SYNAPTIC DEPRESSION IN DEEP NEURAL NETWORKS FOR SPEECH PROCESSING

Wenhao Zhang, Hanyu Li, Minda Yang, Nima Mesgarani

Department of Electrical Engineering, Columbia University, New York, NY, 10027

wz2293@columbia.edu, hl2776@columbia.edu, my2407@columbia.edu, nima@ee.columbia.edu

ABSTRACT

A characteristic property of biological neurons is their ability to dynamically change the synaptic efficacy in response to variable input conditions. This mechanism, known as synaptic depression, significantly contributes to the formation of normalized representation of speech features. Synaptic depression also contributes to the robust performance of biological systems. In this paper, we describe how synaptic depression can be modeled and incorporated into deep neural network architectures to improve their generalization ability. We observed that when synaptic depression is added to the hidden layers of a neural network, it reduces the effect of changing background activity in the node activations. In addition, we show that when synaptic depression is included in a deep neural network trained for phoneme classification, the performance of the network improves under noisy conditions not included in the training phase. Our results suggest that more complete neuron models may further reduce the gap between the biological performance and artificial computing, resulting in networks that better generalize to novel signal conditions.

Index Terms— synaptic depression, neural network, deep learning, phoneme recognition

1. INTRODUCTION

One of the major differences between biological neurons and the neuron models used in artificial neural networks is the ability of synaptic weights to change dynamically [1, 2, 3, 4, 5] in response to the fluctuations in the input to the neuron. One such mechanism is synaptic depression [2, 6, 7], a decrease in presynaptic efficacy from prolonged neurotransmitter release. Synaptic depression is widespread across cortical synapses [3]. Theoretical and experimental work has shown that synaptic depression can play a critical role in neural circuits [8, 9]. For example, our own experimental work has shown an important role for this mechanism in primary auditory cortical neurons in the formation of a robust representation of an acoustic stimulus that remains unchanged in noisy and reverberant conditions [10]. Since current neural network models do not generalize well to conditions not included in their training, it may be important to determine whether these mechanisms can benefit these models, particularly in generalization to unseen conditions. In this paper, we studied ways by which synaptic depression can be modeled efficiently and the distinct computation that a system with dynamic synaptic connection can perform. Furthermore, we incorporated the synaptic depression into the a deep neural network model trained for phoneme recognition in clean speech which resulted in improved classification accuracy in a variety of unseen noisy conditions.

Minda Yang was partially funded by a grant from Wei Family Foundation. This work was funded by a grant from National Institute of Health, NIDCD, DC014279 and the Pew Charitable Trusts, Pew Biomedical Scholars Program.

2. METHODS

To create efficient models of synaptic depression and gain insight into their computational principles, we started with the standard neuron model, which consists of the weighted sum of the inputs followed by a nonlinear transformation (e.g. Rectified Linear [11]) of the form: $y(t) = f(z(t))$, where $z(t) = \sum_i w_i x_i(t) + b$, and x_i is the input to the neuron, w_i is the weight associated with each input, b is the bias of the neuron.

2.1. Modeling synaptic depression

Models of synaptic depression capture the nonlinear dynamics caused by decreased synaptic weights of a neuron [7]. This assumes that depression at a given synapse could be explained by two parameters: the rate of vesicle depletion per presynaptic action potential and the time constant of vesicle recovery [2]. This phenomenon has been functionally modeled in two alternative ways. In the first approach, which we call *weight depression*, each synaptic weight, w, has a multiplicative depression factor. The alternative approach, which we refer to as *bias depression*, dynamically changes the bias parameter of the neuron, b, depending on the overall input to the neuron. While the general effect of these two models is the same, there are important functional differences because each can be more effective under specific signal distortions, such as additive versus multiplicative noises.

