Language and Speech
http://las.sagepub.com/

Dynamic and Target Theories of Vowel Classification: Evidence from Monophthongs and Diphthongs in Australian English

Jonathan Harrington and Stephen Cassidy

Language and Speech 1994 37: 357
DOI: 10.1177/002383099403700402

The online version of this article can be found at:
http://las.sagepub.com/content/37/4/357

Published by:
SAGE
http://www.sagepublications.com

Additional services and information for *Language and Speech* can be found at:

Email Alerts: http://las.sagepub.com/cgi/alerts

Subscriptions: http://las.sagepub.com/subscriptions

Reprints: http://www.sagepub.com/journalsReprints.nav

Permissions: http://www.sagepub.com/journalsPermissions.nav

Citations: http://las.sagepub.com/content/37/4/357.refs.html

>> Version of Record - Oct 1, 1994

What is This?

DYNAMIC AND TARGET THEORIES OF VOWEL CLASSIFICATION: EVIDENCE FROM MONOPHTHONGS AND DIPHTHONGS IN AUSTRALIAN ENGLISH*

JONATHAN HARRINGTON
and
STEPHEN CASSIDY
Macquarie University, Sydney, Australia

Recent studies on the perception of speech have suggested that vowel identification depends on dynamic cues, rather than a single 'static' spectral slice at the vowel midpoint. The experiments reported in this paper seek both to test the extent to which vowel recognition depends on dynamic information, and to identify the nature of the dynamic cues on which such recognition might depend. Gaussian classification techniques, as well as different kinds of neural network architectures, were used to classify some 3000 vowels in /CVd/ citation-form Australian English words, following training on roughly the same number of vowel tokens produced by different talkers. The first set of experiments shows that when vowels are classified from three spectral slices taken at the vowel margins and midpoint, only diphthongs, but not monophthongs, benefit from the additional spectral information at the vowel margins. A further experiment shows that vowels are no better classified from a time-delay neural network than from the three-slice network in which time is not explicitly represented. At least for the citation-form, Australian English vowels in this study, these results are interpreted as being more consistent with a target, rather than a dynamic, theory of vowel perception.

Key words: Vowel classification, diphthongs, Australian English, neural networks

INTRODUCTION

In recent years, various experiments, primarily in the speech perception literature, have suggested that the phonetic identity of vowels is cued by information which is distributed over time in the acoustic speech signal. In the first of a series of experiments by Strange and her colleagues, Strange, Verbrugge, Shankweiler, and Edman (1976) found that listeners identified vowels more accurately in CVC than in isolated vowel syllables. Strange's (1987, 1989a) interpretation of this result is that CVC syllables contain transitions, whereas isolated vowels do not, and so the transitions must be providing listeners with additional information which cues the vowel's phonetic identity. The importance of transitions in cueing the vowel was further demonstrated in a

* Our thanks to Bruno Repp and the two reviewers of this paper, Stephen Zahorian and Terrance Nearey, for many helpful comments. This project was supported by Australian Research Council grant number AC9031073.

Correspondence should be addressed to Jonathan Harrington, Speech, Hearing and Language Research Centre, Macquarie University, Sydney, NSW 2109, Australia. E-mail: jmh@srsuna.shlrc.mq.edu.au.

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

subsequent study by Strange, Jenkins, and Johnson (1983) who presented listeners with various kinds of edited syllables. Two of these were silent center (SC) syllables in which the central section of the vowel, including the vowel target, was discarded, leaving only the initial and final transitions, and variable center (V) syllables in which the transitions were discarded, leaving the central section of the vowel. The results of this experiment showed that listeners' identifications of SC syllables were as accurate as those of unmodified CVC syllables, but that V syllables were identified more poorly. Despite some methodological difficulties with these kinds of experiments (Assmann, Nearey, and Hogan, 1982; Diehl, McCusker, and Chapman, 1981; Macchi, 1980), many subsequent studies have continued to show that vowels are as well identified from modified SC stimuli as from unmodified CVC syllables (e.g., Benguerel and McFadden, 1989; Fox, 1989; Parker and Diehl, 1984; Rakerd and Verbrugge, 1987; Strange, 1989b; Verbrugge and Rakerd, 1986), although the benefit that context provides to vowel identification has also been shown to be affected by the phonetic identity of both the consonant and vowel (Gottfried and Strange, 1980; Rakerd, Verbrugge and Shankweiler, 1984).

