Consonant Recognition with Continuous-state Hidden Markov Models and Perceptually-Motivated Features

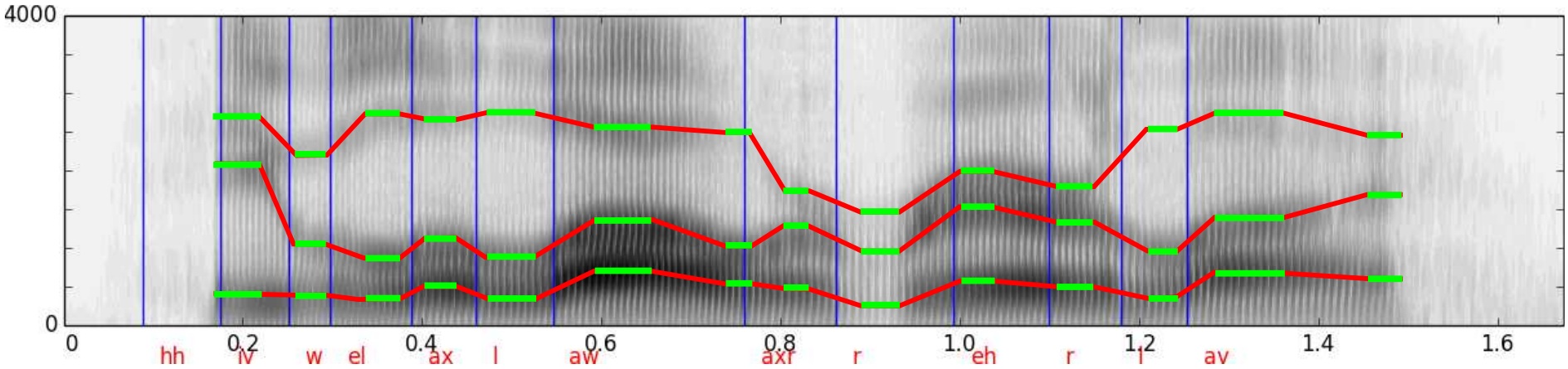
Philip Weber, Colin Champion, Steve Houghton, Peter Jančovič and Martin Russell

Speech Recognition by Synthesis (SRbS)
University of Birmingham, UK

UK Speech, 3 July 2015

Outline

Background – Voiced Sounds



Holmes, Mattingley, Shearme, 'Speech Synthesis by Rule', 1964 (HMS).

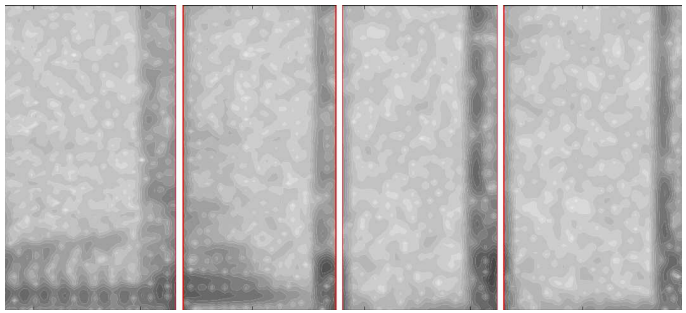
- A sequence of stationary periods linked by smooth transitions.
- Piece-wise linear approximation.
- Dwell (articulator) target frequencies, transitions.

Decode using a Continuous-State Hidden Markov Model (CS-HMM).

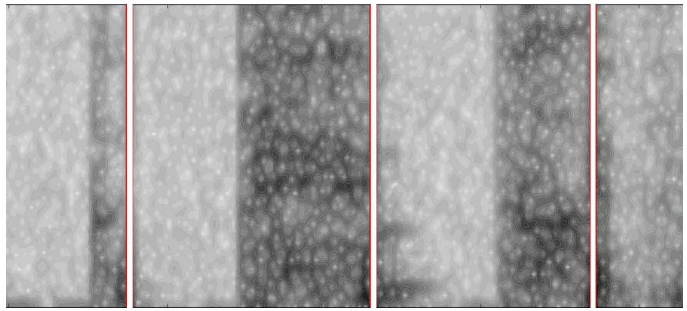
Unvoiced Sounds – Dwell-Only Model

TIMIT Examples of unvoiced phonemes.

