

# Modelling Speech Signals using Formant Frequencies as an Intermediate Representation

Martin J Russell<sup>1</sup>, Xiaoyan Zheng<sup>1</sup> and Philip J B Jackson<sup>2</sup>

<sup>1</sup>Department of Electronic, Electrical and Computer Engineering,

The University of Birmingham, Birmingham, B15 2TT, U.K

m.j.russell@bham.ac.uk, xxz123@bham.ac.uk

<sup>2</sup>Centre for Vision, Speech and Signal Processing,

School of Electronics and Physical Sciences,

University of Surrey, Guildford GU2 7XH, UK

p.jackson@surrey.ac.uk

22nd November 2006

## Abstract

This paper concerns Multiple-level Segmental Hidden Markov Models (M-SHMMs) in which the relationship between symbolic and acoustic representations of speech is regulated by a formant-based intermediate representation. New TIMIT phone recognition results are presented, confirming that the theoretical upper-bound on performance is achieved provided that either the intermediate representation or the formant-to-acoustic mapping is sufficiently rich. The way in which M-SHMMs exploit formant-based information is also investigated, using singular value decomposition of the formant-to-acoustic mappings and linear discriminant analysis. The analysis shows that if the intermediate layer contains information which is linearly related to the spectral representation, that information is used in preference to explicit formant frequencies, even though the latter are useful for phone discrimination. In summary, while these results confirm the utility of M-SHMMs for automatic speech recognition, they provide empirical evidence of the value of non-linear formant-to-acoustic mappings.

Keywords: speech recognition segmental HMMs formants

# 1 Introduction

In the context of acoustic speech pattern modelling, one of the most commonly cited limitations of conventional hidden Markov models (HMMs) is their poor treatment of dynamics. A HMM assumes that a sequence of acoustic feature vectors  $y = y_1, \dots, y_T$  (each  $y_t$  is typically based on the short-term spectrum at time  $t$ ) is generated by a finite sequence of stationary random processes, corresponding to the HMM states. The probability of an individual vector  $y_t$  is assumed to depend on the state at time  $t$  but is otherwise independent of  $y_1, \dots, y_{t-1}, y_{t+1}, \dots, y_T$ . The incompatibility between this independence assumption and the need to take account of temporal dynamics in the acoustic model has resulted in the inclusion of ‘dynamic’  $\Delta$  (velocity) and  $\Delta^2$  (acceleration) parameters in augmented feature vectors. However, the resulting model is inconsistent in its assumption that the static, velocity and acceleration parameters are constant and possibly non-zero throughout a state occupancy. This motivates the ‘trajectory HMM’ [1] where a trajectory which is most consistent with the static and dynamic constraints is synthesised in the (static) feature vector space. Other researchers have adopted a more direct approach to modelling speech dynamics, either through the use of segment-level features [2], or segment models [3] in which states are associated with sequences of acoustic feature vectors. One option is to model a segment as a noisy trajectory in the acoustic feature space. Several types of trajectory have been studied, including constant [4, 5], linear [4], linear dynamical systems [6], ‘smoothed piecewise constant’ [7], non-parametric [8] and exponential [9]. More generally, Graphical Models [29] and Dynamic Bayesian Networks [28] offer a generic framework for characterising dependencies in speech patterns which extends HMMs.

A problem with acoustic-domain segmental models or features is the unsuitability of a

spectrum- or cepstrum-based representation for modelling dynamics. Trajectories which may be described simply in terms of articulator movement or formant transitions are realised as movement between, rather than within, frequency bands, resulting in complex paths in the acoustic feature space. Intuitively, it would be better to model dynamics directly, in a formant-based representation. Of course, the idea of including such an intermediate representation in an acoustic speech model is not new [10, 11, 7, 12, 13, 14, 15, 16].

This paper extends the work on ‘linear/linear’ Multiple-Level Segmental HMMs (M-SHMMs) presented in [15]. Our motivation for studying this system was to confirm that a conventional trajectory-based segmental HMM could be extended to include an intermediate, formant-based representation in which speech dynamics could be represented explicitly, without compromising performance. The linear/linear system is mathematically and computationally tractable and, because of its relationship with a conventional, ‘acoustic’, trajectory-based segmental HMM, an upper-bound on its performance is known [15]. We argued in [15] that a thorough understanding of this relatively simple model is an essential step towards the successful development of more complex multiple-level models incorporating, for example, non-linear formant-to-acoustic mappings.

In this paper new TIMIT phone recognition results are presented for both male and female speakers (the results in [15] refer to phone classification for male subjects only, and the extension to recognition is made computationally feasible by the techniques described in [24]). In addition, we present insights into the way in which M-SHMMs exploit the formant-based information in the intermediate layer, using singular value decomposition of the formant-to-acoustic mappings. This enables us to conclude that if information is available in the intermediate layer that is linearly related to the spectral representation, this is used in preference to formant frequency information, even though the latter contains

information which is useful for phone discrimination. In summary, while these results confirm the utility of M-SHMMs for speech recognition, they provide empirical evidence of the value of non-linear formant-to-acoustic mappings.

## 2 Multiple-level, trajectory-based segmental HMMs

In the M-SHMM described in [15] (figure 1), the relationship between symbolic and acoustic representations of speech is regulated through an intermediate formant-based layer. In what follows we assume that models correspond to phone-level units. In a linear/linear M-SHMM each state  $i$  is associated with a variable duration linear trajectory  $f$  in the intermediate layer  $I$ . A trajectory of length  $\tau$  corresponding to state  $i$  is mapped into the acoustic layer  $A$  by a phone-class-dependent, linear, formant-to-acoustic mapping  $W_i$ . The ‘synthetic’ acoustic feature vectors  $W_i(f(t))$  ( $t = 0, \dots, \tau - 1$ ) are treated as the means of Gaussian probability density functions (PDFs). The probability of a segment of acoustic feature vectors  $y = y_s, \dots, y_e$  of length  $\tau$  ( $e - s + 1 = \tau$ ) given state  $i$  is:

$$b_i(y) = d_i(\tau) \prod_{t=s}^e N_{(W_i(f(t)), \sigma_i)}(y_t) \quad (1)$$

where  $N_{(W_i(f(t)), \sigma_i)}$  is a multivariate Gaussian PDF with mean vector  $W_i(f(t))$  and covariance matrix  $\sigma_i$ , and  $d_i$  is the duration PDF associated with state  $i$ . In practice, the relationship between states and acoustic segments is determined by a segmental version of the Viterbi algorithm [15].

## 3 M-SHMM training

M-SHMM training is presented in [15] and only briefly described here. Training has two stages. First, one or more formant-to-acoustic mappings are estimated. Then the

model trajectory parameters and state duration and transition probabilities are optimised relative to these mappings.

### 3.1 Estimation of the formant-to-acoustic mappings

Linear, phone-class-dependent formant-to-acoustic mappings are estimated using ‘matched’ sequences of formant and acoustic data. Suppose that  $f_1, \dots, f_T$  and  $y_1, \dots, y_T$ , are two such sequences (in other words,  $f_t$  is a formant-based description of  $y_t$ ), and that the dimensions of these vectors are  $d_f$  and  $d_a$ , respectively. The goal is to find a  $d_a \times d_f$  matrix  $W$  such that the error

$$E = \sum_{t=1}^T \|W f_t - y_t\|^2 \quad (2)$$

is minimised. This is a standard problem in linear algebra (see, for example, [17]). In the experiments reported here, the acoustic vectors  $y_t$  consist of 13 cepstral features computed over overlapping 25ms sections of speech waveform, while the corresponding vector  $f_t$  is derived using the Holmes formant analyser [18] (see section 4).

