Prediction of formant frequencies from linear combinations of filterbank and cepstral coefficients

Jesper Högberg

Abstract

The topic of this paper is formant frequency prediction using multiple linear regression. This technique provides robust formant frequency estimates but with limited precision. We apply this approach to predict formant frequencies of one male speaker comparing three different spectral representations. Mel scaled filterbanks are shown to perform slightly better than linear prediction based cepstral coefficients.

The baseline approach uses one linear prediction model for each formant. This approach is extended by using multiple models for each formant. In this case, each model is trained on data from sub-bands of the formant frequency. This method is shown to be very useful for predicting F2, which has a large frequency span. The rms error of F2 can be reduced by up to 45% using multiple models on an independent test set. Moderate sizes of training data suffice to derive the linear prediction models which implies that predictors can be trained for a new speaker with a small formant labelling effort.

Introduction

The use of formant frequencies in speech analysis and modelling is appealing in principle due to the close relation to the vocal tract geometry. Unfortunately, reliable formant frequency estimates are very difficult to extract from the speech wave. However, several studies have shown that there exist approximately linear relationships between formant frequencies and other spectral representations (Plomp et al., 1967; Pols et al., 1969; Klein et al., 1970; Pols et al., 1973; Broad & Clermont, 1989; Bayya & Hermansky, 1990; Hermansky & Cox, 1991).

Pols and colleagues used principal component analysis to derive features that maximally explained the variance in vowel spectra described by the log energy in 18 1/3-octave frequency bands. The relationship between the two most important components was strikingly similar to that of F1 and F2 (Pols et al., 1973).

Broad & Clermont (1989) predicted the first three formant frequencies from linear prediction (LP) based cepstral coefficients using multiple linear regression. Bayya & Hermansky (1990) also showed that formant bandwidths could be predicted using the same method.

Multiple linear regression yield formant predictions that compare very favourable to estimates derived directly from LP in terms of robustness (Hermansky & Cox, 1991). That is, smooth and continuous formant frequency trajectories can be obtained using this technique. The robustness is, however, traded for some loss in precision compared to LP.

Several formant based applications have emerged very recently indicating a revived interest in formant analysis. For instance, Ding & Campbell (1997) used a formant frequency based distance measure in unit selection for concatenation synthesis. Formant estimates have also been shown to be useful in speaker normalisation for automatic speech recognition (Lincoln et al., 1997).

In an effort to increase the naturalness of the KTH text-to-speech system (Carlson et al., 1991), we have applied data driven methods to formant synthesis (Högberg, 1997). The idea is to create a trainable TTS-system that can be easily adapted to a new speaker. A crucial point in this endeavour is, of course, automatic parameter estimation, e.g. formant extraction.

The starting point in this study is hence to use multiple linear regression models to predict formant frequencies, F_i, from some spectral representation, β, with N dimensions,

$$\hat{F}_i = \alpha_{i,0} + \sum_{j=1}^{N} \alpha_{i,j}\beta_{i,j} \,. \qquad (1)$$

The prediction coefficients, $\alpha_{i,j}$ are derived using standard least squares methods. Three different spectral representations are compared in this study.

Pols et al. (1973) found that linear models based on speech spectra from 50 speakers did not provide precise formant predictions for the individual speakers though the pooled results were good. Broad & Clermont (1989) also showed, as could be anticipated, that the prediction errors increased when formant frequencies of more than one speaker were predicted. In view of these results and with the speech synthesis application in mind we are primarily interested in speaker dependent models. Therefore, the main concern of this study is to investigate the amount of data that is needed to train predictors for one speaker obtaining maximum prediction precision for unseen data. We propose a novel approach to increase prediction precision in which multiple models are used for each formant. In this procedure, the final formant frequency prediction is a weighted sum of the contributions from the individual models.

The rest of this paper is organised as follows: The next section briefly describes the speech data used for this investigation. The following section addresses the use of different spectral representations and model orders. Next, the multiple model approach is evaluated and the need for training data is investigated in quantitative terms. A discussion and some conclusions are provided in the last sections of this paper.

