INTEGRATED PITCH AND MFCC EXTRACTION FOR SPEECH RECONSTRUCTION AND SPEECH RECOGNITION APPLICATIONS

Xu Shao, Ben Milner and Stephen Cox

School of Computing Sciences, University of East Anglia, Norwich, UK

{x.shao, b.milner, s.j.cox}@uea.ac.uk

ABSTRACT

This paper proposes an integrated speech front-end for both speech recognition and speech reconstruction applications. Speech is first decomposed into a set of frequency bands by an auditory model. The output of this is then used to extract both robust pitch estimates and MFCC vectors. Initial tests used a 128 channel auditory model, but results show that this can be reduced significantly to between 23 and 32 channels.

A detailed analysis of the pitch classification accuracy and the RMS pitch error shows the system to be more robust than both comb function and LPC-based pitch extraction. Speech recognition results show that the auditory-based cepstral coefficients give very similar performance to conventional MFCCs. Spectrograms and informal listening tests also reveal that speech reconstructed from the auditory-based cepstral coefficients and pitch has similar quality to that reconstructed from conventional MFCCs and pitch.

1. INTRODUCTION

Speech communication from mobile devices has traditionally been made using low bit-rate speech codecs. The low bit-rates at which these codecs operate introduce a slight distortion of the speech signal which becomes more severe in noisy conditions. When input into a speech recogniser, this distortion causes a noticeable reduction in accuracy. To overcome this problem the technique of distributed speech recognition (DSR) [1] has been proposed by the ETSI Aurora group.

DSR replaces the codec on the terminal device with the feature extraction component of the speech recogniser and so removes codec-based distortion from the speech recogniser input. This results in a significant improvement in speech recognition accuracy. However, because speech feature vectors are designed to be a compact representation, optimized for discriminating between different speech sounds, they do not contain sufficient information to enable reconstruction of the original speech signal. In particular, valuable speaker information, such as pitch, is lost. However, several schemes have been proposed recently for reconstructing speech from a combination of MFCC vectors and pitch. These have been based on either a sinusoidal model [2] or a source-filter model [3] of speech production. An extension of this work also considered the reconstruction of clean speech from noise contaminated MFCC vectors and a robust pitch estimate [4].

In these systems, the MFCC vectors and pitch are extracted using separate speech processors. For example in [4] a 128-channel auditory model [11] provided robust estimates of the pitch. The aim of this work is to integrate the MFCC extraction and pitch estimation components into a single speech front-end. For both pitch estimation and MFCC extraction, the speech signal is decomposed into a number of discrete frequency bands either by an auditory model or mel-filterbank. It is therefore reasonable to combine this into a single system and this is described in section 2. A detailed evaluation of the pitch extraction component is described in section 3 and a comparison made with alternative pitch extraction methods. Speech recognition and speech reconstruction results are presented in section 4 and a conclusion is given in section 5.

2. INTEGRATED FRONT-END

This section describes the proposed integrated speech front-end and back-end systems, which are illustrated in figure 1. The front-end comprises three main parts; auditory model, MFCC extraction and pitch estimation. Three features are output across the communication channel; MFCC vectors, pitch and energy. At the remote back-end the MFCC vectors and pitch estimates are used for speech reconstruction. For speech recognition the MFCC vectors and energy are used together with their temporal derivatives.