In [15] five different formant-to-acoustic mapping schemes are considered, ranging from a single ‘phone-independent’ mapping to 49 ‘phone-dependent’ mappings. In what follows the notation  $W_i$  indicates dependency on state  $i$ .

### 3.2 Estimation of the model trajectory parameters

Once the mappings  $W_i$  have been derived the formant training data plays no further role. Given initial estimates of the remaining model parameters and a set of annotated training utterances, the segmental Viterbi algorithm [15] is used to ‘forced align’ the correct sequence of model states with each training utterance. In the acoustic domain, the differences between the trajectory values and the actual feature vectors are used to

estimate the covariance matrices  $\sigma_i$ . An acoustic segment corresponding to a state  $i$  is then ‘pulled back’ into the formant domain using the pseudo-inverse matrix  $W_i^\dagger$  [17], and used to estimate new values for the state trajectory mean and slope vectors and state duration PDF. The process is repeated until the improvement in the probability of the training data falls below a threshold, or a maximum number of iterations is reached.

## 4 The Holmes formant analyser

Formant-based intermediate representations of speech were computed using the J. N. Holmes formant analysis toolkit, which is described in [19] and [18]<sup>1</sup>. A short description is included here. Pitch-synchronous Fourier analysis transforms the speech signal, sampled at 8,000 samples per second, into a sequence of 31-dimensional spectral vectors. Each vector is compared with a set of ‘reference’ spectra, which have already been assigned one or possibly two sets of formant values, to give a ‘shortlist’ of reference spectra. A more careful analysis follows in which constrained, non-linear frequency warping is used to match the test spectrum with each shortlisted reference. Formant frequencies for the best matching reference, adjusted to take account of the frequency warp, are then assigned to the current spectrum. The Holmes formant analyser produces two alternative formant-based parameterisations. The first comprises estimates of the first three formant frequencies, together with ‘confidence’ estimates, alternative formant frequency values (when they are available), plus five amplitude measures associated with different frequency bands. Neither the confidence measure nor any alternative formant frequency estimates were used in the current study. The *3FF* and *3FF+5BE* representations consist

---

<sup>1</sup>The Holmes formant analysis and synthesis software is available from Aurix Limited, <http://www.aurix.com>

of the first three formant frequency estimates, and the first three formant frequencies plus the five amplitude measures, respectively. The final representation, *12PFS*, is computed separately and consists of the 12 Parallel Formant Synthesiser (PFS) [20] control parameters (table 1). The Holmes analyser currently returns fixed values for the first and twelfth parameters, and these were retained for compatibility with the parallel formant synthesiser.

Number	Parameter description
1	FN: Frequency of ‘low frequency’ formant (default value 250Hz)
2	ALF: Amplitude of FN in dB
3	F1: Frequency of first formant (in 25 Hz steps)
4	A1: Amplitude of first formant in dB
5	F2: Frequency of second formant (in 50 Hz steps)
6	A2: Amplitude of second formant in dB
7	F3: Frequency of third formant (in 50 Hz steps)
8	A3: Amplitude of third formant in dB
9	AHF: Amplitude in high frequency region (centred on 3500 Hz) in dB
10	V: Degree of voicing (1 = completely unvoiced, 63=fully voiced)
11	F0: Fundamental frequency on logarithmic scale
12	MS: Glottal-pulse mark-space ratio

Table 1: *The 12 Parallel Formant Synthesiser (PFS) control parameters in the 12PFS representation.*

## 5 Phone recognition experiments on the TIMIT speech corpus

The results on the TIMIT speech corpus [21] presented in [15] are limited to phone classification for male talkers. This section presents phone recognition results for male and female subjects.

### 5.1 Speech data

All experiments use the TIMIT corpus [21] downsampled to 8 kHz for compatibility with the formant analyser [18]. The male and female parts of TIMIT were both partitioned into three sets: a training set (speech from all speakers in the standard TIMIT training set except the first speaker in each dialect region), an evaluation set (all of the speech from the first speaker in each of the eight dialect regions), and a test set (speech from all speakers in the standard TIMIT test set (table 2)). Acoustic features (13 MFCCs,

	Training	Evaluation	Test
Male	(318) 121,400	(8) 3003	(112) 42,421
Female	(128) 49,570	(8) 3,107	(56) 21,724

Table 2: *Numbers of subjects (in brackets) and phones in the male and female training, evaluation and test sets.*

including the zeroth) were calculated using HTK (25 ms window, 10 ms fixed frame rate) [22]. The three formant-based parameterisations each have a 10 ms frame rate and hence are synchronous with the acoustic data. The vectors were augmented with an additional ‘bias’ value which was set to 1 (this enables the result of the linear mapping to include

a constant term). Thus the dimensions of the augmented  $3FF$ ,  $3FF+5BE$  and  $12PFS$  representations are  $d_f = 4, 9$  and  $13$  respectively.

Linear articulatory-to-acoustic mappings were estimated as described in section 3.1 and [15].

## 5.2 Phone categories

Following [15], sets of formant-to-acoustic mappings  $\{W_1, \dots, W_K\}$  were obtained for each of the following categorisations of the TIMIT phone set (the number of mappings  $K$  is given in parentheses): A. (1) all data; B. (6) linguistic categories; C. (10) as in [13]; D. (10) discrete articulatory regions [23]; E. (49) individual phones.

The rationale for categorizations B, C and D is given in [15]). The categories are defined in table 3

B	<i>vowels</i> , {hh, l, r, w, y}, <i>nasals</i> , {dh, f, s, sh, th, v, z, zh}, {ch, jh}, {b, cl, d, vcl, dx, epi, g, k, p, q, sil, t}
C	<i>vowels</i> , {epi, q, sil}, {hh, l, r, w, y}, <i>nasals</i> , {sh, f, th}, {ch, s} {dh, v, zh}, {jh, z}, {cl, k, p, t}, {b, d, vcl, dx, g}
D	<i>vowels</i> , {epi, q, sil}, {dx, el, l, r, w, y}, <i>nasals</i> , {vcl}, {cl}, {b, d, g, jh}, {ch, k, p, q, t}, {dh, v, z, zh}, {f, hh, s, sh, th}

Table 3: *Definitions of the phone partitions B, C, and D. The terms ‘vowels’ and ‘nasals’ denote the sets {aa, ae, ah, ao, aw, ax, ay, eh, el, er, ey, ih, ix, iy, ow, oy, uh, uw}, and {en, m, n, ng}, respectively)*

### 5.3 M-SHMM training

Conventional, three-state Gaussian monophone HMMs were created for each symbol in the TIMIT 49-phone set using HTK [22], and used to seed a set of M-SHMMs. The M-SHMM (formant domain) state trajectory means were set to be the pseudo-inverse images of the corresponding HMM state means under the appropriate formant-to-acoustic mapping, the state trajectory slopes were set to zero, the (acoustic) M-HSMM state variance vectors were set equal to the corresponding HMM state variance vector, and the (non-parametric) state duration PDF was uniform. The maximum state duration was set to 15 frames ( $\tau_{\max} = 15$ ), which is sufficient to accommodate all TIMIT phone labels. The appropriate formant-to-acoustic mapping and its pseudo-inverse were then added to the model [15]. The parameters of the resulting M-SHMMs were trained using segmental Viterbi re-estimation.

