MAP Prediction of Pitch 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 method of predicting pitch and voicing from mel-frequency cepstral coefficient (MFCC) vectors. Two maximum a posteriori (MAP) methods are considered. The first models the joint distribution of the MFCC vector and pitch using a Gaussian mixture model (GMM) while the second method also models the temporal correlation of the pitch contour using a combined hidden Markov model (HMM)-GMM framework. Monophone-based HMMs are connected together in the form of an unconstrained (HMM)-GMM framework. Monophone-based HMMs are therefore not possible to simply invert the stages involved in MFCC extraction to re-create the acoustic speech signal.

To enable speech to be reconstructed several schemes have been proposed which complement the feature vector (e.g. MFCCs) with the pitch frequency and voicing associated with an MFCC vector. This enables speech to be reconstructed solely from the MFCC stream and avoids the need for modifications to feature extraction and also reduces transmission overheads. This also allows speech to be reconstructed from MFCC parameterized utterances which have no associated pitch measurement. Pitch prediction 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 with modifications made to the pitch contour [6]. Prediction of the pitch from spectral envelopes modified in a voice conversion system has also made use of this correlation [7].

Section 2 of this work provides a brief overview of speech reconstruction from MFCC vectors and pitch using a sinusoidal model. Section 3 develops a Gaussian mixture model (GMM) for predicting pitch from MFCC vectors and then extends this to a hidden Markov model (HMM) framework. The HMM-GMM structure is also used to determine the voicing associated with an MFCC vector. Measurements of the predicted pitch accuracy are described in section 4 together with an examination of the resultant reconstructed speech signal.

1. Introduction

Distributed speech recognition (DSR) has been shown to give significant improvements in speech recognition accuracy from mobile devices by replacing low bit-rate speech codecs with the front-end processing component of the speech recogniser [1]. Such systems transmit feature vectors (such as MFCCs) directly to the speech recogniser. However, because feature vectors are a compact representation, optimized for discriminating between different speech sounds, they contain insufficient information to enable reconstruction of the original speech signal. In particular valuable speaker information, such as pitch, is lost in the transformation. It is therefore not possible to simply invert the stages involved in MFCC extraction to re-create the acoustic speech signal.

To enable speech to be reconstructed several schemes have been proposed which complement the feature vector (e.g. MFCCs) with the pitch frequency and voicing associated with that frame [2][3]. These schemes require feature extraction on the terminal device to be modified such that pitch estimation is included and also need additional bandwidth in order to transmit the pitch and voicing. Such a system has been included in the latest version of the ETSI Aurora standard [1].

The aim of this work is to predict the pitch frequency and voicing from the MFCC vectors themselves. This enables speech to be reconstructed solely from the MFCC stream and avoids the need for modifications to feature extraction and also reduces transmission overheads. This also allows speech to be reconstructed from MFCC parameterized utterances which have no associated pitch measurement. Pitch prediction 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 with modifications made to the pitch contour [6]. Prediction of the pitch from spectral envelopes modified in a voice conversion system has also made use of this correlation [7].

2. Speech Reconstruction

An effective model for speech reconstruction has been shown to be the sinusoidal model of speech [8]. This synthesises a speech signal, $x(n)$, by summing together $L$ sinusoids of varying amplitude, $A_l$, frequency, $\omega_l$, and phase, $\theta_l$,

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

(1)

To facilitate speech reconstruction from MFCC vectors using this model an estimate of the spectral envelope can be calculated from an MFCC vector by zero padding and applying an inverse discrete cosine transform. An exponential operation applied to the resulting log mel-filterbank estimate, followed by interpolation, gives a smoothed magnitude spectral estimate, $\hat{X}(\omega)$ [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, $\omega_l$, can be estimated from the pitch frequency, $\omega_p$, by assuming a harmonic relationship, i.e. $\omega_l = l \omega_p$. The amplitude, $A_l$, of the sinusoids can then be computed from the smoothed magnitude spectral estimate,

$$A_l = |\hat{X}(l \omega_p)|$$

(2)

The phase offset, $\theta_l$, 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.

