Interspeech 2018
2-6 September 2018, Hyderabad

Exploring how phone classification neural networks learn phonetic information by visualising and interpreting bottleneck features

Linxue Bai, Philip Weber, Peter Jančovič, Martin Russell

School of Engineering, The University of Birmingham, Birmingham B15 2TT, UK
{lxb190, p.jancovic, m.j.russell}@bham.ac.uk, dr.philip.weber@ieee.org

Abstract

Neural networks have a reputation for being "black boxes", which it has been suggested that techniques from user interface development, and visualisation in particular, could help lift. In this paper, we explore 9-dimensional bottleneck features (BNFs) that have been shown in our earlier work to well represent speech in the context of speech recognition, and 2-dimensional BNFs directly extracted from bottleneck neural networks. The 9-dimensional BNFs obtained from a phone classification neural network are visualised in 2-dimensional spaces using linear discriminant analysis (LDA) and t-distributed stochastic neighbour embedding (t-SNE). The 2-dimensional BNF space is analysed with regard to phonetic features. A back-propagation method is used to create "cardinal" features for each phone under a particular neural network. Both the visualisations of 9-dimensional and 2-dimensional BNFs show distinctions between most phone categories. Particularly, the 2-dimensional BNF space seems to be a union of phonetic category related subspaces that preserve local structures within each subspace where the organisations of phones appear to correspond to phone production mechanisms. By applying LDA to the features of higher dimensional non-bottleneck layers, we observe a triangular pattern which may indicate that silence, friction and voicing are the three main properties learned by the neural networks.

Index Terms: neural network, interpretation, visualisation, bottleneck features, phonetic features, phone classification

1. Introduction

There is a "growing sense that neural networks need to be interpretable to humans" [1]. Understanding the learning behaviour of neural networks and the internal representations they develop has therefore recently received considerable attention, particularly in visual contexts such as image or handwriting recognition (e.g., [2, 3, 4]). Approaches to interpretation focus on visualisation and attribution [1]. For example, dimensionality reduction and visualisation has been applied to a long short-term memory (LSTM) system for handwriting recognition [2], using t-distributed stochastic neighbour embedding (t-SNE) [5] to cluster cell activations and suggest that particular groups of cells work together to predict pen lifts, horizontal and vertical position. Several researchers [3, 4] try to relate the activations of units, to visualisations, to determine the image patches that an individual neuron detects and how they are combined to make a prediction.

There has been less research on interpreting networks for speech processing, where visual interpretations are not so immediate [6]. Mohamed [7] applies t-SNE to show increasing organisation of the structure at deeper layers (e.g., activations for different speakers become closer) and claims that the network is thus implementing something similar to hand-crafted training sequences such as SAT-fMLLR-discriminative training. RNNs are analysed by Tang et al. [6]. While they visualise and compare the distributions of the activations of LSTM and GRU gates, and also the evolution over time of random subsets of hidden units, we relate the behaviour and activations to human models of speech. Some work [8, 9, 10] suggests that deep neural networks (DNNs) learn phonetic structures in acoustic features and treat different broad phone classes differently ([10] shows multilingual bottleneck features seem to learn phonetic information), whereas others [11] argue that DNNs have to be stimulated to learn proper phonetic structures. We show that the networks appear to simultaneously learn multiple "entangled" representations of speech appropriate for different phone classes, similar to the work of Bau and Zhou [12] who aver, for image-detecting convolutional neural networks (CNNs), that "The emergence of interpretable structure suggests that deep networks may be learning disentangled representations spontaneously".

