AUTOMATIC VOWEL QUALITY DESCRIPTION USING A VARIABLE MAPPING TO AN EIGHT CARDINAL VOWEL REFERENCE SET

*Shuping Ran, Bruce Millar and Phil Rose**

Computer Sciences Laboratory
Research School of Information Sciences and Engineering
* Linguistic Department, Faculties
Australian National University

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 eight cardinal vowels. The paper aims to improve the positioning of vowels in the open-close dimension by experimenting with a parameter in the model α which is the parameter which controls the slope of the sigmoid function employed in the multi-layer perceptrons.

1. 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 requires extensive auditory training, and is not feasible for non-phoneticians.

Is it possible to describe vowel quality without the skills of an experienced phonetician using a method which automatically places a given vowel into a space which is defined by a set of reference vowels and approximates to the phonetic space used by phoneticians?

The eight cardinal vowels (Fig.1) produced by an experienced phonetician trained in the British tradition represent the extremities of the dimensions “front-back”, “open-close”, and “rounded-unrounded”, and together form an external framework for the vowel space of that speaker. An automatic method [3] for placing the English vowels produced in stop consonantal context by the same speaker has already been reported [4]. Good performance for front vowels but much poorer performance for back vowels focused attention on the acoustic impact of rounding on these reference vowels.

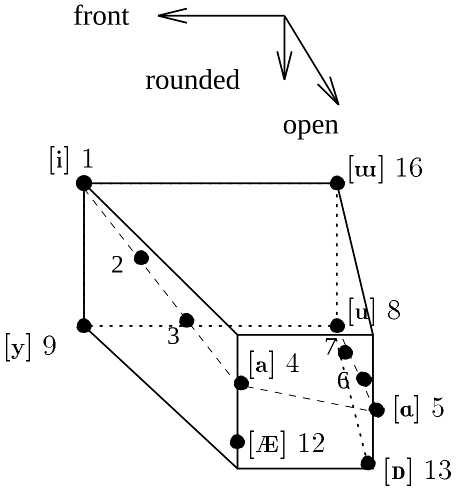

Figure 1: A three dimensional model of the vowel space (after Ladefoged [2])

The four primary cardinal vowels (vowel 1 [**i**], 4 [**a**], 5 [**ɑ**] and 8 [**u**]) were then selected from the eight cardinals to overcome the lip rounding effect introduced by some of the reference vowels. The placement of the same test vowels showed some improvement when using just these four reference vowels. The mean relative positioning of the vowels (across all consonantal contexts) approximated to the relative positioning of the vowels when measured acoustically and placed on an F_1 by F_2-F_1 plane [5].

Critical observations were made in two previous studies [4,5] that the resolution of the vowel positioning appears to be rather sensitive to differencies in the consonantal context, and that in individual contexts (especially in study [5]) some test vowels were often placed the extremities of the “closeness” dimension where only the cardinal vowels would be expected. It was hypothesised that this might be related to non-linearity in the output stage of the automatic process.

The present study was designed to examine this hypothesis.

2. METHOD

The vowel space which this study is attempting to model is described by two dimensions that are correlated with the articulatory dimensions “front-back” and “open-close”. An artificial neural network with a Multi-Layer Perceptron (MLP) architecture was used to model each of the dimensions. MLPs with one hidden layer were used because of their ability to encode relationships of any complexity [6].

All the spoken vowel data were analysed 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. The MLP training data comprised those parts of four repetitions of the cardinal vowels where F0 remained constant. The MLPs were trained using the back-propagation algorithm in which MLP outputs generated by frames of LPCCs were compared with the “back” and “close” articulatory labels as shown in Table 1.

After training the MLP models became detectors for the articulatory feature for which they were trained. In the testing process, analysed frames of the English vowels were presented to the input of each detector which generated the probability that its feature was present in the input data.

