AUTOMATIC VOWEL QUALITY DESCRIPTION USING FOUR PRIMARY CARDINAL VOWELS

Shuping Ran, Phil Rose, Bruce Millar and Iain Macleod*
Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University, Canberra, ACT 0200, Australia
** Linguistics Department, Faculties, ANU*

ABSTRACT

This paper investigates the possibility of describing vowels phonetically using an automated method. Models of the phonetic dimensions of the vowel space are built using two multi-layer perceptrons trained using four primary cardinal vowels. Test vowels processed by these perceptrons are placed onto a cardinal-like vowel chart. These automatically derived positions are compared with the positions of these vowels in a similar space as judged by a phonetician, and with the acoustic space derived from these vowels. The differences observed are discussed.

INTRODUCTION

Vowels are described in phonology and traditional phonetics with the three major parameters of height, backness and rounding, as well as additional parameters like nasality and tenseness. Although backness, height and rounding are often defined articulatorily, it is now widely assumed following Ladefoged [1] that the labels are primarily acoustic or perceptual, and relate to perceptually motivated transforms of F_1 (height) and effective F_2 (backness and rounding).

Vowels are traditionally described by phoneticians by listening to the vowels, and then placing a vowel symbol onto the cardinal vowel chart or assigning it appropriate diacritics according to learned auditory models. Figure 1 illustrates a three dimensional cardinal vowel system. This traditional method is very tedious, and is not feasible for non-phoneticians. This paper investigates the possibility of describing vowel quality without the skills of an experienced phonetician, using a novel method which automatically places a given vowel into a space which is defined by a set of reference vowels.

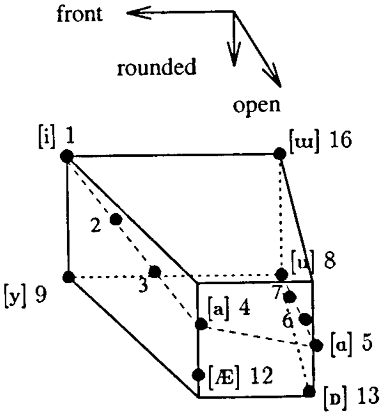

Figure 1. A three dimensional model of the vowel space (after Ladefoged [2])

A preliminary study [3] was carried out in which the vowels of four speakers of Australian English were analysed by this method. Models of each speakers' vowel space were trained using three reference vowels from an existing data corpus to encode the form of acoustic evidence for phonetic features which correlate with the dimensions of the vowel space (e.g. open-close, front-back). The reference vowels were chosen according to their relatively extreme positions on the cardinal vowel chart and their stability within Australian English. While the results of this study were encouraging, it was clear that the choice of the reference vowels was crucial for more accurate positioning of the vowels on the vowel chart.

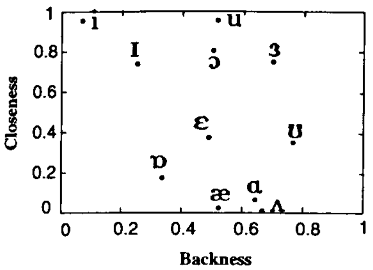

Figure 2. Test results averaged over six stop contexts of eight reference vowel model: 11 pseudo steady state vowels on a Closeness versus Backness plane.

In a further study [4], eight cardinal vowels which represented the extremes of the dimensions: front-back, open-close, rounded-unrounded, produced by an experienced phonetician were used for the model training. English vowels in stop consonantal context produced by the same speaker were used for testing. The results showed that the method worked well with respect to the front vowels, but badly for the back vowels (see Figure 2). It was suspected that this result was due to the lip rounding of some reference vowels introducing some misleading information into the models.

In the present study, we aim to minimise this potentially misleading information by choosing a different set of reference vowels.

REFERENCE VOWELS

The reference vowels used in this study were derived from the vowel model expressed by Figure 1. The aim was to use primary cardinal vowels that were maximally extreme on the two dimensions of front-back and open-close. The four primary cardinals (vowel 1 [i], 4 [a]; 5 [ɑ] and 8 [u]) fit this specification.

