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

Interpretation of Low Dimensional Neural Network Bottleneck Features in Terms of Human Perception and Production

Philip Weber, Linxue Bai, Martin Russell, Peter Jančovič and Stephen Houghton

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

Abstract

Low-dimensional 'bottleneck' features extracted from neural networks have been shown to give phoneme recognition accuracy similar to that obtained with higher-dimensional MFCCs, using GMM-HMM models. Such features have also been shown to preserve well the assumptions of speech trajectory dynamics made by dynamic models of speech such as Continuous-State HMMs. However, little is understood about how networks derive these features and how and whether they can be interpreted in terms of human speech perception and production.

We analyse three-dimensional bottleneck features. We show that for vowels, their spatial representation is very close to the familiar F_1:F_2 vowel quadrilateral. For other classes of phonemes the features can similarly be related to phonetic and acoustic spatial representations presented in the literature. This suggests that these networks derive representations specific to particular phonetic categories, with properties similar to those used by human perception. The representation of the full set of phonemes in the bottleneck space is consistent with a hypothesized comprehensive model of speech perception and also with models of speech perception such as prototype theory.

Index Terms: Neural Networks, Human Speech Perception, Recognition, Segmental Models, Bottleneck Features

1. Introduction

Significant progress has been made in automatic speech recognition using large statistical models [1, 2], but at the expense of interpretable models. It is very difficult to relate a model with many thousands of parameters to the relatively small number of moving parts of the human vocal tract involved in producing speech. Current mainstream models for speech recognition also largely ignore the dynamics of speech and the human articulators, i.e. relatively slowly changing acoustics and features strongly correlated in time [3].

Various authors have postulated that speech lies on low-dimensional manifolds embedded in high-dimensional space [4, 5]. Vowels are well distinguished phonetically by comparing tongue 'backwardness' with height, represented by the ideal 'vowel quadrilateral' phonetic space [6]. A similar spatial representation is obtained when vowel sounds are analysed either acoustically (the F_1:F_2 'vowel space diagram') or in terms of human perception [7, 8]. Two-dimensional representations cannot capture the full range of vowel quality, but Pols et al. showed that 81.6% of the variance in the acoustics could be explained by just 3 dimensions, rising to 94.0% with 5. Other speech sounds have been less investigated in this manner, but for fricatives [9, 10] just 2 dimensions have been found to account for over 95% of the variance. The interpretation of the dimensions was less clear, but again very similar spaces were found by phonetic, acoustic and perceptual analyses.

Segmental [11, 12, 3] and Continuous-State HMM [13] models of speech aim to be faithful to the true nature and dynamics of speech. Recognition accuracies have been hampered by poor ability to model all the variation found for example in formants, due to speaker differences, co-articulation, and so on. Further, the analyses reported above suggest that different classes of phonemes lie on distinct low-dimensional spaces, and thus might be best modelled by different types of features (e.g. formants for vowels [14], spectral energies for consonants [15]).

Switching feature representations during decode is not trivial, and also some potentially useful features such as formants are difficult to estimate [16]. In previous work [17, 20] we explored using neural networks to automatically derive low-dimensional 'bottleneck' feature (BNF) representations of speech. With 9-dimensional (9d) features we obtained 29.4% phoneme recognition error on TIMIT [18], using a standard monophone GMM-HMM recogniser [19], significantly less than the 49.9% using formants plus time derivatives (also 9d).

However, from aiming at models and features which can be related to knowledge of human production and perception of speech, we ended up with features which do not seem to be so interpretable. What do BNFs represent? Can they be related to human perception or production of speech? Also, how does the network learn them, and if speech is inherently low-dimensional, why is a network with many parameters needed to generate a low-dimensional representation adequate for automatic recognition? If we can relate the 'blind learning' from data by the network to human speech recognition, we would gain confidence that the network has optimised for speech rather than some spurious data characteristics.

