EVALUATION OF THE ROBUSTNESS OF PERCEPTUAL LINEAR PREDICTION ANALYSIS USING MULTI-SPEAKER AUSTRALIAN ENGLISH VOWEL DATA

J.Bruce Millar and Xue Yang

Computer Sciences Laboratory
Research School of Physical Sciences
Australian National University

ABSTRACT - Perceptual linear prediction analysis is a recently developed extension of standard linear predictive analysis which takes into account the characteristics of human hearing. It has been shown that its application to American English data presented to a simple automatic speech recognition system improves the speaker-independent performance of that system. In our studies, we use multi-speaker Australian English vowel data to evaluate the relative sensitivity to speaker difference and to phonetic difference of perceptual linear prediction analysis and standard linear prediction analysis. This work extends that of Hermansky by its detailed analysis of the vowel space for a moderately large number of speakers.

INTRODUCTION

Hermansky (1990) draws together the results of a number of experiments that he has reported over recent years in which he evaluates the benefits of modifying the standard form of linear prediction analysis in order to make it sensitive to the characteristics of human hearing. The information resulting from this analysis would then correspond more closely to that which is available to the human speech decoding system rather than simply a linear transform of the signal measured by a microphone. The a priori assumptions of this approach include the expectation that speech information presented in this form will enable automatic speech recognition systems to focus on the features of speech acoustics that are similar to those which enable the superior performance of the human perceptual system.

Hermansky evaluated his perceptual linear prediction (PLP) and the standard linear prediction (LP) methods as preprocessing for an automatic speech recognition (ASR) system for a corpus of keyboard character names spoken by two sets of four speakers. Results were reported in terms of percent correct, yielding a maximum value when 5th order PLP was used. He also reported the strong correlations between the performance of the PLP representation and several aspects of human performance in the perception of vowels. We were therefore intrigued to discover how low order PLP analysis manages speaker and phonetic information in the vowel space, and to apply it to Australian English vowels.

In this study Australian English monophthongs from 33 speakers are analysed using both PLP and LP analysis in order to compare the robustness of these methods to speaker variability across the vowel space. Accordingly an objective function, designed to reflect the relative contributions of speaker variance and phonetic variance, is evaluated for each vowel at a number of different orders of all-pole analysis, and for different distance measures in the cepstral domain.

PERCEPTUALLY BASED LINEAR PREDICTION ANALYSIS

PLP analysis was first published by Hermansky (1985). It calculates the spectrum that is assumed to be available to natural speech perception by implementing observed psychoacoustic transforms relating to frequency resolution and the dependence of loudness on frequency and intensity. It then models this assumed "auditory spectrum" by a low-order all-pole model. Specifically, the auditory spectrum, scaled in Bark, is derived from the speech waveform by filtering (using 18 critical-bands in the 0-5kHz range), followed by equal-loudness preemphasis and cubic-root amplitude compression. A set of autoregressive coefficients are derived from the auditory spectrum using autocorrelation LP techniques.

SPEECH CORPUS AND ANALYSIS

The speech corpus for our evaluation comprised the speech of 15 male and 18 female Australian English speakers, each of whom produced one example of 11 monophthong hVd words (Millar et al, 1989). The most energetic 40ms of each utterance was extracted for analysis in the experiments. The 363 extracted segments were processed using the PLP algorithm operating on 20ms windows with a 50% overlap, thus generating three auditory spectral frames from each 40ms segment. Each

of these segments was represented by one set of coefficients derived by averaging the three auditory spectral frames in the autocorrelation domain. The averaged spectral frame was then expressed as a set of autoregressive coefficients which were then transformed into low-order cepstral coefficients by a recursive calculation. In this way 363 vectors were formed in the cepstral domain representing the 363 utterances.

