A Comparison of Estimated and MAP-Predicted Formants and Fundamental Frequencies with a Speech Reconstruction Application

Jonathan Darch and Ben Milner

School of Computing Sciences, University of East Anglia, Norwich

jonathan.darch@uea.ac.uk, b.milner@uea.ac.uk

Abstract

This work compares the accuracy of fundamental frequency and formant frequency estimation methods and maximum a posteriori (MAP) prediction from MFCC vectors with hand-corrected references. Five fundamental frequency estimation methods are compared to fundamental frequency prediction from MFCC vectors in both clean and noisy speech. Similarly, three formant frequency estimation and prediction methods are compared. An analysis of estimation and prediction accuracy shows that prediction from MFCCs provides the most accurate voicing classification across clean and noisy speech. On clean speech, fundamental frequency estimation outperforms prediction from MFCCs, but as noise increases the performance of prediction is significantly more robust than estimation. Formant frequency prediction is found to be more accurate than estimation in both clean and noisy speech. A subjective analysis of the estimation and prediction methods is also made by reconstructing speech from the acoustic features.

Index Terms: formant estimation, fundamental frequency estimation, speech reconstruction, speech synthesis, DSR

1. Introduction

Acoustic speech features, namely formants, fundamental frequency (f0) and voicing, are traditionally estimated from time-domain waveforms of speech or some representation such as short-time spectra or autocorrelation derived from waveforms. Within a distributed speech recognition (DSR) environment, the time-domain waveform is not transmitted to the remote back-end. Instead, mel-frequency cepstral coefficient (MFCC) vectors are transmitted which are designed for speech recognition. In order to reconstruct the original speech waveform from MFCCs an estimate of f0 is required because the MFCC extraction process aims to remove fundamental frequency information. The ETSI Extended Front End (XFE) and Extended Advanced Front End (XAFE) require an extra 800 bps to transmit voicing and f0 information [1]. Earlier work has demonstrated how to predict f0 from MFCCs using a model which describes the joint density of MFCC vectors and f0 [2]. Previous work has shown how this technique can be extended to predict formant frequencies in unvoiced and voiced speech [3]. Due to a lack of large hand-corrected formant frequency data, the technique has so far not been evaluated against reliable references. The motivation of this paper is to present a thorough comparison of acoustic speech feature prediction from MFCCs within a DSR environment against estimation techniques using the original waveforms, using hand-corrected references for evaluations.

In section 2 several methods for estimating acoustic speech features from time-domain waveforms are described. In a DSR environment, traditional estimation techniques cannot be used as the time-domain waveform is not available. Instead, a method of predicting acoustic speech features from MFCC vectors is described in section 3. Experimental results are presented in section 4. Section 5 considers the reconstruction of speech from reference, estimated and predicted acoustic speech features. Conclusions are drawn in section 6.

2. Estimation of Acoustic Speech Features

This section briefly describes the set of f0 and formant estimation methods considered in this work. All tools are freely available. The sampling frequency of the data is 8 kHz and the analysis frame rate is 100 Hz.

2.1. Fundamental Frequency Estimation Methods

- AMDF: The AMDF (average magnitude difference function) [4] routine included as part of the free Snack Toolkit[1] is a simple autocorrelation-based method of estimating f0.
- ANAL: The SFS (Speech Filing System)[2] 'fxanal' routine estimates f0 using the normalised cross correlation function on linear prediction coefficient residuals [5]. Dynamic programming is used to find the best f0 estimates at each frame.
- RAPT: The Snack Toolkit also includes the ESPS 'get_f0' function which is the same as the RAPT (robust algorithm for pitch tracking) method. Here, f0 is estimated from the normalised cross correlation function. Dynamic programming is used to improve the estimation accuracy [6].
- YIN: This method combines autocorrelation and AMDF techniques. Several stages of further processing are applied which reduce the effect commonly found autocorrelation techniques such as estimates at half the true f0 [7].
- ETSI: The ETSI Extended Front End (XFE) f0 estimator [1] uses both frequency-domain spectral peaks analysis and time-domain correlation scores to compute f0 estimates using a heuristic decision tree.

