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

Philip Weber, Linxue Bai, Martin Russell, Peter Jančovič, Steve Houghton

School of Engineering, Department of EESE, University of Birmingham, UK

UNIVERSITY OF BIRMINGHAM

What This Work is About

Analysis of very low-dimensional bottleneck features.

- spatial representation of vowels is very close to $F_1{:}F_2$;
- consonants are analysed in the same framework;
- neural networks seem to derive representations specific to particular phonetic categories —
- properties similar to those used by human perception.

Bottleneck Features (BNFs) and Models of Speech

Speech suggested to lie on low-dimensional manifolds (e.g. [1]).

- for Vowels: 81.6% acoustic variation is explained by 3 dimensions (94.0% with 5) (*Pols et al.* [2]) — *cf* formants $F_1 - F_3$;
- for Fricatives: 95% by 2 dimensions (*Choo et al.* [3]).

Segmental and Continuous-State HMM models for ASR

- aim to be faithful to the nature and dynamics of speech,
- but had difficulty adequately modelling (e.g.) the all variability in natural low-dimensional representations such as formants.
- 9d Bottleneck features found to perform well in CS-HMMs [4]:

Network	Autoencoder		MLP		DBN	
Dim./Layer	**3d/3**	**9d/3**	**3d/3**	**9d/3**	**3d/4**	**9d/4**
all phones	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

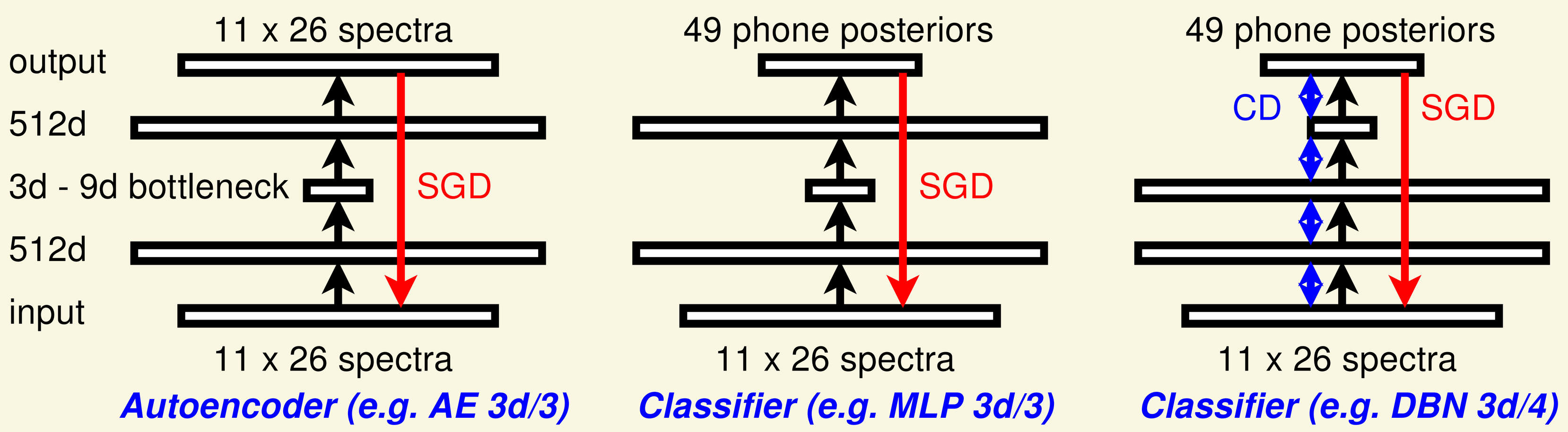
CS-HMM phone recognition % Error using bottleneck features (BNFs).

But what do the BNF features represent?
Can they be related to human perception or production?

Bottleneck Networks

Bottleneck Features (BNFs) from 3 types of networks (from TIMIT).

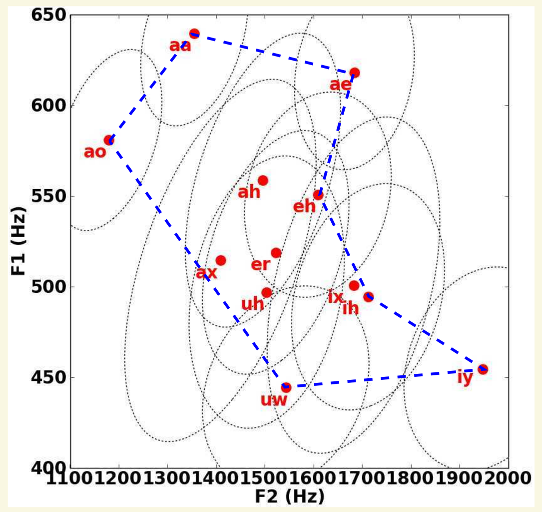
1. Autoencoders: reconstruct input.
2. MLP Classifiers (SGD-trained): predict posteriors for 49 phones.
3. DBN Classifiers (CD pre-trained with SGD fine-tuning).