In the following we begin to answer some of these questions, using a spatial analysis of three-dimensional BNFs. For vowels, we show surprisingly close correspondence with familiar vowel-space diagrams, and also find similar correspondences for other classes of phonemes. We conclude by considering these findings in the light of hypothesized comprehensive models of speech perception [9, 21], also considering models of speech perception such as Prototype theory [22].

2. Background

We first describe bottleneck features, TIMIT recognition results, and measures for comparing feature spaces.

2.1. Bottleneck Features (BNFs)

Bottleneck features are obtained from the activations of neurons in a narrow (3 to 9 neuron) layer in 5-layer neural networks, trained using Theano [23]. The majority of these networks were multi-layer perceptrons (MLPs) with the bottleneck in the cen-

http://dx.doi.org/10.21437/Interspeech.2016-124

Network	Autoencoder		MLP		DBN	
Dim./Layer	3d/3	9d/3	3d/3	9d/3	3d/4	9d/4
all phonemes	64.8	60.8	47.8	37.9	43.8	35.2
voiced only	72.7	65.6	50.5	43.8	47.7	40.3
unvoiced	40.5	37.7	30.2	20.6	26.6	19.9

Table 1: CS-HMM phone recognition % Error using BNFs from 5 layer Autoencoder and classifier networks.

tral (third) layer, trained discriminatively on the TIMIT training set to predict posterior probabilities for 49 phonemes [24]. Inputs were 11-window log Mel-frequency filterbanks. Full details may be found in Bai et al. [20]. We also trained two other types of networks: firstly, Autoencoder networks with the same structure, where the task of the network was to reconstruct the input features; secondly, 5-layer classifier networks with the bottleneck in the final (fourth) layer before the Softmax output. Although the latter were pre-trained as deep belief networks, then fine-tuned discriminatively, recognition experiments showed that the effect of the changed training methodology was negligible compared with the effect of moving the bottleneck layer deeper in the network.

2.2. CS-HMM Phoneme Recognition using BNFs

The CS-HMM is a 'parsimonious' model of speech which aims to reflect speech structure and dynamics more faithfully than conventional large statistical models. The results in Table 1 were obtained with just 535 parameters in the recogniser for the experiments using 9-dimensional (9d) features and the full phoneme set. The model assumes features which embody the smooth, constrained movement of the human articulators, such as formants [25, 26] or low-dimensional BNFs [17]. The trajectories of these features during speech are recovered using a continuous state which describes a distribution over the current feature values, given the observations seen and estimated trajectory and phonetic history. The space of possible trajectories is explored through branching and pruning a set of hypotheses. Full details may be found elsewhere [13, 14, 26].

Table 1 shows phoneme recognition error rates using the CS-HMM with 9d and 3d BNFs from various networks. These error rates are significantly lower [17] than obtained with formants estimated using WaveSurfer [27] (73.7% for the full phoneme set). This suggests that the network training removed much of the variability from the features not needed to predict phonemes. Features from multiple random initialisations of the networks are different, but give very similar results.

Notably, BNFs derived from the Autoencoder (AE-BNFs) give considerably higher recognition error than those from the classifiers. Intuitively, AE-BNFs retain more unwanted variation for the CS-HMM to deal with (using few parameters), since the Autoencoder must compress as much as possible of the information in the input features into the BNFs in order to reconstruct its input. The classifier is free to remove any variation unnecessary for predicting phonemes. The features derived from the bottleneck in layer 4 give slightly better accuracy than those from layer 3. This is again as expected, since deeper layers progressively remove more unwanted variation [28].

2.3. Metrics to Compare Phonetic Spaces

We represent features spatially using two-dimensional plots showing the centroids of clusters of phoneme realisations for a given set of n phonemes $\Phi = \phi_1 \dots \phi_n$ (e.g. vowels, Fig. 1). Let X, Y be $n \times 2$ matrices giving the coordinates of points $\boldsymbol{x}_i = (x_{i1}, x_{i2})$, $\boldsymbol{y}_i = (y_{i1}, y_{i2})$ in two such plots ($1 \le i \le n$). X and Y are assumed ordered according to Φ.