In our previous work [13, 14], we demonstrated that very low-dimensional bottleneck features (BNFs) extracted from phone discrimination bottleneck neural networks contain sufficient information to support high-accuracy phone recognition. Specifically, 9-dimensional (9D) BNFs extracted from a phone discrimination bottleneck neural network provided better ASR phone accuracies than 39-dimensional Mel-frequency cepstral coefficients (MFCCs) in conventional GMM-HMM ASR systems. In [15], we report visualisations and interpretations of 3-dimensional (3D) BNFs and argue that the bottleneck neural networks derive representations specific to particular phonetic categories, with properties similar to those used by human perception. In this paper we extend this research and try to explore how these bottleneck neural networks learn phonetic information. We focus on 9D and 2D BNFs in this paper. We first visualise the 9D BNFs using linear discriminant analysis (LDA) and t-SNE, then narrow the bottleneck layer to 2 nodes to have a direct view of a BNF space. With a set of "cardinal" BNFs, we analyse the organisation of BNFs and interpret it with regard to phonetic features (e.g., [16]). Finally, we use LDA again to visualise the non-bottleneck layers to see how phonetic information is passed from input to output.

2. Methodologies

2.1. Neural network structure used for BNF extraction

We use neural networks having five layers, as in our previous work [13]. A bottleneck layer is used as the second hidden layer to extract BNFs. The neural networks are trained on the TIMIT corpus with TIMIT labels. Logarithm filter-bank energies with context are used as input to the network (26 logFBEs with context of 5 frames before and 5 frames after the current frame in the input layer, i.e., the input layer is of size 286). The

10.21437/Interspeech.2018-2462

output of the neural networks are phone posterior probabilities of the 49 phones [17]. We denote the network structure as 286-H-B-H-49, where ‘B’ stands for the bottleneck layer, ‘H’ other hidden layers.

The neural networks are trained with standard stochastic gradient descent back-propagation using Theano [18, 19].

2.2. Phone categorisation

When interpreting the features, we use the phone categorisation based on [20, 21], which is listed in Table 1.

Table 1: *Phone categorisation used.*

Phone category	Phone label
Plosive	/g/, /d/, /b/, /k/, /t/, /p/
Strong fricative	/s/, /z/, /sh/, /zh/, /ch/, /jh/
Weak fricative	/f/, /v/, /th/, /dh/, /hh/
Nasal/Flap	/m/, /n/, /en/, /ng/, /dx/
Semi-vowel	/l/, /el/, /r/, /w/, /y/
Short vowel	/ih/, /ix/, /ae/, /ah/, /ax/, /eh/, /uh/, /aa/
Long vowel	/iy/, /uw/, /ao/, /er/, /ey/, /ay/, /oy/, /aw/, /ow/
Silence	/sil/, /epi/, /q/, /vcl/, /cl/

2.3. Visualisations of BNFs

We choose linear discriminant analysis (LDA) and t-distributed stochastic neighbour embedding (t-SNE) [22] to visualise 9D BNFs. In the case of LDA, the first and second dimensions of the LDA-based projections are selected and plotted on a 2D graph. The LDA process is a linear supervised process and the projections are learned with the 49 phone labels shown in Table 1. In the case of t-SNE, perplexity and training iterations are set to 50 and 2000, respectively. These are chosen empirically after exploring suitable values. t-SNE is a non-linear unsupervised process and does not use any labels during the training. The phone label information is only used when plotting the 2D mapping space.

To find at the limit what the network is trying to do, we narrowed the bottleneck layer down to 2 nodes so that the BNF space can be visualised straightforwardly. A phone discrimination network of structure 286-512-2-512-49 is trained and 2-dimensional (2D) BNFs extracted. We develop a way to obtain “cardinal” BNFs, as described in Section 2.4.

2.4. Optimised neural activations

We use an approach to obtain the “best” or “cardinal” phone representations under a neural network. In other words, to find what pattern of activation in the hidden layers would be optimal to maximise the probability of the network predicting each phone, given a neural network.

