PERCEPTUAL SPACE OF MALE AND FEMALE AUSTRALIAN ENGLISH VOWELS

R. H. Mannell
Speech, Hearing and Language Research Centre
Macquarie University

ABSTRACT - This study investigates the phonemic space of synthetic male and female vowel tokens as perceived by native speakers of Australian English. The data was also examined for evidence of vowel normalisation.

The acoustic specification of the male production space of Australian English (Aus.E.) vowels was carried out by Bernard (1970). To date, no exhaustive studies have been performed on the acoustic specification of the female Aus.E. production space. Further, there have been no studies on the perceptual space of male and female vowel phonemes. Such studies have been performed on several languages including American English (House and Stevens, 1957). A perceptual map of Aus.E. long and short vowels is a desirable aid to studies of perceptual confusions by Aus.E. listeners and may assist in the prediction of confusions of synthetic vowels produced by various synthesiser configurations. The House and Stevens (ibid) study of American English vowel perceptual space examined a series of synthetic vowels which were uniformly spaced in a three dimensional articulatory space. The vowels were produced on an electric vocal tract analog and the perceptual space was plotted onto both articulatory and acoustic (F1/F2) diagrams with individual vowel spaces indicated by iso- contours linking points of equal percentage correct identifications. In the present study all of the vowels are specified acoustically and are uniformly spaced on the F1/F2 plane. The male test space was derived from the production space derived by Bernard whilst a female space had to be defined in a more artificial manner due to the lack of female articulatory data.

It is well known that the acoustic signal of a single phoneme before entering the peripheral auditory system is highly variable (depending on speaker, context, etc...) and yet it is converted by the processes of speech perception into a single phoneme with considerable constancy. This problem of acoustic inconstancy versus perceptual constancy is one of the central issues which confronts the study of speech perception. The notion of "normalisation" has been an important feature of many models and is particularly favoured today by researchers into the problem of computer recognition of speech. Normalisation is in essence a transformation process which converts the incoming patterns into patterns which are more able to match up with those patterns (templates) stored by the listener. The presence of both male and female test tokens provided an opportunity for an extension of the study into a examination of the question of vowel normalisation.

Fant (1966) found that the formants of women are on average about 20% higher than those of men. He also demonstrated non-uniformity in the scaling of male and female area functions with the proportional length of the pharynx varying to a greater extent than that of the mouth cavity with the result that the female pharynx is proportionally shorter than that of males. This would have the effect, it was claimed, of causing a non-uniform scaling of formant values with F1 varying to a greater extent from males to females than does F2. Nordstrom and Lindblom (1975), using a normalisation procedure based on F3 were able to remove a lot of the differences between male, female and child vowels. They argued that the success of their normalisation procedure challenges the notion that non-uniform variation in area functions need not necessarily produce non-uniform variation in formant values. In response to the above study, Fant (1975) presented evidence based on six languages for what he described as "universal tendencies of departure from a simple uniform scaling" (ibid, p1) which were described to be due in part to "non-uniform scaling of vocal tract dimensions" as well as to "sex-specific articulation" (ibid, p1). He showed that the percentage difference between male and female F1 and F2 values increases with increasing formant frequency whilst the reverse appears to be the case for F3. Based on the data produced in this study, Fant (ibid) developed a non-uniform normalisation procedure which was shown to out-perform the uniform procedure of Nordstrom and Lindblom. Similarly, Matsumoto and Wakita (1986) also showed that linear scaling accounted for most of the required normalisation from female vowels to male reference vowels in a speech

PERCEPTION I

PERCEPTUAL SPACE OF MALE AND FEMALE AUSTRALIAN ENGLISH VOWELS

R. H. Mannell
Speech, Hearing and Language Research Centre
Macquarie University

ABSTRACT - This study investigates the phonemic space of synthetic male and female vowel tokens as perceived by native speakers of Australian English. The data was also examined for evidence of vowel normalisation.

The acoustic specification of the male production space of Australian English (Aus.E.) vowels was carried out by Bernard (1970). To date, no exhaustive studies have been performed on the acoustic specification of the female Aus.E. production space. Further, there have been no studies on the perceptual space of male and female vowel phonemes. Such studies have been performed on several languages including American English (House and Stevens, 1957). A perceptual map of Aus.E. long and short vowels is a desirable aid to studies of perceptual confusions by Aus.E. listeners and may assist in the prediction of confusions of synthetic vowels produced by various synthesiser configurations. The House and Stevens (ibid) study of American English vowel perceptual space examined a series of synthetic vowels which were uniformly spaced in a three dimensional articulatory space. The vowels were produced on an electric vocal tract analog and the perceptual space was plotted onto both articulatory and acoustic (F1/F2) diagrams with individual vowel spaces indicated by iso- contours linking points of equal percentage correct identifications. In the present study all of the vowels are specified acoustically and are uniformly spaced on the F1/F2 plane. The male test space was derived from the production space derived by Bernard whilst a female space had to be defined in a more artificial manner due to the lack of female articulatory data.

