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

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

School of EESE, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK
phil.weber@bcs.org.uk,{c.champion,s.houghton,p.jancovic,m.j.russell}@bham.ac.uk

Abstract

Research into human perception of consonants has identified phoneme-specific perceptual cues. It has also been shown that the characteristics of the speech signal most useful for recognition depend on the specific speech sound. Typical ASR features and recognisers however neither vary with the type of sound nor relate directly to perceptual cues.

We investigate classification and decoding of non-sonorant consonants using basic perceptually-motivated features – phoneme durations and energy in a few broad spectral bands. Our classification results using simple classifiers suggest that features optimal for human perception also perform best for machine classification. We show how characteristics of the models learned relate to knowledge of human speech perception.

Recognition results using a continuous-state HMM (CSHMM) show accuracy similar to a discrete-state HMM with similar assumptions. We conclude by outlining how the CSHMM provides a mechanism to make use of other perceptually-important features by integration with similar models for recognition of voiced sounds.

Index Terms: Perceptual Features, CSHMM, Speech Analysis, Consonant Classification

1. Introduction

It has been shown that humans use a relatively small number of basic features of the auditory signal as perceptual cues to perceive and distinguish consonantal sounds (see e.g. [1]). These include rapid changes in spectra or amplitudes, presence or absence of voicing or aspiration, and global spectral properties such as broadband energy at specific frequencies. A key distinguishing characteristic of non-sonorant speech sounds, plosives, fricatives and affricates in particular, is the amplitude and duration of broadband noise in particular frequency bands.

Li et al. used perceptual experiments to identify cues for human recognition of plosives [2] and fricatives [3], in consonant-vowel (CV) pairs such as 'ta'. These cues are predominantly expressed as bursts of energy of phoneme-specific duration, frequency and amplitude. Similar cues were identified by Stevens et al. (e.g. [4, 5]), who also reported additional critical cues such as formant transitions and duration or absence of voicing. Other cues to consonant perception have also been described, such as duration of preceding vowel [6], gap to succeeding vowel [2, 3], formant transitions into or from the consonant [7], and correlations between features of succeeding phonemes [8].

These results raise the question: if a basic set of features – phoneme duration and mean log energy in a small set of spectral bands – largely enable human discrimination among non-sonorant consonants, could such perceptually-motivated features serve well for machine recognition of these consonants?

We explore this question through classification and recognition experiments on TIMIT [9], on a limited set of non-sonorant consonants, limiting ourselves to this basic set of features. Through the classification experiments we show that the features optimal for human perception also perform best for machine classification. We also discover interesting relations between parameters of the models learned by the classifiers, and prior knowledge of human speech perception and production.

Our motivation is to develop parsimonious models of speech and appropriate modelling algorithms which are faithful to the dynamics of the speech signal and what is known of the mechanisms of the production of different types of speech sound. Previous work [10, 11] proposed Continuous-State HMMs (CSHMMs) with formant features, as the appropriate modelling framework for the dynamics of voiced sounds, following a simplified Holmes-Mattingly-Shearme (HMS) [12] model of speech. In the HMS model, continuous, smoothly-varying parameters such as formants are approximated by alternate dwell (stationary) phases and linear transitions.

In contrast, a sequence of consonants can be described by a series of dwells, with abrupt transitions between phonemes, and decoded using a 'dwell-only' CSHMM, a simplification of the 'dwell-transition' model [11]. Using the same basic 'perceptual' features, we show similar recognition accuracy to that obtained from an equivalent (single-state, single Gaussian) discrete-state HMM built using HTK [13]. The work described in this paper therefore provides a basis for integration with the dwell-transition model into a fully CSHMM-based speech decoder. This framework will enable incorporation of other features identified as important for perception, in particular formant transitions between voiced and unvoiced regions, and correlations between features of different phonemes.

2. Features and Classifiers

The CSHMM requires a single model for each phoneme, and assumes mutually independent features. We are therefore interested in features containing information solely for the consonant in question. For this reason and because transitions between unvoiced consonants are abrupt (e.g. [1]), we use Fourier transforms with short windows. The resulting lower frequency resolution is not a concern since we further reduce resolution by summing several contiguous Fourier bins into each band. Our best results were obtained with no windowing function.