The resulting monophone M-SHMMs were used to seed a set of triphone MSHMMs, selected according to a ‘backoff’ scheme described in [15]. These were re-estimated using further iterations of Viterbi training.

### 5.4 Language model

A phone-level probabilistic bigram language model was estimated using all of the TIMIT label files in the training set. A language model scale factor regulated the effect of the language model on recognition.

### 5.5 Recognition

Phone recognition experiments were conducted using the segmental Viterbi decoder and the phone-level bigram language model from section 5.4. The computational load was

made viable using segmental ‘beam pruning’ [24]. The unsmoothed, triphone-dependent state duration PDFs include many duration probabilities of zero. Since segments with zero probability durations are automatically discarded during Viterbi decoding, the more sophisticated ‘duration pruning’ described in [24] is not needed.

Language model and duration scale factors and the beam pruning threshold were determined empirically on the evaluation set.

## 5.6 Experiments

Separate, gender-dependent sets of ‘male’ and ‘female’ formant-to-acoustic mappings were estimated for each of the formant-based representations (section 5.1) and each phone categorisation schemes (section 5.2). In the first experiment, these were used in separate male and female triphone M-SHMM sets. With a backoff threshold of 30, this resulted in 1364 male and 660 female triphone M-SHMMs. Phone recognition experiments were conducted on the male and female test data using these gender-dependent models. We refer to these conditions as  $M-M-M$  and  $F-F-F$ , where  $X-Y-Z$  denotes the use of the ‘gender  $X$ ’ training set for model parameter estimation (3.2), the ‘gender  $Y$ ’ formant-to-acoustic mapping (3.1), and testing on the ‘gender  $Z$ ’ test set.

We also investigated conditions  $M-M-F$ ,  $F-F-M$ ,  $M-F-M$  and  $F-M-F$ . The first two are ‘fully cross gender’ experiments, while the remaining conditions combine gender-dependent training with a ‘cross gender’ formant-to-acoustic’ mapping. These experiments are interesting because they indicate how much of the ‘gender specificity’ of the models is due to the formant-to-acoustic mappings and how much is due to the formant trajectory parameters.

### 5.6.1 Within-gender experiments

Figure 2 presents phone recognition results for the ‘within-gender’ experiments ( $M-M-M$  and  $F-F-F$ ). The ‘baseline’ phone error rates for fixed-linear-trajectory SHMMs [25] with no intermediate layer are 36.6% for the male test data and 38.9% (a relative increase of approximately 5.5%) for the female data. As explained in [15], these are of interest because they are lower-bounds for the M-SHMM error rates. The results for M-SHMMs with the three different intermediate representations (3FF, 3FF+5BE and 12PFS) and different phone categorization schemes (A - E) follow a similar pattern to the phone classification results for male speech presented in [15], with performance improving either as the dimension of the intermediate representation or the number of mappings is increased.

In the case of female speech, performance is similar to the baseline for all combinations of the two higher-dimensional intermediate representations  $3FF+5BE$  and  $12PFS$  and phone categorization schemes A to E. This indicates that the inclusion of either of these intermediate representations as an intermediate layer in a M-SHMM does not compromise performance for female speech. In particular, there is no indication of problems due to formant analysis errors for female speech. For the male test data the  $12PFS$  intermediate representation gives better results than the  $3FF+5BE$  representation, with phone categorization scheme E giving the best result (36.8%). For comparison, the best phone recognition performance achieved on the full TIMIT core test set (male and female speakers) using optimised conventional HMMs is 27.1% phone errors [27].

### 5.6.2 Cross-gender experiments

As one would expect, the results of the ‘fully cross-gender’ experiments are uniformly poor, varying between 55% and 59.2% (average 57%) phone error rate for the  $M-M-F$

experiments and between 58.3% and 61.3% (average 59.9%) for *F-F-M* experiments.

Figure 3 presents the results of the *M-F-M* and *F-M-F* experiments. The white columns, *M-F-M*, show the percentage increase in phone error rate when formant-to-acoustic mappings trained on female speech are used in M-SHMMs trained and tested on male speech. The black columns show the corresponding results when formant-to-acoustic mappings trained on male speech are used in M-SHMMs trained and tested on female speech. In the majority of cases the phone error rate increases when the ‘wrong’ mapping is used. However, in the single mapping cases (scheme A) the ‘male’ mapping works well for female speech. The same is not true when the single ‘female’ mapping is used for male speech, but this may be due to the smaller female training set. Figure 3 also shows that the phone error rate difference increases as the number of mappings increases. It seems that increasing the number of phone classes raises the utility of the phone class information, leading to the reductions in phone error shown in figure 2. However, figure 3 suggests that the way in which this phone-class information is exploited is gender-dependent, and does not generalise to the ‘cross-gender’ mappings. For example, in the case of the *M-F-M* experiment, as the number of mappings is increased from 1 to 49 the error rate increases from 7.3% to 20.8%.

## 6 Analysis of the formant-to-acoustic mappings

The Singular Value Decomposition (SVD) of a formant-to-acoustic mapping  $W$  is  $W = USV^T$ , where  $V$  is a  $d_f \times d_f$  orthogonal matrix,  $S$  is a  $d_a \times d_f$  diagonal matrix whose elements, called the singular values of  $W_i$  are real and non-negative, and  $U$  is a  $d_a \times d_a$  orthogonal matrix. It is usual to write the singular values in decreasing order. Suppose that  $u_i$  and  $v_i$  denote the  $i^{th}$  column of  $U$  and  $V$  respectively. Then  $\{u_1, \dots, u_{d_a}\}$  and

$\{v_1, \dots, v_{d_f}\}$  are orthonormal bases for the acoustic and intermediate vector spaces, respectively, and  $Wv_i = s_i u_i$  for each  $i = 1, \dots, d_f$ . In particular, the columns of  $V$  which correspond to zero-valued singular values form a basis for the null space of  $W$ , which is the subspace of the intermediate space comprising vectors  $f$  such that  $Wf = 0$ . In other words, the null space of  $W$  is the part of the intermediate space which makes no contribution to the construction of acoustic vectors from formant-based vectors. The rank of  $W$  is the number of non-zero singular values of  $W$ , which is equal to  $d_f$  minus the dimension of the null space of  $W$ . The rank of  $W$  is the dimension of the image of the intermediate space under  $W$  in the acoustic space.

SVD was applied to the male formant-to-acoustic mappings for each of the three formant-based representations, and mapping scheme A. For this analysis, each parameter in the 3FF, 3FF+5BE and 12PFS representations was normalised to have zero mean and unit variance prior to estimation of the mapping. Figure 4 shows the singular values of the formant-to-acoustic mappings for the 3FF (a) and 3FF+5BE (b) and 12PFS (c) intermediate representations.

Figure 5 shows the basis vectors  $v_1, \dots, v_4$  corresponding to figure 4(a) for the 3FF representation. From the first graph in figure 5, labelled SV1, it is clear that basis vector  $v_1$ , corresponding to the largest singular value, is aligned with the 3FF constant 4<sup>th</sup> parameter. From graphs SV2 and SV3,  $v_2$  and  $v_3$  lie in the subspaces spanned by F1 and F2 and F1 and F3, respectively, and, from SV4,  $v_4$  corresponds approximately to F3. Thus in the case of the 3FF intermediate representation the mapping is based on a constant MFCC vector  $Wv_1$ , which is modified according to the values of the three formant frequencies.