We want to compare the BNF space with, for example, formant space. Since the only constraint on training BNFs is for them to lie in $[0, 1]$, we consider the shape $\mathcal{D}_x$ described by connecting points $\boldsymbol{x}_i$ to be significant, but not its location or rotation in BNF space. We define two distance measures:

1. $d_2(X, Y)$ is the Euclidean distance between points $\boldsymbol{y}_i$ and the $\hat{\boldsymbol{y}}_i$ found by attempting to affine transform X to Y,

$$\begin{aligned} \hat{Y} = &\quad AX, \text{ where } A = YX^{-1}, \text{ and} \\ d_2(X, Y) = &\quad \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \sqrt{\sum_{j=1}^{2} (y_{ij} - \hat{y}_{ij})^2}. \end{aligned} \tag{1}$$

2. $d_s(X, Y)$ allows greater shape deformation by comparing matching edges and angles of $\mathcal{D}_x$ and $\mathcal{D}_y$ (*cf* [21]). Let vector $\boldsymbol{e}_{i+}^{(x)}$ be the edge from $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i+1}$, $\boldsymbol{e}_{i-}^{(x)}$ from $\boldsymbol{x}_i$ to $\boldsymbol{x}_{i-1}$, and $\theta_i^{(x)}$ the angle between $\boldsymbol{e}_{i+}^{(x)}$ and $\boldsymbol{e}_{i-}^{(x)}$ (let $n + 1 \triangleq 1$).

$$\begin{aligned} d_s(X, Y) = &\quad \tfrac{1}{2}\big(d_e(X, Y) + d_a(X, Y)\big), \text{ where} \\ d_e(X, Y) = &\quad \tfrac{1}{n\sqrt{2}} \textstyle\sum_{i=1}^{n} |l_i^{(x)} - l_i^{(y)}|, \text{ for} \\ l_i^{(x)} = &\quad \sqrt{\textstyle\sum_{j=1}^{2} (x_{ij} - x_{(i+1)j})^2}. \\ d_a(X, Y) = &\quad \tfrac{1}{2n} \textstyle\sum_{i=1}^{n} |\cos\theta_i^{(x)} - \cos\theta_i^{(y)}|. \end{aligned} \tag{2}$$

($l_{i+}^{(x)}$ is the length of vector $\boldsymbol{e}_{i+}^{(x)}$, and the normalising constants ensure that $d_s(X, Y) \in [0, 1]$).

3. Representations in BNF Space

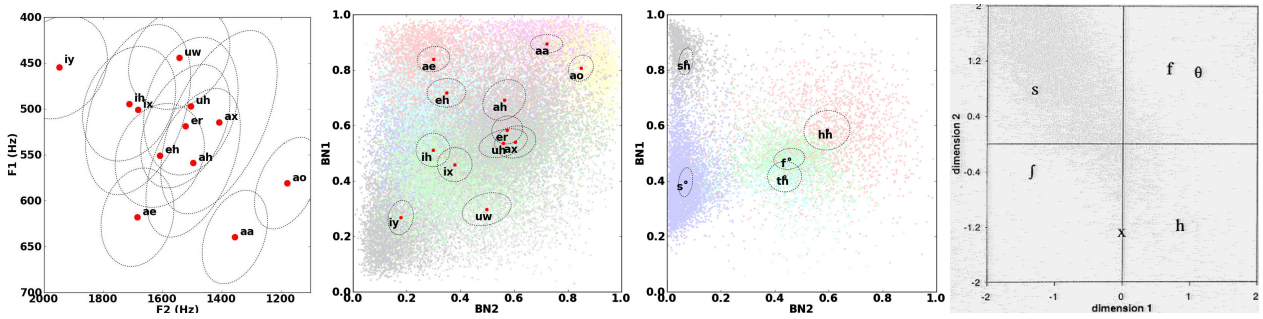
In this section we investigate and relate the representations in BNF space learned for vowels and for fricatives.

