A generative auditory model embedded neural network for speech processing

Yu-Wen Lo [1], Yih-Liang Shen [1], Yuan-Fu Liao [2], and Tai-Shih Chi [1]

[1] Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan

[2] Department of Electronic Engineering, National Taipei University of Technology, Taipei, Taiwan

INTRODUCTION

- The generative auditory model consists of two stages, the stage of spectrum estimation in the logarithmic-frequency axis by the cochlea and the stage of spectral-temporal analysis in the modulation domain by the auditory cortex.
- The NN is evaluated in a simple speaker identification task. Experiment results show that the auditory model embedded NN is still more robust against noise, especially in low SNR conditions, than the randomly-initialized NN in speaker identification.

THE GENERATIVE AUDITORY MODEL

- **1. The first stage: Cochlear filtering**
  In short, the output of this stage is referred to as the auditory spectrogram, which represents neuron activities along the time and the logF axes. Intuitively, the auditory spectrogram is similar to the magnitude response of the STFT spectrogram presented along the logF axis. The extracted local envelope approximates the magnitude of the STFT spectrogram.

- **2. The second stage: Cortical filtering**
  The second stage models the spectro-temporal selectivity of A1 neurons. Briefly speaking, the auditory spectrogram is further analyzed/ decomposed by A1 neurons which are modeled as two-dimensional filters tuned to different spectro-temporal modulation parameters .

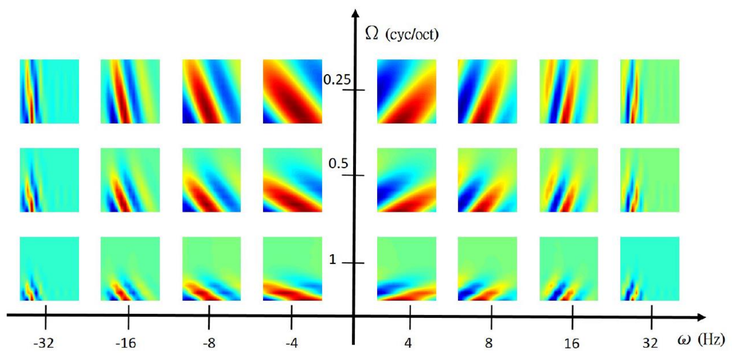

Fig. 1. Spectro-temporal impulse responses of sample modulation filters in the cortical stage.

PROPOSED NEURAL NETWORK

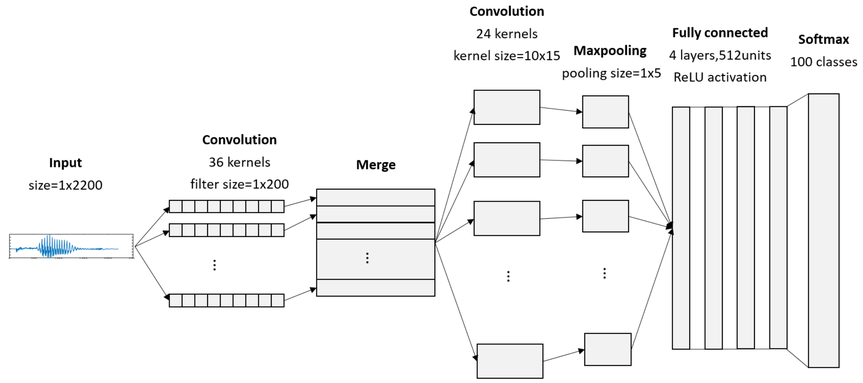

Fig. 2. Architecture of the proposed NN for speech processing on discriminative tasks.

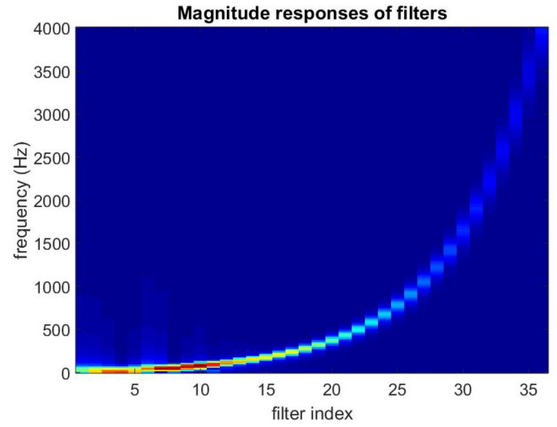

Fig. 3. Magnitude responses of 36 gammatone filters.

- The generative 2-stage auditory model consists of two major operations to decompose speech waveforms: the 1-D cochlear filtering and the 2-D spectro-temporal modulation filtering. Each filtering can be implemented by convolution. Therefore, we construct the NN based on the convolutional neural network (CNN) for discriminative tasks. Fig. 2 shows the proposed NN which includes an input layer, a 1-D convolution layer, a merge layer, a 2-D convolution layer, a pooling layer, and four fully-connected layers. The input to the NN is the time-domain waveform without any pre-processing.

EXPERIMENT RESULTS

Table 2. Speaker identification rates for all test conditions

Scenario	SNR		
	-5 dB	0 dB	5 dB
GammaFix_A1Init	74.75%	81.50%	93.75%
GammaFix_A1Rand	72.50%	80.75%	94.50%
GammaInit_A1Init	73.50%	81.25%	93.75%
GammaInit_A1Rand	67.00%	80.25%	91.75%
BothRand	47.75%	62.50%	83.00%
i-vectors/GMM [10]	39.05%	62.50%	68.69%

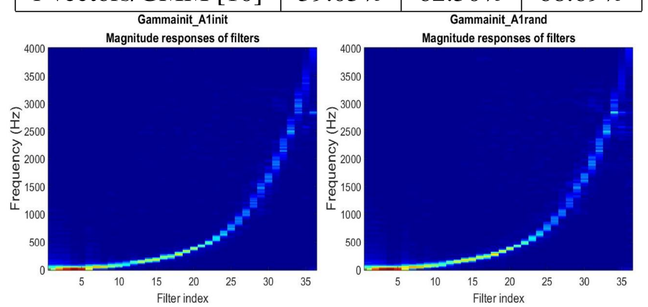

Fig. 4. Magnitude responses of 1-D kernels of GammaInit_A1* methods after training.

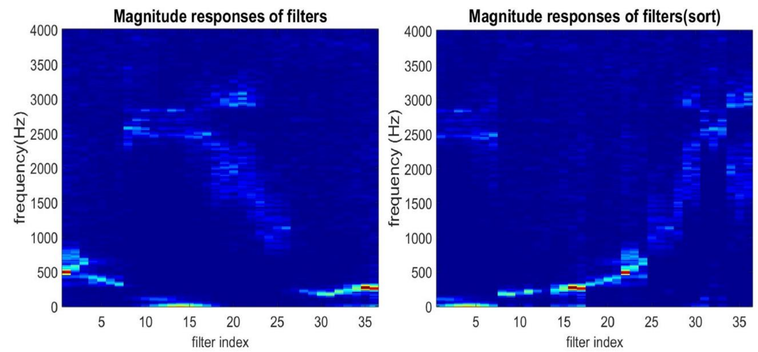

Fig. 5. Magnitude responses of 1-D kernels of BothRand method after training. The left panel shows the original responses and the right panel shows rearranged responses.