Proceedings of the Institute of Acoustics

THE PERCEPTOGRAM – A PSYCHOACOUSTIC SCALING OF THE SPECTROGRAM

M.A. BROWNE AND M.J. HENERY

Department of Instrumentation and Analytical Science, University of Manchester Institute of Science and Technology Manchester, M60 1QD, UK

1. INTRODUCTION

The principle of acoustic invariance states that phonetic features are decoded by the human auditory system from complex invariant properties of the speech waveform. In search of means to produce this invariance various researchers have used techniques related to current knowledge of the peripheral auditory system [1, 2, 3]. We propose a method of representing acoustic data with psychoacoustic scaling to produce a new graphical output which is related to the spectrogram.

The spectrogram, which was for many years produced by analogue means, provides a graphical description of a speech waveform in which the ordinate describes frequency in Hz, the abscissa describes time and the grey-scale intensity (dBA) at a particular coordinate (darker means higher intensity). Nowadays, digital techniques are replacing the analogue systems and Figures 1 (a) and (b) present spectrograms produced on such a system [9]. Experienced readers can interpret a great deal of information from spectrograms, but for automatic analysis their variability due to speaker gender, dialect, health and so forth causes many difficulties. Furthermore since it is well known that the primary recognition device in use, the human auditory system, does not perceive acoustic signals according to linear physical scales, the spectrogram offers a poor representation of the perceived information. In fact, the situation is worsened by digital techniques, because when the discrete Fourier transform is applied in order to achieve adequate spectral resolution temporal resolution must be sacrificed, thereby smearing and losing temporal cues for use in segmentation.

For these reasons we have developed alternative means of describing the speech waveform, which is based on psychoacoustic scaling, as perceived by the human auditory system, and which utilises temporal data to provide segmentation cues within words. The motivation for this work has been the assertion that psychoacoustic scaling should contribute to achieving acoustic invariance which will aid phonetic decomposition and phoneme classification. This analysis and subsequent syntactic pattern recognition should constitute a viable route to at least word recognition and might also act as a lower level processing step for connected speech interpretation. To aid the first two steps in this process we have used a limited model of the peripheral auditory system in which spectral and intensity data undergo transformation into psychoacoustic scales and temporal analysis, using a modification of Baker's [4] log inverse period (LIP) plots, is used to provide phoneme segmentation cues.

2. PSYCHOACOUSTIC SCALING

Psychoacoustics is the branch of acoustics which deals with the relationship between acoustic stimulus and subjective response. Subjective responses have been quantified in perceptual scales and these scales have been used to transform raw acoustic data into the psychoacoustic domain. Intensity, which is normally measured in decibels (SPL – sound pressure level) relative to 2×10^{-5} Pa, is transformed into loudness and measured in *sones*, details of which can be found in BS3045:1981

and BS3383: 1988. It is found that perceived loudness depends on frequency and the relationships provide a non-linear mapping from intensity to sones.

The perception of frequency or pitch is also found to be at odds with physical scaling. The relationship between the physical and perceived frequencies is a non-linear function and converts Hz to Barks (the perceptual unit).

In the peripheral auditory system the process generally referred to as lateral inhibition is evident and leads to the phenomenon of masking [6]. By this means dominant tones from complex acoustic signals tend to be emphasised. As with all such physical processes a memory of these events exists. This same process of lateral inhibition gives rise to *critical bands*, which describe the extent in frequency terms over which the masking is active. Zwicker [5] proposed that an empirically defined critical-band scale be adopted as a standard. His proposed scale divides the human auditory range below 16kHz into 24 critical band units which he termed "barks" after Barkhausen, the creator of the loudness level. Since this proposal much work and debate has followed and various bark scales now exist, though with broadly similar characteristics.

In addition to non-linear scalings and lateral inhibition or masking the PAS is also known to exhibit dynamic selection and adaptation to the envelope of acoustic signals. Rate intensity functions are used to describe the response to onset and offset and adaptation to continuous or background tones [2]. These factors should be included in a representative computational model of the PAS, although to date we have implemented only the non-linear scaling and re-scaled LIP plots.

3. THE PAS MODEL AS A PREPROCESSOR

The PAS model is used to pre-process the digitised speech signal and the output, which consists of temporal and spectral data, is presented graphically in the first instance and used in the subsequent steps of segmentation and analysis.