3.1. BNF Representation of Vowels

Fig. 1(a) shows F_1 plotted against F_2 for formants estimated using WaveSurfer [27], for TIMIT. For each instance of each vowel, the average feature was calculated between the phoneme boundaries given by the TIMIT transcriptions. The figure shows the centroids and ellipses indicating 0.5 standard deviation of the instances for each phoneme. The structure of the space is as expected from the literature (e.g. [30]), but there is considerable overlap between phonemes.

In a corollary with the first three formants F_1, F_2, F_3, we plot two-dimensional spatial representations of instantiations of vowels for the three pairs of 3d BNFs $(1, 2), (1, 3), (2, 3)$ from several initialisations of the networks reported in Table 1. For each initialisation, one of the pairs corresponds strikingly well with the vowel space diagram. An example is shown in Fig. 1(b). This shows that the network has learned by itself to identify a set of parameters that have similar properties to formant frequencies, and hence can be related to tongue position, and that these features may be optimal for discriminating between vowels. It is remarkable that these features should be so close to those known for the human auditory space.

Comparing Figs. 1(a) and 1(b), we see that this BNF space is inverted and rotated in comparison with the vowel space diagram. This is not significant, because network training places no constraints on the learned feature space. But the structure, or 'shape' defined by the centroids, is significant, as it is indicative of the structure of the underlying acoustic space.



(a) Formants (WaveSurfer [27]). (b) BNF MLP 3d/3 Vowels. (c) BNF MLP 3d/3 Fricatives. (d) Fricatives from [10].

Figure 1: (a,b) TIMIT vowels showing correspondence between (a) formants, (b) one BNF pair from MLP 3d/3 (Table 1). (c) BNF fricatives, corresponding to (d) [10]. Each cluster is shown by its centroid and an ellipse showing 0.5 standard deviation.

		Best Matching			Avg. Worst	
Feature	**Pair**	**Formant**	$\mathbf{d_2}$	$\mathbf{d_s}$	$\mathbf{d_2}$	$\mathbf{d_s}$
MLP 3d/3 a	$(1,3)$	F_1:F_2	0.142	0.147	0.278	0.201
MLP 3d/3 b	$(1,2)$	$\mathbf{F_1}$:$\mathbf{F_3}$	0.195	0.139	0.336	0.224
MLP 3d/3 c	$(1,2)$	F_1:F_2	0.216	0.139	0.330	0.206
MLP 3d/3 d	$(1,2)$	F_1:F_2	0.119	0.153	0.339	0.231
DBN 3d/4	$(2,3)$	F_1:F_2	0.176	0.155	0.363	0.226
A/E 3d/3	$(2,3)$	F_1:F_2	0.192	0.187	0.388	0.250
MFCC	*In-conclusive* (mean $d_2 = 0.348, d_s = 0.217$)					

Table 2: Distances between BNF and formant spaces.

The BNF clusters have lower variance and less overlap than their formant space counterparts, and make fuller use of the available space compared with the theoretical frequency space for formants, which is limited by the human vocal tract[1]. While preserving the vowel space layout, the network has at the same time removed much of the variation unimportant for its task of predicting phoneme outputs from input features.

In Table 2 we report distances (Section 2.3) between the formant space and the matching BNF pairs, together with average distance between the formant space and non-matching BNFs. For comparison, the average distance between two sets of 'vowel space' BNFs is d_2=0.065, d_s=0.041 (non-matching: d_2=0.243, d_s=0.192), and average distance from the BNFs to a set of random points in $[0,1]^2$ is $d_2 = 0.431$, $d_s = 0.282$. The metrics (and visualisation) highlight a single anomaly, for network 'c', for which phoneme /er/ was apparently out of place. In this case it was found that the network had learned an alternative representation, matching F_1:F_3 more closely, which gives a plot similar to F_1:F_2 with /er/ moved (for American English).

