# SPECTRAL DISTORTION AND SPECTRAL DISTANCE MEASURES

R. H. Mannell
Speech, Hearing and Language Research Centre
Macquarie University

ABSTRACT - Several different spectral distance measures have been compared in order to see which measures most closely correlate with the intelligibility of speech systematically distorted by various channel vocoder configurations.

## INTRODUCTION

Algorithms that measure the acoustic difference between two signals are of interest in the areas of speech processing and speech recognition. The comparison may be between the smoothed spectral envelopes of the test and reference items (Gray & Markel, 1976) or they can be made directly on other parameters such as LPC coefficients or LPC-derived cepstral coefficients (Atal, 1974, Ikatura, 1975, Gray & Markel, 1976, Tohkura, 1987).

Most attempts at the comparison of the performance of spectral distance measures have entailed comparing the success rates of the speech recognition systems containing them. Klatt (1982), on the other hand, examined the ability of auditorily-weighted (critical band) spectral distance measures to predict differences in the human perception of various spectrally distorted synthetic speech tokens. He varied such parameters as formant amplitudes, spectral tilt and formant centre frequencies and then examined the effect on listener identifications. This enabled him to rank the perceptual effects of the spectral distortions.

The present study examines several smoothed log spectra distance metrics. The distance measures compare the input natural speech tokens with tokens output from various digitally simulated channel vocoder configurations which have been designed to systematically distort the input speech in various ways. The input natural speech and output synthetic speech are precisely time aligned and so the distance measures can be compared without the confounding influence of other aspects of a speech recognition system such as time warping. It also has the advantage over Klatt's experiment in that distorted synthetic speech can be directly compared to natural speech rather than to reference synthetic speech.

The present experiment experiment is similar, in some ways, to a study by Bladon and Lindblom (1981), however they examined the correlation between a different (but overlapping) set of spectral distance measures with measures of perceived vowel quality differences whilst this study examines the correlation between spectral distance and vowel and consonant intelligibility.

## METHODOLOGY

All of the test material was produced from natural speech using the Speech, Hearing and Language Research Centre channel vocoder designed by the author and colleagues (Clark, Mannell & Ostry, 1987, Clark & Mannell, 1988) and the intelligibility results used in this study are reported on in the above papers. The test speech comprised of a set of 30 nonsense syllables (11 /h_d/ vowels and 19 CV consonants) uttered by a single male speaker of Australian English. The speech was systematically distorted in the frequency domain by varying the filter bandwidths of the channel filters and all filters in each filterbank had the same bandwidth in either Hertz (non-auditory) or Bark (auditory) scales. There were 14 different vocoder filterbank configurations used in this study having bandwidths of 100 Hz, 200 Hz, 400 Hz, 800 Hz and 0.75 Bark, 1.0 Bark, 1.5 Bark, 2.0 Bark and 3.0 Bark. The Bark-scale vocoders were of two types. One type simply output the synthetic speech without adjusting amplitude for the greater bandwidth of the higher filters (uncorrected), whilst the other type corrected each filter's output amplitude by a factor proportional to the ratio of the base-bandwidth over that channel's bandwidth. This provided two sets of auditorily

spaced vocoder outputs with differing spectral slopes. The output of the uncorrected vocoders had, effectively, a high frequency post- emphasis. All vocoder configurations considered in the current experiment had the same time resolution (10 msecs) as defined by the bandwidth of the system's filters.

Two types of Cepstral smoothing were utilised. The first method was the familiar cepstral smoothing algorithm in which the cepstrum is liftered somewhere below the first rahmonic and then an FFT is obtained giving a smoothed log spectrum. This method has the disadvantage of modeling the average spectral level of the original FFT rather than modeling the peak levels as does the LPC. The LPC is a model of the vocal tract filter plus the source and radiation slope characteristics whilst the normal cepstrally smoothed spectrum is a closer model of the vocal tract filter alone. It is desirable to examine the interaction between the slope characteristics of the speech signal and intelligibility as well as the effects of the vocal tract filter function and so a method that retains both characteristics of the speech signal would be desirable. The LPC does this of course, but, being an all-pole model, it cannot model the spectral zeroes accurately (but see Markel & Gray, 1976, pp 271-275, for a discussion of LPC derived pole-zero estimations). There would seem to be an advantage in an FFT based method which could model both spectral zeroes and also the source and radiation characteristics of the speech signal. For this reason an "improved" cepstral smoothing method was devised. This method estimates the pitch from the cepstrum's first rahmonic (if the signal is unvoiced the pitch is set nominally to 100 Hz). The log spectrum is then divided into equal bands each F0 wide and the highest peak in each band is identified. Spectral points between these points are given new values by interpolation. An inverse FFT is then performed on this now partially-smoothed log spectrum. The resulting cepstrum is then liftered in the normal way and an FFT is performed to produce a fully smoothed spectral envelope which hugs the spectral peaks of the log spectrum and appears to faithfully model the zeros.

