USE OF THE ERB SCALE IN PERIPHERAL AUDITORY PROCESSING FOR VOWEL IDENTIFICATION

D. H. DETERDING*

Department of Linguistics
Sidgwick Avenue
Cambridge CB3 9DA
England

ABSTRACT

Some previous systems for using knowledge of peripheral auditory processing in speech recognition have used the Bark scale. Here, the use of the ERB scale is compared with the Bark scale.

Vowel spectra are transformed in the manner suggested by Bladon and Lindblom. The resulting vowel representations using the two different scales are then compared for a whole-spectrum approach to speaker-independent vowel recognition.

The success rate for correct identification is quite high with either scale; but it is unlikely that the remaining errors could be overcome using this kind of whole-spectrum approach.

INTRODUCTION

In recent years, many researchers have investigated the use of models of the peripheral auditory system as the first stage in automatic speech recognition systems. It is argued that, if the speech can be transformed in a manner similar to the processing of the ear, the task of recognition will be made easier.

If such a transformation is to be used, it is important that it be as accurate as possible. In their suggested auditory transform, Bladon and Lindblom [1] use a Bark scale. Moore and Glasberg [2] suggest that their ERB scale (standing for Equivalent Rectangular Bandwidth) is more accurate. In this paper, a comparison is made of the effectiveness of using these two scales in producing auditorily-transformed spectra for speaker-independent vowel recognition.

BARK SCALE vs ERB SCALE

Plots of the two scales against a log Hertz scale are shown in Figures 1 and 2.

The principal differences between the two scales are: the width of the critical band estimated by Moore and Glasberg is smaller, so there are more ERBs below 5000

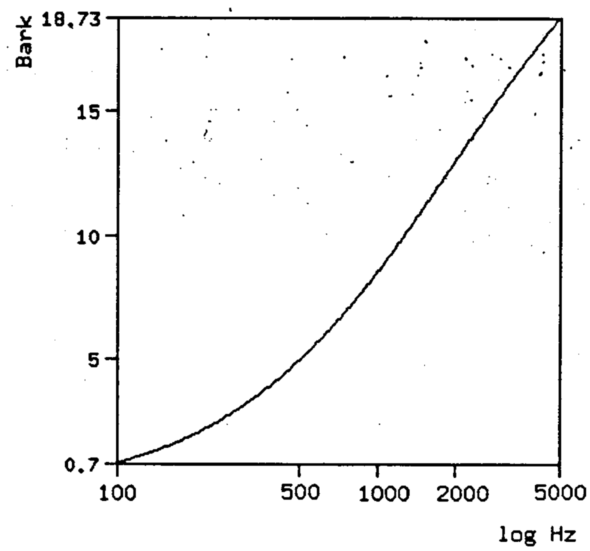

Figure 1. Plot of Bark scale against log Hz scale.

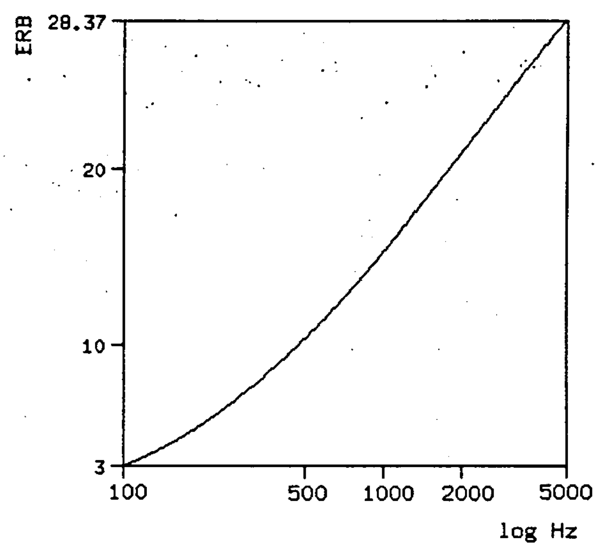

Figure 2. Plot of ERB scale against log Hz scale.

Hz than there are Bark; and the ERB scale deviates less from a logarithmic scale below 500 Hz.

One consequence of these differences is that, when vowel spectra are transformed to simulate aspects of peripheral auditory processing, the lower harmonics tend to be resolved on an ERB scale, whereas they are smoothed out on the Bark scale.

