AN INVESTIGATION INTO THE CORRELATION AND PREDICTION OF ACOUSTIC SPEECH FEATURES FROM MFCC VECTORS

Jonathan Darch[1]*, Ben Milner*[1]*, Ibrahim Almajai*[1] *and Saeed Vaseghi*[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, i.almajai}@uea.ac.uk saeed.vaseghi@brunel.ac.uk

ABSTRACT

This work develops a statistical framework to predict acoustic features (fundamental frequency, formant frequencies and voicing) from MFCC vectors. An analysis of correlation between acoustic features and MFCCs is made both globally across all speech and within phoneme classes, and also from speaker-independent and speaker-dependent speech. This leads to the development of both a global prediction method, using a Gaussian mixture model (GMM) to model the joint density of acoustic features and MFCCs, and a phoneme-specific prediction method using a combined hidden Markov model (HMM)-GMM. Prediction accuracy measurements show the phoneme-dependent HMM-GMM system to be more accurate which agrees with the correlation analysis. Results also show prediction to be more accurate from speaker-dependent speech which also agrees with the correlation analysis.

Index Terms— Formants, fundamental frequency, voicing, GMM, HMM

1. INTRODUCTION

In distributed speech recognition (DSR) systems only the speech features (typically mel-frequency cepstral coefficients (MFCCs)) are transmitted to the remote recogniser. As no time-domain speech signal is transmitted, techniques such as Fourier and LP analysis cannot be applied to extract acoustic speech features such as voicing, fundamental frequency or formant frequencies. These acoustic speech features are often important to obtain and have application in speech analysis, reconstruction, enhancement and recognition. Therefore, to obtain acoustic speech features directly from MFCC vectors, alternative methods of extraction need to be developed. It is the aim of this work to develop methods that enable acoustic speech features to be predicted from the MFCC representation.

This work builds on previous work that predicted either the voicing and fundamental frequency [1] or formant frequencies [2] from MFCC vectors. These methods operated by modelling the joint density of MFCCs and either fundamental frequency or formant frequencies. Using the joint density and an input MFCC vector, a prediction could then be made of the fundamental or formant frequencies. This work now combines prediction so that for the first time, voicing, fundamental frequency and formant frequencies are predicted jointly. Additionally, the voicing of the speech is also considered explicitly to further improve prediction accuracy.

This work also analyses the correlation between acoustic features and MFCCs to gain understanding in order to increase prediction accuracy. A comparison of measuring correlation globally and then within phoneme classes is made. Investigation is also made into the effect on correlation of moving from speaker-independent speech to speaker-dependent speech. This correlation analysis is carried out in section 2. Section 3 describes the proposed method of joint prediction of acoustic features from MFCCs. First the global correlation is exploited within a Gaussian mixture model (GMM) framework and then secondly the phoneme-specific correlation is used within a combined hidden Markov model (HMM)-GMM framework. Predictions are also made as to the voicing. Experimental results are presented in section 4 which compare the accuracy of global prediction of acoustic features from phoneme-specific prediction. The effect on prediction accuracy of using speaker-dependent and speaker-independent speech is then examined. The results of prediction are compared to the correlation analysis.

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

2. CORRELATION ANALYSIS

This section investigates the correlation that exists between acoustic speech features and MFCCs. The aim of this analysis is to confirm the existence of correlation and then to see how correlation can be increased which should lead to better prediction of acoustic features from MFCC vectors. The analysis first compares the correlation when measured globally across all speech sounds to that measured within individual phoneme classes. Second, correlation is compared when measured on speaker-independent speech and speaker-dependent speech.