The LPC method utilised was the autocorrelation method as recommended for this type of signal by Markel and Gray (ibid, p152). 30 coefficients were used. In general, the minimum number of coefficients used must be twice the number of poles in the signal. For adult male vowels band-limited between 0-5 kHz there are usually five major poles and so 10 coefficients would seem a reasonable minimum. But as Markel and Gray (ibid, p 154) point out at least 15 coefficients are required to permit the resolution of closely spaced formants. Even more are required if the fine spectral detail is to be resolved although if too many are used the individual harmonics begin to be resolved and the the comparison of two spectra will also be a function of pitch synchrony. For the purposes of the present study the number of coefficients was systematically

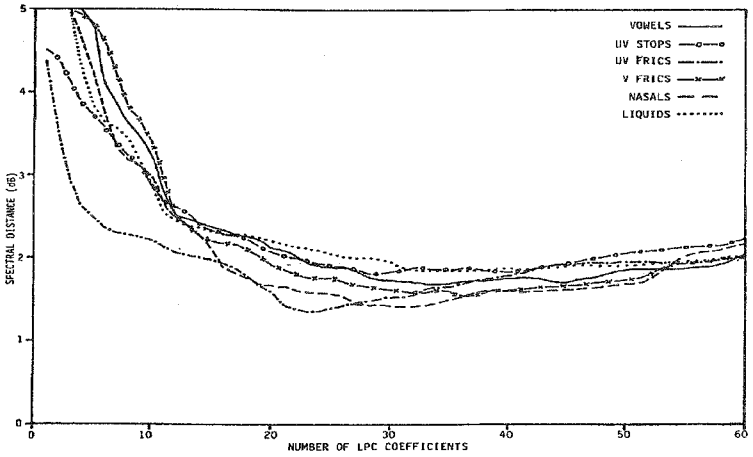


*Figure 1. Spectral distance between an "improved" Cepstrally-smoothed log spectrum and an LPC derived log spectrum of natural speech tokens with the number of LPC coefficients being varied from 1 to 60*

varied and the resultant smoothed spectrum compared with a Cepstrally smoothed spectrum of the same frame of the same natural token. The Cepstrally smoothed spectrum was produced using the "improved" cepstral method described above as it has approximately the same spectral slope as the LPC spectra. The differences between each spectral pair was computed using the weighted Euclidean measure described below and an average value was obtained for each of 6 phonetic classes. The position of the minimum spectral distance was found to occur between 20 and 30 coefficients (depending on the phonetic class) with a gradual deterioration above about 30.

The difference between the spectra at each spectral point was weighted in proportion to the inverse of the auditory system's critical bandwidth before the points were averaged together. This would give a distance measure which would be equivalent to the distance between two Bark-scaled spectra. Thus a Hertz-scale (unweighted) and a Bark-scale (auditorily weighted) version of each of the Cepstrum, "improved" Cepstrum and LPC measures was calculated.

The distances between the smoothed spectra of two signals are usually derived using the Euclidean (root mean square) method although sometimes the simple Chebyshev or city-block distance measure (arithmetic mean of the absolute differences at each point) is utilised (O'Shaughnessy (1987)). These differences are usually measured between the log spectra of the two signals giving an average distance in dB. Both of these methods were compared in the present study. The Chebyshev distance measure treats all point by point differences equally, whereas the Euclidean distance measure emphasises large deviations at the expense of smaller deviations so that whilst two pairs of spectra may have the same Chebyshev value the pair with a few points of great deviation will have a higher Euclidean distance than a pair without such great deviations.

A single distance value was then calculated for each pair of tokens by averaging the individual frame distance measures across the entire segment. This was readily achieved as the natural speech was segmented prior to the operation of the distance measure algorithms and since the natural and vocoded tokens were already precisely time aligned.