In the case of the 3FF+5BE intermediate representation, figure 4(b) shows that the final 3 singular values are close to zero. Figure 6 shows clearly that the corresponding

singular vectors lie in the subspace of the  $3FF+5BE$  representation spanned by the formant frequencies ( $3FF+5BE$  parameters 1, 2 and 3). This indicates that the formant frequency values are essentially ignored in the reconstruction of the acoustic representation. The most significant singular vector  $v_1$  corresponds to the fixed 9th ‘bias’ parameter, indicating that, as with the previous representation, the offset provided by the constant parameter is a major component of the formant-to-acoustic mapping, and that the five band energies are used to construct acoustic vectors relative to this offset.

Turning to the  $12PFS$  representation, figure 4(c) shows that singular values 9 and above are essentially zero. Hence the singular vectors  $v_9, \dots, v_{13}$  play little part in formant-to-acoustic mapping. Figure 7 shows the basis vectors  $v_1, \dots, v_8$ . As with the  $3FF$  and  $3FF+5BE$  representations, the first singular vector corresponds to the constant ‘bias’ parameter, which is used to establish a basic acoustic feature vector ‘template’ around which the formant-to-acoustic mapping is built. However, in this case the formant frequencies (parameters 3 (F1), 5 (F2) and 7 (F3)) do contribute to the mapping. The second singular vector,  $v_2$  includes components in the directions of F1 and A1 (and, to a lesser extent, F2 and A2), although its most significant component is in the direction of the amplitude of the low-frequency region. Vectors  $v_4, v_5$  and  $v_6$  include components aligned with F2, F2, A2, F3 and A3, and A1, F2 and A2 respectively.

In summary, if sufficient frequency band energy information is available, as in the  $3FF+5BE$  representation, then no use is made by the formant-to-acoustic mappings of explicit formant frequency data. However, in the case of the more abstract  $12PFS$  representation (which gives the best phone recognition results), the formant frequency and amplitude information is being used together with the other information to construct the acoustic representation. However, even in the case of the  $12PFS$  representation the mapping makes significant use of the low frequency band energy.

It is also interesting to note that the rank of the formant-to-acoustic mappings, which is the dimension of the image of the intermediate space in the acoustic space, is between 6 and 8.

## 7 Linear discriminant analysis (LDA) of the formant data

The results from the previous section question the utility of the explicit formant data used in the experiments for phone classification. To investigate this further we applied Linear Discriminant Analysis (LDA) to the *3FF+5BE* data in the male training set to determine which components contribute most to phone discrimination. This requires an eigenvector decomposition of the matrix  $\Sigma_b \Sigma_w^{-1}$ , where  $\Sigma_b$  and  $\Sigma_w$  are the between-class and average within-class covariance matrices of the data, respectively (here ‘class’ means monophone class in the 49 phone TIMIT set). For this experiment the bias term was removed, leaving an 8 dimensional representation, and the remaining parameters were normalised as in the previous section. Figure 8 shows the values of the eigenvalues associated with the LDA discriminative vectors and figure 9 shows the vectors themselves. Recall that the first three of the *3FF+5BE* components are formant frequencies and the remaining five are the band energies.

The first, most significant, LDA eigenvector in figure 9 (whose eigenvalue is approximately 27% greater than the second eigenvalue) is very similar to the second singular vector in figure 6, except that while the components in the directions of the three formant frequency parameters in the singular vector are close zero, the corresponding values for F2 and F3 in the first LDA vector are non-zero. However, both vectors are dominated by

the first three band energies. Each of the first six LDA eigenvectors contains at least one non-zero component in the direction of one of the formant frequencies, whereas the first singular vector to include a significant formant frequency component is the sixth, which contains a large component in the direction of F2. However, we have observed previously from figure 4(b) that the corresponding singular value is close to zero and hence this vector is close to the null space of the formant-to-acoustic mapping. Note that in the first three LDA eigenvectors there is a correspondence between the selected formant and the dominant frequency band energy.

In summary we now have two pieces of evidence regarding the utility of formant-based information as an intermediate representation in our model. The LDA results of this section for the  $3FF+5BE$  parameterisation confirm, to some extent, that the formant frequency values contain information which is useful for phone classification. However, the results of section 6 indicate that the strategy employed by the formant-to-acoustic mappings, to construct acoustic feature vectors from the  $3FF+5BE$  intermediate representation, essentially ignores the formant frequency information completely. Instead the mapping uses the band-energy information in the  $3FF+5BE$  parameterisation, which is linearly related to the acoustic representation. This suggests that use of formant information is compromised by the linearity assumption in our formant-to-acoustic mappings. We interpret this as empirical evidence of the need for non-linear formant-to-acoustic mappings in our model.

Of course, the utility of the formant frequency information may be compromised by errors in the automatic formant analysis. However, results presented in [15] suggest that this is not the case, since discarding formant data values with low ‘confidence’ had little effect on recognition accuracy. A standard, hand-corrected, dataset for evaluating vocal tract resonance frequency estimation has recently been announced [26]. When this

becomes available it will provide a benchmark against which our formant data can be compared.

## 8 Conclusions

In a linear-linear M-SHMM [15] the relationship between symbolic and acoustic representations of speech is regulated by an intermediate, formant-based representation. Dynamics in the intermediate representation are characterised by linear trajectories, and the relationship between the formant-based and acoustic representations is assumed to be linear.

This paper presents the results of M-SHMM phone recognition experiments on the male and female TIMIT test sets. The increased search space, for recognition compared with classification, necessitates the use of the pruning techniques described in [24]. The results for male and female test data follow similar patterns to those reported in [15], with best performance obtained by increasing the ‘richness’ of the intermediate layer or the formant-to-acoustic mapping. When these are chosen appropriately recognition accuracy achieves its theoretical upper bound. In particular, the introduction of an intermediate layer does not compromise performance and there is no evidence that performance for female speech is affected by difficulties with formant analysis.

The paper also reports the results of ‘cross-gender’ experiments, in which M-SHMMs trained and tested on male speech incorporate female formant-to-acoustic mappings ( $M-F-M$ ), and M-SHMMs trained and tested on female speech use male mappings ( $F-M-F$ ). In general using the ‘wrong’ mapping results in more errors. However, for each representation and gender, error rate increases as the number of formant-to-acoustic mappings is increased, indicating that the phone-class-dependent information which is made available

by increasing the number of mappings does not generalise across genders.

The goal of the remainder of the paper is to understand how M-SHMMs encode the relationships between formant-based and acoustic representations of speech. Section 6 presents an analysis of the single-mapping systems (scheme A) for each intermediate representation and male speakers, based on singular value decomposition of the mappings. The results show that different strategies are being used for the *3FF+5BE* and *12PFS* representations. In the case of the *3FF+5BE* representation, no use is made of the explicit formant frequency information, and the mapping relies instead on frequency band energy. In the case of the *12PFS* representation, more use is made of the explicit formant data, though the mapping still appears to prefer band-energy type data (such as the low frequency energy parameter) where it is available. Two possible explanations are that the explicit formant data does not contain information which is sufficiently useful for phone classification (for example, due to formant analysis errors) or that the use of linear formant-to-acoustic mappings precludes the use of the formant information. In order to clarify this, Linear Discriminant Analysis was applied to the *3FF+5BE* parameters. The results are presented in section 7. The most significant singular vector (excluding the one which corresponds to the constant ‘bias’ term) in the decomposition of the ‘formant-to-acoustic’ mapping is very similar to the most significant LDA vector, except that the latter makes use of formant frequency information while the former does not. In fact, while none of the significant singular vectors in the decomposition of the mapping include formant frequency components, this information is present to some extent in most of the LDA vectors.

