Analysis of a low-dimensional bottleneck neural network representation of speech for modelling speech dynamics

Linxue Bai, Peter Jančovič, Martin Russell, Philip Weber

School of EESE, University of Birmingham

Low Dimensional Bottleneck Features

We analyse a low-dimensional representation of speech, extracted using bottleneck neural networks, for modelling speech dynamics.

Segmental Models of Speech

Conventional HMMs model speech as a sequence of piece-wise constant segments, associating states with individual acoustic feature vectors which are assumed independent.

Segmental (e.g. [1, 2]) and Continuous-State [3] models instead associate states with sequences of conditionally interdependent features.

Such models aim to be more faithful to the dynamics of the speech signal arising from slow, continuous movement of a few human articulators between target positions for the various speech sounds.

For example the Holmes, Mattingly, Shearme (HMS) Model [4] models speech as:

- smooth trajectories (in a suitable space), with
- piece-wise linear 'dwell-transition' approximation, consisting of
- alternating stationary periods connected by smooth transitions,

Representations Preserving Speech Dynamics

MFCCs are less suitable for segmental models since they only indirectly manifest the articulator dynamics of speech.

Formant or articulatory parameters directly describe the processes of speech production and preserve dynamics, but are difficult to estimate reliably or not well defined for all speech sounds.

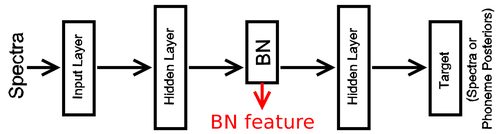
Therefore, there is a need for compact representations of speech, that can be reliably estimated for all speech sounds.

References

1. M. Ostendorf, V. V. Digalakis, and O. A. Kimball, *"From HMMs to segment models: A unified view of stochastic modeling for speech recognition"*, *IEEE Trans. Speech Audio Process.*, 4(5), pp. 360–378, 1996.
2. W. J. Holmes and M. J. Russell, *"Probabilistic trajectory segmental HMMs"*, *Computer Speech and Language*, 13(1), 1999.
3. C. J. Champion and S. M. Houghton, *"Application of Continuous State Hidden Markov Models to a classical problem in speech recognition"*, *Accepted to CSL*, 2015.
4. J. N. Holmes, I. G. Mattingly, and J. N. Shearme, *"Speech synthesis by rule"*, *Language and Speech*, 7(3), pp. 127–143, 1964.

Experimental Setup

Hidden unit activations are extracted from the bottleneck third layer of five layer networks, and used to train and test a standard HMM-GMM phoneme recogniser.



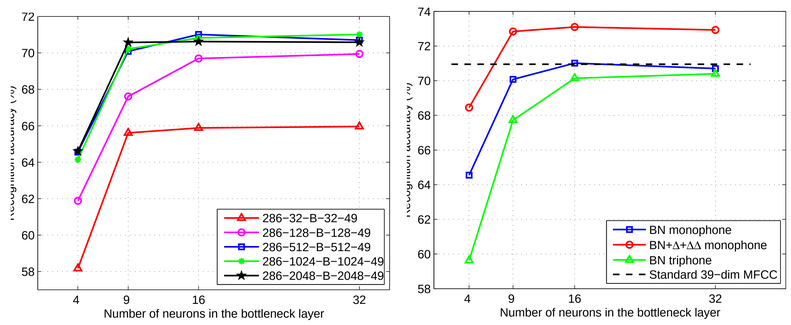
Neural network for feature extraction:

- Input: 26-dim Mel filterbanks, 16 kHz audio, 25ms Hamming window, 10ms frame rate, zero mean, unit variance.
- Training set: 90% of TIMIT Train (remainder for validation).
- Training algorithm: stochastic gradient descent with Theano.
- Training targets: (i) reconstruct input, (ii) estimate posterior probabilities for 49 TIMIT phone classes.

Recognition system:

- Monophone GMM-HMMs built in HTK, 49 3-state models with bigram language model.
- Triphone system also built for some experiments.
- Trained on features from TIMIT Train; evaluated on Core Test.

Results with Various Networks and Feature Dimensions



Results using features from posterior-probability trained networks:

- Best performance (71.0%) with 9 to 16-dimensional features and 512 hidden units in layers 2 and 4 (left figure).
- Monophone results better than triphone, 9-dim features without deltas better than 4-dim $+\Delta+\Delta\Delta$ (i.e. 12-dim) (right figure).

Features from reconstruction networks did not perform well: achieving 56.8%Acc., with 16-dim bottleneck and 128 neurons in layers 2 and 4, *vs* 69.7% from the equivalent posterior network.

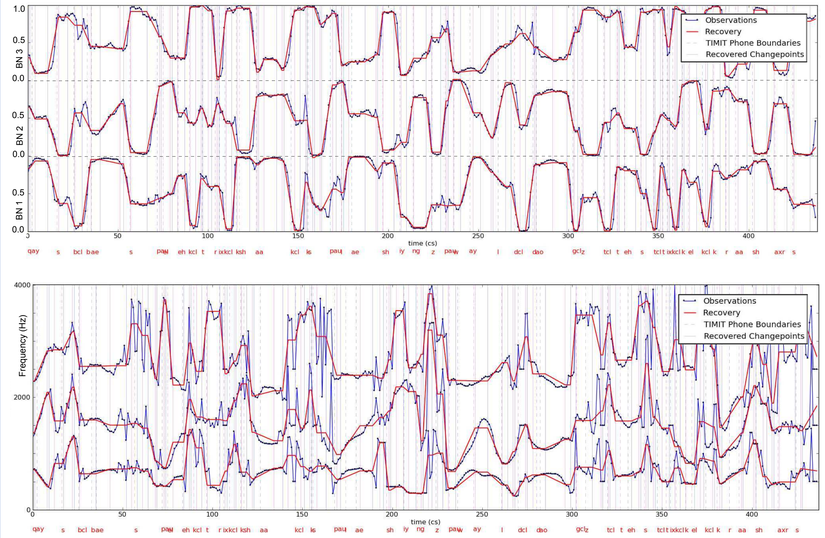
Phone Recognition with Bottlenecks or Formants

Phone errors were 33.7% lower using bottleneck features than with same-dimension formant features, using the same HMM system.

Feature representation	Dim.	Corr (%)	Acc (%)
Baseline: MFCC + Δ + ΔΔ	39	76.2	71.0
3 formants	3	49.3	40.7
3 freq + Δ + ΔΔ	9	56.3	51.1
3 freq & amp & bw	9	56.0	52.0
3 formants & amp & bw+ Δ + ΔΔ	27	65.1	60.4
3 BN features	3	65.0	60.9
3 BN features + Δ + ΔΔ	9	70.9	65.7
9 BN features	9	74.4	70.6
9 BN features + Δ + ΔΔ	27	76.8	73.1

Analysis of the Dynamics of the Features

Example dwell-transition (HMS) trajectories (red) recovered by a CSHMM [3] using bottleneck features (blue, top) and formants (below) suggest that bottleneck features preserve the trajectory continuity better, and fit the CSHMM modelling better, than formants. This is particularly clear in unvoiced regions.



Conclusions

Bottleneck features provide a compact representation and seem well suited for segmental models of speech dynamics. Further research will seek to interpret these features and apply them to recognition with CSHMMs.