This is done by back-propagating layer activations by keeping the network weights fixed and calculating the derivatives of errors with respect to the layer activations. Assume a trained l-layer neural network with $l-2$ hidden layers between the input and output layers. Let L_m denote the layer m after the input ($m \geq 0$), such that L_0 is the input layer, L_{l-1} is the output layer. We use cross-entropy C as the loss function. The derivative of the error with respect to the activations at arbitrary layer L_m can be derived as

$$\frac{\partial C}{\partial \mathrm{a}_m} = \frac{\partial C}{\partial \mathrm{a}_{l-2}} \frac{\partial \mathrm{a}_{l-2}}{\partial \mathrm{a}_{l-3}} \cdots \frac{\partial \mathrm{a}_{m+1}}{\partial \mathrm{a}_m}, \quad (1)$$

and

$$\begin{aligned} \frac{\partial \mathrm{a}_{m+1}}{\partial \mathrm{a}_m} &= \frac{\partial \mathrm{a}_{m+1}}{\partial \mathrm{o}_{m+1}} \frac{\partial \mathrm{o}_{m+1}}{\partial \mathrm{a}_m} \\ &= \mathrm{a}_{m+1}(\mathbf{1} - \mathrm{a}_{m+1}) W_m, \end{aligned} \quad (2)$$

where a_m is the activations at layer L_m, W_m is the weight matrix between layer L_m and L_{m+1} of the trained neural network and o_m is the linear output at layer L_m.

In our experiments, when calculating the “cardinal” bottleneck layer activations, i.e. BNFs, we first back-propagate to the input layer as a pre-training process (with the maximum epoch being 1000), and then use the bottleneck layer activations resulting from this pre-training as the start point, to apply back-propagation to the bottleneck layer (with the maximum epoch being 100).

3. Experimental Results

3.1. Visualisation of 9D BNFs

Figure 1 shows the 1^{st} and the 2^{nd} dimension of the LDA-based projections of 9D BNFs from a phone classification DNN. The visualisations of the training set and the test set show broadly the same pattern. Therefore to improve the clarity of the visualisations, we plot only 10% of the frames randomly sampled from the training set. Figure 2 shows 2D t-SNE visualisations of the same BNFs as in Figure 1 (10% of TIMIT training data). In both Figure 1 and Figure 2 the plotted points are coloured by their broad phone categories.

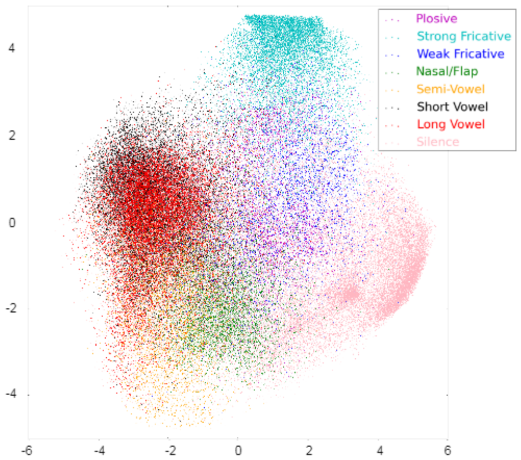

Figure 1: *Visualisations of LDA-based projections (1^{st} vs. 2^{nd} dimension) of 9D BNFs from a phone classification DNN of structure 286-512-9-512-49. Horizontal axis: the 1^{st} dimension of LDA projections; vertical axis: the 2^{nd} dimension of LDA projections.*

From both figures we can see that vowels, consonants and silences are fairly well separated. Also there are many overlaps among the sub-categories of vowels, especially long vowels and short vowels, and among plosive and fricatives, especially plosive and weak fricatives. These overlaps indicate that the overlapping data are alike in some way and may lead to confusions in speech recognition using these features, between the broad phone categories.

