AUTOMATIC VOWEL QUALITY DESCRIPTION USING A CARDINAL VOWEL REFERENCE MODEL

Shuping Ran, Phil Rose*, J.Bruce Millar and Iain Macleod
Computer Sciences Laboratory
Research School of Information Sciences and Engineering
Australian National University

*Department of Linguistics (Faculties), A.N.U.

ABSTRACT

This paper presents an analysis of the vowel space of a single speaker using an automated method of deriving the phonetic dimensions of the constituent vowels. The boundaries of a three dimensional space are defined by a set of eight extreme vowels. These vowels are used to train two multilayer perceptrons such that they encode in their inter-nodal weights the relationship between the dimensions of the vowel space and the acoustic characteristics of the extreme vowels. The English vowels of the speaker are then processed by the perceptrons. The activation levels of the output nodes are used to represent the position of each vowel within the vowel space. These automatically derived positions are then compared with the positions of these vowels in a similar space as judged by a phonetician, and the acoustic space derived from these vowels. The differences observed are discussed in the light of possible improvements in the procedure.

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 (e.g. 1982:201) 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).

This paper investigates the possibility of describing vowel quality without the skills of a well-trained phonetician, using a novel method which automatically places a given vowel into a space which is defined by a set of reference vowels.

A preliminary study (Ran et. al., 1994) was carried out in which the vowels of four speakers of Australian English were analysed using a set of Multi-Layer Perceptrons (MLPs) where each perceptron had been trained using three reference vowels from an existing data corpus to detect the presence of the acoustic evidence for the corresponding articulatory label. The reference vowels were chosen according to their relatively extreme positions on the cardinal vowel chart and their stability within Australian English. The results of this study suggested that the choice of the reference vowels might be crucial for more accurate positioning of the vowels on the vowel chart. For the present study, we carefully selected the reference vowels and developed the models as described below.

REFERENCE VOWELS

The eight cardinal vowels of Jones (1956) are assumed to span the whole vowel space of a speaker (Ladefoged, 1975). They range from an extreme front-close vowel produced with maximally spread lips to an extreme back-close vowel produced with maximally rounded lips, via six intermediate vowels of monotonically increasing roundedness comprising three pairs which exhibit one-third, two-thirds and complete degrees of openness. This three dimensional view is illustrated in Figure 1.

The reference vowels used in this study were derived from the vowel model expressed by Figure 1. The aim was to use vowels that were maximally extreme on the three dimensions of front-back, open-close,

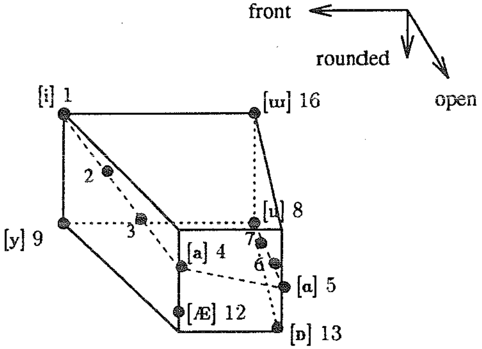


Figure 1: A three dimensional model of the vowel space (after Ladefoged (1975))

and rounded-unrounded. These dimensions are well understood by phoneticians trained in the tradition of Jones and of Ladefoged. Some of these extreme vowels are a part of the teaching and analysis repertoire of such phoneticians. Others such as the maximally rounded open vowels are not and are thus used experimentally in this study.

Five tokens of each extreme vowel were recorded by our speaker, who is an experienced phonetician trained in the British tradition. These tokens were the primary cardinals 1 [i], 4 [a]; 5 [ɑ] and 8 [u], and the corresponding secondary cardinals 9 [y], 12 [Æ], 13 [ɒ] and 16 [ɯ]. They were collected in low noise conditions and were prompted in random order by a computer program which also digitised the utterances and stored them in separate data files. The reference utterances were hand segmented. The parts of the signal where F_0 remained stable were used for this study. An $F_1/(F_2-F_1)$ plot was made of these vowels from conventional wideband spectrograms, and is shown in Figure 2. Note that the relationship between the corresponding unrounded and rounded vowels involves similar differences in F_2, except for the low back pair [ɑ] and [ɒ], which differ considerably in F_1.

