Learning tonotopically organized auditory feature-map from speech by an intermap pooling layer in a deep CNN

Hwaran Lee, Geonmin Kim, Ho-Gyeong Kim, Soo-Young Lee
Dept. of Electrical Engineering, KAIST
E-mail: {hwaran.lee, gmkim90, hogyeong, sylee}@kaist.ac.kr

Abstract:

Convolution neural networks (CNNs) originated from the physiological findings of simple cells and complex cells in the primary visual cortex (V1). For the primary auditory cortex (A1), however, researches on how the auditory neurons localized to frequency- and/or time-responses and how they organize topographic structures have not yet finished. With the assumption that A1 and V1 share the underlying mechanisms of the primary cortical receptive fields, we trained the auditory receptive fields with deep CNNs for speech data. Furthermore, in order to model the tonotopic map of the receptive fields, we propose an intermap pooling of convolution filters along time-axis. The layer groups spectrally variant filters with common frequency characteristics and pools the feature maps of filters in each group. Consequently, the adapted filters by the supervised learning forms 1D or 2D topological map in which the tonotopy is disordered. Moreover, the intermap pooling layer makes the CNN to be robust to spectral variability of speech. In our extensive experiments on WSJ corpus, the 9-layer CNN with an intermap pooling layer resulted WER 3.93% on Eval'92 dataset, which is competitive with a state-of-the-art method, even without speaker adaptation techniques.

Keywords:

tonotopic map, intermap pooling layer, convolutional neural networks, acoustic modeling

Acknowledgement

This work was supported by the ICT R& D program of MSIP / IITP. [R0126-15-1117, Core technology development of the spontaneous speech dialogue processing for the language learning]