THE PREDICTION OF "PERCEPTUAL DISTANCE" FROM SPECTRAL DISTANCE MEASURES BASED UPON AUDITORY AND NON-AUDITORY MODELS OF INTENSITY SCALING

Robert H. Mannell
Speech, Hearing and Language Research Centre
Department of Linguistics
Macquarie University

ABSTRACT - Spectral and Perceptual distances of channel vocoded speech from the original natural speech tokens are examined. Perceptual distances are defined as the differences between the intelligibility of each natural speech token and the intelligibilities of each of its derived synthetic forms. The spectral distances are second order Euclidean distances between the Bark-scaled spectra of each natural speech token and its derived synthetic tokens (summed over 10 ms frames). A number of different spectral distances based on various auditory and non-auditory intensity scales were compared by examining the correlations between the spectral and perceptual distances for each intensity scale. Correlations were compared globally and across various phonetic categories and the intensity scale(s) that provided the best correlations for each phonetic class were determined. The auditory and perceptual consequences of the results are discussed.

INTRODUCTION

Mannell (1992) examined the effects on monosyllable intelligibility of passing natural speech tokens through a number of channel vocoder configurations. These vocoder configurations were identical (1 Bark filter banks and 10 ms frames) except for their intensity processing strategies. In each configuration the demodulated bandpassed output of each channel was quantised according to one of three intensity scales: dB, sones and a scale derived from intensity-jnds. The degree of quantisation coarseness on each of these scales was also varied (eg. 1, 2, 4 ... dB quantisation steps). Further, the quantisation procedure needed to anticipate which of the four presentation levels (40, 50, 70 and 90 dB spl) was to be utilised for the resultant speech token in the perceptual tests (eg. a particular quantisation coarseness on the sone scale needs to be recalculated for each intended presentation level). The results showed that dB quantisation resulted in identical intelligibility curves across the four presentation levels (ie. intelligibility began to deteriorate only when the quantisation coarseness exceeded 8 dB and this deterioration point was the same for all four presentation levels[1]). Surprisingly, the speech derived from quantisation based on the two scales derived from careful psychoacoustic measurements (sones and intensity-jnds) were shown to have intelligibility curves which varied with presentation level (eg. speech intelligibility deteriorated for quantisation coarseness greater than 0.8, 1.6, 6.4 and 12.8 sones at 40, 50, 70 and 90 dB presentation levels respectively). It was argued that phonetically significant information is auditorily encoded in the intensity dimension according to a scale similar to the dB scale rather than a scale such as the sone loudness scale. It was further argued that a scale based on the log of loudness ("logsones") is very closely approximated by dB SL (ie. above threshold).

[1] Note that the dB results in Mannell (1992) were consistently out by a factor of two, so that the paper showed the point of intelligibility deterioration to be 16 dB rather than the correct value of 8 dB. (nb. All other results presented in that paper were accurate.)

The above experiments were designed to elicit the intensity scale which was assumed to be most closely related to auditory scaling of intensity. In other words, the scale which produces identical points of intelligibility deterioration (in terms of quantisation coarseness) across a number of presentation levels was taken to be the major (or perhaps only) phonetically-relevant *auditory* representation of the intensity dimension. This auditory representation is assumed to form the input to the phonetic processing centre(s) in the brain. Once central phonetic processing commences it is possible that the scale selected by the above experiments may no longer be the most relevant intensity representation utilised phonetically. It seems possible that the more central phonetic processes may utilise various non-linear transformations of the intensity dimension. Repp (1987) argues that in the perception of degraded, context-free speech signals, of the kind utilised in these experiments, "...the perceiver must make a decision based on the *perceptual distances* of the input from possible phonetic alternatives (prototypes) stored in his or her permanent knowledge base." (ibid, p16) He asks "... what is the phonetic *distance metric*, what are the dimensions of the perceptual space in which it operates, and what are the perceptual weights of these dimensions." (ibid, p16). If the auditory intensity scale determined by the previous experiments is used by the phonetic processors without non-linear transformation then spectral distances based on that scale will always produce a good approximation to phonetic or perceptual distance. If, on the other hand, different phonetic (cue-extraction) processes involve different transformations of the auditory representation, then it is possible that other intensity scales may sometimes provide a better approximation of phonetic distance.

METHODOLOGY

In the present study, a second-order Euclidean distance metric based on 1 Bark-scaled, cepstrally-smoothed FFT spectra was used to determine the effect of various auditory and non-auditory intensity scales on their ability to predict intelligibility. The speech utilised in this study was derived from 29 natural speech tokens (11 vowels in /h_d/ context and 19 consonants in CV /_a/ context) passed through 14 vocoder configurations in which the speech was systematically distorted in either the frequency or the time dimension (but not the intensity dimension which was consistently subjected to 16 bit intensity quantisation).

