PREDICTING FORMANT FREQUENCIES FROM MFCC VECTORS

Jonathan Darch[1], *Ben Milner*[1], *Xu Shao*[1], *Saeed Vaseghi*[2] *and Qin Yan*[2]

[1]School of Computing Sciences, University of East Anglia, Norwich, U.K.
[2]Dept. of Electronic and Computing Engineering, Brunel University, U.K.
{jonathan.darch, b.milner, x.shao}@uea.ac.uk {saeed.vaseghi, qin.yan}@brunel.ac.uk

ABSTRACT

This work proposes a novel method of predicting formant frequencies from a stream of mel-frequency cepstral coefficients (MFCC) feature vectors. Prediction is based on modelling the joint density of MFCCs and formant frequencies using a Gaussian mixture model (GMM). Using this GMM and an input MFCC vector, two maximum a posteriori (MAP) prediction methods are developed. The first method predicts formants from the closest, in some sense, cluster to the input MFCC vector, while the second method takes a weighted contribution of formants predicted from all clusters. Experimental results are presented using the ETSI Aurora connected digit database and show that predicted formant frequencies are within 3.2% of reference formant frequencies.

1. INTRODUCTION

Formants are a useful acoustic parameter in speech processing and methods for their robust estimation have been the subject of much research [1]. Traditional methods have used frequency-domain techniques such as peak-picking to identify formants from the spectral envelope. Parametric techniques such as linear predictive coding (LPC) analysis have also been successful in identifying poles corresponding to formant resonances [2].

This work differs from these techniques in that it aims to predict formant frequencies from features designed for speech recognition purposes, specifically MFCC vectors. The motivation for formant prediction is twofold. First, it is considered that MFCCs are a relatively robust representation of speech and in particular the effect of external influences such as noise or channel distortion can be reduced through a variety of processing techniques [3]. Therefore, if a statistical mapping can be derived which enables formants to be predicted from MFCCs, the resulting estimate may be more robust than traditional methods which are more susceptible to distortion. Secondly, the distributed speech recognition (DSR) framework proposed by the ETSI Aurora standard [4] only transmits MFCC vectors to the remote back-end. If acoustic parameters of the speech signal are required, some kind of prediction from the received MFCC vector stream is necessary.

The procedure for extracting MFCC vectors from speech involves much loss of information which is integral to the structure of the original speech. The ETSI Aurora standard [4] for obtaining MFCCs is shown in figure 1. MFCCs are extracted using 25ms frames at 10ms intervals. For every frame of speech, an MFCC vector comprising coefficients zero to twelve and a log energy term is computed. Phase information is lost in the magnitude operation, while spectral detail is lost during mel-filtering from 128

The work is funded by EPSRC grant GR/S30238/01.

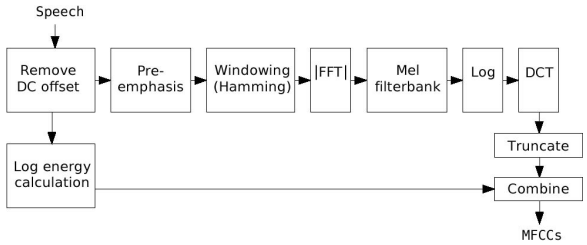

Fig. 1. Outline of ETSI Aurora standard for MFCC extraction [4]

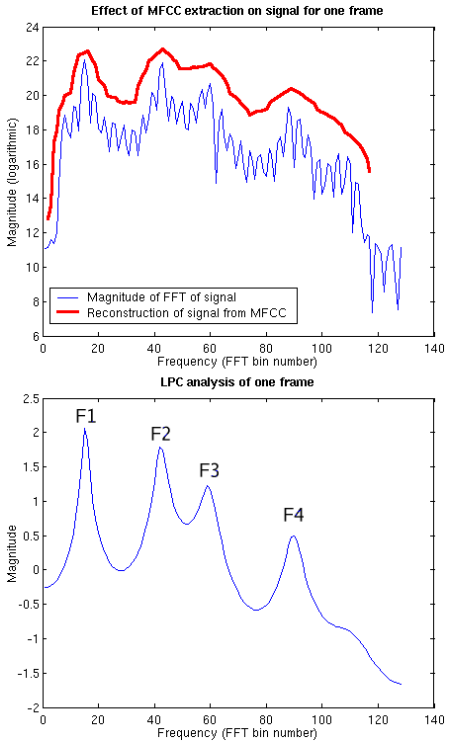

Fig. 2. a) Comparison of original (thin line) and MFCC-derived (bold line) magnitude spectra. b) LPC-derived magnitude spectrum of same frame