- 5 layers (2d–9d bottleneck in layer 3 [e.g. **'3d/3'**] or 4 [**'3d/4'**]).
- Input: 11 frames of 26-dimensional log Mel filterbanks.
- Training: SGD and/or CD using Theano, cross-entropy error.

Vowels: Formants *vs* Autoencoder 3-dimensional BNFs

1. Pairwise plots of 3d BNFs in comparison with formants.
2. Visualisation and comparison using metrics to compare 'shape'.

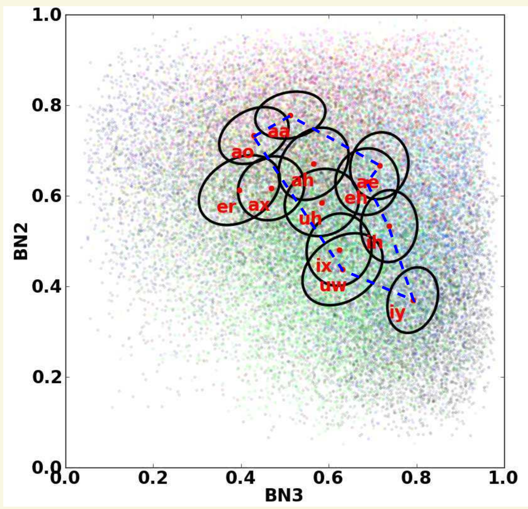


$F_1{:}F_2$ Vowel Space

3d BNFs from Autoencoder

- preserve formant vowel space structure,
- exhibit large variance
- poorly separated phone clusters,
- compression into part of space mimics limits of human articulation.

(TIMIT features)



BNF pair from Autoencoder

Visualisation: Features are represented spatially using 2d plots showing centroids and 0.5 s.d. of clusters of phone realisations (average feature between TIMIT boundaries) for a set of phones.

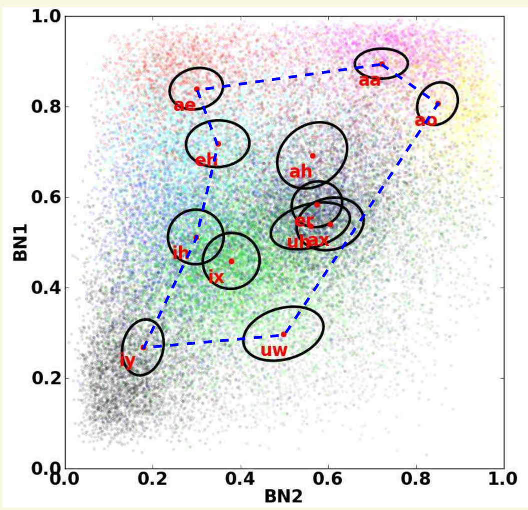
Metrics: Let X, Y be $n \times 2$ matrices of points $\mathbf{x}_i = (x_{i1}, x_{i2})$, $\mathbf{y}_i = (y_{i1}, y_{i2})$ in two such plots.

The shape $\mathcal{D}_x$ described by connecting points $\mathbf{x}_i$ is significant, but not its location or rotation.

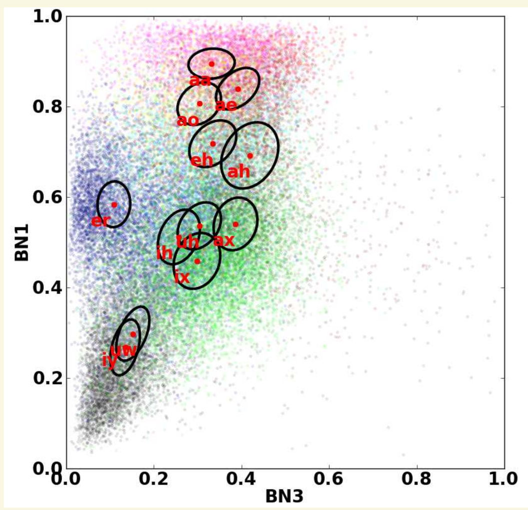
$d_2(X, Y)$ is the Euclidean distance between points $\mathbf{y}_i$ and the best-fit $\hat{\mathbf{y}}_i$ found by affine transforming X towards Y,

