SUPERVISED CEPSTRUM-TO-FORMANT ESTIMATION: A NEW PIECEWISE-LINEAR MODEL

Simon Hawkins and Frantz Clermont

Computer Sciences Laboratory
Research School of Physical Sciences
Australian National University

ABSTRACT - A multiple-linear regression model of the relationship between low-order LPC-cepstral coefficients and vowel formant contours has been proposed by Broad and Clermont (1989). However, less-than-perfect formant estimates generated using this method suggest that the assumption of linearity underlying this model is questionable. In the present study, a neural net, with the potential for developing mappings that provide a nonlinear partitioning of large multi-dimensional space, is shown to produce substantially more accurate formant estimates than is possible using Broad and Clermont's multiple linear regression model. Because the neural net provides no indication as to the nature of the nonlinearity it has discovered, we propose a new piecewise multiple-linear regression model of the cepstrum-to-formant relationship. This parametric nonlinear model is thought to capture the quintessential nonlinearity in the cepstrum-to-formant relationship because it produces first and second formant estimates which are almost as accurate as those generated using the neural net.

INTRODUCTION

Pols, Plomp, and Tromp (1973) demonstrate that the first and second formant frequencies of steady-state vowel data could be obtained directly from *linear* combinations of the log energy output levels of a bank of eighteen 1/3 octave bandpass filters. Pols et al report correlations (based on 12 vowel types averaged over 50 speakers) of near 0.98 between these linear combinations and the formant frequencies. Pols et al warn, however, that their method is only capable of generating accurate formant estimates for vowel formants which have been *averaged* over many speakers.

Pols' finding inspired Broad and Clermont (1989) to find a model capable of deriving formant estimates from *linear* combinations of the low-order LPC-cepstral coefficients on the grounds that these are linearly related to the low-resolution log spectrum. Unlike Pols et al, Broad and Clermont attempt to estimate the vowel formants of *individual* speakers rather than estimating vowel formants which have been averaged over multiple speakers.

To produce an equation capable of estimating an individual speaker's formants from the speaker's low-order cepstral coefficients, Broad and Clermont compute a separate multiple-linear regression (MLR) equation for each formant, F_j, j = 1 to 3, of the form:

$$\hat{F}_j \approx \alpha_0 + \sum_{i=1}^{M=14} \alpha_i c_i \qquad (1)$$

where c_i is the ith LPC-cepstral coefficient, and α_i is the set of regression coefficients computed to minimize the mean square error in estimating F. These MLR equations are found to produce relatively robust but only moderately accurate formant estimates for the formants of individual speakers.

A possible problem with Broad and Clermont's MLR model, however, is that it assumes that there is a *linear* cepstrum-to-formant relationship for each *individual* speaker. Yet Pols' work only demonstrates a strong linear relationship between filter-bank outputs and formants which have been *averaged* over many speakers. It seems entirely possible, therefore, that a *nonlinear* relationship exists between the filter-bank outputs and formants of each *individual* speaker in the study by Pols et al. When these filterbank outputs and formants are averaged over many speakers, the nonlinear relationship between filterbank outputs and formants which exist for each individual speaker cancel one another out to produce a relationship between average filterbank outputs and average formants which is approximately linear. In estimating the formants of *individual* speakers, one should therefore question the assumption of a *linear* cepstrum-to-formant relationship which underlies the MLR model.

In the present study, a *continuous-valued output neural net (NN)* is used to investigate whether the cepstrum-to-formant relationship might be better described as *nonlinear*. The NN is capable of learning virtually any nonlinear cepstrum-to-formant relationship which may exist. However, the mapping learnt by the NN is nonparametric and is not, therefore, readily interpretable. In the present study, formant estimates generated using the NN are compared to those generated by the MLR model which assumes a strictly linear cepstrum-to-formant relationship. If more accurate estimates of the formants are obtained using the NN, then this will be taken as evidence that the assumption of linearity underlying Broad and Clermont's MLR model is untenable.

PRESENT STUDY

SPEECH MATERIAL

The speech database used in the present study was developed by Clermont (1990) and is the same as that used by Broad and Clermont (1989). The database is composed of nine vowel types produced by four Australian males in CVd context, where C=/ h, b, d, g, p, t, k /, and V is as in heed, hid, head, had, hard, hod, who'd, hudd, and herd. There are five repetitions of each CVd monosyllable from each speaker. The formant frequencies were obtained for 11 frames equally spaced through each vowel for a total of 3465 frames for each speaker. The first three formant contours were measured in three stages: (1) 14th order LPC autocorrelation analysis on Hamming windowed 256-sample frames, (2) peak-picking on the LPC spectrum followed by formant tracking (McCandless, 1974), and (3) hand editing of the resulting candidates on the basis of estimated bandwidth, formant ranges, and continuity. Following this, the values are averaged frame by frame across the five repetitions of each CVd syllable to form a training set of 693 samples per speaker.