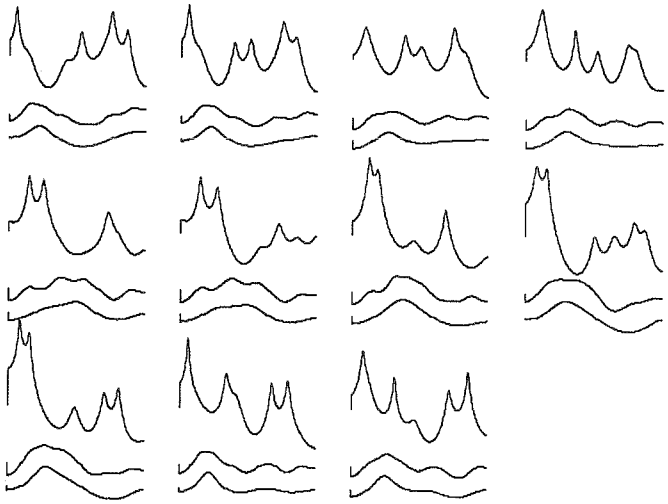

Figure 1. Comparative plots of 11 vowels (left to right, then top to bottom in the order of table 1) of one male speaker where upper trace is standard 14th order LP on a Hertz scale, the middle trace is 14th order PLP on a bark scale, and the bottom trace is 5th order PLP on a bark scale.

CRITERIA FOR EVALUATION

Objective function

The objective function relating to the phonetic and speaker related information in the vowel representations is defined by determining distances in the N-dimensional space of the 363 analysis vectors.

For vowel j, the centroid of the space occupied by the analysis vectors for all the speakers is established,

$$\vec{C}_j = \frac{1}{n}\sum_{i=1}^{n}\vec{c}_{ji} \quad j = 1, 2, \ldots, 11$$

where $\vec{c}_{ji}$ is the vector in N^{th} order cepstral space representing vowel j spoken by speaker i, and n is the total number of speakers. The mean distance from each vector representing the utterance of vowel j by one speaker to the centroid of vowel k for all speakers is defined as,

$$D_{jk} = \frac{1}{n}\sum_{i=1}^{n} d(\vec{c}_{ji}, \vec{C}_k) \quad j,k = 1,2,...,11$$

where d is a distance measure in the N^{th} order cepstral space. Speaker related information is measured for each of the eleven vowels by setting k=j. D_{jk} then becomes the "inter-speaker distance" for vowel j,

$$ISD_j = D_{jk} \quad k = j; \quad j = 1,2,...,11$$

If k ≠ j, D_{jk} is called an "inter-vowel distance" of vowel j, and is roughly equivalent to the distance between centroids of the vowels j and k.

$$IVD_{jk} = D_{jk} \quad k \neq j; \quad j = 1,2,...,11$$

The objective function for vowel j is then expressed as the ratio of its smallest inter-vowel distance to its inter-speaker distance.

$$MIVD_j = MIN(\, IVD_{jk}, \quad 1 \leq k \leq 11;\ k \neq j)$$

$$OF_j = \frac{MIVD_j}{ISD_j}$$

This objective function measures the ratio of the separation of each vowel from its nearest neighbour, which is the critical distance in the vowel space on which to monitor potential confusion, and the degree of speaker variation with the vowel itself. Each evaluation of the objective function is therefore specific to a particular vowel pair which are nearest neighbours in the vowel space. This objective function should be large for a representation of speech which maximises phonetic information and minimises speaker variance. In order to achieve this the spectral representation should not only be similar for utterances with different acoustic qualities but with identical phonetic quality, but also different for the utterances with different phonetic quality.

Distance measurement

Distance measurements d were made in the cepstral domain. A 'standard cepstral distance' gives equal weight to each cepstral dimension in a Euclidean space. An 'index-weighted cepstral distance' gives increasing weight to cepstral dimensions of increasing order in a Euclidean space. This is the 'group delay' metric (Yegnanarayana and Reddy, 1979) used by Hermansky (1990). The index-weighted cepstral distance is designed to give greater weight to component spectral patterns of greater spectral resolution, and consequently to reduce the weight given to low spectral resolution features such as spectral tilt.

EXPERIMENTS

Evaluation 1

The first evaluation examined the effect of the PLP and LP models, over a range of orders N, on the inter-speaker distances using the standard cepstral distance measure. Table 1 shows that the inter-speaker distances measured using LP analysis are an order of magnitude greater than those measured using PLP analysis. This indicates that the similarity between the PLP-modelled auditory spectra from different speakers is much greater than that of the LP- modelled power spectra. With increasing order, the inter-speaker distances with both models increase. This implies that as more spectral detail is modelled, more speaker variance can be included in the model. It is also noticed that as the analysis order is increased the percentage increase for LP-analysis is much greater than for PLP-analysis. This implies that the PLP representation is intrinsically less susceptible to speaker-related variance than is the LP representation.