2.2. Formant Estimation Methods

- ESPS: The ESPS 'formant' tool estimates formant frequencies by solving for the roots of a 12^{th} order linear predictor polynomial. Dynamic programming is used to find the optimal formant tracks. For a given frame, all mappings of the complex roots to the estimated formant frequencies for the previous frame are calculated and a cost, based on formant frequencies and bandwidths, is obtained. The optimum formant track is given by the path with the lowest cost. The tool is available as part of the Snack Toolkit.

[1] http://www.speech.kth.se/snack/

[2] http://www.phon.ucl.ac.uk/resource/sfs/

- LPC: The second time-domain method uses 10^{th} order LPC analysis followed by Kalman filtering as described in [8].
- MFCC: This method provides a way of estimating formants from MFCCs and provides an alternative to predicting formants from MFCCs, described in section 3. It is possible to estimate formants from a magnitude spectrum found by transforming a MFCC vector to a spectral magnitude representation, even though magnitude spectra obtained from MFCC vectors are spectrally smoothed [3]. 10^{th} order LPC analysis is performed on magnitude spectra to provide initial formant candidates. Estimates are the candidates with the four smallest bandwidths. A five point median filter is used to smooth each formant track by removing discontinuities.

3. Prediction of Acoustic Speech Features from MFCC Vectors

Regression analysis of MFCCs and formants and f0 shows that correlations exist between MFCCs and acoustic speech features [9]. For example, mean formant frequency correlations for voiced speech were found to be 0.7. Acoustic feature prediction from MFCCs exploits these correlations through the modelling of the joint density of acoustic features and MFCCs [3].

Acoustic feature prediction from MFCCs comprises two parts. First, three Gaussian mixture models (GMMs) are created to model the joint density of acoustic features and MFCCs for non-speech, unvoiced, and voiced speech. Second, for predicting acoustic features from a stream of MFCC vectors, a voicing decision is made for each vector using prior and posterior voicing probabilities. According to the predicted voicing class (voiced, unvoiced or non-speech), acoustic features are predicted from the appropriate GMM using maximum a posteriori (MAP) prediction. Each acoustic feature track is smoothed using a five point median filter.

4. Evaluation of Prediction Against Hand-Corrected References

The use of a large database of hand-corrected formants enables a comparable evaluation of predicted and estimated formants. Fundamental frequency predictions and estimations are also evaluated using hand-corrected references. This section describes the databases used, the evaluation measures employed and presents the results.

4.1. Databases

Two databases of hand-corrected data are used to compare acoustic feature estimation and prediction. The VTR (vocal tract resonance) database comprises hand-corrected vocal tract resonance information for a subset of the TIMIT database [10] and is used here to evaluate formant frequency prediction. TIMIT is a large vocabulary database containing speech from eight US dialects. For this work the audio data were downsampled to 8 kHz. Although the VTR database provides the frequencies and bandwidths of the first four formants for 512 utterances, only the frequencies of the lowest three formants are hand-corrected. The database is split into testing (192 utterances from 24 male and female speakers) and training (324 utterances from 173 different speakers).

Fundamental frequency prediction is evaluated using reference data from a female US English speaker for which laryngograph-based f0 tracks were hand-corrected. The speech was downsampled to 8 kHz and the data split into training and testing sets comprising 579 and 246 sentences, respectively.

To evaluate the effectiveness of the estimation and prediction methods in noisy speech, 'exhibition hall' noise, extracted from the ETSI Aurora database, was added to the clean speech of both databases at SNRs of 20 dB to −5 dB in 5 dB steps.

4.2. Evaluation Measures

Initial experiments use two simple error measures: voicing class error and percentage f0 and formant frequency error. The voicing class error is defined as the proportion of non-voiced frames wrongly estimated or predicted as voiced and voiced frames wrongly classified as non-voiced:

$$E^v = \frac{N_{\text{v|nv}} + N_{\text{nv|v}}}{N_{total}} \times 100\% \quad (1)$$

where $N_{\text{v|nv}}$ is the number of non-voiced frames incorrectly classified as voiced, $N_{\text{nv|v}}$ is the number of voiced frames incorrectly classified as non-voiced and N_{total} is the total number of frames in the test data. The percentage f0 error is defined as:

$$E^{f0} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\hat{f0}_i - f0_i}{f0_i}\right| \times 100\% \quad (2)$$

A mean formant frequency percentage error, E^F, is similarly calculated, but is the mean of all three formant frequencies.

