Abstract

This study introduces a framework for effective phone-level segmentation for Mandarin speech and singing voice corpora. To perform initial phonetic segmentation, we employ hidden Markov models (HMM) for the forced alignment of speech data. On the other hand, for singing voice data, we adopt both HMM and DTW (dynamic time warping). Since the initial estimates are usually inaccurate, we need to perform boundary refinement to improve the segmentation accuracies. In this dissertation, we proposed two methods to refine the initial boundaries, ones is based on a hybrid approach and the other is based on a score predictive model. The boundary refinement based on a hybrid approach combines the statistical pattern recognition and heuristic rules. 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. However, it suffers from two drawbacks, namely, unsuitable binary decision for crisp classification and a fixed search range in the boundary refinement. In view of this, we propose the concept of score predictive model (SPM) instead. Under the framework of SPM, we can predict the scores of candidate boundaries effectively with a set of acoustic features. The optimum boundary with the highest score can be chosen accordingly. Several experiments are designed to verify the feasibility of the proposed SPM. The experimental results indicate that the proposed SPM method outperforms the hybrid approach. Finally, these identified boundaries of speech/singing voice corpora can then be used for corpus-based speech/singing voice synthesis.