to 23 channels and through truncation from 23 to 13 coefficients after the DCT stage. Returning the truncated MFCC vector to a magnitude spectral representation through zero padding, inverse DCT, exponential operation and interpolation results in a spectrally smoothed estimate. This loss of spectral detail causes the accuracy of traditional frequency-based formant estimation methods to degrade. Figure 2a illustrates this by comparing the orig-

inal magnitude spectrum of a frame of speech to the magnitude spectrum obtained by inverting an MFCC vector extracted from the same frame of speech. The LPC-derived magnitude spectrum in figure 2b shows clearly the position of the four formants. In this example, comparing the MFCC-derived magnitude spectrum with the LPC-derived magnitude spectrum shows that the first and fourth formants are well defined, but it is significantly more difficult to distinguish between the second and third formants.

This indicates that it is not possible to obtain accurate formant tracks by inverting the MFCC extraction procedure. However, recent work has shown that pitch can be predicted from MFCCs within a statistical framework by modelling the joint density of MFCCs and pitch using a GMM [5]. The work presented here applies such a statistical technique to predict formant frequencies from MFCCs. There is reason to suggest that formant prediction will be at least as accurate as pitch prediction because MFCCs are a compressed representation of the spectral envelope with a much weaker representation of pitch information.

The proposed formant prediction system is described in section 2. An evaluation of the predicted formant accuracy is made in section 3 and conclusions drawn in section 4.

2. FORMANT FREQUENCY PREDICTION

This section describes the training of a GMM to model the joint density of MFCCs and formant frequencies and subsequent prediction of formant frequencies from the GMM.

2.1. Training

Training begins with the creation of a set of augmented feature vectors, $\mathbf{y}$, which are defined as:

$$\mathbf{y}_i = [\mathbf{x}_i, \mathbf{F}_i]^T \tag{1}$$

where $\mathbf{x}$ is a static MFCC vector, $\mathbf{x} = [x_0, x_1, \ldots x_{12}, ln(e)]$, $\mathbf{F}$ is a vector of the first four formant frequencies, $\mathbf{F} = [F_1, F_2, F_3, F_4]$ and i indicates the frame number. LPC analysis is used to obtain initial estimates of the first four formant frequency tracks, $\mathbf{F}$. These are appended to the fourteen dimensional MFCC vector, $\mathbf{x}$, resulting in an eighteen dimensional vector.

Not all feature vectors represent speech with a clearly defined formant structure, such as those from silence and some unvoiced sounds. At present, a feature vector is classified as containing formant structure according to a voicing decision. This is not an ideal classification but it does constrain the region from which formants are predicted. To classify frames as voiced a simple energy-based threshold is used. The voicing decision is subsequently hand-corrected to eliminate speech with no formant structure. It should be noted that accurate annotation of formant tracks is non-trivial and to a certain extent subjective, due to processes occurring in speech production. The procedure for obtaining the final augmented feature vectors is illustrated in figure 3.

Given a set of augmented feature vectors, a GMM can be used to model the joint density of MFCCs and formant frequencies. Using training data associated with voiced speech, the expectation-maximisation (EM) algorithm is used to perform unsupervised clustering to produce a GMM with K clusters. Each cluster, c_k, models the localised joint probability density function (PDF) of MFCCs and formant frequencies with mean and covariance:

$$\mu_k^y = \begin{bmatrix} \mu_k^x \\ \mu_k^F \end{bmatrix} \quad \text{and} \quad \mathbf{\Sigma}_k^y = \begin{bmatrix} \mathbf{\Sigma}_k^{xx} & \mathbf{\Sigma}_k^{xF} \\ \mathbf{\Sigma}_k^{Fx} & \mathbf{\Sigma}_k^{FF} \end{bmatrix} \tag{2}$$

while the set of K clusters models the joint density across the entire feature vector space.

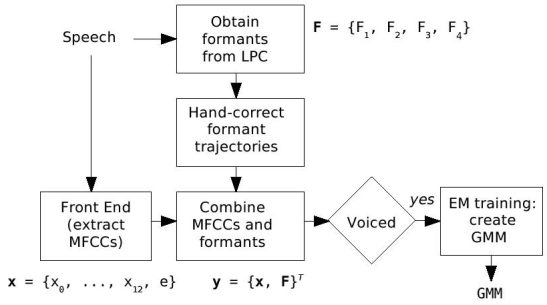

Fig. 3. Training the formant predictor

2.2. Prediction

Using the relationship between MFCCs and formant frequencies modelled by the GMM, prediction of a formant vector, $\hat{\mathbf{F}}$, can be made from an input MFCC vector, $\mathbf{x}$. This prediction can be made from the closest cluster to the MFCC vector, in some sense, or using a weighted contribution from all clusters. For both cases, the procedure is shown in figure 4.

