Auto segmentation for Malay Speech Corpus

Tan Tian Swee, Ting Chee Ming, Chin Wee Lip, Lau Chee Yong, Sh-Hussain Salleh

Centre of Biomedical Engineering (CBE), Transportation Research Alliance (TRA), Universiti Teknologi Malaysia, 81310, Skudai, Johor, Malaysia

***Abstract*—Abstract-This paper deals with the automatic segmentation of Malay continuous speech database. Auto segmentation is a process of producing a sequence of discrete utterance with particular characteristics remaining constant within each one. In terms of quality, hand crafted segmentation would be the best method. However, due to the large database size, manual speech segmentation and labeling become tremendous. It is time consuming and error prone. Besides, even if the database is segmented by an expert, the segmentation rule may become subjective and not reproducible. Inconsistency result may occur from different linguistic experts. Thus, an automated segmentation rule was drawn to consistently segment the large scale database with satisfactory level of quality. Automated segmentation of Malay Language syllable is not a tough task because all syllables in Malay Language are pronounced almost equally and moreover it is not a tonal language like English. The manipulation and identification of the segment boundaries of Malay Language is straight forward and easy to understand. For the segmentation, the HMM based approach with adapted Viterbi force alignment technique is used. Composite HMM with Baum Welch reestimation was utilized to ease the process of phonetic segmentation. All the data from the database was fed into the segmentation tool directly without prior trained sample for pre-training purpose. For the design of the sentence coverage of the database, the scripts are consisting of 1000 sentences. 620 sentences are selected from primary school Malay Language text book and 380 sentences were computed using the 70% highest frequency words that appear in the 10 million words online digital text. This configuration of Malay Language script already promises a phonetically balanced database which covers all the vowels and consonants. The objective evaluation method is used to identify the performance. The result from the auto-segmentation was verified to obtain the accuracy degree and overall quality. The result was tested perceptually and it is proven to have satisfactory high quality.**

I. INTRODUCTION

In the speech technology, auto segmentation is a crucial step to process the database. Before the database can be feed into the system for another processing, the database should be segmented according to the needs of the system. Auto segmentation is a process of producing a sequence of discrete utterance with particular characteristics remaining constant within each one [3]. It is the first step in any tools or systems like morphological analyser, POS tagger, syntactic parser, etc., and applications, like machine translation, information extraction, information retrieval, etc. Finding the boundary of either a sentence or a token is nontrivial [1]. To obtain the best result of segmentation, manually craft the database is the best solution. But hand crafted process is time consuming and expensive [4]. Moreover, due to the large database size, manual speech segmentation and labeling become tremendous and error prone. Besides, even if the database is segmented by an expert, the segmentation rule may become subjective and not reproducible [2]. Inconsistency result may occur from different linguistic experts. Thus, an automated segmentation rule was drawn to consistently segment the large scale database with satisfactory level of quality. Automated segmentation of Malay Language syllable is not a tough task because all syllables in Malay Language are pronounced almost equally [5]. The segmentation of Malay Language has some similarities if compared to English. Firstly, Malay language is a phonetic language and it is also written in Roman characters like English. Secondly, it is not a tonal language because all the syllables in Malay are pronounced almost equally. As a review of Malay Language, there are six (6) main vowels and 29 consonants in standard Malay (SM). SM have a total of nineteen of the consonants, where /m/, /n/, /f/, /l/, /s/ and /y/ are pronounced almost the same way as in English. In Malay language, the syllabic structure is well-defined and can be unambiguously derived from a phone string. The basic syllable structure of the Malay language is generated by an ordered series of three syllabication rules. The linguists claimed that Malay is a Type III language, namely Consonant-Vowel (CV) and Consonant-Vowel- Consonant (CVC) are the most common and they can be found almost in every Malay primary word (Noraini et al. 2008). To perform the segmentation process, hidden Markov model (HMM) is used. HMM is a statistical model which is able to model patterns and sequence. It is widely used in speech technology like recognition and speech synthesis. The basic idea of HMM is the transition of states and observations. All the states inside the HMM consist of a mean and variance which is able to release the desire output with high probability after trained. The model can be illustrated as the figure below.

