# 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

## ABSTRACT

Before the era of the neural network (NN), features extracted from auditory models have been applied to various speech applications and been demonstrated more robust against noise than conventional speech-processing features. What's the role of auditory models in the current NN era? Are they obsolete? To answer this question, we construct a NN with a generative auditory model embedded to process speech signals. 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.

***Index Terms—*** generative auditory model, convolutional neural network, multi-resolution, speaker identification

## 1. INTRODUCTION

During past few years, neural networks (NNs) have been successfully applied to many difficult engineering problems, especially in image processing and speech processing, thanks to their great discriminative power based on backpropagation using piecewise linear units. In speech related applications, variant NNs have been proposed in different topics and brought significant performance improvement over past methods. For instance, the deep NN (DNN) and the recurrent NN (RNN) were used in speech recognition [1][2], speech separation [3], and dereverberation [4][5] tasks with great success. In these NN-based studies, speech signals are processed in the forms of raw data in the time domain or in the short-time Fourier transform (STFT) domain. With the great discriminative power of the NNs, certain representations or features of speech signals seem no longer required. However, this is not the way human process speech. The open question is whether machine perception has to be similar to human perception. This study does not try to force a definite answer to that question but to investigate if human perception can still help in the NN era.

To address human perception of speech, a two-stage auditory model based on neuro-physiological data was proposed in [6]. The first stage mimics the peripheral function of the auditory system to transform the sound into an internal neuron-activity representation. The second stage mimics the function of the auditory cortex (A1) to analyze and decompose the internal representation for further cognitive functions. This model has already been successfully used in many applications, such as in assessing speech intelligibility [7], identifying speaker [8], and separating singing voice from background music [9]. From the functional point of view, the most important function in the first stage is the cochlear filtering which decomposes the sound using a bank of constant-Q filters and produces a 2-D auditory spectrogram in the joint time and logarithmic-frequency (logF) domain. The second stage models A1 as a bank of 2-D spectro-temporal modulation filters which decompose the auditory spectrogram in a 2-D multi-resolution fashion for further analysis [6]. Overall speaking, this auditory model behaves like a generative model using fixed functions (kernels) to decompose representations of speech in the time and the auditory-spectrogram domains.

To investigate the benefits provided by the auditory model to NNs, we construct a NN for speech processing with the auditory model embedded. To evaluate the proposed NN, we conduct simulations in the task of speaker identification. The rest of this paper is organized as follows. In Section 2, we give a brief introduction of the generative auditory model. In Section 3, we describe the proposed NN and test scenarios in simulations. Experiment results are demonstrated in Section 4 with discussions. Lastly, the conclusion is given in Section 5.

## 2. THE GENERATIVE AUDITORY MODEL

### 2.1. The first stage: Cochlear filtering

The first stage consists of several modules to model functions of the peripheral auditory system. The time-domain sound waveform first passes through a bank of constant-Q bandpass

This research is supported by the Ministry of Science and Technology, Taiwan under Grant No MOST 105-2221-E-009-152-MY2.

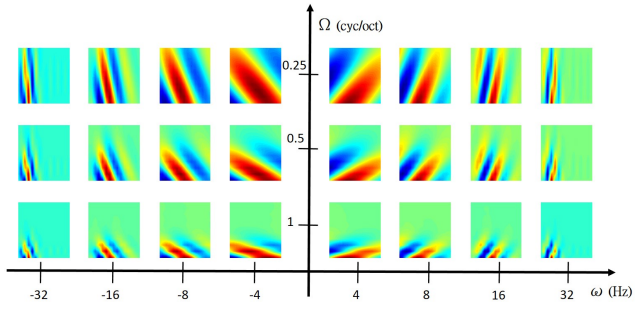


**Fig. 1**. Spectro-temporal impulse responses of sample modulation filters in the cortical stage.

filters, then through a non-linear compression module and a lateral inhibitory network (LIN), and finally through an envelope extractor. The non-linear compression models the saturation caused by inner hair cells, and the LIN models the frequency masking of hearing. More detailed descriptions of the auditory process of this cochlear stage and corresponding mathematical formulations can be accessed in [6].

