PITCH PREDICTION FROM MFCC VECTORS FOR SPEECH RECONSTRUCTION

Xu Shao and Ben Milner

School of Computing Sciences, University of East Anglia, UK
{x.shao, b.milner}@uea.ac.uk

ABSTRACT

This work proposes a technique for reconstructing an acoustic speech signal solely from a stream of mel-frequency cepstral coefficients (MFCCs). Previous speech reconstruction methods have required an additional pitch element, but this work proposes two maximum a posteriori (MAP) methods for predicting pitch from the MFCC vectors themselves. The first method is based on a Gaussian mixture model (GMM) while the second scheme utilises the temporal correlation available from a hidden Markov model (HMM) framework. A formal measurement of both frame classification accuracy and RMS pitch error shows that an HMM-based scheme with 5 clusters per state is able to correctly classify over 94% of frames and has an RMS pitch error of 3.1Hz in comparison to a reference pitch. Informal listening tests and analysis of spectrograms reveals that speech reconstructed solely from the MFCC vectors is almost indistinguishable from that using the reference pitch.

1. INTRODUCTION

In recent years the performance of speech recognition systems from mobile devices has been improved through the use of distributed speech recognition (DSR) [1]. Such systems replace low bit-rate speech codecs with the front-end processing component of the speech recogniser and transmit feature vectors (such as MFCCs) directly to the speech recogniser. The removal of the speech codec gives increased recognition accuracy, particular in the presence of acoustic noise or channel errors. However, because feature vectors are designed to be a compact representation, optimized for discriminating between different speech sounds, they have been considered as containing insufficient information to enable reconstruction of the original speech signal. In particular valuable speaker information, such as pitch, is lost. It is therefore not possible to simply invert the stages involved in MFCC extraction to re-create the acoustic speech signal.

However, several schemes [2,3] have been proposed recently which enable speech to be reconstructed from MFCC vectors through the inclusion of pitch information. This pitch value is included in the information extracted from the speech signal on the terminal device and is transmitted as an additional element of the feature vector.

The aim of this work is to predict the pitch frequency from the MFCC vector which will therefore enable speech reconstruction to be achieved solely from the MFCC stream. This is motivated by several studies which have indicated that class-dependent correlation exists between the spectral envelope and pitch [4,5,6,7]. This correlation has been exploited to provide improved phoneme recognition accuracy through class-based normalisation of the spectral envelope by the pitch [4,5]. The correlation has also been utilised to increase the perceptual quality of concatenative text-to-speech synthesis by adjusting the magnitude spectrum of speech units in accordance to pitch modifications [6]. Prediction of the pitch from modified spectral envelopes has also made use of this correlation for voice conversion applications [7].

A brief description of speech reconstruction from MFCC vectors and pitch using the sinusoidal model is presented in section 2. Section 3 introduces two methods for predicting pitch from MFCC vectors. The first is based on a Gaussian mixture model (GMM) while the second utilises the temporal correlation of pitch through a hidden Markov model. Measurements of the predicted pitch accuracy are described in section 4 together with an examination of the resultant reconstructed speech signal. A conclusion is made in section 5.

2. SPEECH RECONSTRUCTION

The sinusoidal model [8] synthesises a speech signal, $x(n)$, by summing together L sinusoids of varying amplitude, A_l, frequency, ω_l, and phase, θ_l,

$$x(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \theta_l) \quad (1)$$

An estimate of the spectral envelope can be calculated from an MFCC vector by zero padding and applying an inverse discrete cosine transform (IDCT). An exponential operation applied to the resulting log mel-filterbank estimate, followed by interpolation, gives a smoothed magnitude spectral estimate, $\left|\hat{X}(\omega)\right|$ [2]. Normalisation must then be applied to remove the effect of pre-emphasis and the non-linear filterbank channel bandwidths [3]. The frequency of the sinusoidal components, ω_l, can be estimated from the pitch frequency, ω_0, by assuming a harmonic relationship, i.e. $\omega_l = l\,\omega_0$. The amplitude of the sinusoidal components, A_l, can be computed from the smoothed magnitude spectral estimate,

$$A_l = \left|\hat{X}(l\,\omega_0)\right| \quad (2)$$

The phase offset, θ_l, is calculated from two components; one relating to the speech excitation and the other to the vocal tract [8]. Therefore, given an MFCC vector and pitch estimate, a frame of reconstructed speech can be generated. It is clear from this analysis that accurate pitch estimation is vital for synthesizing realistic sounding speech.

