ACOUSTIC PROCESSING OF PHONETICALLY CONTROLLED VOWELS

Hiroaki Oasa and J.Bruce Millar

Computer Sciences Laboratory
Research School of Physical Sciences
Australian National University

ABSTRACT - A phonetically controlled vowel database, derived from 594 vowel samples spoken by adult males, adult females, and children, was analysed using LPC techniques to obtain a formant description of the database. The measured formants were uniformly transformed using various scaling factors derived from averaged acoustic features or from anatomical features. The effectiveness of these transformations as a first-stage normalisation procedure is evaluated, and the residual inter-speaker variation discussed.

CONCEPT

The work reported in this paper constitutes a part of a new approach to vowel formant normalization which is based on a theoretical model of inter-speaker formant variability. The model views observed formant variations within phonemically equivalent data in three major categories of (1) phonetic variation, (2) uniform variation, and (3) non-uniform variation. The phonetic variation includes any variation in vowel quality that is auditorily discernible. Uniform variation is that which is common to all vowels of a speaker and is attributed to the anatomical size variation that exists between speakers. Non-uniform variation is that variation that differs between vowels of a given speaker and is the remaining variation after the effects of both of the preceding factors have been removed. It may be attributed to vocal tract proportionality variation between speakers as well as specific articulatory changes made to meet perceptual requirements for different pitches.

This type of non-uniformity in speaker scaling factors has been investigated, most notably, by Fant (1966, 1973, 1975). Fant's 'k factors' were derived by averaging the non-uniform factors collected from up to six languages. The original scale factors for the six languages, however, were rather greatly at variance with one another. It was suggested that there were dialectal (i.e., phonetic) variations in the data (Fant 1975:4-5). Hence the k factors therein are considered to be phonetically 'contaminated'.

The present approach uses a database which is dialectally optimally homogeneous to begin with, and applies a rigorous phonetic control over it. The resultant phonetically screened data, which are expected to contain both uniform and non-uniform elements of variation, are processed through a range of uniform normalizations. The residual variation will then serve as optimal material for understanding the reasons for non-uniform formant variation between speakers.

DATA

A total of 594 vowel samples, consisting of 6 repetitions of 11 phonemically monophthongal vowels in the /hVd/ word context, uttered by 9 speakers of Australian English (2 male children, 4 female adults and 3 male adults), constitute the data. The speakers are members of a single large family from a sociolinguistically stable area of Adelaide. The choice of the speakers from this single large family was considered to have the following advantages. Firstly, the family's life style has been such that the members have spent much time together at home in an established valley community. This sociolinguistic environment would be conducive to minimizing inter-speaker phonetic variability. Secondly, as children and male and female adults of varying ages were sampled, there is a good contrast across the speakers in the major aspects of the anatomical structure of the vocal tract (e.g., size, proportionality), while other more peripheral, genetically idiosyncratic differences are likely to be small as compared with a sample from a population in general.

The reading list of 11 isolated /hVd/ words presented in orthographic forms (Bernard, et al. 1982) was arranged in two different sequences: for the first three repetitions, a phonetically contiguous sequence (anti-clockwise around the vowel quadrilateral), and for the last three repetitions, a phonetically contrastive sequence in which any neighbouring vowels contrasted sharply in height and frontness. This arrangement was adopted as a basic control to balance - or to elicit - possible contrast effects arising from a temporary juxtaposition of different vowels in a word list context.

Photographs showing each subject's profile were taken with a ruler held upright in front of the face by the subject to facilitate quantitative measurements of anatomical features. X-ray data would have been ideal for such measurements but were impracticable.

ACOUSTIC ANALYSIS

The data were digitized at the sampling frequency of 10 kilo-samples per second, and each word was manually segmented, labelled, and stored in an individual file. Linear prediction analysis was performed (order 14 for male samples and 12 for female and child samples); other numbers of coefficients were occasionally used to improve the accuracy of formant extraction. Formants were extracted by FFT analysis on the autoregressive coefficients which were obtained from the reflection coefficients for each frame of the sample. F0 was extracted by a cepstrally based periodicity estimation procedure.

The first four formant frequencies were measured at a single target for each vowel token. The target location was selected using the following criteria.

```
1) determine appropriate element of the vocalic nucleus
   if diphthongisation was present;

2) eliminate regions below an energy threshold and regions of
   apparent coarticulatory transition into the following /d/;

3) IF "first two formants have a steady state",
          THEN "select target during that time";
          ELSE IF "F1 contour has a maximum",
                    THEN "select target at F1 maximum location";
                    ELSE "select target at F2 minimum location".
```

This strategy was used to ensure that all tokens for a particular vowel category were measured at equivalent locations.