For classification, we use $m+1$ dimensional feature vectors consisting of m band energies plus duration. Consider example $\varphi^{(i)}$ from n_φ examples of phoneme φ (as labelled by TIMIT). We window the audio segment for $\varphi^{(i)}$ into N_i frames and take the square magnitude of the Fourier transform, giving N_i vectors $\boldsymbol{x}_j$ of real-valued spectral energy bins, $0 \leq j \leq N_i - 1$.

Band energies $\hat{\boldsymbol{y}}_j$ are obtained by dividing $\boldsymbol{x}_j$ into bands according to frequency boundaries given by a vector $\boldsymbol{b}$, taking the log of the sum of the bin energies in each band, and averaging across the N_i frames. Thus the feature vector for $\varphi^{(i)}$ is

$$\boldsymbol{y}_\varphi^{(i)} = \frac{1}{N_i}\sum_{j=0}^{N_i-1}\hat{\boldsymbol{y}}_j, \text{ whose elements } \hat{y}_{jk} = \log\sum_{l=b_k}^{b_{k+1}-1} x_{jl}, \quad (1)$$

for $0 \leq k \leq m-1$. (x_{jl} and b_k indicate elements of $\boldsymbol{x}_j$ and $\boldsymbol{b}$.) Finally, log duration of $\varphi^{(i)}$, $d_\varphi^{(i)} = \log N_i$, is appended to $\boldsymbol{y}_\varphi^{(i)}$.

Assuming the logs of the band energies and phoneme durations are approximately Normally distributed, we build a Gaussian model $\Phi \sim \mathcal{N}(\boldsymbol{\theta}_\varphi, \boldsymbol{\Sigma}_\varphi)$ of the log band energy and duration for phoneme φ, estimating the mean and co-variance matrix from feature vectors $\boldsymbol{y}_\varphi^{(i)}$ from n_φ training examples,

$$\boldsymbol{\theta}_\varphi = \frac{1}{n_\varphi}\sum_{i=0}^{n_\varphi-1}\boldsymbol{y}_\varphi^{(i)}, \quad \boldsymbol{\Sigma}_\varphi = \frac{1}{n_\varphi}\sum_{i=1}^{n_\varphi-1}(\boldsymbol{y}_\varphi^{(i)})^2 - \boldsymbol{\theta}_\varphi^2. \quad (2)$$

Test example t is classified to the model for class ψ that minimises the negative log probability of its feature vector $\boldsymbol{y}_t$.

$$\psi = \underset{\varphi}{\operatorname{argmin}}\left(\log|\boldsymbol{\Sigma}_\varphi| + (\boldsymbol{y}_t - \boldsymbol{\theta}_\varphi)'\boldsymbol{\Sigma}_\varphi^{-1}(\boldsymbol{y}_t - \boldsymbol{\theta}_\varphi)\right). \quad (3)$$

We built three types of classifiers. Naïve Gaussian treats the features as independent; Gaussian Naïve Bayes gives the class probability as the product of the posterior probability and class prior probability $\frac{n_\varphi}{n_T}$, (n_T phonemes in the training set), equivalent to using a unigram language model; and the Full Covariance Gaussian classifier drops the assumption of independent features. Basic linear shrinkage [14, 15] is used to improve the estimate of $\boldsymbol{\Sigma}_\varphi$ to account for limited training data for some classes. This weights the covariance towards the diagonal by a parameter $\alpha \in [0, 1]$, which we optimised empirically,

$$\hat{\boldsymbol{\Sigma}}_\varphi = (1-\alpha)\boldsymbol{\Sigma}_\varphi + \alpha\frac{\mathrm{Tr}(\boldsymbol{\Sigma}_\varphi)}{m+1}\mathbf{I}_{m+1}. \quad (4)$$

We used several banding schemes, described next. These descriptions (and subsequent experimentation on TIMIT) assume audio sampled at 16kHz.

