A Study on Automatic Speech Segmentation Method Based on Human Perception Characteristics

Keyi Zhou (1110204)

School of Information Science,
Japan Advanced Institute of Science and Technology

August 8, 2013

Keywords: automatic speech segmentation, human perception characteristics, spectrum target prediction model, errors list file.

Nowadays, there have been more and more applications with the requirement for highly accurate and reliable speech segmentation. Given a specific example, based on the unit-selection technique, concatenative speech synthesis has become an essential approach to text-to-speech (TTS) systems. The quality of synthesized speech by this technique is dependent on the quality of the speech corpus, to a large extent.

Traditionally, manual segmentation has been considered the most reliable and precise method to get the segments. However, it will be time-consuming and labor-intensive, especially when the required size of the speech database is huge. Therefore, an appropriate automatic method for segmentation is more feasible and practical.

Some methods for automatic speech segmentation have been proposed based on Hidden Markov Model (HMM). For segmentation, these methods divide continuous speech into segments (e.g. phoneme) based on the variety of pattern distance (e.g. likelihood in HMM). Although these methods are certified useful to train model parameters utilized for speech recognition and automatic speech segmentation can be implemented very efficiently using the Baum-Welch algorithm and Viterbi algorithm, it is suspected for highly natural speech synthesis or some other application with the requirement for highly accurate and reliable speech segmentation, by using

automatic segmented speech databases. The reason is that segmentation at points-in-time when output of the HMM changes to a following phoneme, according to Maximum Likelihood (ML) model, makes imprecise boundaries for the segments compared to professional manual segmentation.

Some methods have been proposed to solve the problem mentioned above, by refining the initial HMM-based segmentations and Some research has achieved great improvement for standard HMM-based segmentation. However, considering the principle of the HMM-based segmentation, there has been some problems remained to be solved for further research. One of the problems is that HMM is a probability model without considering human perception characteristics. Based on the human perception characteristics, it may be helpful to obtain more accurate boundaries for the segments.

Referring to target prediction model simulating the human speech perception mechanism, especially a compensating mechanism which presumably exists in the speech perception mechanism, the purpose of this research is to propose a speech segmentation method based on human perception characteristics. The model predicts the stable spectral target in each short-term interval and compensates for phonemic characteristics from coarticulation. By choosing segmentation points-in-time when the estimated phoneme target changes in human perception, it is possible to solve the problem that segmentation based on HMM makes mismatching to human perception characteristics.

To achieve the purpose, there are several sub-goals for this research. First of all, the candidates for the precise boundaries should be calculated out by the spectrum target prediction model. How to extract the boundaries from the candidates is the next sub-goals. Afterwards, based on the strategies for extracting the phoneme boundaries, the schemes for labelling methods should be constructed. The last, there should be evaluation for the proposed method compared to the HMM-based method.

To achieve the first sub-goal, simulation experiments for the spectrum target prediction model were carried out firstly, then experiments for segmentation using ATR database were implemented to verify the effectiveness of the model for the purpose of this research. By calculating the Euclidean distances of two adjacent targets, some candidates for extracting the phoneme boundaries can reveal. Compared to manual labelling, around

the hand-made segmentation boundaries in human perception, there are always peaks of Euclidean distances of the adjacent targets to be selected as the automatic segmentation boundaries. However, because the spectrum target prediction model is highly sensitive to the variation of targets, there are too much peaks whose values are small, making difficulty for determining the boundaries.

Comparing advantages and disadvantages of HMM and the spectrum target prediction model, for the HMM-based segmentation and labelling, the number of boundaries can be determined, however, the boundaries are not accurate enough compared to hand-made boundaries, on the opposite, there are always candidates close to the hand-made boundaries by the spectrum target prediction model, however, there have been no suitable rules for choosing these accurate automatic segmentation boundaries from a number of candidates directly. Thus, to refine the HMM-based boundaries to the accurate boundaries obtained by the spectrum target prediction model is a possible strategy for extracting the phoneme boundaries from all candidates. Besides, according to the results of experiments for the spectrum target prediction model, around the hand-made segmentation boundaries in human perception, there are always candidates to be selected as the automatic segmentation boundaries. Thus, if the estimation for the boundaries in human perception is feasible using the existing labelling files by the HMM-based method, it is foreseeable to obtain more precise boundaries compared to HMM-based segmentation results.

To implement this strategy, the labelling files obtained by the HMM-based method are needed, and one more file called "errors list file" is required. The "errors list file" records the mean error, maximum error, and minimum error for each boundary, as well as the phoneme before the boundary and the phoneme after the boundary by statistically calculating the errors between all automatic segmentation boundaries by the HMM-based method and the corresponding manual boundaries. When a HMM-based boundary is provided to the proposed method, firstly the mean error should be found for this boundary by checking in the errors list file according to the former phoneme and the later phoneme, and then the absolute value of the mean error will be judged if it is less than 20 milliseconds. If the answer is Yes, a estimated manual boundary is determined by moving

the HMM-based boundaries according to the found error and the nearest peak will be selected to be the boundary by the proposed method. Otherwise, a range $[hmm\_boundarymax\_error, hmm\_boundarymin\_error]$ for estimating manual boundary will be used for selecting the peak with biggest value as the automatic segmentation boundary, where $hmm\_boundary$ is the HMM-based boundary, meanwhile, max_error and min error are the maximum error and the $minimum\_error$ for this boundary compared to the manual boundary checked from the errors list file. After the phoneme boundaries by the proposed method are obtained using the speech signal and the corresponding HMM-based labelling file, the labelling procedure can be treated as the adjustment for HMM-based boundaries to achieve the third sub-goal.

Experiments were carried out using TIMIT database. Considering the language of TIMIT database is English, the Hidden Markov Model Toolkit (HTK) is used to obtain rough automatic labelling files. Then, the "errors list file" is obtained by the HMM-based labelling files and the corresponding manual labelling files of the whole training subset. Finally, the proposed method is utilized to get automatic labelling files with accurate boundaries. The obtained labelling files by the proposed method indicate the effectiveness of the proposed method.

According to analysis for the results by the HMM-based segmentation method, an objective evaluation method is designed. Then, the theoretical upper-limitation evaluation results are shown, assuming that manual labelling files in TIMIT database could be used as references for determining the boundaries from all candidates obtained by spectrum target prediction model and the nearest peaks from the manual labelling files are always selected as the boundaries. The next, the objective evaluation will be carried out for the segmentation results by the proposed method, compared to the results by the HMM-based segmentation method. According to the evaluation results, the proposed method improve the accuracy of segmentation boundaries compared to the standard HMM-based segmentation method, which indicates the achievement of the purpose of the proposed method.

However, there have still be plenty of work for further research, such as the subjective evaluation and the comparison to other automatic segmentation methods. Especially the errors list file, used to estimate the boundaries

in human perception, is the key factor to improve the accurate of phoneme boundaries compared to the HMM-based method. Thus, to prove the possibility to obtain a reliable commonly-used errors list file for all possible boundaries is a essential and hard task for extending the proposed method to all databases. Besides, compared to the theoretical upper-limitation evaluation results, how to optimizing the proposed method is also the future work to achieve higher performance in the domain of automatic speech segmentation.