Automatic Phonetic Segmentation of Malay Speech Database

Chee-Ming Ting, Sh-Hussain Salleh, Tian-Swee Tan, A. K. Ariff.
Center for Biomedical Engineering, Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81300 Skudai, Johor, Malaysia.

***Abstract*— This paper deals with automatic phonetic segmentation for Malay continuous speech. This study investigates fast and automatic phone segmentation in preparing database for Malay concatenative Text-to-Speech (TTS) systems. A 35 Malay phone set has been chosen, which is suitable for building Malay TTS. The segmentation experiment is based on this phone set. HMM based segmentation approach which uses Viterbi force alignment technique is adapted. We use continuous density HMM (CDHMM) with Gaussian mixture which is performs well in speech recognition to prevent large segmentation errors. Besides, this paper presents an implicit boundary refinement method that is incorporated in the Viterbi phonetic alignment. In this approach, the HMM model is trained with phone tokens with their boundaries extended to the be-side phones. This increases the ability of the HMM in modeling phone boundaries and provides effect of implicit boundary refinement when used in phonetic alignment thus reduce segmentation errors. This approach improves increase the performance of baseline HMM segmentation from 42.39%, 74.83%, 84.34% of automatic boundary marks within error smaller than 5, 15, and 25ms to 47.75%, 76.38%, 85.55%.**

***Keywords*— Speech recognition, Speech synthesis.**

I. INTRODUCTION

Segmenting the phone boundaries in speech waveforms is essential for a corpus based concatenative TTS system. The most precise way of phonetic segmentation is manually. However, manual segmentation is very costly and requires much time and effort. Thus, it is desirable to have an automatic approach for segmentation, especially when the speech corpus is very large. Many phonetic segmentation tools have existed, but not suitable for building inventories for Malay TTS because these tools are based on other language. Speech in different language domain is characterized by different phone set, thus it is improper to represent Malay speech with other language phone set. Identifying own phone set for Malay speech is crucial for accurate phone modeling. This study provides a basis in developing an automatic Malay phone segmentation system for Malay TTS.

In this study, forced alignment, an HMM-based approach [1,2,3] is used. This approach which adopted from automatic speech recognition (ASR) is most widely used for automatic segmentation in speech synthesis, providing consistent and accurate phone segmentation. In this approach, forced alignment using Viterbi algorithm is applied to find out the most probable boundaries for the known sequences of phone units. However, such boundaries are not necessary the best concatenation points for these units. Its limited ability to remove discontinuities at concatenation points is because of the Viterbi alignment tries to find the best HMM sequence when given a phonetic transcription and a sequences of HMM parameters, not the optimal boundaries between adjacent phones. Usually a post-refinement technique [1,4] is performed to search for the most suitable locations for all boundaries, in which a small amount of manually labeled boundaries have to be provided for learning the characteristics of the preferred boundary locations. Thus, extra boundary modeling procedure and boundary feature extraction is needed. This increase the mathematical complexity and computational time.

This paper proposes an implicit boundary refinement (IBR) method that embedded in the Viterbi forced alignment. First the start and end point of the training tokens are extended to their adjacent phones. Thus, the extended training tokens take consideration on a wider range of boundaries located between them and their adjacent phones. The HMM phone models trained on these extended tokens will better modeling the phonetic boundaries. The Viterbi alignment using these models implicitly refines the boundaries and reduces segmentation errors.

This paper is organized as follows: Section 2 elaborates the HMM based approach for phonetic segmentation. Section 3 presents the proposed implicit boundary refinement method. Section 4 presents the evaluation experiments and results. Finally section 5 given conclusions and outlines for the future work.

