INVESTIGATING THE DYNAMIC STRUCTURE OF VOWELS USING NEURAL NETWORKS

Steve Cassidy and Jonathan Harrington

Speech, Hearing and Language Research Centre
Macquarie University

Abstract

The target theory of vowel perception suggests that the vowels are identified from the static spectral characteristics at the vowel target. This has been challenged recently by Strange who claims that dynamic information may be more important than static spectral shape in identifying vowels. In the work described here we attempt to investigate this issue using a neural network trained to identify vowels from bark spectra inputs. If the network is able to better identify vowels which contain natural dynamic information than similar stimuli which do not, then this dynamic information must be characteristic of the vowel, rather than being noise. Our results confirm that dynamic information is useful in categorising eleven monophthongal vowels.

INTRODUCTION

The common view of the defining characteristics of vowels was for a long time the static spectral shape taken in the central, steady state portion of the vowel (see Strange (1987) for a review). Even when, in continuous speech, many vowels do not reach this steady state, many studies in the phonetics literature are based on the premise that the phonetic identity of the vowel is defined by a *target* (which the vowel might not attain). Recent challenges to this theory claim that additional information is used to identify vowels including the vowel duration, the pseudo-steady state central portion of the vowel and the formant transitions into and out of the pseudo-steady state portion (Strange, 1989b).

Strange (1989a) cites a number of experiments which show that subjects were capable of using this dynamic information to identify vowels. She constructed stimuli from CVC syllables with silent centres which retained the original syllable lengths, similar stimuli with the length cue removed and stimuli with just the onsets or offset transitions of the original syllable. The silent centre stimuli were perceived relatively accurately as long as the length cue was retained. Subjects performed poorly on the stimuli which contained only the onset or offset of the original syllable.

In this paper we describe a series of experiments using neural network speech recognisers to investigate the utility of the dynamic information in citation form vowels. We compared network performance on the original vowel which includes the dynamic information with that on artificial vowels with certain aspects of this information removed. The neural network is being used here not as a model of human perception but as an adaptive categorisation tool that can learn to use whatever information is appropriate to classify vowels. If the network can be shown to benefit from the presence of dynamic cues, then this will show that the dynamic information is not noise but is part of the character of the vowels.

TEMPORAL FLOW NEURAL NETWORKS

The application of neural networks to speech problems requires that the problem of the representation of time be solved. A number of approaches to this problem map the temporal dimension onto a spatial dimension; that is, the network is given all of the time-slices of the speech data at one time in a two dimensional array. In the Temporal Flow Model (Watrous, 1990), the temporal relationship between units is represented in the network by explicit propagation delays. A speech signal is applied one frame at a time to the input and the activation due to that input flows along the delayed links to the output layer.

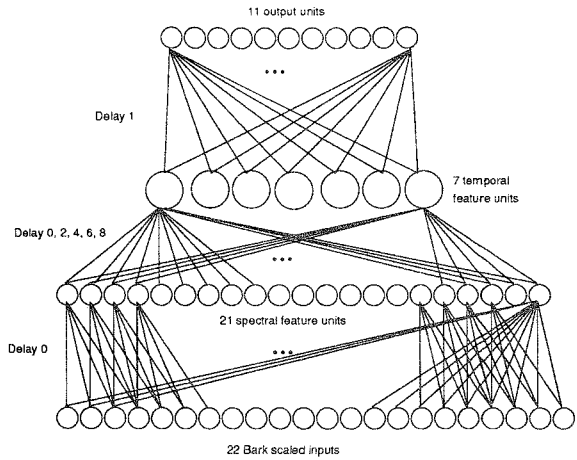

Figure 1: The structure of the networks used in the vowel recognition experiments.

Thus, data is viewed as flowing through the model along the time axis. The output of such a network is also a function of time. The network is trained to respond to an input pattern with, say, a sigmoid output function which rises from 0.5 to 0.95 part way into the vowel. In the networks described here, one output unit was allocated to each vowel and the output function for all 'off' units was the opposite of the 'on' unit.

Network Architecture

The networks used in these experiments consist of four layers of units connected by links of varying delays. The input layer consists of 22 units each of which receives its activation level from the first 22 critical bands of the current spectral frame. These units are connected by links of zero delay to a second layer of 19 spectral feature detectors. Each unit in the second layer is connected to four adjacent input units with pre-wired weights of -3, -8, 8 and -3; thus the second layer units are sensitive to the spectral gradient in the input and tend to enhance the formant structure of the input spectrum. Two units in the second layer are also connected to all 22 input units with links of zero delay.

The third layer consists of seven units connected to all 21 units in the second layer by links of delay 0, 2, 4, 6 and 8. These units are intended to extract second order spectral features based on both the static and temporal properties of the input. This layer is also connected back on itself via a recurrent link on each unit with delay one.

The output layer consists of one unit per vowel, eleven for the experiments reported here, connected to the third layer by links of delay one. The overall structure of the network is shown in figure 1.