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

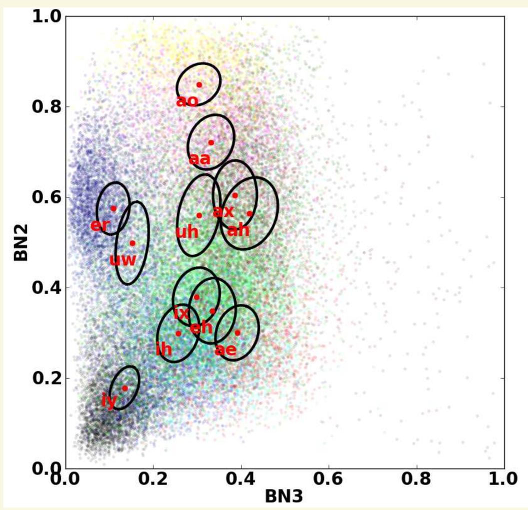
Vowels: Formants *vs* Classifier 3-dimensional BNFs



3d BNF from MLP: pair $(0, 1)$



pair $(0, 2)$



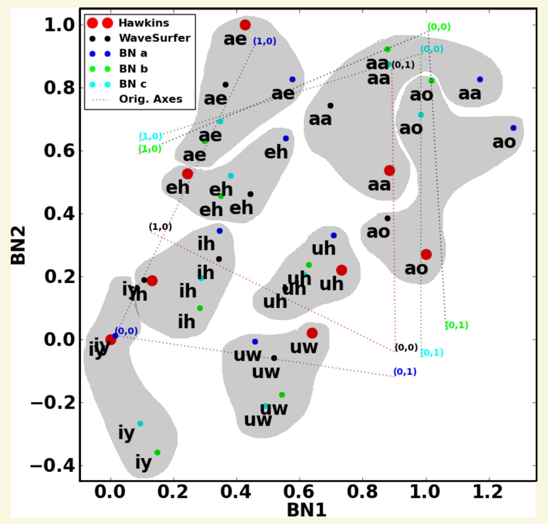
pair $(1, 2)$

Feature	Pair	Best Match Formant	d_2	Avg. Worst d_2
MLP 3d/3 a	$(1, 3)$	$F_1{:}F_2$	0.142	0.278
MLP 3d/3 b	$(1, 2)$	$\mathbf{F_1{:}F_3}$	0.195	0.336
MLP 3d/3 c	$(1, 2)$	$F_1{:}F_2$	0.216	0.330
MLP 3d/3 d	$(1, 2)$	$F_1{:}F_2$	0.119	0.339
DBN 3d/4	$(2, 3)$	$F_1{:}F_2$	0.176	0.363
A/E 3d/3	$(2, 3)$	$F_1{:}F_2$	0.192	0.388
MFCC	*Inconclusive* (mean $d_2 = 0.348$, $d_s = 0.217$)			

Distances between BNF and formant spaces.

Alignment of formant and BNF representations:

- Preserves structure;
- not identical.
- $\Rightarrow$ local optima and/or
- 3 dimensions is too few.

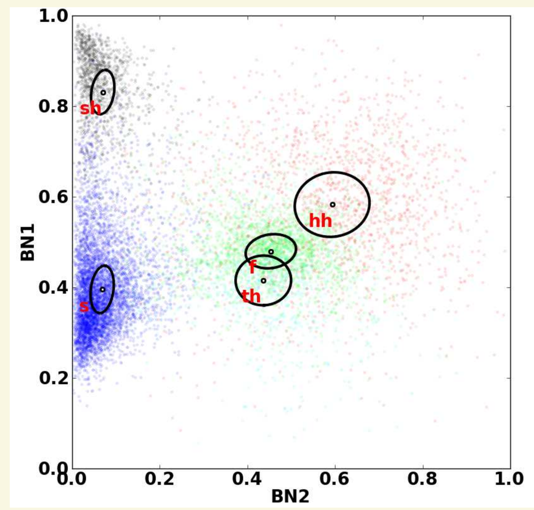


Summary:

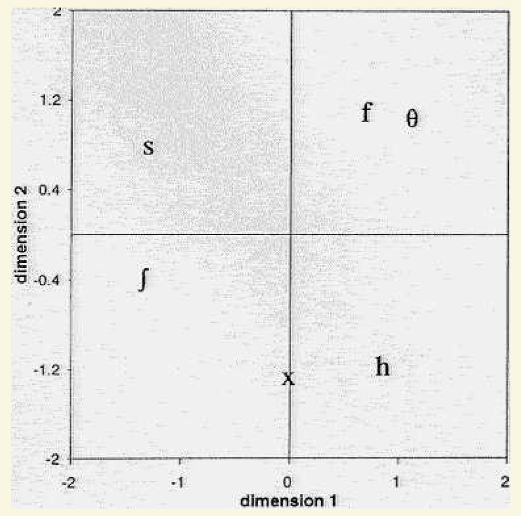
- One BNF pair always matches the vowel space.
- Reduced variance and increased cluster separation.
- 3[rd] dim. appears to relate to consonants or voicing.
- Metrics highlight a single anomaly – network reaches a different local optimum ($F_1{:}F_3$)?