3. MAP Prediction of Pitch

This section proposes two methods for predicting the pitch frequency associated with an MFCC vector and classifying the
vector as to whether it represents voiced or unvoiced speech. These methods are based on modeling the joint density of an MFCC vector, $x$, and the associated pitch frequency, $f$. In the training phase of this modeling an augmented feature vector, $y$, is defined as,

$$ y = [x, f]^T $$

(3)

The MFCC vector used in this work comprises static coefficients 0 to 12. For modeling the pitch is estimated from the time-domain data using autocorrelation methods and is subsequently manually corrected where necessary. For unvoiced frames the pitch frequency is set to zero.

3.1. GMM-based prediction

This system models the joint density 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,

$$ \mu_k^x = \left[ \mu_k^x \right] \quad \text{and} \quad \Sigma_k^x = \left[ \Sigma_k^{xx} \Sigma_k^{xf} \right] \left[ \Sigma_k^{fx} \Sigma_k^{ff} \right] $$

(4)

These clusters can now be used to make a maximum a posteriori (MAP) prediction [9] of the pitch frequency, $\hat{f}$, given an input MFCC vector, $x_i$. A simple method of predicting the pitch is to consider only the closest cluster, in some sense, to the input MFCC vector. However, previous work [10] has found pitch prediction to be more accurate by taking a weighted contribution from all clusters in the GMM,

$$ \hat{f}_i = \sum_{k=1}^{K} h_k(x_i) \left( \mu_k^x + \Sigma_k^x \sum_{i=1}^{K} \frac{(x_i - \mu_i^x)^T}{\Sigma_i^x} \right) $$

(5)

The term $h_k(x_i)$ weights the pitch prediction contribution made by each cluster in the GMM by the posterior probability of $x_i$ belonging to it,

$$ h_k(x_i) = \frac{\alpha_i \sum_{l=1}^{K} \alpha_l \frac{(x_i - \mu_l^x)^T}{\Sigma_l^x}}{\sum_{l=1}^{K} \sum_{i=1}^{K} \alpha_i \frac{(x_i - \mu_i^x)^T}{\Sigma_i^x}} $$

(6)

$\alpha_i$ is the prior probability of the $k$th cluster in the GMM which is determined from the proportion of vectors assigned to that cluster in training. $p(x_i | k)$ is the marginal distribution of the MFCC vector for the $k$th cluster.

3.2. HMM-based prediction

The unsupervised training used to create the GMM does not allow the temporal correlation in the pitch contour to be modeled. To overcome this, earlier work [10] utilized a set of digit-based HMMs to model the joint density of pitch and MFCC vector within the framework of a connected digit grammar. This enabled the temporal correlation of pitch to be modeled both at the state level, within the digit HMMs, and at the model level through the connection of the HMMs. This system gave more accurate pitch prediction than the GMM but was specific to the connected digit task. This work extends the modeling to a general speech vocabulary by using a set of 3-state phoneme HMMs which are connected as an unconstrained monophone network. This allows pitch to be predicted from unconstrained speech input rather than being restricted to digit strings. Figure 1 illustrates the joint modeling of the feature space using in figure 1a the GMM and in figure 1b the HMM-GMM.

![Figure 1: a) GMM clustering, b) HMM-based states](image)

The clustering in the creation of the GMM does not consider the temporal position of vectors and partitions the feature space into a set of unconnected clusters. However, the HMM modeling of the joint density partitions the feature space into a series of 3-state HMMs, $\lambda_{w}$, which are linked according to the unconstrained monophone network. Applying Viterbi decoding to a stream of MFCC vectors to determine their model and state sequence enables pitch to be predicted from a localized region of the joint feature space. Additionally, the predicted pitch values will also exhibit temporal correlation through the connection of the states in the models. The solid line in figure 1b illustrates a feature vector trajectory through the HMMs from which pitch can be predicted.