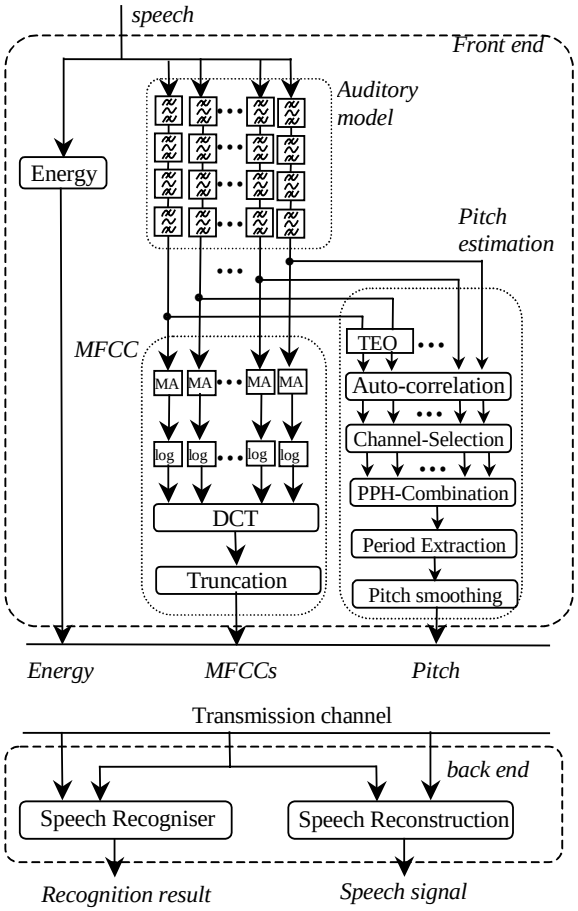

Figure 1: Integrated front-end and back-end systems

Decomposition of the input speech signal into frequency bands is performed by the auditory model. The output of this is used by both the MFCC extraction and pitch estimation components. The original pitch estimation system proposed in [7] used a 128 channel auditory model. However most MFCC extraction algorithms use significantly fewer channels (e.g. 23 for the Aurora standard). One of the aims of this work is to vary the size of the auditory filterbank to produce a compromise that gives both robust pitch estimates and MFCCs which result in accurate speech recognition.

2.1. Auditory Model

The auditory model upon which the speech is decomposed into frequency bands was proposed in [11]. Auditory models have been successfully used for robust pitch estimation [6][7] and therefore form the first stage of this integrated front-end. Decomposition of the speech signal into a number of frequency bands is achieved using a series of non-linearly spaced and overlapping bandpass filters. The spacing of these bandpass filters is determined by an equivalent rectangular bandwidth (ERB) scale [10] and is similar to mel-scale spacing.

In the original system a set of 128 channels was used. These give sufficient frequency response detail which the subsequent pitch estimation component uses. However for MFCC extraction, the Aurora standard defines just 23 channels. Work shown in later sections examines the effect of reducing the number of channels from 128 to 23 in terms of the resulting speech recognition performance and pitch estimation accuracy.

2.2. Feature Extraction

The output of the auditory model takes the form of a series of time-domain samples from each of the bandpass filters. In conventional MFCC feature extraction a windowing function captures a short-time frame of speech. From this a Fourier transform determines the magnitude spectrum and this is then quantised in frequency using a mel-spaced filterbank. To generate a filterbank vector from the time-domain filter outputs of the auditory model a mean amplitude (MA) filter is employed. This outputs the root mean square amplitude, c_k, from each bandpass filter, k, at 10ms intervals from a 25ms buffer of time-domain samples, where

$$c_k = \sqrt{\frac{1}{N}\sum_{n=0}^{N-1} x_k(n)^2} \qquad (1)$$

$x_k(n)$ is the n^{th} time-domain sample from the k^{th} bandpass filter in the 25ms buffer, N is the buffer length ($N=200$ samples for the 8kHz sampling frequency). This is consistent with the frame width and frame rate used in the Aurora MFCC standard. The final three stages are logarithm, discrete cosine transform and truncation. These are identical to the last three stages in conventional MFCC extraction. It should be noted that the positioning of the auditory filters is close to, but not exactly, mel-scaled. Therefore the features extracted by this system are not strictly MFCCs. However, for the purpose of this work they are referred to as auditory model-based MFCCs.

2.3. Robust Pitch Estimation

Auditory models have been demonstrated as being one of the most successful methods for accurately estimating pitch [6][7]. For speech reconstruction, especially in noisy conditions, it is vital to have a robust pitch estimate. Previous work in this area successfully used a 128 channel auditory model to achieve this [4]. To estimate pitch, the bandpass filter outputs from the auditory model are divided into two components; a high frequency part, where center frequencies are greater than 800Hz, and a low frequency part. An energy envelope is then extracted from the high frequency part using a Teager energy operator (TEO) [6]. An auto-correlogram is obtained from the energy envelope of the high frequency component and the remaining low frequency signals. Channels contaminated by noise are removed by discriminative algorithms [6][7] which analyse the structure of the auto-correlogram. Finally the pitch contour is extracted using a pseudo-periodic histogram (PPH) from the summation of the remaining clean channels [6]. This is subsequently smoothed to produce a robust pitch estimate.