Temporal processing computes the time between consecutive up-crossings of the raw signal, and uses linear interpolation to improve the estimate. Hysteresis is employed to reduce corruption by noise. For each detected cycle, the magnitude, inverse period (frequency) and time of occurrence are computed. Using the former two measurements, the phone value is calculated and then the sone value as detailed below. The bark value is then found for frequency from [6].

$$Bark = ((\ 26.81 * Freq\ ) / (\ 1960 + Freq\ ))$$

The number of zero crossings per 10ms is also evaluated.

In spectral processing the raw signal is applied to a bank of 20 digital filters, each of which has the characteristics of a critical band as evaluated in the PAS, i.e. a bank of 20, 1 bark digital filters. The twenty outputs are then subject to the same zero-crossing analysis described above and the bark, sone and time output is produced. The temporal and spectral data is then used to produce the perceptogram plots as described in Section 4.

3.1 Amplitude Scaling to Sones

Standard tables are given in BS3383:1988 which graphically depicts a set of equi-loudness contours. Loudness L_s in phones is first calculated, assuming the raw digitised signal is linearly

related to SPL, by the following:

$$L_n = 4.2 + \frac{a_f (L_f - T_f)}{1 + b_f (L_f - T_f)}$$

where a_f, b_f and T_f are tabulated coefficients at given spot frequencies and L_f is SPL of a tone in dB relative to 2×10^{-5} Pa.

To obtain intermediate values with 3 variables, linear interpolation is inaccurate, and so a cubic spline fitting technique, utilising a matrix formulation [7] is used. Sones can then be computed from the phone value, L_n as described in BS3045:1981

$$\text{Loudness (N) Sones} = 2^{0.1\,(L_n - 40)}$$

3.2 The Bark Digital Filter Bank

The basic design equation is taken from Sekey and Hanson [8] and allows construction of arbitrary 1-bark filters with centre frequencies spaced at one bark and with a roll-on of +15dB/bark and roll-off -25dB/bark and 3dB bandwidth of 1 bark.

The filter function F(z) used for all filters is

$$10 \log_{10} F(z) = 7.00 - 7.5\,(Z - 0.215) - 17.5\,(0.196 + (Z - 0.215)^2)^{1/2}$$

and frequency in Hz, f is related to barks, z as follows: $f = 600 \sinh [(z + 0.5)/6]$, $z = 1, 2, \ldots 20$.

The frequency response values were calculated over a range of 0 to -60dB for each bark filter and then passed to a digital filter design program [9] as a template. A best fit was generated using a ten-stage bi-quad bandpass implementation and the resultant coefficients for each bark filter were stored and used in subsequent processing. The characteristics of a selected group of templates from the bark is shown in Figure 2.

4. THE PLIP AND PERCEPTOGRAM PLOTS

The temporal and spectral data obtained as described above from the raw digitised signal are used to produce two forms of perceptually scaled graphical outputs. The first is a perceptually scaled log inverse period (PLIP) plot after Baker [4]. In this plot the abscissa is time and the ordinate frequency in barks, with the spot located at the corresponding coordinate having a diameter proportional to loudness in sones. An example for a tri-phone spoken by male and female subjects is shown in Figure 3 (a) and (b). The zero-crossing count in each 10ms period is overlaid on this plot and clearly shows the potential of this information as a phoneme segmentation cue, as noted in previous studies [2, 4]. If the results are compared with a standard LIP plot (Figure (c) there is clearly less information present and comparison between Figures 3 (a) and (b) suggest that a certain amount of invariance has been achieved, though further intensive study must be undertaken to establish the observation for a range of speakers.

The perceptogram itself is obtained by plotting the PLIP information output from each of the twenty channels of the bark filter bank. Results are shown in Figures 4 (a) and (b) for the same tri-phone as Figure 3. It is clear that despite the vast discrepancy between the digitised raw signals a degree of invariance has been achieved and with segmentation cues offered by the zero-crossing template the steps towards phoneme identification can begin.

Acoustic invariance must be investigated especially in respect of the relationship between formants for different speakers for vowel classification. Plosives and fricatives show clear differences and voiced and unvoiced plosives were clearly distinguishable in other studies.

5. CONCLUSIONS

Processing of acoustic speech signals by the PAS is thought to provide acoustic invariance which enables recognition of utterances, as specific phonemes. To study the contribution of psychoacoustic scaling to acoustic invariance we have developed two rescaled graphical presentations of speech data. The perceptual scaling of log inverse period plots and similar analysis of data after processing by a 20 channel 1-bark filter bank have been presented.

