ON THE ROBUST INCORPORATION OF FORMANT FEATURES INTO HIDDEN MARKOV MODELS FOR AUTOMATIC SPEECH RECOGNITION

Philip N. Garner, Wendy J. Holmes

Defence Evaluation and Research Agency,
St. Andrews Rd, Malvern, WORCS. WR14 3PS, UK

ABSTRACT

A formant analyser is interpreted probabilistically via a noisy channel model. This leads to a robust method of incorporating formant features into hidden Markov models for automatic speech recognition. Recognition equations follow trivially, and Baum-Welch style re-estimation equations are derived. Experimental results are presented which provide empirical proof of convergence, and demonstrate the effectiveness of the technique in achieving recognition performance advantages by including formant features rather than only using cepstrum features.

1. INTRODUCTION

Formant frequencies are known to be important in determining the phonetic content of speech sounds. Formants, however, are not generally used as features for automatic speech recognition as they may be ambiguous or badly defined and do not provide the necessary information for making certain distinctions (such as identifying silence). A new method of formant analysis has recently been presented [1] which includes techniques to overcome the difficulties normally associated with extracting and using formant information. Firstly, in cases of ambiguity, alternative sets of formant frequencies are offered to the recognition process. Secondly, a novel feature of the new formant analyser is that each formant frequency estimate is assigned a measure of confidence. The confidence measure is important because it allows for cases where formants are poorly defined in the signal (e.g. fricatives) so that any single estimate of frequency is likely to be unreliable. In such cases, it is essential that the estimated frequencies are given little weight in the recognition process, and that the recognition decision is based on signal level and general spectral shape information.

Whilst it is clear that the confidence measures have implications when the formants are used as features in speech recognition, it is not obvious how to include such measures in, for instance, an HMM based system. In this paper, we present a method for interpreting confidence estimates which can then be rigorously incorporated into a probabilistic model.

2. INTERPRETATION OF THE CONFIDENCE MEASURE

The formant analyser produces a confidence value for each formant for each time frame. This value represents and estimate of the confidence in the accuracy of the formant frequency measurement, and is derived automatically based on spectral level and curvature. The confidence values are represented as standard deviations which, when squared, can be thought of as variances of normal distributions centred upon the formant estimates. Interpreted in this way, the formant analyser emits the parameters of a normal distribution representing its belief about the position of each formant. When the confidence is high, the variance is low, representing strong belief in the estimate, and weak belief outside it. At the other extreme, a low confidence represents a high variance representing almost equal belief in all possible frequencies. This belief oriented interpretation is necessarily Bayesian.

3. MATHEMATICAL FORMULATION

3.1. Recognition

In conventional hidden Markov modelling, a state is assumed to emit an observation, $\boldsymbol{y}_t$, according to some output distribution. In this paper, we will assume that the output probability distribution for state j is a single multivariate normal with mean $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. The required probability at time t is

$$\Pr\left(\boldsymbol{y}_t \,\middle|\, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right) = \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}).$$

With the formant analyser, the observation comprises both a formant vector, $\boldsymbol{f}_t$, and a confidence vector, $\boldsymbol{c}_t$. The actual feature vector, being the real values of the formant frequencies, is unknown. The confidence measure of the formant analyser is assumed here to take the form of variance.

Given that we observe a distribution, the required expression for the output probability of the state is now

$$\Pr\left(\boldsymbol{f}_t, \boldsymbol{C}_t \,\middle|\, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right),$$

where $\boldsymbol{C}_t$ is the (diagonal) matrix of formant variances. The most informative way to proceed is to expand this expression thus

$$\Pr\left(\boldsymbol{f}_t, \boldsymbol{C}_t \,\middle|\, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right) = \Pr\left(\boldsymbol{f}_t \,\middle|\, \boldsymbol{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right) \Pr\left(\boldsymbol{C}_t \,\middle|\, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right),$$

and then to make the assumption that the confidence measure produced by the formant analyser is a reliable estimate, hence $\Pr\left(\boldsymbol{C}_t \,\middle|\, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right) = \Pr\left(\boldsymbol{C}_t\right) = 1$, since the confidence

measure is clearly independent of the output distribution parameters.

It can now be argued that the model proposed so far is mathematically the same as a noisy channel model, and that it in practice it is easier to think of it in these terms. The state output distribution emits a value, $\boldsymbol{y}_t$, which then passes through a noisy channel with zero mean and covariance $\boldsymbol{C}_t$; $\boldsymbol{f}_t$ is then the noisy observation. The expression of interest is clearly

$$\Pr\left(\boldsymbol{f}_t \,\middle|\, \boldsymbol{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right),$$

which is the same as before, but without the prior on $\boldsymbol{C}_t$.