2.1. Bands Interpreted from Li et al. [2, 3] ('LiAllen')

Li et al. identify perceptual cues to human recognition of stops and fricatives, in CV pairs with /aa/, e.g. 'ta', 'pa'. They describe cues in terms of the energy within specific bounds of frequency and time (Table 1). These frequency bands are consistent with results by Stevens et al. [1] on stops [4] and fricatives [5]. Based on these cues, we create 9-dimensional feature vectors: frequency bands between 0, 300, 500, 1000, 1500, 2000, 3000, 4000 and 8000 Hz, appended with log phoneme duration.

2.2. 'Greenwood' Perceptually Motivated Scale [16]

Li et al. identify the frequency cues by high- and low-pass filtering the speech signal at cut-off frequencies given by the 'Greenwood function' [16]. Perception experiments have shown these frequencies to correspond to equal lengths on the Basilar Membrane. We combine these values to give 13 bands between 0, 250, 363, 509, 697, 939, 1250, 1649, 2164, 2826, 3678, 4775, 6185 and 8000 Hz, and again append log phoneme duration.

The Greenwood function is similar to functional approximations to the Mel [17, 13], Bark [18] and ERB scales [19] (Figure 1). The Mel scale was popularised by Davis and Mermelstein [20], who note that the differences between these scales are not significant. These scales are approximately logarithmic, with greater resolution at lower frequencies. Figure 1 shows that our banding scheme derived from Li et al.'s work ('LiAllen') is roughly a linear approximation to these scales. Frequencies above 4000 Hz are aggregated into a single band, no distinguishing subdivisions being identified in this range (Table 1), but the frequency range is represented by fewer features, which may enable more accurate estimation of classifier parameters.

CV Pair	Main Perceptual Cues Identified
/t aa/	High-frequency burst above 3 kHz (15 ms).
/d aa/	High-frequency burst above 4 kHz.
/k aa/	Mid-frequency burst around 1.6 kHz.
/g aa/	Half-octave burst from 1.4 to 2 kHz.
/p aa/	Wide-band 'click' $0.3 - 7.4$ kHz, formant resonance at $1 - 1.4$ kHz.
/b aa/	Wide-band 'click' $0.3 - 4.5$ kHz.
/sh aa/	Frication noise above 2 kHz (200 ms).
/s aa/	Frication noise above 3.2 kHz (140 ms).
/f aa/	Wide-band frication $0.3 - 7.4$ kHz (120 ms).
/th aa/	Cues could not be reliably identified.

Table 1: Frequency and duration cues for human perception of stop and fricative consonants, reported by Li et al. [2, 3].

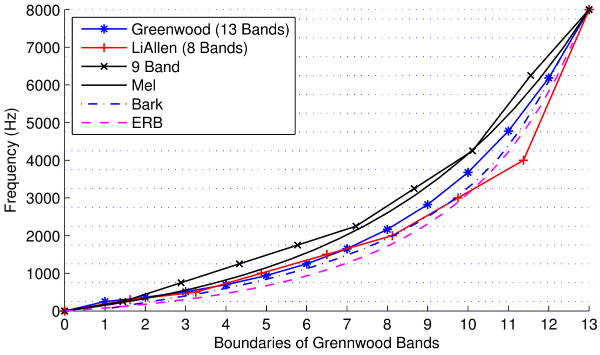

Figure 1: Frequency banding schemes alongside approximate Mel [20], Bark [18] and ERB [19] scales. The x axis indexes the 13 Greenwood bands; schemes with fewer bands are plotted at points interpolated over the same range.

2.3. Bands from 17-Bin Short Window FFT ('9 Band')

For the previous banding schemes we obtained best classification results with features derived using 5 ms FFT windows with 1 ms offset (see Table 3, Section 4). To further reduce the effect on the features of windows overlapping the phoneme boundaries, and to improve their statistical independence, we created features using a 2 ms FFT window, no overlap, no padding, giving 17 bin energies per window. We aggregated these into 9 frequency bands at hand-crafted boundaries (0, 250, 750, 1250, 1750, 2250, 3250, 4250, 6250 and 8000 Hz) designed to roughly linearly approximate the other scales (Figure 1).

