INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

On the Role of Nonlinear Transformations in Deep Neural Network Acoustic Models

Tasha Nagamine[1], *Michael L. Seltzer*[2], *Nima Mesgarani*[1]

[1]Department of Electrical Engineering, Columbia University, New York, USA
[2]Microsoft Research, Redmond, USA

tasha.nagamine@columbia.edu, mseltzer@microsoft.com, nima@ee.columbia.edu

Abstract

Deep neural networks (DNNs) are widely utilized for acoustic modeling in speech recognition systems. Through training, DNNs used for phoneme recognition nonlinearly transform the time-frequency representation of a speech signal into a sequence of invariant phonemic categories. However, little is known about how this nonlinear mapping is performed and what its implications are for the classification of individual phones and phonemic categories. In this paper, we analyze a sigmoid DNN trained for a phoneme recognition task and characterized several aspects of the nonlinear transformations that occur in hidden layers. We show that the function learned by deeper hidden layers becomes increasingly nonlinear, and that network selectively warps the feature space so as to increase the discriminability of acoustically similar phones, aiding in their classification. We also demonstrate that the nonlinear transformation of the feature space in deeper layers is more dedicated to the phone instances that are more difficult to discriminate, while the more separable phones are dealt with in the superficial layers of the network. This study describes how successive nonlinear transformations are applied to the feature space non-uniformly when a deep neural network model learns categorical boundaries, which may partly explain their superior performance in pattern classification applications.

Index Terms: Deep neural networks, deep learning, automatic speech recognition

1. Introduction

In recent years, deep neural networks (DNNs) have come to dominate both research and industry in the field of automatic speech recognition [1, 2, 3]. This is due to their significant performance advantage over previous models such as GMMs [4], made possible by advances in training algorithms and GPU computing [5]. Despite these advances, state-of-the-art ASR systems cannot match human-level performance in a variety of speech recognition tasks [6], motivating a number of recent studies aimed at better understanding deep learning in the hope of gaining intuitions that may lead to improved models [6, 7, 8, 9, 11, 12].

One aspect of deep learning that has not been extensively studied is how the multiple layers of nonlinearity used in a DNN aid in the formation of invariant phonemic categories. It is a well-known fact that networks with at least one hidden layer are universal approximators [13, 14], but this does not provide specific insight regarding the advantage of using multiple nonlinear transformations of features in ASR tasks. In our previous work, [11], we characterized the representational properties of nodes in the hidden layers of a DNN acoustic model. In this study, we focus on the question of what nonlinear transformations are applied to the features as they are mapped from one layer to another, and how do these transformations create the complex categorical boundaries needed to separate them apart? Answering these questions is crucial in explaining the computational principles of deep neural network models, revealing their limitations, and providing a link between the complexity of a given task and the required network architecture.

2. DNN Architecture

The DNN analyzed in this study was trained for phoneme recognition on the clean training set of the WSJ Aurora 4 corpus. 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 five hidden sigmoid layers with 256 nodes each and an output layer with 41 nodes corresponding to the HMM emission probability of one of 40 English phonemes and silence. The model parameters were initialized using unsupervised restricted Boltzmann machine (RBM) layerwise pretraining and then fine-tuned using 25 epochs of backpropagation with a cross-entropy objective function. While we chose to use a small, context-independent network to limit the number of parameters in the analyses, the questions raised in this study are general and likely to generalize to larger, more complex models.

3. Results

We begin by quantifying the degree of nonlinearity in each hidden layer of the network. Next, we characterize the properties of the nonlinear transformations in hidden layers of the DNN and describe how the complex phonemic boundaries are created. Finally, we study the successive transformations applied to the instances of phones that are responsible for the successful discrimination of confusable phones.