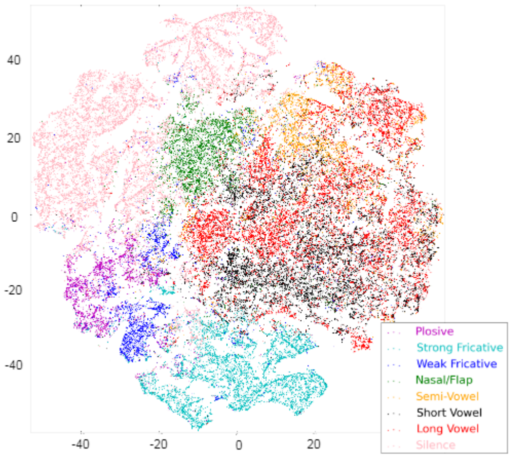

Figure 2: *2D t-SNE visualisations of 9D BNFs from a phone classification DNN of structure 286-512-9-512-49.*

In Figure 1, the 1^{st} LDA dimension (horizontal axis) seems to indicate voicing, with voiced phones on the left and unvoiced on the right. Moving from left to right, we observe vowels, nasals, then fricatives and plosives, and finally silences. In Figure 2, the sizes of clusters and distances between them are not directly related to the size or importance of clusters in the original high-dimensional space, due to the properties of the t-SNE algorithm.

3.2. Visualisation of 2D BNFs and analysis of 2D BNF space

The plot of the 2D BNFs is shown in Figure 3. To keep consistency with the plots in the previous sections, we again plot a random 10% sample of the TIMIT training set. The definition of phone categories and colours are the same as used in Section 3.1.

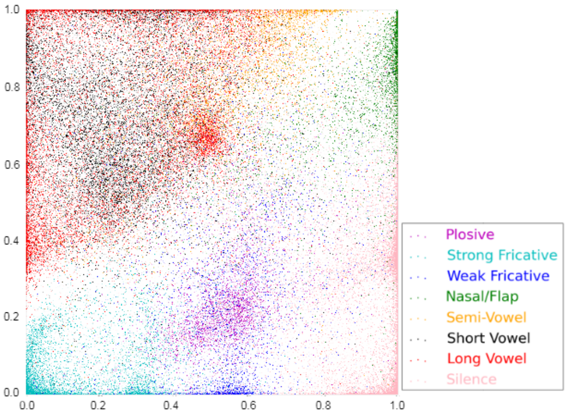

Figure 3: *2D BNFs from a phone classification DNN of structure 286-512-2-512-49.*

From Figure 3 we can see fairly clear organisations of phone categories: vowels (red, black, and orange) are distributed at the left top half, nasals (green) at the right top corner, strong fricatives (cyan) at the lower left, plosive (purple) somewhere at lower middle, some weak fricatives (blue) mixing up with plosive and some at the mid lower edge, and silence takes the right lower part of the figure. The BNFs are constrained within the range of [0,1], due to the "squashing effect" of the sigmoid function. "Concentrated" edges along the four sides of the square appear to indicate "hard" or "certain" decisions made by the sigmoid for those BNFs.

Using the method described in Section 2.4, we obtain 49 2D BNF vectors representing the 49 phones for the DNN. We plot them in Figure 4 (solid dots). We also plot the centroids of the 2D BNFs (i.e. feature means) of each phone in the training set in Figure 4 using open circles. In the figure we link every pair of dot and circle points that correspond to the same phone for a clearer view.

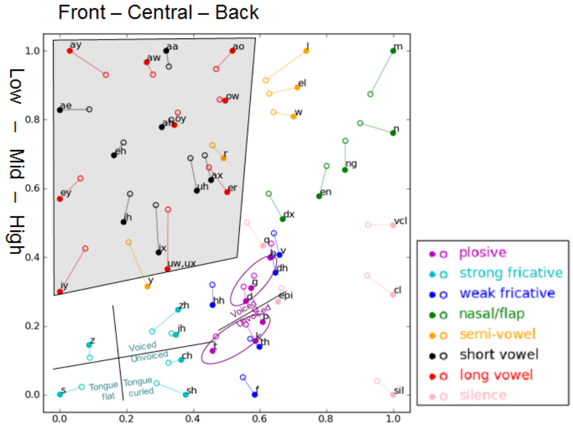

Figure 4: *Optimised 2D BNFs (dots) and feature means of 2D BNFs (circles) for each phone for a phone classification DNN of structure 286-512-2-512-49.*

