PERCEPTUAL MAPPING AND VOWEL NORMALISATION

Robert H. Mannell
Speech, Hearing and Language Research Centre
Macquarie University, Sydney, Australia

ABSTRACT

A large number of vowel tokens in an /h_d/ frame were synthesised using a formant synthesiser. Each experimental condition consisted of a simulated "speaker" for whom three parameters, vowel space size, F0 range and higher formant frequencies, were characteristic of a male, female or "neutral" voice. These three parameters were either matched or mis-matched with each other in terms of their target vocal gender. For each "speaker" the tokens were evenly spaced on the F1/F2 plane. For each condition 20 speakers were asked to identify the English vowel that each token most sounded like. The resulting vowel phoneme perceptual maps were compared to examine the interacting effects of each of the parameters on vowel normalisation.

INTRODUCTION

Mannell [1] examined the perceptual mapping of Australian English long and short monophthongs. In that experiment there were two conditions which simulated a male and a female speaker. The male speaker's characteristics were based on measurements of 19 Australian English vowels spoken in /h_d/ context by 172 male speakers [2][3]. From these measurements it was possible to determine the limits of the male Australian English vowel acoustic space (henceforth referred to as the male frame) on the F1/F2 plane as well as appropriate F3 values for each F2 value. From these measurements a vowel plane was derived in F1/F2/F3 space and two sets of vowel tokens were derived on this plane representing long (300 ms) and short (150 ms) monophthongs in an /h_d/ frame. Each set of "male" vowels consisted of all possible vowel qualities on the F1/F2 plane separated by 100 Hz in both dimensions and within the constraints imposed by the defined male frame. F4 and F5 were fixed at 3500 Hz and 4500 Hz respectively. The female frame was derived from the male frame by multiplying the F1 and F2 maxima and the higher formant values by 1.2 to produce a larger frame size. The resulting frame size was compared with measurements obtained from 12 female speakers of Australian English to confirm that the derived vowel space was a valid representation of actual female data. The F0 contour was held constant for all tokens (male and female) at an average "gender neutral" value of 160 Hz. All of the tokens were generated by a parallel formant synthesiser [4] using specialised synthesis-by-rule software written especially for this experiment.

Perceptual contours (25%, 50% and 75% identification) were derived for every vowel phoneme and the resulting contour maps for the male and female long and short vowel spaces were compared. Particular attention was paid to the 50% identification contours or "predominance boundaries" [5] within which the identification of a particular phoneme predominated (ie. ≥ 50%). The male and female perceptual spaces were very similar in shape, differing mainly in the size of the spaces. The female spaces were shown to closely match the male spaces when both the long and short vowel female spaces were uniformly divided by a factor of 1.2. The match was even closer when a -50 Hz correction was made for the normalised female F1 values (possibly correcting for differences in male and female oral/pharyngeal tract lengths). Normalisation to a particular vocal type (in this case, simulated male and female voices) had clearly occurred as there were numerous vowel pairs that were identical in terms of their F1/F2 values but which resulted in consistently different vowel phoneme identifications for the male and female voices. For example, the "long" (300 ms) vowel with an F1 of 600 Hz and an F2 of 1900 Hz lies within the female /ɜ:/ predominance boundary and in the male /æ/ predominance boundary. Clearly some vocal factor or combination of factors has triggered different normalisation strategies for these two vowels based on the listeners' perception of differences between the two "speakers". Since F0 has been held constant for this experiment the trigger for the different normalisation strategies must depend upon one or both of the only two parameters which differentiate the two "speakers" in this experiment, vowel frame size and higher formant frequencies.

To examine whether the vowel frame size or the higher formant values had the stronger effect on vowel normalisation, Mannell [1] presented a series of "male" vowels representing a selection of vowels across the entire male vowel space. This was then followed by a series of 33 vowels which had "female" higher formant values but which had F1/F2 values which wholly fit within the male vowel frame. These 33 vowels were then followed by the female version of the vowel referred to above, which had an F1 of 600 Hz and an F2 of 1900 Hz and typically female higher formant values. It was assumed that 33 preceding vowels would be sufficient to alert the listeners to the new voice and to familiarise them with the "female" voice. This familiarisation would not, however, be based on the vowel frame size information as the test vowel would not be preceded by any vowels with F1/F2 values outside the male frame and thus exclusive to the larger female frame. If the listeners normalised fully to the female voice then this vowel should be heard as /ɜ:/, if the normalisation was based on the preceding male voice then the vowel should be heard as /æ/, and if the higher formants were responsible for partial shifting of the normalisation strategy towards that for the female voice then a mixture of /ɜ:/ and /æ/ responses should occur. The result was that 17 out of 20 subjects perceived /æ/ and there were no /ɜ:/ responses (the remaining three subjects heard /e/). The insertion of one high F2 (non-male-frame) vowel in the list of vowels preceding the test vowel reversed this effect, with more than 50% of the subjects perceiving the vowel as /ɜ:/ as would be appropriate for a female voice (this effect will be examined in more detail in future experiments). What seemed clear from is result was that normalisation appears to be strongly influenced by the listener's determination of the vowel frame size. Further, only one high F2 front vowel appears to be necessary to establish appropriate normalisation procedures. This last observation is consistent with the point normalisation hypothesis of Nearey [5].

These experiments, whilst pointing out the importance of vowel frame size in the normalisation of vowels, did not examine the effect of F0 on normalisation, nor did they examine the ways in which vowel-frame, F0 and higher-formant parameters interact during the process of vowel normalisation.