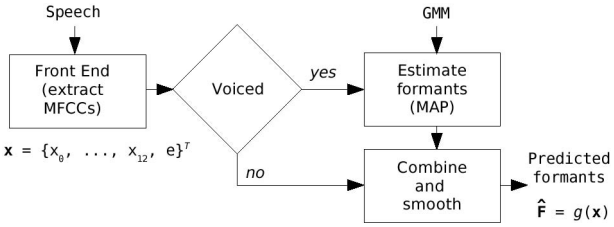

Fig. 4. Predicting formants from the GMM

2.2.1. *Closest cluster*

The closest cluster, $\tilde{k}$, to the input MFCC vector, $\mathbf{x}_i$, can be defined as:

$$\tilde{k} = \arg\max_k \left\{ p\left(\mathbf{x}_i | c_k^x\right) \alpha_k \right\} \tag{3}$$

where $p(\mathbf{x}_i|c_k^x)$ is the marginal distribution of the MFCC vector for the k^{th} cluster and α_k is the prior probability of that cluster [6]. The marginal distribution is the likelihood of the MFCC vector, given that it belongs to the k^{th} cluster. Using the joint density of MFCCs and formant frequencies described by equation 2 for the closest cluster, a maximum a posteriori (MAP) prediction of the formant vector can be made from an input MFCC vector [7]. The predicted formant vector, from the closest cluster, $\tilde{k}$, is given by:

$$\hat{\mathbf{F}}_i = \mu_{\tilde{k}}^F + \mathbf{\Sigma}_{\tilde{k}}^{Fx} \left(\mathbf{\Sigma}_{\tilde{k}}^{xx}\right)^{-1} \left(\mathbf{x}_i - \mu_{\tilde{k}}^x\right) \tag{4}$$

2.2.2. *Weighted combination of clusters*

To avoid making a hard-decision as to the cluster from which prediction is made, an alternative is to take a weighted contribution from all clusters. The formant vector is now given by:

$$\hat{\mathbf{F}}_i = \sum_{k=1}^{K} h_k(\mathbf{x}_i) \left\{ \mu_k^F + \mathbf{\Sigma}_k^{Fx} (\mathbf{\Sigma}_k^{xx})^{-1} (\mathbf{x}_i - \mu_k^x) \right\} \tag{5}$$

The weighting term, $h_k(\mathbf{x}_i)$, scales the formant prediction contribution from each of the K clusters by the posterior probability of the MFCC vector, $\mathbf{x}_i$, belonging to the k^{th} cluster [5]:

$$h_k(\mathbf{x}_i) = \frac{\alpha_k p(\mathbf{x}_i|c_k^x)}{\sum_{k=1}^{K} \alpha_k p(\mathbf{x}_i|c_k^x)} \quad (6)$$

3. EXPERIMENTAL RESULTS

This section analyses the effectiveness of the formant prediction techniques described in section 2. Evaluation measures are defined which consider both frame classification error and percentage frequency error. Using these criteria, formant prediction is measured.

3.1. Evaluation measures

Performance was measured using two criteria. Formant classification error for the j^{th} formant track is defined as:

$$E_j^c = \frac{N_{v|u} + N_{u|v} + N_{>20\%}^j}{N_{total}} \times 100 \quad (7)$$

where $N_{v|u}$ is the number of unvoiced frames classified as voiced, $N_{u|v}$ the number of voiced frames classified as unvoiced and $N_{>20\%}$ the number of frames where the j^{th} formant error was more than 20%. N_{total} is the total number of frames in the test data (88,121). For frames which do not class as classification errors, percentage formant frequency errors are calculated. The overall percentage formant frequency error for the j^{th} formant is given by:

$$E_j^{\%} = \frac{1}{N_j} \sum_{i=1}^{N_j} \left| \frac{\hat{F}_j(i) - F_j(i)}{F_j(i)} \right| \times 100 \quad (8)$$

where N_j is the number of frames not classed as classification errors for the j^{th} formant.