2.1.1. Weight Depression

This model assigns a multiplicative depression factor to every weight in the network [12] (Fig. 1a). The depression factor, $d_{i,w}(t)$, is estimated using a recursive equation:

$$d_{i,w}(t) = (1 - \frac{1}{\tau})d_{i,w}(t-1) + vx_i(t-1)(1 - d_{i,w}(t-1)) \quad (1)$$

where the depression factor d, bounded between 0 (fully recovered) and 1 (fully depressed), is the time constant of vesicle recovery and represents the rate of vesicle depletion per presynaptic action potential. The equation of a standard neuron is then modified accordingly:

$$y(t) = f(\Sigma_i w_i(t)x_i(t)(1 - d_{i,w}(t)) + b) \quad (2)$$

Effectively, this operation modulates each weight of the standard neural network dynamically based on the short-term history of its corresponding input. Therefore, if one input becomes very large, the neuron automatically decreases the contribution of that input to maintain the same computation. If we assume a constant input, the depression coefficient will converge to $\frac{\tau v x}{1+\tau v x}$, resulting in the effective input of $\frac{x}{1+\tau v x}$. Therefore, the equation of the neuron in equilibrium is $y(t) = f(\sum_i w_i(t)\frac{x_i}{1+\tau v x_i})$. If we choose τv such that $\tau v x_i \ll 1$ in normal signal condition, then the small inputs remain almost unchanged and the network operates as if synaptic depression is not present. However, large inputs to the neuron will become compressed (Fig. 1.b). This mechanism is particularly effective in

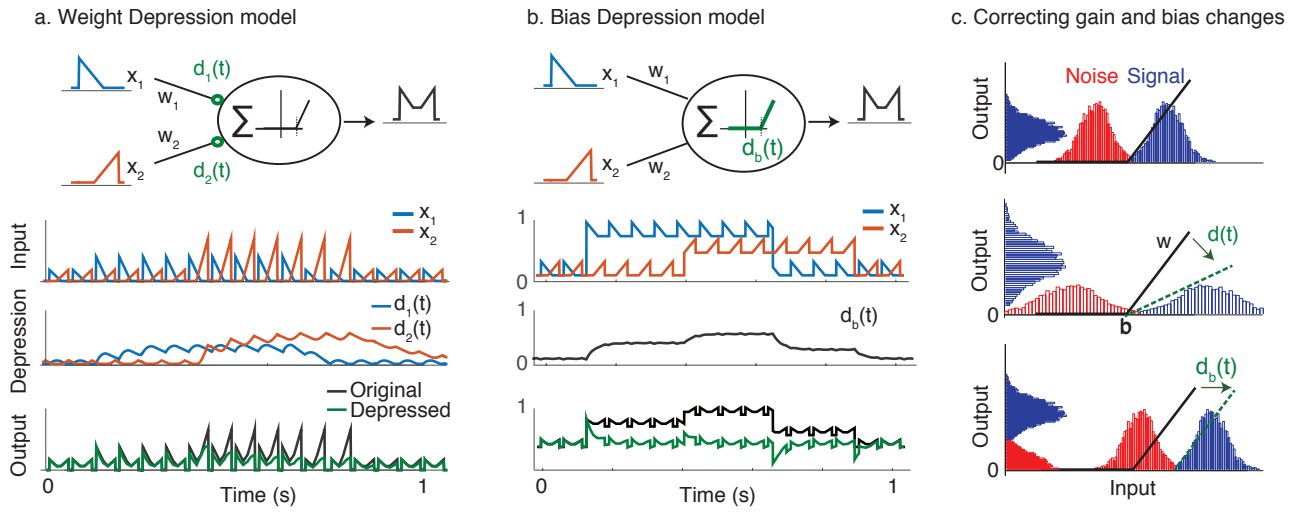

Fig. 1. Charaterizing the effect of synaptic depression in a model neuron. a) A neuron with weight depression is able to normalize the variations in input gain, b) A neuron with bias depression is able to normalize the deviations of the summed input from its expected mean. c) Interaction between neuron's nonlinearity and the synaptic depression in separating signal (blue) and noise (red).

suppressing multiplicative distortions where each x_i is replaced with $m_i x_i$. Figure 1 shows the simulation result of a simple neuron with two inputs. Comparing of the output of the neuron with and without depression shows increased robustness to unwanted gain variations of the input.