2.4. Uniform Bands

To give some insight into the effect of the number of bands *vs* varying the frequency resolution according to perceptual knowledge, we constructed banding schemes which divide the frequency range uniformly. With 500 Hz bands, the feature dimension is greater than than for the perceptually-motivated features. Frequency resolution is lower at low frequencies and higher at high frequencies. 200 Hz bands give much higher dimensional features and frequency resolution similar to Greenwood at low frequencies, but greatly increased at higher frequencies.

3. Dwell-Only CSHMM

The dwell-only CSHMM for consonant recovery simplifies the dwell-transition model [10, 11]. It is equivalent to a segmental HMM [21] but takes advantage of the inductive calculation provided by the CSHMM framework. The main difference from the conventional (discrete-state) HMM (DSHMM) is the factoring of variability into realisation and observation variances. Thus instances of the same sound differing in loudness will be accounted for by realisation variance, once per instance, whereas in a DSHMM these differences can only be viewed as element observation variability, the effect of which will be brought in on every frame.

A hypothesised explanation for a set of observations $\boldsymbol{y}_0 \dots \boldsymbol{y}_{t\text{-}1}$ up to time t consists of Baum-Welch alpha values for all possible realisations (mean features) of the current phoneme, plus discrete state components: φ, the assumed identity of the current phoneme, h, the number of features in the current phoneme, and as many preceding phonemes as needed for the language model. The alphas take the form of a scaled Gaussian distribution over the means, whose mean $\boldsymbol{\mu}_t$, precision $\boldsymbol{p}_t$, and scale factor k_t are updated on every observation.

The parameters of the model are $\boldsymbol{\theta}_\varphi$, the canonical (target) mean of each phoneme, $\boldsymbol{c}_\varphi$, the (matrix) variance of realised means about the canonical mean for each phoneme, $\boldsymbol{e}_\varphi$, the (matrix) variance of observations about the realised mean, and a timing model, which in this paper we assume to be Lognormal.

The m-dimensional spectral means $\boldsymbol{\theta}_\varphi$ and $\boldsymbol{\Sigma}_\varphi$ are as estimated by the classifiers, with durations marginalised out. From (1) and (2), the observation variance is estimated as

$$\boldsymbol{e}_\varphi = \frac{1}{n_\varphi} \sum_{i=0}^{n_\varphi - 1} \Big[\frac{1}{N_i} \sum_{j=0}^{N_i - 1} (\hat{\boldsymbol{y}}_j)^2 - (\boldsymbol{y}_\varphi^{(i)})^2 \Big]. \tag{5}$$

and used to estimate $\boldsymbol{c}_\varphi$ from $\boldsymbol{\Sigma}_\varphi$.

The CSHMM provides an inductive algorithm to extend hypotheses from one observation to the next. At $t = 0$, a hypothesis is initialised for each of the M phonemes in the inventory, from the inventory means and realisation variances. At time $t - 1$ a hypothesis may be extended either by continuing in the current phoneme or stepping to a new one.

Staying in the current phoneme φ, we evolve the Gaussian parameters to account for the latest observation $\boldsymbol{y}_t$ and the timing model which gives the probability of staying within the same phoneme. If we move to a new phoneme φ', we split the hypothesis to M successors, initialising the parameters from the inventory, and factoring in a language model probability.

At each observation, every hypothesis has M successors, so the number of hypotheses grows rapidly. They may be thresholded on k_t, and hypotheses with the same histories merged.

4. Experiments and Results

We use 13 classes for classification and recognition: stops /b/, /p/, /d/, /t/, /g/, /k/, /dx/ (/d/ 'flap'); a single closure class /cl/ aggregating the stop closures /dcl/, /bcl/, /gcl/, /pcl/, /tcl/, /kcl/ and glottal stop /q/; unvoiced fricatives /s/, /sh/, /f/, /th/ and the affricate /ch/. These were chosen as best described by spectral energy features (opposed to formant-like representation). They are predominantly unvoiced, but include 'voiced' stops, for which degree of voicing can vary considerably (Section 5).