As Fox (1989) and more recently Andruski and Nearey (1992) have noted, there are various possible interpretations of the results of these experiments. Strange's interpretation is closely related to theories of the control of speech production which view phonetic segments as articulatory gestures that unfold in time. As Strange (1989b) comments:

> "According to this view... vowels are conceived of as characteristic *gestures* having intrinsic timing parameters (Fowler, 1980). These dynamic articulatory events give rise to an acoustic pattern in which the changing spectrotemporal configuration provides sufficient information for the unambiguous identification of the intended vowels."

A second possible interpretation is that even 'monophthongal' vowels are characterized by a changing acoustic pattern which is not attributable to context effects, but which forms part of the inherent structure of the vowel, in much the same way that the transition between two targets is an inherent aspect of diphthongs. This interpretation of 'dynamic' is consistent with studies by Nearey and colleagues (Nearey and Assmann, 1986; Andruski and Nearey, 1992), who show that at least some monophthongal vowels are characterized by inherent formant movement that may also be important for their perceptual identification. Compatibly, both the acoustic classifications and perceptual identification scores in Huang (1992) show an improved performance when vowels are identified from three spectral slices (25%, midpoint, 75%) compared with a single slice; and in a study of static and dynamic representations of 11 vowels by Zahorian and Jagharghi (1993), higher classification scores were obtained from the dynamic representations in both types of parameterization (formant and discrete cosine) of the acoustic signal.

Whatever the exact interpretation of 'dynamic', there is considerable evidence from these studies in the last 10 – 15 years that the vowel's phonetic identity is not just cued by formant values at the vowel midpoint, as suggested by the more 'traditional' method of vowel classification, dubbed the 'target theory of speech perception' (Jenkins, 1987). Instead, under the dynamic theory of vowel perception, speech production is inherently dynamic and, compatibly, listeners extract dynamic information distributed throughout the segment in decoding the speech signal.

The aim of the present set of experiments was to use various acoustic classification techniques to test the extent to which vowels are dynamic, and to try to characterize more precisely the nature of the dynamic information. As an initial hypothesis, we reasoned that

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

TABLE 1

Australian English vowels used in the study. Note: The transcription system, which is based on Mitchell and Delbridge (1965) indicates /ɛə/ to be diphthongal. However, /ɛə/ is monophthongal for most Australian English speakers, and certainly for the speakers used in this study.

Monophthongs (long)	Example	Monophthongs (short)	Example	Diphthongs	Example
i	heed	ɪ	hid		
u	who'd	ʊ	hood		
ɛə	hair	ɛ	head	eɪ	say
ɜ	heard			oʊ	so
ɔ	saw	ɒ	pod	ɔɪ	boy
a	hard	æ	had	aɪ	high
		ʌ	mud	aʊ	how

if the vowel's phonetic identity is cued by information that is temporally distributed in the acoustic signal, then classifications from combined spectral slices in the transitions and close to the vowel target should be better than a comparable classification from a single spectral slice at the vowel midpoint. On the other hand, under target theories of speech perception, since monophthongal vowels are sufficiently specified by information at the vowel midpoint, only diphthongs, but not monophthongs, should benefit from the spectral information close to the vowel margins. The predictions of the two theories were tested in the first experiment in which Australian English monophthongs and diphthongs were classified from single and multiple spectral slices.

EXPERIMENT 1

Method

Materials. Ten adult speakers of Australian English (five men, five women) read a list of CVd words in a carrier phrase, where C was /p t k b d g f θ s ʃ tʃ/ combined with 17 vowels (Table 1). The speakers' accents can be described as intermediate between Cultivated and General Australian (Bernard, 1970; Mitchell and Delbridge, 1965). Five of the vowels are 'traditionally' considered to be diphthongal in Australian English, i.e., to consist of two targets with a transition between them. Most of the other vowels in the speakers' accents are monophthongal; there is a tendency for the Australian English close vowels /i u/ (*beet, boot*) to be produced with a long onglide starting from a position close to a schwa vowel, but this is predominantly characteristic of a Broad Australian accent, rather than of the speakers' accents in this study. Some further details on the phonetic characteristics of Australian vowels are given in the appendix.

The word lists were read twice on separate days thereby giving a total of 10 talkers × 12 contexts × 17 vowels × 2 readings = 4080 tokens. These materials, which form part

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

of the Australian National Database of Spoken Language (Croot, Fletcher and Harrington, 1992) were recorded in a sound-treated room using high quality recording equipment at Macquarie University and digitized at 20 kHz. Phonetic segmentation and labeling of all the materials were carried out using Waves by trained phoneticians at Macquarie. The boundaries of the vowels were marked at the first and last glottal pulses respectively, as judged from a combination of displayed speech waveforms and spectrograms. After removing a number of tokens which had been misread, 1889 tokens remained from the first reading and 1892 from the second.