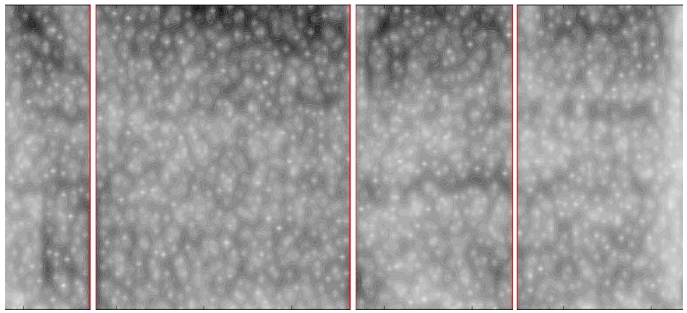
/b/



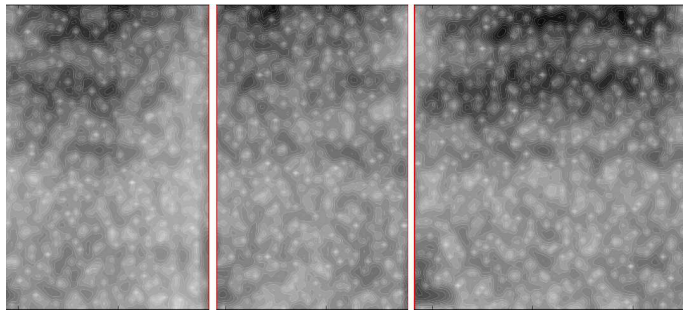
/p/



/s/



/sh/

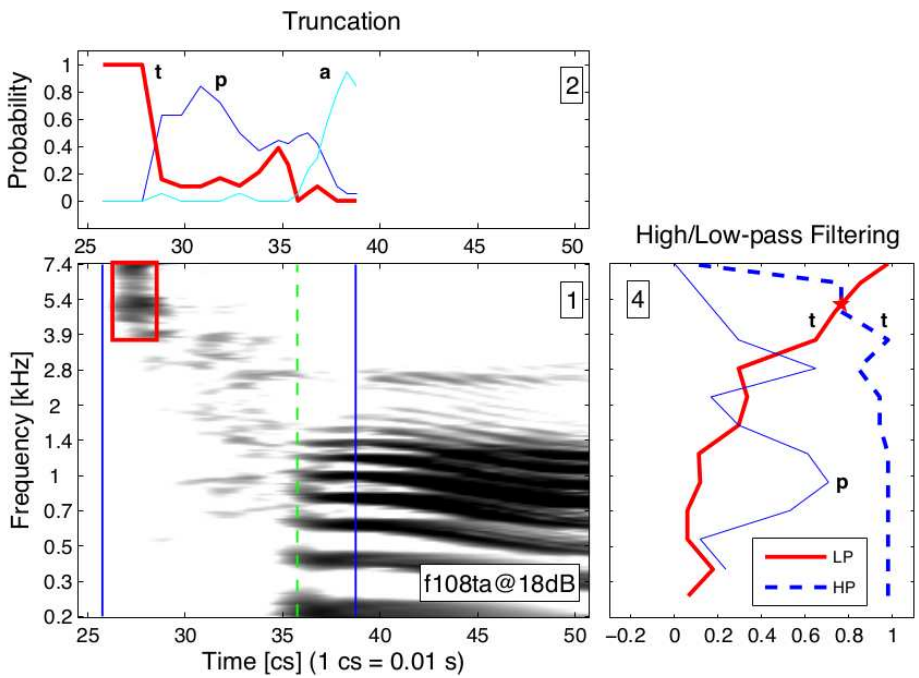


- Broadband noise at phoneme-specific frequencies.
- A sequence of stationary periods, abrupt 'transitions'.