II. HMM FOR PHONETIC SEGMENTATION

The most frequent approach for automatic phonetic segmentation is to modify an HMM based phonetic recognizer to adapt it to the task of automatic phonetic segmentation. The main modification needed consists in letting the recognizer know the phonetic transcription of the sentence to segment by building a recognizer's grammar for that transcription and performing forced alignment. The segmentation system used in this study consists of two phases; training HMMs, and phone segmentation using Viterbi alignment of shown in Figure 1. The system uses speaker-dependent HMMs – SDHMMs which built from labeled and segmented training data set. The model is used to segment the speech waveform from the same speaker. SD HMMs are generally used for automatic segmentation in speech synthesis, but have the drawback of consuming much time to prepare.

1-4244-0983-7/07/$25.00 ©2007 IEEE

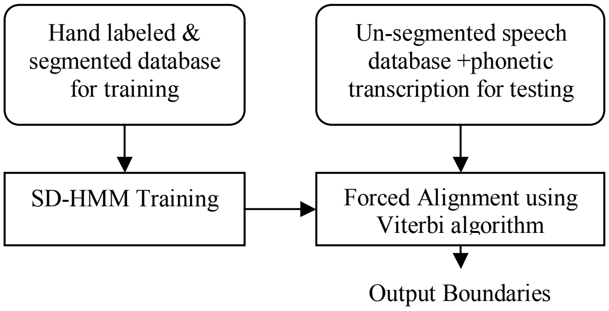

Figure 1. Block diagram of automatic segmentation system.

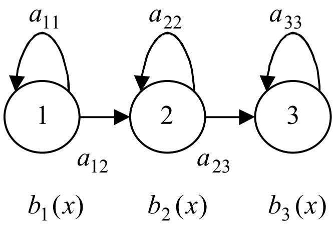

Figure 2. Representation of left-to-right HMM.

For acoustic modeling, 3-state left-to-right with no skips CDHMM with Gaussian mixture density [6] as shown at Figure 2 is used to represent a phone. The parameters which characterize the CDHMM of Figure 2 are the following: (1) $A = \lfloor a_{ij} \rfloor, 1 \le i, j \le N$, the state transition matrix where a_{ij} is the probability of making a transition from state *i* to state *j*; and (2) B, the observation probability function associated with each state j. The observation probability distribution of CDHMM is modeled by continuous probability density function, $B = \{b_j(x)\}$, for $1 \le j \le N$ and x represents continuous observations of K-dimensional random vectors. The most general representation of the pdf of CDHMM, is a finite Gaussian mixture density of the form

$$b_j(x) = \sum_{m=1}^{M} c_{jm} \mathrm{N}(x, \mu_{jm}, \Sigma_{jm}), \qquad 1 \le j \le N \qquad (1)$$

,where x is vector being modeled, c_{jm} is the mixture coefficient for the *m*th mixture component in state *j* and N is Gaussian density, with mean vector $\mu_{jm} = [\mu_{jmd}]$ and covariance matrix $\Sigma_{jm} = [\Sigma_{jmde}]$ for the *m*th mixture component in state *j*, for $1 \le d, e < D$, where *D* is the number of dimensions in feature vectors. Diagonal matrices are used.

The reason for using 3 states is to incorporate coarticulatory effects implicitly into the model. The first part and third part are assumed to account for coarticulary effects due to the transitions to the neighboring phonemes, whereas the second part stands for the middle part of the phoneme which is known to be less affected by phonetic context. The essential advantage of the mixture density is that several maxima in the density function can be modeled, which may correspond to different acoustic realizations of the same phoneme due to coarticulary effects. This model which provides good recognition accuracy in ASR is used to prevent very large segmentation errors. It is common practice to use context independent HMMs for speech segmentation [2,5]. Context-independent HMMs, trained with realizations of phone in different context are able to discriminate between the phone to model and its context which varies. They produce more precise segmentation than its context-dependent counterparts which are trained with phone in same context and unable to discriminate between the phone and its context. 8 iterations of Viterbi re-estimation procedure is used to train the acoustic models.

III. Implicit Boundary Refinement

Most of the segmentation systems introduce the post refinement method on the initial phone boundaries obtained from the Viterbi alignment. This paper proposes an implicit refinement method when doing phonetic alignment; this will save time and reduce complexity. The proposed automatic segmentation system with implicit boundary refinement is shown in Figure 3.