PHONETIC CONTROL

The samples were screened for phonetic variation by a pseudo-transcriptional method. In a software-controlled procedure, a set of 99 digitized samples (11 vowels x 9 speakers) were presented to a trained phonetician transcriber[1] in 10 different random sequences - this multiple transcription offered an opportunity to assess the extent of transcription variability. As there were six sets of 99 samples, a total of 5,940 transcriptions were obtained. The transcriptions were made by pressing a magnetic pen in a large cardinal vowel quadrilateral placed on a graphics tablet in response to each stimulus which was repeated three times. The primary transcription datum was the pair of coordinates of the pen's position when pressed on the tablet. The coordinates were stored in files together with other occasional information such as marked nasality and roundedness, and an estimate of confidence in the transcription, which was entered by pressing the pen in the appropriate menu box at another location on the tablet. The transcriptional records, therefore, were of a continuous nature, in contrast to traditional transcriptions, which are symbolic and therefore necessarily discrete. In the first stage of the processing of these data, transcription variation was reduced by a software-controlled interactive procedure whereby transcriptional outliers were identified and removed from the data, which were then averaged to obtain the final 594 representative transcriptions. Outlying variants in these representative transcription data were considered to be phonetically variant and were removed from the data in the subsequent phonetic control stage. The acoustic data corresponding to the surviving tokens, now considered to be phonetically homogeneous within each vowel category, were then used for the subsequent investigation of acoustic variation. Of the 594 samples, 469 were passed on to the acoustic processing.

It has been demonstrated that one of the weaknesses of phonetic transcription (with its human operator) is the lack of agreement in the judgement of the roundedness of vowels that are far away from the primary cardinal plane (Ladefoged 1967). This disagreement was most pronounced in the high vowel region, where a degree of roundedness different from that of a primary cardinal vowel was perceived by the transcribers as varying degrees of frontness and roundedness in combination.

In the present study, another area of relative insensitivity in transcription appears to be nasalization. Although some samples were occasionally labelled as nasalized, these judgements were not sufficiently consistent to warrant the removal of the suspected samples. However, acoustic spectra of some samples - e.g., low vowels of one female speaker - showed consistently lowered and broadened F1, sometimes with split peaks - a sign of nasalization where one or more pole-zero pairs which appear in the vicinity of F1, and sometimes of F2, interact with the formant pattern (cf. Hawkins and Stevens 1985). The exact pattern of this interaction is elusive as it is determined by the size of the coupled nasal orifices which differs between individuals.

Although the difference between the two elicitation sequences mentioned earlier was generally not reflected in the formant description, there were several exceptional instances in which two distinct clusters were found within a vowel category corresponding to the two elicitation list styles. The distribution of these instances indicates that the sensitivity to the list context is both speaker and vowel specific. This list context dependent perturbation descernible in the formant data was not always discernible auditorily.

The implication of these three areas of possible transcription insensitivity is that some finer variables of phonetic quality of vowels such as roundedness and nasalization cannot be reliably determined in auditory terms alone, and a further screening of data for these variables using acoustic phonetic criteria should be incorporated in the phonetic control process.

UNIFORM TRANSFORM

Parameters used for the uniform transformations of the first two formant frequencies were (1) F3 average, (2) F4 average, (3) F0 average, (4) F1 average and F2 average, and (5) photographically inferred anatomical size. F3 average over open vowels was used by Nordstrom and Lindblom (1975) in their uniform normalization, the use of F4 average was suggested by Ladefoged (1975), and the use of F0 average was motivated by the claim that the tonotopical distance between F0 (Bark) and F1 (Bark) is constant for a given degree of openness (Traunmuller 1981). The use of F1 average and F2 average is unreported in the literature, but it would be essentially the same process as Nearey's log-mean method under the Constant Log Interval Hypothesis 2, if it were performed in the log domain (Nearey 1978). (5) is an experimental attempt at an inference of the vocal tract length through a direct measurement of visible anatomical features. (1) through (4) were obtained by straightforward calculations from the available acoustic data. (5) imposed difficulty. Profile photographs were used to estimate the length of the vocal tract for each speaker, but the estimation of the location of the glottis by this method was not reliable even though the subjects' heads were held level and their jaws closed. An alternative anatomical measurement which can be made more reliably and is less susceptible to the changes of facial expressions, is the measurement of the distance between the side edge of the eye and the bottom of the chin in a straight vertical line (provided the subject's jaw is closed).

The uniform transformations were performed by the following equation:

$$Fn = F / Pi$$

where Fn = normalized formant frequency,
F = raw formant frequency,
Pi = parameter value for speaker i.