The correlation between each acoustic feature and the MFCC vector is measured using multiple linear regression [3]. A linear model is computed to describe the relation between the MFCC vector (independent variable) and each of the acoustic features (dependent variable). This allows each acoustic feature, $\mathrm{F}(j)$, to be represented in terms of the MFCC vector, $\mathbf{x}$, using a set of $M+1$ regression coefficients, $[b_{j,0}, \ldots, b_{j,m}, \ldots, b_{j,M}]$ which are specific to the j^{th} acoustic feature:

$$\mathrm{F}(j) \;=\; b_{j,0} + b_{j,1}\mathbf{x}(1) + b_{j,2}\mathbf{x}(2) + \ldots + b_{j,M}\mathbf{x}(M) + \varepsilon \qquad 1 \le j \le M \tag{1}$$

where ε is an error term. Using a set of training data, least squares estimation can determine the regression coefficients. These allow a prediction, $\hat{\mathrm{F}}(j)$, of the j^{th} acoustic feature to be made from the MFCC vector, $\mathbf{x}$. The correlation between the j^{th} acoustic feature and the MFCC vector, $\mathbf{x}$, can finally be determined from the R-squared term which is defined as:

$$R(j)^2 \;=\; 1 - \frac{\sum_i \left(\mathrm{F}_i(j) - \hat{\mathrm{F}}_i(j)\right)^2}{\sum_i \left(\mathrm{F}_i(j) - \overline{\mathrm{F}}_i(j)\right)} \tag{2}$$

where $\overline{\mathrm{F}}(j)$ is the mean of the j^{th} acoustic feature.

MFCC vectors are computed as specified in the ETSI Aurora DSR front-end [4], which leads to a stream of 14-D MFCC vectors at a rate of 100 vectors per second. Acoustic features are extracted at the same rate as the MFCC vectors and using the same 25ms frames of speech. The acoustic feature vector, $\mathbf{F}$, comprises fundamental frequency, F0, and the first four formant frequencies, F1 to F4. The databases used in these tests are described in section 4.

2.1. Global and Phoneme-specific Correlation

The global correlation between acoustic features and the MFCC vector is measured by pooling features from all speech sounds in the set of training utterances. Using multiple linear regression, the correlation of each acoustic feature to the MFCC vector can then be measured. The phoneme-specific correlation is measured by first segmenting the training data into phoneme classes using hand annotations of the data. Within each phoneme class, multiple linear regression is used to measure the correlation between each acoustic feature and the MFCC vector. To obtain a single correlation measure, the correlation of each acoustic feature is averaged across the set of phonemes.

Global and phoneme-specific correlations of fundamental frequency, F0, and formant frequencies, F1 to F4, are shown in table 1, measured from male speaker-independent speech taken from the WSJCAM0 database - see section 4 for details. The correlation measures are broken down into those from unvoiced and voiced speech.

	voicing	F0	F1	F2	F3	F4
Global	u	–	0.740	0.708	0.733	0.796
	v	0.101	0.721	0.626	0.693	0.766
Phoneme - specific	u	–	0.781	0.799	0.754	0.653
	v	0.390	0.805	0.886	0.795	0.704

Table 1. Correlations between acoustic features and MFCCs calculated globally and by phoneme for unvoiced and voiced speech

The results show that when measuring correlation within individual phoneme classes, higher correlation is observed than when considering the global correlation across all speech sounds. This observation is to be expected as restricting the multiple regression to model correlation from a small cluster of related sounds is more likely to produce good modelling than when generalised across all speech sounds. This suggests that predicting acoustic features from phoneme-specific models should be more accurate than prediction from global models.

Of the acoustic features being measured, the formants show considerably higher correlation to the MFCC vector than fundamental frequency. This is attributed to the shape and spacing of the mel-filterbank used in MFCC extraction. The filterbank allows a reasonable spectral envelope to be reproduced which shows formant positions but lacks much of the finer spectral structure which conveys fundamental frequency information. A significant increase in fundamental frequency to MFCC correlation is observed when comparing the global to phoneme-specific measurement. This indicates that fundamental frequency is influenced by the phoneme which has also been reported in [5].