Acoustic parameters. For each vowel, summed energy values in critical (Bark) bands 4 – 19 (roughly 300 – 5250 Hz) were calculated from 512 point FFTs (25.6 msec window) centered at 20%, 50%, and 80% time points relative to the vowel boundaries. The frequency ranges for the Bark bands were taken from Zwicker (1961). Hamming windows were used. No pre-emphasis was applied. The energy above the upper limit of 5250 Hz was ignored, since it was considered unlikely to benefit vowel separation. The motivation for the lower cut off frequency of Bark band 4 was to avoid, as far as possible, the confounding effects due to the male – female differences in fundamental frequency.

It is of course difficult to separate transitions from vowel targets on a principled basis (see, e.g., the remarks in Benguerel and McFadden, 1989, and Nearey and Assmann, 1986), and so the assumption was made that 20% and 80% would be sufficiently close to the vowel margins to be influenced by the surrounding context and to form part of the consonant-vowel and vowel-consonant transitions. The 50% time point was assumed to be closest to the vowel target for most monophthongal vowels.

Total segment duration was also determined for each vowel since it is an important parameter in distinguishing long from short vowels of similar quality (e.g., /i/ *vs.* /ɪ/).

Since the data included male and female speakers, a basic form of speaker normalization was carried out by subtracting speaker centroids from the data. A speaker centroid is a vector of mean values, one per Bark band, calculated across all the vowel tokens for each speaker separately. Subtracting speaker centroids in this way has been shown to be an effective and simple way of reducing speaker specific effects (Chan and Cheung, 1986; Klein, Plomp, and Pols, 1970). In addition, to ensure that the variance of the duration dimension was comparable to the variance of the spectral dimensions, duration was expressed in tenths of a second. This normalization is necessary to ensure that duration does not dominate the principal components analysis used in the Gaussian classification experiment. Duration was not included in the per-speaker normalization.

Training, testing, and evaluation All of the results reported in this paper are based on dividing the data into training, evaluation, and testing sets. The training set was used to build a model, or set of models, to identify the 17 vowels using one of two classification techniques (Gaussian classification and neural networks). The evaluation set was used to choose one of the possible models generated in the training stage: the one which identifies most vowels correctly. The test set was used for the final evaluation of this chosen model. In this way, the data used to train the model and those used to test it were kept separate: The model was not tuned in any way for its performance on the test set. The test results should therefore represent a true open test of the model's performance.

In order to provide as much data as possible to evaluate the model, each experiment was performed twice on different sets of training and test data. The 10 talkers of this study were allocated to three groups (groups A, B, and C). Groups A and C consisted of four

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

talkers each (two males and two females) and were used as both training and testing sets. Group B consisted of the remaining two talkers (one male, one female) and was used as the evaluation set. Each experiment was run twice: firstly, by training on Group A, evaluating on Group B, and testing on Group C; and secondly, by training on Group C, evaluating on Group B; and testing on Group A. The results from both runs were combined to give results for eight talkers in each experiment.

Therefore, in all of the experiments reported in this paper, training and testing were carried out on different tokens produced by different sets of talkers. The purpose of including two runs with different talkers in the training and testing sets is to ensure, as far as possible, that the results are not biased by speaker-specific effects.

Classification algorithms. Two types of classification algorithms were used. The first is based on principal components analysis and Gaussian classification. The second uses a neural network architecture. Each of these two classifications was applied to two different spaces, defined as *midpoint*, and *concatenated.* The *midpoint* space is based on critical band values at the 50% time point plus total segment duration (17 parameters); the *concatenated* space includes critical band values at the three time points under consideration plus total segment duration (49 parameters).

In Gaussian classification, the centroid and covariance matrix of the training set are estimated for each vowel class. A token from the testing set is then classified based on the Mahalanobis distance to each of the class centroids. The Mahalanobis distance is defined as the Euclidean distance weighted by the covariance matrix (see O'Shaughnessy, 1987, for a more detailed mathematical discussion).

The space in which Mahalanobis distance calculations are made is the critical (Bark) band space transformed by principal components analysis (PCA). The transformed spaces are obtained by matrix multiplying (weighting) the critical band values of a token from the testing set with the eigenvectors calculated from the training set. The motivation for the PCA on the critical band space is that at least as good, and sometimes better, classification scores can be obtained from a smaller number of transformed dimensions than from the original, untransformed dimensions.