A more conventional method of uniform transformation is expressed as follows.

```
Fnn = F / SF
        where SF = Pi / Pref = scale factor,
              Pref= parameter value for the reference speaker,
              Fnn = normalized formant frequency.
```

It can be seen that the relationship between the two methods is as follows:

```
Fnn = Fn * Pref.
```

As Pref is a constant for a given transform parameter, Fnn is a linearly expanded version of Fn. In fact, once Fn has been obtained, the reference speaker (the speaker, real or idealized, to whom all the data are normalized) can be arbitrarily selected.

The transformation using F0 average was performed in the Bark domain (derived using the equation of Traunmuller (1983)).

EVALUATION OF NORMALIZATION PROCESS

The effectiveness of each of the uniform transformations was evaluated in two stages.

(1) Because of the presence of the non-uniform elements of variation in the data, the results of the uniform transformations are not expected to have zero variance. If it is assumed that the goal of uniform normalization is the maximal removal of the uniformly varying elements from the data, it is valid to regard the maximal reduction in the residual variance as the criterion for evaluating the uniform processes. Hence the first stage is to compute the reduction in variance due to each uniform transformation (Table 1). This is expressed as the ratio of the standard deviation of the residual variance to the mean of the residual variance, otherwise known as its coefficient of variation.

Table 1. Mean reduction in coefficient of variation after uniform transformation.

Parameters	F1	Parameters	F2
F1 average	38.9 %	F2 average	47.3 %
Anatomical	34.1 %	Anatomical	37.3 %
F4 average	29.8 %	F3 average	36.6 %
F3 average	19.1 %	F0 average	35.5 %
F0 average	5.0 %	F4 average	31.0 %

The most effective transforming parameter in stage 1 in terms of the mean reduction of the coefficient of variation in the residual variance was the F1 average for F1, and the F2 average for F2 with mean reduction rates of 38.9 % and 47.3 % respectively (Table 1). This result might have been expected as these parameter values are intrinsic to the formant data concerned.

It is noteworthy that the relatively ad hoc measurement of anatomical size yielded a result which compares favourably with all the other extrinsic parameters. It is also interesting to note that the reductions achieved by the F3 and F4 parameters seem related to their proximity to the formant being normalized. The F4 parameter has an almost identical effect on F1 and F2 of approximately 30 %. F3 has a significantly stronger effect on F2 variance than on F1 variance. The opposite is true of F0 which reduces F2 variance more than seven times as much as it reduces F1 variance. The underlying reason for this pattern of variance that is global to all vowels needs further study by examining the influence of subgroups of the speakers.

(2) Explanations are sought for each of the residual variance patterns after uniform transformation. If there is a satisfactory model for the residual, that uniform-transform/residual-model pair is a candidate

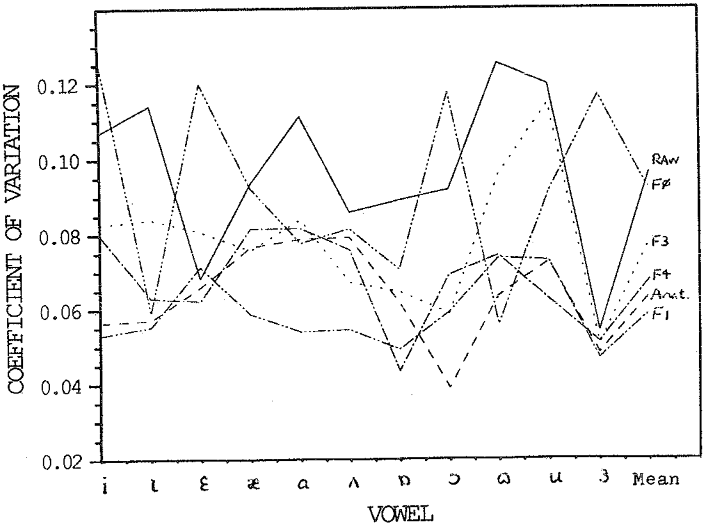

Figure 1. Coefficients of variation of normalized F1 values for all speakers for each normalization parameter and for each set of phonetically homogeneous vowels. Line types for both Figure 1 and 2 are as follows: raw data = solid; anatomical = dashed; F3 = dotted; F4 = dash and dot; F1 or F2 = dash and two dots; F0 = dash and three dots.

