Exploring How Phone Classification Neural Networks Learn Phonetic Information by Visualising and Interpreting Bottleneck Features

Linxue Bai, Philip Weber, Peter Jančovič, Martin Russell

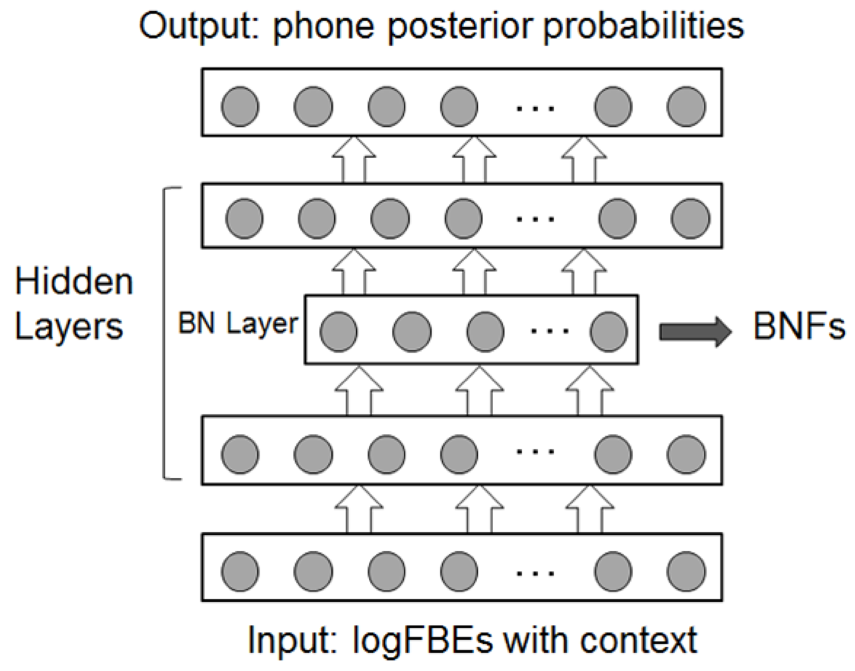
School of Engineering, University of Birmingham

Motivation

- Deep Neural Networks (DNNs) are criticized for being "black boxes"
- Very low-dimensional bottleneck features (BNFs) extracted from phone discrimination bottleneck DNNs contain sufficient information to support high-accuracy phone recognition - 9D BNFs better than 39D MFCCs [1,2]
- Can visualization of BNFs explain DNN strategies (see Weber, Bai 2016)?
- Paper explores representation of speech sounds in: very low dimensional "bottleneck" layers & non-bottleneck hidden layers

Experiments

Very low dimensional BNFs



- Conventional DNN
  - Input: 26 logFBEs + context (±5)
  - Output: phone posterior probabilities
- 286-512-2/9-512-49
  - 512 hidden units (1st and 3rd hidden layers)
  - 2 or 9 BN units
- TIMIT speech corpus

Fig 1. DNN structure used to extract BNFs.

Visualisation of 9D BNFs: LDA and t-SNE

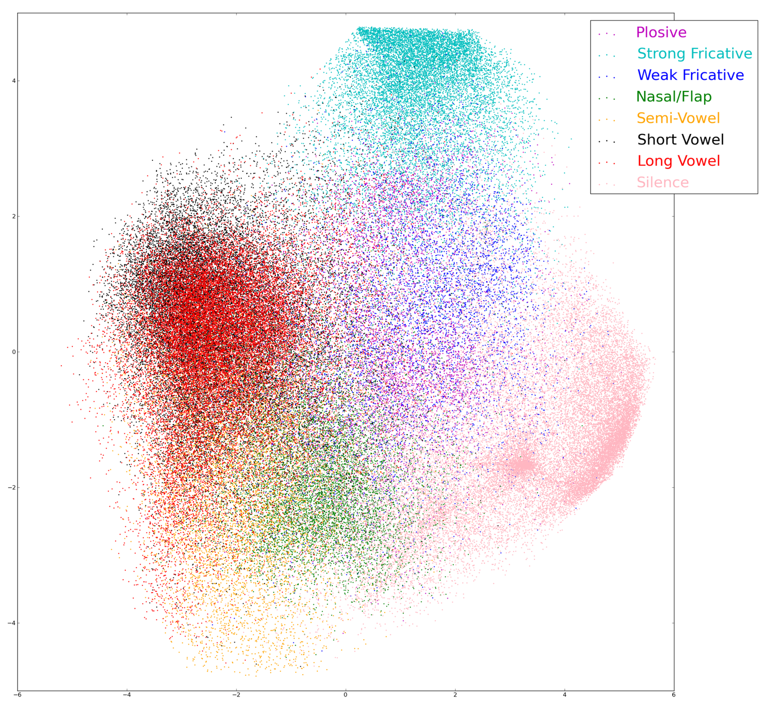


Fig 2. LDA-based projections of BNFs (1st vs 2nd dimension), coloured by broad phone categories (linear discrimination analysis (LDA) applied to 9D BNFs; *Monophone labels used* as targets in training)