In this study the probabilities generated by the “closeness” detector are subjected to analysis to investigate to what extent non-linearity in the output stage of the “closeness” MLP is responsible for the performance observed in the earlier studies reported above.

cardinal vowel	articulatory description	back	close	round
i1	front-close-unround	0	1	0
y9	front-close-round	0	1	1
ɯ16	back-close-unround	1	1	0
u8	back-close-round	1	1	1
ɑ5	back-open-unround	1	0	0
ɒ13	back-open-round	1	0	1
a4	front-open-unround	0	0	0
Æ12	front-open-round	0	0	1

Table 1: Articulatory labels for the reference vowels.

Figure 2 is an example of a MLP with one hidden layer. Every node is fully connected to every node in the adjacent layers. The output $a_{i,j}$ of the node j of layer i is input to every node of the next layer $i+1$. The output $a_{i+1,j}$ is calculated as:

$$a_{i+1,j} = f(\sum_{j=1}^{N} w_{i,j} * a_{i,j})$$

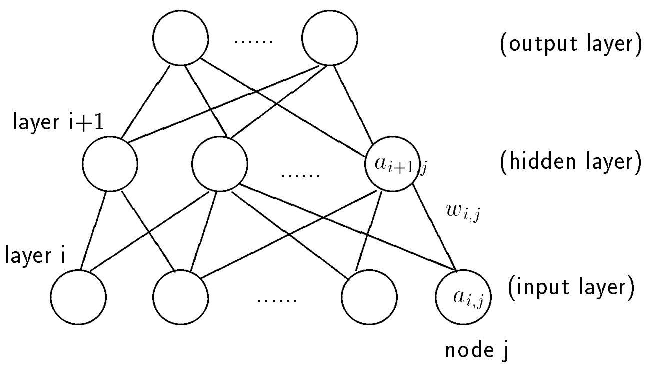

Figure 2: An example of MLP

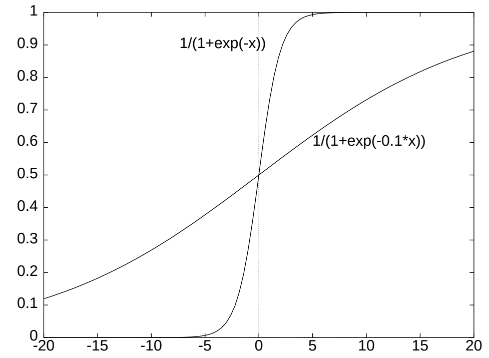

Figure 3: Sigmoid function with $\alpha = 1.0$ and $\alpha = 0.1$

where function $f(.)$ is called the activation function.

The sigmoid function:

$$1/(1+e^{-\alpha x})$$

is popularly used for classification problems. The non-linearity of the sigmoid function increases with α. $\alpha = 1.0$ was used for the previous studies producing a highly non-linear activation function. Figure 3 shows a figure of sigmoid function with $\alpha = 1.0$ and $\alpha = 0.1$.

Intuitively, one can understand that it is desirable to have a highly non-linear sigmoid function as activation function for classification problems as it will map most of the input to an output which is close to maximum or minimum indicating the class membership of that input. In our application we are looking for mapping that will provide graded interpolation between the cardinal vowel extremities. Thus a more linear output stage mapping using a small α would appear to be appropriate.

In the present study, we experimented with four different values of α, namely: $\alpha = 0.1$, $\alpha = 0.25$, $\alpha = 0.5$ and $\alpha = 1.0$ to test our hypothesis. For each value of α a new architecture of MLP had to be determined as more or less non-linearity was available in the output stage. As in previous studies the number of hidden units was increased until no further improvement in the modelling of the training data was observed. The resulting architecture was then trained 100 times using different initial conditions so that suboptimal training solutions could be eliminated. For each α the MLP giving the best classification of the training data was used to process the English vowels.