Preliminary results suggest that new scalings show some promise in the normalisation of disparate acoustic speech signals. An intensive evaluation of the limitations of such scaling must now be undertaken.

6. REFERENCES

[1] R. GOLDHOR, "A Speech Signal Processing System Based on a Peripheral Auditory Model", ICASSP 83, Boston, pp. 1368-71, IEEE, 1983.

[2] M. COOKE, "Speech Analysis Using a Model of Hearing", Proc. 2nd Int. Conf. Machine Intelligence, 26-28 November 1985, London, IFS.

[3] A.K. SYRDAL, H.S. GOPAL, "A Perceptual Model of vowel Recognition Based on the Auditory Representation of American English Vowels", J.A.S.A., 79(4), p. 1086, April 1986.

[4] J.M. BAKER, "A New Time-Domain Analysis of Human Speech and Other Complex Waveforms", PhD Thesis Carnegie-Mellon University, 1975.

[5] E. ZWICKER, "Sub-Division of the Audible Frequency Range Into Critical Bands (Frequenzgruppen)", J.A.S.A., 33(2), p. 248, 1961.

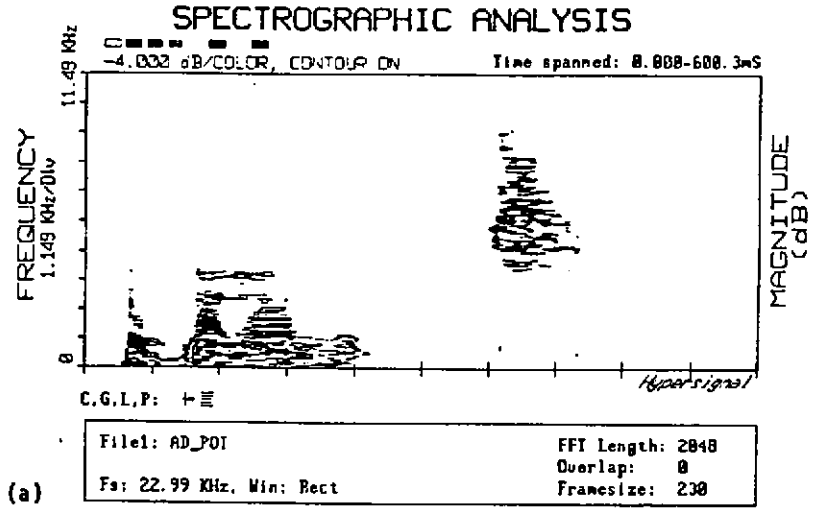
[6] E. ZWICKER, "Masking and Psychological Excitation as Consequences of the Ear's Frequency Analysis", in Frequency Analysis and Periodicity Detection in Hearing, Ed. R. Plomp and G. Smoorenburg, Leiden, 1970.

[7] M.A. BROWNE AND P.A. GAYDECKI, "High Speed Spline Fitting - With Application to Boundary Tracing in Low Contrast Digital Images", Comput. Biol. Med., 17(2), pp. 109-116, 1987.

[8] A. SEKEY AND B.A. HANSON, "Improved 1-Bark Bandwidth Auditory Filter", J.A.S.A., 75(6), pp. 1902-04, 1984.

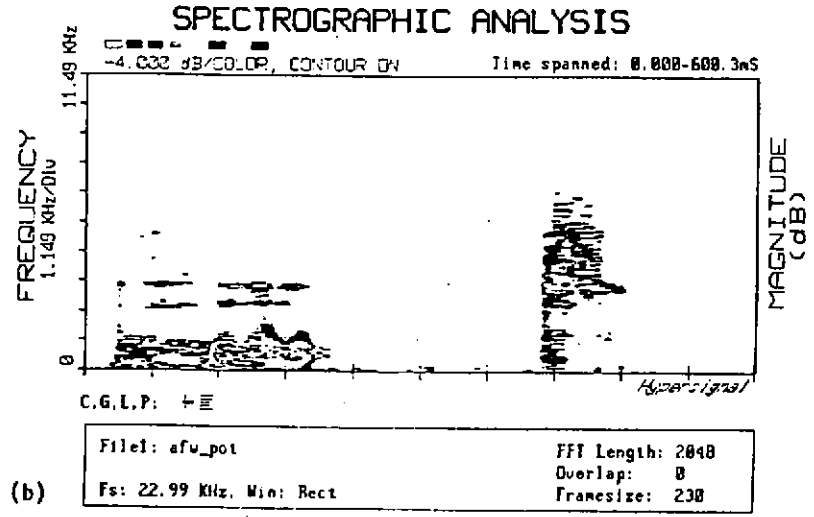
[9] Hypersignal - Supplied by LSI Ltd., Loughborough, UK.

