Progress on Phoneme Recognition with a Continuous-State HMM

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

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

UNIVERSITY OF BIRMINGHAM

What This Work is About

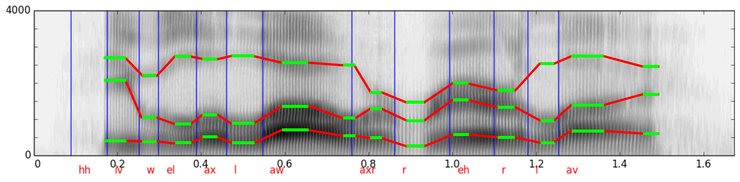
Phoneme recognition with a segmental model of speech, a continuous-state algorithm, very few parameters, using a low-dimensional representation of speech derived from a bottleneck neural network.

Low-Dimensional Models of Speech

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

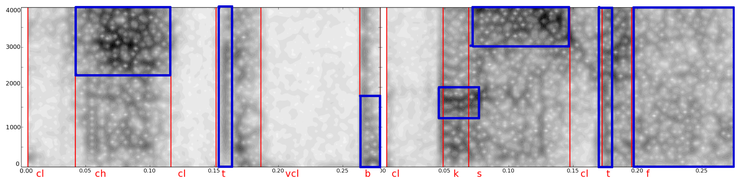
Since speech is generated by the relatively slow, constrained and smooth movement of a small number of articulators in the human vocal tract.
Therefore features are strongly correlated in time and typically exhibit smooth, slowly-varying dynamics.

e.g. Voiced sounds described by smoothly-varying resonances [2],



Piecewise linear continuous model with dwell-transition dynamics, inspired by Holmes, Mattingly and Shearme.

or Consonants by stationary features and abrupt changes [3].



Piecewise constant disconnected model with dwell-only dynamics: 2 consonant sequences highlighting discriminatory energy bands.

This poster concentrates on application of the dwell-transition model to low-dimensional bottleneck features.

CS-HMM Model and Recovery [2]

Assuming continuous features with smoothly-varying dynamics, generated according to the 'dwell-transition' model:

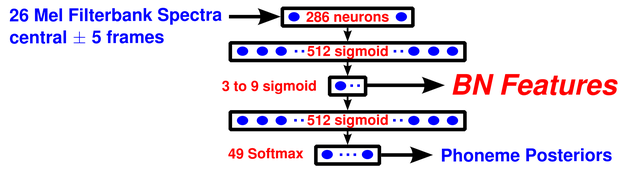
1. fit a continuous sequence of trajectories to them, to recover the sequence of phonemes.
2. Estimate an inventory of phoneme targets $\boldsymbol{\theta}_{\varphi}$ and realisation (instance) variances A_{φ}.
3. Initialise hypotheses for each starting phoneme, with continuous and discrete components.
4. Continuous components: $\boldsymbol{x}_t$: current estimate of current realisation based on phonetic history and observations $\boldsymbol{y}_1 \ldots \boldsymbol{y}_t$, stored as a scaled Gaussian $K_t \mathcal{N}(\boldsymbol{x} - \boldsymbol{\mu}_t, \mathrm{P}_t)$.
5. Discrete components $\boldsymbol{d}$: phonetic history and time in current dwell or transition.
6. Update the continuous components $\boldsymbol{x}$ based on observation $\boldsymbol{y}_t$.
7. Realisation A_{ϕ} and observation variance E are accounted for separately.
8. Branch and prune hypotheses to explore all explanations of the data.
9. Score hypotheses based on K_t.

This scheme requires only a small number of parameters.

'Natural' and Bottleneck Features

'Natural' Features: Formants estimated using WaveSurfer, Vocal Tract Resonances (VTRs), and Spectral Features measuring mean energy in perceptually-motivated frequency bands.

Bottleneck Features: activations of a 3-9 neuron hidden layer, in a 5-layer MLP classifier trained by Stochastic Gradient Descent to predict phoneme posteriors from 11-frame filterbank inputs [4].



Recognition Results: Bottleneck vs 'Natural Features'

Features	Dimension	Phone Set	Train	Test	%Corr	%Err	Parameters
MFCC+δ+$\delta\delta$	39	all	Train	Core Test	76.2	29.1	1.4×10^7
9D Bottleneck	9	all	Train	Core Test	74.4	29.4	2.3×10^5
3D Bottleneck	3	all	Train	Core Test	65.0	39.1	7.6×10^4
3 Formant	3	all	Train	Core Test	49.3	59.3	7.6×10^4
3 Formant+δ+$\delta\delta$	9	all	Train	Core Test	56.3	48.9	2.3×10^5

Discrete-state monophone multiple mixture GMM-HMMs

9D bottleneck features (BNFs) give similar accuracy to MFCCs (13 plus deltas and delta-deltas). 3D BNFs considerably out-perform equivalent-dimension estimated formants.