Speech material

The speech data (16 kHz sampling frequency) used in this study pertains to one male subject reading eleven short stories and newspaper text. The first part of the material has been used in several other studies, e.g. Carlson & Nord, 1993; Högberg, 1994. Both corpora have been manually segmented and labelled and contain more than 30 minutes of speech together. In all, the speech data comprise some 12,000 vowels. The first four formant frequencies have been hand corrected and used in a previous study (Högberg, 1997). Figure 1 shows the formant frequency distributions. The mean frequencies are 442, 1451, 2470 and 3424 Hz for F1, F2, F3 and F4, respectively. The corresponding standard deviations are 94, 356, 204 and 256 Hz. Thus, the relative variation in the first two formant frequencies are much larger than the corresponding variation in F3 and F4. The standard deviations are 21% and 25% of the mean frequencies for F1 and F2. The corresponding variation in the higher formants amounts to 8% for F3 and 7% for F4.

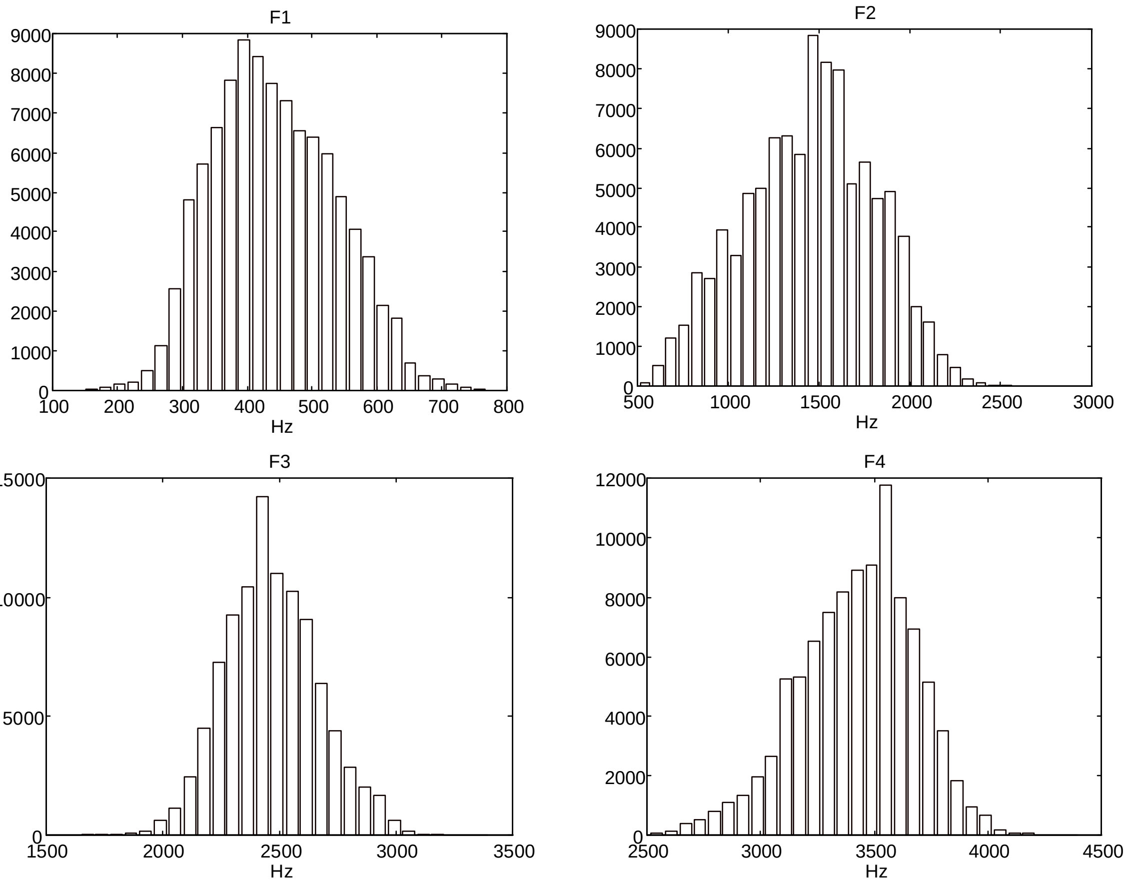

Figure 1. Formant frequency histograms based on 100,000 frames.

Effect of model order and spectral representation

Speech spectrum representations

The speech spectrum representations were mel-scaled filterbank coefficients (MFFB), cepstral coefficients derived from the mel scaled filterbank (MFCC), and cepstral coefficients derived from linear prediction parameters (LPCC). This analysis was performed using the HTK-toolkit (Young & Woodland, 1993).

The MFFB coefficients (filter outputs) were calculated as weighted sums of FFT magnitudes in N frequency bands. The frequency bands were equally spaced on a mel scale defined by

$$Mel = 2595 * \log_{10}\left(1+\frac{f}{700}\right), \qquad (2)$$

where f is the frequency in Hz.