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.

### 2.2. 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 [6]. The tuning parameters include rate ($\omega$, in Hz), scale ($\Omega$, in cycle/octave), and the directivity of the changing pattern. The rate parameter catches how fast the local envelope of the auditory spectrogram varies along the time axis. For example, the speaking rate of a speaker can be captured by a specific rate parameter. The scale parameter catches how broad the envelope distributed along the logF axis. Therefore, the formant and the harmonic structures of speech can be characterized by the scale parameter. The directivity represents the sweeping direction of the envelope and is encoded in the sign of the rate parameter (negative/positive for upward/downward sweeping direction).

The frequency response of the modulation filter tuned to ($\omega_c$, $\Omega_c$) can be written as:

$$STMF_{+\omega_c,\Omega_c}(\omega,\Omega) = \begin{cases} |F\{h_{rate}(t;\omega_c\} \otimes F\{h_{scale}(x;\Omega_c\}|\,0, \leq \omega; \Omega \leq \pi \\ 0, otherwise \end{cases} \quad (1)$$

$$STMF_{-\omega_c,\Omega_c}(\omega,\Omega) = \begin{cases} |F\{h_{rate}(t;\omega_c\} \otimes F\{h_{scale}(x;\Omega_c\}|\,, -\pi \leq \omega \leq 0; 0 \leq \Omega \leq \pi \\ 0, otherwise \end{cases} \quad (2)$$

where $F$ is the 1-D Fourier transform, $\otimes$ is the outer product, and $x$ means the logF axis. The rate ($\omega$) and the scale ($\Omega$) are respectively the frequency domains of time and logF. The $h_{rate}$ and $h_{scale}$ are the 1-D temporal and spectral impulse responses derived from gammatone filters centered at $\omega_c$ and $\Omega_c$ as:

$$\begin{cases} h_{rate}(t;\omega_c) = t^4 e^{-2\pi BW_{rate} t}\cos(2\pi\omega_c t) \\ h_{scale}(x;\Omega_c) = x^4 e^{-2\pi BW_{scale} x}\cos(2\pi\Omega_c x) \end{cases} \quad (3)$$

where the bandwidth $BW_{rate}$ and $BW_{scale}$ increase according the central frequency $\omega_c$ and $\Omega_c$. Fig. 1 shows 24 impulse responses of the 2-D modulation filters, with parameters of $\omega_c$ = {4, 8, 16, 32} Hz, $\Omega_c$ = {0.25, 0.5, 1} cycle/octave, and both sweeping directions encoded by the sign of $\omega_c$.

## 3. PROPOSED NEURAL NETWORK

### 3.1. Network architecture

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.

The 1-D convolution layer consists of 36 1-D kernels to perform the time-domain convolution in a similar way to cochlear filtering. The outputs of these 36 kernels are then merged into a 2-D acoustic scene, which is similar to a spectrogram. In the following 2-D convolution layer, we choose 24 2-D kernels for decomposing the spectrogram. Then the pooling layer is used to lower output dimensions while preserving important information. The fully-connected layers play the role in organizing and analyzing incoming information as the cognitive function induced beyond A1.

### 3.2. Auditory model initialized kernels and test scenarios

The 1-D kernels are used to simulate the cochlear filters on the logF axis. To cover the critical bandwidth of the cochlear filter, we use the resolution of 5 filters per octave in the proposed NN. Therefore, the 36 kernels (filters) span about 7 octaves to cover the frequency range of 30 Hz to 4000 Hz for speech