3. PITCH PREDICTION

Two methods are proposed for predicting the pitch frequency from a stream of MFCC vectors. A method is also introduced for classifying MFCC vectors as representing either voiced or unvoiced speech. These scheme are based on modeling the joint density of the MFCC vector, **x**, and pitch frequency, *f*. From a set of training data, a series of augmented feature vectors, **y**, are extracted,

$$\mathbf{y} = [\mathbf{x}, f]^T \tag{3}$$

Currently in this work the MFCC vector comprises static coefficients 0 to 12. The pitch frequency is estimated with a comb function [9] and is subsequently manually corrected. For unvoiced frames the pitch frequency is set to zero. Specific training set and feature extraction details are given in section 4.

3.1 GMM-based prediction

This system models the joint distribution of the MFCC vector and pitch using a GMM. From the training set of augmented vectors, unsupervised clustering is implemented using the expectation-maximisation (EM) algorithm to produce a set of *K* clusters. Each of these clusters is represented by a Gaussian probability density function (PDF) with mean and covariance,

$$\mu_k^y = \begin{bmatrix} \mu_k^x \\ \mu_k^f \end{bmatrix} \quad \text{and} \quad \Sigma_k^y = \begin{bmatrix} \Sigma_k^{xx} & \Sigma_k^{xf} \\ \Sigma_k^{fx} & \Sigma_k^{ff} \end{bmatrix} \tag{4}$$

This set of *K* clusters models the joint density of pitch and MFCC. Using these cluster-based correlations a prediction of the pitch frequency, $\hat{f}_i$, can be made from an input MFCC vector, $\mathbf{x}_i$. The prediction can be made from the closest cluster, in some sense, to the input MFCC vector or by taking a weighted contribution from all clusters.

The closest cluster, *k*, to the input MFCC vector, $\mathbf{x}_i$, is given,

$$k = \arg\max_k \left\{ p\left(x_i \middle| c_k^x\right) \alpha_k \right\} \tag{5}$$

where $p\left(x_i \middle| c_k^x\right)$ is the marginal distribution of the MFCC vector for the k^{th} cluster and α_k is the prior probability of that cluster. Using the joint density of pitch and MFCC a MAP [10] prediction of the pitch can be made,

$$\hat{f}_i = \mu_k^f + \Sigma_k^{fx}\left(\Sigma_k^{xx}\right)^{-1}\left(x_i - \mu_k^x\right)^T \tag{6}$$

To avoid the problem of identifying the most appropriate cluster, an alternative method combines the MAP pitch prediction from all *K* clusters in the GMM,

$$\hat{f}_i = \sum_{k=1}^{K} h_k(x_i)\left(\mu_k^f + \Sigma_k^{fx}\left(\Sigma_k^{xx}\right)^{-1}\left(x_i - \mu_k^x\right)^T\right) \tag{7}$$

The term $h_k(x_i)$ weights the contribution from each cluster in the GMM by the posterior probability of $\mathbf{x}_i$ belonging to it,

$$h_k(x_i) = \frac{\alpha_k\, p\left(x_i \middle| c_k^x\right)}{\sum_{k=1}^{K} \alpha_k\, p\left(x_i \middle| c_k^x\right)} \tag{8}$$

where α_k and $p\left(x_i \middle| c_k^x\right)$ are as defined for equation (5).

3.2 HMM-based prediction

The unsupervised training used in the GMM does not fully exploit the class-based correlation between the MFCC vector and pitch or the temporal correlation of the pitch contour. To better model the inherent correlation of the feature vector stream, and to select a more appropriate cluster from which to predict the pitch, an HMM-based extension to the GMM scheme is made. This is illustrated in figure 1a where the joint MFCC and pitch feature space, as occupied by the clusters of the GMM, is shown. Figure 1b shows the same feature space but modeled by a series of HMMs, λ_w, each comprising a number of states from which pitch is predicted.