4.3. Experimental Results

4.3.1. Fundamental Frequency

The aim of the experiments reported in this section is to compare the accuracy of f0 estimation and prediction when evaluated using hand-corrected references for the single female speaker.

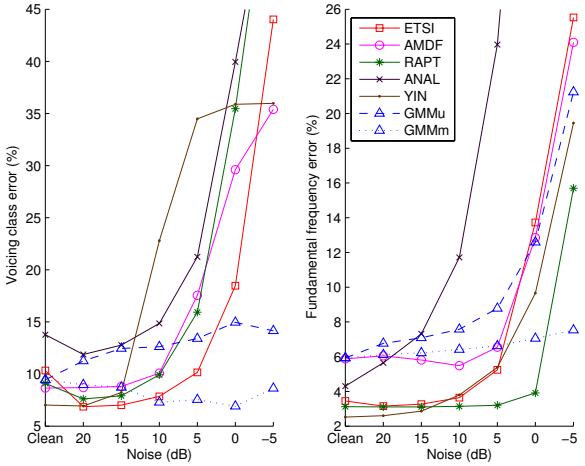

Figure 1: *a) Voicing class and b) f0 errors evaluated using hand-corrected references for one female speaker*

Figure 1a shows voicing class error (E^v) for the f0 estimation and prediction methods as noise increases. For GMM-based prediction, results are shown for both unmatched and matched condition testing. In unmatched conditions, predictions are made from models trained using clean speech. For matched condition testing, models are created using noisy training data with the same SNR as the testing data. Unmatched

and matched conditions provide the lower and upper bounds of GMM prediction performance accuracy.

The GMM prediction method provides good voicing class predictions in noisy speech because voicing is largely dependent on energy which is contained within MFCCs. The signal-processing based estimation techniques are less accurate and analysis shows that they tend to underestimate the number of voiced frames. Except for the ETSI estimator, the f0 estimators are not designed to provide voicing class decisions in noise. As the amount of noise increases, these estimators class fewer frames as voiced.

Figure 1b presents f0 estimation and prediction percentage errors (E^{f0}) as noise increases. At high SNRs the estimation methods provide more accurate f0 estimates than the prediction method. Accurate f0 prediction from MFCCs is more difficult because MFCCs contain little f0 information. As noise increases, the estimation methods deem fewer frames as voiced, so the percentage f0 errors become less reliable. Analysis shows that noise masks less harmonically well-defined speech and so f0 estimates are only made from frames with the greatest extent of harmonicity. The matched condition results for GMM prediction demonstrate the potential accuracy of f0 prediction in noise.

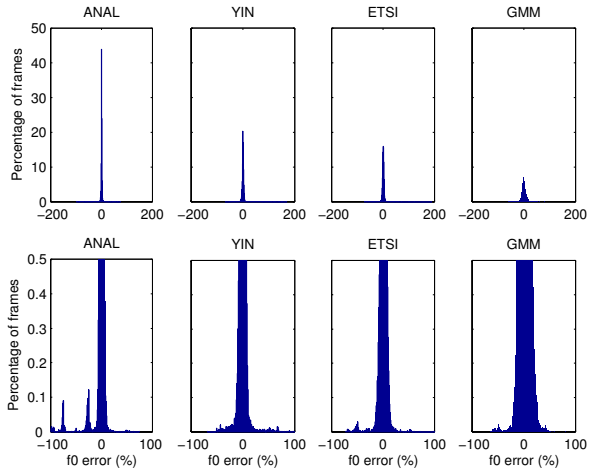

Figure 2: *f0 error distributions for single female clean speech*

An analysis of the distribution of f0 errors in clean speech is presented in figure 2 for three estimation methods (ANAL, YIN and ETSI) and the GMM prediction method. The lower plots show the lowest 0.5% of the histograms in more detail. For each distribution, the percentage f0 error is concentrated around 0%, although the distribution is noticeably wider for the GMM prediction method.