Most of the "cardinal" features (dot points) are close to the corresponding centroids obtained from the data (circles), and the organisations of both are similar. The direction from top left to bottom right seems to indicate voicing, with vowels distributed at the top left half of the graph, and silences being at the right bottom corner. Compared to the feature means (circles), for the "cardinal" features (dots) the various categories seem to be pushed to the edges of a local space. The reason may be that "cardinal" features are trained to provide more certain phone decisions than random BNFs, which forces the hidden layer to make harder decisions. The edges of these local spaces also seem to be hinted at in Figure 3 by the denser congregations of points.

We now focus on long vowels and short vowels (shaded area) in Figure 4. The arrangement of centroids and cardinal features strongly resembles a "traditional" F1:F2 vowel space diagram used by phoneticians. Vertically from top to bottom, we observe /ay/, /ey/, /iy/ (left side) and /ao/, /uh/, /uw/ (right side) – roughly corresponding to the places of articulation (tongue positions) from low to high. Horizontally from left to right, we observe /ey/, /ah/, /ow/ – roughly front to back with respect to place of articulation.

Strong fricatives are displayed in cyan. /s/ and /z/, produced with a flat tongue, are distributed at the left bottom corner. /zh/, /jh/, /ch/ and /sh/, produced with the tip of the tongue curled up, are distributed in mid-lower region. Strong fricatives at the top (/z/, /zh/, /jh/) are voiced, and those at the bottom (/s/, /sh/, /ch/) are unvoiced. A similar pattern is seen for the plosives (in purple) – voiced at the top (/d/, /g/, /b/) and unvoiced at the bottom (/t/, /k/, /p/). For both voiced and unvoiced plosives, phones are placed horizontally in an order that reflects their place of articulation (from left to right: teeth, soft palate and lips).

The 2D BNF space therefore can be seen to show distinct regions used for each phonetic category. Within each category, the organisation of phones appears to correspond to phone production mechanisms. However, the interpretation of the axes of one phone category cannot be simply applied to other categories, and so the BNF space seems to be a union of phonetic category related subspaces that preserve local structures within each subspace.

3.3. Visualisation of non-bottleneck layer activations

We next apply LDA to non-bottleneck layers to investigate what information is carried through these layers. We use the same DNN as in Section 3.1. The DNN hidden layer size was 512-9-512, thus we now visualise the 1^{st} and the 2^{nd} dimensions of the LDA projection for the two 512-node layers.

Figure 5 shows the 1^{st} and the 2^{nd} dimension of the LDA-based projections of the activations of the first hidden layer, plotted on the 10% of the training set like before. The same process is applied to the 3^{rd} hidden layer, giving Figure 6.

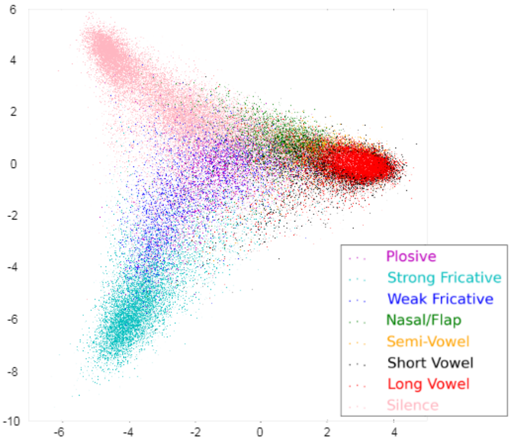

Figure 5: *Visualisation of LDA-based projections (1^{st} vs. 2^{nd} dimension) of the 1^{st} hidden layer activations from a phone classification DNN of structure 286-512-9-512-49. Horizontal axis: the 1^{st} dimension of LDA projections; vertical axis: the 2^{nd} dimension of LDA projections.*