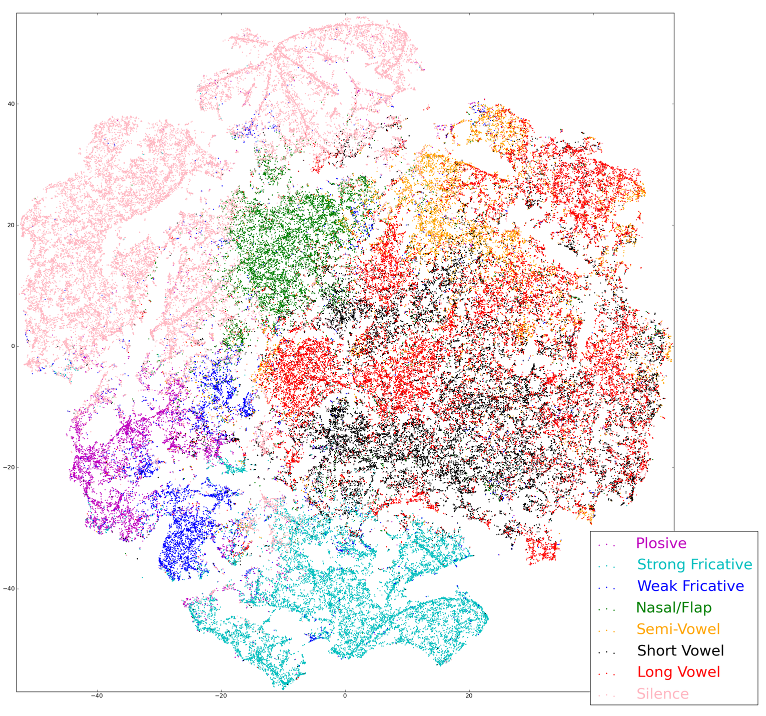


Fig 3. 2D t-SNE plots, colour-coded by broad phone categories (2D t-distributed stochastic neighbour embedding (t-SNE) visualisation of 9D BNFs; *unsupervised - monophone labels NOT given* during training)

Fig 2:

- Vowels, consonants and silences fairly well separated
- Overlaps among sub-categories of vowels; plosives & fricatives
- Horizontal axis corresponds to voicing?

Fig 3:

- Similar to LDA visualisation wrt space separation of broad classes
- "Leafs" within a broad class usually correspond to different monophones
- Sizes of clusters and distances between them NOT directly related to the size or importance of clusters in original space

Visualisation of 2D BNFs & Optimized neural activations

- ☐ For a given phone and hidden layer, a "cardinal" vector is a pattern of activations in the hidden layer that maximizes the posterior probability of the phone in the output layer
- ☐ Obtained by back-propagating layer activations
- ☐ Pre-training: Back-propagate to input layer. Use resulting BN layer activations as start-point to apply back-prop to the BN layer

References

[1] L. Bai, P. Jančovič, M. Russell, and P. Weber, *"Analysis of a low dimensional bottleneck neural network representation of speech for modelling speech dynamics"*, *Proc. Interspeech 2015*, pp. 583–587.

[2] P. Weber, L. Bai, S. Houghton, P. Jančovič, and M. Russell, *"Progress on phoneme recognition with a Continuous-State HMM"*, *Proc. ICASSP 2016*, pp. 5850–5854.

[3] 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"*, *Proc. Interspeech 2016*, pp. 3384–3388.

[4] G. E. Hinton, *"Training products of experts by minimizing contrastive divergence"*, *Training*, vol. 14, no.8, 2016

[5] A. Mohamed, G. E. Hinton, and G. Penn, *"Understanding how deep belief networks perform acoustic modelling"*, *Proc. ICASSP 2012*, pp. 4273–4276.

