10:30 AM - 12:10 PM

[3Rin2-31] Speech extraction from conversation based on image-to-image translation using deep neural networks

Kosuke Takaichi, Yoshio Katagami, Yoshiaki Kurosawa, Kazuya Mera, Toshiyuki Takezawa (1. Graduate School of Information Sciences Hiroshima City University, 2. School of Information Sciences Hiroshima City University)

Keywords: sound source separation, deep neural networks, pix2pix

We aim to separate sound sources by deep neural networks which has been active in recent years. We attempt to extract a certain human voice from usual conversation using the networks. We focus on image-to-image translation: pix2pix. The algorithm of pix2pix bases on purely procedure of the image processing. Therefore, we need an additional procedure, that is, we convert voice to spectrogram once. After that we perform to learn the networks to separate human voice, we especially pay attention to segmentation between the same sex and opposite sex. From this point of view, we conducted two experiments using the sounds overlapped both sexes in this paper. Structure-Similarity (SSIM) index and color map representation were used as evaluation criteria. As a result, we confirmed the good extraction of the female voice from the one synthesized both sexes. However, we did not extract the female voice from same sex. Although we reached the conclusion that the separation did not work well, the generated voice seemed to be played naturally. This is not objective judgment. It is our future work.