3.1. The representation of speech is characterized by increasing nonlinearity in deeper layers

The representational power of a DNN comes from the successive nonlinear transformations of the input [14]. In order to confirm that the deeper layers of the network represent the speech signal more nonlinearly, we compared the linear approximation of the activation of each node with the actual node activation. As such, we trained a linear model for each hidden layer mapping one 792-dimensional input frame to one frame of hidden layer activations (ridge regression, ridge parameter λ determined by grid search with 10-fold cross-validation). These linear activations serve as a basis of comparison to understand

http://dx.doi.org/10.21437/Interspeech.2016-1406

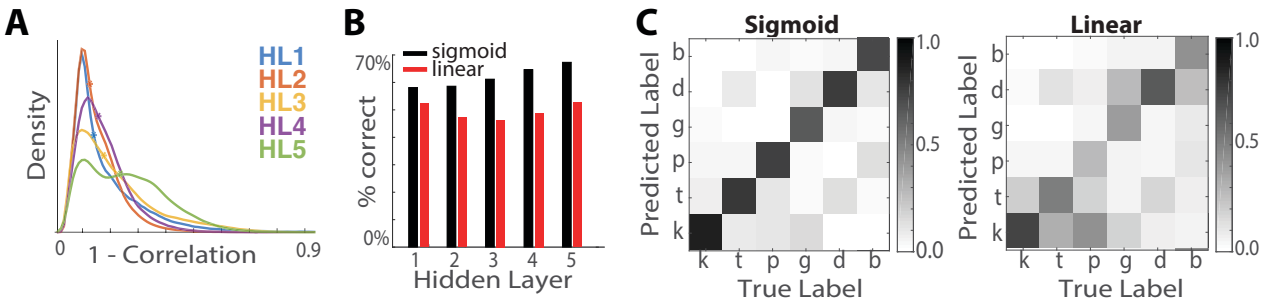

Figure 1: *(A) Non-parametric distribution fitted to the histogram of correlation errors between the actual and linearly predicted DNN activations for validation sentences in each hidden layer. Asterisks show median value of the distribution. (B) Softmax regression classification accuracy over frames for sigmoid DNN and linear model. (C) Confusion matrices for plosives for classifiers trained on hidden layer 5 of sigmoid network and linear model.*

what parts of the node activations cannot be captured linearly and thus what must be encoded by the nonlinear transformations in the network. Linear activations were obtained on 900 sentences from dialects 3 and 7 of the TIMIT speech corpus (51 male and 49 female speakers) [15]. After regression, linear activations with values above and below 0 and 1 were set to 0 and 1, respectively.

To quantify and visualize the amount of nonlinearity present in the hidden layers of the DNN, we calculated the correlation coefficient ρ (Pearson's r) between actual and linearly predicted DNN activations. For each hidden layer, a non-parametric distribution was fitted to the histogram of estimation errors $(1 - \rho)$ between the actual and linearly estimated activations of the 256 node to each sentence (Figure 1A). We find that the transformed representation of the input becomes increasingly nonlinear in deeper layers, evidenced by the greater degree of error between nonlinear and linearly estimated activations.

To confirm an increased discriminability of phoneme classes in deeper layers of the network, we measured the phone classification accuracy in each hidden layer by training a softmax regression for frame-wise recognition using the activations of the nodes in that layer. This was done without changing the original weights of the network; as a result, the output of these classifiers reflects the separability of phones in the corresponding hidden layer. The same was done with activations from the linear approximations of the hidden layer activations. Comparing the frame-wise classification results in subsequent hidden layers reveals an increased classification accuracy for the sigmoid network. This trend was not observed in the linear model (Figure 1B), confirming that transformations that result in increased separability of phones in deeper layers of the network are increasingly nonlinear. These nonlinear transformations also decrease the overlap between more similar phonemes, as can be seen in the confusion matrices of Figure 1C.

3.2. Nonlinearity expands the acoustic dimensions non-uniformly

While the previous analysis confirmed a more nonlinear and separable representation of phoneme classes in deeper layers of the network, it does not provide an explanation for how the network achieves this task. To explicitly characterize the nonlinear transformations that are applied to the spectrotemporal features in each layer, we trained softmax classifiers using the entire TIMIT training set on each hidden layer of the network (HL1 to HL5) using Theano [16]. Drawing on the intuition gained from the confusion matrices in Figure 1, we also performed softmax regression on a subset of selected consonants (Table 1) with target labels determined by either manner or place of articulation [17]. As expected, classification accuracy was much greater for manner. However, we can see that for place, which has weaker acoustic correlates and should thus be more difficult to discriminate, the relative improvement (bold percentages in Table 1) compared to the accuracy in hidden layer 1 was much greater. This suggest a hierarchical classification scheme in the network, where the superficial layers transform the features into a space where manners are easily separable, and the nonlinear transformations in the deeper layers of the network selectively warp the acoustic features that aid in a successful discrimination of place features.