2.1.2. Bias Depression

The alternative model for synaptic depression is the dynamic spike threshold model [13, 14], in which the threshold of spiking for each neuron can vary over a broad range depending on the sum of the weighted input to the neuron, $z(t)$:

$$d_b(t) = (1 - \frac{1}{\tau})d_b(t-1) + vz(t-1) \tag{3}$$

This model is then integrated into the standard neuron model as a dynamic change of the bias:

$$y(t) = f(\Sigma_i w_i x_i + b - d_b(t)) \tag{4}$$

This mechanism effectively tracks the long-term average of the overall input to a neuron and can correct any deviation from this mean. For example, if each input is corrupted with additive noise c_i, then the new input to the neuron will be $\sum_i w_i(x_i+c_i)+b = \sum_i w_i x_i + b + \sum_i w_i c_i$. If the bias b adapts dynamically to this change, it can effectively cancel out the noise term $\sum_i w_i c_i$. Unlike the weight depression (equation 1), bias depression (equation 3) is linear and can be analyzed accordingly: $\frac{D_b(\omega)}{Z(\omega)} = \frac{v}{1-(1-\frac{1}{\tau})e^{j\omega}}$. The simulation shown in Fig. 1 shows how a change in the bias of inputs can be normalized using the dynamic bias depression model.

2.2. Interaction between synaptic depression and the nonlinearity

Neural network models can warp the input feature space nonlinearly and implement complex decision boundaries using many canonical nonlinear projections (e.g., sigmoid or rectified linear nonlinearity) [15]. Nonlinear computations, however, are very sensitive to the range of the input (operating point), otherwise the nonlinearity may not operate as it was intended [16, 17]. Dynamic synaptic changes proposed here could play a crucial role in keeping the system within

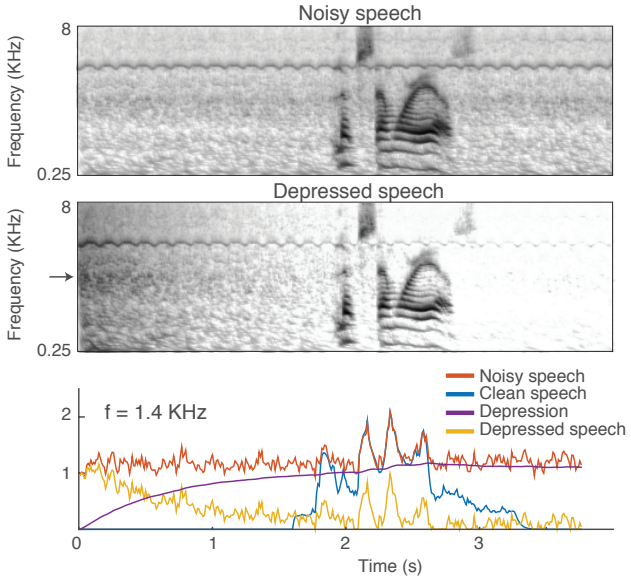

Fig. 2. (top) Noisy speech passes through a standard auto-encoder neural network model with one hidden layer, (mid) adding synaptic depression in the hidden layer reduces the background noise in the output. (bottom) The dynamic of depression parameters for a single frequency channel shown by arrowhead.

the correct operating range. As a result, these dynamic mechanisms combined with the nonlinearity of neurons cannot be seen merely as a mean normalization. This point is illustrated in Fig. 1.c, showing a rectified-linear neuron with one input and bias, b . The neuron performs a nonlinear filtering by passing the signal (blue) through, while completely eliminating the noise (red). A change in the bias of the input can result in non-optimal filtering (Fig. 1.a, bottom). An adaptive threshold however, $b - d_b(t)$, can track this change, and restore the intended nonlinear computation of this example neuron. An increase in the gain on the other hand can also produce non-optimal output (1.c), a distortion that can be compensated for by

