# USING FORMANT FREQUENCIES IN SPEECH RECOGNITION

*John N. Holmes (1), Wendy J. Holmes (2) and Philip N. Garner (2)*
(1) Speech Technology Consultant, 19 Maylands Drive, Uxbridge, UB8 1BH, U.K.
Tel: +44 1895 236328, E-mail: jnh@jnholmes.demon.co.uk
(2) Speech Research Unit, DRA Malvern, St. Andrews Road, Malvern, Worcs.,WR14 3PS, U.K.
Tel: +44 1684 894104/894157, E-mail: holmes/garner@signal.dra.hmg.gb

## ABSTRACT

Formant frequencies have rarely been used as acoustic features for speech recognition, in spite of their phonetic significance. For some speech sounds one or more of the formants may be so badly defined that it is not useful to attempt a frequency measurement. Also, it is often difficult to decide which formant labels to attach to particular spectral peaks. This paper describes a new method of formant analysis which includes techniques to overcome both of the above difficulties. Using the same data and HMM model structure, results are compared between a recognizer using conventional cepstrum features and one using three formant frequencies, combined with fewer cepstrum features to represent general spectral trends. For the same total number of features, results show that including formant features can offer increased accuracy over using cepstrum features only.

## 1. INTRODUCTION

It has been known for many years that formant frequencies are important in determining the phonetic content of speech sounds. Several authors have therefore investigated formant frequencies as speech recognition features, using various methods for basic analysis, such as linear prediction [1], [2], analysis by synthesis with Fourier spectra [3], and peak picking on cepstrally smoothed spectra [4]. However, using formants for recognition can sometimes cause problems, and they have not yet been widely adopted. It is obvious, for example, that formant frequencies cannot discriminate between speech sounds for which the main differences are unrelated to formants. Thus they are unable to distinguish between speech and silence or between vowels and weak fricatives. Whenever any formants are poorly defined in the signal (e.g. in fricatives), measurements will be unreliable, and it is therefore essential that their estimated frequencies should be given little weight in the recognition process.

To be useful as features for automatic speech recognition, formant frequencies must be supplemented by signal level and general spectral shape information, such as provided by low-order cepstrum features, for example. However, whenever the speech spectrum has a peaky structure, the phonetic detail is better described by formant frequencies than by the more usual higher-order cepstrum features, which have no simple relationship with formant frequencies.

It is impossible to determine from the spectrum of some speech sounds whether a particular peak should be associated with one formant or with a pair, and sometimes a formant may be so weak as a consequence of weak excitation that it causes no peak in the spectrum. Either of these situations can cause all higher-frequency formants to be wrongly labelled, with disastrous effects on the recognition. In such cases alternative labellings must be produced, and any uncertainties that cannot be resolved in other ways must be resolved within the recognition algorithm. The decisions are thus delayed until the words have been recognized [1]. However, many labelling uncertainties of single frames can be safely resolved merely by applying formant continuity constraints [2], which are a general property of speech. First applying continuity constraints is actually better for the standard HMM formalism, which does not exploit continuity of features.

This paper presents a new method of formant analysis which has provision for dealing with ambiguous labelling and with indistinct formants. The method has been used to supplement low-order cepstrum features for speech recognition.

## 2. NEW METHOD FOR FORMANT ANALYSIS

### 2.1 Human interpretation of formants

When supplied with a wide-band spectrogram of a speech signal, an expert in experimental phonetics can usually estimate fairly well where the formant trajectories are for all parts of the signal for which such an interpretation would be useful. For those parts of the signal where the formant peaks of a particular spectral cross-section are not well defined, an expert can normally still make a reasonable interpretation by using phonetic knowledge about the normal properties of speech sounds and by interpolation between neighbouring sounds for which the formant structure is clearer. It is generally more difficult to estimate formant frequencies automatically, given the same short-term spectral analysis that is the basis of spectrographic display. However, the task is easy if the spectral cross-section of the signal has a small number of clearly defined peaks. Provided that each of the three lowest-frequency peaks is in the frequency range typical of one of the three lowest formants, only one sensible formant interpretation of the spectral shape is possible.