Figure 4 shows the concepts of start and end point extension of the training tokens. The figure shows the waveform segment of three phonemes in adjacent, /a/-/n/-/a/, which was manually segmented and labeled (indicated by the solid lines). Actually there is fuzziness in determining the location of the boundary between adjacent phones, even the manual-boundary is an approximation. The boundaries of the manually segmented phone token is extended from the manual-boundary to the beside phones (the new start and end point of token shown by dotted lines in Figure 4) to take consideration of a wider range of probable boundaries. The HMMs trained from the extended training tokens are more capable in modeling the phone boundaries with its first and final states. There is a smooth transition form one phone model to another for the concatenated sentences HMM which is used for alignment of the speech feature patterns. Thus the Viterbi phonetic alignment using these models will implicitly refine the phone boundaries and reduce segmentation errors. This technique also provides better modeling of contextual effects.

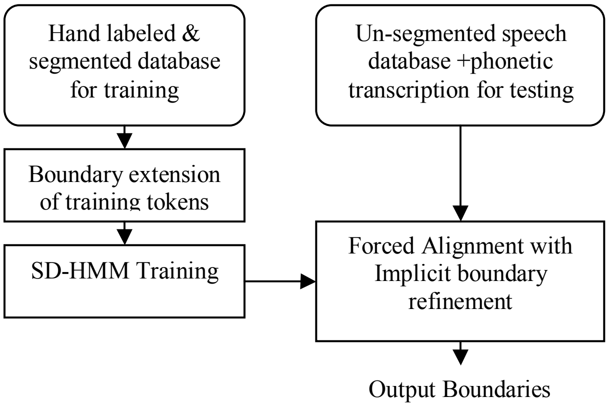

Figure 3. Block diagram of automatic segmentation system with implicit boundary refinement.

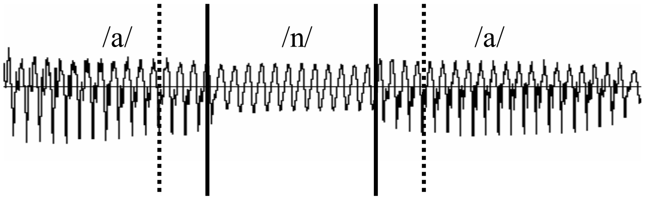

Figure 4. Start and end point extension of the training phone tokens for three adjacent Malay phonemes /a/-/n/-/a/.

IV. Experiments and Results

This section describes a series of segmentation experiments to evaluate the effectiveness of HMM-based approach in segmenting Malay phone units. Experiments were also done to compare the baseline system with the newly proposed system which incorporates implicit boundary refinement in terms of segmentation accuracy.

A. Database and Analysis Condition

The database consists of 71 phonetically balanced Malay continuous speech sentences, uttered by a Malay female speaker. 39 utterances were used for training purpose. The results presented are evaluation on the other 32 utterances. All the sentences have been hand labeled and segmented according to the chosen Malay phone set in Tabel 1.

TABLE I. List of Malay Phones According to Categories

Category	Malay Phones
Vowels	/a/, /e/, /eh/, /i/, /o/, /u/
Plosives	/b/, /d/, /g/, /p/, /t/, /k/
Affricates	/j/, /c/
Fricatives	/s/, /h/, /f/, /z/, /sy/, /kh/, /gh/, /v/.
Nasal	/m/, /n/, /ng/, /ny/
Trill	/r/
Lateral	/l/
Semi-vowel	/w/, /y/

In Malay language, there are 24 pure phonemes and about 6 borrowed phonemes, divided into 8 categories [7]. Among the pure phonemes, there are 18 consonants and 6 vowels. The borrowed consonantal phonemes are /f, z, sy, kh, gh, v/. There are 5 diphthongs in Malay language: /ai/, /au/, /oi/, /ua/, /ia/. This Malay phone set cover all Malay phoneme unit in Malay language and thus suitable for characterize Malay speech corpus for building TTS system. The segmentation experiment is based on the 35 phones above and a garbage model for pausing /pau/. The /gh/ was folded to /g/ due to limited training tokens.