Five repetitions of each primary cardinal were recorded in a sound booth by our speaker, who is an experienced phonetician trained in the British tradition. The reference utterances were hand segmented. The parts of the signal where F_0 remained stable were used for this study. An $F_1/(F_2\text{-}F_1)$ plot was made of these vowels from conventional wide band spectrograms, as shown in Figure 3.

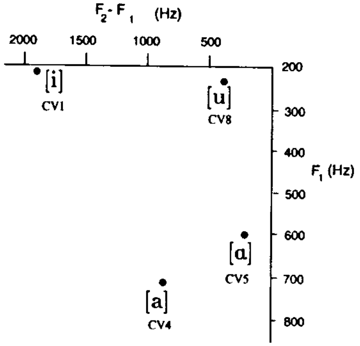

Figure 3. F_1 vs F_2-F_1 plot for phonetician's cardinal vowels CV 1 4 5 8.

ENGLISH VOWELS

Five repetitions of English vowels in the context of [stop][vwl]d utterances were produced by our speaker, where: [stop] represents one of the six phonemically voiced and voiceless labial, alveolar, and velar plosives of English (/b, p, d, t, g, k/); [vwl] represents one of the eleven nominally monophthongal phonemes (/i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ/); and d is /d/. The [stop][vwl]d utterances were manually segmented and labelled according to the procedures described by Ran [5]. Only the pseudo steady-state vowel interval was of interest for this study.

These vowels were transcribed by the phonetician, and placed on a traditional chart showing height and backness, with rounding indicated separately -- see Figure 4. This figure shows an unremarkable auditory configuration typical for the British English accent of the speaker, with some apparent influence from Australian English. Thus the /u/ is considerably fronted ([ʉ] >); the /ɔ/ is a close-mid [o]; the /ɛ/ is closer than open-mid, and the /ɜ/ is closer and more front. An F_1/(F_2-F_1) plot of the English vowels from conventional wide band spectrograms also reflects this pattern (Figure 5).

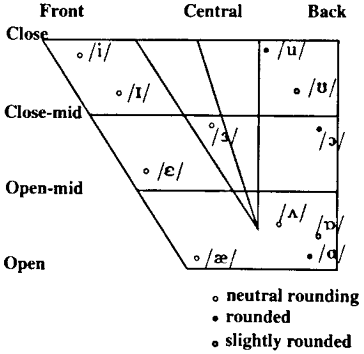

Figure 4. English vowel description by a phonetician.

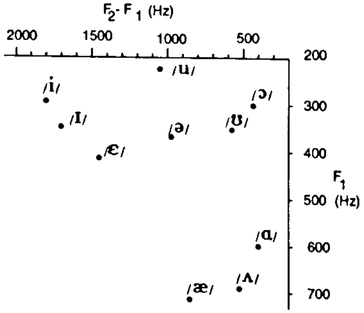

Figure. 5. F_1 vs F_2-F_1 plot for phonetician's English vowels in b-g *context.*

DATA PRE-PROCESSING

The data, including the pseudo steady-state English vowel intervals and the reference vowels, were processed in 'frames' of 12.8ms, with adjacent frames having a 6.4ms overlap, by passing them through a Hamming window, and then deriving 13 Linear Predictive Cepstral Coefficients (LPCCs) for each frame.

MODEL TRAINING

Two Multi-Layer Perceptrons (MLPs) were used to model the articulatory dimensions of front-back and open-close in order that they may be used as articulatory descriptors for backness and closeness. Each MLP was implemented with one hidden layer of two nodes and was trained by using the back-propagation algorithm. The inputs for this training comprised frames of four repetitions of the four reference vowels, and comparator outputs were their articulatory labels as shown in Table 1.

cardinal vowel	articulatory description	back	close
i1	front-close	0	1
u8	back-close	1	1
ɑ5	back-open	1	0
a4	front-open	0	0

Table 1. Articulatory labels for the reference vowels.