To evaluate this expression, we must acknowledge that the measured vector, $\boldsymbol{f}_t$, depends upon the unknown output vector $\boldsymbol{y}_t$, and this vector must be integrated out:

$$\Pr\left(\boldsymbol{f}_t \,\middle|\, \boldsymbol{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right) = \int_{\Re^n} d\boldsymbol{y}_t \Pr\left(\boldsymbol{f}_t \,\middle|\, \boldsymbol{y}_t, \boldsymbol{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right) \Pr\left(\boldsymbol{y}_t \,\middle|\, \boldsymbol{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right),$$

where $\Re^n$ denotes the n-dimensional Euclidean space of possible observations. Observing some obvious independencies and substituting normal distributions,

$$\Pr\left(\boldsymbol{f}_t \,\middle|\, \boldsymbol{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right) = \int_{\Re^n} d\boldsymbol{y}_t \mathcal{N}(\boldsymbol{f}_t; \boldsymbol{y}_t, \boldsymbol{C}_t)\mathcal{N}(\boldsymbol{y}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}).$$

This form is very intuitive, it just states that the output probability should be evaluated for all possible values of the feature vector, weighted by the formant analyser's belief of each value. Given that $\mathcal{N}(\boldsymbol{f}_t; \boldsymbol{y}_t, \boldsymbol{C}_t) \equiv \mathcal{N}(\boldsymbol{y}_t; \boldsymbol{f}_t, \boldsymbol{C}_t)$, the integral is the convolution of two normal distributions. It can be shown that the variances simply add, the result being

$$\Pr\left(\boldsymbol{f}_t \,\middle|\, \boldsymbol{C}_t, \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t}\right) = \mathcal{N}(\boldsymbol{f}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t} + \boldsymbol{C}_t).$$

So, to incorporate the formant variances in recognition, we simply add the appropriate confidence variance to that of the output distribution. This result is intuitively pleasing: For high confidence (low variance), the usual expression applies, and for low confidence the output distribution widens to equally favour all output values.

3.2. Re-estimation

The re-estimation problem is to find a set of parameters $\boldsymbol{\lambda}$ which maximises the likelihood $\Pr(\boldsymbol{O} \,|\, \boldsymbol{\lambda})$ of an observation sequence $\boldsymbol{O} = \boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_T$. $\boldsymbol{\lambda}$ consists of an $S \times S$ transition probability matrix $\boldsymbol{A}$, and means and covariance matrices $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$, where $i = 1, \ldots, S$. Substituting the pair $\{\boldsymbol{f}_t, \boldsymbol{C}_t\}$ for $\boldsymbol{o}_t$, the probability of the observation is

$$\Pr(\boldsymbol{O} \,|\, \boldsymbol{\lambda}) = \sum_{\boldsymbol{s}} a_{s_0} \prod_{t=1}^{T} a_{s_{t-1}s_t} \mathcal{N}\left(\boldsymbol{f}_t; \boldsymbol{\mu}_{s_t}, \boldsymbol{\Sigma}_{s_t} + \boldsymbol{C}_t\right).$$

Following Liporace's interpretation of Baum's method [2], we define an auxiliary function $Q(\boldsymbol{\lambda}, \overline{\boldsymbol{\lambda}})$:

$$Q(\boldsymbol{\lambda}, \overline{\boldsymbol{\lambda}}) = \sum_{\boldsymbol{s}} \Pr(\boldsymbol{O}, \boldsymbol{s} \,|\, \boldsymbol{\lambda}) \log \Pr\left(\boldsymbol{O}, \boldsymbol{s} \,\middle|\, \overline{\boldsymbol{\lambda}}\right),$$

which has the property that

$$Q(\boldsymbol{\lambda}, \overline{\boldsymbol{\lambda}}) > Q(\boldsymbol{\lambda}, \boldsymbol{\lambda}) \Rightarrow \Pr\left(\boldsymbol{O} \,\middle|\, \overline{\boldsymbol{\lambda}}\right) > \Pr(\boldsymbol{O} \,|\, \boldsymbol{\lambda}).$$

