SPEAKER NORMALISATION FOR A VOWEL IDENTIFIER BASED ON ROBUST FORMANT EXTRACTION

D.H. Deterding *

ABSTRACT

A speaker-independent method of recognising vowels by means of matching sets of formants against templates is introduced. The formants are derived by solving for the roots of linear prediction polynomials. There is a single template for each vowel. The method works well for male speakers but is not so successful with females.

INTRODUCTION

It has been shown that the phonetic identity of vowels is largely determined by the frequency of their formants (ref 2). However, there are two basic problems with a formant-based vowel identifier: formants are not always easy to locate; and their frequencies for different speakers may differ considerably. A method of overcoming these two problems is described in this paper.

FORMANT EXTRACTION

When formant frequencies are used for vowel recognition, there can be gross errors if the formants are mis-labeled. For example, it is sometimes difficult to locate both F1 and F2 in back vowels; and with nasalised vowels, there is usually an extra "spurious" formant in the F1 region. Many vowel recognition schemes avoid this problem by using the whole spectrum (ref 11), but in these schemes, it is difficult to isolate those acoustic features that contribute to the identity of the vowel from those that are part of the speaker's personal voice quality.

Hunt (ref 6) proposes a formant-based comparison measure that uses dynamic programming to find the best match between the formants of the test and those of the templates. This method allows formants to be deleted, and also for two formants in the template to be assigned to a single formant in the test (and vice versa).

Hunt derives formants by solving for the roots of short-frame linear prediction, and he determines the placement of each frame on the basis of the minimum normalised error. Unfortunately, with this method some of the frames may be located in the open phase of the glottal cycle, and this introduces considerable unreliability to the linear prediction analysis.

An improved method of achieving pitch-synchronous short-frame linear prediction is described in ref 3. Briefly, the method is as follows: a robust pitch extractor supplies information on pitch, cross-correlation is used to locate the same point in each pitch period, and the best offset from this point is found by the minimum residual error throughout the voiced section of speech. Using this method, the frames were consistently located in the closed-glottis phase, but the stability of the resulting spectral analysis was found to be disappointing when compared with a long-frame asynchronous analysis.

This pitch-synchronous method was used for the experiment described here, to allow a quantitative comparison with an asynchronous long-frame method of the success rate in vowel identification. For both pitch-synchronous and asynchronous analysis, 10th order covariance method linear prediction was used. For the pitch-synchronous analysis, the frame size was 30 samples (at 10 kHz sampling rate); for the asynchronous analysis, the frame size was 200 samples, with 100 samples overlap. For both methods, the beginning and end of voicing was determined in each word by means of the pitch extractor, and a set of roots and their magnitudes was derived for each frame starting at 20% into the stretch of voicing and ending at 65%.

* Department of Linguistics, Cambridge University, Sidgwick Avenue, Cambridge, CB3 9DA, UK
SPEAKER NORMALISATION

The acoustic variability in vowels arising from differences in personal voice quality is largely eliminated if formants are used for vowel identification. However, the formant frequencies of the same vowels spoken by different speakers vary widely (ref 10). When measuring the formants of a number of trained phoneticians, Ladefoged (ref 7) found that the [o] of one speaker had identical values for the first two formants as the [ɔ] of another speaker. It is clear that some kind of normalisation of formant values is required.

Ladefoged (ref 7) concluded that the phonetic quality of a vowel can only be determined by comparison with the speaker's other vowels, and many vowel normalisation schemes require prior knowledge of the vowel space used by the speaker (refs 5, 8, 9). To achieve this kind of normalisation for automatic speech recognition, the speaker would be required to produce a calibration phrase which would contain sufficient information to define the speaker's vowel space.

However, listeners are able to identify isolated vowels without any prior information about the speaker's vowels (ref 1), so the information required for achieving normalisation must exist within the vowel itself.

In the vowel identifier described here, speaker normalisation is attempted by a uniform shift along a log Hertz scale. This allows compensation for varying vocal tract lengths, though not for the non-uniformities between male and female vowels (ref 4). A log Hertz scale is preferred to a Bark scale, because compensation for vocal tract length can be achieved by a simple shift on the former but not the latter, and a simple shift on a Bark scale does not account for the male-female non-uniformities. In fact, if the average values from Peterson and Barney (ref 10) are plotted on a Bark scale, these non-uniformities are increased.