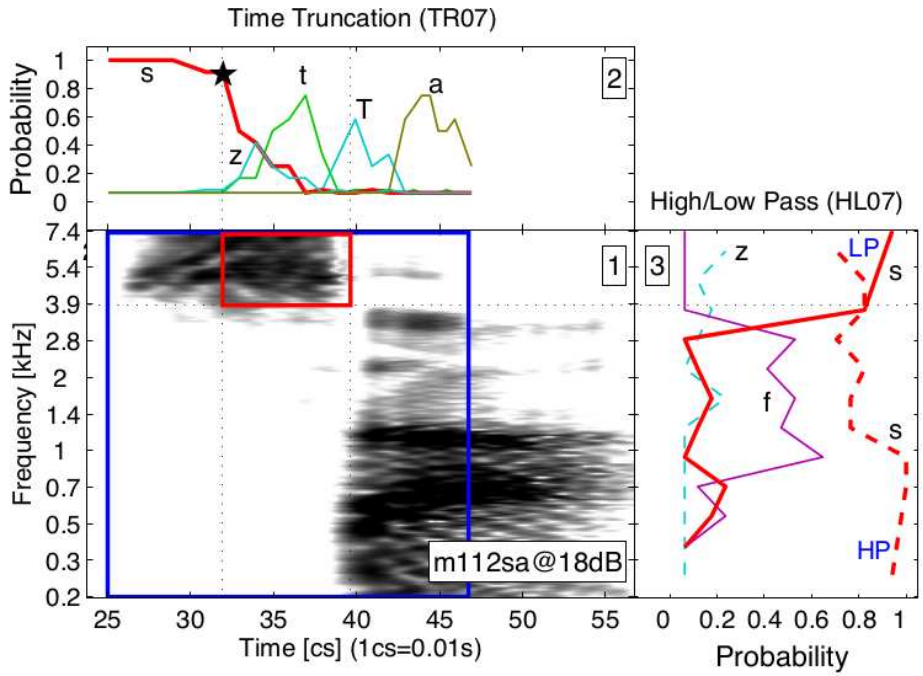
Li and Allen [1,2]: Perceptual '3DDS' Experiments

Human perception of plosives and fricatives.
Identified in frequency × time × amplitude space.

/ta/



/sa/



Figures taken from

1. **F. Li, A. Menon, and J. B. Allen,** *"A psychoacoustic method to find the perceptual cues of stop consonants in natural speech,"* The Journal of the Acoustical Society of America, 127(4), pp. 2599–2610, 2010.
2. **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,"* The Journal of the Acoustical Society of America, 132(4), pp. 2663–2675, 2012.