To investigate and visualize the mechanisms by which the nonlinearity facilitates improved recognition accuracy for easily confused phonemes, we used multidimensional scaling (MDS) [18] to visualize the relative location of different phones in each hidden layer as reflected in the softmax output vectors. A two-dimensional MDS analysis was performed on the pairwise Euclidean distances between all phones, time-averaged over phone duration. Figure 2A visualizes the average distance between the centroids of all phonemes, which shows that the separation between different manners of articulation emerges early in the network. Place of articulation distinctions however are apparent only in the deeper layers, as suggested by Table 1.

To determine whether the network nonlinearity is applied selectively to more confusable phones, we examined the relative location between three example phonemes when they were chosen from either the same or different manners of articulation. We used the MDS analysis (Figure 2B) to visualize the relative location of all individual phones of three phoneme classes across different manners of articulation: /t/ (plosive), /s/ (fricative), and /n/ (nasal), and three phonemes within the same manner of articulation (Figure 2C) (/p,t,k/, plosive). These groups represent varying degree of confusability. In Figure 2B-D, correctly classified phones are shown in colors corresponding to each manner of articulation, phones confused between the three classes are shown in red, and otherwise incorrectly classified phones are shown in gray. The character displayed corresponds to the centroid of the phones (average relative distance between the three classes). Comparison of the relative distances between the phonemes of the two groups shown in Figure 2B versus 2C shows that the network nonlinearities expand the space between the more confusable phonemes (/p,t,k/) more than for phonemes that are more separated in the acoustics (/t,s,n/). This effect can

Layer	All phonemes	Manner		Place		Manner	Place
Features	50.97%	–		–		fricative	labial
HL 1	58.92%	90.55%	**0.0%**	77.69%	**0.0%**	{/f/, /s/, /sh/, /v/, /z/, /zh/}	{/b/, /f/, /m/, /p/, /v/}
HL 2	61.39%	90.99%	**0.49%**	79.87%	**2.81%**	plosive	coronal
HL 3	63.55%	91.92%	**1.51%**	82.56%	**6.27%**	{/b/, /d/, /g/, /k/, /p/, /t/}	{/d/, /t/, /n/, /s/, /sh/, /z/, /zh/}
HL 4	65.43%	92.05%	**1.66%**	83.24%	**7.14%**	nasal	velar
HL 5	66.90%	92.92%	**2.62%**	84.82%	**9.18%**	{/m/, /n/, /ng/}	{/g/, /k/, /ng/}

Table 1: *Frame classification accuracy using softmax regression to decode phonemes and phonetic features (manner of articulation and place of articulation for selected consonants, shown at right) from features and hidden layer activations for TIMIT core test set. For manner and place, percentages in bold show relative classification improvement compared to hidden layer 1.*