Features from other networks are rotated differently, and may not be inverted, but show the same structure. In Fig. 2(a) we show the results of using affine transformations to align the matching BNF pair and formant spaces (in calculating d_2 (Eq. 1)) with manually-selected centroids of vowel F_1:F_2 clusters reported by Hawkins [30]. Evidently the various representations preserve the vowel space structure, but are not identical. The networks may have learned only local optima, albeit good ones (by the recognition rate). Alternatively three dimensions may not be enough and each dimension accounts for more than a single underlying feature. Plots of the remaining two pairs of BNFs exhibit some of the 'vowel space' structure, but compressed into part of the space. This explains why the values in columns 6 and 7 of Table 2 are not higher. It seems likely that the third dimension is more relevant to consonants. We examine this further in the next section.

Visualisations of BNFs from Autoencoders (AE-BNFs) and from bottlenecks in layer 4 (DBN 3d/4), confirm that in both cases the vowel space structure is recovered. The intuitions mentioned in Section 2.2 to explain the variations in phoneme recognition error are also confirmed. For the AE-BNFs, the variance in each cluster is much larger, whereas for BNFs from the deeper bottleneck layer, variance is slightly lower and the clusters are also slightly better separated. The distances in Table 2 do not show these differences. This is likely because they take no account of cluster variance, and suggests that alternative measures could be found to relate 'goodness' of clustering with recognition performance.

When all pairs of 9d BNFs are plotted in the same way, some are again similar to F_1:F_2, but it is not clear which pair matches best, either visually or from the distances. This may indicate that some dimensions are superfluous, so information is duplicated between features, or that 9 dimensions allows the network to create a finer analysis than possible in 3d. Further work is needed to understand and interpret the optimal number of dimensions to represent vowels, and to find more discriminating distance measures. Finally, similar plots for the first three of 13 MFCCs do not match the F_1:F_2 structure in the same way, confirmed by no match being identified using distances.

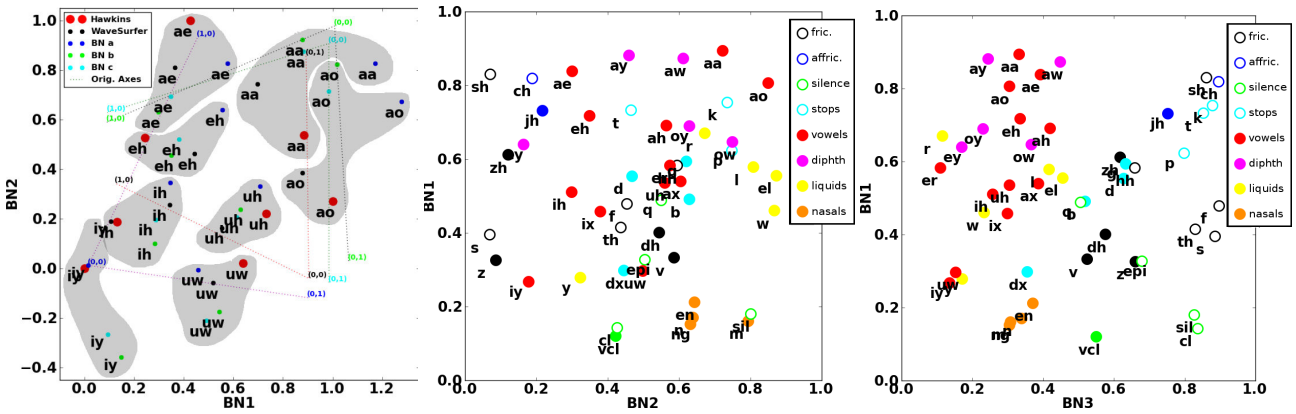
3.2. BNF Representation of Fricatives

Perceptual, acoustic and articulatory spaces have been less investigated for consonants, but the ideas for vowels have been successfully extended to fricatives [10], showing strong correspondence between, acoustic, phonetic and production spaces for fricatives /s/, /sh/, /f/, /th/ and /hh/ (Fig. 1(d)). The BNF space for fricatives is similar (Fig. 1(c)). As for vowels, the network has found a representation which is interpretable in terms of human perception and production.