	analysis =	5th-PLP	5th-LP	7th-PLP	7th-LP	10th-PLP	10th-LP	14th-PLP	14th-LP
1.	/ i /	.046	.41	.049	.51	.053	.71	.056	.78
2.	/ ɪ /	.037	.38	.041	.44	.046	.65	.049	.74
3.	/ ɛ /	.036	.33	.044	.43	.048	.59	.051	.68
4.	/ æ /	.037	.34	.044	.47	.050	.61	.051	.71
5.	/ ɑ /	.039	.36	.045	.44	.048	.53	.050	.63
6.	/ ʌ /	.037	.29	.042	.37	.045	.49	.047	.59
7.	/ ɒ /	.031	.31	.036	.40	.042	.50	.045	.59
8.	/ ɔ /	.026	.27	.032	.36	.035	.41	.038	.50
9.	/ ɷ /	.027	.29	.032	.36	.036	.42	.038	.47
10.	/ u /	.031	.30	.035	.35	.039	.48	.042	.57
11.	/ ɜ /	.032	.30	.036	.36	.040	.51	.041	.59
	mean	.034	.32	.040	.41	.044	.54	.046	.62

Table 1. The inter-speaker distances

Evaluation 2

The second evaluation examined the effect of the style of processing (PLP or LP), the order of the cepstral representation (4,5,6,7,8,10,12,14), and the distance measure (standard, index-weighted) on the objective function. Tables 2 to 5 supply the results.

	order =	4th	5th	6th	7th	8th	10th	12th	14th
1.	/ i /-/ ɪ /	1.03	1.02	1.02	1.02	1.02	1.02	1.02	1.02
2.	/ ɪ /-/ i /	1.04	1.03	1.03	1.03	1.03	1.03	1.03	1.03
3.	/ ɛ /-/ ɪ /	1.79	1.72	1.62	1.61	1.58	1.56	1.53	1.53
4.	/ æ /-/ ʌ /	1.99	2.07	1.94	2.09	2.08	2.01	2.00	2.00
5.	/ ɑ /-/ ʌ /	1.07	1.09	1.07	1.07	1.07	1.07	1.07	1.07
6.	/ ʌ /-/ ɑ /	1.07	1.07	1.07	1.08	1.08	1.07	1.08	1.07
7.	/ ɒ /-/ ʌ /	1.87	1.96	2.00	2.00	2.02	1.99	1.98	1.94
8.	/ ɔ /-/ ɷ /	1.19	1.23	1.29	1.27	1.27	1.29	1.29	1.28
9.	/ ɷ /-/ ɔ /	1.19	1.22	1.28	1.27	1.26	1.28	1.28	1.27
10.	/ u /-/ ɜ /	1.95	1.87	1.82	1.78	1.71	1.66	1.65	1.62
11.	/ ɜ /-/ u /	1.87	1.84	1.80	1.76	1.69	1.65	1.64	1.63
	mean	1.46	1.46	1.45	1.45	1.44	1.42	1.41	1.40

Table 2. The objective function for PLP model with standard cepstral metric

	order =	4th	5th	6th	7th	8th	10th	12th	14th
1.	/ i /-/ ɪ /	1.02	1.01	1.02	1.02	1.02	1.02	1.02	1.02
2.	/ ɪ /-/ i /	1.02	1.01	1.03	1.02	1.02	1.02	1.02	1.02
3.	/ ɛ /-/ ɪ /	1.18	1.26	1.29	1.32	1.32	1.30	1.29	1.27
4.	/ æ /-/ ʌ /	1.35	1.39	1.47	1.49	1.45	1.55	1.54	1.55
5.	/ ɑ /-/ ʌ /	1.05	1.03	1.04	1.03	1.04	1.04	1.06	1.06
6.	/ ʌ /-/ ɑ /	1.06	1.04	1.05	1.04	1.04	1.05	1.06	1.06
7.	/ ɒ /-/ ʌ /	1.24	1.42	1.47	1.51	1.57	1.66	1.59	1.58
8.	/ ɔ /-/ ɷ /	1.06	1.11	1.10	1.10	1.12	1.15	1.16	1.19
9.	/ ɷ /-/ ɔ /	1.06	1.10	1.10	1.10	1.13	1.15	1.16	1.20
10.	/ u /-/ ɜ /	1.12	1.15	1.19	1.18	1.23	1.27	1.28	1.28
11.	/ ɜ /-/ u /	1.12	1.15	1.17	1.17	1.22	1.25	1.27	1.27
	mean	1.12	1.15	1.17	1.18	1.20	1.22	1.22	1.23