Expanding the $\overline{\boldsymbol{\lambda}}$ portion of Q and rearranging the final term to isolate the parameters to be re-estimated,

$$\begin{aligned} Q(\boldsymbol{\lambda}, \overline{\boldsymbol{\lambda}}) = & \\ & \sum_{\boldsymbol{s}} \Pr(\boldsymbol{O}, \boldsymbol{s} \,|\, \boldsymbol{\lambda}) \times \Bigg[\log \overline{a}_{s_0} \\ & + \sum_{t=1}^{T} \Bigg\{ \log \overline{a}_{s_{t-1}s_t} - \frac{n}{2} \log(2\pi) - \frac{1}{2} \log \left| \overline{\boldsymbol{\Sigma}}_{s_t} + \boldsymbol{C}_t \right| \\ & - \frac{1}{2} \left(\boldsymbol{f}_t - \overline{\boldsymbol{\mu}}_{s_t}\right)' \left(\overline{\boldsymbol{\Sigma}}_{s_t} + \boldsymbol{C}_t\right)^{-1} \left(\boldsymbol{f}_t - \overline{\boldsymbol{\mu}}_{s_t}\right) \Bigg\} \Bigg]. \end{aligned}$$

It is clear that the re-estimation equations for the transition probabilities will be unchanged from the standard ones. The means and covariances, however, are likely to be different. First the means: Given that

$$\frac{\partial}{\partial \boldsymbol{x}} (\boldsymbol{y} - \boldsymbol{x})' \boldsymbol{A} (\boldsymbol{y} - \boldsymbol{x}) = -\boldsymbol{A}(\boldsymbol{y} - \boldsymbol{x}) - \boldsymbol{A}'(\boldsymbol{y} - \boldsymbol{x})$$

for any general matrix $\boldsymbol{A}$, and the covariance term is symmetric,

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\lambda}, \overline{\boldsymbol{\lambda}})}{\partial \overline{\boldsymbol{\mu}}_j} = & \\ & - \sum_{\boldsymbol{s}} \Pr(\boldsymbol{O}, \boldsymbol{s} \,|\, \boldsymbol{\lambda}) \sum_{\{t:\, s_t = j\}} \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \left(\boldsymbol{f}_t - \overline{\boldsymbol{\mu}}_j\right), \end{aligned}$$

Interchanging the order of summation and equating to zero,

$$\sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t = j\}} \Pr(\boldsymbol{O}, \boldsymbol{s} \,|\, \boldsymbol{\lambda}) \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \left(\boldsymbol{f}_t - \overline{\boldsymbol{\mu}}_j\right) = 0,$$

Following Liporace, we would be able to pre-multiply by the inverse of the matrix term. Here, however, $\boldsymbol{C}_t$ is frame dependent and must remain. Rearranging yields the re-estimation formula for the mean:

$$\begin{aligned} \overline{\boldsymbol{\mu}}_j = & \left(\sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t = j\}} \Pr(\boldsymbol{O}, \boldsymbol{s} \,|\, \boldsymbol{\lambda}) \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \right)^{-1} \\ & \times \sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t = j\}} \Pr(\boldsymbol{O}, \boldsymbol{s} \,|\, \boldsymbol{\lambda}) \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \boldsymbol{f}_t. \end{aligned}$$

We assume that the current value of the covariance matrix can be used, instead of the re-estimate. We also note that, in the fully multivariate case, this expression requires a matrix inversion for each frame.

Now consider the covariance re-estimation. Liporace differentiates with respect to the inverse, but here it is more

convenient to use the matrix itself:

$$\frac{\partial Q(\boldsymbol{\lambda},\overline{\boldsymbol{\lambda}})}{\partial \overline{\boldsymbol{\Sigma}}_j} = -\frac{1}{2}\sum_{\boldsymbol{s}} \Pr(\boldsymbol{O},\boldsymbol{s}\,|\,\boldsymbol{\lambda}) \sum_{\{t:\, s_t=j\}} \left\{ \frac{\partial}{\partial \overline{\boldsymbol{\Sigma}}_j} \log|\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t| + \frac{\partial}{\partial \overline{\boldsymbol{\Sigma}}_j} \left(\boldsymbol{f}_t - \overline{\boldsymbol{\mu}}_j\right)' \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \left(\boldsymbol{f}_t - \overline{\boldsymbol{\mu}}_j\right) \right\}.$$

Taking each term separately,

$$\frac{\partial}{\partial \overline{\boldsymbol{\Sigma}}_j} \log|\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t| = \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1}.$$

Strictly, when differentiating with respect to a symmetric matrix, the off diagonal elements of the result should be doubled [4]. In this case, however, this result is to be combined with another where the same effect happens, and it is both consistent and more readable to 'ignore' this effect. Liporace's derivation omits this caveat, though his results are valid for the same reason. It can be shown with reference to [3] that

$$\frac{\partial}{\partial \boldsymbol{A}} \boldsymbol{x}'(\boldsymbol{A}+\boldsymbol{B})^{-1}\boldsymbol{x} = -\left[(\boldsymbol{A}+\boldsymbol{B})^{-1}\right]' \boldsymbol{x}\boldsymbol{x}' \left[(\boldsymbol{A}+\boldsymbol{B})^{-1}\right]'.$$