The results in this section compare the formant prediction accuracy of the two MAP techniques described in section 2. Male clean speech utterances were taken from a subset of the ETSI Aurora connected digit corpus, providing 633 utterances (108,140 vectors) for training and 501 (88,121 vectors) for testing. Fifty-five speakers were used for training and a separate fifty-two for testing.

3.2. Formant prediction accuracy

Classification accuracy, E^c, obtained by the closest cluster (equation 4) and weighted combination (equation 5) techniques is shown in figure 5a using from 1 to 32 clusters in the GMM. Similarly, figure 5b compares the percentage predicted formant frequency error, $E^{\%}$, averaged across all four formants. The results show a significant reduction in classification error when increasing the number of clusters from 1 to 2. However, less significant reductions in classification error are observed for further increases in the number of clusters. Figure 5b shows that the increase in clusters gives consistent reductions in predicted formant frequency error from 6.25% with 1 cluster to around 3.2% with 32 clusters. The results also show that using a weighted combination of predicted formants from all clusters gives consistently superior performance to just using the closest cluster.

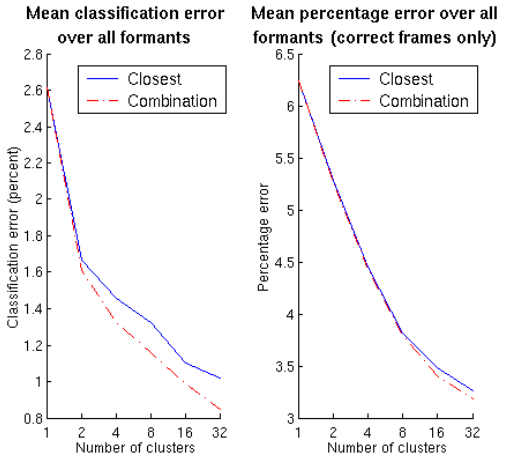

Fig. 5. Comparison of closest cluster and weighted combination of clusters: a) classification error b) formant frequency error

The previous analysis showed classification and prediction errors averaged across all four formants. More detailed analysis is shown in tables 1 and 2 by considering the classification and prediction errors of the four formants individually, using weighted combination prediction.

Clusters	F1	F2	F3	F4	Mean
1	2.92	5.21	1.49	0.83	2.61
2	2.69	1.51	1.42	0.82	1.61
4	1.88	1.18	1.36	0.86	1.32
8	1.25	1.13	1.31	0.93	1.16
16	0.84	0.91	1.23	0.97	0.99
32	0.53	0.76	1.19	0.91	0.85

Table 1. Classification decision error using weighted combination of clusters by formant

Clusters	F1	F2	F3	F4	Mean
1	6.98	7.77	5.33	4.88	6.24
2	6.40	4.98	4.89	4.76	5.26
4	5.37	3.75	4.07	4.53	4.43
8	3.78	3.12	3.84	4.44	3.79
16	3.39	2.60	3.42	4.23	3.41
32	3.12	2.33	3.15	4.15	3.19

Table 2. Percentage formant error using weighted combination of clusters by formant

Again, the results show that increasing the number of clusters leads to reductions in the individual formant classification and prediction errors. In particular, using 32 clusters, formant 2 is most accurately predicted with a percentage formant frequency error of 2.33%. Formants 1 and 3 have prediction errors of around 3.12% and 3.15%, while formant 4 is least accurate, with an error of 4.15%. To illustrate the effectiveness of formant prediction, figure 6 compares reference formant tracks with those predicted from a stream of MFCC vectors. Accurate prediction of the lower formants is shown clearly, while the variability of prediction for higher formants is noticeable, particularly for formant 4. This can be attributed to the non-permanent nature and lower energy of higher formants for both prediction and the creation of reference tracks.