AUDITORY TRANSFORMS

In the experiment reported here, frames of 25.6 msec of speech were extracted from vowels uttered by a number of speakers. FFTs of these frames of speech then underwent transformations derived from models of the peripheral auditory system, and the final representations were used for attempts at automatic vowel recognition.

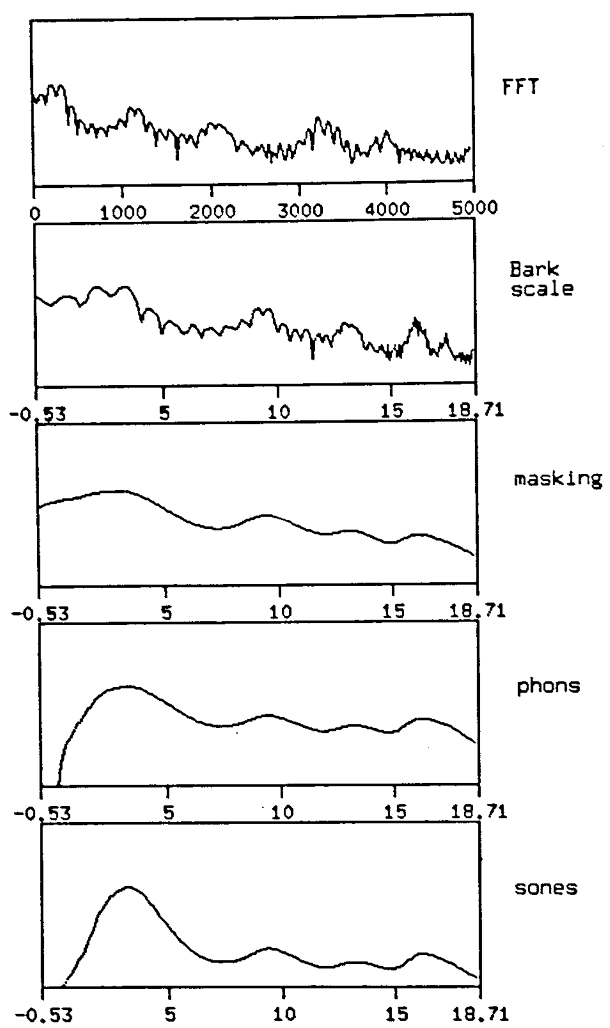

Figure 3. The effect of the various stages of the Bark transform for one token of "who'd".

For the Bark scale representations, the various stages of the Bladon and Lindblom transform were performed according to the formulae in [1].

To derive comparable representations for the vowels on the ERB scale, the formula for calculating excitation patterns from Moore and Glasberg [2] was used in place of the convolution of the masking filter in the Bladon and Lindblom model; but Moore and Glasberg provide no formulae for db-to-phons or phons-to-sones conversions, so these were taken directly from the Bladon and Lindblom model.

Examples of the various stages of the two transforms on the FFT spectrum of a frame of speech are shown in Figures 3 and 4.

It can be seen that the final ERB scale representation is less smooth than the final Bark scale representation. This is a consequence of the narrower masking filter suggested by Moore and Glasberg. It is possible that any harmonic ripple that has not been smoothed out could interfere with vowel identification; so a wider masking filter was also tried with the ERB scale. However, the success rate for vowel recognition using this wider filter was worse, so the results presented here for the ERB scale are for the narrower filter.