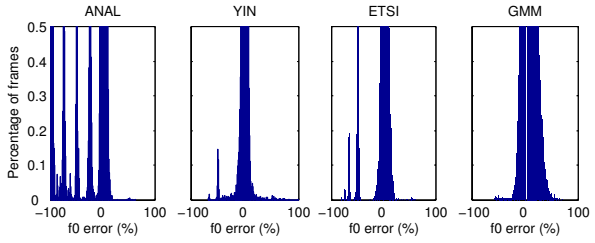

Figure 3: *f0 error distributions for single female speech (SNR 5 dB)*

Figure 3 shows the distributions of f0 errors in speech corrupted by noise at a SNR of 5 dB. The histograms are only shown from 0 to 0.5% and the GMM prediction results were obtained using unmatched condition testing. Despite the GMM prediction resulting in a higher percentage f0 error as shown in figure 1, these errors remain closely distributed around 0%. The estimation methods increase the number of larger errors which lead to more noticeable distortions in reconstructed speech.

4.3.2. Formant Frequencies

Figure 4 shows that the formant frequencies closest to the hand-corrected references are those obtained through GMM-based prediction from MFCCs. The distributions of the formant frequency estimation and prediction errors from clean speech are presented in figure 5. Compared with f0, formant frequency errors are more normally distributed, but cover a wider range. Similar histograms for formant frequency errors from noisy speech are shown in figure 6 for the lowest 0.2% of histogram bin populations. At 5 dB, the error distributions become broader.

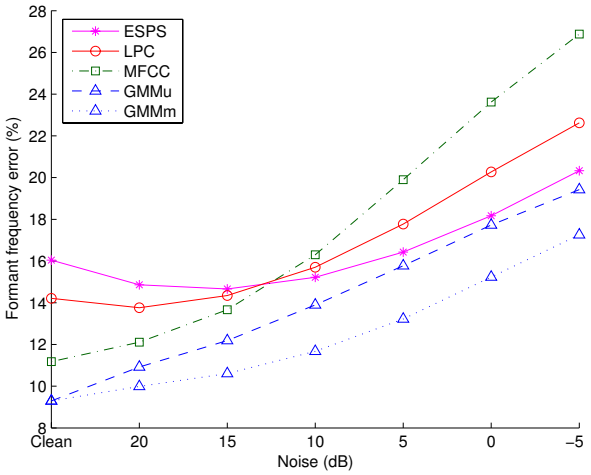

Figure 4: *Mean formant frequency errors evaluated using 24 male and female speakers*

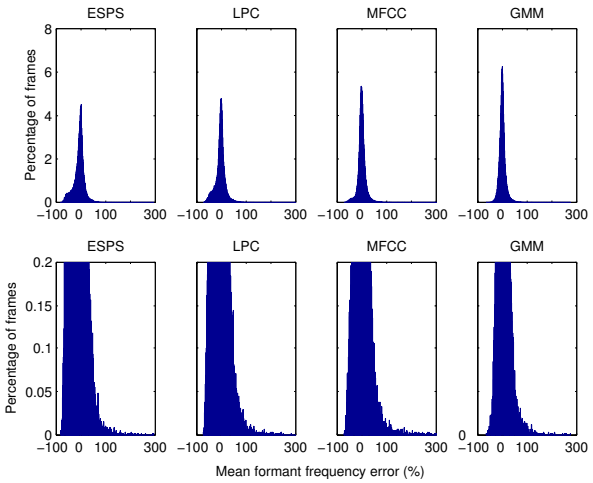

Figure 5: *Distributions of mean formant errors using 24 male and female speakers (clean speech)*

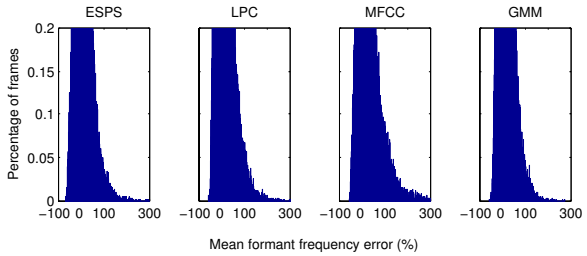

Figure 6: *Distributions of mean formant errors using 24 male and female speakers (SNR 5 dB)*

5. Reconstruction of Speech from Acoustic Speech Features

One application of the method of predicting acoustic speech features from MFCC vectors is the reconstruction of speech within a DSR environment without transmitting voicing and f0. This section examines how f0 and formant frequencies and bandwidths can be used to reconstruct speech. Comparisons are made between reconstructing speech using reference, estimated and predicted f0 and formants through the use of PESQ mean opinion scores (MOS).