Fig. 1 shows a spectral cross-section which has clear peaks, with the positions of the formants marked. On these occasions a single spectral cross section is all that is required to make a reliable estimate. Sometimes, however, two formants may be so close in frequency that they give rise to only a single spectral peak. There can also often be occasions where a total of three spectral peaks are visible, but the frequencies and intensities might be such that the middle peak could plausibly be F2 by itself and the third peak be F3, or the middle peak could be F2 and F3 together, with the third peak being F4. In this case even a human expert would be incapable of making a reliable choice,

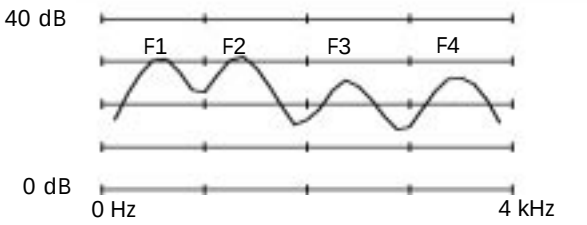


Fig. 1. Spectrum with clear formants

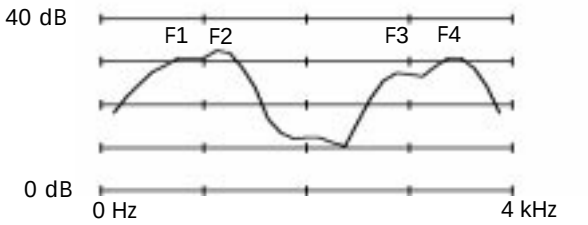


Fig. 2. F1 and F2 in a single spectral peak

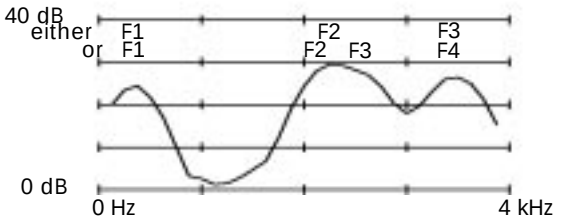


Fig. 3. Ambiguous formant labelling

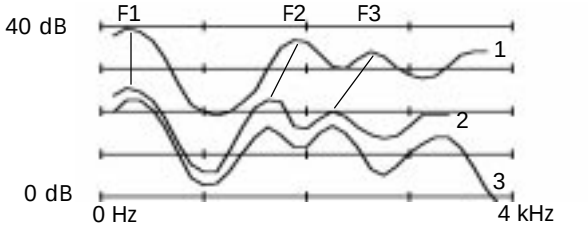


Fig. 4. Frequency warping of pattern (trace 1) into warped pattern (trace 2) to align with input (trace 3)

given only a single spectral cross-section. However, the expert would be able to postulate a small number of plausible alternatives, where in most cases all but one of these alternatives could subsequently be rejected by using continuity constraints. Thus unambiguous formant trajectories would be obtained for a substantial proportion of any utterance. Fig. 2 shows a spectral cross-section for which F1 and F2 are obviously both associated with the lowest-frequency peak, whereas the spectrum shown in Fig. 3 is an example where there is uncertainty about the correct formant labelling, and both of the marked formant allocations would be plausible.

An important novelty of the formant estimation method described in this paper is that it exploits this human ability to apply formant labels to spectral cross-sections, giving alternative formant allocations to peaks where appropriate.

**2.2 Preliminary formant estimates**

The formant analysis uses log power spectra derived from 64-point FFTs of a signal sampled at 8 kHz. To ensure that the cross-sections represent the formants as well as possible, the FFTs are taken from regions immediately after points of excitation of the vocal tract, selected on the basis of a local power maximum. There is a store of about 150 typical spectral cross-sections, each of which is associated with one or more sets of plausible labellings of the lowest three formants, provided by a human expert. Each input spectral cross-section is first compared with all the stored patterns, to select a few which have the most similar general spectral shape. These few patterns are then compared with the input using a dynamic programming (DP) technique in the frequency domain to find the frequency scale warping of the stored patterns which gives the best match to the input. Fig. 4 illustrates a typical warping operation. The DP cost function includes components dependent on spectral level, spectral slope and extent of frequency warping. The pattern with the best DP score and any close competitors are selected for further consideration. The frequency warping of each such pattern is applied to the formant frequencies stored with the pattern, to give preliminary formant frequency estimates. These estimates are quantized at the 125 Hz spacing of the FFT, and more finely quantized formant frequencies are derived by matching typical formant shapes to the spectrum in the region of the chosen FFT points.