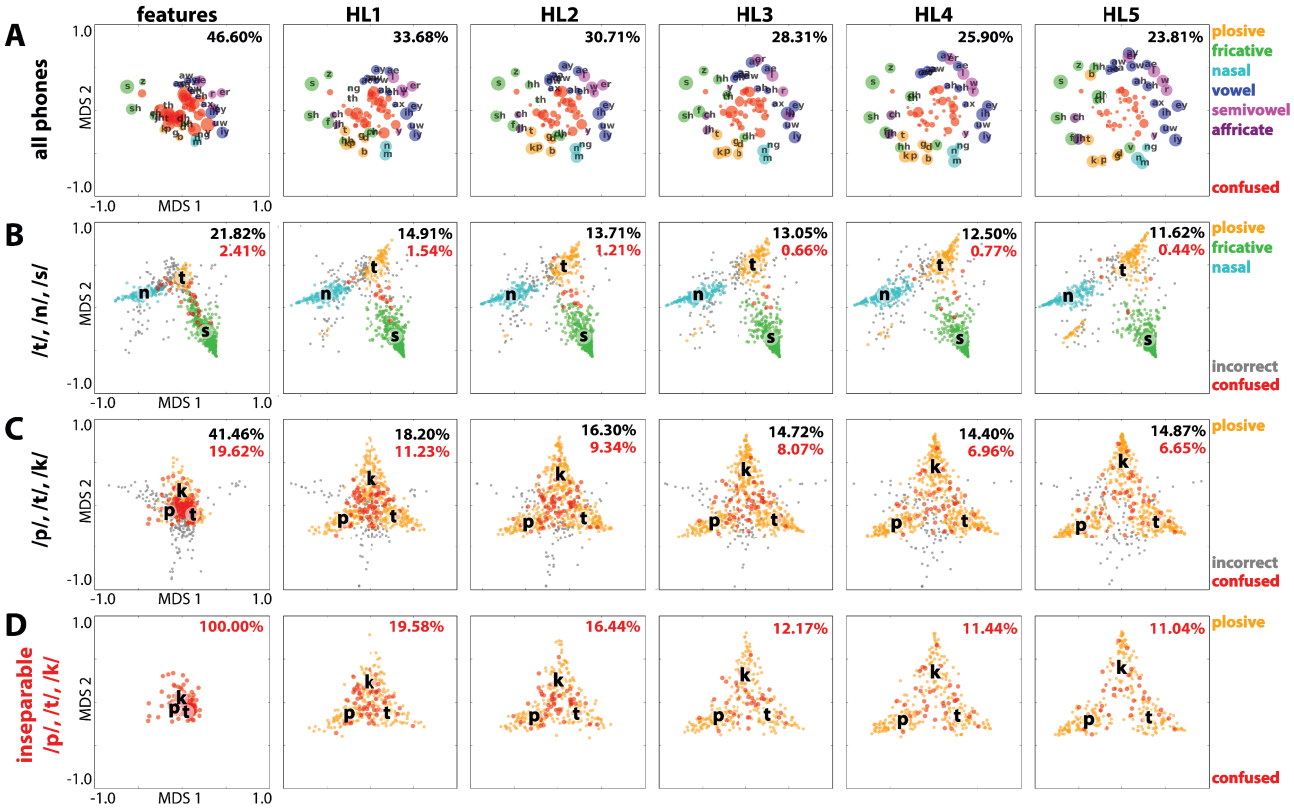

Figure 2: *(A-D) First two MDS dimensions of softmax output for instances of selected phoneme subsets (rows). Columns show decoding from features and all hidden layers of the network. Correctly classified phones are shown by colors corresponding to manner of articulation (right), phones confused with one of the other two in the row heading are shown in red, and otherwise incorrectly classified phones are shown in gray. Overall prediction accuracy for each phoneme subset in is shown in the upper right corner in black, and percentage of phonemes that are confused within the subset are shown in red. Centroids for correctly classified phones are shown in black text with the phoneme label. (A) All phonemes, with colored dots showing the centroid of correctly classified phones and red dots showing centroids of incorrect phones. Area of circle is proportional to number of phones in the category. (B) Phonemes = /t,s,n/, same place of articulation (alveolar). (C) Phonemes = /p,t,k/, same manner of articulation, voicing (unvoiced plosives). (D) Phonemes = /p,t,k/, only inseparable phone instances misclassified (red) in (C), left panel.*

be seen more explicitly in Fig. 3A-B, where the centroids of the phonemes in consecutive layers of the network are overlaid in one figure. This effect is also quantified in Fig. 3D, where the relative expansion of the distance rises much faster in deeper layers for the /p,t,k/ phonemes compare to /t,s,n/.