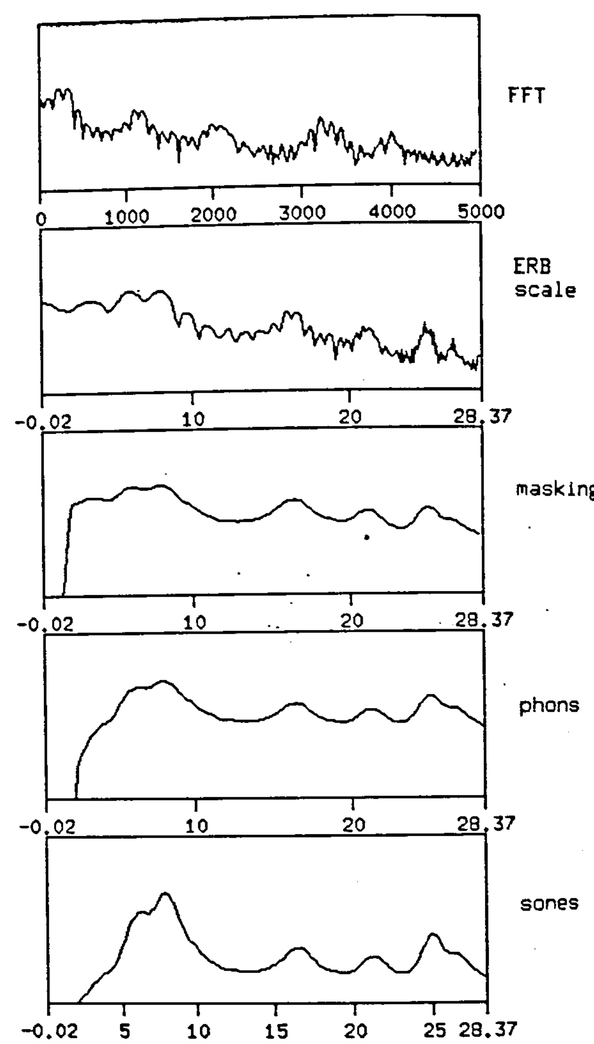

Figure 4. The effect of the various stages of the ERB transform for one token of "who'd".

NORMALIZING AUDITORY REPRESENTATIONS

Blomberg et al [3] find that, for vowel identification, the various stages in their auditory transform are actually destructive except for the last (DOMIN) stage; but they investigate recognition for each speaker independently, without attempting any kind of cross-speaker normalization. It is possible that an auditory representation only becomes important when speaker-independent recognition is attempted.

In the experiment reported here, identification of the vowels of each of thirteen speakers was based on templates derived from the vowels of the other speakers, so some kind of normalization was needed.

If speaker normalization can be achieved by a simple shift along an auditory scale to account for different vocal tract lengths [4], the shift required for adapting to one speaker from a set of templates should be appropriate for all the vowels of that speaker. Derivation of an appropriate shift can therefore be done on the basis of a single calibration vowel: the shift that allows the two representations of the calibration vowel to become most similar can be used for normalizing all the other vowels. This is comparable to the normalization scheme proposed by Nearey [5], though it uses an auditory scale instead of the logarithmic scale that he suggests.

Various vowels were tried as the calibration vowel for normalization, and the vowel from "hard" was found to provide the highest success rate. For the results presented, the calibration vowel was always "hard".

VOWEL RECOGNITION EXPERIMENT

Eight male and five female speakers, all using a Standard Southern British accent, each produced the words "heed", "hid", "head", "had", "hard", "hud", "hod", "hoard", "hood", "who'd", and "heard" in isolation. The frame of speech for use in the recognition was extracted from about one third of the way along each vowel. The location of this frame was determined manually, by examining the speech with a speech editor.

For identification of the vowels of each speaker, templates were derived by averaging the vowel representations of all the other speakers. For each vowel, identification was done by finding the template with a representation (after displacement by the normalizing shift) most similar to that of the vowel. The similarity of two vowel representations was determined by the Euclidean space between them.

RESULTS

The percentage of correct vowel identifications under various conditions is shown in Table 1. It is hard to draw clear conclusions about the superiority of either auditory scale from these results.

The success rate for vowel recognition after each of the various stages of the transforms is shown in Table 2. These figures suggest that each of the stages improves the recognition success rate, with the possible exception of the last stage. These findings differ from those of Blomberg et al [3].

The results in Table 1 show that the recognition performance for the female

	BARK	ERB
Normalized		
Male Only	89	92
Female Only	76	78
All	86	86
Un-normalized		
Male Only	90	94
Female Only	74	78
All	84	83

Table 1. Percentage of correct identifications under various conditions: in the "normalized" conditions, a normalizing shift was derived as described; in the "un-normalized" condition, no normalizing shift was used; in the "male" condition, the vowels of the male speakers were recognized using templates derived from the vowels of only the other male speakers; similarly for the "female" condition; in the "all" condition, the vowels of each of the speakers were used for identification of all the other speakers.