3. REFERENCE VOWELS

The reference vowels used in this study were derived from the vowel model expressed by Figure 1. The aim was to use cardinal vowels that were maximally extreme on the two dimensions of front-back and open-close. The eight cardinal vowels are 1 [**i**], 4 [**a**], 5 [**ɑ**], 8 [**u**], 9 [**y**], 12 [**Æ**], 13 [**ɒ**] and 16 [**ɯ**].

Five repetitions of each cardinal were recorded in a sound booth by our speaker. An $F_1/(F_2\text{-}F_1)$ plot was made of these vowels from conventional wide band spectrograms, as shown in Figure 4.

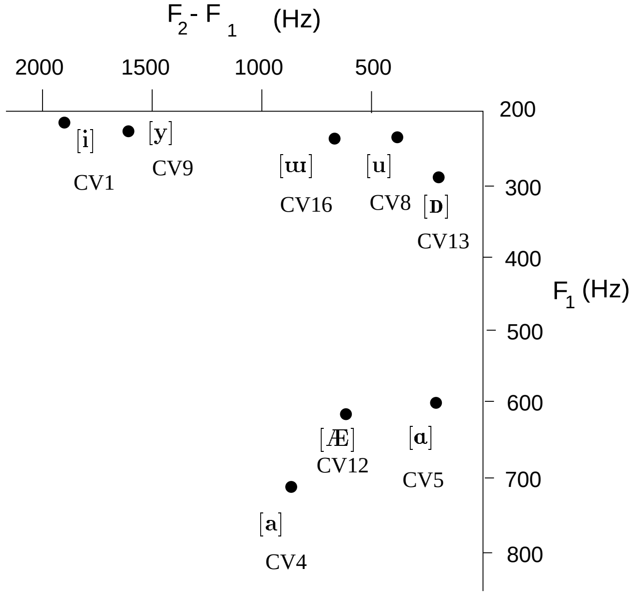

Figure 4: F_1 vs F_2-F_1 plot for phonetician's cardinal vowels CV 1 4 5 8 9 12 13 16.

4. 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 [7]. 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 5. 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 $F_1/(F_2\text{-}F_1)$ plot of the English vowels from conventional wide band spectrograms also reflects this pattern (Figure 6).

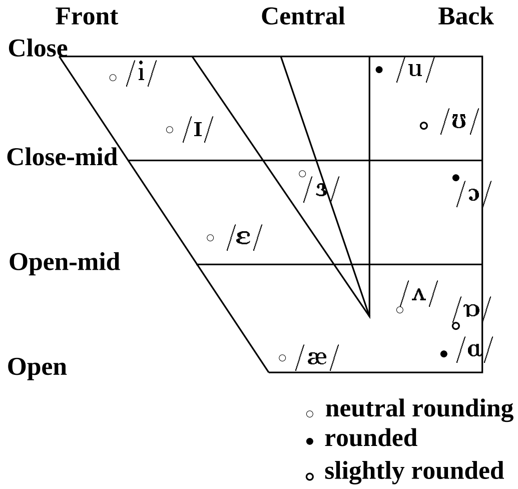

Figure 5: English vowel description by a phonetician.

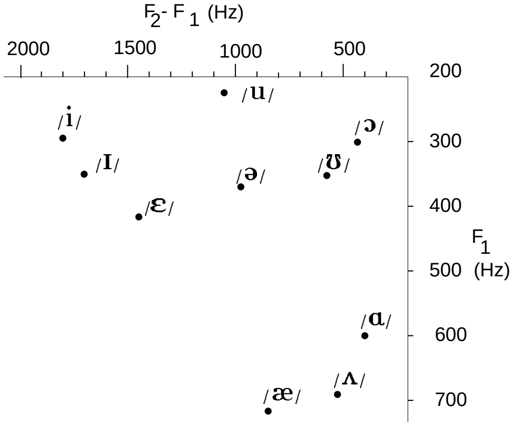

Figure 6: F_1 vs F_2-F_1 plot for phonetician's English vowels in **b-d** context.

