Phoneme classification tree based on multi resolution analysis and appling to speech sound super-resolution

T14M003 Yuki Sugii
Adviser Qing Ma
Graduate Course of Applied Mathmatics and Informatics
Graduate School of Science and Technology
Ryukoku University

Abstract

HMM and statistical techniqus are popular in speech recognition. These methods are effective in recognition in noiseless environments and with well-learned dataset. Otherwise these are not in noisy environments and withnovel dataset. Futermore, recognition rate of HMM and the statistical techniqus deteriorate with limited learning data. We propose a method that classify the individual waveform of a Japanese phoneme into 50 Japanese syllabaries or into list of syllabary candidates.

It is extremely difficult to classify the phonemes to any one of 50 syllabaries, because the phoneme has various frequency components. In this study, we propose a classification tree for phonemes which is made with the multiple resolution analysis and PCA. An individual phoneme can be represented in L-D spaces which are depend on a frequency level L of multiple resolution analysis and a reduced dimension D by PCA. Our phoneme classification tree are generated by deciding on a most separatable L-D space for each phoneme.

In this article, we have compared our phoneme classification tree with SVM and MRA template matching on Japanese syllabaries classification. We have evaluated the identification by the number of the correctly classified phonemes. There is not much difference in the number of the correctly among each of the method. However, the correctly identification rate in the syllabary candicates with our phoneme classification tree at depth 1 gives about 40%. We have studied about the correctly identification rate of the phoneme classification tree under the restrict to frequency level of the multi resolution analysis. This restriction is equivalent to the decrease in sampling frequency of the input phoneme waveform. The correctly identification rate with 1000Hz sampling frequency which beyond human recognizable as a syllabary is equals to or greater than with 16000Hz.

Proposed phoneme classification tree is also effective in identification of downsampled phoneme waveforms. “Speech sound super-resolution” which interpolates appropriate high frequency sound components into a deteriorated phoneme using our phoneme classification tree is proposed in this article. We have tried to interpolate the high frequency component obtained our phoneme classification tree of speaker A into the downsampled phoneme waveform of another speaker B, that is our speech sound super-resolution method. The results have been compared with the frame interpolation methods(NN, Linear), and with subjective quality evaluation. On the classical frame interpolation mathods, unnecessary high frequency components has been added and any syllabaries have not been able to recognized. On our speech sound super-resolution, plausible components has been added and we can recognized a syllabary. However, individual characteristics of the speakers cloud not be reconstructed with our speech sound super-resolution interpolation.