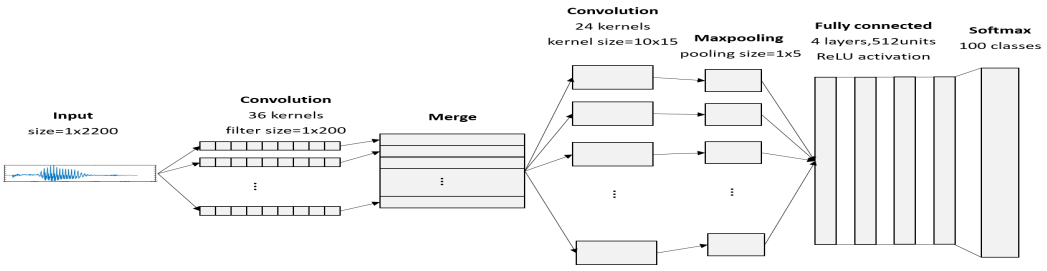


**Fig. 2**. Architecture of the proposed NN for speech processing on discriminative tasks.

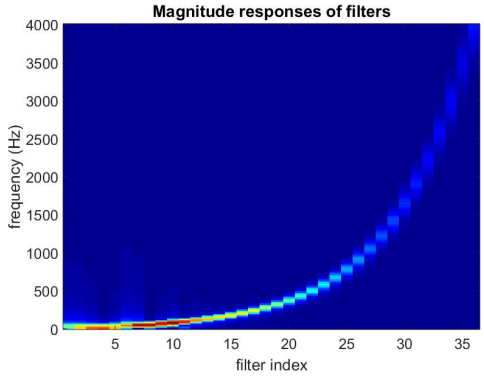


**Fig. 3**. Magnitude responses of 36 gammatone filters.

sampled at 8 kHz. In this study, we use impulse responses of a bank of 36 gammatone filters to implement the 1-D kernels. The magnitude responses of the gammatone filters are shown in Fig. 3.

For implementing the 2-D kernels, we use the 24 2-D impulse responses tuned to $\omega_c$ = {4, 8, 16, 32} Hz, $\Omega_c$ = {0.25, 0.5, 1} cyc/oct, as shown in Fig. 1. The selection of $\omega_c$ is to cover the speaking rate of a regular speaker and some finer temporal structures of his/her speech. The selection of $\Omega_c$ is basically to cover the formant structure of speech. Each 2-D kernel is with the size of 10x15, 10 being the frequency span on the logF axis covering 2 octaves and 15 being the number of frames on the temporal axis.

To investigate the effects of the auditory-model inspired kernels on system performance, we consider test scenarios from combinations of test conditions of the 1-D and the 2-D kernels. There are two test conditions for the 1-D kernels, fixed with gammatone filters and initialized by gammatone filters. There are also two test conditions for the 2-D kernels, initialized by A1 filters and randomly initialized. Except in the "fixed" condition, the kernels are allowed to change during training due to the backpropagation. All test scenarios are listed in Table 1 with the baseline system, BothRand, whose kernels are all randomly initialized.

**Table 1**. Five test scenarios for comparison

| Test Scenarios | | |
|---|---|---|
| **1-D kernels** | **2-D kernels** | **abbreviation** |
| Gammatone Fix | A1 Initial | GammaFix_A1Init |
| Gammatone Fix | A1 Random | GammaFix_A1Rand |
| Gammatone Initial | A1 Initial | GammaInit_A1Init |
| Gammatone Initial | A1 Random | GammaInit_A1Rand |
| Both Random | | BothRand |

## 4. EXPERIMENT RESULTS

### 4.1. Setting of the proposed NN

The proposed NN can be used for speech-related discriminative tasks. For evaluation, we conduct simulations on speaker identification using the 2008 NIST SRE (Speaker Recognition Evaluation) dataset. We used audio files from randomly selected 100 people in the short2 category of the training set for our simulations. We extracted active parts of each audio clip and divided them into 24 5-second long sections. Two sections with the highest energies were used for test and the other 22 sections were used for training.

