Improved Phone Recognition on TIMIT using Formant Frequency Data and Confidence Measures

N.J. Wilkinson and M.J. Russell.

Electronic, Electrical & Computer Engineering, University of Birmingham, Birmingham B15 2TT, UK

Email: wilkinsn@eee.bham.ac.uk, m.j.russell@bham.ac.uk

ABSTRACT

This paper presents a novel approach to integration of formant frequency and conventional MFCC data in phone recognition experiments on TIMIT. Naive use of format data introduces classification errors if formant frequency estimates are poor, resulting in a net drop in performance. However, by exploiting a measure of confidence in the formant frequency estimates, formant data can contribute to classification in parts of a speech signal where it is reliable, and be replaced by conventional MFCC data when it is not. In this way an improvement of 4.7% is achieved. Moreover, by exploiting the relationship between formant frequencies and vocal tract geometry, simple formant-based vocal tract length normalisation reduces the error rate by 6.1% relative to a conventional representation alone.

1. INTRODUCTION

Formants are the resonance frequencies of the vocal tract. In general, different phones correspond to different vocal tract configurations and thus to different sets of formant frequencies. It is well known that formant frequencies are important for human vowel perception. Also, they provide a potentially useful representation for direct modelling of speech dynamics, due to their relationship with articulatory parameters. These arguments, plus the visual prominence of formants in regions of a spectrogram which correspond to voiced sounds, make a strong argument for the use of formant data as features for automatic speech recognition.

However, while conventional mel frequency cepstral coefficients (MFCCs) are obtained as a simple linear transformation of the short-term log power spectrum, estimation of formant frequencies is a significant pattern processing problem. Formants are most visible during voiced sounds, and are thus most prominent in vowels and glides. In some stops and fricatives they may be absent, and estimating a frequency value for the feature vector becomes difficult. A further problem ([1]) derives from the need to label the formants, rather than simply identify resonance frequencies. Often the higher formants are so close to each other that it is unclear whether one is looking at F2, F3 or a combination of both. This problem may be resolved by tracking the evolution of the spectrum in time.

Errors introduced during formant extraction will propagate into the classification process. Hence, if formant parameters simply replace more conventional parameters in a recognition system, an increase, rather than decrease, in error rate will result. However, use of formants can improve spoken digit recognition accuracy, provided that classification incorporates a measure of confidence in each formant frequency estimate [1]. This is developed further in [2], where a 1.2% reduction in error rate on the same task is achieved, compared with standard MFCCs.

This paper extends the techniques described in [1] and [2]. New methods for incorporating formant frequency estimate confidence measures into the recognition process are presented, which trade-off formant frequency estimates against conventional MFCC parameters. These methods allow the system to choose the most appropriate parameterisation depending upon the type of speech sound which is being hypothesised. A novel, formant-based approach to vocal tract shape normalisation is also presented. The results of phone recognition experiments on TIMIT show that these techniques can result in relative reductions in phone recognition error rate of up to 6%.

2. EXPERIMENTAL METHOD

2.1. Speech Data

All experiments use the TIMIT corpus, which comprises phone-labelled speech from 8 US dialect regions. The database is partitioned into disjoint training (462 speakers) and test sets (162 speakers). Only male speakers
were used in these experiments: 326 from the training set and 8 from the test. Each subject spoke 10 sentences, thus giving 3,260 training and 80 test sentences.

2.2. Speech Analysis

The TIMIT data was downsampled to 8kHz and mel-frequency cepstral coefficients 0 to 8 were computed across a 4 kHz bandwidth. A discrete Fourier transform was used (30ms Hamming window, 10ms overlap), the resulting complex spectrum converted into a log power spectrum, and mel scale filtering applied [3]. A cosine transform was then applied to generate cepstral coefficients. The first (\(\Delta\)) and second (\(\Delta^2\)) time-differences were calculated for each parameter, to give a 27 dimensional feature vector, corresponding to the ‘MFCC_0_D_A’ parameterisation from HTK [3]. This forms the ‘8MFCC baseline’ system.