While this analysis shows that the feature space between more overlapping categories is warped more nonlinearly, we wanted to also investigate whether this non-uniform and selective nonlinear transformation occurs within each phoneme category as well. Toward this goal, we separated the phone instances in the /p,t,k/ group by whether they were correctly classified in the features space, or if they were confused with another class in the subset (Fig. 2C, first column, red samples). Figure 2D tracks the relative location and classification results for these inseparable samples as they propagated through the network. Figure 2D shows a rapid nonlinear increase in the relative distance between these most difficult samples, which occurs more strongly compared to the more separable samples of the same classes (comparison of Fig. 3B to 3C, where the triangles denote the centroid of the corresponding three classes in

each subsequent layer). This observation is also quantitatively confirmed in Fig. 3E, where the expansion of the feature space grows more rapidly in the hidden layers for the samples that were less separable. The observed non-uniform, nonlinear, and focused stretching of the feature space illustrates the power of a multilayer neural network to nonlinearly expand specific parts of the feature space that are critical for discrimination of overlapping categories, while at the same time applying more linear transformations to the parts of the feature space which are less overlapping.

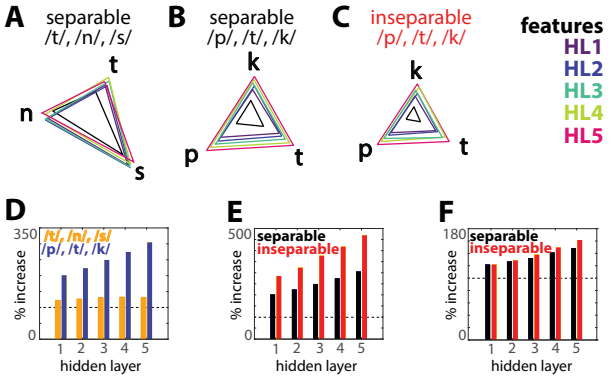

Figure 3: *(A-C) Comparison of centroids of classes (MDS projection) in each hidden layer for subsets: correctly classified /t,n,s/, correct /p,t,k/, and inseparable /p,t,k/ (incorrect in features). (D-F) Relative increase in mean distance between group centroids compared to the feature space for (D) separable /t,n,s/ vs. separable /p,t,k/, (E) separable vs. inseparable /p,t,k/, and (F) separable vs. inseparable phones (all classes).*

3.3. Evidence of hierarchical processing

Although we have demonstrated nonlinear warping of particular regions of the feature space as the signals are mapped from one layer to the next, the exact function of each layer in the formation of phonemic categories remains unanswered. In this section, we tested the hypothesis that the complex boundaries separating phonemic categories are created gradually and in a piecewise manner, where each layer of the network focuses on only a part of the complex boundary. To test this hypothesis, we separated the phones into different subsets, defined by whether they were correctly classified in a particular layer of the network $S_{\text{FEAT,HL1,...,HL5}}$ (Figure 4A). For each hidden layer L, S_L is defined as the subset of phones that were incorrectly classified in all the previous layers, but were correctly classified in layer L. This particular partitioning of the phones allows us to track the separability of each subset from the first to the last hidden layer of the network and thus study their processing in the space created by different layers of the neural network model.

Figure 4B shows classification accuracy for all subsets S in each layer. We observed a particular pattern in the classification accuracies, where the subset that becomes separable in a particular layer stays separable afterward. This result indicates that the subsequent layers of the network focus on different archetypes of problematic phones. Alternatively, one could imagine a scenario where the whole space is transformed from layer to layer, resulting in classification results for the phone subsets that do not necessarily stay correct in subsequent layers. These suggest that each layer of a neural network transforms a specific subset of phones, where the remaining problematic phones are simply passed on to the next layers, where the network can now only focus on creating the required boundary specific to those samples. As a result, the individual layers in a deep neural networks may be utilizing a piecewise approximation strategy to the creation of the nonlinear categorical boundaries, where highly complex partitions can be made layer by layer.

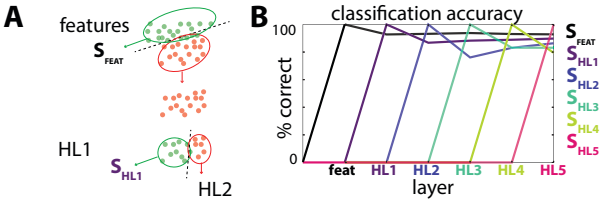

Figure 4: *(A) Schematic showing how phone subsets S are defined. In each subsequent layer L, S_L is the subset of phones that are incorrect $L-1$ and become correct in L. (B) Classification accuracy for phones in subset S_L for all layers L.*