ARCHITECTURE OF THE NEURAL NET

The NN is a modified form of a fully interconnected, feedforward, multi-layer perceptron (Lapedes and Farber, 1987). It is designed to learn a function which derives the first, second, and third formants simultaneously from a set of 14 low-order LPC-cepstral coefficients. The net has an input layer, a single hidden layer, and an output layer. The input layer contains 14 inputs which correspond to the 14 unscaled LPC-cepstral coefficients in each sample. The hidden layer contains 25 hidden units. The activation level of each hidden unit is a nonlinear sigmoidal function of the weighted sum of inputs to this unit. The complexity of the nonlinear mapping which can be learnt by the NN is determined both by the number of units in the hidden layer of the net and also by the number of hidden layers. A single layer of 25 hidden units is used in the present study because this is found to produce optimal formant estimates when the NN is trained on a set of three speakers and then tested on a fourth speaker. The output layer contains three output units corresponding to the first, second, and third formants. The activation level of each output unit ranges between 0 and 1 and is a *linear* rather than sigmoidal function of the weighted sum of the inputs to this unit.

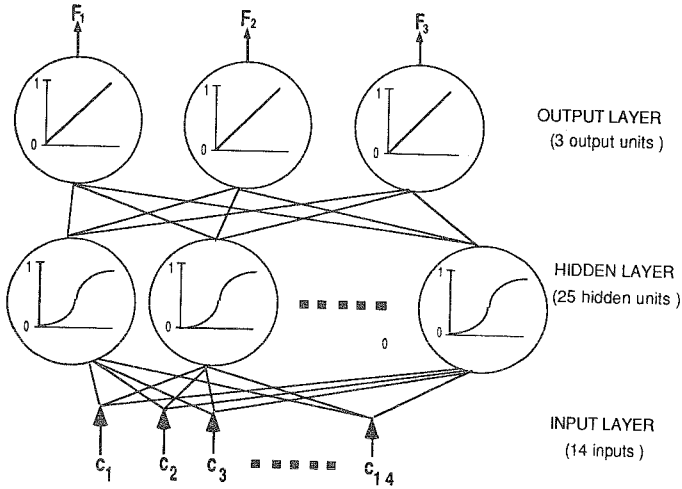

Figure 1. Architecture of the continuous-valued output neural net used in this study

Prior to training the neural net, the first, second, and third formant frequencies of samples in both the training and test sets are linearly scaled to lie on a continuous-valued scale between 0.1 and 0.9. The NN is then trained using the method of back-propagation to update the inter-connection weights in order to minimize the system error. The system error is defined as the square of the difference between the actual and estimated formant frequencies summed over the three output units and then averaged across all training samples. Training of the neural net terminates when the error for each

individual pattern falls within acceptable limits and when further iterations do not result in a further decrease in the system error. After the NN has been trained, a set of test samples is presented to the trained neural net. The activation levels of the three output units are converted back to formant frequencies for each test sample and the root-mean-squared (rms) error between the actual formant frequencies produced by the NN and the desired formant frequencies is calculated for the entire test set.

EXPERIMENT 1: Determining whether the cepstrum-to-formant relationship is linear or nonlinear.

EXPERIMENTAL DESIGN

This database is divided between a training and test set in two ways: (1) *representation* (in which the training set and test set are the same) versus *prediction* (in which the training and test sets differ), and (2) *single-speaker* versus *multiple-speaker* settings. These two distinctions result in four experimental conditions.

In each experimental condition, the models are trained on a set of training data and then evaluated on a set of test data. Formant estimates generated by the models on the *test* data are then pooled over all trials in the experimental condition. A single rms error measure for each formant is then calculated on this pooled data.

In the two *representation* conditions, the MLR and NN models are trained and then tested on the same data set. In the *single speaker representation* condition, the models are trained and tested on all frames from a single speaker. There are four experimental trials in this condition corresponding to the four speakers. The rms error, shown in the left hand side of Table 1 below, indicates the ability of the models to represent the formant data of a single speaker. In the *multiple-speaker representation* condition, the data from all four speakers is pooled to form the data set on which the models are trained. The same pooled data set is then used to test the models. There is thus only a single experimental trial in this condition. The rms error measure from this pooled data, shown in the left hand side of Table 1 below, indicates the ability of the models to represent the pooled formant data of four speakers.

