*INTERSPEECH 2019*
September 15–19, 2019, Graz, Austria

# A study of a cross-language perception based on cortical analysis using biomimetic STRFs

*Sangwook Park*[1], *David K. Han*[2], *and Mounya Elhilali*[1]

[1]Department of Electrical and Computer Engineering, Johns Hopkins University
[2]Computaional and Information Sciences Directorate, Army Research Laboratory
{spark190, mounya}@jhu.edu, ctmkhan@gmail.com

## Abstract

For those in the early stage of learning a foreign language, they commonly experience difficulties in understanding spoken words in the second language, while they have no problem in recognizing words spoken in their mother tongue. This paper examines this phenomenon using biomimetic receptive fields that can be interpreted as a transfer function between acoustic stimulus and cortical responses in the brain. While receptive fields of individual subjects are often optimized to recognize unique phonemes in their mother language, it is unclear whether challenges associated with acquiring a new language (especially in adulthood) is due to a mismatch between phonemic characteristics in the new language and optimized processing in the system. We explore this question by contrasting biomimetic systems optimized for four different languages with sufficiently different characteristics. We perform English phoneme classification with these language-optimized systems. We observed distinctive characteristics in receptive fields emerging from each language, and the differences of English phoneme recognition performance accordingly.

**Index Terms**: auditory system, biomimetic STRFs, spectro-temporal receptive fields, cortical analysis, and phoneme classification

## 1. Introduction

Learning a new language by the time we reach over ten is often very difficult; and only few learners are able to achieve native accents [1]. Part of the challenge with new language acquisition for young and older adults is recognizing unique phonemes of a foreign language if these phonemes don't exist in their mother tongue. This difficulty gets amplified in noisy environments such as cocktail parties or in subway stations; even if native speakers may not find such environments as difficult. There has been some investigations into this phenomenon, focused on how nonnative phones are perceived into the categories for native phonemes [2]–[5]. According to [2], Guion, et. al. explored contrasts of consonants between Japanese and English and argued that foreign phonemes are perceived as the closest phoneme in their mother language based on a discrimination test. So and Best summarized that one's ability of recognizing phonemes from their mother language may not always be helpful in recognizing non-native phonemes [3]. Studies have shown that the difficulty with foreign phonemes can be mitigated as the person continues in learning the language by expanding the cumulative auditory exposure and experience [2], [4]. As stated earlier, young children typically below the age of ten have shown plasticity in their brain tuning with potential of recognizing a wide variety of phonemes regardless of language [5].

Our understanding of auditory processing in the brain suggests that sounds entering our ears are analyzed along a hierarchical set of stages that gradually extract acoustic attributes from the signal, guided by cognitive feedback from memory, prior knowledge and past experiences. Of particular interest in the current work is the analysis that takes place at the level of auditory cortex, where there is evidence of plasticity effects that are shaped by learning or guided attention [6], [7]. While processing in the mammalian auditory system is not fully understood, it has successfully been approximated using a linear systems approach where each neuron is modeled as a spectro-temporal convolution of the input signal and a system transfer function [8]–[12]. The spectro-temporal convolution model is represented in

$$r(t) = \iiint s(t,f)H(\tau - t,\ \zeta - f)d\tau d\zeta df \qquad (1)$$

where $t$ and $f$ are time and frequency indices respectively, and $H(t, f)$ is the transfer function of the model; also referred to as a Spectro-Temporal Receptive Field (STRF) [8]. As it is obvious from (1), the same stimulus $s(t,f)$ analyzed through an array of neurons with varying receptive fields $H_k$ will result in different mappings of the input signal. Here, we hypothesize that STRFs tuned to analyze statistical structure of one's native language may be suboptimal for some phonemes in other foreign languages. To demonstrate the emergence of distinctive STRFs as a function of language, our work is focused on cortical analysis denoted in (1) with language specific biomimetic STRFs.

To demonstrate this hypothesis, we first develop a training method for biomimetic STRFs based on Restricted Boltzmann Machine (RBM) and train biomimetic STRFs according to different languages. Then, we compare characteristics of the biomimetic STRFs in terms of Best Frequency (BF), Best Scale (BS), and Best Rate (BR) [13]. Additionally, we perform English phoneme classification using these language sensitive STRFs. From these investigations, we find that the biomimetic STRFs have different characters depending on the language, and their phoneme classification of a specific language varies depending on the language they were optimized for.