The mel scale filterbank cepstral coefficients, c_n, were calculated from the N filterbank coefficients, m_j, using the Discrete Cosine Transform

$$c_n = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi n}{N}(j-0.5)\right), \qquad (3)$$

and the LP-cepstral coefficients are given by

$$c_n = -a_n + \frac{1}{n}\sum_{i=1}^{n-1}(n-i)\,a_i c_{n-i}\ , \qquad (4)$$

where a_i are the coefficients in the all pole model of the vocal tract transfer function

$$H(z) = \frac{1}{1+\sum_{j=1}^{p} a_j z^{-j}}. \qquad (5)$$

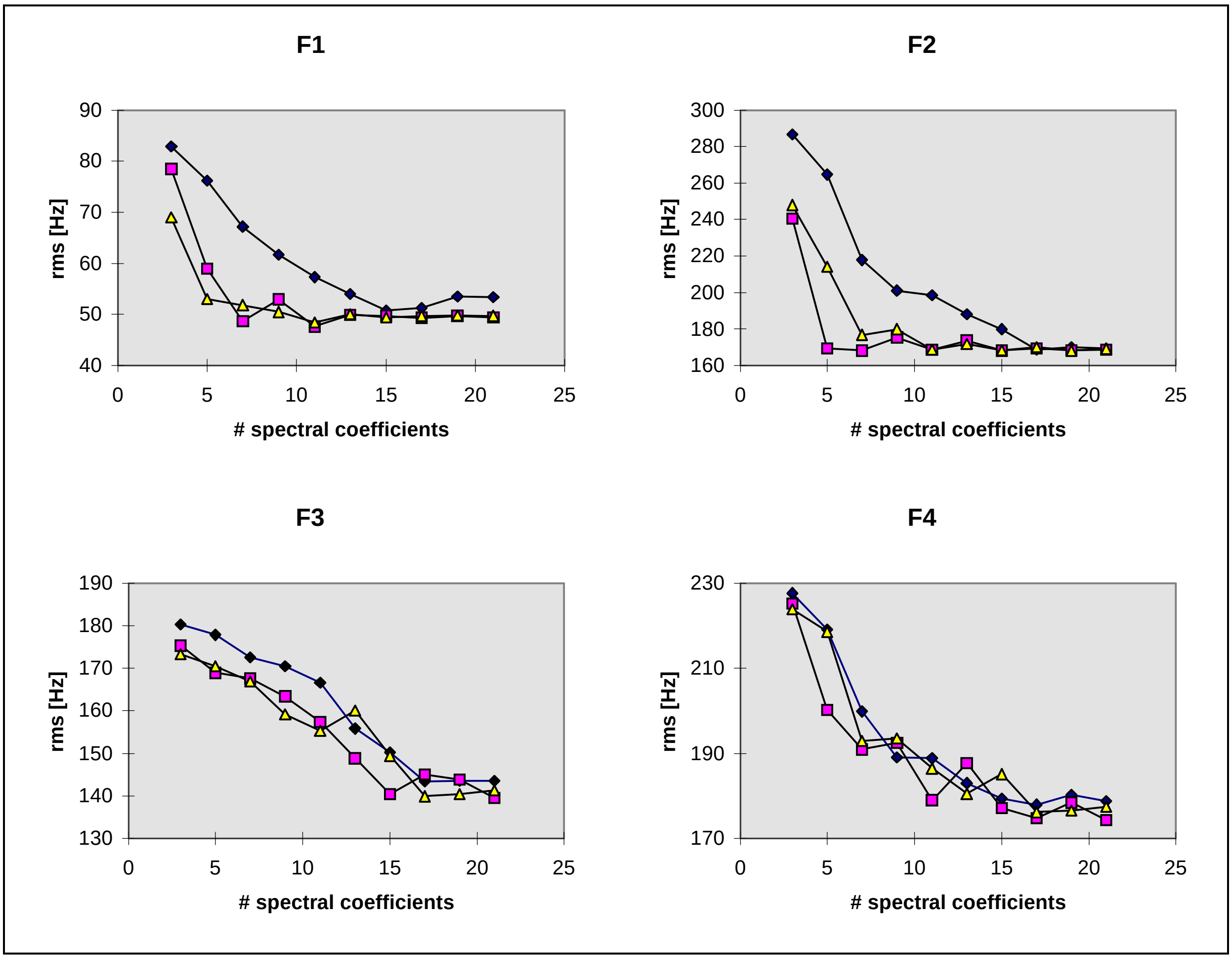

Figure 2. The root mean square errors in predicted formant frequencies for different number of spectral coefficients (model order.) Diamonds stand for LPCC based models, triangles and squares represent MFCC and MFFB based models respectively.