2.2. Speaker Dependent and Independent Correlation

To examine the effect that different, and multiple, speakers have on the correlation between acoustic features and MFCC vectors, three speech databases have been used. Two databases are speaker-dependent (one from a male speaker and the other a female speaker) and comprise a set of 246 phonetically rich sentences each. The third database is taken from 10 different male speakers from the WSJ-CAM0 database and comprises 765 utterances. Further details of the speech databases are given in section 4. Using a similar procedure to that described in section 2.1 for phoneme-specific correlation measurement, the fundamental frequency and formant frequency correlation to MFCC vectors of the three databases have been computed. These are shown in table 2, broken down into unvoiced and voiced speech.

	voicing	F0	F1	F2	F3	F4
SI male	u	–	0.781	0.799	0.754	0.653
	v	0.390	0.805	0.886	0.795	0.704
SD male	u	–	0.783	0.826	0.811	0.744
	v	0.714	0.835	0.885	0.837	0.745
SD female	u	–	0.799	0.763	0.790	0.747
	v	0.830	0.827	0.732	0.791	0.812

Table 2. Phoneme-specific correlations between acoustic features and MFCCs for the speaker-independent (SI) and two speaker-dependent (SD) databases

Comparing first the correlation results for the speaker-independent male and speaker-dependent male shows a slight increase in correlation for formant frequencies and a substantial increase in correlation for fundamental frequency. The large increase in fundamental frequency correlation, when moving to speaker-dependent analysis, arises from the reduction in fundamental frequency variation. The more modest increase in formant frequency correlation, when moving to speaker dependent analysis, is explained by the fact that the MFCC vector itself contains a substantial amount of formant information. Moving to speaker-dependent measurement does not provide much extra information, hence the limited increase in correlation.

Comparing correlation results for speaker-dependent male and speaker-dependent female speech shows, in general, higher formant frequency to MFCC correlation for the male speaker. This result is consistent with traditional signal processing methods of formant estimation that perform less well on female speech due to the wider spacing of pitch harmonics which makes the precise localisation of formant frequencies difficult. For fundamental frequency, the reverse is true, with substantially higher correlation observed with the female speaker. This may be due to the higher frequencies associated with female speech spanning a wider range of mel-filterbank channels than with male speech, making their identification more accurate.

Generally, voiced speech produces higher formant frequency to MFCC correlation, compared to unvoiced speech. This is also consistent with formant estimation methods which are more accurate for voiced speech. This is due to better defined spectral structure produced by the harmonic structure of voiced speech in comparison to the noise-like structure from unvoiced speech.

3. PREDICTION OF ACOUSTIC FEATURES

Prediction of acoustic features from MFCCs can either exploit the global correlation between acoustic features and MFCCs or phoneme-dependent correlations. A brief description of the GMM and HMM-GMM methodologies is described here, for more details see [2].

3.1. GMM-based Prediction

Predicting acoustic features from MFCCs comprises two parts. First, three GMMs are created to model the joint density of acoustic fea-

tures and MFCCs for non-speech, unvoiced, and voiced speech. Second, for predicting acoustic features from a stream of previously unseen MFCC vectors, for each vector, a voicing decision is made using prior and posterior voicing probabilities. Depending on the predicted voicing class (voiced, unvoiced or non-speech), acoustic features may be predicted from the appropriate GMM using a maximum a posteriori (MAP) prediction.

Training begins with the creation of augmented feature vectors by concatenating MFCC, $\mathbf{x}$, and acoustic, $\mathbf{F}$, vectors:

$$\mathbf{y}_i \;=\; [\mathbf{x}_i, \mathbf{F}_i]^T \tag{3}$$

where vector $\mathbf{x}_i = [x(0), x(1), \ldots x(12), ln(e)]$ comprises static MFCCs 0 to 12 and log energy for the i^{th} frame of speech. The acoustic feature vector $\mathbf{F}_i = [\text{F0, F1, F2, F3, F4}]$ comprises the fundamental frequency and frequencies of the first four formants of the i^{th} frame of speech.

From pools of non-speech, unvoiced and voiced augmented feature vectors, three Gaussian mixture models (GMMs) are created to model the joint density of non-speech, unvoiced and voiced MFCC vectors and acoustic features across all phonemes, denoted Φ^{ns}, Φ^{u} and Φ^{v}, respectively.

