20

Extracting essential information from sounds

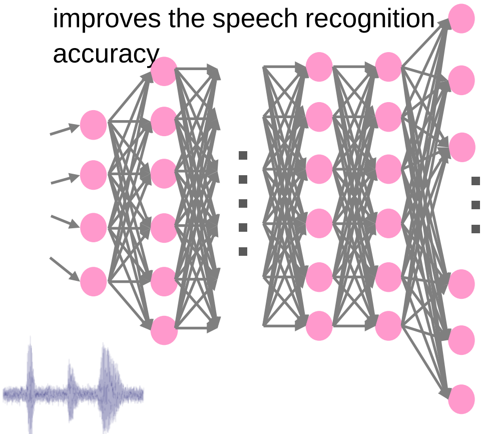
～Advances in distant speech recognition by deep learning～

Abstract

We are working on conversational speech recognition and communication scene analysis in real world sound environments. We have proposed various speech processing methods based on deep learning (DL), which is an essential technique for their realization. In addition to the speech recognition techniques in which DL has been widely employed, we are proposing a variety of DL-based speech processing methods, namely, speech enhancement and acoustic event detection techniques. These DL-based speech processing methods achieve excellent recognition performance for conversational speech. Our DL-based techniques expand the usability of a voice interface in real and noisy daily scenes.

Deep learning (DL)

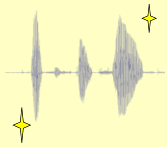
- With "deep" neural networks, DL realizes a high representation ability that can describe complex phenomena.
- DL automatically learns appropriate features and parameters, that improve speech enhancement and/or recognition performance.
- In recent years, DL has attracted a lot of attention, because it greatly improves the speech recognition accuracy



Speech signals recorded in a noisy environment （e.g., conversation, train station, airport…）

Extracting essential information

Speech enhancement



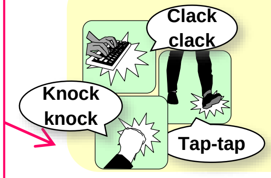
Effectively reduce interference/ reverberant components by automatically learning the parameters for speech enhancement filter design.

Speech recognition

"I, uh, think, we… should…um…ask for da…their recommendation"

Realize high recognition accuracy by discriminating the context of, e.g., spontaneous speech and recording conditions.

Acoustic event detection



Improve recognition accuracy of acoustic events (door, applause etc.) by automatically acquiring the optimal features for event discrimination.

Commercialization

DL-based real-time automatic speech recognition engine has been developed by Media Intelligence Laboratories.

【Application example】
Real-time meeting speech recognition system: RexMeet

Related works

[1] S. Araki, T. Hayashi, M. Delcroix, M. Fujimoto, K. Takeda, T. Nakatani, "Exploring multi-channel features for denoising-autoencoder-based speech enhancement," in *Proc. ICASSP*, 2015.

[2] T. Yoshioka, S. Karita and T. Nakatani, "Far-field speech recognition using CNN-DNN-HMM with convolution in time," in *Proc. ICASSP*, 2015.

[3] D.Q. Truong, S. Nakamura, M. Delcroix, T. Hori, "WFST-based structural classification integrating DNN acoustic features and RNN language features for speech recognition," in *Proc. ICASSP*, 2015.

Contact

Shoko Araki Signal Processing Research Group, Media Information Laboratory

E-mail : araki.shoko(at)lab.ntt.co.jp