In the *prediction* conditions, the MLR and NN models are trained on one training set and then evaluated on an independent test set. The training and test sets are derived from either a single speaker or from multiple speakers. In the *single-speaker prediction* condition, four out of the five repetitions from each speaker are used to train the two models. The fifth repetition from the same speaker is then used to test the models. The process is repeated so that each repetition is used once as the test set. The rms error on the pooled data from the 4 speakers x 5 repetitions = 20 experimental trials indicates how well the models can predict a speaker's formants when trained on independent formant data from that speaker. In the *multiple-speaker prediction* condition, the models are trained on the pooled frames from three speakers and then tested on the frames from a fourth speaker. The rms error on the pooled data from the four experimental trials indicates the ability of the models to predict a speaker's formants in the absence of any training data from that speaker.

RESULTS AND DISCUSSION OF EXPERIMENT 1

TABLE I. The rms errors (in Hz) which result when the MLR and NN models are trained and tested on the same data set (Representation) or on different data sets (Prediction) from either a single speaker or from multiple speakers.

	Representation						Prediction					
	Single speaker			Multiple speaker			Single Speaker			Multiple speaker		
	F1	F2	F3	F1	F2	F3	F1	F2	F3	F1	F2	F3
MLR	27	78	81	34	109	132	41	126	134	46	153	220
NN	21	45	50	19	52	68	31	64	79	35	84	149

From Table 1, it is clear that the NN model generates substantially more accurate formant estimates than the MLR model for all three formants in all four experimental conditions. These results indicate that the NN has found significant nonlinearities in the cepstrum-to-formant relationship which are not accounted for by the MLR model. The question remains, however, as to the exact nature of the nonlinearity in the cepstrum-to-formant relationship.

This nonlinearity may be explicable in terms of a dichotomy between front and back vowels proposed originally in a piecewise-planar model by Broad and Wakita (1977). Broad and Wakita observe that the

first three formant frequencies of vowel steady-states produced by an individual speaker cluster about a two-part, V-shaped, piecewise-planar surface. The formants of the front vowels cluster about one plane; the formants of the back vowel about the other. The intersection between the two planes is a line of nearly constant F_2 corresponding to the F_2 of a uniform vocal tract of the same length as the speaker under investigation. Although this critical F_2 breakpoint between the two planes was originally determined by inspecting a three dimensional plot of the three formants, Broad (1981) suggests an alternative method of finding this breakpoint. He notes that the histogram of an individual speaker's F_2 frequencies is bimodal with the fewest number of F_2 frequencies in the mid-range of F_2. This F_2 frequency corresponds to the constant F_2 value which separates the plane about which the front vowel formants cluster from the plane about which the back planes cluster.

Thus, Broad and Wakita's (1977) piecewise-planar model stipulates that a different linear inter-formant relationship exists for the front and back vowels. If Broad and Clermont's cepstrum-to-formant MLR model is applied to predict front-vowel and back-vowel formants *independently*, then the resulting effect is that of a piecewise multiple-linear regression (PW-MLR) model. In other words, Broad and Wakita's inter-formant piecewise-planar model and Broad and Clermont's cepstrum-to-formant MLR model combine to form the new piecewise-multiple-linear regression (PW-MLR) model proposed in this paper. In this PW-MLR model, separate MLR cepstrum-to-formant equations are calculated for the front and back vowels.

EXPERIMENT 2: Modelling the nonlinear cepstrum-to-formant relationship

EXPERIMENTAL DESIGN

To test the validity of the PW-MLR model proposed above, samples are partitioned into front and back vowel categories. This is achieved using histograms of F_2 values in the steady-state portions of each vowel nucleus (frame 6) to find the critical F_2 breakpoint which separates the front and back vowels. For all speakers in the study, the distribution of F_2 frequencies is found to be bimodal with the fewest F_2 values occurring at a frequency of approximately 1600 Hz. Samples from the entire vowel trajectory (frames 1 to 11) are assigned to the front categories, if their F_2 value is greater than or equal to 1600 Hz, or to the back category if their F_2 values is less than 1600 Hz.