Formant frequency estimation used the Holmes analyser [5]. The speech is transformed into the frequency domain and compared with 129 ‘template’ spectra, each with hand-labelled formant frequencies, using frequency warping. The formant frequencies for the closest template are adjusted according to the frequency warp and used as estimates of the current formant frequencies, \(f_1\), \(f_2\), and \(f_3\). In some cases an alternative set of formant frequencies is also produced, but these were not used in the current experiments. Confidence measures, \(conf_1\), \(conf_2\) and \(conf_3\), are defined as

\[
Conf_i = RSA*Curv
\]

In this equation, the relative spectral amplitude, \(RSA\), is calculated by comparing the amplitude of the closest frequency point to formant \(n\), to the largest amplitude in the local spectral slice. Thus a formant must have sufficient relative amplitude to achieve a good confidence score. This ensures that a “blip” in the spectrum will score low confidence. Curvature is defined by:

\[
Curv = 2S(f) - S(f-2) - S(f+2)
\]

where \(S(f)\) is the amplitude of the spectrum close to the formant. Curvature measures the prominence of a formant peak. It follows that the confidence is a good indicator of the presence or absents of a formant in the spectrum, rather than the accuracy of the frequency estimate.

2.3 Inclusion of formant parameters

As in [1] and [2], the three formant frequency estimates replaced or supplemented the top three mel-frequency cepstral coefficients, however, the use of the confidence weights is different. In ([1], [2]) a formant frequency estimate is treated as a noisy observation, whose variance is derived from the confidence measure. Experiments on TIMIT using this interpretation gave poor results. The derivation of the confidence does not measure the accuracy of the frequency of a formant but indicates whether it is present of not in the spectrum. This implies that the confidence should instead be used to weight the contribution of a formant to the feature vector. So the contribution \(p(f_i)\) to the acoustic vector probability due to a formant frequency estimate \(f_i\) was replaced by \(p(f_i)^{conf_i}\), where \(conf_i\) is scaled such that \(0 \leq conf_i \leq 1\). The following feature combinations were considered:

A. 5 MFCCs plus 3 formant frequencies replacing the top 3 MFCCs, no confidence measures.
B. 5 MFCCs plus 3 formant frequencies replacing the top 3 MFCCs. The contribution \(p(f_i)\) to the acoustic vector probability due to a formant frequency estimate \(f_i\) was replaced by \(p(f_i)^{conf_i}\)
C. Formant frequency estimates and confidence measures used as in B. Top 3 MFCCs re-introduce, with their contribution to the acoustic observation probability scaled by a factor equal to one minus the corresponding formant confidence measure.
D. Vocal tract normalization: Formant frequency estimates, plus confidence, plus MFCCs 6-8 used as in C. However, for each speaker and each formant, the mean formant value was subtracted from each formant frequency estimate, corresponding to a form of vocal tract length normalization.
E. Vocal tract normalization of mean and variance: Formant frequency estimates, confidences and MFCCs 6-8 used as in D. Relative formant frequencies normalized by the subject’s formant frequency variance.

Formally, acoustic vectors are made up as follows:

- Mel frequency cepstral coefficients \(c_0,...,c_7\).
- Estimates of the three formant frequencies \(f_1,f_2\) and \(f_3\).
- Confidence weights \(conf_1, conf_2\) and \(conf_3\).

In the baseline system (8MFCCs), an acoustic feature vector \(y=(y_0,...,y_26)\) comprises cepstral coefficients \(c_0,...,c_7\) plus the corresponding delta and acceleration coefficients:

\[
y_i = c_i, y_{i+9} = \Delta c_i, y_{i+18} = \Delta^2 c_i, y_{i+27}\ (i=0,...,8)
\]

\[
p(y) = \prod_{i=0}^{27} p(y_i)
\]

In (A), MFCCs 6, 7, 8, and their \(\Delta\) and \(\Delta^2\) parameters, are replaced by \(f_1\), \(f_2\) and \(f_3\) and their \(\Delta\) and \(\Delta^2\) parameters.

