gaussian mixtures			
		1	2	4	8
10	45.8	69.9	69.4	68.0	65.9
20	73.6	91.5	91.1	90.3	89.2
30	86.9	96.9	96.7	96.5	96.1

Table 2: refinement accuracy in 20ms tolerance vs. MTI

Training size	Base-line	*MTI*					
		2	5	10	20	40	80
5,000	73.6	81.7	89.8	90.0	89.8	88.8	86.3
20,000	73.6	91.4	91.4	91.5	91.2	91.1	90.3

3.6. Experiment 2: accuracy vs. the number of CART nodes

This experiment is designed to find out how deep the CART should be grown. The number of Gaussian components is set to 1, based on the result of experiment 1. Models are trained on a training set with 5000 instances and 20,000 instances respectively. Through adjusting the *Minimum Training Instances* (*MTI*) per leaf node, CARTs of different scales are obtained. As the *MTI* decreases, the number of leaf nodes on the CART (also the number of *GMMs*) increases. The testing results for models trained with different *MTI* are given in Table2. It is found that, when training with the 20,000 set, the accuracy of the refinement drops if the *MTI* is set to values larger than 40 and the accuracy is almost unchanged for all other settings. However, when the train set is reduced to 5000 samples, the accuracy of refinement increases as the *MTI* decreases until it reaches 10. This implies that the accuracy of refinement will increase when more contextual-dependent models are used as long as a minimum number of instances for training a reliable Gaussian mixture is guaranteed, i.e. fine contextual-dependent model performs better than the kind of rough contextual model described in [1][3].

3.7. Experiment 3: accuracy vs. the size of training corpus

Experiment 3 studies how many manually labeled data should be provided in order to get high refinement accuracy. In this experiment, the number of Gaussian components is again set to 1 and *MTI* is set to 10. From the results shown in Figure3, it is seen that as the size of training set exceeds 5,000, the rate of performance improvement starts to slow down. Of course, more training data is still helpful. The curve becomes saturated after the train set reaches 30,000. Therefore, it is recommended to provide at least 5,000 correct boundaries (approximately 250 utterances) for training the refining models. If 30,000 precise boundaries are provided, the proposed method can achieve the highest accuracy: 91.9%.

Figure 4 shows the improvement on various tolerant thresholds, when 20,000 training instances are provided. Significant improvements are achieved on all thresholds.

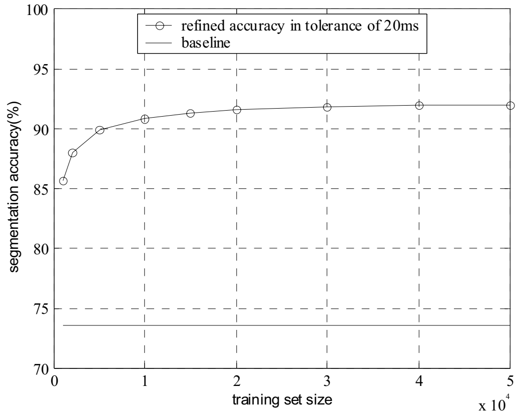

Figure3: Accuracy of refined boundaries vs. size of training set.

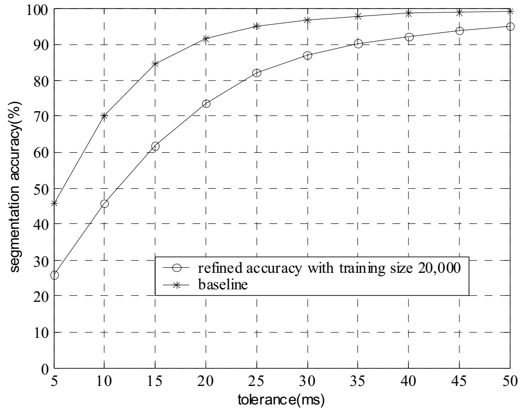

Figure4: Improvement achieved by refinement, when 20000 manually labeled boundaries are provided.

4. CONCLUSION

This paper proposed a post-refining method with fine contextual-dependent GMMs for the auto-segmentation task. The experiment results show that, on the one hand, the distribution of the super feature vectors of boundaries within a fine cluster is pretty simple and can be described with a single Gaussian mixture, on the other hand, the more precise the classification is, the more consistent the boundaries belonged to one class will be, as a result the higher accuracy will be achieved.

The proposed method validates even for a small amount of training data, say containing only 1000-2000 instances (or 50-100 manually labeled sentences). Pretty high accuracy (90%) can be achieved when 5000 instances (or about 250 sentences) are provided for training the refine models. The best performance (91.9% accuracy) is achieved when 30000 samples (or about 1500 sentences) are provided.

In addition, this method is quite generic and makes no inherent assumptions on the language or speaker type. We have applied this method to the language of English and have achieved similar improvement, which is not shown here. Further research may focus on the adjustment of frame number and frame step in the boundary feature extraction and the extension to other languages.

5. REFERENCES

[1] Abhinav Sethy, Shrikanth Narayanam, “Refined speech segmentation for concatenative speech synthesis,” Proc. ICSLP, pp.145-148,2002.

[2] D.T. Toledano, “Neural network boundary refining for automatic speech segmentation,” Proc. ICASSP, pp.3438-3441, 2000.

[3] KI-Seung Lee and JeongSu Kim, “Context-adaptive phone boundary refining for a TTS database,” Proc. ICASSP, pp.252-255, 2003.

[4] Eun-Young Park, Sang-Hun Kim and Jae-Ho Chung, “Automatic speech synthesis unit generation with MLP based postprocessor against auto-segmented phoneme errors,” Proc. ICASSP, pp.2985-2990,1999.

[5] Yeon-Jun Kim and Alistair Conkie, “Automatic segmentation combining an HMM-based approach and spectral boundary correction,” Proc. ICSLP, pp.145-148,2002.

[6] Odell J, Ollason D, Woodland P, Young S, Jansen J, “The HTK Book for HTK V3.0”, Cambridge University Press, Cambridge, UK, 2001.