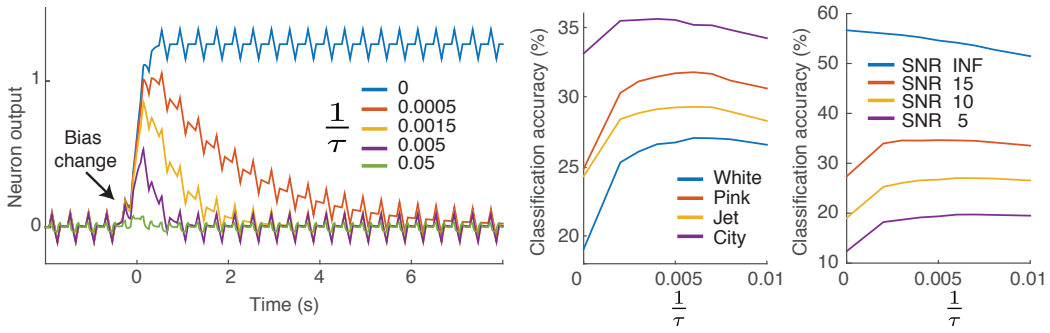

Fig. 3. (left) Dependence of convergence speed of a neuron with bias depression as a function of $\frac{1}{\tau}$. (middle and right) Phoneme classification accuracy with varying $\frac{1}{\tau}$ values in various noise types and different SNR values.

weight depression model which effectively changes the slope of the transformation.

3. NEURAL NETWORK MODELS WITH BIAS DEPRESSION

Since many of the noises that naturally occur in real-world conditions are additive, we focus on bias depression model (equation 3) for the remainder of this paper and examine its effective computation when integrated in an artificial neural network model. First, we test the bias depression model in an auto-encoder network to show how the background activity is suppressed in the reconstructed output with and without depression. Second, we show the effect of integrating the bias depression model into a multilayer neural network model trained for phoneme recognition. To evaluate the performance of the network with synaptic depression, we used a frame-wise phoneme classification task to directly measure the effect of this added mechanism on the acoustic model. The classification accuracy with synaptic depression is then evaluated in noisy conditions that were not included in the training of the networks.

3.1. Autoencoder with Synaptic Depression

To provide an intuitive account of how this process works in a neural network model, we first incorporate the proposed bias depression model in an autoencoder network [18]. The input and output of the network are 128 frequency channels, and the network has one hidden layer consisting of 128 rectified linear neurons. The autoencoder network is trained to reconstruct the time-frequency representation of speech. Fig. 2 shows the simulation results of this network. In the absence of synaptic depression in the hidden layer, additive noise passes through the network and the reconstructed signal is also noisy (Fig. 2, top). However when the bias depression model (equation 3) is added to the hidden layer neurons, this mechanism adapts the activation threshold to the changing statistics of the input and removes the additive background activity (Fig. 2, middle). Figure 2, bottom shows the effect for one frequency channel, where the increased bias due to noise is subtracted out after depression. The auto-encoder network intuitively shows how bias depression could make conventional neural networks more robust to the changes in input signal conditions.

3.2. Phoneme classification using a DNN with Synaptic Depression

Next, we quantified the effect of synaptic depression using a neural network model trained for phoneme classification. We used the frame-wise phoneme classification metric to explicitly measure the