All coefficients were calculated from pre-emphasised speech using 25 ms Hamming windowed speech frames at rate of 100 frames per second. The maximum set of formant frequency vectors amounted to almost 100,000 tokens. These vectors were extracted from vowel segments only.

Each formant frequency was aligned with the corresponding spectral vector. Equation *1.* was solved for the prediction coefficients, $\alpha_{i,j}$, that minimised the total squared error in the predicted formant frequency values. This procedure was repeated for each of the four formant frequencies for each spectral representation and for each model order. In the MFCC case, the number of cepstral coefficients was chosen to be the number of filters minus one, c.f. equation *2*. The number of LP based cepstral coefficients was chosen to equal the order of the LP-analysis.

Results

Figure 2 shows the root mean squared errors in the predicted formant frequencies for the MFFB, MFCC and LPCC-representations as a function of the number of coefficients. The errors are computed from predictions of the entire data set, i.e., in this figure the train and test sets are the same. The different spectral representations all yield similar predictions when the higher orders are compared. The LPCC seems to be less well suited than MFFB and MFCC for lower order modelling of low order formants. The optimum order of the LPCC models is about seventeen. This matches the best theoretical model order for the average male speaker using two poles for each formant and one for the source spectrum ($F_s/2$ = 8kHz).

The relative impact of the formant frequencies on the detail of the speech spectrum is reflected in these curves. Generally speaking, the best model order increases with the formant number.

The best models yield rms errors around 50, 170, 140 and 175 Hz for F1, F2 F3 and F4, respectively. This means that roughly 50% of the standard deviation could be explained for F1 and F2. For F3 and F4, about 30% of the standard deviation could be explained using these models.

The smallest rms errors are larger than those obtained by, for instance, Broad & Clermont (1989) who reported 27, 79 and 82 Hz for F1 to F3 when training and testing on the same material of one speaker. Part of this discrepancy can be explained by the complexity and quality of the speech data bases. One important explanation is probably found in the reference data; some precision was traded for speed in the markup procedure of the current data base. Moreover, the vowel spectra in the current study are influenced by various voice source conditions found in continuous speech, most importantly affecting spectral tilt. Many vowel spectra are substantially influenced by the surrounding context, e.g. adjacent nasals have introduced extra poles and zeros.

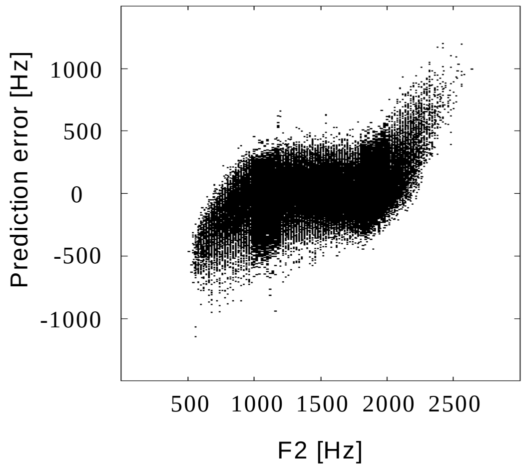

Figure 3. Measured F2 values plotted against the prediction errors. The predictions were made on the entire data set using 17 mel filterbank cepstral coefficients (MFCC).

In Figure 3, the measured F2 values are plotted against the prediction errors. Apparently, the prediction errors are not randomly distributed over the formant frequency range. It seems that the linear mapping does not model the variation over the entire frequency range properly. This is particularly troublesome for F2 which has a very wide frequency span, approximately 500-2500 Hz for the speaker of this study. This figure shows that the prediction errors are clearly frequency dependent, high frequencies above 2000 Hz have been underestimated and low values under 1000 Hz have been overestimated. This seems to be a general problem with the linear regression approach. Hermansky & Cox (1991, page 331) commented on similar F2 errors in predictions on a sample utterance.