The first stage of modeling the joint density involves training a set of monophone-based HMMs, $\Lambda$, on just the MFCC component, $x$, of the augmented vector, $y$. On this task a set of 40 monophone models were produced together with silence and short pause models.

The second stage of training models the joint distribution of the MFCC vector and pitch within each state of the HMM using a state-specific GMM. Using the set of monophone HMMs the training data is forced aligned to the HMMs using Viterbi decoding. Vectors assigned to each state, $s$, of each model, $w$, for the entire training data set are then pooled together. Vectors corresponding to unvoiced speech (as indicated by the pitch component) are removed to ensure that the joint density of MFCC and pitch is not distorted. Clustering is applied to the pooled vectors within each voiced state using the EM algorithm as described in section 3.1. This results in a set of means, $\mu_k^{x,s,w}$, and covariances, $\Sigma_k^{x,s,w}$, corresponding to the $k$th cluster of the GMM associated with state $s$ of speech model $w$.

Pitch prediction, for voiced frames, is made from the MFCC vectors by first determining the model and state sequence from the set of monophone models using Viterbi decoding with the unconstrained monophone grammar. For each MFCC vector, $x$, the allocated model, $m$, and state, $q$, are then used to determine the MAP prediction of the pitch,

$$ \hat{f}_i = \sum_{k=1}^{K} h_k(x_i) \left( \mu_{k,di,m}^x + \Sigma_{k,di,m}^x \sum_{i=1}^{K} \frac{(x_i - \mu_i^x)^T}{\Sigma_i^x} \right) (x_i - \mu_{k,di,m}^x)^T $$

(7)
where \( h_{k,w}(x) \) is computed using equation (6) with \( P(x|c_{ij}) \) specific to state \( q_{ij} \) of model \( m_{k} \). Finally the sequence of pitch predictions is smoothed using a median filter of width 5 to reduce spurious errors.

### 3.3. Voiced/unvoiced classification

In the training phase of the HMM-GMMs, the prior probability of a state being voiced is related to the proportion of voiced vectors and unvoiced vectors associated with each state \( s \) of each model \( w \) as represented by \( \nu_{s,w} \) and \( u_{s,w} \) respectively, where \( \nu_{s,w} + u_{s,w} = 1 \). Previous work [10] classified an MFCC vector, \( x_{i} \), as either voiced or unvoiced according to the model, \( m_{i} \), and state, \( q_{i} \), alignment of the vector, i.e.

\[
\text{voicing}_{i} = \begin{cases} \text{voiced} & \nu_{m_{i},q_{i}}>\theta \\ \text{unvoiced} & \nu_{m_{i},q_{i}}\leq\theta \end{cases} \tag{8}
\]

Figure 2 shows a histogram of the prior voicing probabilities, \( \nu_{s,w} \), taken from each of the 3 states of the 42 HMMs in the monophone recogniser.

![Figure 2: Histogram of voicing proportions for 126 states](image)

The histogram shows that some states are strongly voiced (for example 81 states have a prior voicing probability over 0.9) while for other states the distinction is not so clear. To avoid making a hard voicing decision a soft decision is proposed based on the posterior probabilities. For each state there are \( K \) clusters within the GMM that are created from the pooled voiced vectors and unvoiced vectors associated with each state \( s \) of each model \( w \) respectively, where \( \nu_{s,w} + u_{s,w} = 1 \). For evaluation purposes both the accuracy of classifying vectors as voiced or unvoiced and the RMS pitch prediction error for voiced frames has been measured. Pitch classification error is defined as,

\[
E_{c} = \frac{N_{UV} + N_{UV} + N_{20\%}}{N_{\text{Total}}} \times 100\% \tag{12}
\]

where \( N_{UV} \) is the number of unvoiced frames classified as voiced, \( N_{UV} \) 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%. \( N_{\text{Total}} \) is the total number of frames in the test. For frames correctly classified as voiced, the RMS pitch error is computed as,

\[
E_{p} = \frac{1}{N} \sum_{i=1}^{N} \left( f_{i} - f_{i} \right)^{2} \tag{13}
\]

where \( f_{i} \) 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, \( E_{p} \), and the classification error, \( E_{c} \), using both the threshold method of equation (8) (with \( \theta = 0.2 \)) and the probability-based method of equation (11). Results are shown using the GMM described in section 3.1 using \( K = 64 \) clusters and the HMM-GMMs using from 1 to 16 clusters in each state.