Classifier	**Features**	**Priors**	**%Corr**
Diagonal Gauss.	Grn. 5/4/512	no	60.04
Gauss. Naïve Bayes	Grn. 5/4/512	yes	62.92
Full Covar. Gauss.	Grn. 5/4/512	no	68.09
Full Covar. Gauss.	Grn. 5/4/512	yes	70.36
Robust Covars. ($\alpha = 0.1$)	Grn. 5/4/512	yes	71.87
Gauss. Naïve Bayes	MFCC 25/15/512	yes	68.53
Gauss. Naïve Bayes	*ditto* $+\Delta$+$\Delta\Delta$	yes	67.02
Gauss. Naïve Bayes	MFCC 5/4/512	yes	47.58

Table 2: Consonant classification using 'Greenwood' features and MFCCs. 'Features' *p*/*q*/*r* indicates FFT parameters: p ms windows, q ms offset, padded with zeros to r points.

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

Table 3: Classification with various banding schemes.

In Table 2 we compare classifiers trained using the 'Greenwood' bands, on all 'non-SA' utterances in TIMIT 'TRAIN', using the TIMIT transcriptions including phoneme boundaries. We report average classification accuracy on all examples of phonemes in the 13 classes, extracted from TIMIT 'TEST', a closed-class experiment. The top part of the table shows the classification performance improving from 60.04% to 71.87%, as the classifier is refined (class priors, full covariance matrices, robust estimation of covariances). These all used 5 ms rectangular windows with 1 ms offset, padded with zeros.

For comparison, accuracy with 'standard' MFCC features (25 ms windows, 10 ms offset, 26 filterbanks, 13 MFCC coefficients), even adding deltas, was lower than with spectral features. MFCCs over narrow windows gave very poor results, which we did not investigate further in this study. Using deltas the features are derived from a timespan which exceeds the phoneme boundaries (and do not align meaningfully with TIMIT time marks), hence MFCC features are not limited to information from the consonant in question (e.g. having sight of formants in adjacent vowels). Since this contradicts the CSHMM assumptions we did not pursue use of these features.

Table 3 shows classification accuracy with different banding schemes. The best results, similar for all three perceptually-motivated banding schemes, are from 5ms windows with overlap. We found zero padding before FFT to make no significant difference. We speculate that slightly reduced accuracy with the 'LiAllen' features is due to one fewer dimension or slight differences in the division into bands (which could perhaps be optimised), and with the Greenwood features, due to reduced ability to robustly estimate the greater number of parameters.

Since the only cue to specify frequencies above 4 kHz is for /d/, and Li et al.'s figures [2, Fig. 3] show this cue extends below 4 kHz, we tried discarding bands above 4 kHz. Reduced accuracy suggests useful information for classification is present in higher frequencies. Dividing into uniform 500 Hz bands also

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 (1s)	56.5	28.8	14.7	2.5	46.0
MFCC 2/0/0 (LN)	57.0	27.9	15.1	2.4	45.4
9 Band 2/0/0/LM	73.1	19.5	8.2	3.2	30.8
MFCC 2/0/0/LM (1s)	66.1	26.2	7.9	4.9	38.8

Table 4: Decoding with CSHMM, and single state single Gaussian (1s) DSHMM. 'LM' indicates a bigram language model.

reduced accuracy: reduced resolution in lower frequencies was not compensated by the increased feature dimension. Accuracy with 200 Hz bands (increased resolution) was much worse, again due to problems robustly estimating the parameters.

Table 4 shows recognition results for consonant sequences, using the CSHMM with spectral features, and DSHMM (HTK) with MFCCs calculated over the same window (26 filterbanks, 13 MFCCs). We used the TIMIT phone boundaries to extract 29959 training and 10694 test segments from the 3640 TIMIT 'TRAIN' and 1336 'TEST' 'non-SA' utterances, for sequences of one or more consecutive phonemes from our 13 classes. The '1s' DSHMM used a single-state, single Gaussian — a similar number of parameters to the CSHMM. For 'LN' the trained state was duplicated to create a chain HMM whose transition probabilities simulate a Lognormal duration distribution with the same parameters as the CSHMM model.