This whole process resulted in 12 distance measures (see table 1) for each token pair. Each distance measure for each token pair was then plotted against vocoded token intelligibility (as a percentage of natural intelligibility ie. perceptual distance) and a Pearson's R correlation was calculated for all vowels and for all consonants (nb. standard scores were derived first and all correlations were calculated from them). For each of the two major phonetic classes (ie. vowels and consonants) twelve sets of correlations between spectral and perceptual distances were calculated for i) all vocoders pooled, ii) all Hertz scaled vocoders, iii) all uncorrected Bark-scaled vocoders, iv) all corrected Bark-scaled vocoders. A "t" score of each correlation was also calculated and all correlations were tested for significance. Further, the significance of the differences between each relevant pair of Pearson's R values was also calculated. This last calculation had to take into account the fact that for each pair of correlations one of the parameters (intelligibility) was shared by both correlations and that the second parameter in one correlation (ie. spectral distance) was highly correlated with the similar parameter in the second correlation.

One of the aims of this study was to examine the effectiveness of spectral distance measures on signals containing significant spectral zeros and so the nasals, which showed considerable correlation between intelligibility and spectral distance, also had all of the above tests carried out on them as a separate group.

RESULTS AND DISCUSSION

The results of the correlations between intelligibility and all 12 distance measures for all vocoders are summarised in table 1. It can be seen that all the distance measures show a high correlation (significant at the 0.01 level) between spectral distortion and intelligibility. The results for the Hz-scaled and corrected and uncorrected Bark-scaled vocoders were carried out but are not shown for reasons of space.

Each pair of correlations that differed for one pair of parameters only, was examined for significant difference, but only when at least one member of the pair was a significant correlation. Six pairs were compared to test the difference between auditorily-weighted and unweighted distance measures, six pairs compared Euclidean versus Chebyshev measures, four pairs compared Cepstrum versus "improved"

Cepstrum, four pairs compared Cepstrum versus LPC and four pairs compared "improved" Cepstrum versus LPC. The results for all vocoders are summarised in table 2. Ignoring the cases where there is no significant difference it can be seen that weighted measures are better correlated with intelligibility than unweighted measures, Euclidean measures are better correlated with intelligibility than Chebyshev measures, and LPC measures are better correlated with intelligibility than are Cepstrum measures which in turn perform better than the "improved" Cepstrum measures. Clearly, the Bark-weighted Euclidean measures of spectral distance between LPC spectra is the measure most highly correlated with intelligibility.

The same comparisons when performed on nasal consonants and on the Hz-scaled and Bark-scaled vocoders separately give the same patterns as above with the following exceptions.

1) The Euclidean measures always perform the same as or better than the Chebyshev measures except for the corrected and uncorrected Bark-scale vocoded vowels. In those cases, the unweighted Chebyshev measures perform better than the unweighted Euclidean measures. Weighted Euclidean measures, however, always perform better than the weighted Chebyshev measure and the unweighted Euclidean and Chebyshev measures.

2) The Cepstrum method is always better than or no different to the "improved" Cepstrum method for the vowels. For consonants on the other hand the "improved" Cepstrum method is either no different to or better than the ordinary Cepstrum method and the latter trend is especially strong for the Hz-scaled vocoded consonants.

## CONCLUSIONS

Bark-scale weighted Euclidean distance measures are shown to be superior to the other methods in all cases. Neither the normal nor "improved" Cepstral smoothing methods can clearly be shown to be better than the other, the choice depending upon whether the segments being analysed are vowels or consonants. The preference of Bark-weighted to unweighted spectral distance measures (in agreement with Bladon and Lindblom, 1981) is unsurprising as this weighting models the frequency resolution of the auditory system. The preference for the Euclidean measure is also not surprising as it emphasises major deviations at the expense of minor deviations and so is more likely to emphasise perceptually important spectral distortions at the comparative expense of less important deviations such as overall spectral slope.

The LPC spectrum method is never inferior to the Cepstrally smoothed spectrum methods used and is definitely superior if it is combined with the Bark-scale weighted Euclidean measure. The all-pole LPC method appears to be able to adequately model speech signals even though it cannot accurately model spectral zeros. Whether this is so if the number of coefficients are reduced to 15, the minimum number recommended by Markel and Gray, needs to be the subject of further analysis. When 30 coefficients are used significant spectral dips are modeled but the LPC method is still weighted non-linearly in favour of the spectral peaks. Perhaps the degree of spectral detail thus provided for the modeling of the zeroes is sufficient to meet that required by human speech perception. This seems to be so even for the nasal consonants which contain significant spectral zeroes.