The remainder of this paper is as follows. First, we describe a method for training the biomimetic STRFs. Next, we develop measures for characterizing the STRFs. These STRFs are then compared for their performance in phoneme classification. Finally, the experimental results are summarized in Section 3 followed by a discussion and conclusion.



http://dx.doi.org/10.21437/Interspeech.2019-2507

## 2. Methods

### 2.1. Training biomimetic STRFs

Reverse correlation has been classically used for estimating STRFs [13], [14]. This method requires a pair of a stimulus and corresponding neural response acquired in the brain in order to infer the tuning characteristics of the neuron under study. In this work, we take a converse approach where *infer* STRF structure from the data using RBMs.

An audio waveform is first transformed into an auditory spectrogram that shows time-frequency representation based on a mammalian cochlear model [9]. We use the NSL toolbox to obtain an auditory spectrogram, and the frame length was set to 8 ms without overlaps [15]. A vector $x$ constructed by concatenating 15 frames is fed into the input layer in the RBM. For emulating neuron's activity that seems to be binary, binary hidden nodes are considered in the hidden layer. In here, the number of hidden nodes is set to 100, and this setting is the same in all STRFs.

A training cost is defined by combining a reconstruction error and a constraint on the number of active nodes in the hidden layer as

$$L = \frac{1}{2}\sum_m \left(x_m - \boldsymbol{W}^T g(\boldsymbol{h})\right)^2 + \lambda\left(p - \sum_i h_i\right)^2 \quad (2)$$

where $\boldsymbol{W}$ is a weight matrix of the RBM, $\boldsymbol{h}$ is a feedforward vector applied sigmoid function for reflecting a probability, $m$ is training data index, and $g$ is a binary sampling function based on Bernoulli distribution with $h$ as an onset probability. $p$ represents the average number of activated neurons ($p$=10), and is a regularization coefficient set at 0.0001. The constraint on the hidden nodes prevents overfitting and it emulates sparse neuronal reaction to a stimulus as observed in biological systems. Let $Y$ be a random variable representing the number of activated nodes by the function $g$. Then, the distribution known as the Poisson binomial distribution is denoted as

$$P(Y=k) = \sum_A \left[\prod_{i\in A} h_i \prod_{j\in A^c}\left(1-h_j\right)\right] \quad (3)$$

where $A$ is a set whose elements are possible combinations for choosing $k$ nodes from $N$ hidden nodes. This can be approximated by Binomial($N$, $\mu/N$) where $\mu = \sum_i h_i$ [16]. Thus, the sum of elements in feedforward vector means the average number of activation neurons. After training, we obtain biomimetic STRFs by reshaping the weight $\boldsymbol{W}$ to a 3D tensor in dimension of the number of STRFs, the number of frequency bins, and the number of frame bins, respectively.

Table 1: *Database used in training biomimetic STRFs*

| Language | Database | Subset |
|---|---|---|
| English | LibriSpeech [17] | dev-clean |
| Korean | ETRI 1000 [18] | read 1000 |
| Chinese | Thchs-30 [19] | training set |
| Japanese | Jsut v1.1 [20] | basic5000 |

In our implementation, we train the RBM by using the AdamOptimizer from TensorFlow library. The learning rate, batch size, and maximum number of iterations are set to 0.0001, 250, and 1000, respectively. This paper considers three Asian languages that have different linguistic origins compared to English, and databases used in training RBM are described in Table 1. Under the same parameters, language sensitive biomimetic STRFs are obtained by separately training RBMs according to the language.

### 2.2. Analysis of STRFs characteristics

An STRF includes excitatory and inhibitory regions in time-frequency domain. Since these regions have an effect on neurons' activity by adjusting a membrane potential, their shapes in time-frequency domain are important to characterize STRFs. Inseparability can be considered as one of the measures to characterize the shapes of excitatory and inhibitory regions [21]. To quantify inseparability, Singular Value Decomposition (SVD) is performed on each STRF. From its singular values, the inseparability is defined as

$$Ins = 1 - \sigma_1^2 / \sum_i \sigma_i^2 \quad (4)$$

where $\sigma_1$ is the biggest singular value as $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_K$. If the *Ins* is close to 0, the STRF can be approximated to rank 1 matrix, which can be approximated with a simple temporal function and a simple spectral function.