Features	Dimension	Phone Set	%Corr	%Sub	%Del	%Ins	%Err (S/E)	Parameters
3 Formant	3	all	31.1	35.6	33.4	4.8	73.7	163
3 Formant	3	voiced	20.4	31.2	48.4	1.6	81.2	112
3 VTR	3	all	29.2	36.2	34.6	3.7	74.6	163
3 VTR	3	voiced	29.2	37.0	33.8	3.3	74.2	112
3 VTR	3	unvoiced	32.2	33.4	34.4	2.5	70.3	67
9 Spectral	3	unvoiced	68.0	22.4	9.6	6.2	38.1	283
3D Bottleneck	3	all	55.7	30.1	14.2	3.6	47.9 (0.07)	163
3D Bottleneck	3	voiced	52.5	29.3	18.2	3.4	50.9 (0.09)	112
3D Bottleneck	3	unvoiced	71.9	17.4	10.7	2.3	30.4 (0.01)	67
9D Bottleneck	9	all	66.9	22.9	10.2	5.0	38.1 (0.11)	535
9D Bottleneck	9	voiced	60.9	24.6	14.5	3.9	43.0 (0.01)	382
9D Bottleneck	9	unvoiced	82.8	10.9	6.3	4.2	21.3 (0.25)	247

Continuous-State Hidden Markov Model

9D BNFs approach DS-HMM accuracy but with several orders of magnitude fewer parameters. Low variation between repeated random initialisations. Formants and VTRs perform very poorly.

Note: For all results, models were trained for the TIMIT 49 phoneme set and scored using the TIMIT 40 phoneme set. A bigram language model was used.

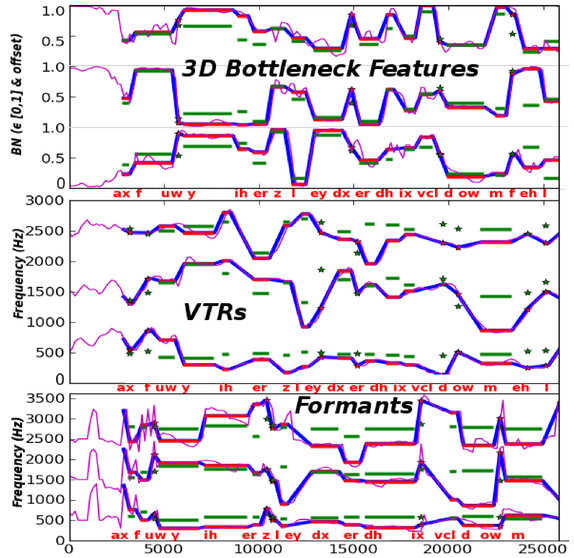
Some Points of Note

These are the first CS-HMM full phoneme recognition results with reasonable error rates. Room for improvement in training algorithm and well-informed addition of parameters. Lower error with unvoiced may give insight into the way the network is trained?

Lack of interpretability of features, and many parameters used to generate the features, although these are not used in training or testing the recogniser.

The dynamics of the bottleneck features are interesting. They seem appropriate for the CS-HMM for all types of sound.

Analysis of the Dynamics of the Features



Example CS-HMM recoveries (**thick blue lines**), showing **realised dwells (red)**, **inventory feature means (green)**. From top: 3D BNs (magenta) $\in [0, 1]$, offset to visualise), VTRs, formants.

Formants and VTR: CS-HMM recovery (**blue**, **red**) fits data (magenta) closely. Inventory frequencies (**green**) lack discrimination, and are far from the realisations (**red**)
$\Rightarrow$ Too much variability for the model/algorithm as configured.
Formants: Ill-defined for unvoiced $\Rightarrow$ tries to fit many short dwells.
Bottlenecks: Inventory is more discriminatory, recovery fits the data well, features are less dependent on part of speech $\Rightarrow$ some variability unnecessary for this task has been removed.

Summary and Questions Arising

Phoneme recognition using a model faithful to human speech production. Low dimensional 'bottleneck' features apparently somehow capture the true dynamics of speech. **But** . . .

- How should we interpret these features?
- How do they capture the dynamics of human speech?
- How and what is the network learning?
- **What can we learn about human speech, its perception or production, neural networks, or to apply to improving ASR?**

Selected References

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

[2] C. J. Champion and S. M. Houghton. *"Application of Continuous State Hidden Markov Models to a classical problem in speech recognition"*. *Computer Speech and Language*, 36(1):347–364, 2016.

[3] P. Weber, C. Champion, S. Houghton, P. Jančovič, and M. Russell. *"Consonant Recognition with Continuous* -State Hidden Markov Models and Perceptually-Motivated Features". *Interspeech 2015*, Dresden, pp. 1893–1897.

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