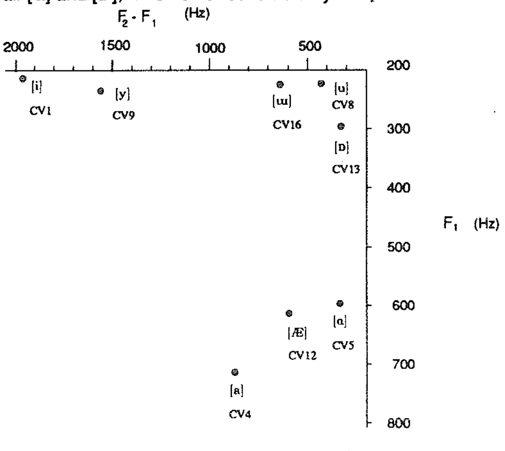


Figure 2: F_1 vs F_2-F_1 plot for phonetician's cardinal vowels CV 1 4 5 8 9 12 13 16.

ENGLISH VOWELS

A set of English vowels comprising five repetitions 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/), and [vwl] represents one of the eleven monophthongal phonemes (/i, ɪ, ɛ, æ, ɑ, ɒ, ɔ, ʊ, u, ʌ, ɜ/); and d is /d/. The [stop][vwl]d utterances were been manually segmented and labelled according to the procedures described by Ran (1994). 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 3). 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 /ɯ/ 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 wideband spectrograms also reflects this pattern (Figure 4).

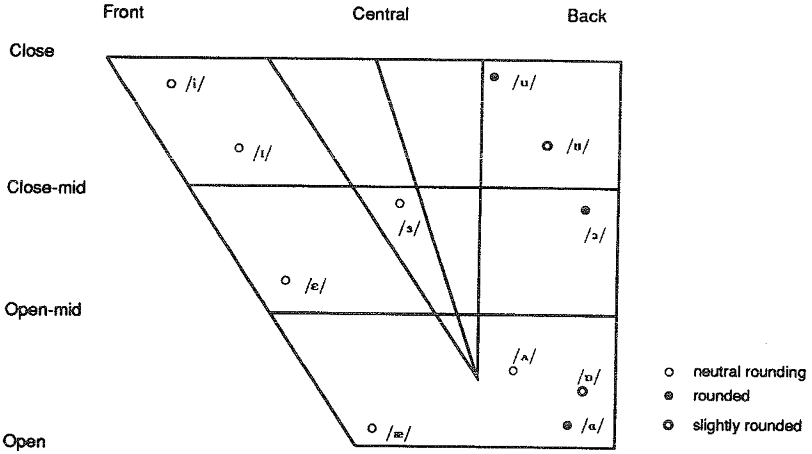


Figure 3: English vowel description by a phonetician.

Cardinal Vwl	Articulatory Description	Front	Open	Round
i1	front-close-unrounded	1	0	0
y9	front-close-rounded	1	0	1
ɯ16	back-close-unrounded	0	0	0
u8	back-close-rounded	0	0	1
ɑ5	back-open-unrounded	0	1	0
ɒ13	back-open-rounded	0	1	1
a4	front-open-unrounded	1	1	0
Æ12	front-open-rounded	1	1	1

Table 1: Articulatory labels for the reference vowels.

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, passing them through a Hamming window, and deriving 13 Linear Predictive Cepstral Coefficients (LPCCs) for each frame. The LPCCs were used as input to each of the multilayer perceptrons (MLPs) on

PERCEPTRON TRAINING

Each MLP served as an articulatory descriptor for a single dimension of the vowel space and was trained using eight reference vowels which were labelled according to Table 1. The descriptors were trained by comparing the output value to the label values, then adjusting their internal weights using the back-propagation algorithm. The inputs for this training comprised four repetitions of the eight reference

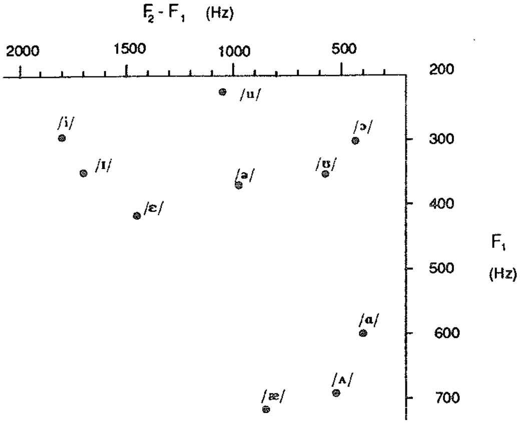


Figure 4: F_1 vs F_2-F_1 plot for phonetician's English vowels in b-g context.

vowels. Training was concluded when the comparison could not be improved further.