The BNF space is more distorted compared with the literature reference, than we saw for vowels, and comparison of the spaces using distance measures and transformations is not so convincing. This may again suggest that three dimensions is inadequate, a view reinforced by reading the literature on human perception of consonants and our previous work [15]. Or it may simply be due to differences in speaker cohorts between TIMIT and the study from which Fig. 1(d) was taken [10].

The study proposes (but does not quantify) that the axes be interpreted as sibilance and place of articulation. Previous stud-

[1]The clusters are not Gaussian as suggested by Fig. 1(b), but rather the networks seek to make full use of the $[0,1]$ BNF space (*cf* Fig. 1(c)).



(a) Aligned BNFs and formants. (b) BNF $(1, 2)$ MLP 3d/3. (c) BNF $(1, 3)$ MLP 3d/3.

Figure 2: (a) Centroids (identified by hand) of formant vowel clusters identified by Hawkins [30] (British English), formants (from WaveSurfer) and BNFs from three network initialisations, optimally aligned, (b,c) centroids for two pairs of BNFs, for all phonemes in the BNF (MLP 3d/3) space; phonetic categories are colour-coded, voiced phonemes indicated by solid circles, unvoiced by open.

ies [31] have shown that humans distinguish fricatives by broadband energy in specific frequencies. We could therefore hypothesise a link between characteristic broadband noise frequencies for fricatives, and formant frequencies for vowels, to explain their co-existence within the same set of low-dimensional bottleneck features. We discuss this further in the next section.

3.3. A Unified Bottleneck Feature Space

The BNFs must represent all speech sounds in a single space, including vowels, naturally represented by features with continuous trajectory dynamics [25]; and unvoiced consonants, characterised by broadband noise in specific frequency bands and piecewise constant [33, 31] dynamics. Voiced consonants combine features of both, so we may anticipate being able to make some interpretations of the space and to find some overlap in representations, but given the different temporal dynamics and longer-term temporal effects [34, 35] it is not obvious how a single space could describe all sounds. However, for fricatives it was found that "auditory processing in the fricative data was adequately modelled by the auditory transformations used in the vowel data" [10], and other studies have mapped vowels and consonants to a single 3d space [21].

Figs. 2(b) and 2(c) show plots of BNF MLP 3d/3 pairs $(1, 2)$ and $(1, 3)$ for one network, for all TIMIT phonemes. Phonetic classes are colour-coded, voiced phonemes shown by solid circles, unvoiced by open. Fig. 2(b) corresponds to 1(b) so the same vowel space structure can be seen. Other structure can be seen: Firstly, the lower right corner is free of vowels, but occupied by nasals. This is is understandable; one characteristic of nasals is lowered formants [36]. Secondly, clustering according to manner of articulation, such as liquids (/w/ close to the diphthong /ow/, /y/ to /iy/), and various groups of fricatives.

Considering how these phonemes are articulated suggests that the BNF dimensions in Fig. 2(b) may relate more generally to tongue position. This explains co-located phonemes in the figure. Within phoneme groups such as {/s/, /iy/, /z/}, {/zh/, /sh/}, {/w/, /l/, /ao/}, the vocal tract is to some extent in a similar configuration. Therefore the location in some dimensions of acoustic space, and thus bottleneck space, may be similar. In Fig. 2(b), vowels and voiced consonants appear in the left half, unvoiced in the right, suggesting that the third feature correlates strongly with voicing. /s/ and /y/ are now far apart, /s/ and /z/ less so, reflecting differences in strength of voicing. Finally, plots (not shown) of the third pair $(2, 3)$ show /s/ and /sh/ very close, reflecting their similar frication characteristics.