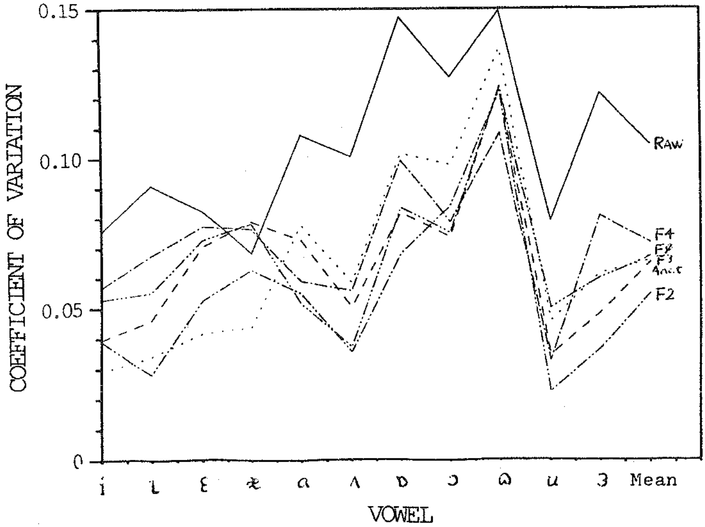

Figure 2. Coefficients of variation of normalized F2 values.

for a full description of the inter-speaker variation. If no plausible explanation can be found for the residual pattern, the uniform normalization by the particular parameter does not lead to a substantive understanding of the acoustic-phonetics of the vowel quality within the speakers sampled.

At the present time we offer some initial observations of the vowel dependent features of the coefficent of variation data (Figures 1 and 2). It should be noted that smoothly varying variance values across the phonetically contiguous vowel space indicate the prospect of phonetically interpretable factors within the residual variance following the uniform transformations.

There are two vowel categories, /ɛ/ and /ɜ/, whose relatively small variance in the raw F1 data is not markedly reduced (Fig 1). F0 is seen actually to increase the variance for these vowels. The major reason for this phenomenon is that, after most of the uniform normalizations, the data of the child and female group for these vowels were over-rescaled, and the relative positions of this group and the male adult group on the formant frequency scale were reversed. It is interesting to note that the two vowels concerned are both mid vowels which are in the same general F1 range.

Generally steadier pattern of reduction across the contiguous vowel categories is seen for F2 than for F1 (Fig 2). One exception is /æ/, whose relatively small variance in the raw data is not reduced by three of the parameters. The only external parameter that had any significant effect in reducing the variance for this vowel is F3. F3 appears to have the most consistent effect in variance reduction across all vowel categories. Low vowels and non-front vowels show particularly good reduction in variance except for /ʊ/ for which the reduction is moderate.

SUMMARY

This study represents a platform from which a phonetically principled study of vowel normalization can proceed. It has quantified performance differences of a variety of normalizing parameters suggested in the literature as a first stage, and has indicated some directions for establishing a substantively complete two-stage process of vowel normalization.

NOTE

(1) For evidence for the justification of using a single trained transcriber, see Laver (1965) and Ladefoged (1967).

REFERENCES

Bernard, J., Blair, D., Clark, J., Fraser, H., Guy, G. and Horvath, B. (1982) *Australian Speech Archive,* Occasional Papers, Speech and Language Research Centre, Macquarie University, December.

Fant, G. (1966) *A Note on Vocal Tract Size Factors and Non-uniform F-pattern Scalings,* STL-QPSR, 4, 22-30.

Fant, G. (1973) *Speech Sounds and Features,* (MIT Press, Cambridge, Mass.).

Fant, G. (1975) *Non-uniform Vowel Normalization,* STL-QPSR, 2-3, 1-19.

Hawkins, S. and Stevens, K. (1985) *Acoustic and Perceptual Correlates of the non-nasal-nasal distinction for vowels,* J. Acoust. Soc. Am. 77(4), 1560-1575.

Ladefoged, P. (1967) *Three Areas of Experimental Phonetics,* (Oxford University Press: London).

Ladefoged, P. (1975) *A Course in Phonetics,* (Harcourt Brace Jovanovich).

Laver, J. (1965) *Variability in Vowel Perception,* Language and Speech, Vol. 8, Pt 2, 95-121.

Nearey, T. (1975) *Phonetic Feature Systems for Vowels,* (Indiana University Linguistics Club, Bloomington, Indiana).

Nordstrom, P.-E. and Lindblom, B. (1975) *A Normalization Procedure for Vowel Formant Data,* Paper 212, VIII International Congress of Phonetic Sciences in Leeds.

Traunmuller, H. (1981) *Perceived Dimension of Openness in Vowels,* J. Acoust. Soc. Am. 69(5), 1465-1475.

Traunmuller, H. (1983) *On Vowels. Perception of Spectral Features, Related Aspects of Production and Sociophonic Dimensions.* Thesis, University of Stockholm.