	BARK	ERB
FFT	64	64
auditory scale	74	73
masking	81	86
phons	83	87
sones	86	86

Table 2. Percentage of correct vowel identifications using the outputs of each of the stages of the transforms.

vowels was considerably worse than for the male vowels. Examination of the pattern of misidentifications showed that on both scales many of the vowels of one female speaker had been incorrectly identified. The possibility that the normalizing shift for this speaker was not optimal was then investigated.

All possible normalizing shifts, from minus 40 to plus 40 points, were tried. (One point represents 1/256 of the total spectrum, ie 0.075 Bark or 0.11 ERBs.) No shift allowed more than six (out of eleven) correct identifications on the Bark scale or seven on the ERB scale.

Even if, for this speaker, the templates were derived from only the other female speakers, the success rate was not perfect: no normalizing shift allowed more than eight correct identifications on either scale.

It seems that no simple normalizing shift will allow all the vowels of this speaker to be identified correctly.

It might be argued that the perception of some vowel distinctions lies mostly in the duration of the vowel, so, for example, for many speakers of Standard Southern British one cannot expect /ɑ:/ and /ʊ/ to be differentiated on the basis of a single extracted frame. But, with the best shift for this speaker using the female only templates, the remaining errors on both scales included:

/ae/ identified as /3:/
/u:/ /I/

These errors could not be resolved by considering the duration of the vowel.

DISCUSSION

Many of the vowel representations looked like that in Figure 5, with much less distinct peaks than those of Figures 3 and 4.

Given the amorphous shape of the vowel in Figure 5, the high success rate of the recognition was surprising. If a single normalizing shift is used with a whole spectral matching, it is doubtful if the success rate could be improved much beyond its present level.

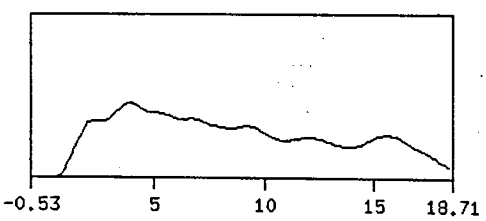

Figure 5. A Bark scale vowel representation of one token of "had".

Psychophonetic experiments [6] indicate that whole-spectrum-based vowel recognition is not likely to succeed because it retains spectral information that is relevant to the speaker's voice quality but not to the phonetic identity of the vowel. Spectral tilt, formant bandwidth, and even substantial changes in relative formant amplitude have little effect on phonetic vowel identity, but they have drastic effects on whole-spectrum matching scores. In obtaining better phoneme recognition scores than achieved here, Suomi [7] attempts to factor out the effects of spectral tilt from his whole-spectrum representations.

It is clear that some attempt must be made to find important features, principally the location of the formant peaks, and to use these for vowel recognition.

ACKNOWLEDGEMENTS

This paper is based on work carried out as part of the Linguistics Department's component (SERC grant GR/D/42405) of Alvey research project MM1/069 on Automatic Speech Recognition, which involves Cambridge University, the MRC Applied Psychology Unit, and Standard Telecommunication Laboratories Ltd.

REFERENCES

[1] R.A.W. Bladon and B. Lindblom, "Modelling the judgment of vowel quality differences", JASA 69 (5) pp. 1414-1422, 1981

[2] B.C.J. Moore and B.R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns", JASA 74 (3) pp. 750-753, 1983

[3] M. Blomberg, R. Carlson, K. Elenius, and B. Granström, "Auditory models as front ends in speech-recognition systems", in "Invariance and Variability in Speech Processes", (ed. J.S.Perkell and D.H.Klatt) Lawrence Erlbaum Assoc. pp. 108-114, 1986

[4] R.A.W. Bladon, C.J. Henton and J.B. Pickering "Outline of an auditory theory of speaker normalization", Proc. of 10th Int. Conf. on Phon. Sciences, Utrecht, pp. 313-317, 1983

[5] T.M. Nearey, "Phonetic Feature Systems for Vowels", Indiana University Linguistics Club, 1978

[6] R. Carlson and B. Granström, "Model predictions of vowel dissimilarity", STL-QPSR 3-4/1979 pp. 84-104, 1979

[7] K. Suomi, "Whole spectrum vowel normalization", Speech Communication 3, pp. 199-209, 1984