**2.3 Selection of smooth formant tracks**

Any alternative formant labellings given by the few best-fitting patterns are used as input to an additional DP process, which finds the best smooth trajectories through the available formant frequency candidates. A second pass of the DP smoothing process is then made, in which the best formant labelling given by the first pass is used as an additional input to the DP cost function. This second pass will give an alternative smooth path through the available formant candidates if the score for such a path is not much worse than the score of the best path.

The formant analysis method usually gives a unique formant interpretation of speech signals, and never gives more than two different interpretations. Whenever it is apparent from a spectrogram where the formants should be, it is extremely rare for the algorithm to fail to give the correct values, and they are nearly always provided by the first choice. For each output formant frequency an estimate of confidence in the measurement is derived based on spectral level and spectral curvature, so that less reliable formant frequencies can be given less weight in recognition decisions.

**2.4 Analysis example**

Fig. 5 shows a typical spectrogram with superimposed formant tracks. During the [ʃ] and the [t] burst F1 has been omitted because there was no confidence in its accuracy. The two alternative interpretations of F2 and F3 are both reasonable, but the first choice obviously provides correct continuity into the nearby phones. Neither F2 nor F3 could be usefully estimated during the [d] closure, and F2 in the [n] was only given any confidence for one frame. The first choice is clearly correct during the first part of the [e**I**]

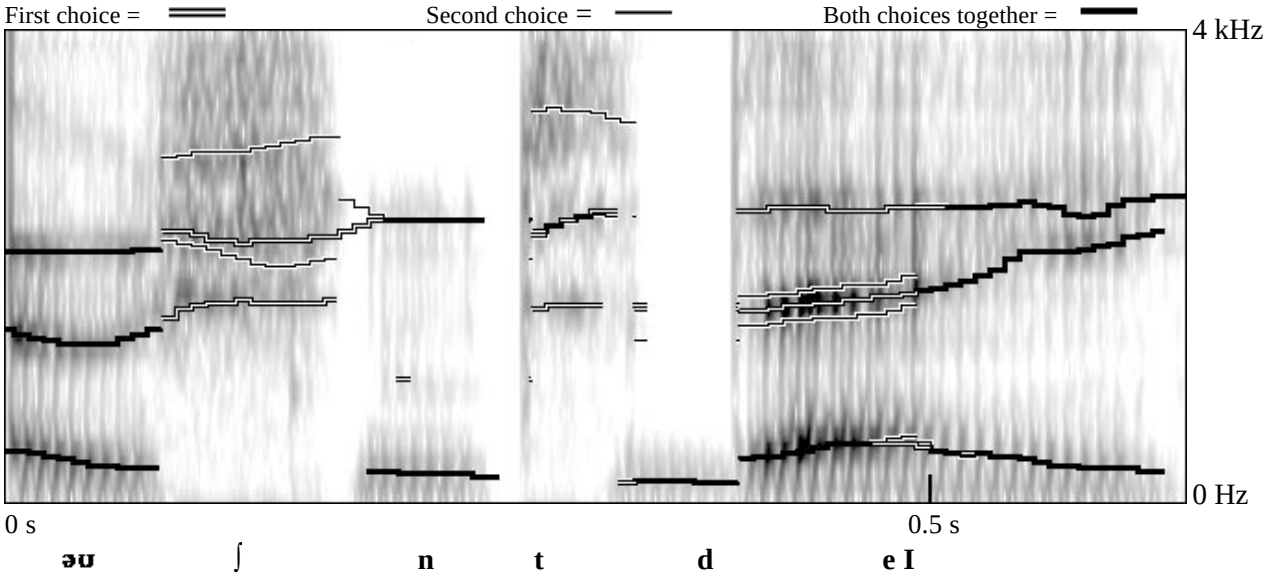


Fig. 5. Spectrogram of the words "ocean today", with superimposed formant tracks. Tracks are not plotted when there is no confidence in their accuracy.

diphthong, but the second choice was initially a plausible interpretation, until the later part of the diphthong had been analysed to reveal the first-choice F2 moving close to F3.