THE PERCEPTOGRAM - A PSYCHOACOUSTIC SCALING OF THE SPECTROGRAM



(a)

Figure 1 Spectrogram of a Triphone (POT) For
(a) A Female Speaker
(b) A Male Speaker



(b)

THE PERCEPTOGRAM - A PSYCHOACOUSTIC SCALING OF THE SPECTROGRAM

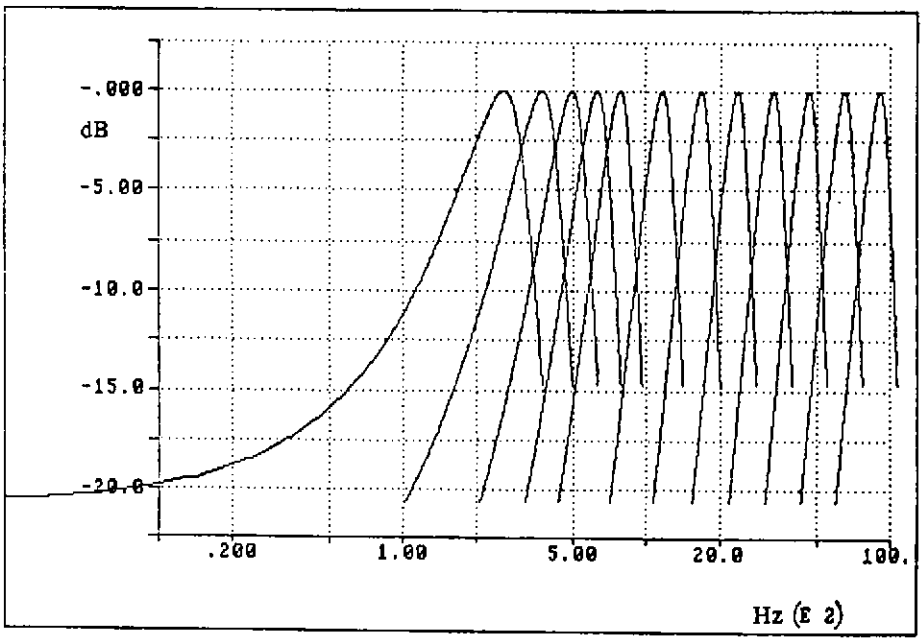


Figure 2 Selected Group of Frequency Response Templates For the 1-Bark Filter Functions