So, denoting $\boldsymbol{f}_t - \overline{\boldsymbol{\mu}}_j$ by $\boldsymbol{x}$, interchanging the order of summation as before and equating to zero yields

$$\sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t=j\}} \Pr(\boldsymbol{O},\boldsymbol{s}\,|\,\boldsymbol{\lambda}) \left[\left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} - \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \boldsymbol{x}\boldsymbol{x}' \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \right] = 0. \quad (1)$$

At this stage, it is clear that $\overline{\boldsymbol{\Sigma}}_j$ cannot be isolated, and it is necessary to make an approximation. Two alternative approximations are proposed, as described below.

3.2.1. Method 1

Equation 1 can also be written,

$$\sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t=j\}} \Pr(\boldsymbol{O},\boldsymbol{s}\,|\,\boldsymbol{\lambda}) \left[\left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t - \boldsymbol{x}\boldsymbol{x}'\right) \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \right] = 0.$$

We now assume that $\boldsymbol{C}_t$ is independent of time for a given state, that is, it can be assumed constant for the duration of the state. This approximation is not unreasonable because the confidence with which the formant frequencies are estimated will generally be similar for all the feature vectors corresponding to any one model state. The two inverse terms can now be brought outside the summation, and the expression can then be pre- and post-multiplied by the inverse of those terms leaving a single instantiation of $\overline{\boldsymbol{\Sigma}}_j$:

$$\sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t=j\}} \Pr(\boldsymbol{O},\boldsymbol{s}\,|\,\boldsymbol{\lambda}) \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t - \boldsymbol{x}\boldsymbol{x}'\right) = 0.$$

Rearranging,

$$\overline{\boldsymbol{\Sigma}}_j = \frac{\displaystyle\sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t=j\}} \Pr(\boldsymbol{O},\boldsymbol{s}\,|\,\boldsymbol{\lambda}) \left[\boldsymbol{x}\boldsymbol{x}' - \boldsymbol{C}_t\right]}{\displaystyle\sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t=j\}} \Pr(\boldsymbol{O},\boldsymbol{s}\,|\,\boldsymbol{\lambda})}.$$

This approximation appears to have a problem: Where $\boldsymbol{C}_t$ is large, the term in square brackets will not be positive definite, which is one of the conditions cited by Liporace for the re-estimation to be valid. A remedy is to simply ignore the contribution of frames for which this term is not positive definite, that is, the sum of the eigenvalues is not positive. The effect of this is that the system is not trained on low confidence frames, which is entirely reasonable. For states where one or more frame elements are always low confidence, we suggest that this will be true in recognition too, and hence the $\boldsymbol{C}_t$ term will dominate there also. In the particular case where the covariance is assumed diagonal, the individual elements of the term in square brackets can be handled individually.

3.2.2. Method 2

Starting again from equation 1, notice that the first term in the squares brackets can be written

$$\left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} = \overline{\boldsymbol{\Sigma}}_j^{-1} \left(\mathrm{I} + \boldsymbol{C}_t \overline{\boldsymbol{\Sigma}}_j^{-1}\right)^{-1}. \quad (2)$$

Hence, by substituting equation 2 into equation 1, and pre- and post-multiplying both sides by $\overline{\boldsymbol{\Sigma}}_j$, a term in $\overline{\boldsymbol{\Sigma}}_j$ can be isolated. This means that the equation can be rearranged thus:

$$\overline{\boldsymbol{\Sigma}}_j = \left(\sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t=j\}} \Pr(\boldsymbol{O},\boldsymbol{s}\,|\,\boldsymbol{\lambda}) \left(\mathrm{I} + \boldsymbol{C}_t \overline{\boldsymbol{\Sigma}}_j^{-1}\right)^{-1} \overline{\boldsymbol{\Sigma}}_j \right)^{-1} \times \sum_{t=1}^{T} \sum_{\{\boldsymbol{s}:\, s_t=j\}} \Pr(\boldsymbol{O},\boldsymbol{s}\,|\,\boldsymbol{\lambda}) \overline{\boldsymbol{\Sigma}}_j \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \boldsymbol{x}\boldsymbol{x}' \left(\overline{\boldsymbol{\Sigma}}_j + \boldsymbol{C}_t\right)^{-1} \overline{\boldsymbol{\Sigma}}_j. \quad (3)$$

If it is assumed that $\overline{\boldsymbol{\Sigma}}_j$ terms on the right hand side can be replaced by their previous values, then equation 3 constitutes a re-estimation equation for $\overline{\boldsymbol{\Sigma}}_j$.

