令和2年度 知能システム学専攻修士論文要旨

工藤 研究室	氏　　名	Shaoxiang Dang
修士論文題目	Improved Speech Separation Performance from Monaural Mixed Speech Based on Deep Embedding Network	

Audio signals account for a large proportion of information exchanged in human communication. Speech separation (Blind source separation), one of the studies that simulate human auditory function, refers to the separation of utterances in which multiple people are speaking simultaneously, and such a situation includes, for instance, conferences, debates, or doctor-patient consultations. With the rapid advancement of technology, many researchers gave their ideas, and most of those gained tremendous success. This thesis puts forward a new model based on a model called deep clustering (DC). The proposed model can overcome the downsides of DC to gain a better result. DC uses a deep embedding network to embed audio data in the underlying manifold space, and data with similar property gathers tightly in the embedding space. Then the model uses a clustering algorithm because the clustering algorithm can easily separate polarized data. Regarding the learning process, the model is supervised by an ideal affinity matrix constructed of binary masks of annotation data. However, the binary mask gives a bottleneck to the entire system since the same position of bins in masks are assigned to 0 and 1 according to the contribution of individual utterances to mixed spectrogram when using binary masks, while the accurate masks could be any real value between 0 and 1. What is worse, the binary mask often leads to defects in separated spectrograms.

Thus, we propose an extended two-stage version based on the deep embedding network that eliminates the shortcomings of the binary mask using various more accurate masks. We employ DC as our first stage. The first stage's output will be cascaded to the mix feature as the latter stage's input. The representation of the affinity matrix can inherently solve the permutation problem because of the clustering algorithm in decoding of DC. Therefore, we have to conduct a permutation invariant training approach to prevent permutation in the second stage. In this way, the proposed model makes it acceptable to employ other masks such as superior ideal ratio mask and phase-sensitive mask. In order to evaluate the proposed model, we establish four data sets from Japanese Newspaper Article Sentences (JNAS): female/male mixture, female/female and male/male mixture, and random mixture.

As a result, we successfully have conducted the proposed model and acquired better results. It outperforms the original DC model by 1.55dB in signal-to-noise ratio by 4.45dB in source-to-distortion ratio,4.41dB in source-to-distortion ratio improvement, 0.16 in short-time objective intelligibility, and 0.3 in perceptual evaluation of speech quality on average. We also observe that the proposed method can repair the defects in the spectrograms brought in by DC.