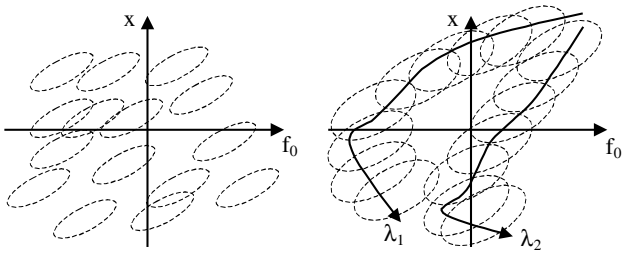

Figure 1: a) GMM clustering, b) HMM-based states

The first stage of training involves the creation of a set of HMM-based speech models, Λ. At present this work uses the ETSI Aurora database which means that the set of models comprises 11 digit models. These models are trained using the MFCC component, **x**, of the augmented vector, **y**, using the standard Baum-Welch algorithm and a diagonal covariance matrix. Once a set of models has been created the training data is aligned to the speech models using Viterbi decoding and vectors belonging to each state, *s*, of each model, *w*, are pooled together. Vectors corresponding to unvoiced speech (as indicated by the pitch component) are removed to ensure that the joint distribution of MFCC and pitch is not distorted by unvoiced vectors. Clustering is applied to the pooled vectors within each voiced state (according to section 3.3) using the EM algorithm. This results in a set of means, $\mu_{k,s,w}^y$, and covariances, $\Sigma_{k,s,w}^y$, corresponding to the k^{th} cluster of the s^{th} state of speech model *w*.

Prediction of the pitch, for voiced frames, is accomplished from the MFCC vectors by first determining the model and state sequence from the set of models using Viterbi decoding. For each MFCC vector, $\mathbf{x}_i$, the allocated model, m_i, and state, q_i, are used to determine the MAP prediction of the pitch,

$$\hat{f}_i = \sum_{k=1}^{K} h_{k,q_i,m_i}(x_i)\left(\mu^{f}_{k,q_i,m_i} + \Sigma^{fx}_{k,q_i,m_i}\left(\Sigma^{xx}_{k,q_i,m_i}\right)^{-1}\left(x_i - \mu^{x}_{k,q_i,m_i}\right)^{T}\right) \quad (9)$$

where $h_{k,qi,mi}(\boldsymbol{x}_i)$ is computed using equation (8) with $P(x_i|c_k)$ specific to state q_i of model m_i.

3.3 Voiced/unvoiced classification

Classification of the MFCC stream into voiced or unvoiced speech is achieved through analysis of the resulting model and state sequences after Viterbi decoding. To determine the voicing associated with each state, the proportion of training data frames aligned to each state, s, of each model, w, which are voiced was calculated, $o_{s,w}$. For illustration, figure 2 shows the proportion of frames, allocated to the 16 states of the digits "six" and "three" which are voiced.

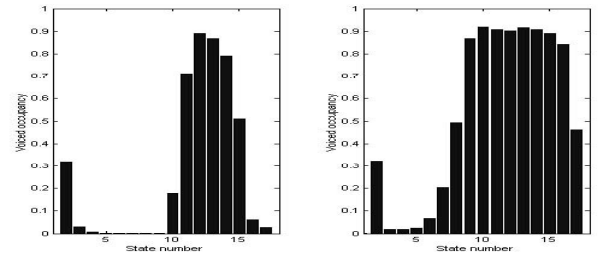

Figure 2: Voiced state occupancy for digits a) six, b) three

The first and last few states of the model "six" contain relatively few voiced frames which corresponds to the unvoiced phonemes /s/ and /k/ /s/. The central part of the model comprises nearly all voiced frames and corresponds to the vowel /ih/. The state occupancy for the model "three" shows similar behavior. Initial states have relatively few voiced frames, corresponding to the /th/ phoneme, while the remaining states are dominated by voiced frames from the phonemes /r/ and /iy/.

To classify an input MFCC vector, $\mathbf{x}_i$, as either voiced or unvoiced Viterbi decoding is used to determine the model, m_i, and state, q_i, allocation. Using the state occupancy, $o_{s,w}$, measured from the training data, the voicing is determined,

$$voicing_i = \begin{cases} voiced & o_{q_i,m_i} > \alpha \\ unvoiced & o_{q_i,m_i} \le \alpha \end{cases} \quad (10)$$

The threshold, α, has been determined experimentally with a reasonable value being $\alpha=0.2$. At this value more errors are likely to come from unvoiced frames being classified as voiced. As energy tends to be low for these frames any errors make little perceptible sound. Conversely, if more errors were made when classifying voiced frames as unvoiced, their higher energy would cause more noticeable noise-like errors.