This analysis confirms that the formant frequencies do contain information which is relevant for phone classification. Therefore, we conclude that the tendency of our model not to make substantial use of explicit formant frequency information is due to

the constraint that the formant-to-acoustic mappings are linear, and that the results presented in this paper provide empirical evidence of the need for non-linear formant-to-acoustic mappings.

We believe that the results presented in [15] and the present paper have not only achieved the goals set out in the introduction, but have also demonstrated empirically the limitations of our model and the need to acknowledge explicitly the non-linear nature of the relationship between formant and spectral representations of speech.

## 9 Acknowledgements

This work was funded partly by GCHQ Cheltenham, who the authors would like to acknowledge and thank for their funding and support, and partly by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant EP/515986 “A unified model for speech recognition and synthesis”.

## References

- [1] Tokuda, K., Zen, H. and Kitamura, T., Trajectory modelling based on HMMs with the explicit relationship between static and dynamic features, Proc. Eurospeech '03, Geneva, Switzerland, 2003.
- [2] Glass, J., A probabilistic framework for segment-based speech recognition, *Comp. Speech & Lang.*, 17, 2-3, April-July 2003, pp 137-152.

- [3] Ostendorf, M., Digalakis, V. V. and Kimball, O. A., From HMM's to segmental models: a unified view of stochastic modeling for speech recognition, *IEEE Trans. on Spch. & Aud. Proc.*, 1996, 4, 5, pp 360-378.
- [4] Russell, M. J., A segmental HMM for speech pattern modelling, *Proc. IEEE-ICASSP*, Minneapolis, MN, 1993, pp 499-502.
- [5] Gales, M. J. F. and Young, S. J., Segmental Hidden Markov Models, *Proc. Eurospeech '93*, Berlin, Germany, 1993, pp 1579-1582.
- [6] Digalakis, V., Segment-based stochastic models of spectral dynamics for continuous speech recognition, PhD thesis, 1992, Boston University, Boston, MA, USA.
- [7] Richards, H. B. and Bridle, J. S., The HDM: a segmental Hidden Dynamic Model of coarticulation, *Proc. IEEE-ICASSP*, Phoenix, AZ, 1999, pp 357-360.
- [8] Ghitza, O. and Sondhi, M. M., Hidden Markov models with templates as non-stationary states: an application to speech recognition, *Comp. Speech & Lang.*, 1993, 2, pp 101-119.
- [9] Wiewiorka, A. and Brookes, D. M., Exponential interpolation of states in a hidden Markov model, *Proc. Institute of Acoustics*, 18, 9, 1996, pp 201-208.
- [10] Deng, L. and Braam, D., Context-dependent Markov model structured by locus equations: Applications to phonetic classification, *J. Acoust. Soc. Am.*, 1994, 108(6), pp 2008-2025.
- [11] Deng, L., A dynamic, feature-based approach to the interface between phonology and phonetics for speech modelling and recognition, *Speech Communication*, 1998, 24(4), pp 288-323.

- [12] Gao, Y., Bakis, R., Huang, J. and Zhang, B., Multistage coarticulation model combining articulatory, formant and cepstral features, Proc. Int. Conf. on Spoken Lang. Proc. Beijing, 1, 2000, pp 25-28.
- [13] Deng, L. and Ma, J., Spontaneous speech recognition using a statistical coarticulatory model for the vocal-tract-resonance dynamics, J. Acoust. Soc. Am., 2000, 108, pp 3036-3048.
- [14] Zhou, J., Seide, F. and Deng, L., Coarticulation modelling by embedding a target-directed hidden trajectory model into HMM - modelling and training. Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. Hong Kong, 2003, 1, pp 744-747.
- [15] Russell, M. J. and Jackson, P. J. B., A multiple-level linear/linear segmental HMM with a formant-based intermediate layer, Comp. Speech & Lang., 2005, 19, 2, pp 205-225.
- [16] Yu, D. Deng, L. and Acero, A., Evaluation of a long-contextual-spen trajectory model and phonetic recognizer using A\* lattice search, Proc. Interspeech, Lisbon, Portugal, 2005.
- [17] Bishop, C. M., Neural networks for pattern recognition, (OUP, Oxford, UK, 1995).
- [18] Holmes, J. N., Speech processing system using formant analysis, US Patent US6292775, 2001.
- [19] Holmes, J. N., Robust measurement of fundamental frequency and degree of voicing, Proc. Int. Conf. on Spoken Lang. Proc., Sydney, Australia, 1998.
- [20] Holmes, J. N., Mattingly, I. G. and Shearme, J. N., Speech synthesis by rule, Language & Speech, 7, 1964, pp 127-143.

- [21] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L. and Zue, V., TIMIT Acoustic-Phonetic Continuous Speech Corpus, LDC Catalog No.: LDC93S1, Linguistic Data Consortium, Univ. Pennsylvania, Philadelphia, PA, USA, 1993.
- [22] Young, S. J., Odell, J., Ollason, D., Valtchev, V. and Woodland, P., The HTK Book, Entropic Camb. Res. Lab., Cambridge, UK, 1997.
- [23] Jackson, P. J. B., Lo, B.-H. and Russell, M. J., Data-driven, non-linear, formant-to-acoustic mapping for ASR, *El. Lett.*, 2002, 38, 13, pp 667-669.
- [24] Russell, M. J., Reducing computational load in segmental HMM decoding for speech recognition, *El. Lett.*, 2005, 41, 25, pp 1408-1409.
- [25] Holmes, W. J. and Russell, M. J., Probabilistic-trajectory segmental HMMs, *Comp. Speech & Lang.*, 1999, 13, 1, pp 3-37.
- [26] Li Deng, Xiaodong Cui, Robert Pruvencok, Jonathan Huang, Safiyy Momen, Yanyi Chen and Abeer Alwan, A database of vocal tract resonance trajectories for research in speech processing, *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Proc. ICASSP 2006*, Toulouse, France, 2006, pp I-369 - I-372.
- [27] Lamel, L. F. and Gauvain, J. L., High Performance Speaker-Independent Phone Recognition Using CDHMM, *Proc. EUROSPEECH'93*, 1993, pp 121-124.
- [28] Jensen, F. V., *An Introduction to Bayesian Networks*, Springer, Berlin, 1996.
- [29] Lauritzen, S. L., *Graphical Models*, Oxford Science Publications, 1996.