For prediction of acoustic features, a voicing decision must first be made in order to determine which GMM is to be used. For MFCC vectors predicted as voiced, fundamental and formant frequencies are predicted. For vectors predicted as unvoiced, only formant frequencies are predicted. No acoustic features are predicted for MFCC vectors predicted as non-speech.

The probabilities of a given MFCC vector, $\mathbf{x}_i$, coming from non-speech, unvoiced and voiced speech are calculated. For voiced speech, this is given by:

$$P\left(v|\mathbf{x}_i\right) \;=\; \frac{P(v)\; p\left(\mathbf{x}_i|v\right)}{p\left(\mathbf{x}_i\right)} \tag{4}$$

where $P(v)$ is the prior probability of the voiced class, $p(\mathbf{x}_i)$ is the prior probability of vector $\mathbf{x}_i$, and $p(\mathbf{x}_i|v)$ is given by the corresponding marginalised GMM $\Phi^{v,x}$:

$$p(\mathbf{x}_i|v) = \Phi^{v,x}(\mathbf{x}_i) = \sum_{k=1}^{K} \alpha_k^v \phi_k^{v,x}(\mathbf{x}_i) = \sum_{k=1}^{K} \alpha_k^v p(\mathbf{x}_i|\phi_k^{v,x}) \tag{5}$$

where $p\left(\mathbf{x}_i|\phi_k^{v,x}\right)$ is the marginal distribution of the MFCC vector for the k^{th} cluster of the voiced GMM, $\phi_k^{v,x}$. Similarly, $P(u|\mathbf{x}_i)$ and $P(ns|\mathbf{x}_i)$ are computed.

The predicted voicing class of a MFCC vector is given as voiced, unvoiced or non-speech according to the highest probability. If a MFCC vector is deemed to be voiced, then the voiced GMM, Φ^{v}, is used to predict fundamental and formant frequencies. For unvoiced MFCCs, formant frequencies are predicted from the unvoiced GMM, Φ^{u}, whilst for non-speech MFCCs, no acoustic features are predicted.

For prediction of acoustic features from voiced MFCC vectors, the maximum a posteriori (MAP) estimation of the i^{th} vector of acoustic frequencies, $\hat{\mathbf{F}}_i$, from $\mathbf{x}_i$ is given by:

$$\hat{\mathbf{F}}_i \;=\; \arg\max_{\mathbf{F}_i} \left\{p\left(\mathbf{F}_i|\mathbf{x}_i, \Phi_k^v\right)\right\} \tag{6}$$

Acoustic feature vector predictions from each cluster are weighted by the posterior probability, $h_k(\mathbf{x}_i)$, of the i^{th} MFCC vector $\mathbf{x}_i$, belonging to the k^{th} cluster:

$$\hat{\mathbf{F}}_i \;=\; \sum_{k=1}^{K} h_k(\mathbf{x}_i) \left\{ \boldsymbol{\mu}_k^{v,F} + \boldsymbol{\Sigma}_k^{v,Fx} (\boldsymbol{\Sigma}_k^{v,xx})^{-1} \left(\mathbf{x}_i - \boldsymbol{\mu}_k^{v,x}\right) \right\} \tag{7}$$

The posterior probability, $h_k(\mathbf{x}_i)$, is given by:

$$h_k(\mathbf{x}_i) \;=\; \frac{\alpha_k^v\; p\left(\mathbf{x}_i|\phi_k^{v,x}\right)}{\sum_{k=1}^{K} \alpha_k^v\; p\left(\mathbf{x}_i|\phi_k^{v,x}\right)} \tag{8}$$

A five point median filter is used to smooth each acoustic feature vector track by removing discontinuities. Segments of speech and non-speech are also forced to have a minimum duration of 30ms.

3.2. HMM-GMM-based Prediction

For HMM-GMM prediction, three GMMs which model non-speech, unvoiced and voiced speech are created for every state of each model of a set of monophone HMMs. During training, a set of W three-state monophone HMMs are created, one HMM for each monophone, through Baum-Welch re-estimation using MFCC vectors, $\mathbf{x}$, and their velocity and acceleration derivatives. The non-speech, unvoiced and voiced GMMs associated with each state of every HMM are created by realigning training data vectors to the HMMs using Viterbi decoding. For each training utterance, model and state allocations are found through forced alignment with annotations [2].