In addition, we analyze the language-sensitive STRF based on three additional metrics; Best Frequency (BF), Best Scale (BS), and Best Rate (BR) [13], [21]. Figure 1 shows an STRF and Scale-Rate (SR) plot as an example to explain the three metrics. The SR plot shows intensity along the spectral variation (i.e. scale) and temporal variation (i.e. rate) of STRF and can be obtained by applying a 2D Fourier Transform to an STRF. From the figure, BF is defined as a frequency which shows the maximum positive value along the frequency axis. In the right panel, BS and BR are defined as the centroid along the scale and the rate axes, respectively.

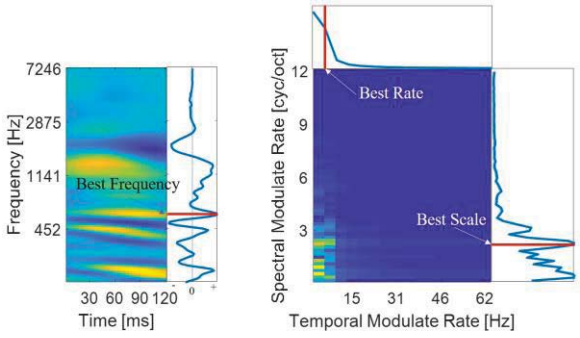


Figure 1: *STRF and SR Plot for explaining three metrics; Best Frequency, Best Scale, and Best Rate*

### 2.3. Phoneme classification

Next, we adapt the phoneme recognition framework used by Thomas et al. [10]. Figure 2 depicts a pipeline that employs the language-sensitive biomimetic STRF in phoneme recognition. As mentioned previously, audio waveforms are transformed into an auditory spectrogram as stimuli, $s(t,f)$, and neural response, $r(t,n)$, is estimated by performing a 2D convolution

and a frequency marginalization for each STRF ($H(t, f)$). By stacking the responses, a 100-dimensional response vector can be obtained per a frame, and Discrete Cosine Transform (DCT) is applied to the response vector for performing a decorrelation. After that, Multi-Layer Perceptron (MLP) is considered as a classifier. An input vector composed by concatenating 9 frames is fed into the input layer and goes to the hidden layer composed by 2000 nodes and the output layer in sequence. Note that the number of output nodes is determined depending on the number of target phonemes.

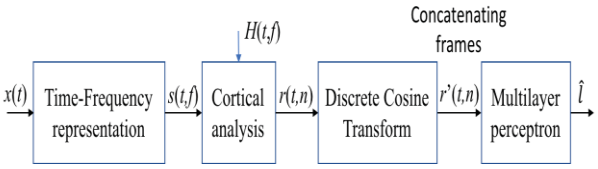


Figure 2: *A flowchart for phoneme classification based on biomimetic STRFs*

In here, two types of target phonemes, vowels and consonants, are considered, consistent with targets considered by Mesgarani et al. [13]. For system training, we use the TIMIT database excluding the 'SA' dialect sentences.

## 3. Results

### 3.1. Training biomimetic STRFs

Each language database in Table 1 is exclusively divided into a training and a validation sets with a ratio of 10:1. Figure 3 shows training and validation costs as defined in (2). The costs are significantly decreased after about 20 iterations. Note that the standard deviation of error between consecutive cost values after 100 iteration is about 0.001 in all cases.