Both figures show a clear "triangular" shape with similar structures, where vowels, strong fricatives and silences each occupy a corner of the triangle. Along the horizontal axis, from left to right, we see silence and fricatives first and then vowels, which could be interpreted as transitioning from unvoiced to voiced, or as the energy in low frequency bands increasing; Along the vertical axis, from upper to lower, we see silence first, and then vowels, finally fricatives – this could be interpreted as energy in high frequency bands increasing. As the horizontal and vertical axes correspond to the first two dimensions of LDA, these interpretations may indicate that the energies in low and high frequency bands are two main pieces of information learned by the DNN. An alternative interpretation of this triangular shape could be that there is some inherent 3-dimensional structure in the high dimensional data, corresponding to 3 properties of phones: silence, frication and voicing.

Comparing Figures 5 and 6, we can see the triangular plot of the third hidden layer is similar but in sharper focus than that of the first hidden layer. As the interpretation of the input progresses through the network, from input to output, the phonetic categories become more specific, seeming to confirm the predictions of deep learning behaviour [23] and prior interpretations of speech recognition networks [7].

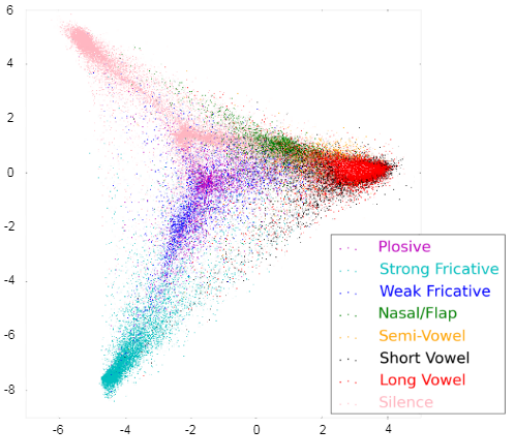

Figure 6: *Visualisations of LDA-based projections (1^{st} vs. 2^{nd} dimension) of the 3^{rd} hidden layer activations from a phone classification DNN of structure 286-512-9-512-49. Horizontal axis: the 1^{st} dimension of LDA projections; vertical axis: the 2^{nd} dimension of LDA projections.*

We find that this triangular visualisation of the 1^{st} and the 2^{nd} dimension of the LDA-based projections is always observed when analysing "bigger" hidden layers (wider than about 30 nodes) within DNNs of a similar 5-layer structure.

4. Discussions and Conclusions

The visualisations of BNFs indicate that the strategy of a phone classification neural network can be interpreted in terms of phonetic categories. For the 9D BNFs, in both LDA and t-SNE experiments we observed phonetically meaningful clusters in the projected 2D spaces. The 2D BNF analysis even suggests that the BNFs can be interpreted in terms of phonetic features, i.e., the organisations of phones in the BNF space appear to correspond to phone production mechanisms.

By visualising non-bottleneck layer activations, we found that as features move through the network, from input to output, phonetic categories become more specific. This is consistent with the observations by Hinton et al. in the deep learning experiment of image processing for digit hand-writing recognition. Also the triangular pattern in the first two dimensions of LDA projection indicates that silence, friction and voicing are the three main properties that are learned by the neural networks.

It holds promise for many research topics when recognising that the internal representations learned by networks for speech recognition can be related to knowledge of human speech structure. For example, to use phonetic knowledge to improve DNN performance (like in [6]), and to use visualisation of DNN structure to gain phonetic insights. Interpretable visualisation may also be used for pronunciation training.

5. References

[1] C. Olah, A. Mordvintsev, and L. Schubert, “Feature visualization,” *Distill*, 2017, https://distill.pub/2017/feature-visualization.

