A study on speech recognition based on sound segregation under noisy environments

HANIU Atushi (210070)

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

Febrary 13, 2004

Keywords: validity of segregation process , validity of segregation result, speech recognition, sound segregation, knowledge, auditory scene analysis, two waves separation model.

Under noisy environments, target sound and background sound exist in time and frequency domain where it is impossible to predict how to overlap there components. Speech recognition system under noisy environments must recognize perfectly what the contents of the target sound are. In order to recognize speech sound under noisy environments, there are two major approarch. One is incorporating preprocessor. The other is deforming acoustic model. However there are no speech recognition system which recognize speech sound well under noisy environments.

When using an adequate segregation method that segregates target sound from mixing sound, if such segregation process using knowledge source is valid for segregating target sound and such segregation result seems to be the target sound, it is considered that the target sound exactly exist in the input sound. Because the origin of the segregated sound is not possible other than the input sound. From this idea, if a system can evaluate validity of process and result in the segregation method, it is considered that the system can recognize target sound in mixing sound.

Thus, in this paper, the method of speech recognition based on sound segregation with evaluating validity of segregation process and results is proposed and then, the proposed method is tested by simulation.

Framework of the proposed model is based on the segregation model proposed by Kubo et al. Kubo et al. have extended the ASA(Auditory Scene Analysis)-based model proposed by Unoki and Akagi to enable to segregate target sound from mixing sound using the knowledge of the target sound. The proposed model is composed in five blocks: the signal analyzer, found mental frequency (F0) estimation, knowledge manager, segregation block and recognition part. In the signal analyzer, a filterbank decomposes the input sound into complex spectra. The F0 estimation block determines the candidates for the F0 and calculate harmonic frequency. The segregation block segregates target sound using harmonicity, knowledge, Bregman's regularities and correlation between knowledge and segregated sound in frequency domain. Segregation process is done by adjusting time with evaluating validity of segregation process when the system segregates target sound from mixing sound. The knowledge manager receives a phonemic transcription that represents the target sound, and selects a suitable spectrum forms of target sound in the frequency domain from knowledge sources. The recognition part judges whether the estimated F0 is similar with the F0 of speaking sound or not. The recognition part monitors process of segregation block, and it judges whether process of segregation block is reasonable. Correlation is calculated between spectrum in frequency domain and knowledge, and the average correlation is used to judge validity of results. Finally, recognition part outputs the results of recognition as symbol(s).

Simulations were carried out to evaluation the segregation block and recognition part. The condition of simulation were one speaker uttered Japanese isolated vowel with Gaussian white noise and the SNR of -20, -10, 0, 10, 20 dB. In order to compare performance between recognition system using proposed method and DTW (Dynamic Time Wrapping), speech recognition system using DTW was used. As results of simulation, it is shown that the proposed method recognize target sound correctly in all condition. It was also shown that the proposed model always got more certain results of recognition than the system using DTW. Another simulation were carried out to evaluate the implemented system. The condition is the same as first simulation. As a result, the implemented system recognized target sound correctly in the SNR of 0, 10, and 20 dB. In the SNR

of -10 dB, the system failed in recognition because F0 estimation block can not estimate valid F0 as human speech. This failure indicates that the F0 estimation block should employ robust method of estimating F0 under noisy environments instead of method using autocorrelation. Many method have been proposed to estimate robustly F0 under noisy environments. The performance of the re-implemented model based on concept of this paper is improved by incorporating robust method of F0 estimation under noisy environments.

These simulations show that the proposed model is effective in a situation of uttered isolated vowel mixing with white noise.

In our future work, there are incorporating robust method of F0 estimation under noisy environments, examining knowledge source and examining segregation process and results.