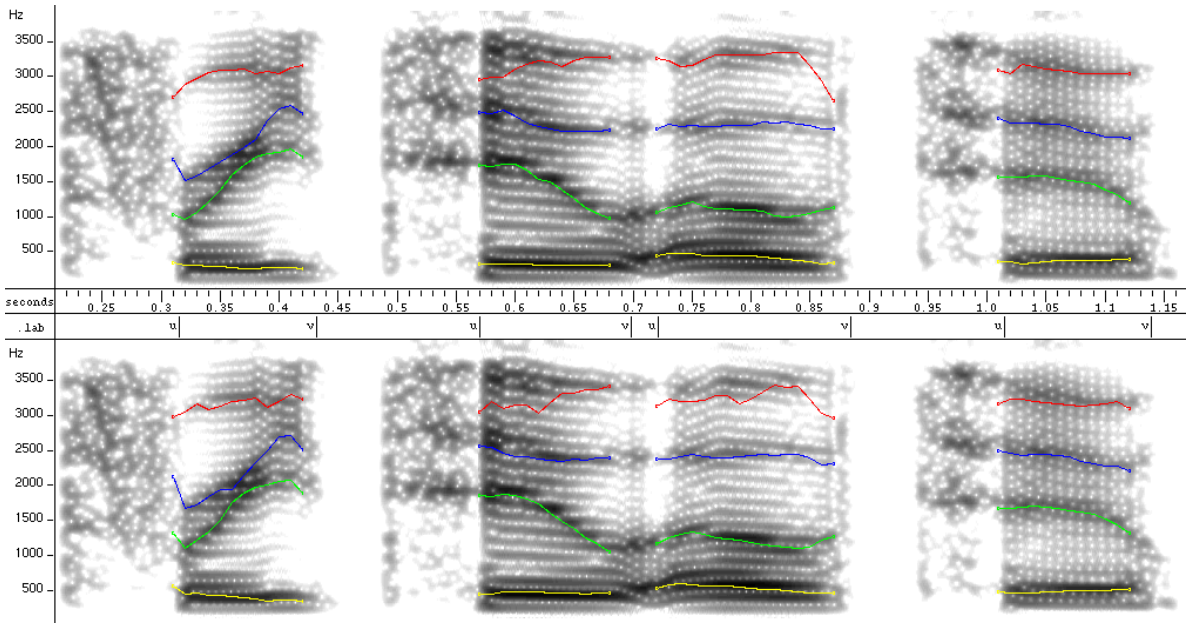

Fig. 6. Spectrograms of "three two oh two" showing formant tracks from a) LPC formant estimation and b) GMM-based formant prediction

3.3. Hand-correction in training

As described in section 2.1, hand-correction of formant tracks for training and test data was carried out in order to train the GMM on voiced speech only. An initial experiment was conducted to determine the importance of hand-correcting the training data. Table 3 shows mean classification and percentage formant frequency errors for weighted combination prediction for both hand-corrected and uncorrected training data. The results show hand-correction of formant tracks for training makes little difference to predicted formant frequency tracks. The results presented in section 3.2 were from GMMs trained on uncorrected data, though testing is always evaluated against hand-corrected formant tracks. Not having to hand-correct training data saves much time and demonstrates robustness to labelling errors in the training process.

mean error	classification, E^c	percentage, $E^{\%}$
hand-corrected	0.914	3.179
uncorrected	0.847	3.187

Table 3. Mean predicted classification and formant frequency errors using hand-corrected and uncorrected training data

4. CONCLUSIONS

This work has used a GMM, trained on a joint feature vector comprising MFCCs and formant frequencies, to enable prediction of formant frequencies from an MFCC vector. The mean predicted formants have been shown to be accurate to within 3.2% of reference formants estimated using LPC analysis and subsequent hand-correction. Two MAP prediction techniques were developed, using either the closest cluster to the input MFCC vector, or a weighted prediction taken from all clusters. Best overall performance was achieved using weighted prediction and a GMM comprising 32 clusters. Further performance increases may be obtained using more clusters in the GMM but the current, relatively small size of the training data set has prohibited this.

The current formant prediction does not utilise the strong temporal correlation which exists in the speech signal. To model this, and hence reduce formant error, further work will investigate the use of state-specific GMMs within the framework of a set of hidden Markov models (HMMs). This has been beneficial for pitch prediction and is likely to also lead to improvements in formant prediction [5].

5. REFERENCES

[1] D. Rentzos, S. Vaseghi, Q. Yan, C.-H. Ho, and E. Turajlic, "Probability models of formant parameters for voice conversion," in *Eurospeech*, Geneva, Switzerland, Sept. 2003, pp. 2405–2408.

[2] L. Rabiner and B.H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993, ISBN: 0-13-015157-2.

[3] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 4, pp. 587–589, Oct. 1994.

[4] A. Sorin and T. Ramabadran, "Extended advanced front end (XAFE) algorithm description, Version 1.1," Tech. Rep. ES 202 212, ETSI STQ-Aurora DSR Working Group, 2003.

[5] X. Shao and B.P. Milner, "Pitch prediction from MFCC vectors for speech recognition," in *ICASSP*, Montreal, Canada, May 2004.

[6] B. R. Ramakrishnan, *Reconstruction of Incomplete Spectrograms for Robust Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, 2000.

[7] C. W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice-Hall, 1992, ISBN: 0-13-217985-7.