3. EVALUATION OF PITCH ESTIMATION

The aim of this section is to examine the effect of reducing the number of channels in the auditory model in terms of pitch estimation accuracy. In particular the number of channels is reduced from 128 to 23 to be comparable with the number of filterbank channels used in conventional MFCC extraction.

3.1. Method of Pitch Evaluation

The pitch extraction system is required to produce two outputs; a flag indicating whether the speech is voiced or unvoiced and, for voiced speech, an estimate of the pitch frequency. It is therefore appropriate to measure the effectiveness of the pitch extraction system using these two criteria.

Before defining these measures it is useful to examine the types of error made in pitch extraction. One form of error is a mis-classification, such as a voiced frame being classified as unvoiced or an unvoiced frame being classified as voiced. Another type of error is a correct classification but a wrong pitch frequency value. To illustrate the second kind of error, a histogram showing the percentage pitch frequency error is shown in figure 2, taken across 75 Messiah sentences. The reference pitch value has been provided by a hand-checked laryngograph signal. For clarity the figure also shows an expanded section of the lower portion of the histogram.

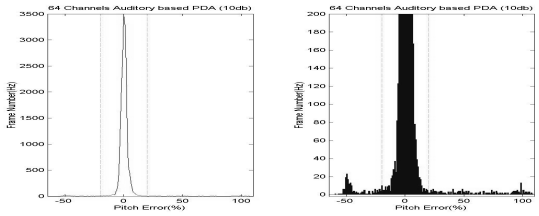

Figure 2: Distribution of percentage pitch errors

The majority of pitch estimates are very close to the measured pitch and apparently have a Gaussian distribution. In fact the dotted vertical line shows the range of pitch estimates that are within +/-20% of the reference pitch - over 97% of pitch estimates are within this range. However a number of errors are concentrated around the –50% and +100% points. These correspond to pitch halving errors and pitch doubling errors which are fairly common mistakes made in pitch estimation.

After consideration of these results, it was decided to label pitch estimation errors of more than 20% as being incorrectly classified [8]. This also means that when calculating the root

mean square (RMS) pitch error, the effect of pitch halving and pitch doubling in the estimation does not dominate the result.

Therefore pitch classification error, E_c, is expressed as

$$E_C = \frac{N_{V/U} + N_{U/V} + N_{>20\%}}{N_{Total}} \times 100\% \quad (2)$$

where $N_{V|U}$ is the number of voiced frames classified as unvoiced, $N_{U|V}$ is the number of unvoiced frames classified as voiced 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.

For frames correctly classified as voiced, the RMS pitch error provides a measure of the accuracy of estimation. The overall RMS error, E_p, is computed as

$$E_p = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\hat{f}_0(i) - f_0(i)\right]^2} \quad (3)$$

where $\hat{f}_0(i)$ is the pitch frequency estimate from the i^{th} frame and $f_0(i)$ is the pitch for the i^{th} frame measured from the laryngograph signal. N is the total number of voiced frames in the test, which is around 23,000 frames for the 75 utterances.

3.2. Assessment of Pitch Estimation

This section evaluates the effectiveness of the pitch estimation scheme using the two performance measures described in the previous section. In particular the effect of reducing the number of channels in the auditory model from 128 down to 23 is examined. The test data used in these experiments is composed of 75 utterances from a set of Messiah sentences. To observe the effect of noise on pitch estimation, examples of office noise from the Aurora database have been artificially added to the speech at a range of signal-to-noise ratios (SNRs) from 30dB to 0dB. Reference pitch measurements come from a laryngograph signal which has been manually checked for accuracy.

The aim of the first experiment is to examine the effect of reducing the number of channels in the auditory model. Tests begin with the original 128 channels and go down to 23 channels (the same number used in the Aurora MFCC standard). Figure 3-a shows the frame classification error, E_c, for 128, 64, 32 and 23 channel auditory models across a range of noise levels. Figure 3-b illustrates the RMS pitch error, E_p, for the different number of channels and noise levels.