<table>
<thead>
<tr>
<th>Model</th>
<th>( E_{c} ) - thresh</th>
<th>( E_{c} ) - prob</th>
<th>( E_{p} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMM</td>
<td>18.29 %</td>
<td>18.01 %</td>
<td>12.2 Hz</td>
</tr>
<tr>
<td>HMM 1</td>
<td>9.66 %</td>
<td>7.93 %</td>
<td>14.1 Hz</td>
</tr>
<tr>
<td>HMM 2</td>
<td>9.52 %</td>
<td>7.75 %</td>
<td>12.9 Hz</td>
</tr>
<tr>
<td>HMM 4</td>
<td>9.44 %</td>
<td>7.54 %</td>
<td>12.0 Hz</td>
</tr>
<tr>
<td>HMM 8</td>
<td>9.36 %</td>
<td>7.36 %</td>
<td>11.2 Hz</td>
</tr>
<tr>
<td>HMM 16</td>
<td>9.46 %</td>
<td>7.35 %</td>
<td>10.4 Hz</td>
</tr>
</tbody>
</table>

Table 1: Pitch prediction and voicing classification.

The results show that in all model configurations the probability-based voicing classification is significantly more accurate than using a voicing threshold. A slight reduction in voicing error is also observed as the number of clusters in the HMM-GMM system is increased. In terms of the RMS pitch error the GMM appears comparable with the 4 cluster HMM-GMM. However, analysing the prediction errors made by the GMM showed that a significant number (>3%) were greater than 20% meaning they are treated as voicing errors. Taking this into account shows the GMM to give the least accurate

### 4. Experimental Results

The experiments in this section first measure the accuracy of pitch prediction and voicing classification. Secondly speech reconstruction using the predicted pitch and voicing is compared to reconstruction using the real pitch and voicing.

#### 4.1. Pitch prediction accuracy

A set of 803 phonetically rich sentences has been used to train the GMM and HMM-GMMs. A further set of 246 phonetically rich sentences, comprising a total of 130,000 vectors, has been used for testing. Each sentence is approximately 5 seconds in duration and from these 13-D MFCC vectors have been extracted at a rate of 100 vectors per second in accordance with the ETSI Aurora standard [1]. For evaluation purposes both the accuracy of classifying vectors as voiced or unvoiced and the RMS pitch prediction error for voiced frames has been measured. Pitch classification error is defined as,

\[
E_{c} = \frac{N_{UV} + N_{UV} + N_{20\%}}{N_{\text{Total}}} \times 100\% \tag{12}
\]

where \( N_{UV} \) is the number of unvoiced frames classified as voiced, \( N_{UV} \) 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%. \( N_{\text{Total}} \) is the total number of frames in the test. For frames correctly classified as voiced, the RMS pitch error is computed as,

\[
E_{p} = \frac{1}{N} \sum_{i=1}^{N} \left( f_{i} - f_{i} \right)^{2} \tag{13}
\]

where \( f_{i} \) 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, \( E_{p} \), and the classification error, \( E_{c} \), using both the threshold method of equation (8) (with \( \theta = 0.2 \)) and the probability-based method of equation (11). Results are shown using the GMM described in section 3.1 using \( K = 64 \) clusters and the HMM-GMMs using from 1 to 16 clusters in each state.