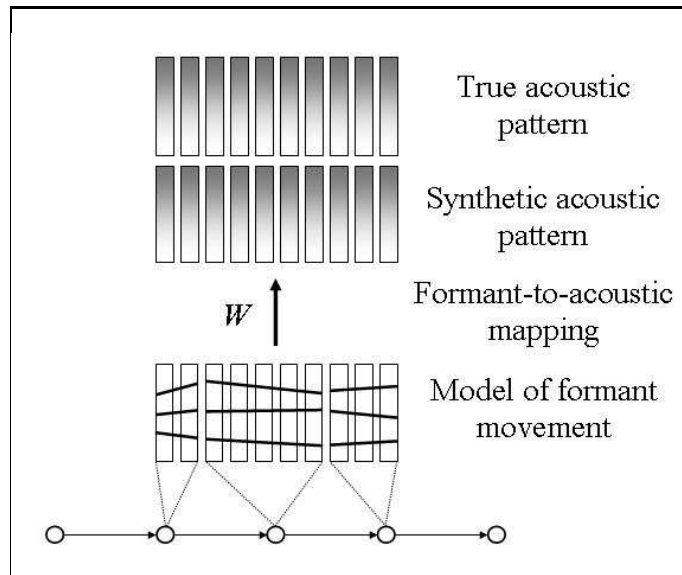


Figure 1: A multiple-level segmental HMM (M-SHMM) in which the relationship between the symbolic and acoustic representations of a speech signal is regulated by an intermediate formant-based layer

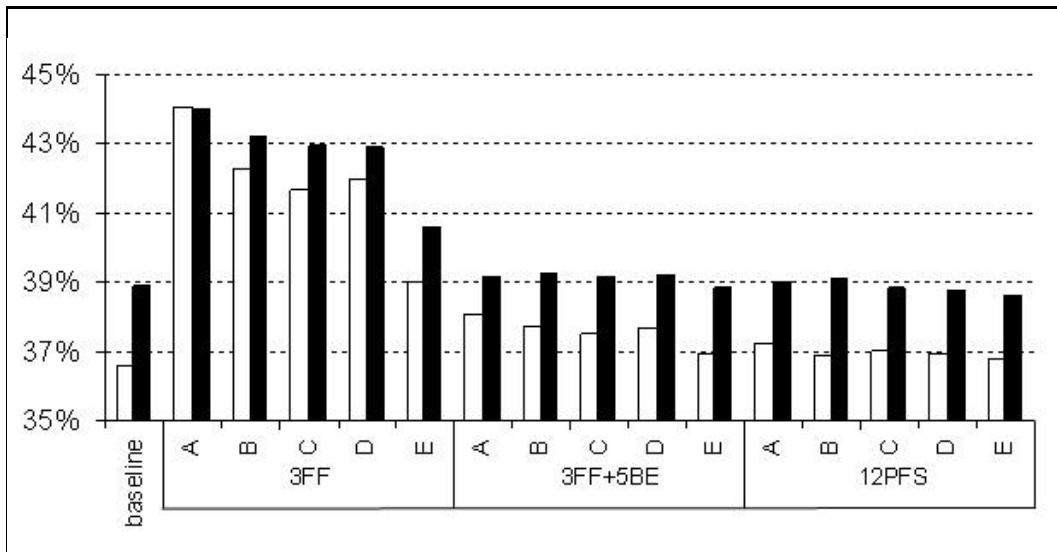


Figure 2: Percentage phone errors for ‘within-gender’ experiments M-M-M (white) and F-F-F (black) for formant-based intermediate representations 3FF, 3FF+5BE and 12PFS, and mapping schemes A-E. The columns on the left are for the baseline with no intermediate representation.

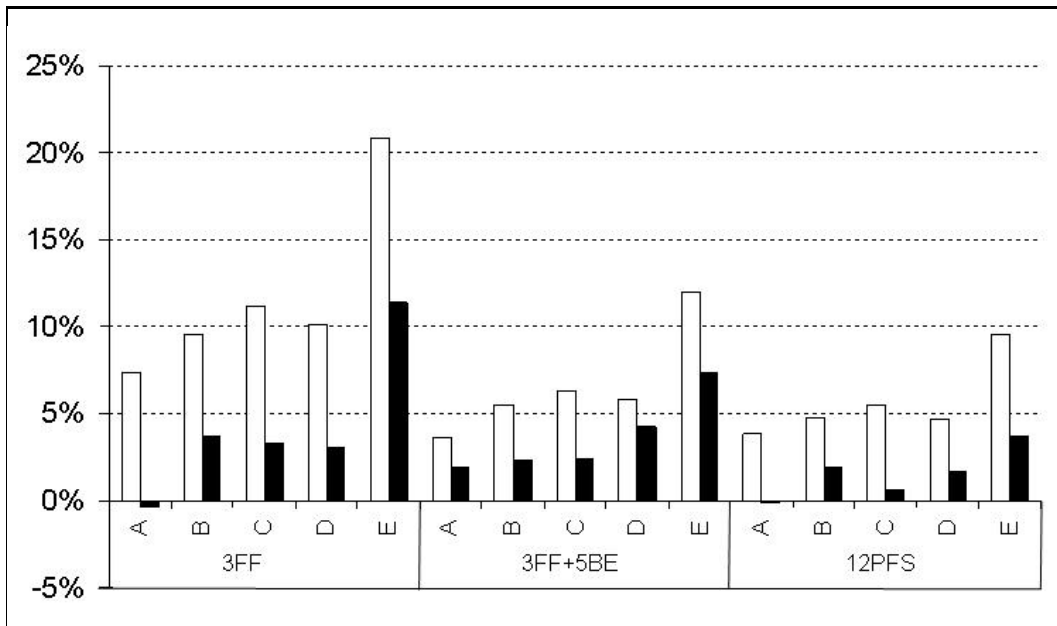


Figure 3: *Absolute percentage increase in phone errors due to the use of a ‘cross-gender’ formant-to-acoustic mapping in experiments M-F-M (white) and F-M-F (black), for formant-based intermediate representations 3FF, 3FF+5BE and 12PFS, and mapping schemes A-E.*

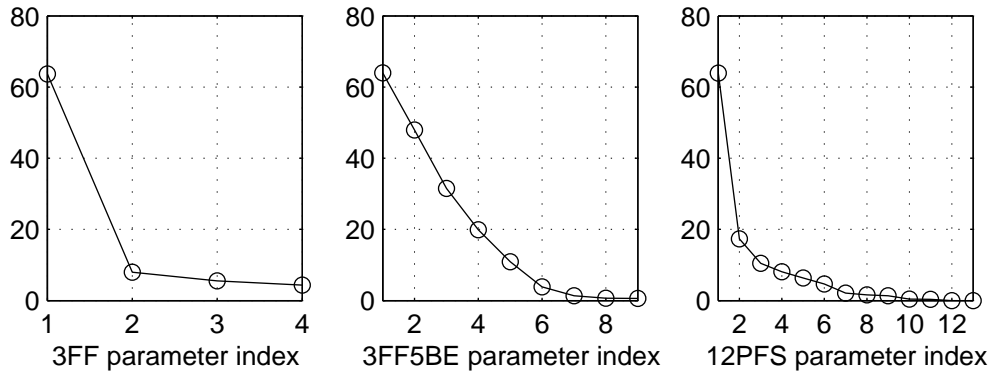


Figure 4: *Singular values of the formant-to-acoustic mappings for the (a) 3FF, (b) 3FF+5BE and (c) 12PFS intermediate representation and phone categorisation scheme A (one mapping), for male data.*