## REFERENCES

Atal, B.S. (1974) "Effectiveness of linear prediction character- istics of the speech wave for automatic speaker identification and verification", J.Acoust. Soc. Am. 55, 1304-1312.

Bladon, R.A. & Lindblom, B. (1981) "Modelling the judgement of vowel quality differences", J.Acoust.Soc.Am. 69, 1414-1422.

Clark, J.E., Mannell, R.H. & Ostry, D. (1987) "Time and frequency resolution constraints on synthetic speech intelligibility", Proc. 11th Int. Congress of Phonetic Sciences, Tallin, Estonia.

Clark, J.E. & Mannell, R.H. (1988) "Some comparative characteristics of uniform and auditorily scaled channel channel synthesis", Proc. 2nd Aust.Int.Conf. SST-88, Sydney.

Gray, A.H. & Markel, J.D. (1976) "Distance measures for speech processing", IEEE Trans. ASSP-24, 380-391.

Klatt, D.H. (1982) "Prediction of perceived phonetic distance from critical-band spectra", Proc. ICASSP-82, 1278-1281.

Markel, J.D. & Gray, A.H. (1976) Linear Prediction of Speech, (Springer-Verlag: Berlin).

O'Shaughnessy, D. (1987) Speech Communication: Human and Machine, (Addison-Wesley: Reading, Mass., USA).

Tohkura, Y. (1987) "A weighted cepstral distance measure for speech recognition", IEEE Trans. Acoust., Speech and Signal Processing, ASSP-35, 1414-1422.

| DISTANCE MEASURE | VOWELS | | CONSONANTS | |
|---|---|---|---|---|
| CEPSTRUM | R | (T) | R | (T) |
| i) Chebyshev (unweighted) | -.5023 | (7.162) | -.1610 | (2.651) |
| ii) Chebyshev (weighted) | -.5362 | (7.832) | -.2623 | (4.236) |
| iii) Euclidean (unweighted) | -.5038 | (7.190) | -.1838 | (3.038) |
| iv) Euclidean (weighted) | -.5693 | (8.537) | -.2775 | (4.693) |
| "IMPROVED" CEPSTRUM | R | (T) | R | (T) |
| i) Chebyshev (unweighted) | -.4936 | (6.998) | -.1632 | (2.688) |
| ii) Chebyshev (weighted) | -.5125 | (7.359) | -.2586 | (4.350) |
| iii) Euclidean (unweighted) | -.4871 | (6.876) | -.1866 | (3.085) |
| iv) Euclidean (weighted) | -.5319 | (7.744) | -.2838 | (4.809) |
| LPC | R | (T) | R | (T) |
| i) Chebyshev (unweighted) | -.5043 | (7.201) | -.2236 | (3.728) |
| ii) Chebyshev (weighted) | -.5273 | (7.651) | -.2902 | (4.928) |
| iii) Euclidean (unweighted) | -.5165 | (7.437) | -.2581 | (4.341) |
| iv) Euclidean (weighted) | -.5768 | (8.705) | -.3197 | (5.481) |

*Table 1. Correlation between the eight spectral distance measures and the intelligibility of all vowel and all consonant tokens for all 14 vocoders. "Student's" T scores for each correlation are given in brackets. df = 152 (vowels) and df = 264 (consonants). All correlations are significant at the .01 level.*

| | VOWELS | CONSONANTS |
|---|---|---|
| Weighted vs Unweighted (6 pairs compared) | W > U (50%)<br>n.s.d. (50%) | W > U (100%) |
| Euclidean vs Chebyshev (6 pairs compared) | E > C (67%)<br>n.s.d. (33%) | E > C (100%) |
| Cepstrum vs "Improved" Cepstrum (4 pairs compared) | C > C+ (75%)<br>n.s.d. (25%) | n.s.d. (100%) |
| Cepstrum vs LPC (4 pairs compared) | n.s.d. (100%) | L > C (100%) |
| "Improved" Cepstrum vs LPC (4 pairs compared) | L > C+ (50%)<br>n.s.d. (50%) | L > C+ (100%) |

*Table 2. Results of tests for significant difference (0.05 level) between pairs of correlations for all 14 vocoders. ("n.s.d." = "no significant difference", "C+" = "improved" Cepstrum method)*