4. Discussion

The goal of this study was to examine the role that multiple layers of nonlinearity play in the acoustic-to-phonetic transformation performed by a DNN trained for phone recognition. We showed that the node activations in the DNN learn a more nonlinear function in deeper layers, and that this nonlinearity is essential for robust phoneme classification. We further demonstrate that these nonlinear transformations, learned through error backpropagation, expand the feature space non-uniformly and selectively in places where the representation of different classes is more similar. Moreover, we observed a piecewise approximation of the categorical boundaries where each layers function is to only correct a subset of phonemes, thereby allowing the subsequent layers to focus only on transforming features that allow discrimination between the remaining confusable instances.

While there have been speculations on the advantages of deep over shallow neural network models, here we provide direct observations on the various types of nonlinear transformations that occur between the hidden layers of a network trained to form invariant categories. Moreover, these findings may provide a more direct link between the complexity of the classification task in hand and the network architecture that is required to model it.

Because of the generality of the principles presented in this paper, these observations are likely to be extended to networks utilizing different types of architecture and nonlinearity, or to categorization tasks other than speech. Moreover, our proposed method allows a direct comparison between various types of nonlinearities (e.g. ReLU networks [19, 20, 21]). We believe that the methods put forth in this study provide valuable insights for better understanding of the strengths and pitfalls of current deep models, as well as motivating strategies for designing models with better generalization power.

5. Acknowledgements

Tasha Nagamine was funded by the From Data to Solutions NSF IGERT grant. This work was funded by a grant from National Institute of Health, NIDCD, DC014279, National Science Foundation CAREER Award, and the Pew Charitable Trusts.

6. References

[1] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 30–42, 2011.

[2] L. Deng, G. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: An overview," in *ICASSP 2013 – IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 26–31, Vancouver, Canada, Proceedings*, 2013, pp. 8599–8603.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, 2012.

[4] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 14–22, 2011.

[5] G. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, pp. 1527–1554, 2006.

[6] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv:1301.3605*, 2013.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intruiging properties of neural networks," *arXiv:1312.6199*, 2013.

[8] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *Techreport*, 2009.

[9] A. Mohamed, G. Hinton, , and G. Penn, "Understanding how deep belief networks perform acoustic modelling," in *ICASSP 2012 – IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 25–30, Kyoto, Japan, Proceedings*, 2012, pp. 4273–4276.

[10] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv:1312.6120*, 2013.

[11] T. Nagamine, M. L. Seltzer, and N. Mesgarani, "Exploring how deep neural networks form phonemic categories," in *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association, September 6–10, Dresden, Germany, Proceedings*, 2015, pp. 1912–1916.

[12] M. D. Zeiler and R. Ferbus, "Visualizing and understanding convolutional networks," *arXiv:1311.2901*, 2013.

[13] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals, and Systems*, vol. 2, pp. 303–314, 1989.

[14] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, pp. 359–366, 1989.

[15] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallet, and N. L. Dahlgren, "TIMIT acoustic-phonetic continuous speech corpus," *LDC93S1*, 1993.

[16] J. Bergstra, O. Breuleux, F. Bastien, P. Lamblin, R. Pascanu, G. Desjardins, J. Turian, D. Warde-Farley, and Y. Bengio, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Jun. 2010, oral Presentation.

[17] P. Ladefoged and K. Johnson, *A Course in Phonetics*. Cengage Learning, 2010.

[18] J. B. Kruskal and M. Wish, *Multidimensional Scaling*. Newbury Park: Sage Publications, 1978.

[19] M. D. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. V. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. E. Hinton, "On rectified linear units for speech processing," in *ICASSP 2013 – IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), May 26–31, Vancouver, Canada, Proceedings*, 2013, pp. 3517–3521.

[20] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *ICML 2013 – 30th International Conference on Machine Learning (ICML), June 16–21, Atlanta, Georgia, Proceedings*, 2013.

[21] T. N. Sainath, B. Kingsbury, A. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for lvcsr," in *ASRU 2013 – Automatic Speech Recognition and Understanding Workshop (ASRU), December 8–12, Olomouc, Czech Republic, Proceedings*, 2013, pp. 315–320.