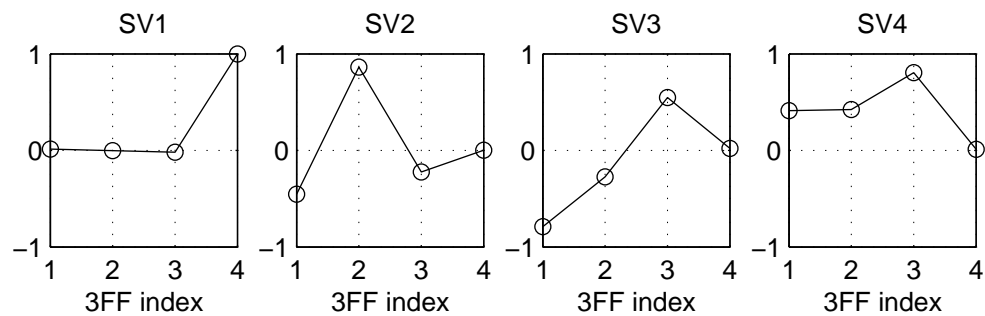


Figure 5: *Singular vectors (ordered according to figure 4(a)) for the ‘male’ formant-to-acoustic mapping, 3FF intermediate representation and phone categorisation scheme A. The horizontal axes correspond to the 4 parameters of the 3FF representation.*

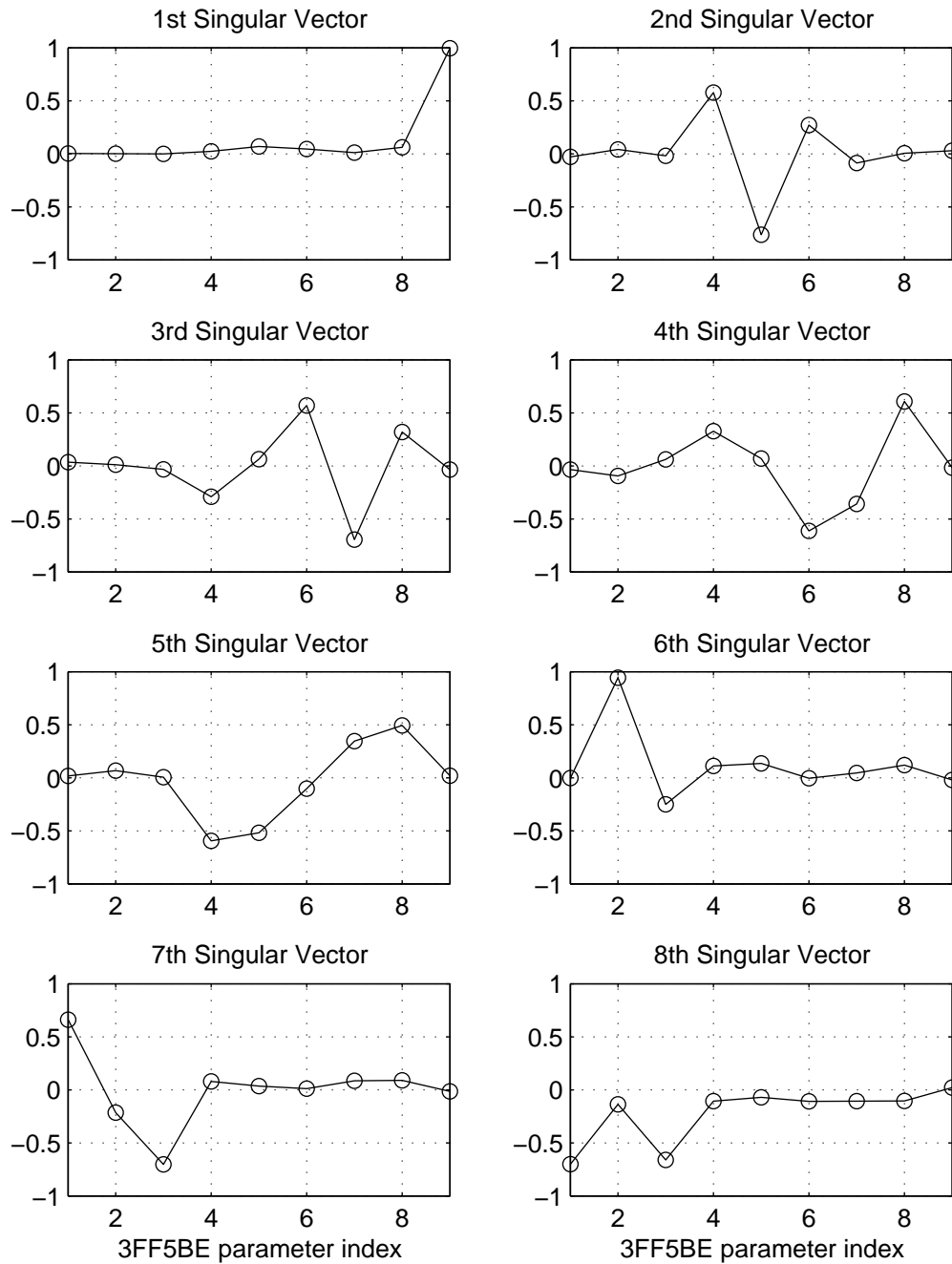


Figure 6: *Singular vectors (ordered left-to-right, top-to-bottom according to figure 4(b)) for the ‘male’ formant-to-acoustic mapping, 3FF+5BE intermediate representation and phone categorisation scheme A. The horizontal axes correspond to the 9 parameters of the 3FF+5BE representation.*