System B makes use of the formant frequency confidence measures. Relative to the previous system the necessary change is

\[
p(y) = \prod_{i=0}^{27} p(y_i)^{w_i}
\]

where, \(w_i = w_{i+9} = w_{i+18} = 1\) if \(0 \leq i \leq 5\) (i.e. \(i\) does not correspond to formant data) and \(w_i = conf_i\) if \(i=6,7,8\) (i.e \(i\) does correspond to formant data)

The problem with this use of confidence is that if the confidence is low, then classification relies only the
bottom 6 MFCCs. This is overcome in system C, where the probability calculations include all 9 cepstral coefficients plus the three formant frequency estimates. The contributions of the formant frequency probabilities are again regulated by the confidence weights. However, if confidence is less than 1 a scaled cepstral coefficient probability is re-introduced:

\[ p(y) = \prod_{i=0}^{5} q(c_i) \prod_{i=1}^{3} q(f_i) \prod_{i=6}^{8} q(c_i) \]

where \( q(x_i) = p(x_i) p(\Delta x_i) p(\Delta^2 x_i) \)

2.3. HMM, Training and Recognition

In all experiments a total of 634 clustered triphone and biphone HMMs were used to represent the TIMIT phone set. Each HMM had 3 emitting states, with 4 Gaussian mixture components.

The HMM parameters were initialized using the TIMIT phone-level annotation, followed by Viterbi alignment to improve the state-time correspondence. The Baum-Welch algorithm was then applied at the sentence level.

The test set comprised 1 male speaker from each dialect region. Each sentence was recognized at the phone level using Viterbi decoding. The phones in the recogniser output and in the corresponding transcription from the TIMIT corpus were then mapped onto the standard TIMIT 39 phone equivalence classes, and the system’s performance measured. A phone level-language model was used. The HTK toolkit was employed throughout [3].

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>A: MFCCs 0-5 + 3 Formant frequencies, no confidence.</td>
<td>56.09</td>
</tr>
<tr>
<td>B: MFCCs 0-5 + 3 Formant frequencies, with confidence</td>
<td>56.87</td>
</tr>
<tr>
<td>Baseline: MFCCs 0-8</td>
<td>59.73</td>
</tr>
<tr>
<td>C: Feature vector compensation.</td>
<td>60.77</td>
</tr>
<tr>
<td>D: Feature vector compensation and vocal tract normalization on the formants.</td>
<td>62.07</td>
</tr>
<tr>
<td>E: As above with vocal tract normalization of both mean and variance.</td>
<td>62.17</td>
</tr>
</tbody>
</table>

Table 1: Experimental results

The experiments conducted altered both the type and number of features in the feature vector. This has a marked effect on the log probabilities produced and so the optimal language scaling factor and insertion penalty must be re-computed for each configuration.

3. EXPERIMENTAL RESULTS

The first two rows of table 1 show that replacing the top 3 MFCCs with formants, with or without confidence, has an adverse effect on recognition accuracy. Figures 1 and 2 show the effect of the use of formants on accuracy for different phone classes. The improvement in vowel and glide recognition is expected, since these phones are voiced and formants are generally visible. The nasals, and particularly the stops, show worse performance. Figure 3 shows the distribution of confidence measures for stops, and indicates that confidence is generally low for this class of phone. Thus, for stop classification the formants are heavily de-emphasised and the recognizer relies, essentially, on MFCCs 0 to 5, reducing the number of features by 33% relative to the baseline. The nasals exhibit a similar but less severe effect.

![Figure 1: % accuracy for vowels, nasals and fricatives](image1)

![Figure 2: % accuracy for glides, stops and affricates](image2)
sounds have an unvoiced sound source with significant spectral energy between ~3Khz and ~6Khz. Since the speech is cut off at 4Khz some of this high frequency information is lost.

The distribution of confidence over all data (figure 5) shows that for about 50% of frames the confidence measure is below 0.5 and is of limited use. For approximately 40% of frames the formant data has either full or no confidence.

Normalization of the formant frequencies (scheme D) results in an improvement in recognition performance of 6% relative to baseline, due mainly to improvements in recognition of nasals and glides (figures 1 and 2).

In the final scheme (E) the variances of the formant frequency estimates are also normalized. This technique offers a small increase of 0.1% over scheme D. Figures 1 and 2 show that the vowels, fricatives, glides, stops and affricates are all improved, with only nasals being adversely effected. Note that only 11% of each feature vector underwent normalization in schemes D and E, and that this 11% is only used for approximately 50% of frames, since formant data is de-emphasised when confidence in it is low.

4. CONCLUSIONS

Careful incorporation of formant data has been shown to result in an improvement in TIMIT phone recognition accuracy of 6% relative to the baseline system, and 14% relative to naïve incorporation of formant data in scheme A.

5. REFERENCES