In this work, speech is reconstructed using the sinusoidal model of speech similar to that found in the ETSI XFE standard [1]. The spectral envelope may be provided from the MFCC vector, or from formants. The cascade implementation of the Klatt formant synthesizer [11] is used to generate spectral envelopes from formant frequencies and bandwidths. The GMM-based prediction technique described in section 3 is extended to also predict formant bandwidths.

Experiments aim to determine the effect of f0 and formant accuracy on speech reconstruction. Table 1 shows how f0 accuracy affects reconstruction for the single female speaker. Here the ETSI XFE method is used which estimates the spectral envelope from MFCCs. In order to evaluate the different methods of obtaining f0, clean speech MFCCs are used, whilst the SNR of the speech used to obtain f0 is varied. Despite not having the lowest percentage f0 error, the GMM prediction method results in intelligible speech even at 10 dB and 0 dB.

Method	Clean	20 dB	10 dB	0 dB
Hand-corrected	1.8686	–	–	–
AMDF	1.8884	1.9373	1.8454	1.8487
ANAL	1.8678	1.8290	1.8371	1.8167
RAPT	1.9135	1.8760	1.8863	1.7962
YIN	1.8632	1.9317	1.8467	1.7814
ETSI	1.7650	1.8692	1.7869	1.9145
GMM (unmatched)	1.8393	1.8947	1.9209	1.8966

Table 1: *MOS by f0 method for single female speaker*

To determine how formant frequency and bandwidth accuracy affects speech reconstruction in noise, the spectral envelope is calculated from the reference, estimated and predicted formants. Clean f0 estimates are used to evaluate reconstruction to isolate the effect of formant accuracy independently. Despite producing the most accurate formant frequencies, the GMM method does not result in the best reconstruction because bandwidths are not predicted accurately compared with using estimation.

Method	Clean	20 dB	10 dB	0 dB
Hand-corrected	1.9911	–	–	–
ESPS	1.9187	1.9250	1.9319	1.9384
LPC	1.9232	1.9444	1.9403	1.9372
MFCC	1.9634	1.9376	1.9315	1.9381
GMM (unmatched)	1.5414	1.6318	1.6785	1.6490

Table 2: *MOS by formant estimation/prediction method for 24 male and female speakers*

6. Conclusions

Formants and f0 from estimation methods and GMM prediction from MFCCs have been evaluated using hand-corrected data. The GMM prediction method is more successful at predicting formants compared with f0, because MFCCs contain a representation of the spectral envelope, but are too coarse to contain f0 harmonics. There is little difference in the quality of reconstructed speech when using reference or predicted f0.

7. References

[1] ETSI, "Extended Front End Algorithm, Version 1.1.1," ETSI STQ-Aurora DSR Working Group, Tech. Rep. ES 202 211, Nov. 2003.

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

[3] J. Darch, B. Milner, and S. Vaseghi, "MAP prediction of formant frequencies and voicing class from MFCC vectors in noise," *Speech Communication*, vol. 48, no. 11, pp. 1556–1572, Nov. 2006.

[4] M. Ross, H. Shaffer, A. Cohen, R. Freudberg, and H. Manley, "Average magnitude difference function pitch extractor," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 22, no. 5, pp. 353–362, Oct. 1974.

[5] B. Secrest and G. Doddington, "An integrated pitch tracking algorithm for speech systems," in *ICASSP*, Boston, MA, Apr. 1983.

[6] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*, W. Kleijn and K. Paliwal, Eds. Elsevier, 1995, ch. 14, pp. 495–518.

[7] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *JASA*, vol. 111, no. 4, pp. 1917–1930, Apr. 2002.

[8] 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.

[9] J. Darch, B. Milner, I. Almajai, and S. Vaseghi, "An investigation into the correlation and prediction of acoustic speech features from MFCC vectors," in *ICASSP*, Honolulu, Hawaii, Apr. 2007.

[10] L. Deng, X. Cui, R. Pruvenok, J. Huang, S. Momen, Y. Chen, and A. Alwan, "A database of vocal tract resonance trajectories for research in speech processing," in *ICASSP*, Toulouse, France, May 2006.

[11] D. Klatt, "Software for a cascade/parallel formant synthesizer," *JASA*, vol. 67, no. 3, pp. 971–995, Mar. 1980.