Training

The training set consists of 325 vowels from four speakers extracted from citation form speech in the SHLRC speech database (Croot, Fletcher and Harrington, this volume) using the mu+ system (McVeigh and Harrington, this volume). Eleven different monophthongal vowels were used: /A/, /E/, /I/, /O/, /U/, /V/, /a:/, /e:/, /i:/, /o:/ and /u:/ with approximately 30 examples of each. The vowels were spoken in

monosyllables of the form /CVd/ where C varied across /b d g p t k h f s S ts/. The digitised waveform extracted from the database (sample rate 20KHz) and a 512 point Fourier transform with a 256 point frame shift was applied. A 22 band Bark scaled spectrum was constructed from each segment. Each vowel token for presentation to the network consists of a sequence of Bark scaled spectral frames, the number depending on the length of the original vowel.

Artificial 'steady state' stimuli were constructed by taking a single spectral frame from the mid-point of the vowel and duplicating it to the full length of the vowel. These stimuli retain the length of the original vowel but do not contain any of the dynamic spectral information.

The networks were trained using a training algorithm which minimises the mean squared error in the network by performing gradient descent in weight space. The algorithm (called BFGS for Broyden, Fletcher Goldfarb and Shanno (Watrous, 1988)) is superior to backpropagation as it fully minimises the error function along the direction of steepest descent in weight space on each iteration. Backpropogation on the other hand, is an iterative algorithm which takes a small step in the direction of steepest descent on each iteration. The BFGS results in faster training times than backpropogation but is neurologically less plausible as it relies on global information and a central process to monitor and update weight values.

Training the entire network on the vowel categorisation task proved impossible because the network would not converge given the training data. This may be because of the small number of tokens in each category or it may be a more general problem in this domain. The solution was to pre-train the first three layers of the network to extract a set of features from the input spectrum. Each vowel was classified according to the features: front/central/back, open/mid/close and long/short as shown in Figure 2. These features were mapped onto the seven units in the third layer of the network. This three layer network was trained on the feature extraction task using the whole vowel stimuli as input for around 400 epochs, by which time it had reached a mean squared error (mse --- the mean squared difference between the desired and achieved outputs) of 0.006. This network was then used as the basis for all of the four layer networks by adding a further layer after the seven feature units.

	front	central	back
close	/i:/ heed /I/ hid	/u:/ who	/U/ hood
mid	/e:/ there /E/ head		/o:/ saw /O/ hot
open	/A/ had		/a:/ hard /V/ mud

Figure 2: The vowels used to train the network categorised for place of articulation.

The full network was then trained to respond with one output node becoming active part way through the vowel (following a sigmoid output function) and all other output nodes showing the opposite pattern. The network was trained until performance on an open test (using a different set of vowel tokens from the same speakers) showed a clear peak in performance. This typically occured at an mse of around 0.005 after 2--3 days of Sparc2 processing time.

EXPERIMENTS

To evaluate the dynamic/target issue, networks were trained on different stimuli which contained more or less dynamic information. Networks were trained on the full vowel and on artificial steady-state vowels. Each network was then tested on a different set of tokens of the same type from the same four speakers.

If dynamic information is useful for vowel categorisation we would expect the network trained on full vowel stimuli to show better performance than the steady-state network. However, if spectral variation within the vowel is in fact noise, and cannot aid vowel categorisation, we would expect the network trained on 'cleaner' steady-state vowels to give better performance.

The networks were evaluated by presenting the test vowels and comparing the actual output with the desired output for that vowel. Each output unit will produce some activation pattern as the stimulus is fed through the network. These activation patterns are compared with the 'correct' pattern for a positive response, in this case a sigmoid curve changing from 0.5 to 0.95 part way through the vowel. The output unit whose output best fits this curve is deemed to be the network's response as long as the fit (measured as the mean squared error) is better than 1.5 times the nearest competitor. In this latter case, the network is said to have rejected the stimulus. Figure 3 shows the response of the eleven output units to an /A/ test stimulus. It can be seen that the response of the (first) /A/ unit corresponds closely to a positive sigmoid whereas all of the other units show negative curves.

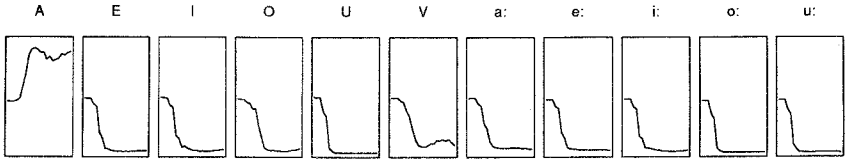

Figure 3: Network response to an /A/ test stimulus.

RESULTS AND DISCUSSION

Table 1 shows the results for the two networks trained on full vowels and steady-state vowels. As predicted by the dynamic theory, the performance of the network trained on the full vowel is considerably better than that of the steady-state network.