It is well known that the acoustic signal of a single phoneme before entering the peripheral auditory system is highly variable (depending on speaker, context, etc...) and yet it is converted by the processes of speech perception into a single phoneme with considerable constancy. This problem of acoustic inconstancy versus perceptual constancy is one of the central issues which confronts the study of speech perception. The notion of "normalisation" has been an important feature of many models and is particularly favoured today by researchers into the problem of computer recognition of speech. Normalisation is in essence a transformation process which converts the incoming patterns into patterns which are more able to match up with those patterns (templates) stored by the listener. The presence of both male and female test tokens provided an opportunity for an extension of the study into a examination of the question of vowel normalisation.

Fant (1966) found that the formants of women are on average about 20% higher than those of men. He also demonstrated non-uniformity in the scaling of male and female area functions with the proportional length of the pharynx varying to a greater extent than that of the mouth cavity with the result that the female pharynx is proportionally shorter than that of males. This would have the effect, it was claimed, of causing a non-uniform scaling of formant values with F1 varying to a greater extent from males to females than does F2. Nordstrom and Lindblom (1975), using a normalisation procedure based on F3 were able to remove a lot of the differences between male, female and child vowels. They argued that the success of their normalisation procedure challenges the notion that non-uniform variation in area functions need not necessarily produce non-uniform variation in formant values. In response to the above study, Fant (1975) presented evidence based on six languages for what he described as "universal tendencies of departure from a simple uniform scaling" (ibid, p1) which were described to be due in part to "non-uniform scaling of vocal tract dimensions" as well as to "sex-specific articulation" (ibid, p1). He showed that the percentage difference between male and female F1 and F2 values increases with increasing formant frequency whilst the reverse appears to be the case for F3. Based on the data produced in this study, Fant (ibid) developed a non-uniform normalisation procedure which was shown to out-perform the uniform procedure of Nordstrom and Lindblom. Similarly, Matsumoto and Wakita (1986) also showed that linear scaling accounted for most of the required normalisation from female vowels to male reference vowels in a speech

PERCEPTION I

PERCEPTUAL SPACE OF MALE AND FEMALE AUSTRALIAN ENGLISH VOWELS

R. H. Mannell
Speech, Hearing and Language Research Centre
Macquarie University

ABSTRACT - This study investigates the phonemic space of synthetic male and female vowel tokens as perceived by native speakers of Australian English. The data was also examined for evidence of vowel normalisation.

The acoustic specification of the male production space of Australian English (Aus.E.) vowels was carried out by Bernard (1970). To date, no exhaustive studies have been performed on the acoustic specification of the female Aus.E. production space. Further, there have been no studies on the perceptual space of male and female vowel phonemes. Such studies have been performed on several languages including American English (House and Stevens, 1957). A perceptual map of Aus.E. long and short vowels is a desirable aid to studies of perceptual confusions by Aus.E. listeners and may assist in the prediction of confusions of synthetic vowels produced by various synthesiser configurations. The House and Stevens (ibid) study of American English vowel perceptual space examined a series of synthetic vowels which were uniformly spaced in a three dimensional articulatory space. The vowels were produced on an electric vocal tract analog and the perceptual space was plotted onto both articulatory and acoustic (F1/F2) diagrams with individual vowel spaces indicated by iso- contours linking points of equal percentage correct identifications. In the present study all of the vowels are specified acoustically and are uniformly spaced on the F1/F2 plane. The male test space was derived from the production space derived by Bernard whilst a female space had to be defined in a more artificial manner due to the lack of female articulatory data.

It is well known that the acoustic signal of a single phoneme before entering the peripheral auditory system is highly variable (depending on speaker, context, etc...) and yet it is converted by the processes of speech perception into a single phoneme with considerable constancy. This problem of acoustic inconstancy versus perceptual constancy is one of the central issues which confronts the study of speech perception. The notion of "normalisation" has been an important feature of many models and is particularly favoured today by researchers into the problem of computer recognition of speech. Normalisation is in essence a transformation process which converts the incoming patterns into patterns which are more able to match up with those patterns (templates) stored by the listener. The presence of both male and female test tokens provided an opportunity for an extension of the study into a examination of the question of vowel normalisation.