For the proposed NN, voice segments of 275 ms (2200 points with the 8k sampling frequency) were used as input with a 10-ms (80-point) jump. The length of each 1-D kernel was set to 200 (i.e., 25 ms). Therefore, the output of the 1-D kernels can be thought as a logF-spectrogram like pattern with the frame duration of 10 ms. The following 2-D kernels were set to the size of 10x15, which covers 2 octaves in frequency and 150 ms in time. The 2-D pooling size used in the max-pooling layer was determined by simulations. We tried pooling sizes of 1x1, 1x5, 1x10, 2x2, and 14x5 with and without temporal overlapping between each pool. The best performance was achieved using the pooling size of 1x5 with the temporal overlap of 4 grids. Finally, these settings were used in the proposed NN for simulations.

### 4.2. Experiment results and discussions

To test the robustness of the proposed NN, we adopted the multi-condition training including two types of back-

**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% |

ground noise (buccaneer and factory noise from NOISEX-92 database [11]) at three SNRs (-5, 0, and 5 dB). The speaker identification rates for all test conditions are listed in Table 2. The performance reported in [10] using i-vectors/GMM in SSN noise is appended to the bottom of the table for reference.

The results clearly show the four methods with auditory-model induced kernels perform better than the method with randomly-initialized kernels, especially in low SNR environments. The identification rates produced by all these five methods, using the proposed NN model inspired by hearing perception, are higher than the reference rates produced by the i-vector/GMM system [10]. We can also observe the GammaFix_* method outperform the GammaInit_* method and the *_A1Init method outperform the *_A1Rand method in low SNR conditions (-5 and 0 dB). To sum up, the GammaFix_A1Init method offers the best performance at low SNRs. Surprisingly, this best method behaves just like auditory attention engaged. When people pay attention to a target sound in a noisy environment, they can recognize it more easily by selectively gating the incoming salient signal [12][13]. Such behaviors play a critical role in triggering task-dependent auditory plasticity of A1 neurons [14]. In other words, the spectro-temporal impulse responses of A1 neurons begin to self-adjust slightly when attention engaged to offer better discriminative ability for the task at hand.

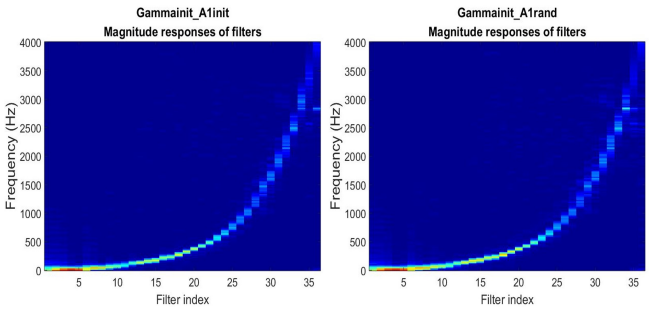


**Fig. 4**. Magnitude responses of 1-D kernels of GammaInit_A1* methods after training.

The 1-D kernels of the GammaInit_A1Init and GammaInit_A1Rand methods after training are plotted in Fig. 4. Compared with the fixed kernels shown in Fig. 3, these 1-D kernels do not change a lot but only show stronger responses in high frequency kernels probably due to adjustments to emphasize high frequency noise. Since these four Gamma*_A1* methods have 1-D kernels similar to gammatone filters, their 2-D kernels can be interpreted as extracting important spectro-temporal patterns for identifying speakers. Although not shown here, some of their 2-D kernels after training do possess simple patterns encoding pitch, speaking rate of the speaker and formant structures of speech. Some other 2-D kernels of course carry complicated patterns.

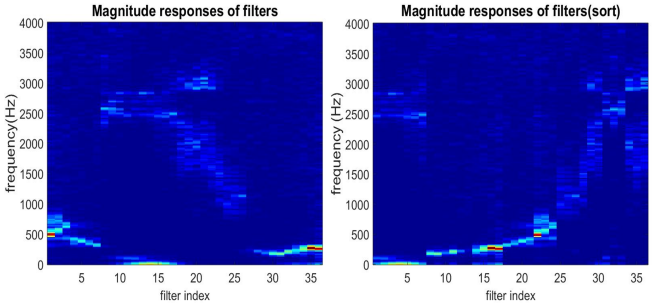


**Fig. 5**. Magnitude responses of 1-D kernels of BothRand method after training. The left panel shows the original responses and the right panel shows rearranged responses.