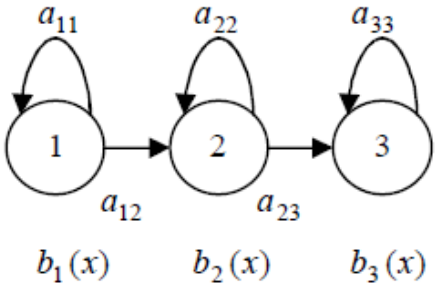

Figure 1. Block diagram of hidden Markov model

Previously proposed idea was adding post refinement to increase the quality of the segmentation result. The post refinement method was called implicit boundary refinement with using Viterbi forced alignment. The method mechanism is to extend the start point and end point of the training data

to the point next to it or its adjacent point. So, each of the training data was embedded with wider range of boundary. That means the training process can be conducted in an easier way and the HMM can be better trained and better model the phonetic boundaries. This post refinement would enable the Viterbi alignment process conducted with better accuracy and less errors. The block diagram of the process can be illustrated in the figure below.

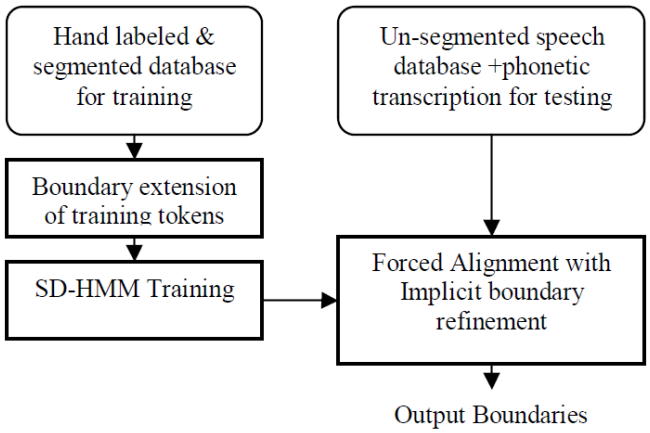

Figure 2. Block diagram of automatic segmentation with implicit boundary refinement

For the implicit boundary refinement method, the starting point and end point of the phoneme were extended until the adjacent point. The solid line represent the original boundary while the dotted line is the extended boundary of the phoneme. The result is shown in the figure below. The problem of this method is the decision to point the boundary. Manual refinement would result in the inconsistency and inaccurate. All the manual refinement is just based on approximation. But the key idea of the refinement is to enable the training process to become easier and reduce error.

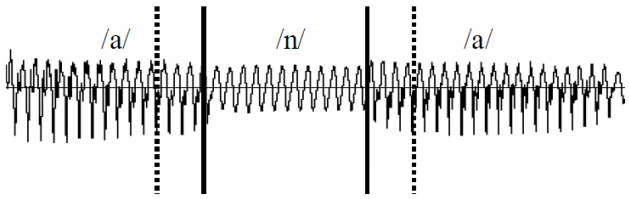

Figure 3. Starting point and end point extension from the original boundary

However, this kind of automatic segmentation requires a pre-manual hand seeded segmented data for the training process. If the language is changed the hand seeded process also has to change. Moreover, manual refinement could result in inconsistency and error prone. So, a fully automated segmentation tool is anticipated to ease the problem. In this project, a fully automated phonetic segmentation is proposed rather than using post refinement method.

II. Methodology

This paper introduces the study of fully automated segmentation of a Malay language database. The theory used to perform the auto segmentation is the composite HMM trained by Baum Welch reestimation and implement the Viterbi alignment on all samples. The advantage of this method is fully automated and also able to obtain satisfactory level of quality. The segmentation system used in this study consists of two phases: Initialization and Iteration. The block diagram of the process is shown below.