To derive formant estimates using the PW-MLR model, least mean squares planes of the general forms given in the following equations are computed separately for frames in the front category (Equation 2a) and back category (Equation 2b). Formant estimates for the formant, F_j, j = 1 to 3, are obtained using the MLR equations:

$$\hat{F}j \approx \alpha_0 + \sum_{i=1}^{M=14} \alpha_i c_i \quad \text{for } F_2 \geq 1600 \text{ Hz} \qquad (2a)$$

$$\hat{F}j \approx \beta_0 + \sum_{i=1}^{M=14} \beta_i c_i \quad \text{for } F_2 < 1600 \text{ Hz} \qquad (2b)$$

where α_0 and β_0 are constants, α_i and β_i are the two sets of regression coefficients for the front and back MLR equations respectively, and the set of c_i's is a single set of 14 cepstral coefficients which is used in both equations.

For each formant, formant estimates derived from the front and back MLR equations are pooled and a single rms error measure is calculated for each experimental trial.

For the purpose of comparison, separate NN's are trained on samples in the front and back categories. For each of the three formants, formant estimates derived from the two NN's are pooled to calculate a single rms error measure in each experimental trial.

The PW-MLR and PW-NN models are evaluated over two experimental conditions. In the *single-speaker representation* condition, the models are trained and tested on the *partitioned* frame data of a single speaker. In the *multiple speaker prediction* condition, the models are trained on the *partitioned* data of three speakers, and tested on the *partitioned* data of a fourth speaker. The rms errors generated by the models over the four experimental trials in each condition are shown in Figure 2 overleaf. For the purpose of comparison, the rms errors produced by the MLR and NN models trained on *unpartitioned* data model are also shown in Figure 2. The rms error produced by the PW-NN model are very similar to those of the NN model and are therefore not shown in Figure 2.

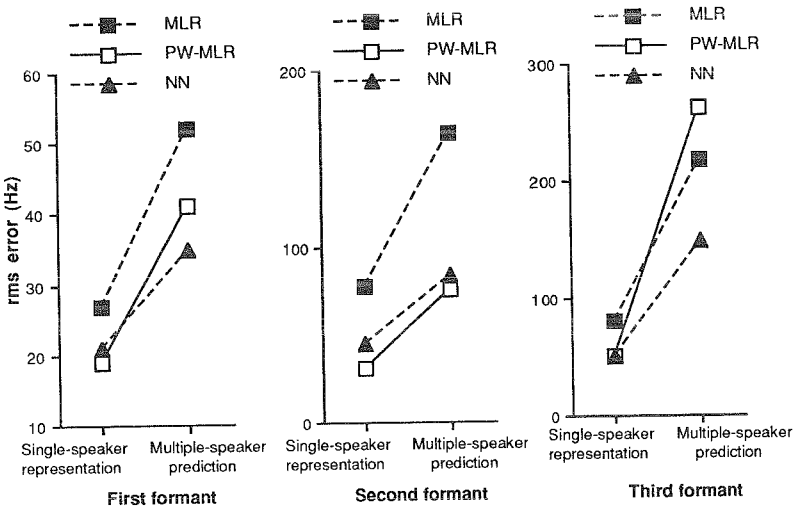

FIGURE 2. The rms errors (Hz) which result when the MLR, PW-MLR, and NN model estimate the first, second, and third formants in both a single-speaker representation condition and a multiple-speaker prediction condition.

The PW-NN model trained on data partitioned into front and back categories generates formant estimates which are only marginally more accurate than those produced by the NN trained on unpartitioned data. This result indicates that the NN trained on unpartitioned data has captured the quintessential nonlinearities in the cepstrum-to-formant relationship. Partitioning of the data into front and back categories does not improve the formant estimates produced by the PW-NN model because the NN model had already discovered the critical difference in the cepstrum-to-formant relationship for front and back vowels.

When the models are required to represent the formant data of individual speakers, the PW-MLR model generates estimates for *all* three formants which are substantially more accurate than those generated by Broad and Clermont's MLR model. This suggests that the quintessential nonlinearity in the cepstrum-to-formant relationship is indeed a distinct difference in the nature of the linear cepstrum-to-formant relationship for front and back vowels. Formant estimates generated by the PW-MLR model for all three formant are about as accurate as those generated by the NN model. This result suggests that *within* the front and back categories, the cepstrum-to-formant relationship is accurately modelled as linear.