The speech was sampled at 16 KHz. Our front end computed mel-frequency cepstral coefficients (MFCCs) and enegy value with 15ms windows at a 5 ms frame rate. We retained 12 MFCC coefficients and a normalized power value for each frame, along with their first and second order derivatives.

B. Performance Evaluation Method

To evaluate the segmentation system, the objective method is used, which measures the agreement with manual segmentation. The automatically segmented boundaries are compared with the manually segmented boundaries. The percentage of boundaries whose error is within a tolerance is measured for a range of tolerances. In this study, we calculated the percentage of boundaries within a set of tolerances which are 5, 10, 15, 20, and 25ms.

C. Effect of Different Feature Set and Varying Number of Gaussian Mixture Components.

A series experiments were done using various combination of features along with varying number of Gaussian mixture components to investigate its effect on segmentation performance. This experiment is based on HMM based segmentation without IBR. The number of Gaussian mixtures, M was varied from 1 to 4 in steps of 2 to investigate its effect on accuracy. The segmentation result is given in Table 2.

Table 2 shows an interesting pattern behavior that at three different tolerance zones, certain HMMs behave better than the others. When the varying number of mixtures is concerned and when the Delta (D) and Delta-Delta (DD) feature is used,, the result shows that, for small tolerance (5-10ms), HMMs with fewer Gaussians perform better. For large tolerances (>20ms) HMMs with more Gaussians perform better generally. For medium tolerances (15ms), the result shows an intermediate change in segmentation result between small and large tolerances where increasing the Gaussian start to perform better. This result is consistent with the result in [1] which uses almost the same feature set of 12MFCC and normalized log energy (E), and first and second order differences. Very large segmentation errors are highly correlated with phone misrecognitions. Therefore, the segmentation results in this range of tolerance are expected to follow the expected results for phonetic recognition: HMM with more Gaussian tend to behave better. The tendency to produce better results with fewer Gaussians in the ranges of small tolerances could be explained by the inherent variance of the spectrum in the vicinity of a phonetic transition [1], which could make a simpler model more adequate. The features without delta features are showed not to follow consistently this trend compared to with delta features. This is because the delta features take consideration of the contextual effect and more Gaussians will provide more maxima in the density function which can model better different acoustic realizations of the same phoneme.

As the different feature set is concerned. Generally, for very small tolerances, incorporation of delta features deteriorates the segmentation performance. For immediate and large tolerances, the use of delta feature increases the accuracy.

D. Implicit Boundary Refinement Result

The HMM based segmentation system described in previous experiment is chosen as baseline system for comparison with segmentation system with implicit boundary refinement (IBR). The training tokens are extended 5ms of its

TABLE II. PERCENTAGE OF BOUNDARIES WITHIN DIFFERENT TOLERANCES (IN MS) WITH VARYING NUMBER OF MIXTURE COMPONENTS USING DIFFERENT FEATURE SET S.

HMM Set	<5	<10	<15	<20	<25
MFCC+E - 1 Gaussian	45.24	62.80	72.14	77.85	82.18
(MFCC+E)+D - 1 Gaussian	50.87	66.52	73.36	76.99	80.97
(MFCC+E)+D+DD - 1 Gaussian	46.19	66.78	74.22	78.37	82.70
MFCC+E - 2 Gaussian	46.63	64.01	73.36	78.98	83.30
(MFCC+E)+D - 2 Gaussian	43.43	64.45	75.35	80.62	84.86
(MFCC+E)+D+DD - 2 Gaussian	42.39	64.79	74.83	80.54	84.34
MFCC+E - 4 Gaussian	47.15	64.01	72.92	78.29	82.70
(MFCC+E)+D - 4 Gaussian	37.54	60.47	73.18	79.84	84.17
(MFCC+E)+D+DD - 4 Gaussian	37.89	62.02	73.62	80.54	84.86

TABLE III. PERFORMANCE COMPARISON BETWEEN BASELINE SEGMENTATION AND WITH IBR WITHIN DIFFERENT TOLERANCES (MS).