Multiple models and training effects

In the foregoing, it was noted that the linear model did not adequately model the whole frequency range. There are a number of plausible explanations and potential remedies to this problem, one being that the models would benefit from higher order prediction terms. However, adding quadratic and cubic terms to the regression equation *1* yields only very small improvements in the rms errors. Instead, we suggest that the relation is only piece-wise linear, i.e., the relation is a well defined function only for limited frequency ranges. In the following section, we propose a method that is based on this hypothesis. This approach utilises statistics to calculate the probability that a spectral vector has a formant in a certain frequency band. Phone-specific statistics can also be accommodated when formant predictions are made on a labelled data base. Henceforth, seventeen MFCCs are used for prediction. This “model order” was chosen because it provided the smallest errors for all formants. The cepstral representation was preferred to the filterbank because of its statistical properties (variable independence) which allows some simplification of the calculations.

Multiple models

The range of the *i*th formant was divided into *K* frequency bands and one linear model, $M_{i,k}$, was derived for each such sub-band of each formant, k=1,..,K. This divides the prediction process into two steps, the first of which consists of predicting the band in which the formant frequency is located. The second step is a fine tuning of the first prediction. The accuracy in the first step becomes crucial, incorrect decisions will result in discontinuities and hence decreased robustness. Therefore, we have implemented a soft classification strategy to avoid these problems. The formant frequency prediction is a weighted sum of the contributions from all *K* models for that particular formant.

The weight assigned to the prediction generated by the model, $M_{i,k}$, is calculated as the probability, *P*, that the true formant value is found in the *k*th frequency band given the corresponding spectral vector, $\boldsymbol{\beta}$. Hence, a prediction for the *i*th formant is given by

$$\hat{F}_i = \frac{\sum_{k=1}^{K} P\left(M_{i,k} \middle| \beta\right) \hat{F}_{i,k}}{\sum_{k=1}^{K} P\left(M_{i,k} \middle| \beta\right)}. \qquad (6)$$

Bayes’ rule gives that

$$P\left(M_{i,k} \middle| \beta\right) = \frac{P\left(\beta \middle| M_{i,k}\right) P\left(M_{i,k}\right)}{P(\beta)}, \qquad (7)$$

where $P(\beta|M_{i,k})$ is estimated using a multivariate normal distribution with a diagonal covariance matrix

$$P\left(\beta \middle| M_{i,k}\right) = \frac{\exp\left(-\frac{1}{2}\sum_{j}^{N}\left(\frac{\beta_j - \mu_{i,j,k}}{\sigma_{i,j,k}}\right)^2\right)}{2\pi^{N/2}\prod_{j}^{N}\sigma_{i,j,k}}, \qquad (8)$$

with $\mu_{i,j,k}$ and $\sigma_{i,j,k}$ being the sample mean and standard deviation of the *i*th formant in the *j*th spectral dimension in the *k*th sub-band. *N* is, as before, the number of spectral coefficients.

$P(M_{i,k})$ is the fraction of the total count that has been used to train model $M_{i,k}$, i.e., the fraction of the tokens of formant *i* that are located in frequency band *k*.

Table 1. Limits (Hz) of the four frequency bands used for each formant frequency.

	BAND 1	BAND 2	BAND 3	BAND 4
F1	150-350	351-500	501-650	651-1500
F2	500-1100	1101-1500	1501-1900	1901-3000
F3	1200-2000	2001-2500	2501-3000	3001-4500
F4	2000-3200	3201-3500	3501-3800	3800-5000

In this investigation we divided the frequency range of each formant into four sub-bands (K=4). Table 1 shows the frequencies bounding the non-overlapping bands for each formant. The upper and lower bounds of each formant frequency are used to exclude extreme outliers.

Adding phone-specific statistics

In the case when the speech data have been segmented prior to the formant frequency analysis, we have access to the phone identity for each frame to be processed. Under these circumstances we can substitute $P(M_{i,k})$ for $P(M_{i,k}\,|I)$, where *I* is the identity of the segment. The identity need not correspond to the phoneme, it might just as well apply to a broader class like “back vowel” or to a more spectrally well defined unit like a phone in a specified context. The latter kind of statistics would of course demand more training data. In this context, we consider phoneme identities or even broad phone classes appropriate since we strive to keep the amount of manually analysed training data to a minimum.

Phoneme statistics are used in the experiments of the current study.

The importance of the size of the training material

In all regression and classification it is important to assess model performance using unseen data. In this study, the training material relies on manual work which we, needless to say, want to minimise. Therefore, it is particularly interesting to investigate how the prediction performance on independent test data varies with the size of the training set.

The data was consequently divided into one test set and a number of training sets of various sizes. The test set consisted of roughly 24,000 tokens and the full training set consisted of 75,000 tokens. There were thirteen training sets of various sizes that were designed by adding new data to the set last used, thus accumulating more and more training data until the entire training material was used. The test set and the full training set were disjunct.