<table>
<thead>
<tr>
<th>Model</th>
<th>( E_{c} ) - thresh</th>
<th>( E_{c} ) - prob</th>
<th>( E_{p} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GMM</td>
<td>18.29 %</td>
<td>18.01 %</td>
<td>12.2 Hz</td>
</tr>
<tr>
<td>HMM 1</td>
<td>9.66 %</td>
<td>7.93 %</td>
<td>14.1 Hz</td>
</tr>
<tr>
<td>HMM 2</td>
<td>9.52 %</td>
<td>7.75 %</td>
<td>12.9 Hz</td>
</tr>
<tr>
<td>HMM 4</td>
<td>9.44 %</td>
<td>7.54 %</td>
<td>12.0 Hz</td>
</tr>
<tr>
<td>HMM 8</td>
<td>9.36 %</td>
<td>7.36 %</td>
<td>11.2 Hz</td>
</tr>
<tr>
<td>HMM 16</td>
<td>9.46 %</td>
<td>7.35 %</td>
<td>10.4 Hz</td>
</tr>
</tbody>
</table>

Table 1: Pitch prediction and voicing classification.

The results show that in all model configurations the probability-based voicing classification is significantly more accurate than using a voicing threshold. A slight reduction in voicing error is also observed as the number of clusters in the HMM-GMM system is increased. In terms of the RMS pitch error the GMM appears comparable with the 4 cluster HMM-GMM. However, analysing the prediction errors made by the GMM showed that a significant number (>3%) were greater than 20% meaning they are treated as voicing errors. Taking this into account shows the GMM to give the least accurate
pitch predictions. Increasing the number of clusters in the HMM-GMM from 1 to 16 gives almost a 4Hz reduction in RMS pitch error as this enables more accurate modeling of the localized joint density of MFCCs and pitch.

To illustrate the accuracy of the 16 cluster HMM-GMM system, figure 3 compares the predicted pitch contour (solid line) with the reference pitch contour (dashed line) for the sentence “look out of the window and see if it’s raining”. The figure shows that the predicted pitch follows closely the reference pitch throughout the sentence, although it has more variation than the reference pitch. Some of this variability has been removed by the median filter although applying further filtering causes too much detail to be lost. Generally, classification of the vectors as voiced or unvoiced follows closely the reference voicing classification. Analysis of voicing errors from the test set has indicated that most occur in relatively low energy regions of speech which frequently occur at the start and end of words. A typical voicing error of this type can be observed around vector 425 where voiced frames are incorrectly identified as unvoiced.

4.2. Speech reconstruction results

The purpose of pitch prediction has been to enable an acoustic speech signal to be reconstructed from a stream of MFCC vectors. To illustrate the effectiveness of this figure 4a shows the narrowband spectrogram of the original speech utterance “look out of the window and see if it’s raining” – 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 pitch predicted using the 16 cluster HMM-GMM.

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 and these arise from pitch prediction errors. The effect of the incorrect classification of voiced frames as unvoiced at the end of the word “raining” can be seen in figure 4c. This has caused a wideband signal be synthesized as opposed to the continuation of pitch harmonics seen in figure 4b, and can be heard as a burst of white noise.

Informal listening tests revealed very little difference between speech reconstructed from the predicted pitch and speech reconstructed using the reference pitch. In particular small errors in the predicted pitch frequency were inaudible. Many of the voicing errors associated with low energy regions at the end of words also tended to be virtually inaudible.

5. Conclusion

This work has developed a method of predicting the pitch frequency and voicing of a frame of speech from an MFCC vector. A system based on modeling the joint density of pitch and MFCCs using a GMM gave reasonable results, but modeling the temporal correlation of the pitch through a combined HMM-GMM system gave superior results. Voicing classification has been improved over a previous threshold-based system through the use of a probabilistic decision on voicing. Using the predicted pitch contour for speech reconstruction has given almost indistinguishable quality and intelligibility to that reconstructed from the reference pitch contour. In particular, many of the pitch frequency and voicing errors are not audible in the acoustic signal.

6. References

[1] ETSI document - ES 202 212- STQ: DSR; Extended advanced front-end feature extraction algorithm; Compression algorithms; Back-end speech reconstruction algorithm, 2003