When the models are required to perform multiple-speaker prediction, the PW-MLR model generates substantially more accurate *first and second* formant estimates than the MLR model but somewhat less accurate *third* formant estimates. In the case of the first and second formants, the nature of the piecewise-linear cepstrum-to-formant relationship is obviously very similar across different speakers. This makes it possible to use the pooled data of three speakers to derive a piecewise-linear cepstrum-to-formant relationship which will produce accurate first and second formant estimates for a fourth speaker. In the case of the *third* formant however, the nature of the piecewise-linear cepstrum-to-formant relationship appears to differ considerably from one speaker to the next. For this reason, a piecewise-linear cepstrum-to-formant function is capable of producing highly accurate third formant estimates when applied to single speakers. However a piecewise-linear cepstrum-to-formant relationship derived using the *pooled* data of three speakers will not produce accurate formant estimates for a fourth speaker. This explains why third formant estimates generated using the PW-MLR model are more accurate than those of the MLR model when the models are trained and tested on the data of a single speaker, but less accurate when the models are trained on a set of three speakers and tested on a fourth speaker.

SUMMARY AND CONCLUDING DISCUSSION

This paper has two aims. Our first aim is to determine whether the cepstrum-to-formant relationship is linear or nonlinear. Since the cepstrum-to-formant mapping generated by the NN is capable of producing substantially more accurate formant estimates than is possible with the MLR model, the cepstrum-to-formant relationship is best modelled as nonlinear.

Our second aim is to determine the precise nature of this nonlinearity. We find that the cepstrum-to-formant relationship is best described as piecewise-planar; one plane describing the linear relationship between front vowel formants and the LPC-cepstral coefficients, the other plane describing a different linear relationship between back vowel formants and the LPC-cepstral coefficients. The nature of this piecewise-planar cepstrum-to-formant relationship appears to be very similar across speakers in the case of the first and second formants. However, the nature of this piecewise-planar relationship appears to differ substantially across speakers in the case of the third formant.

We propose that the piecewise-planar cepstrum-to-formant surface be computed using a PW-MLR model with a breakpoint at that F2 frequency which best separates the front and back vowels. This PW-MLR model is capable of producing substantially more accurate formant estimates than is possible using Broad and Clermont's MLR model. The PW-MLR model produces formant estimates which are about as accurate as those generated by the NN model. The PW-MLR model is to be preferred to the NN model however. The PW-MLR model requires the estimation of parameters which have a straightforward physical interpretation. On the other hand, the nonparametric cepstrum-to-formant mapping produced by the NN is not open to any form of theoretical interpretation.

The current problem with our new PW-MLR model is that it produces somewhat inaccurate third formant estimates when the model is trained on the pooled data of three speakers and tested on a fourth speaker. Although the relationship between the LPC-cepstral coefficients and the third formant is piecewise-planar, the nature of this piecewise-planar, cepstrum-to-formant relationship appears to differ across speakers. Research is already underway to try and model speaker differences in the nature of the cepstrum-to-formant relationship for the third formant.

APPLICATIONS

The model proposed in this study is *supervised* in that *a priori* knowledge of correct formant estimates is required to establish the cepstrum-to-formant relationship. However, the relationship established through training on a set of known formants and LPC-cepstral coefficients can be used to obtain relatively accurate estimates of unknown formants. These formant estimates could be further used in an ASR system either as features, as parameters of a distance metric, or for speaker normalisation purposes.

MANY THANKS for insightful discussions with: Dr. Z. Aleksic. A.Paice, and especially to Dr. R. Williamson; for technical support from D.Andriolo, J. Elso, A. Loeff, A. McGuffin, and J. Sloan; and for supervision from Prof. R. Brent, Dr. I. MacLeod , Dr. B. Millar, and Prof. A. Tsoi.

REFERENCES

Broad, D.J. (1981) *Piecewise planar vowel formant distributions across speakers* , J.Acoust.Soc. Am. 69,1423-1429.

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

Broad, D.J. & Wakita, H. (1977) *Piecewise-planar representation of vowel formant frequencies* , J. Acoust. Soc. Am. 62, 1467-1473.

Clermont, F. (1990). *Unpublished PhD thesis*, Australian National University.

Lapedes, A.S. & Farber, R. (1987). *Nonlinear signal processing using neural networks: prediction and system modelling.* Technical Report, LA-UR-87, Los Alamos National Laboratory.

McCandless, S.S. (1974) *An algorithm for automatic formant extraction using linear prediction spectra* , IEEE Trans. Acoust. Speech Signal Process. ASSP-22, 135-141.

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