Experiments

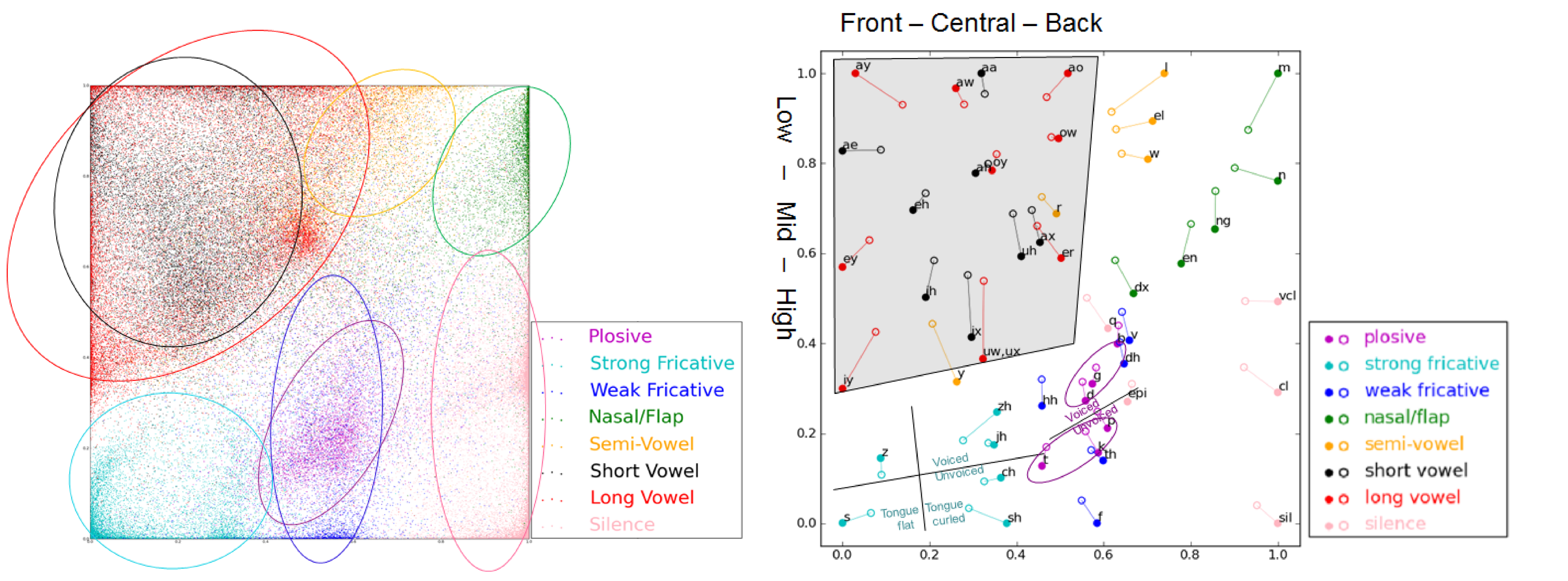


Fig 4. 2D BNF space. The left figure plots 2D BNFs (2 neurons in bottleneck layer, sigmoid activations). The right figure plots representative BNFs for each phone: Solid dots-"cardinal"BNF vectors; Open circles: centroids of the 2D BNFs (i.e. feature means)

- "Cardinal" features (dots) pushed to edges of their local regions
- Shaded area (long and short vowels): positions of centroids and cardinal features resembles phoneticians F1:F2 vowel space.
  e.g. Low to high: /ay/- /ey/- /iy/, /ao/- /uh/- /uw/, Front to back: /ey/- /ah/- /ow/
- Local coordinate systems: Vertical axis = voicing
  Horizontal axis = place of articulation (plosives in particular)

Interpretation

- 2D BNF space shows distinct regions for each phonetic category.
- Organisation of phones in a category appears to correspond to phone production mechanisms.
- Interpretation of axes for one phone category cannot be simply applied to others. BNF space seems to be a union of phonetic category related subspaces that preserve local structures within each subspace.
- Recalls locally-Euclidean topological structure, e.g. topological manifold.