4. Discussion

So the way the 3d BNF spaces map all phonemes is somewhat interpretable, but how does this relate to other acoustic and phonetic characteristics identified as critical for human perception? For example, static features such as voicing, aspiration and total energy; and dynamics such as phoneme duration, formant transitions [37], and correlations between phonemes [35, 38]?

One argument is that this evidence supports the Prototype theory of speech perception [22, 9], which says that speech segments are identified according to their perceived distance from 'prototypes' in perceptual space. Human generation of speech sounds is correspondingly according to articulatory targets (*cf* [25]). This seems to contrast with perception according to complex perceptual cues. However three dimensions is very restrictive, and our networks were simple and had little exposure to speech dynamics, seeing only a window of 11 frames. We cannot on this basis make any judgements about perceptual theory.

It would therefore be interesting to extend this work to analyse low-dimension features produced by recurrent architectures such as RNNs, which are able to learn dynamics over time, to perhaps gain insights into why these networks perform better for speech recognition. Such analysis is also of interest to try to encourage bottleneck features closer to the assumptions of speech dynamics made by segmental and CS-HMM models.

5. Conclusion

We showed that automatically-derived BNF features correspond surprisingly well with representations derived in phonetic, articulatory and acoustic analyses of human speech perception and production. This confirms that the networks generating the features are in some sense learning to recognise speech, rather than spurious characteristics of the data. Further work is necessary to gain a deeper understanding of the unified space described by the BNFs, and to understand and control how they may encode the dynamics of speech in a way that could make them more useful in automatic speech recognition.

6. References

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohammed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep Neural Networks for Acoustic Modeling in Speech Recognition: The shared views of four research groups," *IEEE Sig. Proc. Magazine*, 29(6):82–97, 2012.

[2] A.-R. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic Modeling Using Deep Belief Networks," *IEEE Trans. Audio, Speech, and Language Proc.*, 20(1):14–22, 2012.

[3] L. Deng and J. Ma, "Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics," *JASA*, 108(6):3036–3048, 2000.

[4] G. Fant, *Acoustic Theory of Speech Production*, R. Jakobson and C. H. van Schooneveld, Eds., Mouton, 1970.

[5] A. Jansen and P. Niyogi, "Intrinsic spectral analysis," *IEEE Trans. Signal Proc.*, 61(7):1698–1710, 2013.

[6] D. Jones, *An outline of English Phonetics*. Cambridge University Press, 1975.

[7] L. C. W. Pols, L. J. T. van der Kamp, and R. Plomp, "Perceptual and physical space of vowel sounds," *JASA*, 46:456-467, 1969.

[8] W. Klein, R. Plomp, and L. C. W. Pols, "Vowel spectra, vowel spaces and vowel identification," *JASA*, 48:999–1009, 1970.

[9] W. Choo, "Relationships between phonetic perceptual and auditory spaces for fricatives," Ph.D. dissertation, University College London, 1996.

[10] W. Choo and M. Huckvale, "Spatial relationships in fricative perception," *Speech, Hearing and Language: work in progress*, vol. 10, 1997.

[11] M. Russell and R. Moore, "Explicit modelling of state occupancy in Hidden Markov Models for automatic speech recognition," in *Proc. ICASSP*, New York, 1985, pp. 5 – 8.

[12] M. Ostendorf, V. Digalakis, and O. Kimball, "From HMM's to segment models: a unified view of stochastic modeling for speech recognition," *IEEE Trans. Speech and Audio Proc.*, 4(5):360–378, 1996.

[13] C. J. Champion and S. M. Houghton, "Application of continuous state Hidden Markov Models to a classical problem in speech recognition," *Computer Speech & Language*, 36:347–364, 2016.