LIST 1 /h_d/		LIST 2 /Ca/			
heed	/i:/	tar	/t/	za	/z/
hid	/ɪ/	da	/d/	shah	/ʃ/
had	/æ/	car	/k/	ma	/m/
hard	/a:/	ga	/g/	na	/n/
hud	/ʌ/	cha	/tʃ/	la	/l/
hod	/ɒ/	jar	/dʒ/	ra	/r/
hoard	/o:/	far	/f/	wa	/w/
hood	/ʊ/	va	/v/	ya	/j/
who'd	/ʉ:/	sa	/s/		
heard	/ɜ:/				
hide	/aɪ/				

These speech tokens (plus the original natural speech) were presented to a total of 300 listening subjects (20 subjects per condition) to derive intelligibility measures for each token. These intelligibility results were then converted to a measure of perceptual distance:-

$$D_p = \frac{\left(I_n - I_v\right) \times 100\%}{I_n}$$

where: D_p = Perceptual distance between each synthetic token and its original natural token

I_n = % Intelligibility of natural speech token

I_v = % Intelligibility of the derived synthetic speech token

This perceptual distance was calculated for each token.

Also, for each token, a number of spectral distances were calculated. These spectral distances were the second order Euclidean distance between each original natural token and each of the vocoded tokens derived from it. As the natural and vocoded tokens were perfectly time aligned these calculations were not conflated with the effects of processes such as time warping that were irrelevant to the current investigation. Total spectral distance for a phoneme was determined as the mean of the spectral distances of every 10 ms frame over the duration of that phoneme.

$$D_s = \sqrt[m]{\frac{\sum_{i=1}^{n}\left(abs\left(S_{ni} - S_{vi}\right)\right)^m}{n}}$$

where D_s = Spectral distance

S_{ni} = Amplitude of ith frequency point of natural spectrum

S_{vi} = Amplitude of ith frequency point of vocoded spectrum

m = 2 (for the 2nd order Euclidean distance)

The spectral distances were calculated utilising Pascal, dB, sone, logsone and intensity-jnd scaling. The results of each intensity scaling approach were correlated with perceptual distances (ie. each vocoded token was plotted as a function of its spectral distance against its perceptual distance from the natural speech token).

There were two types of spectral distance determined. The first type was a standard 2nd order Euclidean distance. The second type of distance was identical to the first except that for each frame the two spectra were slid together uniformly along the intensity dimension until the minimum spectral distance was obtained for that frame (so that the shape of the two spectra rather than their relative global intensities were the basis of the resulting spectral distance measure). The latter type of distance is referred to as a "slid" distance below.

RESULTS

Tables 1 to 4 summarise the results of this experiment for vowels, consonants, voiceless stops, voiceless fricatives and approximants.

	Vowels	Consonants	Voiceless Stops
deciBel	0.3864	0.5723	0.8201
logsone	0.3868	0.5825	0.8021
Pascal	0.4518	0.3567	0.0624*
sone	0.4420	0.2829	-0.0911*
intensity-jnd	0.5218	0.5520	0.3122*

Table 1 Pearson's r correlations between Bark-scaled Euclidean spectral distances across five intensity scales. Spectral distances are averaged across the target phoneme (plus transition) only rather than across the entire token. All correlations are significant except those marked with an asterisk. (N_{vwl} = 154, N_{con} = 266, N_{uvstp} = 42)

	Vowels	Consonants	Voiceless Stops
deciBel	0.5241	0.4695	0.2164*
logsone	0.5237	0.4341	0.1959*
Pascal	0.4531	0.3318	0.0977*
sone	0.4695	0.2319	-0.0984*
intensity-jnd	0.5978	0.4462	0.1328*

Table 2 Pearson's r correlations between "slid" Bark-scaled Euclidean spectral distances across five intensity scales. Spectral distances are averaged across the target phoneme (plus transition) only rather than across the entire token and spectral pairs have been slid together to minimise the global distance between them. All correlations are significant except those marked with an asterisk. (N_{vwl} = 154, N_{con} = 266, N_{uvstp} = 42)

	voiceless fricatives	approximants
deciBel	0.5631	0.5612
logsone	0.6141	0.5532
Pascal	0.1938*	0.6339
sone	0.1813*	0.6794
intensity-jnd	0.6038	0.6638

Table 3 Pearson's r correlations between Bark-scaled Euclidean spectral distances across five intensity scales. Spectral distances are averaged across the target phoneme (plus transition) only rather than across the entire token. All correlations are significant except those marked with an asterisk. (N_{vwl} = 154, N_{con} = 266, N_{uvstp} = 42)

	voiceless fricatives	approximants
deciBel	0.5686	0.7352
logsone	0.5672	0.7146
Pascal	0.1389*	0.6043
sone	0.1071*	0.6458
intensity-jnd	0.4973	0.7157

Table 4 Pearson's r correlations between "slid" Bark-scaled Euclidean spectral distances across five intensity scales. Spectral distances are averaged across the target phoneme (plus transition) only rather than across the entire token. Spectral pairs have been slid together to minimise the spectral distance between them. All correlations are significant except those marked with an asterisk. (N_{vwl} = 154, N_{con} = 266, N_{uvstp} = 42)

