year month day
2014 1 14

Department	Electronic and Information Engineering	ID	109302
Name	Narpendyah Wisjnu Ariwardhani		

Supervisor	Junsei Horikawa Kouichi Katsurada

Abstract

Title	A Study on Articulatory Feature-based Phoneme Recognition and Voice Conversion

(800 words)

In this thesis, the behavior of articulatory feature (AF) as linguistic feature representation of the speech waveform in the task of both phoneme recognition (PR) and voice conversion (VC) is studied. Over the past few years, several studies have been conducted on the design of an optimal hidden Markov model (HMM) configuration for automatic speech recognition (ASR). Most of these studies are based on spectral-representation feature vectors. On the other hand, phonetic features, such as articulatory features (AF), have proved their robustness across speakers, against co-articulatory effects, and against noise. Despite these advantages, the literature on the design of an optimal parameter set for AF-based HMM speech recognition is still limited. Subsequent to our previous works of an AF extractor, the first part of this thesis will describe further our experimental studies on the design of an optimal AF-HMM-based classifier.

In the beginning of the thesis, while we also intend to improve the phoneme recognizer performance, the main goal is rather to observe the behavior of AF as the speech representation for PR task. Several strategies for designing the optimal parameter set in AF-HMM-based PR are investigated. These strategies will consider sub-word units extension, number of HMM states addition, vowel group separation, insertion penalty tuning, and HMM topologies selection. Mel-frequency cepstral coefficient (MFCC)-HMM-based PR experiments were also conducted for comparison purpose.

Besides accuracy improvement along the experiments, the analysis showed different behavior between AF-HMM-based PR and MFCC-HMM-based PR in terms of their reaction to insertion penalty (IP) value. Both of the PR systems experienced accuracy degradation during the extension from monophone-based PR to triphone-based PR. This accuracy degradation, which showed that a large insertion error occurred during the process, can be overcome by imposing IP. While MFCC-HMM-based PR only needed IP value around -30 to reach its optimal performance, AF-HMM-based PR needed larger IP value (around -100). This behavior comes from the characteristic of AF that has smaller variance than MFCC, and therefore has more positive likelihood to be suppressed.

For both AF-HMM-based PR and MFCC-HMM-based PR, by tuning insertion penalty and extending monophones to triphones, the phoneme recognition performance (for both accuracy and correct rate) improved. The proposed AF-HMM-based PR with 5-state HMMs, separated vowel, triphone subword, Bakis topology, and optimal insertion penalty provides the best accuracy among the experiments, i.e., 81.38% for the JNAS database.

Furthermore in this thesis, the behavior of AF is also used to realize AF-based VC system. We focus our goal of this section to implement AF-based VC for a small number of target-speaker training data. VC is one of the important technologies in the field of speech processing. VC transforms the voice from the source-speaker onto the target-speaker.

When a source-speaker utters a certain sentence, the converted speech will sound as if a target-speaker is speaking the same sentence. The trend of VC has moved from text-dependent VC, in which it needs parallel utterances between source and target-speakers, into text-independent VC. However, this newer system still needs source speaker utterances as the training data.

The flexibility of AF as speaker independent representation, as showed in PR task, can be used to extend the capability of an AF-based VC application. AF can be used in speaker adaptation technique to develop a VC application which maps features from arbitrary speakers into those of the expected target speakers. During the training process, our approach does not require source-speaker data to build the VC model.

We propose VC based on AF to vocal-tract parameters (VTP) mapping. An artificial neural network (ANN) is applied to map AF to VTP and to convert a speaker's voice to a target-speaker's voice. In order to investigate the effect of ANN architecture and different VTP orders on the performance of AF-ANN-based VC, six ANN architectures correspond to different VTP orders were compared. The architecture that provided the best result compared with other architectures was chosen for the remaining experiments. In addition to the feature vector mapping process, two types of F0 conversions were also conducted. The first F0 conversion was done using time stretching subsequent to sample rate transposing technique. Moreover, the second F0 conversion was done using F0 extraction and re-synthesis technique using MLSA filter.

For comparison, a baseline VC system based on Gaussian mixture model (GMM) approach was conducted. Two types of evaluations were performed, i.e., objective evaluations and subjective evaluations. For objective evaluation, spectrum distortion (SD) is calculated to measure the distance between target-speaker spectrum and converted spectrum. Furthermore, for subjective evaluations, three listening tests were performed, i.e. the similarity test, XAB test, and mean opinion score (MOS) test. For the overall performance, AF-ANN-based VC outperforms MCEP-GMM-based VC for a small number of target-speaker training data. The proposed VC application was also realized for arbitrary source-speakers.