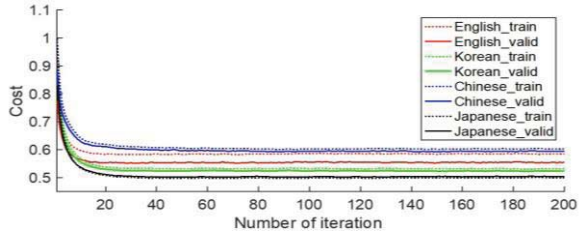


Figure 3: *Training and validation costs.*

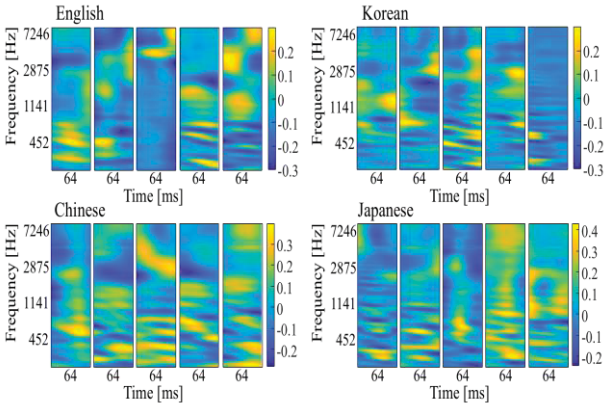


Figure 4: *Examples of biomimetic STRFs.*

Figure 4 gives an example of five STRFs from each RBM optimized for a different language. The frequency domain is represented in octave scale (from 184 Hz to 7246 Hz). In each panel, there are strong positive regions that contribute increasing feedforward excitation for activation of hidden nodes. On the other hand, negative region contribute strong inhibitions. Note that the locations and shapes of these regions are different depending on the language.

### 3.2. Analysis of STRFs characteristics

We look closely at the patterns of trained STRFs. Figure 5 examines the inseparability of each STRF group; and reveals that most distributions resembles a normal distribution, with subtle distinction across languages. Interestingly, none of the STRFs yield an inseparability of 0 indicating the presence of slanted excitatory and inhibitory regions with varying slopes that likely play a role in formant transition detection.

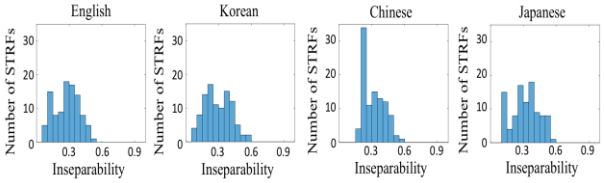


Figure 5: *Histograms of STRF's inseparability.*

Figure 6 shows histograms of STRF characteristics in terms of additional metrics. In the first row, the histograms show BF distributions of the four languages. From the fact that the energy in power spectrum of human voice is concentrated in the frequency band less than 1 kHz, we can expect that BFs are spread in the same band. As expected, BF in about 60% of the STRFs is less than 1.1 kHz in all four languages. Still, the frequency selectivity of some nodes extends beyond 1.1 Khz since the RBM used a fully-connected structure between the input and the hidden layer, though the spread of BF distributions is variable across languages.

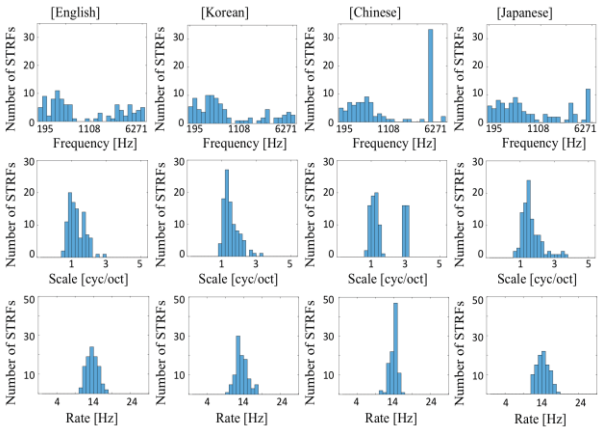


Figure 6: *Histograms of STRFs characteristics.*

The histograms in second and third rows show distributions in terms of BS and BR, respectively. For an assessment of similarity, we performed t-test between English and other languages. In BS, we found statistically significant differences with t-values were 3.8070, 4.6223, and 4.9035 compared to Korean, Chinese, and Japanese (critical value=2.81, when $\gamma = 0.001$ for three cases). On the other hand, t-values were

Table 2: *English phoneme classification rate: 12 vowels.*

| % | /ow | /ao | /aa | /ah | /ae | /eh | /ey | /ax | /iy | /ih | /ix | /ux | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| English | 59.91 | **75.88** | **79.36** | **66.02** | **71.47** | 49.84 | **72.43** | **69.36** | 67.13 | **54.67** | **56.57** | **83.85** | **67.21** |
| Korean | 63.27 | 75.00 | 75.27 | 57.14 | 69.97 | 52.00 | 60.55 | 65.32 | 63.64 | 47.66 | 49.86 | 76.69 | 63.03 |
| Chinese | 54.84 | 70.83 | 73.17 | 59.79 | 64.89 | **52.68** | 58.62 | 66.30 | **68.97** | 45.42 | 44.03 | 82.09 | 61.80 |
| Japanese | **67.11** | 69.04 | 68.40 | 50.26 | 68.72 | 43.11 | 57.59 | 51.09 | 56.37 | 46.02 | 47.70 | 76.72 | 58.53 |