Fant (1966) found that the formants of women are on average about 20% higher than those of men. He also demonstrated non-uniformity in the scaling of male and female area functions with the proportional length of the pharynx varying to a greater extent than that of the mouth cavity with the result that the female pharynx is proportionally shorter than that of males. This would have the effect, it was claimed, of causing a non-uniform scaling of formant values with F1 varying to a greater extent from males to females than does F2. Nordstrom and Lindblom (1975), using a normalisation procedure based on F3 were able to remove a lot of the differences between male, female and child vowels. They argued that the success of their normalisation procedure challenges the notion that non-uniform variation in area functions need not necessarily produce non-uniform variation in formant values. In response to the above study, Fant (1975) presented evidence based on six languages for what he described as "universal tendencies of departure from a simple uniform scaling" (ibid, p1) which were described to be due in part to "non-uniform scaling of vocal tract dimensions" as well as to "sex-specific articulation" (ibid, p1). He showed that the percentage difference between male and female F1 and F2 values increases with increasing formant frequency whilst the reverse appears to be the case for F3. Based on the data produced in this study, Fant (ibid) developed a non-uniform normalisation procedure which was shown to out-perform the uniform procedure of Nordstrom and Lindblom. Similarly, Matsumoto and Wakita (1986) also showed that linear scaling accounted for most of the required normalisation from female vowels to male reference vowels in a speech

recognition procedure. However, when a non- linear ("warped") scaling procedure was added to this, further improvement was obtained.

Ladefoged and Broadbent (1957) on the other hand present evidence which demonstrates that the phonemic identification of a vowel can depend on the "relationship between the formant frequencies for that vowel and the formant frequencies of other vowels pronounced by that speaker" (ibid, p98). In other words an auditory field is built up from the listener's experience of the speaker's vowel system and then any further vowels are identified by being placed within that field. Nearey (1977) contrasts range normalisation with point normalisation. Range normalisation requires that the two (or more) reference vowels must occupy a known position in the speaker's vowel space and that the normalisation procedure scales the unknown vowel's formant values relative to the formants of these reference vowels to give its place in the overall range. Point normalisation only requires one known point in the speaker's vowel system. Shifting the formant frequencies of the reference vowel can cause shifts in the phoneme perceptual boundaries for two-formant synthetic vowels.

METHODOLOGY

A male Aus.E. vowel F1/F2 space was defined using the male vowel production data of Bernard (1970). The limits of the space were defined by the most extreme positions produced when 2 standard deviation ellipses were drawn around each of the monophthong means (Bernard and Mannell, 1986). This produced a non- rectangular vowel area with an F1 range of 200-900 Hz and an F2 range of 800-2600 Hz (see figure 1). A synthetic vowel token was produced at each full multiple of 100 Hz in both the F1 and F2 dimensions giving 112 male data points. Values closer to the formant frequency difference limens measured by Flanagan (1955) were considered. Flanagan's results suggest that it is possible to distinguish changes of formant frequency of about 20 Hz at very low frequencies (300Hz) and this ranges up to a just noticeable difference (j.n.d.) of about 100 Hz at about 2000 Hz. His values for quality j.n.d.'s ranged from 2% to 5% of the formant frequency with an average of about 3%. Using these values to define the step sizes would have greatly increased both the number of data points and the time taken to both present and to analyse the data. Fortunately, the aim of this experiment is not, as it was for Flanagan, to examine the j.n.d.'s for vowel quality but to examine the points at which phonemic judgments change. As Flanagan noted, the difference limens for quality changes are certain to be much smaller than the difference limens for phoneme identification.

A female vowel F1/F2 space was created by multiplying the extreme F1 and F2 values by 1.2, following Fant's (1966) observation that female formants are on average 20% higher than male formants and that formant frequencies are inversely proportional to vocal tract length. This conversion should approximately predict the female vowel production space but should not be used for predicting the actual formant values of individual vowels owing to the non-uniform scaling between male and female vowel formants (Fant, ibid). The above procedure produced a vowel space bounded by F1 values of 200 to 1100 Hz and F2 values of 800 to 3200 Hz and included all the data points that were used in the male vowel space. This gave a total of 180 female data points compared with 112 male data points.

The F3 values for the male vowels were derived graphically from the approximate line of best fit through Bernard's vowel production data and fixing one F3 value for every F2 value. The female F3 values were derived by multiplying the male extrema and mid point F2/F3 values by 1.2 to give F3 values covering the entire female F2 range. Male F4 and F5 values were set to 3500 and 4500 Hz respectively whilst the female F4 and F5 values were 4200 Hz (3500 x 1.2) and 5400 Hz (4500 x 1.2) respectively. The female F5 value was out of the range of the synthesiser and so was switched off. An appropriate female high pole correction was also applied to the female data and the two consonants were modified to represent appropriate female values. All tokens were presented with the same pitch contour (160 Hz initially, falling to 145 Hz) to avoid the effects of pitch variation on vowel perception reported in some studies (eg. Holmes, 1986).

The two sets of data were synthesised on the Speech, Hearing and Language Research Centre parallel synthesiser (Clark, Summerfield and Mannell, 1986) in /h_d/ frames. Each token was produced as a long vowel (300 mS) and as a short vowel (150 mS) and the long and short tokens were randomised and presented to the subjects together. The male and the female tokens were, however, presented separately. All tokens were presented 5 seconds apart and there was a rest pause about every 10 minutes. The total