4. EXPERIMENTAL RESULTS

The experiments first consider the accuracy of acoustic feature prediction from MFCCs from the GMM and HMM-GMM methods to determine whether restricting the sound class that prediction is made from improves accuracy, as suggested by the correlation analysis. Second, the effect on accuracy of using speaker-dependent or speaker-independent speech is investigated. The results are then compared to the correlation analysis made in section 2.

Three speech databases have been used in these experiments. Two are speaker-dependent databases (one US male speech and the other US female speech) which comprise phonetically rich sentences. The speaker-dependent male database comprises 601 sentences for training and 246 for testing. For the female speaker-dependent database, 650 sentences are used for training and 246 for testing. The third database is speaker-independent and comprises 1845 sentences from the male part of the UK English WSJCAM0 database. 1080 sentences, spoken by 54 speakers are used for training and 765 sentences spoken by 10 different speakers are used for testing. All databases are sampled at 8kHz.

Reference formant frequencies have been extracted using LPC analysis followed by Kalman filtering [6]. For the two speaker-dependent databases, reference fundamental frequency tracks were obtained from laryngograph data. The ETSI Aurora DSR front-end [4] was used to give fundamental frequency estimates for the speaker-independent WSJCAM0 database.

To measure the accuracy of acoustic feature prediction a root mean square (RMS)-based error measure has been used. The five acoustic features being measured (F0 to F4) exist at very different regions of the frequency spectrum. For example, the mean of F1 for the speaker-dependent male database is 396Hz compared with 3218Hz for F4. This means that an absolute prediction error, measured in Hertz, has considerably more impact for low frequency formants and fundamental frequency than for higher frequency formants. To normalise prediction errors, the RMS error, measured for each acoustic feature, has been scaled by the standard deviation of the reference frequencies of that acoustic feature. Therefore, the prediction error,

$E(j)$, for the j^{th} acoustic features is computed as:

$$E(j) = \frac{1}{\sigma_j} E_{RMS} = \frac{1}{\sigma_j} \sqrt{\frac{1}{N} \sum_{i=1}^{N} \left(F_i(j) - \hat{F}_i(j) \right)^2} \quad (9)$$

where $\mathrm{F}_i(j)$ and $\hat{\mathrm{F}}_i(j)$ are the frequencies of the j^{th} reference and predicted acoustic feature for frame i and σ_j is the standard deviation of the j^{th} acoustic feature.

4.1. Global and Phoneme-dependent Prediction

The aim of this experiment is to compare the accuracy of acoustic feature prediction from the GMM and HMM-GMM systems. This is in effect using either global or phoneme-specific modelling of the joint acoustic feature and MFCC distribution to make the predictions.

Table 3 shows the normalised RMS prediction error for the acoustic features, computed using the test data of the speaker-independent male speech database. It was shown in table 1 that higher correlations between acoustic features and MFCCs were obtained when measuring correlation by phoneme, rather than globally. Table 3 agrees with the correlation results, as lower normalised RMS errors are obtained for the HMM-GMM prediction method, rather than the GMM method which does not exploit the localised correlations. Except for F4, all the correlation results increase when analysis is localised to the phoneme level and the normalised RMS prediction errors decrease.

Closer analysis reveals that the ranks of the phoneme-specific correlations and HMM-GMM predictions generally match. For example, F2, which has the highest correlation, is also the most easily predicted. Comparing the ranks of the global correlations and GMM predictions do not match, that is the acoustic feature with the highest global correlation, F4, is not the most accurately predicted, but actually has the largest normalised RMS error.

	voicing	F0	F1	F2	F3	F4
GMM	u	–	0.693	0.657	0.717	0.772
	v	1.009	0.670	0.481	0.633	0.717
HMM-GMM	u	–	0.644	0.616	0.687	0.789
	v	0.950	0.617	0.451	0.610	0.739