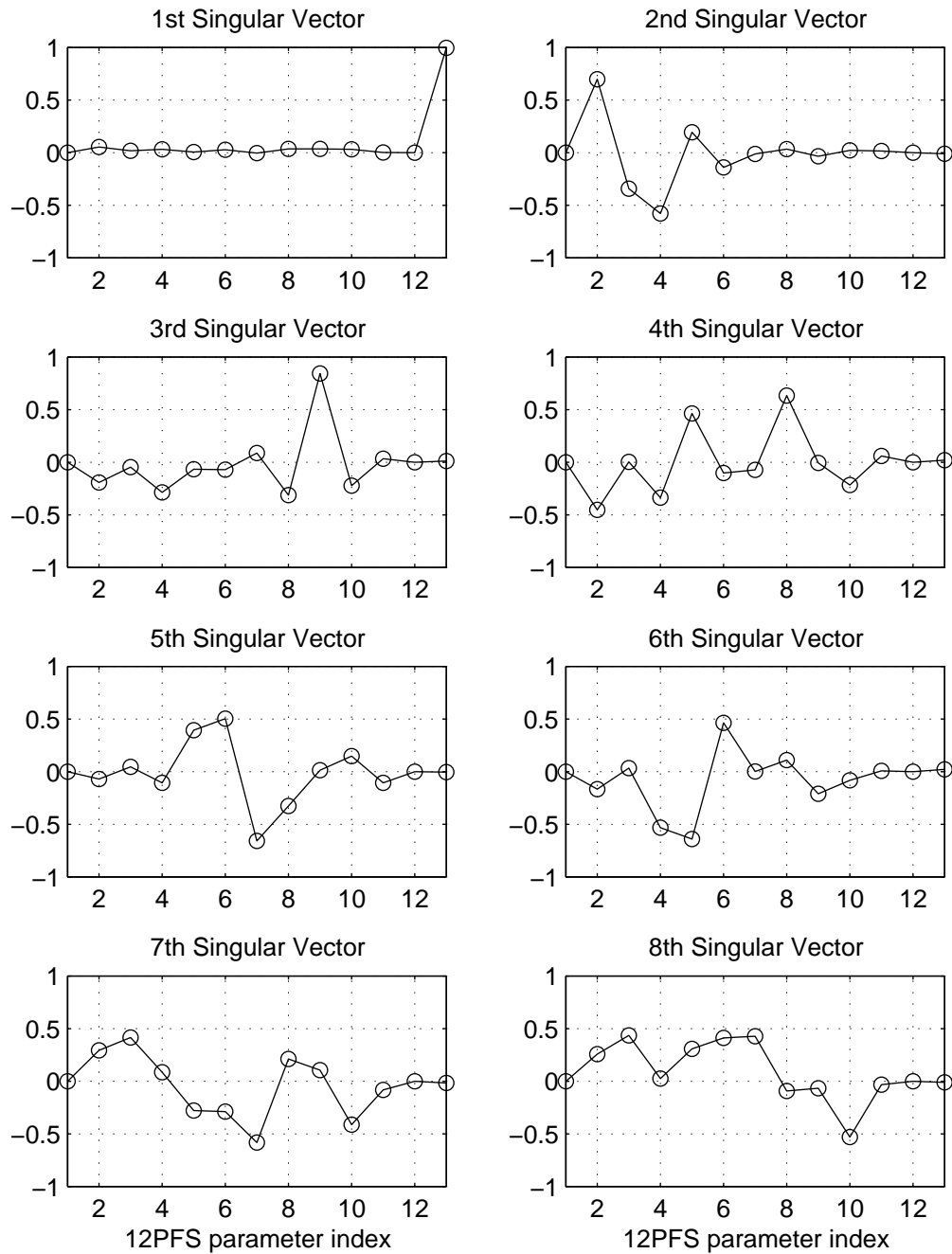


Figure 7: Singular vectors  $v_1$  to  $v_8$  (ordered left-to-right, top-to-bottom according to figure 4(c)) for the ‘male’ formant-to-acoustic mapping, 12PFS intermediate representation and phone categorisation scheme A. The horizontal axes correspond to the 13 parameters of the 12PFS representation.

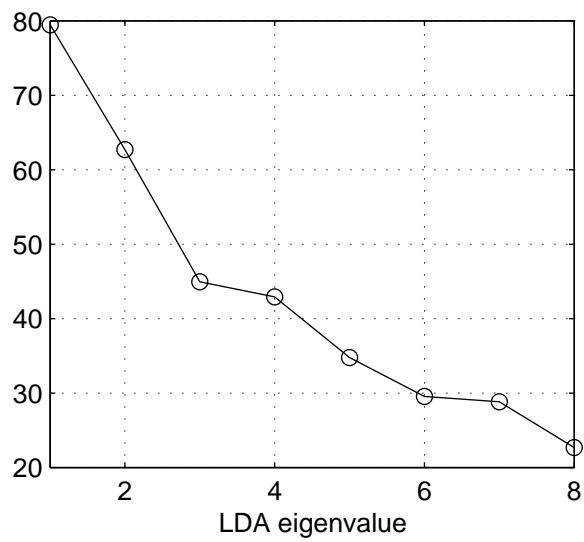


Figure 8: *Eigenvalues for LDA applied to the 3FF+5BE representation and monophone classes(male speakers).*

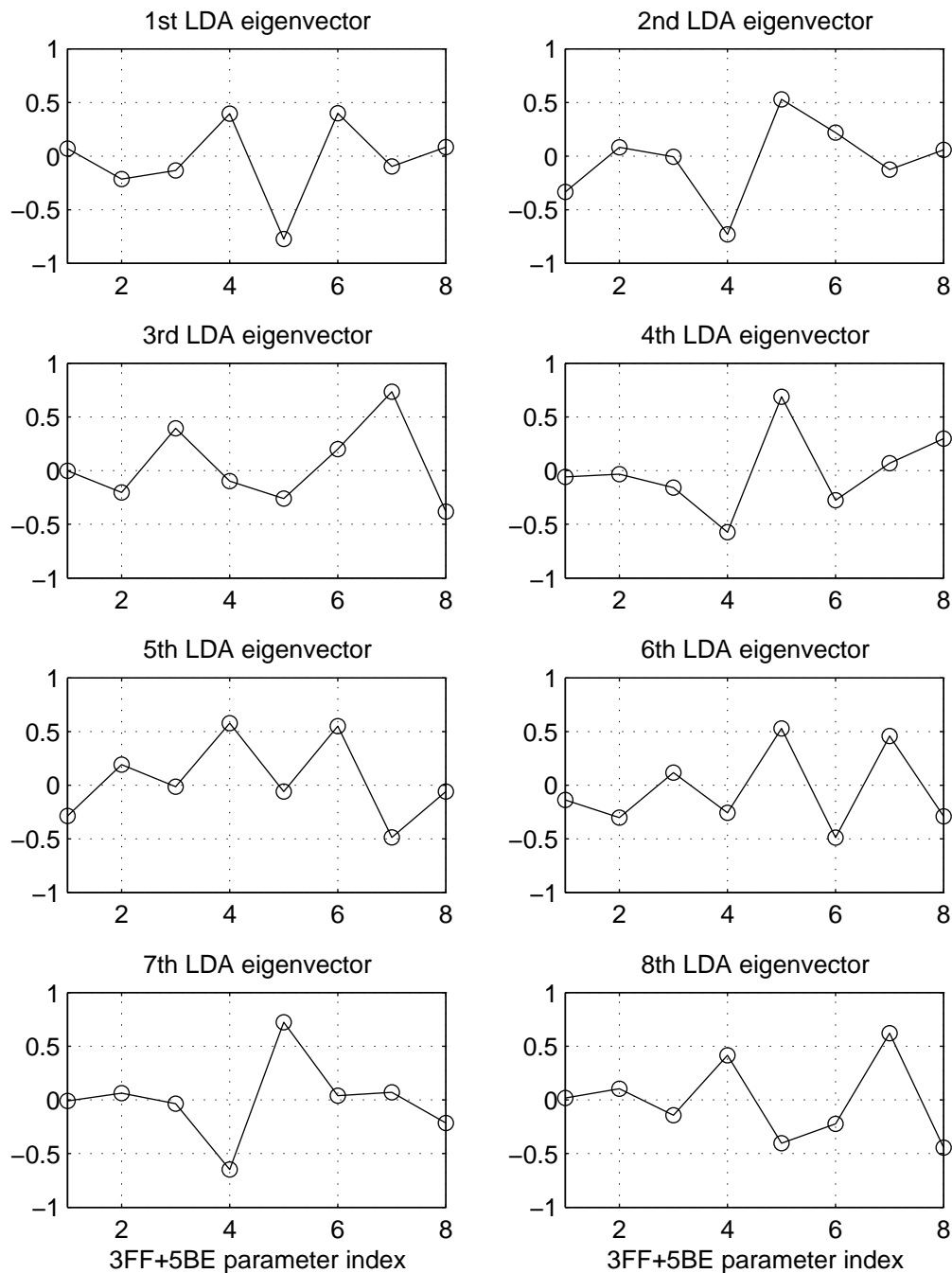


Figure 9: *LDA eigenvectors for the 3FF+5BE intermediate representation and monophone classes. The horizontal axes correspond to the 8 parameters of the 3FF+5BE representation (the 13<sup>th</sup> ‘bias’ parameter has been omitted).*