A study on speech recognition in noise conditions using non-negative matrix factorization

Du Yuxuan (1210033)

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

February 12, 2014

Keywords: Noise conditions, Speech recognition, Auditory scene analysis, Non-negative matrix factorization.

Speech recognition is an important topic in speech processing science. Speech recognition plays a very important role in such as automatic telephone answering, car navigation with speech recognition, speech-to-text areas.Speech recognition methods have been studied since 1950s. In real life, with help of speech recognition, it is possible to make us to easily control a computer using our speech in the situation that we can not use our hands. Until now, speech recognition system can abtain almost 100% correct recognition rate with clean speech input. However, noise is always existing in our speech in real life. If noise is mixed in the input of speech recognition system, the recognition rate degrades significantly. According to this reason, it is hard to use a speech recognition system in noisy conditions.

To solve the problem that speech recognition is easily effected by noise, many methods have been proposed till now, which are mainly divided in 2 types. One is preprocessing before speech recognition such as noise reduction. Another is adapting acoustic model. However, neither of them solved the problem completely. On the other hand, human-beings have ability to perceive speech even through noise is mixed with speech. Having regarded to this ability, Haniu et al. proposed a speech recognition method based on selective sound segregation. In this model, it is assumed one target sound is included in noisy input. According to this assumption, knowledge of the target sound is used to segregate noise and target sound. If it is valid to segregate noise and target sound using the information from assumption, this confirms that target sound in the assumption is included in the input. Haniu's model is a speech recognition model that recognize target sounds by verifying the validity of the segregation using assumption. By using the concept of acoustic processing of human-being, this model showed robustness to noise. However, this model has a high computational cost because the segregation process of the model used a complex selective sound segregation model. Therefore, this method is hard to be used in real conditions.

Having regarded the validity of the concept of acoustic processing of human-being, in this paper, we propose a method to realize the concept by using new tools.

To construct the speech recognition model, Modified Restricted Temporal Decomposition (MRTD) method was used to synthesize assumed templates for recognition and Non-negative Matrix Factorization (NMF) was used to separate noise and target sound. NMF has advantage for computational cost comparing with Haniu's method.

Noise types mixed with speech were set as white, pink, babble, factory noise and the SNRs were set as 0 dB, 10 dB, 20 dB and ∞. One Japanese speaker uttered 100 4-mora Japanese words, and the words were mixed with the noise in each condition as inputs. Speech recognition experiments were carried out in each noise condition. As a well-known template based recognition method, recognition experiments based Dynamic Time Warp (DTW) method were carried out in same noise conditions as a reference method to compare with the proposed method. Experimental results showed the proposed method achieved recognition rates of 80% in 0 dB, that is about 50% higher than that of the method based on DTW. Furthermore, processing time of speech recognition model based on proposed method was much shorter than the method of Haniu et al.

In summary, to solve the problem that speech recognition is easily effected by noise, a speech recognition model is proposed in this paper. This model is expected having robustness to noise because it is based on the concept of acoustic processing of human-being. And this model is constructed by MRTD and NMF method. During recognition experiments, recognition method based on proposed method achieved much higher recognition rates in noisy conditions comparing with that based on DTW method. It confirms that, the proposed method has robustness to noise as expected. Furthermore, the proposed method can reduce processing time significantly comparing with the method of Haniu et al.