Results

Figure 4 shows the test set rms errors for the different type of models as a function of the training set size. The prediction errors using the original one model per formant (single model) are indicated with diamonds. The lowest error rates, using the single model approach, are comparable to those when testing and training on the entire data set. The test set rms errors are less than 10 Hz higher for all formants than the corresponding training set errors depicted in Figure 2. This means that there are very small differences relatively seen except for F1. The models appear to be fully trained, with respect to the current test set, when some 40,000 tokens are used for training. That is, no further improvement is obtained when larger training sets are used.

The multiple model results can also be seen in Figure 4. The multiple model results, with and without the phone-specific statistics, are indicated with squares and triangles, respectively. Using *a priori* knowledge of the segment identity reduces prediction errors as anticipated. We see the largest gains using segment identity information for F1 and F2. The maximally trained models yield a seven to ten percent error reduction relative to the multiple models with no segment identity information.

The multiple model approach performs favourably compared to the single model approach when applied to F1 and, in particular, to F2. In the case of F1, the performance is somewhat unstable for very small amounts of training data. For moderate to large training sets, the best multiple models give about ten percent lower errors than the single models. The most dramatic improvement using multiple models is obtained for F2. The multiple models using phone-specific statistics reduce the mean prediction error with about 35-45% percent for all training set sizes. The best model set used to predict F2 produces a rms error below 100 Hz when applied to the test set.

There seems to be no gain in applying multiple models in prediction of higher formants. Given a reasonable amount of training data, the single models do just as well or better than the multiple models. Single models are probably sufficient because the higher formants have smaller variances relative to the centre frequencies. Also, the more subtle impact on the spectrum of the higher formants, especially on a perceptually motivated scale, also makes it harder to distinguish in what frequency band they are located.

Prediction results that can be anticipated from Figure 2 are obtained when only eleven MFCCs are used (not shown in Figure 4). This representation is better for small and moderate size training sets predicting lower formant frequencies. As the amount of training data increases and the number of formants increases larger numbers of coefficients can be motivated. Prediction errors need not be zero to be negligible; the perceptual limits are more interesting. Therefore, the best results are compared to formant frequency difference limens which amount to 3-5% of the frequency (Flanagan, 1955; Nord & Sventelius, 1979). Table 2 compares the best results obtained with multiple models for each formant using seventeen MFCC coefficients and *a priori* knowledge about segment identities to formant frequency limens. The frequency limens are approximated as five percent of the mean formant frequencies.

Table 2. BEST RMS: The best test set prediction results using seventeen MFCCs, a priori knowledge about segment identity and multiple models for each formant. LIMEN: formant frequency difference limens approximated as 5% of the mean formant frequency values.