5. RESULTS

The results of this study comprise the "closeness" detector's output levels for all the English vowels in each consonantal

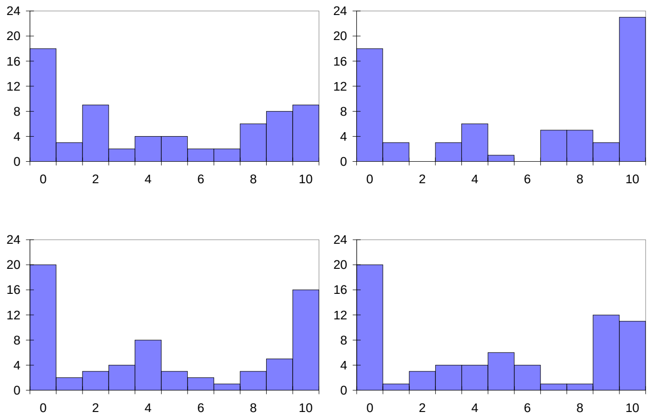

Figure 7: Histogram of output from MLP for all test samples for (a) top left: $\alpha = 0.1$; (b) top right: $\alpha = 0.25$; (c) bottom left: $\alpha = 0.5$; (d) bottom right: $\alpha = 1.0$

context that were processed. These output levels, each lying within the range from 0 to 1, were collected into 11 bins and plotted as a histogram in order to indicate graphically the proportion of output levels existing across their total range. If our hypothesis is true then for large α there should be a clustering of output values at the extremities of the range, but this clustering should be less obvious as α is reduced.

The histograms generated for the four values of α are presented in figure 7. One can observe from the Figure 7 that the number of cases where the input is placed to the extremities does not decrease by decreasing the α. These results disprove our hypothesis.

6. DISCUSSION

The failure of our hypothesis to be sustained by these experiments turns our attention to other factors which could lie behind the unexpected behaviour of our MLP feature detectors.

It should be noted that most contemporary wisdom on the training and testing of MLPs is based on the fact that the population of the training samples is in some way representative of the population of samples used to test the MLP. When this is the case the full input space to be encountered by the MLP is represented in its training input. In this situation, the MLP can interpolate between its training data points to classify its test data. The task that we are giving to the MLP is to interpolate between extremity data by using multiple examples of just two “open”/“close” vowel pairs whose spectral differences are themselves different from each other. While the MLP is known to be quite capable of encoding multiple pathways between input and output, the relative sparseness of the training data space compared to the testing data space may need some special care which we have currently ignored.

Two possible approaches to this problem are suggested. Firstly, we could introduce an additional “standard reference vowel” in the form of the “schwa” or neutral vowel. This has a clear articulatory description as do the primary cardinals and would represent a comprehensively intermediate spectral shape on which to train and which can be labelled 0.5 on both “backness” and “closeness” dimensions. Secondly, the distinctive spectral shape of the primary cardinals could be used to tailor the most appropriate cepstral range on which to base the training. The present range is selected from conventional experience with speech sound classification systems and while it may be appropriate for the test vowels, it may not be appropriate for the training set currently in use.

The results for automated vowel quality description that have already been achieved, based on averaged performance over six consonantal contexts, have indicated that vowels can be placed with reasonable accuracy in certain areas of the vowel space. We have not yet achieved our goal of determining the optimum conditions of acoustic representation, training procedures, and modelling methodologies that will ensure acceptably accurate placement throughout the vowel space. The degree of accuracy ultimately required also needs to be determined with respect to cross-linguistic differences in vowel acoustics (so-called linguistic phonetic differences). We propose to pursue this goal by exploring further refinements to our approach such as those indicated above.

7. 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., P. Rose, P., Millar, B. and Macleod, I. (1995), “‘Automatic vowel quality description using four primary cardinal vowels”, *Proc. XIIIth International Congress of Phonetic Sciences*, Vol. 3, pp. 318-321.

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.

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