Li and Allen: Perceptual Cues

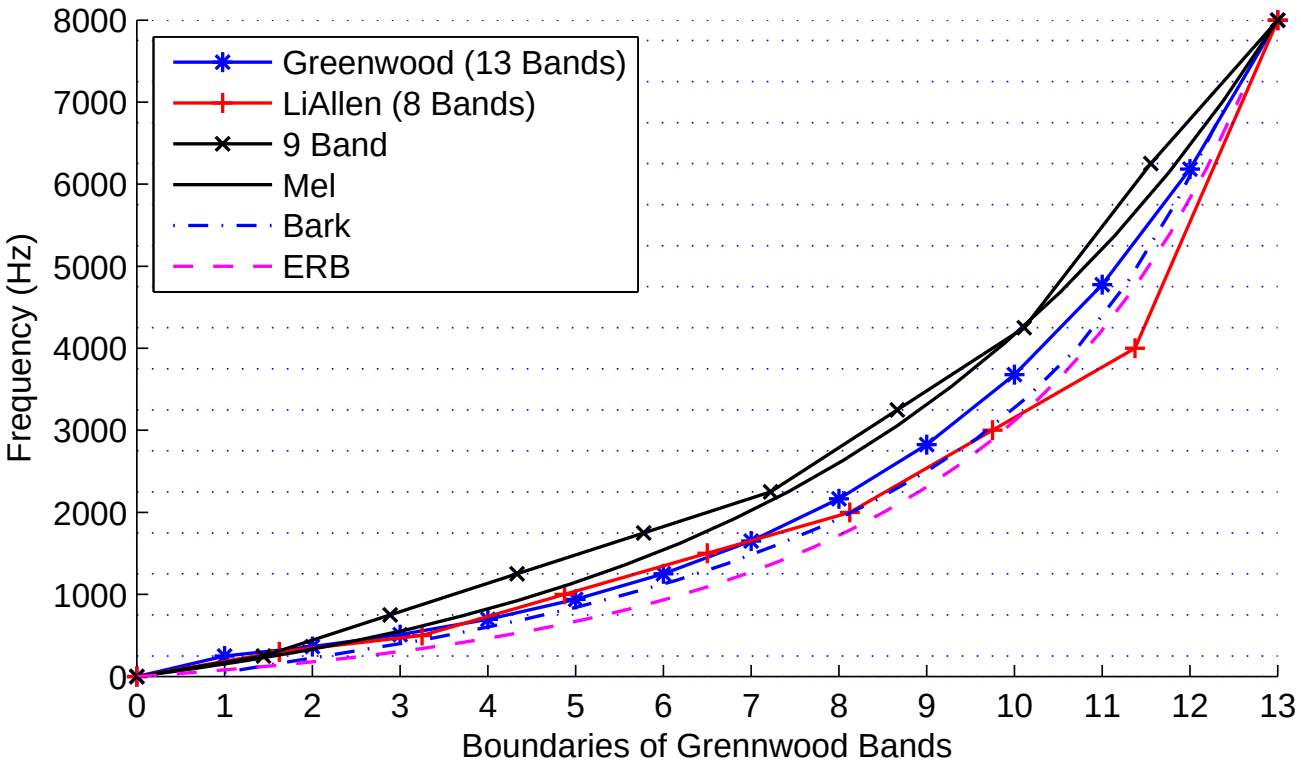
CV Pair	Main Perceptual Cues Identified
/t aa/	High-frequency burst above 3 kHz (15 ms).
/k aa/	Mid-frequency burst around 1.6 kHz.
/p aa/	Wide-band 'click' $0.3 - 7.4$ kHz, formant resonance at $1 - 1.4$ kHz.
...	...
/sh aa/	Frication noise above 2 kHz (200 ms).
/s aa/	Frication noise above 3.2 kHz (140 ms).
/z aa/	Frication region above 2.3 kHz (145 ms), voicing below 0.7 kHz.
...	...

Suggests a 9 dimension feature vector:

energy between selected frequency boundaries + phoneme duration.

0, 300, 500, 1000, 1500, 2000, 3000, 4000, 8000 Hz.