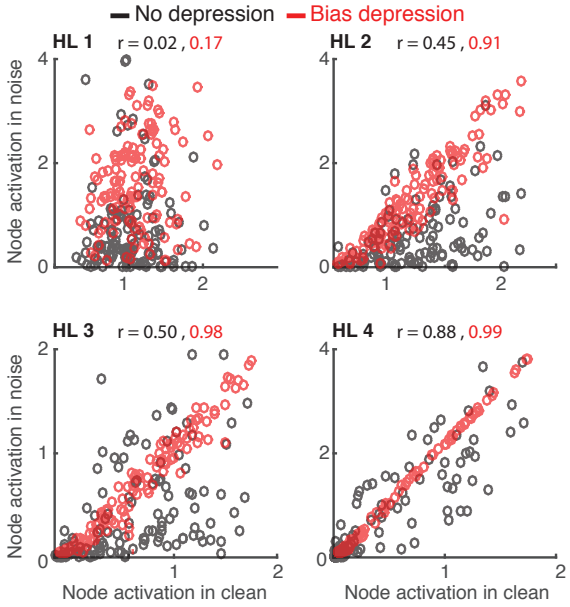

Fig. 4. Average node activation in clean and noisy speech in hidden layers of the deep neural network with and without bias depression (shown in black and red accordingly). Here $\frac{1}{\tau} = 0.003$.

accuracy of the acoustic model with and without synaptic depression. We also studied how parameters such as noise type, SNR, and the speed of convergence $\frac{1}{\tau}$ affect the results.

The phoneme classification network used has four hidden layers with rectified linear nonlinearity and was trained on clean speech on the TIMIT corpus [19]. The network had an input layer with 792 dimensions corresponding to 11 frames of 24-dimensional log-Mel filter bank coefficients, deltas, and double deltas. There were 4 hidden rectified linear layers with 128 nodes each and a sigmoid output layer with 40 nodes corresponding to the HMM emission probability of one of 39 English phonemes and silence. The model parameters were initialized using unsupervised RBM layer-wise pretraining and then fine-tuned using 25 epochs of backpropagation with minimization of mean square error objective function. Various types of noise from the Noisex database [20] were then added to the test samples at different SNR levels to probe the performance of the network in variety of signal conditions. We used frame-wise phoneme classification accuracy to measure the performance, excluding the silence category. The accuracy on Timit test subset in clean is 56.65%.

SNR	White Noise	Pink Noise	Jet Noise	City Noise	Average
INF	56.65% / 54.61%	–	–	–	–
20	35.81% / 41.39%	42.89% / 45.33%	42.13% / 43.83%	47.07% / 46.84%	41.98% / 44.35%
15	27.39% / 34.63%	34.27% / 38.99%	33.13% / 37.18%	41.03% / 42.22%	33.96% / 38.26%
10	19.05% / 26.71%	24.83% / 31.69%	24.28% / 29.23%	33.11% / 35.52%	25.32% / 30.79%
5	12.38% / 19.37%	15.89% / 22.52%	15.48% / 20.94%	23.40% / 27.92%	16.79% / 22.69%
0	7.90% / 14.17%	9.15% / 14.83%	8.45% / 13.85%	14.95% / 19.52%	10.11% / 15.59%
Average	20.51% / 27.25%	25.41% / 30.67%	24.69% / 29.01%	31.91% / 34.40%	–

Table 1. Classification accuracy, number in each entry is obtained by network without depression / with depression

3.2.1. Choosing the depression parameters

The recursive equation used in these networks can be also expressed in the following form:

$$d_b(t) = (1 - \frac{1}{\tau})d_b(t-1) + \frac{\beta}{\tau}(z(t-1) - \bar{z}) \qquad (5)$$

where $\beta = \tau v$. We assume the average node activation $\bar{z}$ in clean speech conditions can be measured during the training and is known. When the network is faced with noisy speech, this mechanism restores the expected average node activation as in the clean condition. Typically β should be slightly smaller than 1 to ensure that depression minimizes the difference between z and $\bar{z}$. The depression parameter $\frac{1}{\tau}$ determines the duration of the signal that affects the computation, and therefore, specifies the rate of convergence for the bias depression model. At one extreme, when $\frac{1}{\tau} = 0$, the effect of z disappears completely; hence, d_b will remain unchanged. On the other extreme, when $\frac{1}{\tau} = 1$, d_b will always stay at $\beta(z - \bar{z})$ and will not adapt the activation threshold dynamically. A simple simulation shown in Fig. 3 illustrates the effect of $\frac{1}{\tau}$ on the convergence speech. Small values of $\frac{1}{\tau}$ result in long adaptation of the bias, where large $\frac{1}{\tau}$ values can potentially distort the stimulus Fig. 3 (left) by following it too quickly. An optimal value of $\frac{1}{\tau}$ can therefore be found to maximize an objective function, e.g., phoneme classification accuracy in noisy conditions. The dependence of optimal $\frac{1}{\tau}$ on phoneme classification in various noise types and signal to noise ratios is shown in Fig. 3. Despite the slight differences between the plots in Fig. 3, the overall trend is the same, confirming the existence of a tradeoff between the speed of convergence and distorting the target signal.