MLPs of one hidden layer were used, because they are theoretically able to solve problems of any complexity (Lippmann, 1987). 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 frontness descriptor was 4 and for the openness descriptor was 3. The total number of frames for the training was 5529.

AUTOMATED ANALYSIS OF ENGLISH VOWELS

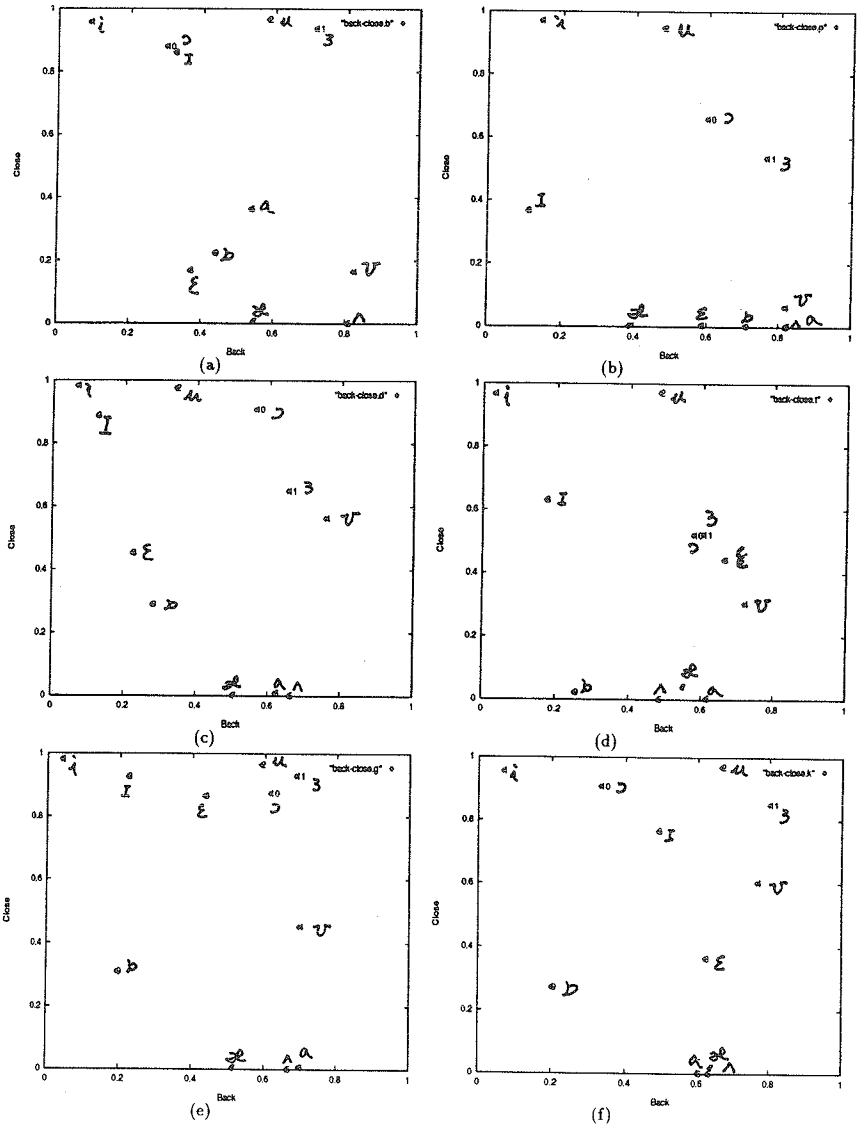
The data of the pseudo-static intervals of the English vowels were processed by the trained articulatory descriptors, namely the openness descriptor and the frontness descriptor, on a frame by frame basis. The outputs from the descriptors were activation scores of the output nodes of the MLPs, which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

Figure 5 reports the results by combining the output from the two descriptors. 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.