Perceptual Scales and Banding Schemes

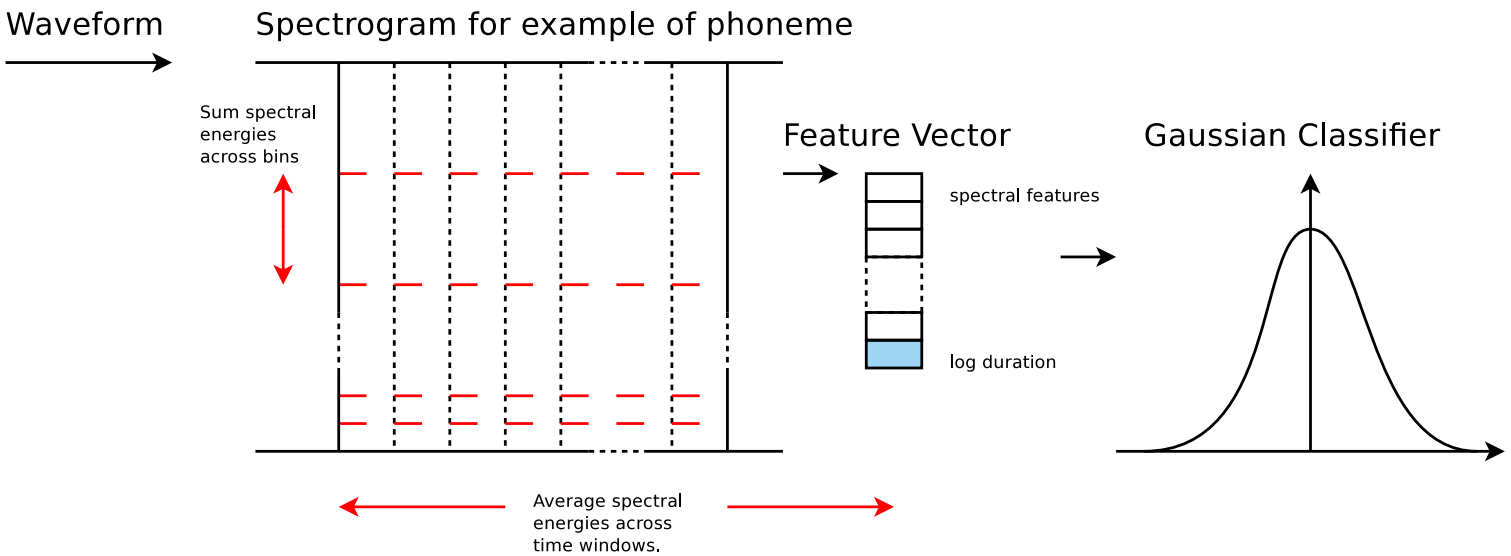


1. 'LiAllen': 8 frequency bands from the cues identified.
2. 'Greenwood': 13 bands, equal lengths on the basilar membrane [1].
3. '9 Band': 9 bands (2ms FFT, no padding – for CS-HMM).

1. **D. D. Greenwood,** *"A cochlear frequency-position function for, several species – 29 years later"* The Journal of the Acoustical Society of America, 87(6), pp. 2592âĂŞ2605, 1990.

1. Classification

Classify 13 plosives and unvoiced fricatives: /p b t d dx k g cl s sh ch f th/.
Assess parametrisation and train models for CS-HMM.



- Window phoneme $\phi^{(i)}$ into N_i (short) frames $\rightarrow$ FFT $\rightarrow$ N_i vectors $\mathbf{x}_j$.
- $\hat{\mathbf{y}}_j$: Log sum of spectral energies in $\mathbf{x}_j$ (FFT bins) within m 'bands'.
- Append log duration $logN_i$.
- $\mathbf{y}_\phi^{(i)}$: Mean of $\hat{\mathbf{y}}_j$ over FFT windows between TIMIT phoneme bounds.

Classification Results: Various Classifiers

Initial classification results using 'Greenwood' features and MFCCs.
(FFT parameters $p/q/r$: p ms window, q ms overlap, r padding).

Classifier	**Features**	**Priors**	**%Corr**
Diagonal Gaussian	Greenwood 5/4/512	no	60.04
Gaussian Naïve Bayes	Greenwood 5/4/512	yes	62.92
Full Covariance Gaussian	Greenwood 5/4/512	no	68.09
Full Covariance Gaussian	Greenwood 5/4/512	yes	70.36
Robust Covariances (Shrinkage)	Greenwood 5/4/512	yes	**71.87**
Gaussian Naïve Bayes	MFCC 25/15/512	yes	68.53
Gaussian Naïve Bayes	$+\Delta+\Delta\Delta$	yes	67.02
Gaussian Naïve Bayes	MFCC 5/4/512	yes	47.58

- Performance progression as expected.
- Poor performance with MFCCs from narrow windows.
- MFCCs with deltas use information from neighbouring phonemes, undesirable for the CS-HMM.

Classification Results: Banding Schemes

Full Covariance Gaussian classifier with Shrinkage.