3.2.2. Average activation of nodes in clean and noise

To examine the effect of bias depression on the average activation of nodes in clean and noisy conditions, we measured the mean activation values of each node in different layers of the network and compared these values under clean and noisy conditions. Fig. 4 shows the scatter plots for the four hidden layers with and without bias depression (shown in black and red, respectively). This figure demonstrates an increasingly similar average activation throughout the network in subsequent hidden layers, which almost normalizes the mean distortions caused by the noise (correlation value with and without depression are shown on top of each plot in Fig. 4). This in effect shows an effective role for bias depression in maintaining a consistent activation that is able to minimize the effect of background noise on speech features, an effect similar to what we observed in the response of biological neurons to speech in noise [10].

3.2.3. Effect of Synaptic Depression in different noise types and SNRs

Phoneme classification accuracies for various noise types and signal-to-noise ratios are shown in Table 1. The accuracy of the networks with and without depression are separated in each column, and the depression parameter $\frac{1}{\tau}$ is fixed at 0.005, determined by the analysis in 3.

Layer with depression	Accuracy
All layers	26.06%
HL 1	24.66%
HL 2	22.71%
HL 3	22.16%
HL 4	20.53%

Table 2. Classification accuracy in DNN with bias depression applied to different layers

As shown in the table, applying synaptic depression has a minimal effect on the accuracy of the network in the clean condition ($\sim 2\%$ reduction). However, when the network is tested in noisy conditions not included in the training, adding the bias depression significantly improves the accuracy. This increased performance is particularly significant for noises that are more stationary with slower temporal dynamics compare to speech. The average gain for different noise types and SNRs are also shown in the last row and column of the table.

4. DISCUSSION

This study shows the feasibility and usefulness of integrating novel bio-inspired mechanisms, specifically synaptic depression, into the current neural network models used for acoustic modeling. We used both qualitative and quantitative methods to describe the effective computation of synaptic depression and how it increases the generalization of the network to unseen conditions. As shown in the example in Fig. 2, incorporating this mechanism in the superficial layers of a network can simply result in adaptation to changing input variations (e.g., baseline change, or channel gain). However, the operation is more convoluted when synaptic depression is added to nodes in deeper layers of a neural network. For example, we have recently shown that the nodes in a DNN trained for phoneme recognition become increasingly more selective to distinct phonetic features of speakers [21]. Therefore, a change in the gain or bias at a deep layer may represent a high-level distortion in the acoustic domain (e.g. the presence of a competing speaker [22]). The parallel efforts in understanding the computational principles of DNNs will be very beneficial in learning more about how the changes in neuron models or network architecture impact the behavior of the network. The future directions of this work include a more realistic benchmark and integration of this mechanism in the state-of-the-art neural network models, as well as exploring whether or not training the depression parameters jointly with the network will benefit the phoneme accuracy.

5. REFERENCES

[1] N. C. Rabinowitz, B. D. B. Willmore, A. J. King, and J. W. H. Schnupp, Constructing Noise-Invariant Representations of Sound in the Auditory Pathway, PLoS Biol., vol. 11, no. 11, p. e1001710, 2013.

[2] L. F. Abbott, J. A. Varela, K. Sen, and S. B. Nelson, Synaptic depression and cortical gain control, Science (80-.)., vol. 275, no. 5297, pp. 221224, 1997.