THE PERCEPTOGRAM - A PSYCHOACOUSTIC SCALING OF THE SPECTROGRAM

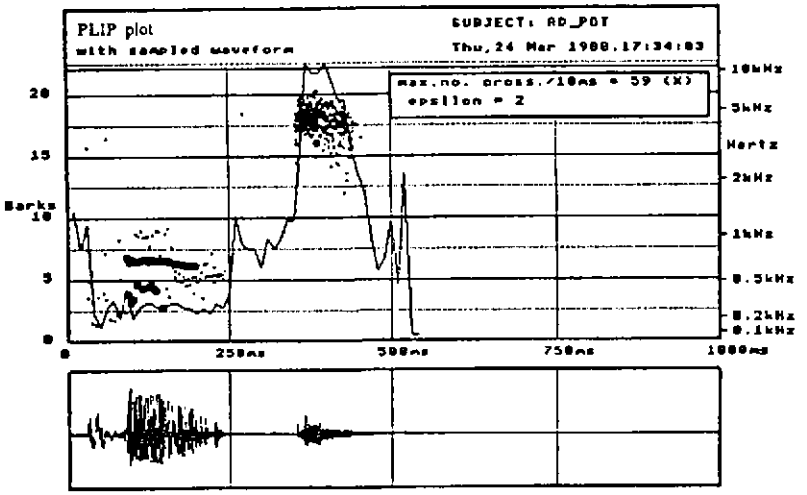
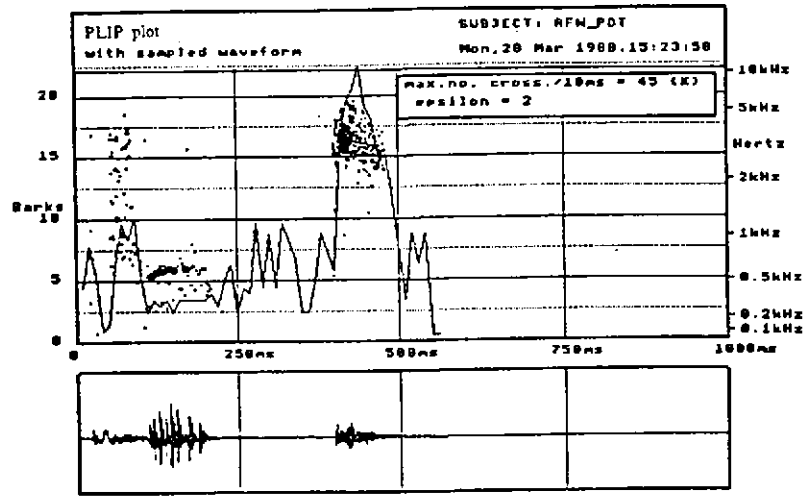
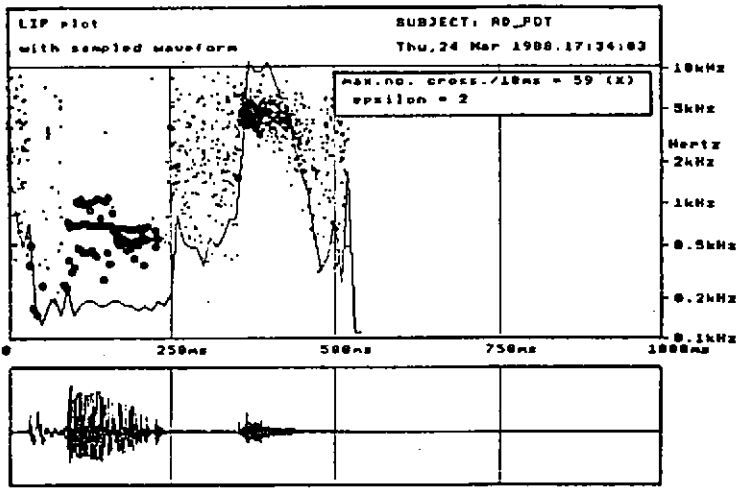


Figure 3 Log Inverse Period Plots of the Triphone
(a) Perceptually Scaled Female Speaker



(b) Perceptually Scaled Male Speaker

THE PERCEPTOGRAM - A PSYCHOACOUSTIC SCALING OF THE SPECTROGRAM



(c) Physically Scaled Female Speaker

THE PERCEPTOGRAM - A PSYCHOACOUSTIC SCALING OF THE SPECTROGRAM

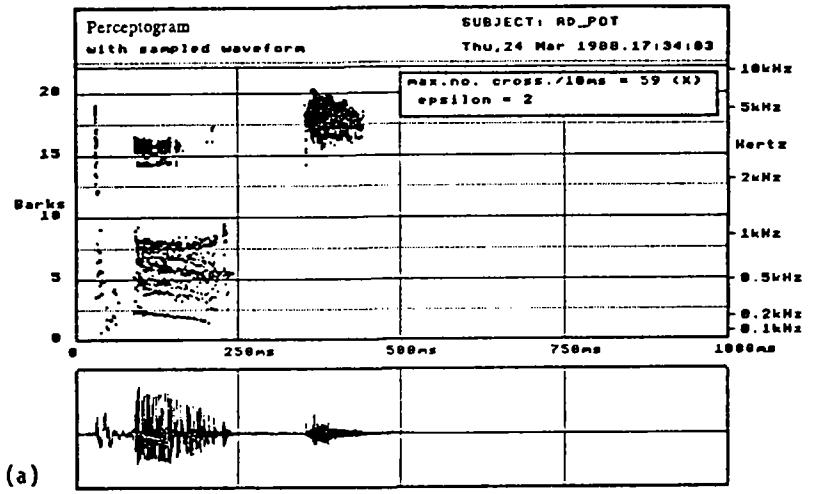


Figure 4 Perceptogram Plots for the Triphone Showing Evidence of Normalisation
(a) Female Speaker
(b) Male Speaker