[14] S. M. Houghton, C. J. Champion, and P. Weber, "Recognition of voiced sounds with a continuous state HMM," in *Proc. Interspeech*, Dresden, pp. 523–527, 2015.

[15] P. Weber, C. Champion, S. Houghton, P. Jančovič, and M. Russell, "Consonant recognition with continuous-state Hidden Markov Models and perceptually-motivated features," in *Proc. Interspeech*, Dresden, pp. 1893–1897, 2015.

[16] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio, Speech & Language Proc.*, 14(2):435–444, 2006.

[17] P. Weber, L. Bai, S. M. Houghton, P. Jančovič, and M. J. Russell, "Progress on phoneme recognition with a continuous-state HMM," in *Proc. ICASSP*, pp. 5850–5854, 2016.

[18] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM," NIST, Tech. Rep., 1990.

[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. A. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book, version 3.4*. Cambridge University Engineering Department, 2006.

[20] 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, pp. 583–587, 2015.

[21] J. Dang, S. Wang, and M. Unoki, "Investigations into vowel and consonant structures in articulatory and auditory spaces using Laplacian Eigenmaps," in *Proc. ICASSP*, pp. 5355–5359, 2016.

[22] P. K. Kuhl, "Mechanisms of development change in speech and language," in *Proc. ICPhS*, 2:132–139, 1995.

[23] 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. the Python for Scientific Computing Conference (SciPy)*, 2010.

[24] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using Hidden Markov Models," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, 37(11):1641–1648, 1989.

[25] J. N. Holmes, I. G. Mattingly, and J. N. Shearme, "Speech synthesis by rule," *Language and Speech*, 7(3):127–143, 1964.

[26] P. Weber, S. Houghton, C. Champion, M. Russell, and P. Jančovič, "Trajectory analysis of speech using continuous state Hidden Markov Models," in *Proc. ICASSP*, Florence, 3042–3046, 2014.

[27] K. Sjölander and J. Beskow, "WaveSurfer - an open source speech tool," in *Proc. ICSLP*, 4:464–467, 2000.

[28] D. Yu, M. L. Seltzer, J. Li, J. Huang, and F. Seide, "Feature learning in deep neural networks - A study on speech recognition tasks," *CoRR*, vol. abs/1301.3605, 2013.

[29] M. Joos, "Acoustic phonetics," *Language*, vol. Monographs 23, no. Suppl. 24, 1948.

[30] S. Hawkins and J. Midgley, "Formant frequencies of RP monophthongs in four age groups of speakers," *J. Int. Phon. Assoc.*, 35(2):183–1999, 2005.

[31] F. Li, A. Trevino, A. Menon, and J. B. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of American English fricative consonants in noise," *JASA*, 132(4):2663–2675, 2012.

[32] L. Deng, X. Cui, R. Pruvenok, Y. Chen, S. Momen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *Proc. ICASSP*, Toulouse, pp. 369–372, 2006.

[33] F. Li, A. Menon, and J. B. Allen, "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," *JASA*, 127(4):2599–2610, 2010.

[34] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *Proc. ICASSP*, 1:289–292, 1999.

[35] K. N. Stevens and S. E. Blumstein, "Invariant cues for place of articulation in stop consonants," *JASA*, 64(5):1358–1368, 1978.

[36] V. Delvaux, "Perception du contraste de nasalité vocalique en Français," *Journal of French Language Studies*, 19:25–59, 3 2009.

[37] J. D. W. Stephens and L. L. Holt, "A standard set of American-English voiced stop-consonant stimuli from morphed natural speech," *Speech Communication*, 53(6)6:877–888, 2011.

[38] A. Khasanova, J. Cole, and M. Hasegawa-Johnson, "Detecting articulatory compensation in acoustic data through linear regression modeling," in *Proc. Interspeech*, pp. 925-929, 2014.