[2] S. Carter, D. Ha, I. Johnson, and C. Olah, “Experiments in handwriting with a neural network,” *Distill*, 2016. [Online]. Available: http://distill.pub/2016/handwriting

[3] M. D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[4] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, “The building blocks of interpretability,” *Distill*, 2018, https://distill.pub/2018/building-blocks.

[5] L. van der Maaten, “Accelerating t-sne using tree-based algorithms,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 3221–3245, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2697068

[6] Z. Tang, Y. Shi, D. Wang, Y. Feng, and S. Zhang, “Memory visualization for gated recurrent neural networks in speech recognition,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2017, New Orleans, LA, USA, March 5-9, 2017*. IEEE, 2017, pp. 2736–2740. [Online]. Available: https://doi.org/10.1109/ICASSP.2017.7952654

[7] A. Mohamed, G. E. Hinton, and G. Penn, “Understanding how deep belief networks perform acoustic modelling,” in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2012, Kyoto, Japan, March 25-30, 2012*. IEEE, 2012, pp. 4273–4276. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2012.6288863

[8] T. Nagamine, M. L. Seltzer, and M. N, “Exploring how deep neural networks form phonemic categories,” in *Interspeech*, 2015, pp. 1912–1916.

[9] T. Nagamine, M. L. Seltzer, and N. Mesgarani, “On the role of nonlinear transformations in deep neural network acoustic models,” in *Interspeech 2016*, 2016, pp. 803–807. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-1406

[10] N. T. Vu, J. Weiner, and T. Schultz, “Investigating the learning effect of multilingual bottle-neck features for asr,” in *Proc. Interspeech,* Singapore, 2014.

[11] S. Tan, K. C. Sim, and M. Gales, “Improving the interpretability of deep neural networks with stimulated learning,” in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*. IEEE, 2015, pp. 617–623.

[12] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, “Network dissection: Quantifying interpretability of deep visual representations,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 3319–3327. [Online]. Available: https://doi.org/10.1109/CVPR.2017.354

[13] L. Bai, P. Jančovič, M. Russell, and P. Weber, “Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics,” in *Proc. Interspeech,* Dresden, Germany, 2015, pp. 583–587.

[14] P. Weber, L. Bai, S. Houghton, P. Jančovič, and M. Russell, “Progress on phoneme recognition with a Continuous-State HMM,” in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP),* New Orleans, Louisiana, USA, 2016.

[15] P. Weber, L. Bai, M. Russell, P. Jančovič, and S. Houghton, “Interpretation of low dimensional neural network bottleneck features in terms of human perception and production,” in *Proc. Interspeech,* San Francisco, CA, USA, 2016, pp. 3384–3388. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2016-124

[16] K. N. Stevens, “Toward a model for lexical access based on acoustic landmarks and distinctive features,” *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.

[17] W. J. Holmes, “Modelling segmental variability for automatic speech recognition,” Ph.D. dissertation, University of London, 1997.

[18] F. Bastien, P. Lamblin, R. Pascanu, J. Bergstra, I. J. Goodfellow, A. Bergeron, N. Bouchard, and Y. Bengio, “Theano: new features and speed improvements,” *Deep Learning and Unsupervised Feature Learning NIPS Workshop*, 2012.

[19] 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 *Proc. of the Python for Scientific Computing Conference (SciPy)*, June 2010.

[20] A. K. Halberstadt and J. R. Glass, “Heterogeneous acoustic measurements for phonetic classification 1,” in *Proc. Eurospeech ’93,* Berlin, Germany, 1997, pp. 401–404.

[21] H. Huang, Y. Liu, L. ten Bosch, B. Cranena, and L. Boves, “Locally learning heterogeneous manifolds for phonetic classification,” *Computer Speech and Language*, vol. 38, pp. 28–45, 2016.

[22] L. van der Maaten and G. Hinton, “Visualizing high-dimensional data using t-SNE,” *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[23] G. E. Hinton, “Training products of experts by minimizing contrastive divergence,” *Training*, vol. 14, no. 8, 2006.