AN *AB INITIO* ANALYSIS OF RELATIONSHIPS BETWEEN CEPSTRAL AND FORMANT SPACES

Simon Hawkins, Iain Macleod and Bruce Millar

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT – Building on earlier work (Hawkins & Clermont, 1990), this paper uses simple Artificial Neural Networks to learn how to map points representing a single speaker's vowels in a 12-Dimensional cepstral space into points in a 3-D formant space. We find that an ANN with only six hidden units can learn this mapping entirely on the basis of the supplied training data. We then analyse the operation of individual hidden units to try to discover the means by which the ANN is able to map input to output points so successfully. The suggestion from this analysis is that the F1 coordinate in the output space is estimated using a linear combination of input coefficients, but that the F2 and F3 coordinates are estimated by piecewise-linear mappings. It appears that selection of one or other of alternative linear mappings is selected in estimating the latter coordinates according to whether F2 is greater or less than 1350 to 1400 Hz or so, corresponding to the traditional front/back distinction.

Having gained some unbiased hints about the possible structure of the relationship between vowel representations in cepstral and formant spaces, we then proceed to evaluate these hints by means of multiple linear regression analysis. The overall results here confirm our somewhat limited analysis of the operation of the trained ANN.

INTRODUCTION

This paper addresses the nature of the relationship between a cepstral representation of a speaker's vowel system and its representation in terms of the first three formant frequencies. Broad and Clermont (1989) have suggested that each of the three formants of a vowel can be estimated from a linear combination of the low-order LPC cepstral coefficients. This implies the existence of a linear mapping between the cepstral and formant domains. If this mapping is in fact linear, then the monophthongal vowels uttered by a single speaker should form a piecewise-planar surface in cepstral space just as they have been observed to do in formant space (Broad, 1981; Broad & Wakita, 1977). The angle between the two planes may well differ in the formant and cepstral domains, but the vowel surfaces should still be piecewise-planar in form. However, Hawkins *et al.* (1994a) have found that the shape of a speaker's vowel surface is not necessarily the same in cepstral and formant spaces. This means that the mapping between these spaces must (at least to some extent) be nonlinear. In cepstral space, a speaker's vowel surface tends to be parabolic in form and symmetrical about the main axis – the front and central vowels predominantly lie on one side of this axis and the back vowels on the other. In formant space on the other hand, the shape of the vowel surface varies from speaker to speaker. Hawkins *et al.* found that some speakers had a curved vowel surface while others had a surface that was near planar in shape.

Broad and Clermont (1989) proposed that each of the first three formants could be estimated in terms of a linear combination of the low-order LPC cepstral coefficients. Broad (Broad, 1981; Broad & Wakita, 1977) had already observed that a speaker's monophthongal vowels form a piecewise-planar surface in formant space, with front and back vowels lying on different planes. F3 of a speaker's front and back vowels thus had to be estimated using different linear combinations of F1 and F2, meaning that by extension when estimating F3 from cepstral coefficients, different linear combinations would also be required for front/back vowels. Hawkins and Clermont (1990) found some evidence for a piecewise-linear relationship between the cepstral and formant domains. The current paper extends their work with the aim of clarifying the extent to which the relationship between these domains can be regarded as intrinsically nonlinear.

A NEURAL NETWORK WHICH PERFORMS CEPSTRAL-TO-FORMANT MAPPING

To investigate this question of nonlinearity, we took advantage of an ANN's ability to learn arbitrary input/output mappings purely on the basis of training data, given only that the ANN has sufficient complexity to encode the required mapping and that the volume of training data is adequate for that complexity. We presented the ANN with 12-D LPC cepstral input data, representing speech frames from the entire vocalic nucleii of the 11 monophthongal and 8 diphthongal vowels of Australian English, as uttered in an *h*[vwl]*d* context. (Further details regarding the speech data are given in Hawkins *et al.* (1994a)). We trained the ANN to map cepstral data points into a 3-D output formant space with minimum error (in terms of average Euclidean distances from the desired output points). The ANN had a single layer of 25 hidden units with sigmoidal transfer functions between the input and output layers. Three output units with linear transfer functions estimated F1, F2 and F3. When trialed on the vowel systems of the four Australian English speakers used in the study by Broad and Clermont (1989), Hawkins and Clermont (1990) found that such an ANN had no difficulty learning to extract highly accurate estimates of F1, F2 and F3 from the cepstral representations of vowels. The ANN, which was capable of learning highly nonlinear input/output mappings, learned mappings which were superior to linear mappings derived using a least squares criterion. This result indicated that the cepstral to formant mapping had nonlinear components.