Table 3: *English phoneme classification rate: 15 consonants.*

| % | /p | /t | /k | /b | /d | /g | /f | /s | /sh | /v | /dh | /z | /m | /n | /ng | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| English | 56.36 | **49.24** | 42.69 | 84.21 | 50.77 | 70.37 | 40.25 | 57.81 | **84.88** | **68.18** | 56.99 | **45.74** | **64.38** | 48.45 | **95.77** | **61.07** |
| Korean | 57.52 | 43.02 | 41.00 | 79.07 | 50.88 | **79.17** | 42.06 | **61.03** | 74.10 | 56.59 | **63.54** | 43.51 | 54.98 | **52.92** | 86.96 | 59.09 |
| Chinese | **66.67** | 31.98 | **43.14** | **87.23** | 46.48 | 70.45 | 42.28 | 57.91 | 76.47 | 63.64 | 50.00 | 43.91 | 64.04 | 48.17 | 85.37 | 58.52 |
| Japanese | 42.86 | 46.27 | 36.65 | 72.73 | **61.43** | 60.98 | **47.89** | 59.06 | 79.29 | 58.65 | 61.11 | 44.08 | 61.32 | 52.87 | 87.34 | 58.17 |

0.8541, 2.0864, and 2.7281 in BR indicating no significant differences. This lack of difference can be explained by the relationship between rates and speech tempo; and noting that the data used in this study included constrained articulation of speech such as broadcast news and read speech.

### 3.3. Phoneme classification

Finally, we perform English phoneme classification with the other three biomimetic STRFs. Because the number of test phoneme samples is different to each other in the TIMIT database (unbalanced data), we consider a class average classification rate, which is obtained by averaging diagonal elements in the confusion matrix, for an assessment. Table 2 and 3 shows the classification results for 12 vowels and 15 consonants, respectively. In both tables, class average classification rates are listed in the right column.

As expected, English sensitive STRFs shows the best results in both tests among the four STRFs. In vowels classification, the case of using English sensitive STRF shows the best in most of the vowels (9/12). On the other hand, the classification results of the other are similar among the four STRFs although the English sensitive STRF show the best average result. The differences of performances in consonant classification are smaller compared to that of the vowels. As shown in figure 4, the excitatory and inhibitory regions represented in STRFs resembles that of harmonics in speech spectrograms, and these shapes help extract formant features represented in vowels rather than consonants. As such, it is not surprising that vowel results are much better than the results of consonant classification.

It may be argued that a weakness of the experiment is the potential correlation between vowels and the speaker. However, the English sensitive STRFs are trained using LibreSpeech that is different from TIMIT; while the language databases included a wide variety of speakers. Many speakers participated in recording for the language database, thus the result supports our hypothesis.

## 4. Discussion

A human perceives speech through a complex process where it uses much higher-level information such as linguistic structures and contexts rather than just acoustic features. In studies of speech recognition, several modules called by language and lexicon model are additionally used for a natural language processing. In the current work, we only considered acoustic features in training of the biomimetic STRFs since the auditory model is the first step in the integrated framework of speech recognition. When an error occurs in the first step, its propagation downstream creates greater degrees of ambiguity.

As far as phoneme classification, the results reported here in table 2 and 3 are less than other state of the art studies. To a large degree, there is a great degree of mismatch between data used to train STRFs and test data. In [10], this issue was circumvented by using STRFs trained with TIMIT dataset by using a reverse correlation method, therefore avoiding issues of mismatch. Additionally, our approach trains STRFs without any phoneme labels and merely aims to infer general statistical structures prevalent in each language. As our focus was on subtle differences between language-sensitive STRFs, the approach used here employ a simple MLP rather than more a complex hierarchical structure.

## 5. Conclusions

This paper demonstrates why people have difficulties of recognizing non-native phonemes when learning a new language. From a cortical model that represents the neural response using a convolution between acoustic stimulus and STRFs, we explored that speech perception can be different depending on the learned STRFs. In order to demonstrate this hypothesis, we first developed an RBM based method for training biomimetic STRFs. After training biomimetic STRFs for four different languages, we found that their BF and BS depend on the language. We then performed English phoneme classification using these language sensitive STRFs. The English sensitive STRFs showed the best performance among the four languages supporting our hypothesis.

## 6. Acknowledgements

This research was supported by Office of Naval Research grants N000141612879, 629 N000141912014 and N000141712736 and National Institutes of Health grants U01AG058532 and R01HL133043.