Table 3. The objective function for LP model with standard cepstral metric

	order=	4th	5th	6th	7th	8th	10th	12th	14th
1.	/ i /-/ ɪ /	1.02	1.02	1.02	1.02	1.02	1.02	1.01	1.01
2.	/ ɪ /-/ i /	1.03	1.02	1.02	1.02	1.02	1.01	1.01	1.01
3.	/ ɛ /-/ ɪ /	1.68	1.70	1.64	1.56	1.45	1.39	1.34	1.31
4.	/ æ /-/ ʌ /	2.25	2.42	1.91	2.02	2.05	1.90	1.84	1.78
5.	/ ɑ /-/ ʌ /	1.13	1.12	1.11	1.10	1.09	1.07	1.07	1.06
6.	/ ʌ /-/ ɑ /	1.14	1.15	1.13	1.11	1.11	1.08	1.08	1.07
7.	/ ɒ /-/ ʌ /	1.84	1.82	1.98	2.04	2.09	2.00	1.99	1.86
8.	/ ɔ /-/ ɷ /	1.33	1.55	1.62	1.48	1.51	1.50	1.44	1.41
9.	/ ɷ /-/ ɔ /	1.41	1.62	1.64	1.53	1.53	1.52	1.48	1.41
10.	/ u /-/ ɜ /	2.07	1.97	2.02	1.91	1.67	1.53	1.52	1.43
11.	/ ɜ /-/ u /	1.92	1.87	1.96	1.78	1.62	1.52	1.50	1.45
	mean	1.53	1.57	1.55	1.51	1.47	1.41	1.39	1.34

Table 4. The objective function for PLP model with index-weighted cepstral metric

	order=	4th	5th	6th	7th	8th	10th	12th	14th
1.	/ i /-/ ɪ /	1.01	1.01	1.02	1.01	1.03	1.04	1.03	1.03
2.	/ ɪ /-/ i /	1.02	1.01	1.01	1.01	1.03	1.04	1.03	1.03
3.	/ ɛ /-/ ɪ /	1.13	1.13	1.18	1.32	1.31	1.28	1.23	1.20
4.	/ æ /-/ ʌ /	1.57	1.55	1.77	1.92	1.93	1.67	1.60	1.51
5.	/ ɑ /-/ ʌ /	1.08	1.04	1.04	1.05	1.08	1.09	1.12	1.09
6.	/ ʌ /-/ ɑ /	1.09	1.05	1.04	1.06	1.09	1.09	1.12	1.09
7.	/ ɒ /-/ ʌ /	1.44	1.59	1.62	1.70	1.65	1.80	1.65	1.58
8.	/ ɔ /-/ ɷ /	1.07	1.12	1.09	1.16	1.32	1.38	1.35	1.46
9.	/ ɷ /-/ ɔ /	1.08	1.13	1.09	1.15	1.31	1.37	1.35	1.51
10.	/ u /-/ ɜ /	1.36	1.35	1.38	1.30	1.42	1.47	1.31	1.30
11.	/ ɜ /-/ u /	1.42	1.41	1.39	1.35	1.42	1.36	1.29	1.30
	mean	1.21	1.22	1.24	1.27	1.33	1.33	1.28	1.28

Table 5. The objective function for LP model with index-weighted cepstral metric

The objective function values using PLP analysis for all vowels at all orders of analysis and for both standard and index-weighted cepstral distance measures were greater than or equal to the equivalent objective functions values derived using LP analysis. The benefit of PLP over LP analysis is very uneven across the vowel space. For vowels / i /, / ɪ /, / ɑ /, / ʌ /, the benefit was non-existent or very weak, whereas for / ɛ /, / æ /, / ɒ /, / u / and / ɜ / it was strong.

Examination of the variation of the objective function with the order of the analysis showed that, averaged across all vowels, PLP analysis tends to give higher values at low orders, whereas LP analysis tends to give higher values at high orders. However no significant result can be obtained for an optimum order for each method owing to the large degree of individual vowel variation.

There was a slight advantage shown in vowel-averaged objective function values when the index-weighted cepstral distance measure replaced the standard cepstral distance measure, but it was statistically non-significant. This was again due to the large degree of individual vowel variation.

Examination of the mean values of the objective function the largest value is obtained with a 5th order PLP analysis using the index-weighted cepstral distance. This result agrees with the conclusions of Hermansky (1990).