Bands	Features	Dim.	kHz Range	%Class.
9 Band	2/0/0	10	0 – 8	70.61
9 Band	2/0/512	10	0 – 8	70.31
9 Band	5/0/512	10	0 – 8	71.03
9 Band	5/4/512	10	0 – 8	**72.51**
LiAllen	5/4/512	9	0 – 8	71.41
Greenwood	5/4/512	14	0 – 8	71.87
Uniform 500 Hz	5/4/512	17	0 – 8	70.43
Uniform 200 Hz	5/4/512	41	0 – 8	57.93
LiAllen	5/4/512	8	0 – 4	68.64

- 'LiAllen' $<$ '9 Band': fewer discriminatory features? Optimise bands?
- 'Greenwood' $<$ '9 Band': too many features to robustly estimate?
- Uniform features lose resolution or cannot be reliably estimated.

‘CS-HMM 101’ – Dwell-Only Model

Assuming a ‘dwell only’ model of speech (no transitions).

- Estimate canonical phoneme targets μ_ϕ (spectral energies).
- Noisy realisations $\mu_r \sim \mathcal{N}(\mu_\phi, A)$ (Gaussian around targets).
- Noisy observations $\mathbf{y}_t \sim \mathcal{N}(\mu_r, E)$ (Gaussian $\sim$ realisations).
- Timing model.

State maintains continuous and discrete components:

- Continuous: current estimate of realisation target.
- Discrete: current dwell/transition, phoneme, dwell time, history, ...

Hypothesis maintains information about infinite set of states (parametric).
Sequential branching algorithm to explore hypothesis space.

See paper in Computer, Speech and Language, 2015.

2. Decoding with Dwell-Only CS-HMM

Dwell-only simplification of the 'dwell-transition' CS-HMM model.

- Use the models trained by the classifier.
- Separate covariance components for observations and realisations.
- Separate duration and spectral energies.
- Acoustics only (flat language model).
- 29,959 TIMIT 'non-SA' sequences of 1 or more unvoiced consonants from Train.
- 10,694 sequences from (full) Test.

CS-HMM Results

Features	%Corr	%Sub	%Del	%Ins	%Err
LiAllen 5/4/512	56.9	25.4	17.8	2.1	45.3
Greenwood 5/4/512	55.1	31.6	13.4	4.5	49.5
9 Band 2/0/0	55.1	24.8	20.1	2.0	46.9
9 Band 2/0/512	54.7	23.3	22.0	1.1	46.5
9 Band 5/0/512	58.2	23.5	18.3	2.6	**44.4**
MFCC 2/0/0 (single state)	56.5	28.8	14.7	2.5	46.0
MFCC 2/0/0 (Lognormal timing)	57.0	27.9	15.1	2.4	45.4
9 Band 2/0/0 – bigram LM	73.1	19.5	8.2	3.2	30.8
MFCC 2/0/0 – bigram LM (single state)	66.1	26.2	7.9	4.9	38.8

- Compare CS-HMM with single-state 'discrete' HMM system.
- 'Chain' HMM (Log-normal timing) slightly out-performs CS-HMM.
- Lowest error for CS-HMM with language model and tuning factors.
- Caution in interpreting results...

So What, Why and What Next (I)?

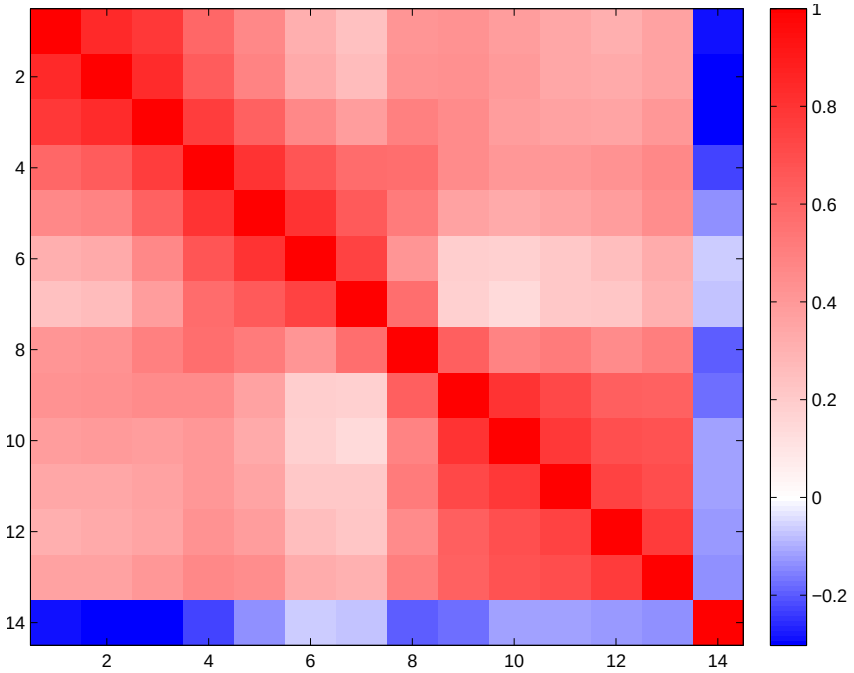
Speech recognition based on faithful, parsimonious models of speech.