Training Set	Test Set	Correct (%)	Rejected (%)	Error (%)
Full Vowel	Full	90.0	5.0	5.0
Steady-State	S-S	73.2	7.5	19.3
Full Vowel	Centres	31.2	38.9	29.9
Steady-State	Half-Length	34.3	37.4	28.3

Table 1: Results of open tests for two networks trained on full vowels and artificial steady-state vowels (percent correct). The networks were tested on both full length and half length stimuli.

Results are also show for a test of the network on half length vowel stimuli. These stimuli were the central part of the vowel in the case of the full vowel network and a half length steady state vowel for

the steady state network. It can be seen that both networks perform poorly in these tests, indicating that the length of the stimulus is an important feature used in identifying the vowels.

The confusion matrices for the open tests of both networks are shown in figures 2 and 3. Much of the error in the full vowel case can be ascribed to confusions between /E/ and /e:/: the network responds /e:/ for /E/ five times, /E/ for /e:/ three times. Most of the rejected responses are also to /E/ (6) or /e:/ (5) targets. The errors made by the steady-state vowel network are more diverse. The confusions between /E/ and /e:/ again accounts for a large number of errors although this time the largest number of rejected responses are due to /V/ targets. There are many confusion between the long and short vowels, for instance /V/ and /a:/ /E/ and /e:/ /I/ and /i:/ etc. These pairs of vowels are minimally separated in the F1/F2 vowel space and so will tend to be more confuseable when only the steady-state spectral shape is available. This indicates, perhaps, that it is in these distinctions that dynamic spectral information is most useful. The steady-state vowel network also makes many more rejections than the full vowel network, this indicates that the information available in the steady-state stimuli has not been sufficient for the network to develop strong categorical boundaries. The consequence of this is that many stimuli excite weak responses from more than one output node, resulting in an ambiguous response which is rejected in this analysis.

	A	E	I	O	U	V	a:	e:	i:	o:	u:	Rejected
A	28	.	.	.	.	.	1	.	.	.	.	1
E	.	21	.	.	.	.	.	3	.	.	.	6
I	.	.	26	.	.	.	.	.	1	.	.	1
O	.	.	.	29	.	.	.	.	.	.	.	.
U	.	.	.	.	28	2	.	.	.	.	.	.
V	.	.	.	1	.	25	.	.	.	.	.	2
a:	.	.	.	.	.	.	29	.	.	.	.	.
e:	.	5	.	.	.	.	.	18	1	.	.	5
i:	.	.	.	.	.	.	.	.	27	.	1	1
o:	.	.	.	.	.	.	.	.	.	29	.	.
u:	.	.	.	.	1	.	.	.	.	.	29	.

Table 2: Confusion matrix for the full vowel network, open test. '.' indicates no responses in this cell.

	A	E	I	O	U	V	a:	e:	i:	o:	u:	Rejected
A	21	.	.	1	.	.	.	.	4	.	.	4
E	1	2.	.	.	.	.	.	6	.	.	.	3
I	.	1	2.	.	.	.	.	.	4	.	.	3
O	.	.	.	24	.	.	1	.	1	2	.	1
U	.	.	.	.	25	.	3	.	.	2	.	.
V	.	.	.	1	1	13	7	.	.	.	.	6
a:	.	.	.	.	.	.	27	.	.	.	.	2
e:	.	9	.	.	.	.	.	14	4	.	.	2
i:	.	.	3	.	5	.	.	.	21	.	.	.
o:	.	.	.	.	1	.	.	2	.	24	.	2
u:	.	.	.	.	.	.	3	.	.	.	26	1

Table 3: Confusion matrix for the steady-state vowel network, open test. '.' indicates no responses in this cell.

CONCLUSIONS

The target theory of vowel perception claims that all of the necessary information for vowel identification is present in a single spectral slice taken from the steady-state portion of the vowel. The dynamic theory on the other hand, claims that spectral variation over the length of the vowel gives an important clue to the vowel's identity. These experiments with neural networks have demonstrated that a single spectral

slice plus total duration is not enough information to accurately identify vowels. The changing spectral shape over the duration of the vowel constitutes valuable information for vowel identification. These experiments therefore lend some support to the dynamic theory of vowel perception.

References

Strange, W. (1987). Information for vowels in formant transitions. *Journal of Memory and Language*, *26*, 550--557.

Strange, W. (1989a). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of The Acoustical Society of America*, *85*(5), 2135--2153.

Strange, W. (1989b). Evolving theories of vowel perception. *Journal of The Acoustical Society of America*, *85*(5), 2081--2087.

Watrous, R. (1988). GRADSIM: a connectionist simulator using gradient optimisation techniques. Technical report, University of Pennsylvania. Included with GRADSIM software package.

Watrous, R. (1990). Phoneme discrimination using connectionist networks. *Journal of the Acoustical Society of America*, *87*(4).