We were interested in the nature of the mapping which had been discovered by the ANN. Complex ANN architectures are notoriously difficult to analyse on the basis of their trained weights, given the nonlinearities usually present. We thus experimented with a much simpler ANN with only six hidden units, analysing the operation of each hidden unit after training. Two of the simpler ANNs were trained to perform the cepstral-to-formant mapping for single speakers (Speakers 24 and 01) not used in the previous study. The first ANN, which was trained on the vowel system of Speaker 24, was found to develop a hidden unit which apparently functioned as a "front/back" feature detector, "firing" in the presence of back vowels but being quiescent otherwise. As shown in Figure 1, this hidden unit used an F2 frequency cutoff in the region of 1350 to 1400 Hz or so to distinguish front and back vowels.

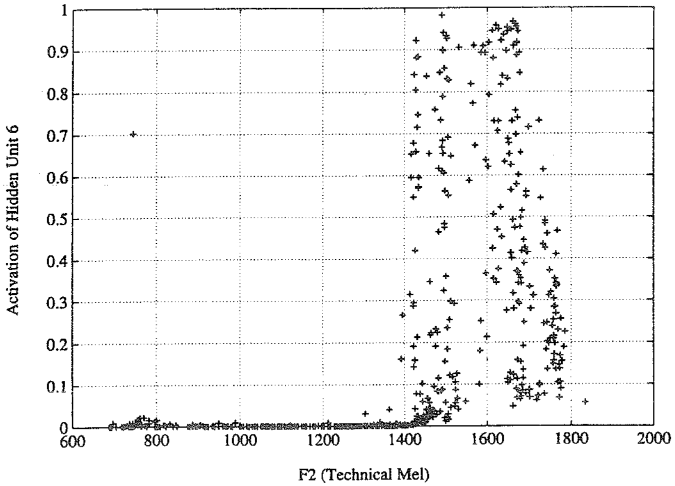

Figure 1: Operation of front/back feature detector.

We inferred that this feature detector was being used to select separate input-output mappings for the front and back vowel classes. We therefore divided the vowel system of each speaker into two classes according to whether their F2 values were above or below the value identified by the feature detectors. We then analysed the inputs to each of the remaining four hidden units (a linear combination of cepstral coefficients) in relation to the outputs from these units (ie. the activation levels after the corresponding weighted sums of inputs had been passed through the sigmoidal transfer functions). Depending on which part of the sigmoidal transfer function is being used within a given hidden unit, the output can relate to the input in either an approximately linear or highly nonlinear manner.

As one aspect of our analysis of the ANN's operation, we assessed the extent to which the three formants of the front and back vowel classes correlated with the hidden units' inputs and outputs. If there is a similarly high correlation between both the input and output of a hidden unit with one of the formant values across all vowels, then we can assume that the unit has discovered an essentially linear mapping for this formant between the distinct representations of a vowel in cepstral and formant space. If on the other hand there is a poor correlation with both the input and output of a hidden unit for a formant, then we can assume that the unit did not discover a mapping which applies to this formant across all vowels (quite possibly because it played another role in the trained ANN). In view of the hint that the ANN was treating front and back vowel classes separately, we examined correlations between formants and the hidden unit outputs separately for each class. Within the two classes, we correlated vlaues for each formant with both the input to each hidden unit and its output. If for a given formant there was a similarly large correlation with both the input and output of a hidden unit, then we can assume that the given unit has discovered a *linear* relationship between the cepstral relationship and this formant: the hidden unit must be passing its input through the central *essentially linear* portion of the sigmoidal transfer function. (Note that the fact that one hidden unit has found such a linear mapping does not necessarily mean that the overall operation of the ANN will lead to such a linear mapping between its inputs and the relevant output, given the possible interactions with other hidden units. The results of the following section tend to suggest that any interactions here are minor.)

If, on the other hand, the correlation between a hidden unit's output and a formant exceeds the correlation with its input, then we assume that the hidden unit has discovered a *nonlinear* cepstral-to-formant mapping: the hidden unit must be passing its input through either the upper or lower *nonlinear* portions of its sigmoidal transfer function.