The mean of the objective function values for PLP analysis using both distance metrics decrease with the model order. From Table 1, 2 and 4, we can say that the finer details of the auditory spectra, modeled by additional poles of the higher-order PLP models carry more speaker-dependent information.

Evaluation 3

The observation that changing from standard to index-weighted cepstral distance did not uniformly increase the objective function for each vowel prompted a more detailed examination of the weighting

of the cepstral space. Both the standard and the index-weighted cepstral measures are special cases of general liftering in cepstral domain.

$$d = \sum_{i=1}^{p} (i^s C_{Ri} - i^s C_{ri})^2$$

where S($\geq$0) is a variable coefficient and p is the number of cepstral coefficients. When S=0, it is the case of the standard cepstral measure, and when S=1, the case of the index-weighted measure.

In order to examine the effect of S on the objective function, the 5th order PLP model was taken as an example, and S was varied from 0 to 2 so that higher resolution details of the spectrum were gradually enhanced and the spectral slope was suppressed in the model.

	S=	0.0	0.3	0.5	0.8	1.0	1.2	1.5	2.0
1.	/ i /-/ ɪ /	1.02	1.02	1.02	1.02	1.02	1.02	1.01	1.01
2.	/ ɪ /-/ i /	1.03	1.03	1.02	1.02	1.02	1.02	1.01	1.01
3.	/ ɛ /-/ ɪ /	1.72	1.79	1.79	1.74	1.70	1.65	1.59	1.54
4.	/ æ /-/ ʌ /	2.07	2.21	2.29	2.38	2.42	2.46	2.49	2.52
5.	/ ɑ /-/ ʌ /	1.09	1.10	1.11	1.12	1.12	1.12	1.12	1.12
6.	/ ʌ /-/ ɑ /	1.07	1.10	1.12	1.14	1.15	1.16	1.17	1.18
7.	/ ɒ /-/ ʌ /	1.96	1.94	1.92	1.87	1.82	1.76	1.67	1.51
8.	/ ɔ /-/ ɷ /	1.23	1.33	1.40	1.49	1.55	1.61	1.69	1.81
9.	/ ɷ /-/ ɔ /	1.22	1.33	1.42	1.54	1.64	1.69	1.80	1.96
10.	/ u /-/ ɜ /	1.87	1.98	2.01	2.00	1.97	1.94	1.89	1.83
11.	/ ɜ /-/ u /	1.84	1.91	1.92	1.90	1.87	1.84	1.80	1.76
	mean	1.46	1.52	1.55	1.56	1.571	1.570	1.567	1.568

Table 6. Objective function for 5th PLP model with variance of S from 0 to 2

From Table 6, we can see that the objective function for individual vowels doesn't always increase with the coefficient S. The values of objective function for vowel / i /, / ɪ /, / ɒ / decrease with S, while the values of objective function for vowel / æ /, / ʌ /, / ɔ /, / ɷ / increase with S. Enhancing spectral peaks of the PLP model is not always beneficial to the objective function of each vowel. An optimal weighting in the cepstral domain needs to be found to increase the objective function for each phonetic value. In terms of the mean of the objective function in Table 6, the best option is S=1.

CONCLUSION

In our experiments, we have observed the advantage of approximating the auditory spectrum instead of the power spectrum in the extraction of phonetic information from vowels at the expense of speaker information. This advantage is however unevenly spread across the vowel space. There is also advantage in applying different weights to the dimensions of the cepstral domain, but the benefit of this is also unevenly spread across the vowel space. Simple linear index-weighting seems to give the best result. The results of Hermansky (1990) are confirmed by these studies and are extended by the evidence of considerable vowel dependence measured over 33 speakers.

REFERENCES

Hermansky, H., Hanson,B.A., Wakita,H. (1985) *Low-dimensional representation of vowels based on all-pole modeling in the psychophysical domain",* Speech Communication 4, 181-187.

Hermansky, H. (1990) *Perceptual linear predictive (PLP) analysis of speech* J. Acoust. Soc. Am. 87 (4) 1738-1752

Millar,J.B., O'Kane,M., Bryant,P. (1989) *Design, collection, and description of a database of spoken Australian English* Australian Journal of Linguistics, 9, pp.165-189

Yegnanarayana,B., Reddy,R. (1979) *A distance measure derived from the first derivative of the linear prediction phase spectra* Proc ICASSP-79 pp.744-747.