4. EXPERIMENTAL RESULTS

The experimental results in this section measure both the accuracy of pitch prediction and the resultant reconstructed speech quality.

4.1 Pitch prediction accuracy

This section determines the accuracy of pitch prediction using a subset of the ETSI Aurora database comprising multiple speakers with 200 utterances used for training and 90 for testing. From this data 13-D MFCC vectors have been extracted together with a pitch estimate at a rate of 100 vectors per second in accordance with the ETSI Aurora standard [1]. The pitch estimate has been made using a comb function approach [9] and has subsequently been manually corrected to form the reference pitch measurement.

The pitch prediction systems from section 3 are evaluated on both their classification of MFCC vectors as voiced or unvoiced and also on the RMS pitch error for voiced frames. Pitch classification error is measured as,

$$E_C = \frac{N_{V/U} + N_{U/V} + N_{>20\%}}{N_{Total}} \times 100\% \quad (11)$$

where $N_{V|U}$ is the number of unvoiced frames classified as voiced, $N_{U|V}$ is the number of voiced frames classified as unvoiced and $N_{>20\%}$ is the number of frames in which the pitch error is greater than 20%. N_{Total} is the total number of frames which was 15651 in these tests. For frames correctly classified as voiced, the RMS pitch error is computed as,

$$E_p = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\hat{f}_i - f_i\right]^2} \quad (12)$$

where $\hat{f}_i$ is the predicted pitch frequency from the i^{th} frame and f_i is the reference pitch for the i^{th} frame.

Table 1 shows the RMS pitch error and classification error for the systems described in section 3. The GMM systems use either the closest cluster to the input MFCC vector (equation 6), or the posteriori weighted MAP prediction (equation 7). In both cases it was found that using $K=64$ clusters gave best performance. Results for HMM-based prediction are shown using from 1 to 5 clusters within each state with posteriori weighted MAP prediction of the pitch (equation 9).

	Classification error	*RMS error*
GMM (closest)	13.9 %	11.7 Hz
GMM (posteriori)	12.4 %	10.8 Hz
HMM 1 cluster	6.9 %	9.7 Hz
HMM 2 clusters	5.8 %	5.4 Hz
HMM 3 clusters	5.7 %	4.1 Hz
HMM 4 clusters	5.7 %	3.5 Hz
HMM 5 clusters	5.7 %	3.1 Hz

Table 1: Classification and RMS errors for pitch prediction

The performance of the two cluster-based systems shows that attempting to identify the best cluster from which to predict the pitch is not as effective as taking a posteriori weighted prediction from all the clusters. This can be attributed to the difficulty in identifying the "correct" cluster. In fact in a preliminary test which was artificially supplied with correct cluster information

the performance was comparable to that using the posteriori weighted prediction. The HMM-based prediction is shown to give considerably more accurate frame classification and a lower RMS pitch error. Increasing the number of clusters in each state of the HMM enables more detailed modeling of the joint distribution of MFCC and pitch and this results in a reduction of frame classification error to 5.7% and in RMS error to 3.1Hz. The accuracy of the recogniser was 97% which means that 3% of digits were aligned to incorrect models from which voicing and pitch were extracted. It should be noted that the significant majority of frame classification errors arise from incorrect voiced/unvoiced decisions which occur in low energy regions at the start and end of speech.

To illustrate the effectiveness of the 5 cluster HMM-based system, figure 3 compares the predicted pitch contour (solid line) with the reference pitch (dashed line) for the digit sequence "6-5-5-6-3-2-0".

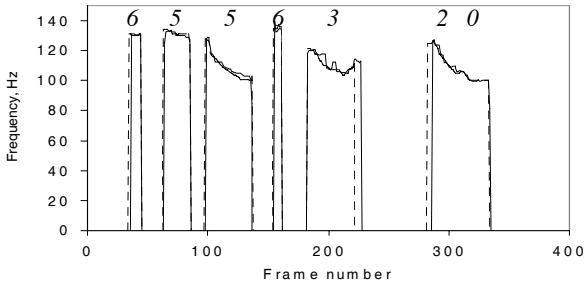

Figure 3: Comparison of predicted and reference pitch contours

The figure shows that the predicted pitch closely tracks the reference pitch throughout the spoken digits. Pitch classification errors can be observed around frame 225 where unvoiced frames are labeled voiced at the end of the digit *three* and around frame 280 where voiced frames just before the start of the digit *two* are labeled as unvoiced. These correspond to very low energy signals which indicates that pitch classification may benefit from an energy term in the feature vector in addition to MFCC(0).