Robust and repeatable analysis without formant analysis.

Spatial Representations of Fricatives



3d BNF Fricatives

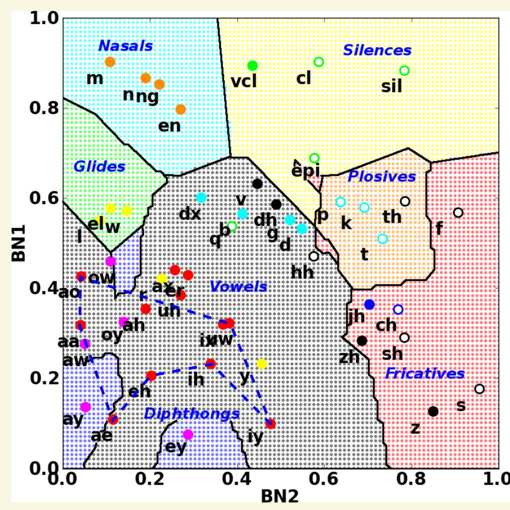


From Choo et al. [3]

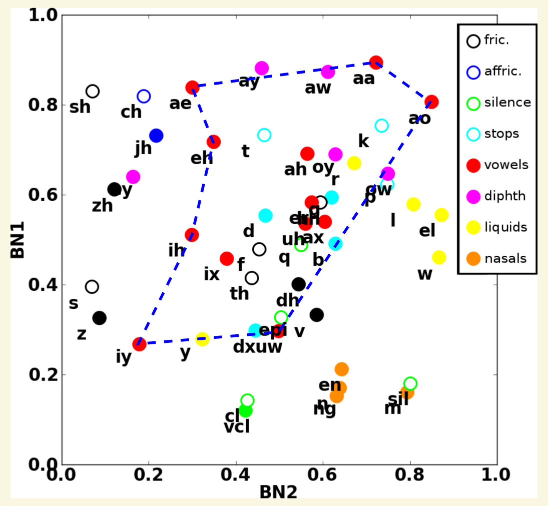
Analysis of fricatives in the same framework.

- Again close correspondence with phonetic, perceptual, and acoustic studies.
- Greater distortion.
- 3 dimensions too few?
- What are the axes?

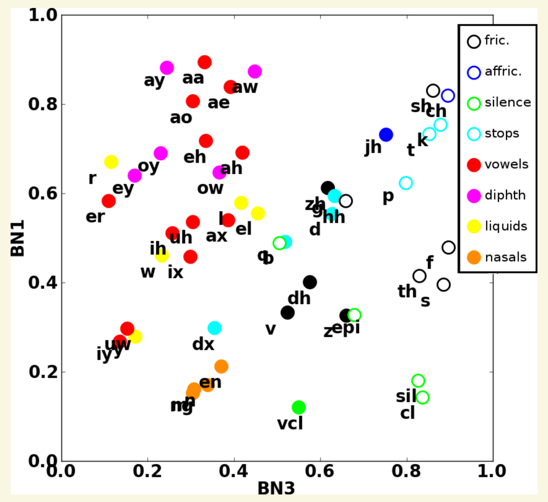
Unified Space for All Phones



2d BNF



3d BNF pair $(0, 1)$



3d BNF pair $(0, 2)$

2D BNF

- Space shows distinct regions used for each phonetic category.
- Divides voiced and unvoiced – voiced consonants on the border.
- Distorted vowel space structure – lost information.

3D BNF

- Co-location of phonetic categories in some dimensions.
- Some clustering according to place and manner of articulation.

Discussion, Conclusions and Questions

Robust and unified analysis framework for all phones.

Unified space may be of benefit to speech scientists and therapists.

- Is it interpretable globally e.g. according to tongue position or acoustics (*cf* groups {/s/, /iy/, /z/}, {/zh/, /sh/}, {/w/, /l/, /ao/})?
- Can we gain insights into theories of perception?
- How would BNFs from RNNs better capture dynamics and other features of known perceptual importance?

Selected References

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

[2] 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.

[3] W. Choo and M. Huckvale, *"Spatial Relationships in Fricative Perception"*, *Speech, Hearing and Language: Work in Progress*, vol. 10, 1997.

[4] P. Weber, L. Bai, S. M. Houghton, P. Jančovič, and M. J. Russell. *"Progress on Phoneme Recognition with a Continuous-State HMM"*. *ICASSP 2016*, Shanghai, pp. 5850-5854.