All results except the final two rows are reported with optimised 'language model scale factor' and 'word insertion penalty', with a similar optimisation for the CSHMM, a flat language model, and used the TIMIT boundaries with no forced alignment. The CSHMM results are similar (but no significance test has been applied), except for the Greenwood features, but the increased insertions suggests this may be improved. Disappointingly, the error for the CSHMM is slightly higher than for the DSHMM (reduced substitutions, increased insertions).

The final two rows show a larger improvement for the CSHMM with a bigram language model. There is room for caution here since the models are not fully equivalent, however, the CSHMM with perceptually-motivated features performs at least as well as the DSHMM trained on comparable MFCCs.

5. Discussion

Our experiments show that the best features for classification are those derived from experiments of human perception of speech ('perceptually-motivated'). This link between human and machine perception is interesting and raises questions such as, why should human and machine 'perception' of the speech signal behave in a similar way? Has speech developed such that its discriminatory features are adapted to human perception, and thus are inherently the information-bearing features for machine recognition? How can this knowledge improve ASR?

Figure 2 shows two examples of correlation matrices for /g/ and /sh/ for the Greenwood frequency bands. Spectral features are labelled 0 to 13 from low to high frequency, the right-hand column, bottom row is log duration. In colour, red signifies positive correlation, blue negative. Such matrices illustrate several interesting characteristics of the consonantal features. All spectral features are positively correlated, indicating that variations in loudness affect all frequencies. This is seen for all phonemes. Durations of /g/, /b/, /d/ and closures are negatively correlated with spectral energies, while /t/ duration and energy are positively correlated. In /p/ and /k/ only some spectral bands are negatively correlated with duration. These results concur with Khasanova et al. [8] who report similar strong correlations between burst duration and loudness, except for /p/ and /k/.

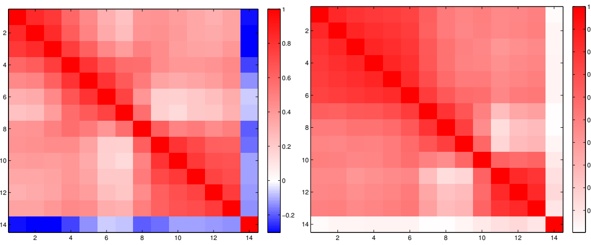

Figure 2: Correlation matrices (spectral and duration features) for /g/ (left panel) and /sh/ (right panel).

'Blocking' in Figure 2 is suggestive but inconclusive, roughly corresponding with the perceptual cues (Table 1). A distinct block in /sh/ in bands 10-13 suggests that energy in these bands varies together, independently of other bands, corresponding to the frequency cue for /sh/. However, there are also blocks in bands 1-6 and 7-10. In /g/ a large block around bands 4-8 (500-1649 Hz) is rather below and wider than the cue, but perhaps includes energy for the F_2 transition into the following vowel, also known to be important [2, Fig. 4].

Many other cues within and between speech sounds are important for perception and could be exploited for recognition. These include aspiration (e.g. [1]), duration of voicelessness in fricatives [22] and gaps between consonant and vowel onset [2, 3]. Correlations between sounds include duration of vowel preceding a consonant [6] and durations and amplitudes of closures and corresponding bursts. Formant transitions into following vowels have been shown to be critical for perception of stops [4, 5, 7, 23]. Syntactic phenomena such as a consonant's position in a word can also affect its acoustic features [8].

Within the CSHMM framework, correlations between phonemes may be naturally accommodated during recognition by modifying the next phoneme's targets based on features of the currently hypothesised phoneme. At present we assume one dwell phase per phoneme; multiple states could account for spectral changes within a phoneme. Finally, we propose to combine the CSHMM described here with the dwell-transition model for voiced sounds. Multiple hypotheses under both models would be probabilistically combined, the highest scoring hypothesis best explaining the observations including division into segments best explained by formant-like or spectral-like features. The parameters of the unvoiced model at the end of an unvoiced segment would influence the priors for the next segment, and vice versa. In particular, this allows the key effect of formant transitions to be modelled and to inform the decode.