Table 1 shows the result of the correlation analysis of the hidden unit inputs and outputs, for the ANN trained on the first speaker (Speaker 24). Hidden unit 6 appeared to function as a back/front feature detector (as described above). Hidden unit 3 appeared to have discovered a linear relationship between the cepstral representation of vowels and F1, with correlations across all vowels of 0.84 with its input and 0.85 with its output.

Hidden unit		All vowels (616)			Front vowels (258)			Back vowels (358)		
		F1	F2	F3	F1	F2	F3	F1	F2	F3
hu1	input	-0.45	0.71	0.59	-0.81	0.70	0.53	-0.40	-0.21	-0.00
	output	-0.45	0.71	0.58	0.79	0.74	0.53	-0.37	-0.24	-0.05
hu2	input	-0.68	-0.02	0.41	-0.54	0.63	0.72	-0.78	-0.74	0.34
	output	-0.24	-0.31	-0.22	-0.27	0.34	0.23	-0.32	-0.44	-0.21
hu3	input	0.84	-0.20	-0.49	0.87	-0.90	-0.69	0.86	0.81	-0.33
	output	0.85	-0.10	-0.49	0.81	-0.71	-0.66	0.92	0.79	-0.44
hu4	input	0.37	-0.45	-0.64	0.31	-0.34	-0.71	0.55	0.49	-0.18
	output	0.31	-0.56	-0.69	0.35	-0.46	-0.75	0.44	0.38	-0.23
hu5	input	-0.45	0.67	0.47	-0.94	0.82	0.59	-0.05	0.03	-0.22
	output	-0.56	0.56	0.41	-0.83	0.82	0.45	-0.33	-0.19	-0.11
hu6	input	-0.25	0.71	0.33	-0.25	0.41	0.21	-0.36	-0.49	-0.30
	output	-0.18	0.34	0.23	-0.20	0.32	0.18	-0.25	-0.39	-0.29

Table 1: Correlations between (i) the inputs and (ii) the activation levels (outputs) of the six hidden units (hu1–hu6) of the ANN and the formants across all vowels uttered by Speaker 24.

For F2 and F3, however, we found evidence that the ANN was applying separate linear cepstral-to-formant mappings for the front and back vowels. For F2, hidden unit 5 gave correlations across all front vowels of 0.82 for both its input and output. This unit is therefore using a strongly linear mapping to extract F2 from the cepstral representation of front vowels. Again for F2, hidden unit 3 gave correlations across all back vowels of 0.81 with its input and 0.79 with its output. This suggests that this hidden unit is using a strongly linear mapping to extract F2 of the back vowels from their cepstral representations. (Note that hidden unit 3 appears to serve two functions, mapping from cepstral space to F1 for all vowels and from cepstral space to F2 for the back vowels.) For F3, hidden unit 4 gave correlations across the front vowels of 0.71 with its input and 0.75 with its output. This hidden unit appears to have learnt a moderatley strong linear mapping which extracts F3 of the front vowels from their cepstral representations. With back vowels, we found no correlation of any magnitude between F3 and the activity of any hidden unit.

The results with the other ANN, trained with Speaker 01, were comparable. We found evidence that hidden unit 2 in the ANN had discovered a single linear cepstral-to-F1 mapping for all vowels. For F1 there were strong correlations of 0.91 with its input and 0.85 with its output. For F2 and hidden unit 1, we found strong correlations across the front vowels of 0.83 with its input and 0.76 with its output. With F3 and hidden unit 5, there were similarly strong correlations across the front vowels of 0.74 with its input and 0.84 with its output. For F2 and hidden unit 3, there were large correlations across the back vowels of 0.92 with its output and 0.92 with its input. As with the first ANN, we found no evidence that the second ANN had been able to discover a linear mapping which extracted F3 of the back vowels from their cepstral representations.

Summarising the results of our experiments with simplified ANNs, it appears that F1 can be estimated from the cepstral representations of all vowels using a predominantly linear mapping. Separate linear mappings were used for front and back vowels when estimating F2. The ANNs discovered a single linear mapping for each speaker with which to estimate F3 for the front vowels, but no such mapping was found for F3 and the back vowels. The relationship between the cepstral representation of a vowel and F3 would thus appear to be rather complex and nonlinear.

Having gained certain data-driven hints from our ANNs about the nature of relationships between cepstral and formant space representations of different vowels, we now proceed to investigate the suggested relationships in a systematic fashion, using vowel data from our 14 male speakers of Australian English.