METHOD

In the present experiment the same procedure was followed as outlined on the first page of this paper, but with the following differences. Firstly, whilst the points on the male spaces were still separated by 100 Hz, on the female spaces the individual points were separated by 120 Hz, resulting in similar numbers of tokens for the male and the female spaces. Secondly, and most importantly, there was a much larger number of "male" and "female" conditions. The conditions varied with respect to F0, vowel-frame-size and higher formant values. The F0

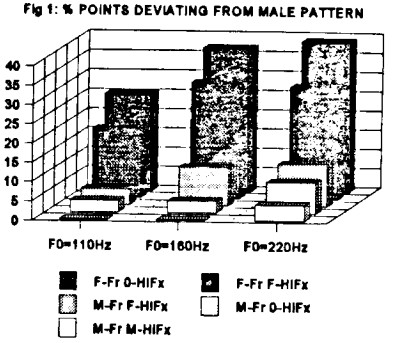
parameter was one of three pitch contours with mean F0 values of 110Hz ("male"), 160Hz ("neutral") and 220Hz ("female"). The vowel frame size parameter was either a male frame, or a female frame (as described above). The higher formants (F3/F4/F5) were either typically "male" or "female" or entirely absent (ie. F1/F2 two formant synthetic vowels). The conditions tested are summarised in table 1.

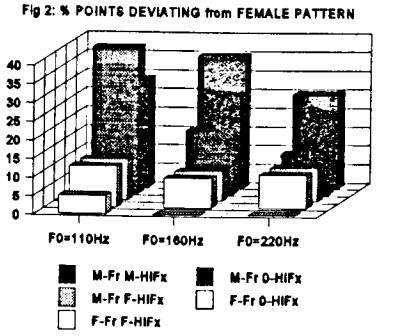
#	Frame	F0	Hi Fx
1	M	110	M
2	M	110	absent
3	M	160	M
4	M	160	absent
5	M	220	M
6	M	220	absent
7	F	110	F
8	F	110	absent
9	F	160	F
10	F	160	absent
11	F	220	F
12	F	220	absent
13	M	110	F
14	M	160	F
15	M	220	F

Table 1. Summary of experimental conditions.

300 listening subjects (phonetically naive, native speakers of Australian English, 20 subjects per test condition) were asked to identify the each token orthographically from a closed set of possible responses. The tokens were presented individually via headphones in a sound treated room. Predominance boundaries were determined for each vowel monophthong phoneme for each condition to produce a set of 30 perceptual maps (15 conditions, long vs short vowels). Data-point-by-data-point χ^2 comparisons of points within the area common to the male and female spaces were made for each relevant pair of conditions. Differences between conditions were determined as the number of data-points significantly different at p=0.01 and are expressed below as the percentage of total data points.

RESULTS





In the above figures "M-Fr" and "F-Fr" refer to male and female frames respectively, whilst "M-HiFx", "F-HiFx" and "0-HiFx" refer to male, female and missing F3/F4/F5 respectively.

DISCUSSION

The vowel frame size has the greatest effect on normalisation and thus vowel perception. This is most clearly seen when female frame data is measured relative to the most male condition (110Hz, male frame, male higher formants: see figure 1). This effect is also strong when male frame data is measured relative to the most female condition (220Hz, female frame, female higher formants: see figure 2) but in this case higher F0 values pull the percept for male-frame data much more strongly in the direction of the female pattern than low F0 pulls the percept for female-frame data in the direction of the male pattern. This can presumably be explained by the fact that all male-frame F1/F2 values could also be female values whilst the extreme front and low vowels in female-frame data cannot be perceived as male and so strongly mark the speaker as female.

An F0 of 110Hz tends to pull the perceptual pattern in the direction of the male pattern. Conversely an F0 of 220Hz tends to pull the perceptual pattern in the direction of the female pattern. This effect is strongest when F0 is reinforced by matching frame-size or higher formant values.

Appropriate male higher formant values reinforce the male perceptual pattern, missing higher formants weakens that pattern somewhat whilst inappropriately female higher formants for male-frame tokens has a strong effect on the perceptual space, pulling it in the direction of the female pattern. The male-frame tokens with female higher formants result in a perceptual pattern intermediate between the male and the female pattern (the perceptual patterns are as distant from the male pattern as they are from the female pattern). On the other hand, missing higher formants appears to consistently enhance the perceived femaleness of female-frame tokens relative to tokens with "female" higher formant values. This may be because the female high formant model utilised in this experiment is not a good model of female vowel productions and so confuse the listening subjects.

The maximum deviations between male and female perceptual patterns of no more than 40% is due to the overlap of the vowel spaces of central and back vowels. The deviations tend to occur at front and low vowel boundaries and result in the shifting of the boundaries to higher (female) or lower (male) frequencies.

CONCLUSION

All three parameters have some effect on normalisation processes during the perception of vowels. The frame-size parameter has the strongest effect but generally requires the support of at least one other factor (F0 or high formants) to produce the strongest male or female patterns. The effect of F0 on vowel perception is greatest when the vowel is otherwise ambiguous.

REFERENCES

[1] Mannell, R.H. (1988), "Perceptual space of male and female Australian English vowels", *Proc. 2nd. Australian International Conf. Speech Science and Technology*, Sydney, Nov. 1988, 22-27.

[2] Bernard, J.B. (1970), "Toward the acoustic specification of Australian English", *Zeitschrift fur Phonetik, Sprachwissenschaft und Kommunikations-forschung*, Band 23, Heft 2/3.

[3] Bernard, J.B. and Mannell, R.H. (1986), "A study of /h_d/ words in Australian English", *Working Papers*, SHLRC, Macquarie University.

[4] Clark, J.E., Summerfield, C.D. and Mannell, R.H. (1986), "A high performance digital hardware synthesiser", *Proc. 1st. Australian Conf. Speech Science and Technology*, Canberra, November 1986, 342-347.

[5] Nearey, T.M. (1977), *Phonetic feature systems for vowels*, PhD dissertation, Univ. of Connecticut.