The 1-D kernels of the BothRand method after training are plotted in the left panel of Fig. 5. The right panel shows the responses after manually rearranging the orders of the kernels. After the rearrangement, the responses look a bit similar to the responses shown in Fig. 3 and Fig. 4. However, the rearrangement is not learned in the proposed NN such that the output of the 1-D kernels does not carry valid spectro-temporal joint patterns but only valid information in each frequency bin. Therefore, the meaning of the 2-D kernels cannot be interpreted intuitively.

## 5. CONCLUSION

In this paper, we propose a generative auditory model embedded NN for speech processing. The NN consists of a 1-D and a 2-D convolution layers, which simulate the filtering operations by the cochlea and the cortex, respectively. The generative basis functions of the auditory model are used to initialize the NN for a discriminative task. Simulation results show that these generative bases can boost speaker identification rates in noisy environments. The most robust method GammaFix_A1Init, whose 1-D kernels are fixed as gammatone filters and 2-D kernels are initialized by spectro-temporal bases of the auditory model, behaves like a generative-discriminative intertwined attention-engaged auditory model. In our opinion, the generative auditory model can still play the supportive role in building a better discriminative NN in this NN era.

## 6. REFERENCES


[1] Z. Q. Wang and D. Wang, "A joint training framework for robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 4, pp. 796–806, 2016.

[2] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," *in Proceedings of ICASSP*, pp. 4280–4284, 2015.

[3] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 3, pp. 483–492, 2016.

[4] B. Wu, K. Li, M. Yang, and C. H. Lee, "A reverberation-time-aware approach to speech dereverberation based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 102–111, 2017.

[5] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *Trans. Audio, Speech, and Language Processing*, vol. 23, no. 6, pp. 982–992, 2015.

[6] Taishih Chi, Powen Ru, and Shihab A Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *The Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[7] Mounya Elhilali, Taishih Chi, and Shihab A Shamma, "A spectro-temporal modulation index (stmi) for assessment of speech intelligibility," *Speech communication*, vol. 41, no. 2, pp. 331–348, 2003.

[8] Tai-Shih Chi, Ting-Han Lin, and Chung-Chien Hsu, "Spectro-temporal modulation energy based mask for robust speaker identification," *The Journal of the Acoustical Society of America*, vol. 131, no. 5, pp. EL368–EL374, 2012.

[9] Frederick Z Yen, Yin-Jyun Luo, and Tai-Shih Chi, "Singing voice separation using spectro-temporal modulation features.," in *ISMIR*, 2014, pp. 617–622.

[10] J. Chang and D. Wang, "Robust speaker recognition based on dnn/i-vectors and speech separation," *in Proceedings of ICASSP*, pp. 5415–5419, 2017.

[11] A. Varga and H. J.M. Steeneken, "Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, pp. 247–251, 1993.

[12] Ervin R Hafter, Anastasios Sarampalis, and Psyche Loui, "Auditory attention and filters," in *Auditory perception of sound sources*, pp. 115–142. Springer, 2008.

[13] Jonathan B Fritz, Mounya Elhilali, Stephen V David, and Shihab A Shamma, "Auditory attention: focusing the searchlight on sound," *Current opinion in neurobiology*, vol. 17, no. 4, pp. 437–455, 2007.

[14] Mounya Elhilali, Jonathan B Fritz, Tai-Shih Chi, and Shihab A Shamma, "Auditory cortical receptive fields: stable entities with plastic abilities," *Journal of Neuroscience*, vol. 27, no. 39, pp. 10372–10382, 2007.