Chapter 7. Conclusions and Future Work

In this study, we have introduced the proposed approaches to automatic phonetic segmentation for speech/singing voice synthesis and addressed the basic difference between speech and singing voices. As most of the studies on automatic phonetic segmentation, we employ HMM with MFCCs for the forced alignment of speech data. On the other hand, for singing voice data, we adopt both HMM with MFCCs and DTW with pitch contours.

Our experimental results indicated that the HMM-based alignment slightly outperformed the DTW-based alignment if we only consider the cases with larger segmentation errors (ex. error range > 30 ms). Nevertheless, either the HMM-based alignment or the DTW-based alignment has satisfactory performance in the phonetic segmentation. In view of this, we have developed a post-processing scheme to refine the initial estimates obtained by HMM and the DTW. In this study, we have introduced two methods to process the boundary refinement, one is based on a hybrid approach and the other is based on a score predictive model.

The primary framework of the proposed hybrid approach is based on the heuristic rules and statistical pattern recognition. Most of the boundaries are identified via statistical pattern recognition, while the most difficult cases (phone transitions with strong co-articulation) are handled via heuristic rules. For the statistics-based method, we applied KNNR as the classifier and LOO as the performance criterion. In order to determine the most influential features for each phonetic transition category, we employed SFS for feature selection. For the heuristic method, we utilized two acoustic features, formants and log energy, to perform an effective boundary refinement. In our experiments, we used the SKL (SFS+KNNR+LOO) boundary refinement to refine most of boundaries and used the

heuristic method to refine the other boundaries (“FINAL + first INITIAL type categories”). The experimental results indicated that the hybrid approach can improve boundary refinement. However, the hybrid approach has two drawbacks:

1) It is unnatural to have a binary decision for crisp classification.
2) A fixed search range used in the boundary refinement is inappropriate.

We thus used soft classification based on the concept of the score predictive model (SPM) to substitute the previous hybrid approach. The principal advantage of the proposed SPM is its capability to predict the scores of candidate boundaries reliably according to the 58-dimensional feature vectors. In the process of boundary refinement, the scores of the initial boundaries (identified by HMM or HMM+DTW) are computed via the SPM. A dynamic search range is designed to determine the suitable range of candidate boundaries. Finally, a boundary with the highest score will be selected as the refined boundary. Experimental results indicate that the performance of SPM is better than that of the previous hybrid approach.

In the future, we will attempt several possible directions to improve the performance of automatic phonetic segmentation. For example, other acoustic features can be included in the proposed framework for better segmentation. Moreover, more effective feature selection schemes, such as sequential floating selection method or stepwise regression, can be investigated to construct a more accurate regression model for SPM.