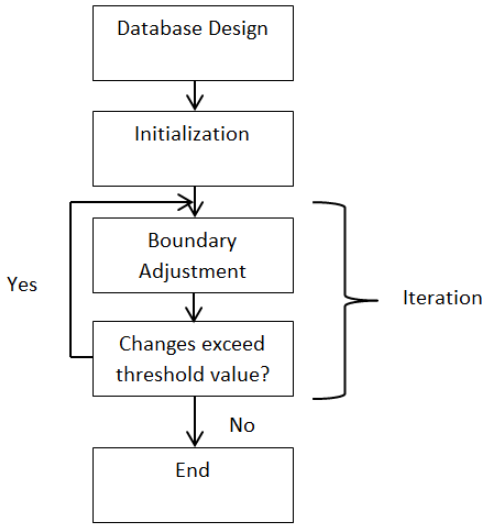

Figure 4. Block diagram of auto segmentation

For the initialization phase, all the data from the Malay Language database will go through the feature extraction process. A normal speech can be model using 39 element feature vector model. The 39 element consist of 13 of static features, 13 delta vector and 13 delta-delta vector. The 13 static features contain 12 MFCC computed from 24 filter banks and log energy. The features extraction can be illustrated as the figure below.

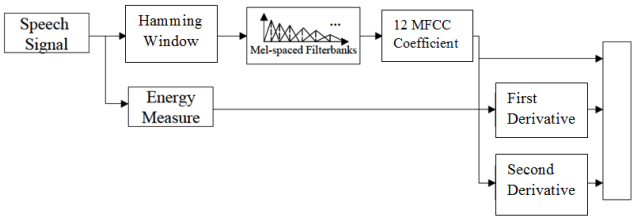

Figure 5. Features extraction process

After that, the HMM was trained using the features extracted from the process above, then the HMM now knows how to segment the rest of the data. All the data were fed into the system to perform the initialization. The initialization means to equally distribute the frame number for all the phoneme by

dividing the total frame to the total number of phoneme in the sentence. This is the first entry step for every data so that the iteration can be conducted based on this initialization result. After that, the iteration takes place to adjust the boundary of the phoneme according to the clusters. The first iteration is based on the initialization result and slowly move the boundary according to the frame characteristic. The iterations after that were proceed if the changes of the boundary exceed the stopping threshold of the segmentation process. The process can be illustrated in figure below.

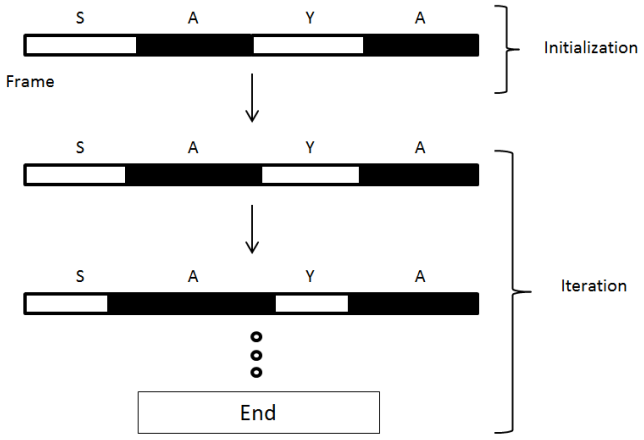

Figure 6. Iteration process of auto segmentation

After the iteration stops, that means the auto segmentation process is done and the result is ready.