## 3. USING CONFIDENCE ESTIMATES AND AMBIGUITY IN RECOGNITION

Alternative formant sets arising from labelling ambiguity have so far been accommodated in recognition just by choosing the formant set which gives the highest HMM emission probability for each frame and model state.

During silence or background noise, and whenever there is no obvious spectral peak near to the estimated formant frequency, there will be no confidence in the formant frequency estimate, which should not then be used at all in the recognition. In this case, the appropriate formant information to use in the recognizer should be specified by prior information about its likely position. During peaky vowel spectra on the other hand, the measured frequencies will be given high confidence, although there may be occasional labelling ambiguity. There is a continuum of possibilities between these two extremes that can most suitably be accommodated by regarding the uncertainty of formant position as the variance of a notional Gaussian distribution of the true frequency about the estimated value.

The probabilistic interpretation leads naturally to the incorporation of prior knowledge about formant positions when the confidence is low. This prior knowledge is used by shifting the mean of the formant distribution away from the measured value, towards some suitable prior value for that formant. A heuristic procedure has been devised for using the estimated confidence computed from the spectrum to derive a formant measurement standard deviation and bias towards a prior distribution, both expressed in Hz. Although this process is *ad hoc*, it has been found to give plausible values and experimentation has shown that the precise values are not critical to recognition performance.

Assuming that variances are associated with all formant measurements, the HMM emission probability calculation needs to be modified to allow for a continuum of possible variance values for each formant. It can be shown that in the case of Gaussian models this modification corresponds to a convolution of the formant and model distributions, so that the variances simply add. The use of variance thus provides a sound theoretical framework to represent confidence associated with formant estimates, which is an improvement over an earlier version [5] of the formant-based recognizer, whereby the confidence was simply used as a weight to multiply log probabilities.

## 4. EXPERIMENTS

The aim was to compare recognition results using formant features for describing fine spectral detail with those obtained using a more conventional mel-cepstrum representation. In order to directly assess the usefulness of the formants, the same total number of features was used for both representations, and exactly the same low-order cepstrum features were used for describing general spectral shape. Thus the only difference was in the use of formants versus higher cepstral coefficients for representing detailed spectrum shape. The experiments were performed for the simple task of connected-digit recognition. While the details of the front-end processing and the modelling task have not been optimized to maximize performance, the system provides a good basis for comparative experiments.

### 4.1 Experimental set-up

The test data were four lists of 50 digit triples spoken by each of 10 male speakers. The training data were from 225 different male speakers, each reading 19 four-digit strings taken from a vocabulary of 10 strings. The output of the FFT was used both to estimate formant frequencies with associated confidence measures and to compute the mel-cepstrum. Experiments were then carried out to compare a representation using the first eight cepstrum coefficients and an overall energy feature, with a feature set in which cepstrum coefficients 6, 7 and 8 were replaced by the three formant features. To provide a basis for comparison, an experiment was also carried out using a representation

| Experimental condition | % Correct | % Subs. | % Del. | % Ins. | % Error |
|---|---|---|---|---|---|
| 5 cepstrum features + energy | 95.5 | 3.5 | 1.0 | 0.3 | 4.8 |
| 8 cepstrum features + energy | 96.0 | 3.0 | 1.0 | 0.3 | 4.3 |
| 5 cepstrum features + energy + 3 formants | 94.0 | 4.8 | 1.2 | 11.6 | 17.6 |
| Include confidence measure with formants | 96.9 | 2.3 | 0.8 | 0.3 | 3.4 |
| Also include second choice formants | 97.1 | 2.2 | 0.7 | 0.3 | 3.2 |

Table 1. Connected-digit recognition performance for front-end representations using only cepstrum features compared with a representation with the higher-order cepstral coefficients replaced by formant features.

which simply omitted cepstrum coefficients 6, 7 and 8, so using a total of only six features.