EVALUATION OF PIECEWISE-LINEAR MAPPINGS VIA LINEAR REGRESSION

Two strong suggestions from our ANN studies were that the relationships between cepstral and formant space representations of a vowel could be approximated in either a linear or piecewise-linear fashion, depending on the formant and the class of vowels, and that in the piecewise-linear case the traditional back versus front vowel distinction formed an appropriate division into two classes (each with a linear mapping). We evaluated these suggestions by assessing the relative errors when estimating each formant, using both linear and piecewise-linear mappings. In the latter case, we tried a range of possible divisions of the vowel space into two classes with linear mappings to assess the utility of the back/front distinction.

The "pivot" of a speaker's vowel distribution in the F1-F2 plane is defined here as the intersection of the line separating the speaker's front and back vowels with the line separating their high and low vowels. We rotate a line about this pivot point to separate the speaker's vowels into various groups. At certain angles this separation will correspond to traditional divisions such as high versus low, back versus front or rounded versus unrounded. A line through the pivot of a vowel distribution provides a simple but effective means for splitting the vowel distribution into two phonemic classes. To find which split provided the best basis for performing a piecewise-linear mapping between cepstral and formant spaces, we rotated this "hinge" line full circle about the pivot point in one degree increments. At each increment, we determined the best overall mapping between 12-D cepstral space and 3-D formant space which could be obtained by applying separate linear mappings to vowels on either side of the hinge line. The overall accuracy of the mapping was measured by the average Euclidean distance between the actual location of vowels in 3-D formant space and the location to which they were mapped. Table 2 reports the results of these and the following experiments.

In the light of Broad and Clermont's (1989) results, we also examined the extent to which a single linear mapping could be used to map vowels between cepstral and formant spaces. A piecewise-linear mapping achieved an average error (in Hz) which was only 60% of that achieved with a linear mapping. On an individual speaker basis, substantially better fits with piecewise-linear mappings were evident except with Speaker 4, Speaker 10 and Speaker 11. For all but one of our fourteen speakers, we found that the best piecewise-linear mapping was achieved by applying separate linear mappings to the speaker's front and back vowels. In the case of Speaker 13, however, the best piecewise-linear mapping involved separate linear mappings for rounded and unrounded vowels. (Note that in Australian English, all front vowels are unrounded and all but one of the back vowels are rounded.)

The need to use a piecewise-linear cepstral-to-formant mapping indicates that the shape of a speaker's vowel surface in the cepstral and formant spaces is different. Inspection of the surfaces (Hawkins *et al.* 1994a) shows that a primary difference in shape lies in the degree of symmetry of the surfaces about their main axis. In cepstral space, a speaker's vowel surface is symmetrical about the main axis. In formant space, it tends to be highly asymmetrical: the speaker's front and central vowels occupy a surface with much more curvature away from the main axis than the speaker's back vowels. This difference in symmetry about the main axis explains the need to "split" the surface down the main axis when mapping between the two spaces.

Spkr. Nmbr	Error mapping vowels from C-to-F space (Hz)			Correlations between actual formants & those estimated by C-to-F mapping						Hinge-line splitting vowels for PWL mapping		
				$r(F_1, \hat{F}_1)$		$r(F_2, \hat{F}_2)$		$r(F_3, \hat{F}_3)$		pivot	angle	splits
	lin.	PWL	Δ	lin.	PWL	lin.	PWL	lin.	PWL	F1,F2 (Hz)	to F1	feature
01	262	122	140	.970	.986	.903	.994	.793	.901	577,1438	-30	front-back
04	129	92	37	.962	.975	.973	.991	.874	.905	505,1357	+50	front-back
10	127	75	52	.957	.984	.978	.994	.912	.962	498,1400	-30	front-back
11	101	60	41	.967	.979	.982	.996	.891	.947	559,1437	+0	front-back
12	216	131	85	.955	.975	.958	.994	.757	.900	525,1359	+30	front-back
13	271	162	109	.942	.976	.912	.978	.796	.915	512,1460	-70	rounding
19	155	79	76	.948	.974	.955	.993	.916	.960	491,1444	+10	front-back
20	202	99	103	.912	.955	.931	.990	.849	.940	483,1452	-50	front-back
24	255	154	101	.959	.971	.922	.987	.750	.853	479,1477	-30	front-back
30	234	135	99	.908	.931	.912	.986	.848	.883	485,1458	-20	front-back
31	174	105	69	.949	.977	.964	.994	.801	.994	507,1502	+10	front-back
34	245	178	67	.929	.943	.936	.988	.711	.751	491,1540	-10	front-back
35	197	125	72	.967	.980	.965	.991	.786	.900	475,1300	+10	front-back
36	148	92	56	.957	.967	.961	.988	.896	.935	478,1501	-10	front-back
Mean	194	114	79	.949	.970	.947	.990	.827	.910	504,1437	-10	13 backness 1 rounding