A. Database Design

For the design of the sentence coverage of the database, the scripts are consisting of 1000 sentences. 620 sentences are all the words in primary school Malay Language text book and 380 sentences are the 70% highest frequency words that appear in the 10 million words online digital text. This configuration of Malay Language script already promises a phonetically balanced database which covers all the vowels and consonants.In Malay language, there are 24 pure phonemes and 6 borrowed phonemes, divided into 8 categories. Among the pure phonemes, there are 18 consonants and 6 vowels. The borrowed consonantal phonemes are /f, z, sy, kh, gh, v/. Five diphthongs can be found in Malay language which are /ai/, /au/, /oi/, /ua/, /ia/. This Malay phone set cover all Malay phoneme unit in Malay language. The segmentation experiment is based on the 35 phones above and a silent model for pausing /pau/. The /gh/ was folded to /g/ due to limited training tokens. All the sentences have been hand labeled and segmented according to the chosen Malay phone set in Tabel 1.

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/

Table I
LIST OF MALAY PHONES ACCORDING TO CATEGORIES

III. RESULT

The auto segmentation process was taken place to segment all the Malay sentences. The result shows the consistency of the segmentation output which can avoid the inconsistency due to manual segmentation. The result was verified according to its waveform and the result is shown in the figure below.

Figure 7. Result of auto segmentation

This method is purely automated and no manually hand crafted input needed. Once the HMM was trained using the features extracted from the same database, then the HMM knows how to segment the waveform. The result shows the waveform was segmented according to its phoneme. The perceptual evaluation is carried through by selecting the phoneme segment and playback to verify whether the segment is correct or not. The data was randomly picked and listen as an evaluation. The result shows that the segment of all the data was correct according to its phoneme. So can be concluded that the segmentation tool is reliable. The result was also compared with the auto segmentation with implicit boundary refinement perceptually. For the segmentation with implicit boundary refinement, the phonetic segmentation was first carried out without the implicit boundary refinement and the result obtained was regard as the standard segmentation result to be compared to another method. The phonetic segmentation with implicit boundary refinement also carried out and the result was compared to the previous result without implicit boundary refinement. The segmentation result was improved by implicit boundary refinement and the post refinement is more accurate in the zone of small tolerances. It can be said that the implicit boundary method is capable to increase the precision of segmentation. However, the comparison between the method proposed in this paper and the method with post refinement, the quality and precision of both of the result shows no much difference and changes using the subjective evaluation. So can be concluded that the method proposed in this paper is capable to obtain the same result with the auto segmentation with implicit boundary refinement. However the

main advantages of this method is it is purely automated without hand crafted data as training data.

IV. Conclusion

In this study, automatic Malay Language segmentation is described. This provides the basis for preparing segmented speech database for Malay TTS. Composite HMM based with Baum Welch reestimation approach using Viterbi alignment is used for the segmentation. From the obtained result, it can be proven that the auto segmentation tool is able to get the satisfactory level of result. The syllable based segmentation can perform as good as the segmentation with implicit boundary refinement. For the future work, the testing can be conducted with more feature set and different number of Gaussian number to verify the reliability of the system.

V. Acknowledgement

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 GUP grant Tier 1 vot number 01H49 with the project title Development of Malay Speech Technology for Preschool Vocabulary Learning System.

References

[1] Bali Ranaivo-Malancon (2011). Building a Rule-based Malay Text Segmentation Tool. International Conference on Asian Language Processing.

[2] F. Brugnara, D. Falavigna and M. Omologo (1993). Automatic segmentation and labeling of speech based on Hidden Markov Models. Speech Communication 12 (1993) 357-370

[3] Matthew A. Siegler, Uday Jain, Bhiksha Raj, Richard M. (1997). Automatic Segmentation, Classification and Clustering of Broadcast News Audio. Proc. DARPA Speech Recognition Workshop page 97-99

[4] Noraini Seman, Kamaruzaman Jusoff (2008). Automatic Segmentation and Labeling for Spontaneous Standard Malay Speech Recognition. International Conference on Advanced Computer Theory and Engineering.

[5] Zaharah Othman and Sutanto Atmosumarto (1995). Colloquial Malay: a complete language course for beginners.