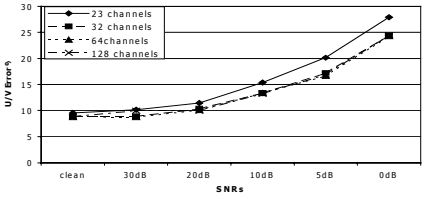

Figure 3-a: Frame classification error, E_c

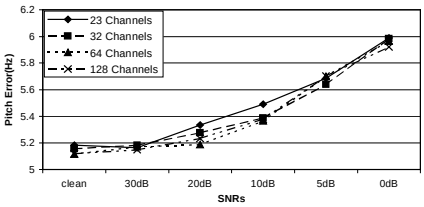

Figure 3-b: RMS pitch error, E_p

The result shows that errors for both frame classification and pitch measurement increase as the SNR decreases, as expected. Pitch measurements from the 128, 64 and 32 channel auditory model give almost identical performance. Reducing the number of channels to 23 causes a slight reduction in performance for more noisy speech.

A second set of tests were performed to compare the performance of the 32-channel auditory model-based pitch measurements with those obtained by alternative algorithms. These were the comb-function [5] and LPC-based pitch estimation through inverse filtering [9]. Figure 4 shows comparative results for both frame classification and RMS pitch error.

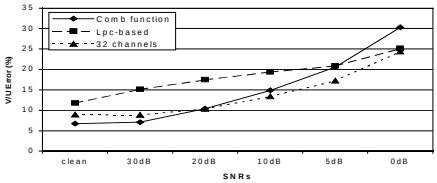

Figure 4-a: Comparative frame classification error, E_c

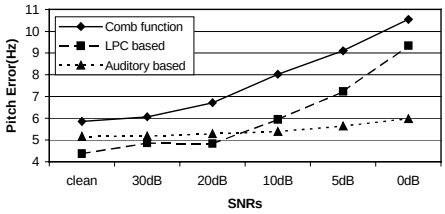

Figure 4-b: Comparative RMS pitch error, E_p

The pitch estimate from the LPC algorithm is the most accurate measurement for voiced frames under 20dB but deteriorates at SNRs below this. However, frame classification error from the LPC algorithm is the worst of the three algorithms. The comb function algorithm gives the best frame classification above SNRs of 20dB but gives the most inaccurate pitch estimates of the three algorithms. The auditory-based algorithm gives close to best performance for clean speech and is significantly more accurate for noisy speech.

4. EXPERIMENTAL RESULTS

The experimental results in this section test both the recognition accuracy of the auditory model-based MFCC vectors and the resultant speech quality after reconstruction.

4.1. Speech recognition performance

Speech recognition accuracy has been evaluated on the Aurora TI digits database which comprises 28000 digit strings for testing and 8440 for training. The digits are modeled using 16-state, 3-mode, diagonal covariance matrix HMMs, trained from 8440 digits strings. The training data covers a range of noises and from clean to an SNR of 0dB (as outlined in the Aurora test specification).

Three feature vector configurations have been tested; conventional MFCC vectors [1], MFCCs extracted from a 23-channel auditory model and MFCCs extracted from a 32-channel auditory model. In each case the final speech vector comprised static MFCCs 1 to 12 and log energy together with velocity and acceleration derivatives. Figure 5 shows

recognition accuracy for the three configurations for both clean and noisy speech.

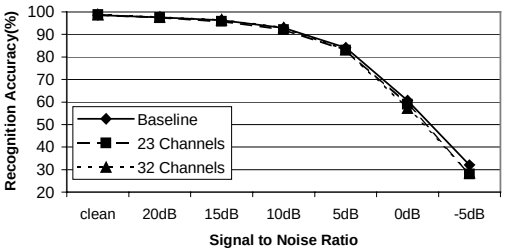

Figure 5: Comparative speech recognition accuracy

For clean speech, the recognition rate from the auditory-based features is slightly higher than that with conventional MFCCs - 98.72% to 98.57%. At lower SNRs the performance of the auditory-based MFCCs falls slightly below that of conventional MFCCs. For example at an SNR of 0dB the MFCCs derived from the 23-channel auditory model attain 59.03% while conventional MFCCs attain 60.69%. Changing from a 23-channel auditory filterbank to a 32-channel filterbank had negligible effect on accuracy.