Table 3. Normalised RMS errors for prediction of acoustic features from MFCCs exploiting global (GMM) and phoneme-dependent (HMM-GMM) correlations

4.2. Speaker-dependent and Speaker-independent Prediction

Table 4 presents errors for HMM-GMM prediction using the speaker-independent and two speaker-dependent databases. Results are given for two types of HMM Viterbi decoding: forced and unconstrained. With forced recognition the model and state sequences are found by forced alignment with the correct model sequence and so are more accurate than unconstrained recognition where the model sequence is unknown. The forced-aligned results provide an upper bound of recognition accuracy. The results produced with unconstrained recognition are more realistic as they have no prior information about the phoneme sequence. Instead, the HMM network decodes the input vectors into the most likely sequence of phonemes for prediction. The word recognition accuracies are 70.7% for the speaker-dependent female database, 69.6% for the speaker-dependent male database and 56.2% for the speaker-independent male database.

	voicing	F0	F1	F2	F3	F4
SI male (unconstr.)	u	–	0.644	0.616	0.687	0.789
	v	0.950	0.617	0.451	0.610	0.739
SI male (forced)	u	–	0.646	0.630	0.685	0.806
	v	0.911	0.612	0.455	0.612	0.742
SD male (unconstr.)	u	–	0.668	0.581	0.604	0.693
	v	0.637	0.563	0.469	0.551	0.672
SD male (forced)	u	–	0.675	0.583	0.602	0.690
	v	0.649	0.574	0.475	0.552	0.665
SD female (unconstr.)	u	–	0.603	0.667	0.640	0.698
	v	0.455	0.577	0.712	0.637	0.597
SD female (forced)	u	–	0.607	0.676	0.645	0.698
	v	0.465	0.578	0.712	0.641	0.596

Table 4. Phoneme-specific normalised RMS errors for prediction of acoustic features from MFCCs with unconstrained and forced HMM recognition

The results in table 4 follow what would be expected from the correlation analysis shown in table 2, in that higher correlation generally leads to lower normalised RMS errors. There is little difference between prediction using forced and unconstrained recognition, with forced prediction actually resulting in greater errors more often than for unconstrained recognition. This demonstrates that the correct phoneme sequence is not vital, as where the recognition decoding makes errors it generally outputs similar sounding phonemes whose acoustic features are close to the correct phoneme.

5. CONCLUSIONS

The analysis of the correlation between acoustic features and MFCCs in section 2 suggested that prediction of acoustic features from MFCCs would be made more accurate by exploiting phoneme-dependent correlations. This was achieved through a HMM-GMM framework which produced greater prediction accuracy compared with a global prediction method.

Analysis of correlation results in relation to prediction accuracy shows that, in general, as correlation increases, prediction error decreases. This indicates that searching for methods to improve correlation would decrease prediction errors.

6. REFERENCES

[1] X. Shao and B. Milner, "Predicting fundamental frequency from mel-frequency cepstral coefficients to enable speech reconstruction," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 1134–1143, Aug. 2005.

[2] J. Darch and B. Milner, "HMM-based MAP prediction of voiced and unvoiced formant frequencies from noisy MFCC vectors," in *Interspeech*, Pittsburgh, PA, Sept. 2006, pp. 1005–1008.

[3] S. Chatterjee and A.S. Hadi, *Regression Analysis By Example*, John Wiley and Sons, 4th edition, 2006, ISBN: 0-471-74696-7.

[4] A. Sorin and T. Ramabadran, "Extended Advanced Front End Algorithm Description, Version 1.1," Tech. Rep. ES 202 212, ETSI STQ-Aurora DSR Working Group, Apr. 2003.

[5] T. Hirahara, "On the role of fundamental frequency in vowel perception," The 2nd joint meeting of ASA and ASJ, Nov. 1988.

[6] Q. Yan, E. Zavarehei, S. Vaseghi, and D. Rentzos, "A formant tracking LP model for speech processing in car/train noise," in *ICSLP*, Jeju, Korea, Oct. 2004, pp. 2409–2412.