Visualisation of non-bottleneck layers with LDA

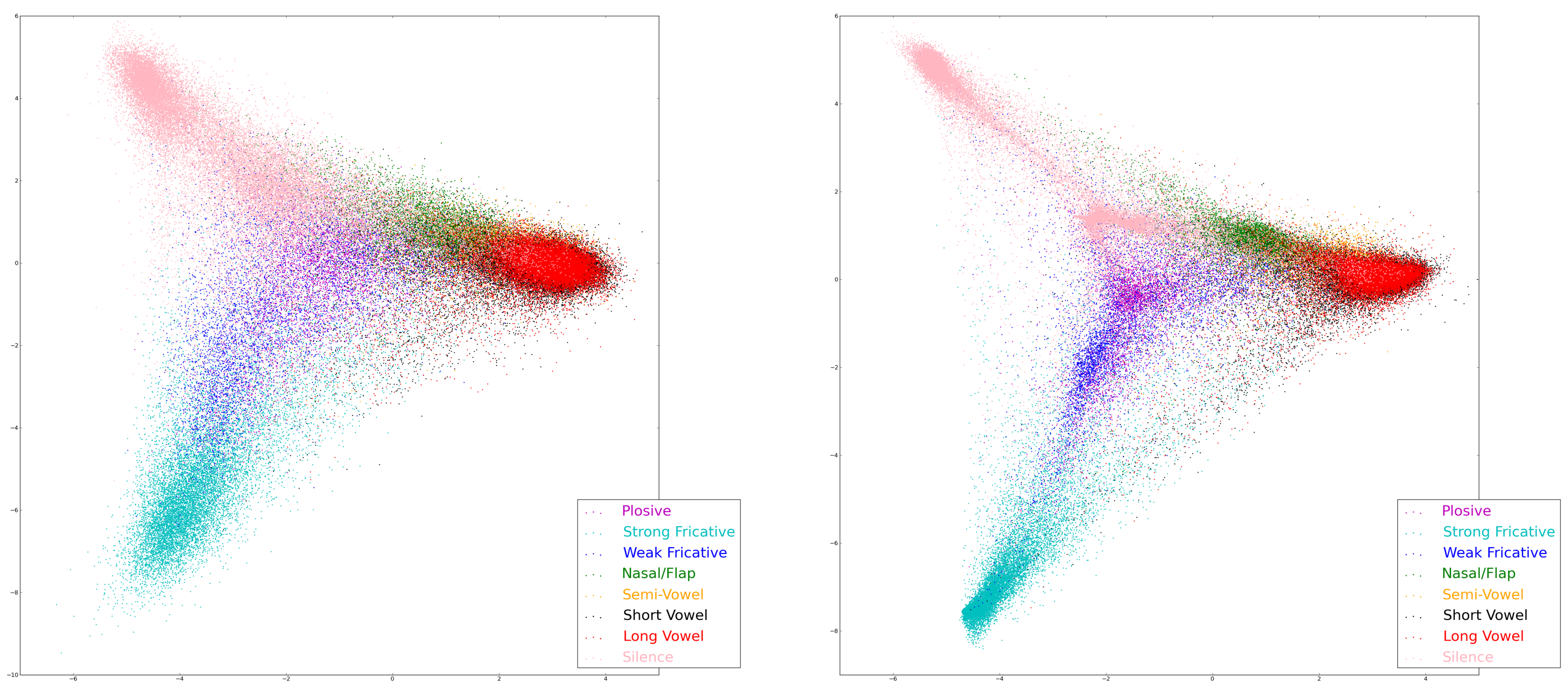


Fig 5. LDA-based projections of activations in other hidden layers (1st vs 2nd dimension) 1st (left) and 3rd (right) in the 286-512-9-512-49 DNN

- "Triangular" shape with similar structures - vowels, strong fricatives and silences each occupy a corner of the triangle.
- Horizontal axis: transitioning from unvoiced to voiced, or increasing energy in low frequency bands (left-to-right)
- Vertical: increasing energy in high frequency bands (top-to-bottom).
- Triangular plot of the 3rd hidden layer is similar but sharper
- "Triangular visualisation" always observed when analysing "bigger" hidden layers (> about 30 nodes) within DNNs of a similar structure.

Conclusions

- Visualisations of BNFs suggest phone classification strategy in DNNs can be interpreted in terms of phonetic categories (see discussion in each subsections).
- In non-bottleneck layers, as data moves through the network, from input to output, phonetic categories become more specific. Consistent with previous interpretations of DNNs [4,5].
- Triangular pattern in 1st two LDA dimensions suggests that silence, friction and voicing are three main properties learned by the DNNs.
- Relationship between internal representations learned by DNNs for speech recognition and phonetic descriptions of speech has potential impact. E.g.
  - Use of phonetic knowledge to improve DNN performance (like in [6]),
  - Use visualisation of DNN structure to gain phonetic insights.
  - May also be useful for pronunciation training.