	F1	F2	F3	F4
BEST RMS [Hz]	52	97	155	186
LIMEN [Hz]	22	72	124	171

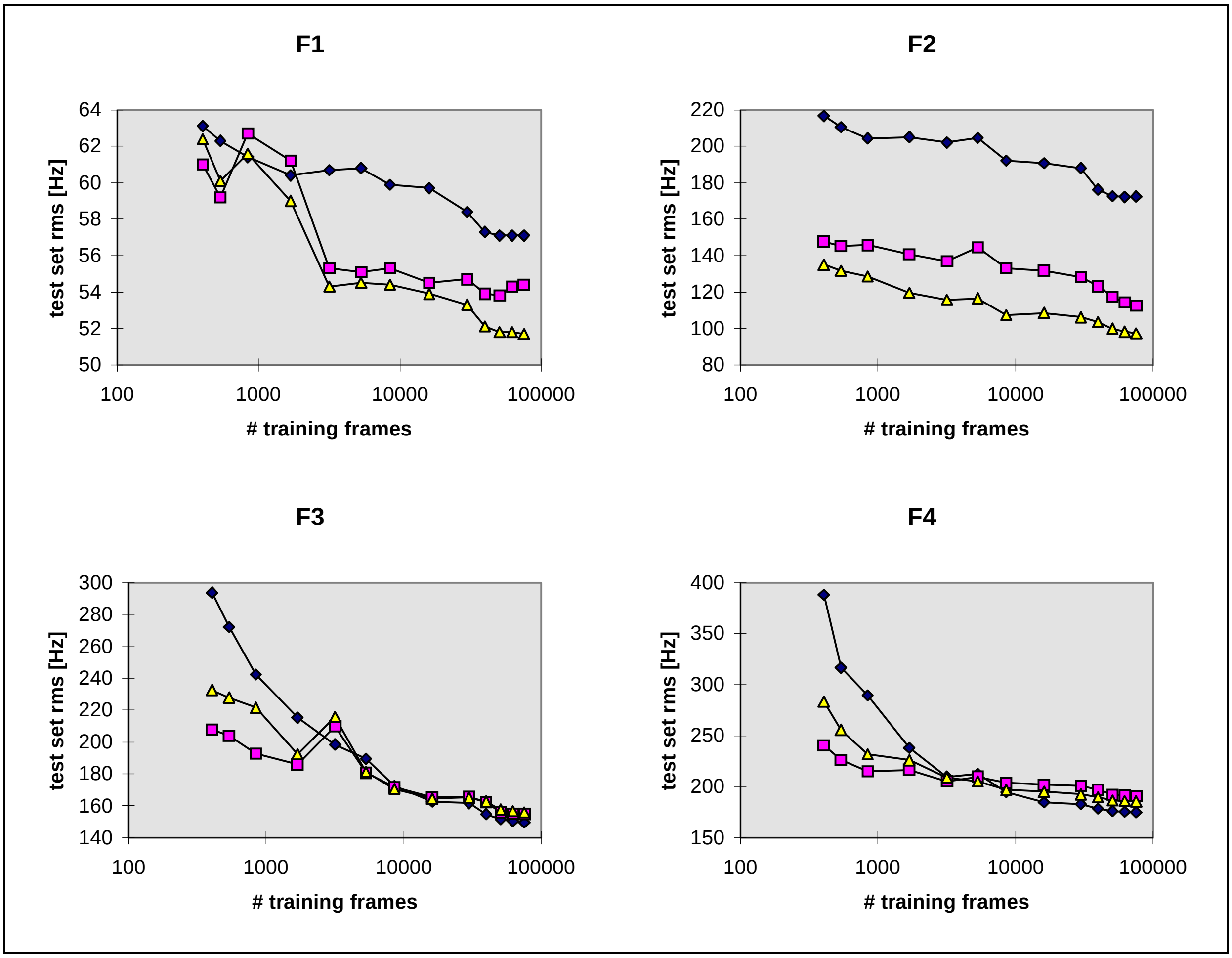

Figure 4. The root mean square errors in predicted formant frequencies of the test set as a function of the size of the training material. The formant frequencies were predicted from 17 MFCCs. Diamonds stand for single model results, squares represent the multi model results and triangles correspond to multi models complemented with phone-specific statistics.

Discussion

As mentioned in the introduction, we are currently developing a trainable formant based TTS system (Högberg, 1997). Therefore, our primary interest has been to find out if multiple linear regression can provide formant predictions that are precise enough to be useful for speech synthesis. Table 2 indicates that the method still needs some refinement to reach a performance where the errors are perceptually insignificant. If the precision could be somewhat improved this method is potentially very useful considering its robustness and the fact that only moderately sized training corpora are needed. From Figure 4 we can see that some 3000-5000 tokens are sufficient to train models that will explain a great deal of the formant frequency variability. This means that, in the current data base, only 20 - 40 utterances would have to be manually corrected. Using limited amounts of data, more attention could be paid to precision than has been done in the correction of the current data base. This would probably decrease the prediction errors. The training data should be controlled for phonetic balance to ensure sufficient coverage of the acoustic space.

A potential problem is the influence of the voice source. For instance, the tilt of the speech spectrum changes with the closing characteristics of the vocal folds. Therefore, variation contributed by this factor most certainly influences the formant frequency prediction. However, an analysis dedicated to quantify the voice source influence in the current prediction framework is needed before we know the importance of this specific problem.

As regards the precision, some modifications could be applied to the proposed multiple model approach. For instance, the probabilistic modelling could be extended to substitute the linear models entirely by creating a code book of Gaussian probability density functions. The formant frequency would then be the sum of the code book centroids weighted by their output probabilities for an observed spectral vector.

An alternative to the probabilistic choice of model would be to adopt an analysis-by-synthesis procedure. Each model would be evaluated by mapping the predicted formant frequency back to a spectral vector and, subsequently, picking the model yielding the spectral vector with the smallest distance from the original.

Figure 3 shows a prediction error that is frequency dependent. It could be argued that this is a consequence of the uneven distribution over the formant frequency range as depicted in Figure 1. It is possible that the under-representation of the extremes means that these cases are less well trained. However, some preliminary experiments using *a priori* knowedge of the distribution to weight the training tokens did not provide an improvement. These experiments were rather crude, so more elaborate attempts using robust prediction should not be ruled out.