Table 2: Mapping the vowels of 14 male speakers from 12-D cepstral space into 3-D formant space

A piecewise-linear mapping will in general be more accurate than a single linear mapping. This begs the question as to whether this advantage of piecewise-linear mapping applies equally to all three formants in formant space. To investigate this issue, we correlated each of the formants of a vowel with the formant frequencies obtained via two types of mapping: a piecewise-linear mapping in which separate linear mappings were applied to the front and back vowels, and a single linear mapping which was applied to all vowels. With F1 and F2, use of a piecewise-linear mapping gave only a marginal improvement over a single linear mapping from 12-D cepstral space into the F1-F2 plane. With a single linear mapping, the correlations between actual and estimated F1 and F2 were both 0.95. This figure improved only marginally with a piecewise-linear mapping to 0.97 for F1 and 0.99 for F2 (the results with the single linear mapping were so good that little further improvement was possible). On the other hand, the correlation between the actual and estimated F3 of a speaker's vowels was 0.83 across 14 speakers with a single linear mapping. Using a piecewise-linear mapping, the improvement to 0.91 obtained with F3 was relatively much greater than tah obtained with F1 and F2.

CONCLUSIONS

Our trained ANNs strongly suggested that F1 could be estimated across all vowels by a linear mapping from cepstral to formant space. They also indicated that in estimating F2 and F3, the best results were obtained with a piecewise-linear mapping comprising two separate linear mappings. The remarkable result was that in a purely data-driven manner, the ANNs appeared to choose one or other of the linear mappings on the basis of whether F2 was greater or less than 1350 to 1400 Hz or so, corresponding to the traditional front/back distinction!

On using correlation analysis to systematically investigate the nature of the relationships between cepstral and formant space suggested by the ANNs, we found the following results across our 14 speakers (as shown in Table 2):

- In comparisom with a linear mapping, a piecewise-linear mapping gave substantially better accuracy when estimating vowel locations in 3-D formant space from their 12-D cepstral representations.
- When estimating F1 there was very little improvement obtained with a piecewise-linear mapping compared with a linear mapping. We therefore conclude that the relationship between cepstral and formant space is essentially linear for F1.
- When estimating F2 there were more-noticeable improvements when using a piecewise-linear mapping, but these were still small enough for us to conclude that the relationship between cepstral and formant space for F2 was approximately linear.
- With F3, however, there was a substantial improvement obtained through use of a piecewise-linear

mapping. We conclude that in the case of F3, the relationship between cepstral and formant spaces is basically non-linear.

The finding through our detailed analysis that the relationship between cepstral and formant spaces is approximately linear in the case of F1 and F2 opens up the possibility of extracting a plane approximately corresponding to F1 and F2 from the 12-D cepstral representation of a speaker's vowel system. An unsupervised algorithm for the extraction of such a plane is developed in Hawkins *et al.* (1994b). The non-linear relationship between F3 and the cepstral representation precludes the possibility of extracting accurate F3 estimates using linear transformation methods.

REFERENCES

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

Broad, D. J. and Clermont, F. (1989), "Formant estimation by linear transformation of the LPC cepstrum," ***Jnl. Acoust. Soc. Am.,*** Vol. 86, 2013–2017.

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

Hawkins, S. and Clermont, F. (1990), "Supervised cepstrum-to-formant estimation: A new piecewise-linear model," ***Proc. Int. Conf. on Speech Science and Technology***, Melbourne, 310–315.

Hawkins, S., Macleod, I. and Millar, B. (1994a), "Modelling individual speaker characteristics by describing a speaker's vowel distribution in articulatory, cepstral and formant space," ***Proc. Int. Conf. on Speech Science and Technology***, Perth (this proceedings).

Hawkins, S., Macleod, I. and Millar, B. (1994b), "An unsupervised algorithm for the extraction of formant-like features from LPC cepstral space," ***Proc. Int. Conf. on Speech Science and Technology***, Perth (this proceedings).