- Perceptually-motivated features gave the best results for classification.
  - Similar performance for CS-HMM.
- Why should human and machine 'perception' behave similarly?
- Perhaps
  - discriminatory features of speech are adapted to human perception ...
  - $\Rightarrow$ inherently the information-bearing features for machine recognition?
- This is highlighted in the learned correlation matrices.
- How can this knowledge improve ASR?

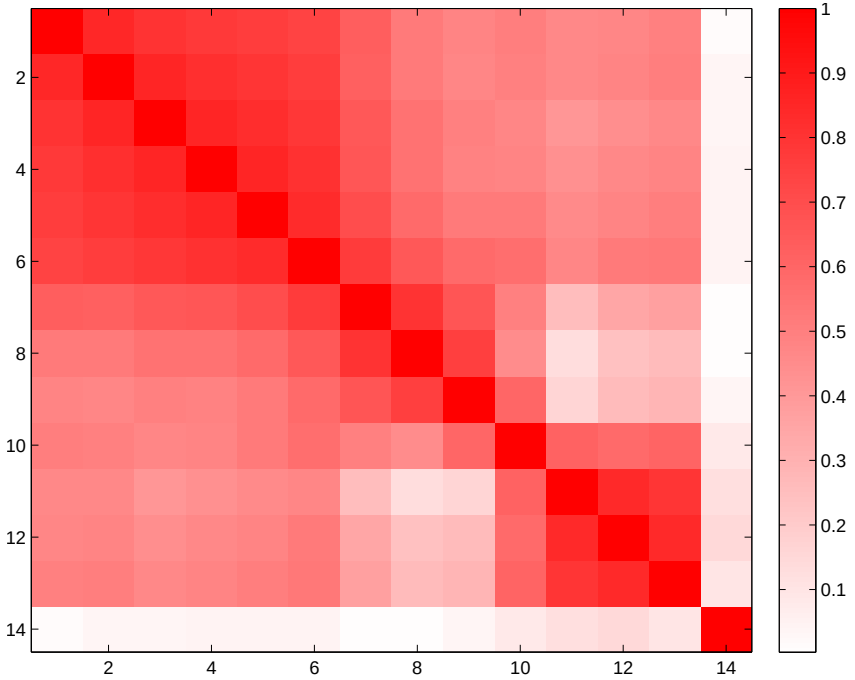
Correlation Matrices *vs* Perceptual Cues

Correlation matrices (learned spectral and duration features).

/g/



/sh/



- Positive correlation in spectral features – uniform variation in loudness.
- Negative correlation with duration in burst – concurs with literature.
- Blocking suggestive of perceptual cues.

So What, Why and What Next (II)?

Speech recognition based on faithful, parsimonious models of speech.

Suitability of linguistically meaningful features.

Motivation to pay attention to features and knowledge of HSR.

Natural framework to include other features of known perceptual importance,

1. sub-phonetic, e.g. aspiration, voiceless periods,
2. Correlations between phonemes: e.g. closure/burst, vowel duration/voicing,

Particularly when the dwell and dwell-transition models are combined,

1. e.g. formant transitions.

Automated methods to craft perceptually-meaningful features...?

Thank you!
Any questions?

http://www.birmingham.ac.uk/SRbS/
http://www.birmingham.ac.uk/philweber/