In all cases, three-state context-independent monophone models and four single-state non-speech models were used, all with single-Gaussian pdfs and diagonal covariance matrices. The model structure was a simple left-to-right one which included self-loop transitions. Model means were initialized from a very small quantity of hand-annotated training data (twelve digits from each of two speakers), with all model variances initialized to the same arbitrary value. All model parameters were trained with ten iterations of Baum-Welch re-estimation. During training, an appropriate lower limit was imposed on all the model variance parameters, to prevent them training to unrealistically low values which could prevent generalisation to the test data.

### 4.2 Treatment of formant features

As a pre-processing stage for both training and recognition, each observed formant value was moved towards its prior by an amount determined by the observation's confidence measure. The result of this stage was that high-confidence formant values were unchanged but, as the confidence decreased, the formant was moved further towards its prior. When there was no confidence, the prior value was used.

The main benefit of the confidence measure and multiple formant hypotheses was expected to be in the recognition stage, as the training process is much more constrained. Therefore, in training, the second choice formant values have not yet been used and no further use has so far been made of the confidence measure. Both were optionally included in the recognition phase, as described in Section 3.

### 4.3 Results and discussion

The results given in Table 1 show that, provided the degree of reliability in the formant estimation is taken into account, recognition performance is better when using formant features than when using only mel-cepstrum features. When compared with the results using just six cepstrum features, the benefit from adding the three formant features is three times greater than that obtained by adding the three additional cepstrum features.

When alternative formant sets were also included, there was a further small improvement in performance. Only a small improvement was expected because the first-choice values given by this algorithm are usually the correct ones. When they are correct, allowing the second choice could only increase recognition errors. It is therefore clearly desirable to find some way of using an estimate of the relative probabilities of correctness of the first and second choice in the recognition, and this will be included in future research.

The recognition results demonstrate the importance of using formant measurement accuracy in order to obtain good recognition performance. When the formant features were not given special treatment, there were significant problems with insertion errors. These errors were caused by mismatches between the formant frequencies in the non-speech models with those measured for the non-speech regions of the test data. A simple word-insertion penalty did not reduce these errors, but they disappeared when the formant confidence measure was incorporated.

## 5. CONCLUSIONS

These simple experiments have already demonstrated that a recognition system using formant features can provide better performance than one using mel-cepstrum features alone, for the same total number of features. We now need to confirm that similar benefits are obtained on a more challenging task with a larger database. The next stage of algorithm development is to incorporate both the variance representing confidence in formant measurement and the multiple formant hypotheses in an extended Baum-Welch re-estimation process. It is also possible to incorporate the shift of uncertain formant measurements towards their priors within the probabilistic formalism itself, in place of the heuristic approach used here.

Other issues to investigate include the use of time derivative features, which ought to be more valuable for smoothly-changing formants than for high order cepstrum features, particularly because formant transitions are known to be important cues for place of articulation of consonants.

## 6. REFERENCES


[1] M.J. Hunt, *"Delayed Decisions in Speech Recognition - The Case of Formants"*, Pattern Recognition Letters, Vol. 6, pp. 121-137, July 1987.

[2] P. Schmid and E. Barnard, *"Robust, N-Best Formant Tracking"*, Proc. EUROSPEECH'95, pp. 737-740, Madrid, 1995.

[3] L. Welling and H. Ney, *"A Model for Efficient Formant Estimation"*, Proc. IEEE ICASSP, pp. 797-800, Atlanta, 1996.

[4] Y. Laprie and M.-O. Berger, *"Active Models for Regularizing Formant Trajectories"*, Proc. ICSLP, pp. 815-818, Banff, 1992.

[5] J.N. Holmes and W.J. Holmes, *"The Use of Formants as Acoustic Features for Automatic Speech Recognition"*, Proc. IOA, Vol. 18, part 9, pp. 275-282, Nov. 1996.