[3] M. V Tsodyks and H. Markram, The neural code between neocortical pyramidal neurons depends on neurotransmitter release probability, Proc. Natl. Acad. Sci., vol. 94, no. 2, pp. 719723, 1997.

[4] N. C. Rabinowitz, B. D. B. Willmore, J. W. H. Schnupp, and A. J. King, Contrast gain control in auditory cortex, Neuron, vol. 70, no. 6, pp. 11781191, 2011.

[5] R. C. Moore, T. Lee, and F. E. Theunissen, Noise-invariant Neurons in the Avian Auditory Cortex: Hearing the Song in Noise, PLoS Comput. Biol., vol. 9, no. 3, p. e1002942, 2013.

[6] S. V David, N. Mesgarani, J. B. Fritz, and S. A. Shamma, Rapid synaptic depression explains nonlinear modulation of spectrotemporal tuning in primary auditory cortex by natural stimuli, J Neurosci, vol. 29, no. 11, pp. 33743386, 2009.

[7] M. Tsodyks, K. Pawelzik, and H. Markram, Neural networks with dynamic synapses, Neural Comput., vol. 10, no. 4, pp. 821835, 1998.

[8] M. Carandini, D. J. Heeger, and W. Senn, A synaptic explanation of suppression in visual cortex, J. Neurosci., vol. 22, no. 22, pp. 1005310065, 2002.

[9] D. V Buonomano and W. Maass, State-dependent computations: spatiotemporal processing in cortical networks, Nat. Rev. Neurosci., vol. 10, no. 2, pp. 113125, 2009.

[10] N. Mesgarani, S. V David, J. B. Fritz, and S. A. Shamma, Mechanisms of noise robust representation of speech in primary auditory cortex, Proc. Natl. Acad. Sci., vol. 111, no. 18, pp. 67926797, 2014.

[11] X. Glorot, A. Bordes, and Y. Bengio, Deep sparse rectifier neural networks, in International Conference on Artificial Intelligence and Statistics, 2011, pp. 315323.

[12] H. Markram and M. Tsodyks, Redistribution of synaptic efficacy between neocortical pyramidal neurons, Nature, vol. 382, no. 6594, pp. 807810, 1996.

[13] R. Azouz and C. M. Gray, Dynamic spike threshold reveals a mechanism for synaptic coincidence detection in cortical neurons in vivo, Proc. Natl. Acad. Sci. U. S. A., vol. 97, no. 14, p. 8110, 2000.

[14] W. B. Wilent and D. Contreras, Stimulus-dependent changes in spike threshold enhance feature selectivity in rat barrel cortex neurons, J. Neurosci., vol. 25, no. 11, p. 2983, 2005.

[15] K. Hornik, M. Stinchcombe, and H. White. "Multilayer Feedforward Networks are Universal Approximators." *Neural Networks*, vol. 2 pp. 359-366, 1989.

[16] G. A. F. Seber and C. J. Wild, Nonlinear regression. Wiley-Interscience, 2003.

[17] G. B. Christianson, M. Sahani, and J. F. Linden, The consequences of response nonlinearities for interpretation of spectrotemporal receptive fields, J Neurosci, vol. 28, no. 2, pp. 446455, 2008.

[18] Hinton, Geoffrey E., and Ruslan R. Salakhutdinov. "Reducing the dimensionality of data with neural networks." Science 313.5786 (2006): 504-507.

[19] J. S. Garofolo Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgrena, N. L., Zue, V., TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguist. Data Consort., 1993.

[20] 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 Commun., vol. 12, no. 3, pp. 247251, 1993.

[21] T. Nagamine, M. L. Seltzer, and N. Mesgarani, Exploring How Deep Neural Networks Form Phonemic Categories, in Sixteenth Annual Conference of the International Speech Communication Association, 2015.

[22] N. Mesgarani and E. F. Chang, Selective cortical representation of attended speaker in multi-talker speech perception, Nature, vol. 485, no. 7397, pp. 233236, 2012.