DISCUSSION

When each of the intensity scales is considered in terms of its ability to encode detailed spectral shape it becomes clear that the dB and logsone scales capture a great deal of spectral detail whilst the Pascal and the sone scales are relatively insensitive to detailed spectral shape but are particularly sensitive to peak (eg. formant) position. The five intensity measures can be seen to lie on the following continuum:-

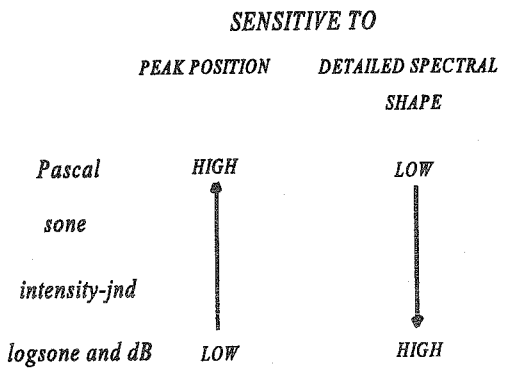

Figure 1 Each of the intensity scales has a different sensitivity to peak position and to detailed spectral shape.

The correlations for vowels and approximants were highest for spectral distances based on the intensity-jnd, Pascal and sone scales, with intensity-jnd correlations being significantly higher than those for the logsone and dB scales. When spectral pairs (natural and vocoded) were slid together uniformly on the intensity axis to their smallest global distance before calculating the spectral distances the performance of the dB and logsone scales for vowels and approximants increased markedly, the performance of the intensity-jnd scale increased a little, whilst the performance of the Pascal and sone scales remained constant. After sliding, the intensity-jnd scaled spectral distance predicted perceptual distances significantly better than the Pascal and sone scales. These results are interpreted as indicating that the

peak position sensitivity of the Pascal and sone scales produced good predictions of intelligibility but that this could be improved by supplying extra information on detailed spectral shape. The intensity-jnd scale provides that extra detail mainly in the region of the spectral peaks and the superior performance of measures based on the scale suggests that peak shape detail (eg. peak bandwidth) provides significant information in the identification of vowels and vowel-like consonants.

The consonant results were quite complex and varied according to phonetic class. The approximants, as suggested above, behaved very much like vowels. The least vowel-like consonant classes were also examined in detail (ie. voiceless stops and voiceless fricatives). The Pascal and sone scaled spectral distances were very poor predictors of both stop and fricative intelligibility. The intensity-jnd scaled distance measure was a good predictor of fricative intelligibility but a relatively poor predictor of stop intelligibility. The dB and logsone distances were good predictors of both stop and fricative intelligibility. Further, when distances were calculated after the sliding together of spectral pairs along the intensity axis there was no effect on the predictive power of the dB, logsone and intensity-jnd distances for fricatives. The sliding of stop pairs resulted, however, in a complete loss of significant predictive power for all intensity-scaled spectral distances.

Logsone, dB and intensity-jnd measures all provide a good representation of the shape of the broad major peaks in the fricative spectra. The intensity-jnd measure is much less sensitive than dB and logsones, however, to the spectral shape detail in the spectral dips. These results suggest that measures that accurately model the shape of the more intense regions of the broad fricative peaks are good predictors of fricative intelligibility.

The failure of all measures other than those based on dB and logsone scales suggests that the encoding of both low and high intensity components of stops is important for the prediction of stop intelligibility. The loss of predictive power of all intensity scaled distances derived after the sliding of spectral pairs to their minimal spectral distance has as one of its effects the distortion of frame to frame intensity differences. In other words, natural versus synthetic differences in the intensity slope at points of rapid intensity change (such as stop bursts) is no longer being adequately modelled when the spectral pairs for each frame are slid together to their minimum distance. Detailed spectral analysis of the natural and synthetic stops (Mannell, 1994) shows that it is this burst intensity slope that most differentiates the natural and their derived synthetic tokens and so would be the most likely cause of variations in intelligibility. It is argued that this slope needs to be modelled in a way that matches the auditory representation of phonetically-relevant intensity changes. The results of Mannell (1992) suggest that the dB and logsone scales most closely match this internal representation and these two scales prove to be the best predictors of stop intelligibility.

Reference:

Mannell, R.H. (1992) "The effects of presentation level on sone, intensity-jnd and deciBel quantisation of channel vocoded speech", SST-92, 802-807

Mannell, R.H. (1994) *The Perceptual and Auditory Implications of Parametric Scaling in Synthetic Speech*, Unpublished Ph.D. dissertation, Macquarie University: Sydney

Repp, B.H. (1987) "The role of psychophysics in understanding speech perception", In M.E.H. Schouten (ed) *The Psychophysics of Speech Perception*, Martinus Nihoff: Dordrecht, 3-27