MLPs with one hidden layer were used because they are theoretically able to encode relationships of any complexity [6]. The number of hidden nodes was chosen by experiment starting with one hidden node, then incrementing the number one by one. The architecture which gave best performance on the training data was chosen. The number of hidden nodes for the backness and closeness descriptors was two.

VOWEL DESCRIPTION RESULTS

The cepstral data of the pseudo-static intervals of the English vowels were processed by the trained articulatory descriptors (i.e. the closeness descriptor and the backness descriptor), on a frame by frame basis. The outputs from the descriptors were the activation scores of the output nodes of the MLPs, which indicated with what probability a given input frame can be labelled with the articulatory label of the descriptor.

Figure 6 reports the results by combining the output from the two descriptors.

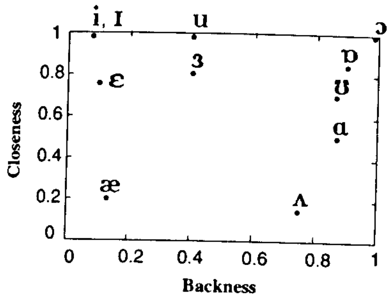

Figure 6. Test results averaged over six stop contexts of four reference vowel model: 11 pseudo steady state vowels on Closeness versus Backness plane.

The horizontal axis represents the backness, where the left represents maximal frontness and the right represents maximal backness. The vertical axis represents the closeness, where the top end represents maximal closeness and the bottom end represents maximal openness.

Analysis of Figure 6 reveals that, compared with the phonetician's auditory judgements (Figure 4) and the F_1/F_2-F_1 plot (Figure 5), the automatic method using the four reference vowels resolves the English vowels well. The positioning of the vowels approximates more closely to the positioning in the F_1/F_2-F_1 plot than to the positioning of the auditory judgements, especially for the back vowels.

Because of the restrictions of space, test results in individual contexts are not included here. The resolutions appear to be rather sensitive to differences in the consonantal frame. It can only be assumed that differential consonantal assimilation is occurring which is currently being studied.

Comparing the test results of eight reference vowel models [4] with that of four reference vowel models, the latter has improved substantially the description of the vowels, specially with respect to the back vowels. One noticeable problem is that some vowels (/i, ɪ, u, ɔ/) are positioned on the extremity of the maximum closeness which is unrealistic.

CONCLUSIONS

This study arose from our concern to improve the reference vowel set over that used in [4]. The results have clearly shown improved vowel positioning by choosing four primary cardinal vowels as reference vowels instead of all the eight cardinal vowels. The method provides a normalised system of automatic phonetic quality description. The challenges that remain include further understanding of the impact of consonantal context on the method and ways of accounting for it. It is also important to find ways of training naive speakers to produce reference vowels which may then be used to normalise automated phonetic description of their vowels.

REFERENCES

[1] Ladefoged, P. (1982) *A Course in Phonetics*, Second Edition, (Harcourt Brace Jovanovich:New York).

[2] Ladefoged, P. (1975), *Three Areas of Experimental Phonetics* (Fourth edition), (Oxford University Press, London).

[3] Ran,S., Millar,J.B., Macleod,I. (1994), "Vowel quality assessment based on analysis of distinctive features", *Proc. International Conference on Spoken Language Processing*, Yokohama, pp. 399-402.

[4] Ran,S., Rose, P., Millar, J. B. and Macleod, I. (1994), "Automatic vowel quality description using a cardinal vowel reference model", *Proc. of the Fifth Australian International Conference on Speech Science and Technology*, pp. 387-392.

[5] Ran, S. (1994), *Speech Knowledge Modelling for Speech Recognition: A Study Based on Distinctive Features*, PhD thesis, The Australian National University.

[6] Lippmann, R. P. (1987), "An introduction to computing with neural nets", *IEEE Trans. on Acoustics, Speech and Signal Processing*, 4(2), pp. 4-22.