4.2. Speech reconstruction

To examine the quality of reconstructed speech a set of Messiah sentences has been used. These are sampled at 8kHz and have then been contaminated by wideband noise from the Aurora database. Speech is reconstructed using a sinusoidal model of speech, with MFCC vectors being inverted to the filterbank domain and then interpolated to provide an estimate of the speech spectral envelope [2]. The pitch estimate is used to provide the finer harmonic detail. Spectral subtraction has also been applied to provide a clean speech spectral estimate from noise contaminated MFCCs [4].

Figure 6-a shows the spectrogram of the sentence “Look out of the window and see if it’s raining” spoken by a female speaker and contaminated by wideband noise at an SNR of 10dB. Figure 6-b illustrates the spectrogram of speech reconstructed from conventional MFCC vectors [4]. Figures 6-c and 6-d show spectrograms of speech reconstructed from 23 and 32 channel auditory-based MFCCs respectively.

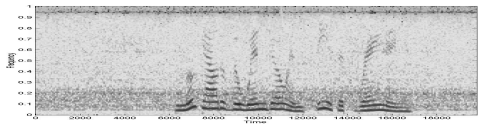

Figure 6-a: Original noisy signal (10dB SNR)

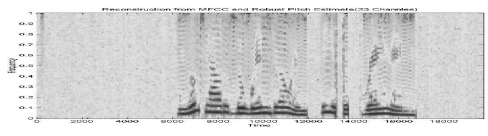

Figure 6-b: Reconstructed speech from MFCCs

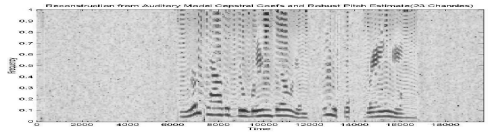

Figure 6-c: Speech from 23 channel auditory-MFCCs

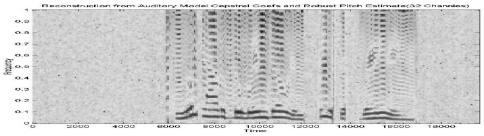

Figure 6-d: Speech from 32 channel auditory-MFCCs

The spectrograms show that speech reconstructed from the auditory-based MFCCs is almost identical to speech reconstructed from conventional MFCCs. A series of informal listening tests revealed this to be true across the range of Messiah sentences.

5. CONCLUSION AND DISCUSSION

This work has proposed an integrated speech front-end capable of generating features for both speech recognition and speech reconstruction. Evaluation of pitch estimation has shown that good performance can be obtained using significantly fewer filterbank channels than the original auditory model used. In combination with this, speech recognition tests have shown that auditory model-based MFCC vectors attain performance almost identical to conventional MFCCs. Using either a 23-channel or 32-channel filterbank has little effect on performance. In addition, speech reconstruction from the auditory model-based MFCCs gives very similar speech quality. Using a 32-channel auditory model gave slightly better pitch estimation, which is more important for speech reconstruction. These results conclude that a single front-end, based on an auditory model using either 23 or 32 channels, is feasible for both speech recognition and speech reconstruction.

6. REFERENCES

1. ESTI document - ES 201 108 – STQ: DSR – Front-end feature extraction algorithm; compression algorithm, 2000.
2. D. Chasan 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. X. Shao and B. P Milner, “Clean speech reconstruction from noisy MFCCs using a sinusoidal model”, Proc. ICASSP, 2003
5. D. Chasan et al, "Efficient periodicity extraction based on sine-wave representation and its application to pitch determination of speech signals", Proc Eurospeech, 2001.
6. J. Rouat, Y. C. Liu and D. Morissette, "Pitch determination and voiced/unvoiced decision algorithm for noisy speech", Speech Communication Journal, pp. 191-207., 1997
7. M. Wu, D. L. Wang and G. J. Brown, "A multi-pitch tracking algorithm for noisy speech", Proc. ICASSP, 2002.
8. L.Van Immerseel and J.P. Martens, “Pitch and voiced/unvoiced determination with an auditory model”. JASA,Vo1.9l, pp. 35l1-3526, 1992
9. L.R. Rabiner and R.W. Schafer, “Digital Processing of Speech Signals”, Prentice-Hall, 1978.
10. M. Slaney, “Auditory toolbox version 2,” Tech. Rep. 1998-010, Interval Research Corporation, 1998.
11. R.D. Patterson et al, SVOS Final Report: The Auditory Filterbank, APU Report 2341, 1988.