4.2 Speech reconstruction results

The purpose of pitch prediction in this work is to enable an acoustic speech signal to be reconstructed from a stream of MFCC vectors with no additional pitch information. To illustrate the effectiveness of this approach, figure 4a shows the narrowband spectrogram of the original speech utterance "6-5-5-6-3-2-0" – as used in figure 3. Figure 4b shows the spectrogram of the speech signal reconstructed from MFCC vectors and reference pitch using the sinusoidal model described in section 2. Figure 4c shows the spectrogram of the speech signal reconstructed solely from MFCC vectors with the pitch predicted using the 5 cluster HMMs.

Comparing figures 4a and 4b shows the spectral smoothing which MFCC extraction has introduced as a result of the mel-filterbank and truncation of DCT coefficients. Only slight differences are observed between figures 4b and 4c which arise from pitch prediction errors. It is interesting to note that the voiced/unvoiced classification errors observed in figure 3 have very little effect in the reconstructed speech as they are associated with very low energy regions of the speech.

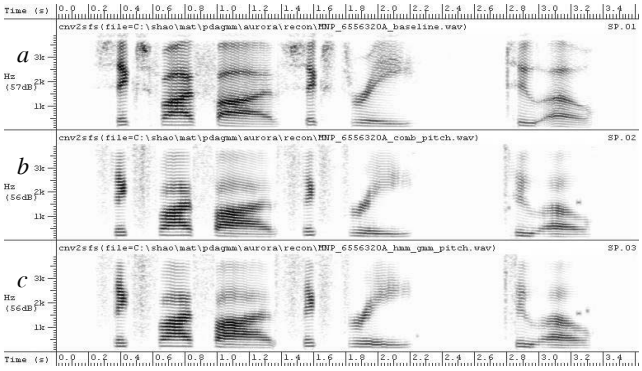

Figure 4: Comparison of narrowband spectrograms

A series of informal listening tests also revealed that speech reconstructed from the predicted pitch is virtually indistinguishable from that reconstructed from the reference pitch.

5. CONCLUSION

This work has introduced a system which enables an acoustic speech signal to be reconstructed solely from MFCC vectors. To accomplish this two methods have been developed which enable the pitch frequency to be predicted from an MFCC vector. A formal evaluation of both the RMS pitch error and voiced/unvoiced decision shows that the system can deliver reliable measurements. Speech reconstructed from the predicted pitch, using a sinusoidal model, is almost indistinguishable from that reconstructed using the reference pitch. This work has been based around a set of digit models, but further work will consider the use of phoneme models and unrestricted grammars.

6. REFERENCES

[1] ESTI document - ES 201 108 – STQ: DSR – Front-end feature extraction algorithm; compression algorithm, 2000.

[2] D. Chazan et al, "Speech reconstruction from MFCCs and pitch", Proc. ICASSP, 2000

[3] B.P. Milner and X. Shao, "Speech reconstruction from MFCCs using a source-filter model", Proc. ICSLP, 2002

[4] K. Fujinaga et al, "Multiple regression hidden Markov model", Proc. ICASSP, 2001

[5] H. Singer and S. Sagayama, "Pitch dependent phone modeling for HMM based speech recognition", Proc. ICASSP, 1992

[6] A. Kain and Y. Stylianou, "Stochastic modeling of spectral adjustment for high quality pitch modification", Proc. ICASSP, 2000

[7] En-Najjary et al, "A new method for pitch prediction from spectral envelope and its application in voice conversion", Proc. Eurospeech, 2003

[8] R. McAulay and T. Quatieri, "Sinsoidal Coding," Ch. 4, Speech Coding and Synthesis, Elsevier, 1995

[9] D. Chazan et al, "Efficient periodicity extraction based on sine-wave representation and its application to pitch determination of speech signals", Proc Eurospeech, 2001

[10] B.R. Ramakrishnan, "Reconstruction of incomplete spectrograms for robust speech recognition", PhD thesis, Carnegie Mellon University, 2000