No automatic method has yet been implemented for estimating the shift required. Currently the best shift for each speaker is determined by trial and error and held in a file. Two possible methods of deriving this shift automatically may be implemented in the future: it can be found by using one or more calibration vowels and finding which shift allows their representation to become closest to the templates; or it might be derived from direct estimation of vocal tract length (ref 12).

TEMPLATE DERIVATION

A single template was derived for each vowel. For each word to be recognised, all the frames were compared with each of the templates in turn. The template that provided the lowest total matching score was output as the vowel identified.

For the calculation of a score for the match between a set of formants and a template, costs of assignment and deletion of formants have to be supplied. In this experiment, the assignment cost was proportional to the log Hertz frequency difference between the two formants. The deletion cost was the fourth power of the magnitude of the formant on the unit circle. The relationship between assignment and deletion costs was set so that the cost of assignment of a 1000 Hz formant to a 2000 Hz formant was twice the cost of deleting a formant with magnitude of 1.0. All these costs were ad hoc, but they seemed to work well.

The values of the template formants were initially set at published average values and were then adjusted manually to allow the best recognition rate. This manual adjustment was done separately for the asynchronous and synchronous methods. An automatic method of training the templates is currently being investigated.

SPEECH DATA

Eight male and five female speakers of Standard Southern British each produced one token of the following words: "heed", "hid", "head", "had", "hud", "hard", "hod", "hoard", "hood", "who'd", "heard". These were digitised and held on the computer.

Five subjects listened to the tokens in random order and selected one of the key words after listening to each token. They could repeat each token as many times as they liked. The random order was different for each listener. A confusion matrix for these listeners is shown in Table 1.
Many of these errors reflect slightly deviant pronunciations. For example, one speaker had very open versions of "hid" and "head", and it is not surprising that these were mis-identified as "head" and "had" respectively.

The scheme for automatic vowel recognition presented in this paper is based only on spectral information and does not make use of any dynamic information to help differentiate long and short vowels. Therefore, it would be surprising if the automatic method could consistently perform as well as the humans.

RESULTS

The percentages of successful recognition are shown below:

<table>
<thead>
<tr>
<th></th>
<th>normalisation</th>
<th>synchronous</th>
<th>asynchronous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male speech, male templates</td>
<td>no</td>
<td>94.3 %</td>
<td>97.7 %</td>
</tr>
<tr>
<td>Female speech, male templates</td>
<td>no</td>
<td>32.7 %</td>
<td>47.3 %</td>
</tr>
<tr>
<td>Female speech, male templates</td>
<td>yes</td>
<td>65.5 %</td>
<td>63.6 %</td>
</tr>
<tr>
<td>Female speech, female templates</td>
<td>yes</td>
<td>74.6 %</td>
<td>78.2 %</td>
</tr>
</tbody>
</table>

RESULTS FOR MALE SPEECH

For recognition of the male speech, the templates were derived by considering the male speech only. No normalising shifts were needed for a very high rate of correct identification. This was surprising, as the speakers differed widely in height and the pitch they used.

In the asynchronous test, out of the 88 tokens, only two were mis-identified:

- "hard" identified as "hud"
- "who'd" identified as "heed"

The first of these would need dynamic information to correct. For the second, the speaker had a very fronted /u/, almost [y]. This might be regarded as a deviant pronunciation in Standard Southern British, though only one listener made this error.

For the synchronous test, three additional errors were made:

- "hid" identified as "head"
- "head" identified as "had"
- "hud" identified as "had"

Of these errors, three subjects in the listening test made the same mis-identification of that token of "hid", two mis-identified the token of "head" as "had", and two mis-identified the token of "hud" as "had". Therefore, all these errors can be explained as deviant pronunciations.

RESULTS FOR FEMALE SPEECH

No simple normalising shift allowed a high recognition rate for the female speech using the male templates.

Under the "female speech, female templates" condition, female templates were derived separately; but even then it was not possible to find any templates that would allow a high recognition rate. The confusion matrix for this condition is shown in Table 2.

Visual examination of the female data showed perfectly clear formant tracks in nearly all of the tokens. It is not yet known whether there is greater variability in the formants of the female speakers or whether the formant matching procedure has for some reason failed. If the former explanation were true, it would be hard to explain why listeners have no more difficulty identifying the female tokens.
CONCLUSION

The success rates shown above do not represent true recognition rates, because only the training data have been tested. However, it seems likely that the method of recognition outlined above will allow a high success rate with male vowels.

Further work is required to determine how female vowels are to be recognised. A way must be found of allowing female vowels to be recognised using female templates before normalisation from male templates can be attempted.

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 STC Technology Ltd.

REFERENCES