4. COROLLARY

The problem as described is applicable to any feature set which is subject to additive, time varying Gaussian noise. A particular special case is that where the uncertainty (noise) can be assumed constant with time. Practically, this means that $\boldsymbol{C}$ is no longer dependent upon t, and certain matrix terms in the re-estimation equations become independent of the summation and cancel. In particular, the first re-estimate of the covariance above ceases to be an approximation, and the re-estimate of the mean reverts to the same as that for the conventional noiseless case.

5. EXPERIMENTS

5.1. Method

The new method for incorporating formant confidence measures in both training and recognition was tested using the same speaker-independent connected-digit recognition task with three-state phone models as was used in earlier studies [1]. As with the previous experiments, the baseline feature set comprised the first eight mel-cepstrum coefficients and an overall energy feature. The performance of this feature set was compared with one in which coefficients 6, 7 and 8 were replaced by three formant features for describing fine spectral detail. In the case of the formant features, the confidence measures were incorporated first in recognition and then also in training, testing both of the approximations suggested in the previous section for the re-estimation of the model variances. For both training algorithms, it was verified experimentally from the training-set probabilities that the re-estimation process converged after a few iterations. For all model sets, a total of ten iterations were performed before testing the models in recognition.

Alternative formant sets arising from labelling ambiguity were optionally accommodated in training and recognition, simply by choosing the formant set which gave the highest HMM emission probability for each frame and model state. Results using the confidences and alternative formant sets were compared with those obtained when no special treatment was given to the formant features.

5.2. Results and Discussion

From the results shown in Table 1 it can be seen that the formant features gave very poor performance unless the degree of confidence in their measurement accuracy was taken into account. When the formant features were not given special treatment, there were serious 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. These errors disappeared when the confidence measure was incorporated in recognition.

A small additional benefit was obtained by also incorporating the confidence measure in training, with very similar results being obtained for the two suggested approaches to training the model variances. In all cases, further small improvements in recognition performance were obtained by including alternative formant sets. The lowest error-rate of 2.5% that was achieved with the formants demonstrates a substantial improvement over the figure of 4.0% that was obtained when using only mel-cepstrum features, for the same total number of features.

These digit-recognition experiments have provided a good basis for initial comparisons, and experiments are now in progress to evaluate performance on the more demanding task of phone recognition using the TIMIT database.

6. CONCLUSIONS

We have shown that formant frequency estimates with confidence levels can be interpreted probabilistically, and that this interpretation leads to theoretically justifiable variants of the standard HMM recognition and re-estimation equations. Further, the theoretical results have been evaluated experimentally and shown to work in practice. Considerable recognition performance advantages have been demonstrated from incorporating formant features in this way, in comparison with using only cepstrum features.

It is planned to incorporate the formant representation into a segmental modelling paradigm to model formant trajectories, and then to progress towards developing an appropriate underlying model of time evolving speech characteristics.

7. REFERENCES

[1] John N. Holmes, Wendy J. Holmes, and Philip N. Garner. Using formant frequencies in speech recognition. In *Proceedings EUROSPEECH'97*, volume 4, pages 2083–2086, September 1997.

[2] Louis A. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Transactions on Information Theory*, IT-28(5):729–734, September 1982.

[3] Jan R. Magnus and Heinz Neudecker. *Matrix differential calculus with applications in statistics and econometrics.* Wiley series in probability and mathematical statistics. John Wiley and sons, 1988.

[4] Shayle R. Searle. *Matrix algebra useful for statistics.* Wiley series in probability and mathematical statistics. John Wiley and sons, 1982.

Model Set	%Subs.	%Del.	%Ins.	%Err.
8 cepstrum features+energy	2.8	1.0	0.2	4.0
5 cepstrum features+energy+3 formants	5.2	1.0	10.2	16.4
Add formant confidence measure (recognition only)	2.1	0.7	0.2	3.0
Also include second choice formants in recognition	2.1	0.4	0.4	2.9
Add confidence measure in training (method 1)	2.0	0.6	0.2	2.8
Also include second choice formants (training and recognition)	1.9	0.6	0.1	2.6
Add confidence measure in training (method 2)	2.0	0.6	0.2	2.8
Also include second choice formants (training and recognition)	1.8	0.6	0.1	2.5

Table 1: Connected-digit recognition performance for different feature sets.