# 7. References


[1] P. K. Kuhl, "Brain Mechanisms in Early Language Acquisition," *Neuron*, vol. 67, no. 5, pp. 713–727, 2010.

[2] S. G. Guion, J. E. Flege, R. Akahane-Yamada, and J. C. Pruitt, "An investigation of current models of second language speech perception: The case of Japanese adults' perception of English consonants," *J. Acoust. Soc. Am.*, vol. 107, no. 5, pp. 2711–2724, 2000.

[3] C. K. So and C. T. Best, "Cross-language perception of non-native tonal contrasts: Effects of native phonological and phonetic influences," *Lang. Speech*, vol. 53, no. 2, pp. 273–293, 2010.

[4] C. Best and W. Strange, "Effects of phonological and phonetic factors on cross-language perception of approximants," *J. Phon.*, vol. 20, no. 3, pp. 305–330, 1992.

[5] J. F. Werker and R. C. Tees, "Phonemic and phonetic factors in adult cross-language speech perception," *J. Acoust. Soc. Am.*, vol. 75, no. 6, pp. 1866–1878, 1984.

[6] S. Shamma and J. Fritz, "Adaptive auditory computations," *Curr. Opin. Neurobiol.*, vol. 25, pp. 164–168, 2014.

[7] R. C. Froemke, M. M. Merzenich, and C. E. Schreiner, "A synaptic memory trace for cortical receptive field plasticity," *Nature*, vol. 450, no. 7168, pp. 425–429, 2007.

[8] M. Elhilali, A. S. Shamma, Z. J. Simon, and B. J. Fritz, "A linear systems view to the concept of STRF," in *Handbook of modern techniques in auditory cortex*, D. Depireux and M. Elhilali, Eds. Hauppauge: NY: Nova Science Publishers, 2013, pp. 33–60.

[9] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 887–906, 2005.

[10] S. Thomas, K. Patil, S. Ganapathy, N. Mesgarani, and H. Hermansky, "A Phoneme Recognition Framework based on Auditory Spectro-Temporal Receptive Fields," in *INTERSPEECH*, 2010, no. September, pp. 2458–2461.

[11] M. A. Carlin, K. Patil, S. K. Nemala, and M. Elhilali, "Robust phoneme recognition based on biomimetic speech contours.," in *INTERSPEECH*, 2012, vol. 2, no. 3, pp. 1348–1351.

[12] T. Chi, Y. Gao, M. C. Guyton, P. Ru, and S. Shamma, "Spectro-temporal modulation transfer functions and speech intelligibility," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2719–2732, 1999.

[13] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, "Phoneme representation and classification in primary auditory cortex," *J. Acoust. Soc. Am.*, vol. 123, no. 2, pp. 899–909, 2008.

[14] D. J. Klein, D. A. Depireux, J. Z. Simon, and S. A. Shamma, "Robust Spectrotemporal and Reverse Correlation and for the Auditory and System: and Optimizing Stimulus and Design," *J. Comput. Neurosci.*, vol. 9, pp. 85–111, 2000.

[15] T. Chi and S. Shamma, "NSL Matlab Toolbox," University of Maryland, Colleage Park, 2005.

[16] K. P. Choi and A. Xia, "Approximating the number of successes in independent trials : Binomial versus Poisson," *Ann. Appl. Probab.*, vol. 12, no. 4, pp. 1139–1148, 2002.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2015, vol. 2015–Augus, pp. 5206–5210.

[18] "ETRI databse: Read speeches of 1000 speakers for speech recognition," *Electronics and Telecommunications Research Institute*. [Online]. Available: https://www.etri.re.kr/eng/main/main.etri.

[19] D. Wang and X. Zhang, "THCHS-30 : A Free Chinese Speech Corpus," 2015. [Online]. Available: http://arxiv.org/abs/1512.01882.

[20] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: free large-scale Japanese speech corpus for end-to-end speech synthesis," 2017. [Online]. Available: http://arxiv.org/abs/1711.00354.

[21] D. A. Depireux, J. Z. Simon, D. J. Klein, and S. A. Shamma, "Spectro-Temporal Response Field Characterization With Dynamic Ripples in Ferret Primary Auditory Cortex," *J. Neurophysiol.*, vol. 85, no. 3, pp. 1220–1234, 2017.