EFFECTIVENESS OF THE METHOD

Analysis of Figure 5 reveals that, compared with the phonetician's auditory judgements (Figure 3) and the F_1/F_2-F_1 plot (Figure 4), the automatic method using the eight reference vowels resolve the English vowels with three degrees of accuracy. The vowels /i, u, æ, ɑ/ and /ʌ/ are generally resolved very well: they show a reasonable degree of clustering across the six contexts - especially /i/ - and they are positioned correctly with respect to both backness and height. /u/ and /ʌ/ even show some expected backness assimilation: tokens flanked by alveolar consonants are located fronter. Three of these vowels - /i, æ, ɑ/ - are very similar auditorily to the reference vowels CV1, CV4 and CV5 respectively, which may account for their relatively good resolution. The vowels /ʊ, ɒ/ and /ɔ/ show less clustering (more sensitive to context), but do show some agreement in one of the two dimensions: /ɒ/ and /ɔ/ are correct with respect to height; /ʊ/ with respect to backness. Finally /ɪ, ɜ/ and /ɛ/ have the greatest spread (or context sensitivity) and do not resolve well in either dimension. Thus the resolutions appear to be rather sensitive to differences in the consonantal frame. It can only be assumed that differential consonantal assimilation is occurring but the particular pattern of this effect requires further study.

In spite of this differential resolution, if the mean position of a vowel is plotted from all six contexts (Figure 6) a much better approximation to the relative auditory positions of the vowels results. It can be seen that all except the lax round vowels /ʊ/ and /ɒ/ are well resolved with respect to height. Resolution in the backness dimension is not quite as good: /ɔ, ɒ, and ɑ/ are too front - /ɔ/ and /ɒ/ unacceptably so -



(a) (b) (c) (d) (e) (f)

Figure 5: Test result based on 8 reference vowels: 11 pseudo steady-state vowels on Close versus Back plane in the context of six stop consonants: (a) /b/; (b) /p/; (c) /d/; (d) /t/; (e) /g/; (f) /k/.

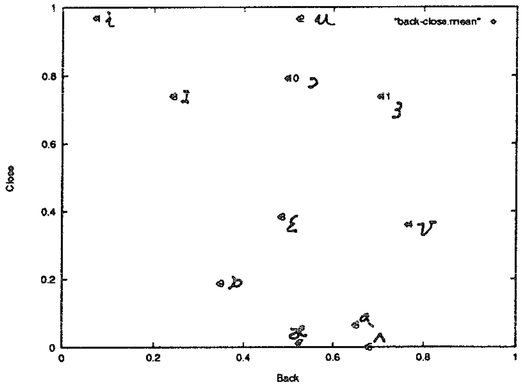


Figure 6: Test results averaged over six stop contexts: 11 pseudo steady state vowels on Close versus Back plane.

and /ɜ/ is too back. Nevertheless the configuration still accords fairly well with the auditory impression, especially as far as the front vowels and /u/ are concerned.

It is likely that at least some of the discrepancies in resolution are attributable to the variable relationship between the rounded and unrounded pairs of reference vowels (see Figure 2). Future research can examine whether the resolution is improved by chosing a different set of reference vowels, particularly without CV13 (/ɒ/).

CONCLUSIONS

This study arose from the need expressed in Ran et. al. (1994) for a rigorously selected and produced set of reference vowels for use in automated vowel quality description. The results have clearly shown the benefit of the set of cardinal vowels produced by an experienced phonetician. While some vowels have not responded entirely well to this approach, a substantial numbers of vowels have. For these vowels we have a normalised system of automatic phonetic quality description. The challenges that remain include further understanding of the reasons for failure with some vowels, methods for accounting for consonantal context, and ways of training naive speakers to produce reference vowels which may then be used to normalise automated phonetic description of their vowels.

REFERENCES

Jones, D. (1956), ***An Outline of English Phonetics***, (W. Heffer & Sons Ltd., Cambridge, MA).

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

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

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

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.

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