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 Hz.
He than are bark; and the EBB 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 smoothed out on the Bark scale.

**AUDITORY TRANSFORMS**

In the experiment reported here, frames 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.

For the Bark scale representations, the various stages of the Bladon and Lindblom transform were performed according to the formulas in [11].

To derive comparable representations for the vowels on the EBB 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 in the Bladon and Lindblom model, the Bark transform for one token of I was derived from the calibration vowel for normalization, and in the EBB scale, the Bark transform for one token of 1 was derived from the calibration vowel for normalization, and for each vowel, identification was done by finding the template by the normalizing shift (after displacement of that of the vowel). The similarity of two vowel representations was measured by the Euclidean space between them.

**NORMALIZING AUDITORY REPRESENTATIONS**

Blomberg et al. [3] find that, for vowel identification, the various stages of the Bark auditory transformation 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 a single template to a single speaker could 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 to become most similar can then be used for normalizing all the other vowels. This is comparable to the normalization scheme proposed by Neavey [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 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 100 tokens of "hat", "had", "hard", "hut", "hod", "hood", "who", and "whod" in isolation. The frame of speech for use in the recognition was extracted from 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 by the normalizing shift (after displacement of that of the vowel). The similarity of two vowel representations was measured 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 speaker was best, with a success rate of 89% for "hat" and 86% for "had". For the male speaker, the success rate was 92% for "hat" and 86% for "hard". The difference in performance between the sexes was statistically significant, with a p-value of 0.01. This result supports the hypothesis that auditory representations only become important when speaker-independent recognition is attempted.
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, i.e. 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 /æ:/ and /o:/ 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:

/æ/ identified as /3:/
/u:/ identified as /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".