Complementary techniques could be used for further fine tuning of the formant frequencies in applications demanding high prediction precision. Carlson & Glass (1992) proposed a method to determine speech parameters adopting an analysis-by-synthesis approach. Synthetic spectra were matched against real speech spectra searching for optimal parameter values in an iterative fashion. The starting guess for the parameter values of the synthetic spectra were provided by a rule based formant synthesiser. Instead of using rules, we suggest that the initial formant frequency estimates could be provided by multiple linear regression.

Further work might also be pursued on speaker independent linear prediction of formant frequencies.

Conclusions

In this study, we have confirmed that formant frequencies can be predicted robustly from coefficients of different spectral representations using multiple linear regression. The experiments we have described addressed issues concerning the choice of spectral representation and model order, modelling extensions and the influence of training set size.

Mel scaled filterbank coefficients, the corresponding cepstral coefficients and cepstral coefficients derived from linear prediction were compared. All three spectral representations provided similar results when a relatively large number of coefficients (>15) were used. The mel based features provided better prediction models for F1 and F2 when a small number of coefficients was used.

Further, we have extended the original regression method to account for multiple models for each formant frequency in order to increase the precision in the predictions. Using this approach we could reduce the F2 rms error on an independent test set with more than 40%. The performance of the method was not as impressive when applied to the other formant frequencies. We hypothesise that the relative success of applying the multiple model approach depends on the width of the formant frequency ranges.

We have also taken a special interest in relating the amount of training data to prediction performance. It seems that moderate amounts of training data (20-40 utterances) suffice to obtain reasonable models. This means that only a small amount of manual labour is needed to create models tailored to a new speaker.

Acknowledgements

This work has been financed by grants from The Swedish National Language Technology Program and by the Centre for Speech Technology.

References

Bayya A & Hermansky H (1990). Towards feature based speech metric. *Proc of ICASSP'90*; 781-784.

Broad D & Clermont F (1989). Formant estimation by linear transformation of the LPC cepstrum. *J Acoust Soc Am 86/5:* 2013-2017.

Carlson R & Glass J (1992). Vowel classification based on analysis-by-synthesis. *STL-QPSR, KTH,* 4/1992: 17-27.

Carlson R & Nord L (1993). Vowel dynamics in a text to speech system - some considerations. *Proc of Eurospeech'93*, Berlin; 1911-1914.

Carlson R, Granström B & Hunnicutt S (1991). Multilingual text-to-speech development and applications. In: Ainsworth A (ed.), *Advances in speech, hearing and language processing.* London: JAI Press, UK.

Ding W & Campbell N (1997). Optimising unit selection with voice source and formants in the CHATR speech synthesis system. *Proc of Eurospeech'97,* Rhodes; 537-540.

Flanagan J (1955). A difference limen for vowel formant frequency. *J Acoust Soc Am* 27/3: 613-617.

Hermansky H & Cox L (1991). Perceptual linear predictive (PLP) analysis-resynthesis. *Proc of Eurospeech'91,* Genova; 329-332.

Högberg J (1994). A phonetic investigation using binary regression trees. *Papers from the Eighth Swedish Phonetics Conference,* Lund, 1994.

Högberg J (1997). Data driven formant synthesis. *Proc of Eurospeech'97,* Rhodes; 565-568.

Klein W, Plomp R & Pols L (1970). Vowel spectra, vowel spaces and vowel identification. *J Acoust Soc Am*, 48/2(2): 999-1009.

Lincoln M, Cox S & Ringland S (1997). A fast method of speaker normalisation using formant estimation. *Proc of Eurospeech'97*, Rhodes; 2095-2098.

Nord L & Sventelius E (1979). Analysis and prediction of difference limen data for formant frequencies. *STL-QPSR, KTH*, 3-4/1979: 60-72.

Plomp R, Pols L & van de Geer (1967). Dimensional analysis of vowel spectra. *J Acoust Soc Am* 41/3: 707-712.

Pols L, Tromp H & Plomp R (1973). Frequency analysis of Dutch vowels from 50 male speakers. *J Acoust Soc Am* 53/4: 1093-1101.

Pols L, van der Kamp L & Plomp R (1969). Perceptual and physical space of vowel sounds. *J Acoust Soc Am* 46/2(2): 458-467.

Young S, Woodland P & Byrne W (1993). *HTK: Hidden Markov Model Toolkit V1.5*. Entropic Research Laboratory.