6. Conclusion

Our aim is speech recognition based on faithful, parsimonious models of speech. We showed that linguistically meaningful features are suitable for a dwell-only continuous-state HMM to recognise unvoiced segments of speech. This model is appropriate for integration with the dwell-transition CSHMM for sonorant speech [10] and provides a natural framework for including other features of known perceptual importance, particularly when the two models are combined.

7. References

[1] K. N. Stevens, “Acoustic correlates of some phonetic categories,” *The Journal of the Acoustical Society of America*, 68(3), pp. 836–842, 1980.

[2] 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.

[3] 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.

[4] K. N. Stevens and S. E. Blumstein, “Invariant cues for place of articulation in stop consonants,” *The Journal of the Acoustical Society of America*, 64(5), pp. 1358–1368, 1978.

[5] J. M. Heinz and K. N. Stevens, “On the properties of voiceless fricative consonants,” *The Journal of the Acoustical Society of America*, 33(5), pp. 589–596, 1961.

[6] L. J. Raphael, “Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English,” *The Journal of the Acoustical Society of America*, 51(4B), pp. 1296–1303, 1972.

[7] L. F. Wilde, *Analysis and synthesis of fricative consonants*, Ph.D. thesis, Massachusetts Institute of Technology, 1995.

[8] A. Khasanova, J. Cole, and M. Hasegawa-Johnson, “Detecting articulatory compensation in acoustic data through linear regression modeling,” in *Proc. Interspeech*, Singapore, 2014.

[9] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “Darpa TIMIT acoustic-phonetic continuous speech corpus CD-ROM,” Tech. Rep., NIST, 1990.

[10] P. Weber, S. M. Houghton, C. J. Champion, M. J. Russell, and P. Jančovič, “Trajectory analysis of speech using continuous state hidden Markov models,” in *Proc. ICASSP*, pp. 3042–3046, Florence, Italy, 2014.

[11] C. J. Champion and S. M. Houghton, “Application of Continuous State Hidden Markov Models to a classical problem in speech recognition,” Accepted to *Computer Speech and Language*, 2015. doi:10.1016/j.csl.2015.05.001.

[12] J. N. Holmes, I. G Mattingly, and J. N. Shearme, “Speech synthesis by rule,” *Language and speech*, 7(3), pp. 127–143, 1964.

[13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, 2006.

[14] J. Schäfer and K. Strimmer, “A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics,” *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[15] O. Ledoit and M. Wolf, “A well-conditioned estimator for large-dimensional covariance matrices,” *Journal of Multivariate Analysis*, 88(2), pp. 365 – 411, 2004.

[16] 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. 2592–2605, 1990.

[17] D. D. O’Shaughnessy, *Speech communications - human and machine (2. ed.)*, IEEE, 2000.

[18] E. Zwicker, “Subdivision of the audible frequency range into critical bands (frequenzgruppen),” *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[19] B. C. J. Moore and B. R. Glasberg, “Suggested formulae for calculating auditoryfilter bandwidths and excitation patterns,” *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[20] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech and Signal Process.*, 28(4), pp. 357–366, 1980.

[21] M. Russell. A segmental HMM for speech pattern modelling. In *Acoustics, Speech, and Signal Processing, 1993. ICASSP-93., 1993 IEEE International Conference on*, vol. 2, pp. 499–502. 1993.

[22] K. N. Stevens, S. E. Blumstein, L. Glicksman, M. Burton and K. Kurowski, “Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters,” *The Journal of the Acoustical Society of America*, 91(5), pp. 2979–3000, 1992.

[23] J. D. W. Stephens and L. L. Holt, “A standard set of American-English voiced stop-consonant stimuli from morphed natural speech,” *Speech Communication*, 53(6), pp. 877–888, 2011.