Segmentation system	<5	<10	<15	<20	<25
MFCC+E - Baseline (BL)	46.63	64.01	73.36	78.98	83.30
(MFCC+E)+D - Baseline (BL)	43.43	64.45	75.35	80.62	84.86
(MFCC+E)+D+DD - Baseline (BL)	42.39	64.79	74.82	80.54	84.34
MFCC+E - BL with IBR	48.10	65.31	73.88	79.41	83.56
(MFCC+E)+D - BL with IBR	50.26	68.08	76.21	80.28	85.03
(MFCC+E)+D+DD - BL with IBR	47.75	67.73	76.38	81.75	85.55

both end manual-labeled boundary to the adjacent phone, which corresponding to one feature frame extension. The models trained from these tokens used in Viterbi alignment. The Number of Gaussians is fixed to 2. The comparison results for different feature set are shown in Table 3.

The result shows improvement by using IBR compared to the baseline for all range of tolerances and different feature set. It can also be seen that improvement produced by IBR tend to be more important in the zone of small tolerances. This means that IBR is capable of increasing the precision of segmentation (reducing small errors). There is a slight increase of performance in range of large tolerances. This can be can be explained by the improved modeling of contextual effect of context-independent HMM trained from the extended tokens improve the recognition accuracy.

V. CONCLUSION AND FUTURE WORK

In this study, automatic Malay phone segmentation is described. This provides the basis for preparing segmented speech database for Malay TTS. HMM based approach using Viterbi alignment is used for the segmentation. This paper also proposed an implicit boundary refinement method for auto-segmentation tasks. The refinement ability of IBR is due to increasing capability of HMM in finely modeling boundaries which trained from extended tokens. The method is simple and save computational time compared to the conventional post-refining method. The IBR is shown to be able to increase the precision of the segmentation without increase or even decrease the gross segmentation errors. This approach improves increase the performance of baseline HMM segmentation from 42.39% to 47.75% in 5ms tolerances. Future work will combine the IBR with those post-refining method [1,4] to increase segmentation precision. Context-dependent will be used. The extension of the training tokens will be tested on more frames.

ACKNOWLEDGMENT

This research project is supported by CBE (Central of Biomedical Engineering) at Universiti Teknologi Malaysia and funded by Minister of Higher Education (MOHE), Malaysia under grant “Support Vector Machines (SVM) As a Large Margin Classifier For Machine Learning Algorithm “ Vot 78029.

REFERENCES

[1] D. T. Toledano, A. Hernandez Gomez, and Luis Villarrubia Grande, “Automatic Phone Segmentation,” IEEE Transactions on Speech and Audio Processing, pp. 617-625, 2003.

[2] C. W. Wigthman, and D. T. Talkin, The aligner: Text to speech alignment using Markov models. Progress in Speech Synthesis, Ed: Spinger, 1997.

[3] A. Ljolje and M. D. Riley, “Automatic Segmentation and Labeling of Speech,” in Proc of the ICASSP, pp 473-476, 1991.

[4] D. T. Toledano, “Neural Network Boundary Refining for Automatic Speech Segmentation,” in Proc of the ICASSP, pp 3438-3441, 2000.

[5] S. Cox, R. Brandy, and P. Jackson, “Techniques for accurate automatic annotation of speech waveforms,” in Proc of the International Conference on Spoken Language processing, Sydney, pp. 1947-1950, 1998.

[6] H. Ney., A. Noll., “Phoneme modeling using continuous mixture densities,” Proceedings of the ICASSP, pp. 437-440, 1988.

[7] Nik Safiah Karim, Farid M. Onn, Hashim Haji Musa, and Abdul Hamid Mahmood. Tatabahasa Dewan. New Ed. Dewan Bahasa dan Pustaka, Kuala Lumpur, 1995.