The training, evaluation, and testing stages of the Gaussian classifier were carried out as follows. Models were constructed using the first two, first three..., first *n* transformed dimensions, where *n* is the number of original dimensions (17 and 49 in the *midpoint* and *concatenated* conditions respectively). Each model was then tested on the evaluation data; the model which gave the best classification score on these data was chosen and tested on the test data.

The technique of Gaussian classification from transformed dimensions is limited in various ways: Specifically, each class is modeled by a *single* centroid and covariance matrix, and the transformed space is obtained from a *linear* combination of the original dimensions. In a neural network with one hidden layer, the transformation from input to hidden nodes (via the weight matrix) performs an operation similar to PCA. However, it performs a nonlinear computation and is not limited to extracting orthogonal dimensions. The hidden nodes can be said to extract *features* from the input spectrum which are then combined by the hidden-to-output weight matrix to give the output classification. Since the most active output node is chosen as the response, the output layer can be said to be performing a distance measurement based on the hidden layer feature set and choosing the closest category as the winner. The neural network may, if it is appropriate, model a

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

single category as a disjoint region in its feature space.

The networks used in this experiment consisted of a set of input units (17 for the *midpoint* condition, 49 for the *concatenated* condition) fully connected to a set of hidden units which were again fully connected to one output unit per vowel class. During training, the correct output unit for each input vector was turned on while all other units were turned off. The networks were tested periodically on an independent set of items (the evaluation set) and when performance on these items began to degrade, training was halted. This technique was used to avoid over-learning of the training set which results in poor generalization to other examples.

The neural network simulations for these experiments were performed with the GRADSIM package (Watrous, 1988, 1990). GRADSIM provides a number of training algorithms including the well-known back-propagation algorithm. We chose to use one of the quasi-Newton optimization algorithms, known as BFGS, which performs the same function as back-propagation, but which is much more efficient, because it takes a global view of the error surface which is being minimized. BFGS, as with all neural network learning algorithms, adjusts the weights in the network to minimize the error between the true output and that produced by the network for a set of training examples.

An important variable in neural network experiments is the number of hidden units in the network. For these experiments we initially built three networks with 7, 14 and 20 hidden units in each. There was an increase in performance across all conditions as the number of hidden units was increased. Since the change between 14 and 20 hidden units was small, and since the absolute performance of the networks was very good, we decided not to increase the number of hidden units further. We will report only the results from the 20 hidden unit network as those with fewer hidden units follow the same general pattern.

Results

As mentioned earlier, all results in this paper are based on pooling the classification scores from the two separate runs (run 1: training on talker set A, testing on talker set C; run 2: training on talker set C, testing on talker set A).

The results of these experiments were compared using a *t*-test on the data broken down on a per speaker basis.

Gaussian classification. The comparisons between the midpoint and concatenated spaces were made on the peak scoring dimensions, as determined from the evaluation set. For the first run, peak scores were obtained on 10 (midpoint) and 13 (concatenated) dimensions; for the second run, peak scores were obtained on 11 (midpoint) and 12 (concatenated) dimensions.

Based on the total number of vowels correctly classified, classifications were more accurate in the concatenated than in the midpoint condition. In the concatenated space, 88.6% of vowels were correctly classified, which is significantly greater [$t(7) = -7.3$, $p < 0.001$] than the corresponding score from the midpoint space (73.2%).

A two-way condition (midpoint *vs.* concatenated) by vowel type (monophthong *vs.* diphthong) repeated measures ANOVA was performed on the vowel scores. The results show significant effects of both condition [$F(1,7) = 9.2$, $p < 0.005$] and type [$F(1,7) = 5.0$, $p < 0.05$] and an interaction between condition and type which just fails to attain significance at the 0.05 level [$F(1,7) = 4.0$, $p < 0.1$]. The interaction between condition and type suggests that the improvement between the midpoint and

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

TABLE 2

Scores for Gaussian and neural network classification methods. Stars (**) mark vowels where the two sets of scores differ significantly ($p < 0.05$)

	GAUSSIAN			NETWORK		
	Single	Concat		Single	Concat	
æ	77.6	92.7	**	75.0	95.8	**
ɛ	96.6	96.6		92.0	98.9	
ɪ	96.4	93.8		92.7	81.2	
ɒ	87.5	92.0		79.5	88.1	**
ʊ	91.5	96.6		92.0	97.7	
ʌ	87.4	75.9		74.7	87.9	**
ɜ	59.7	85.6	**	54.1	75.1	**
a	52.0	92.1	**	71.8	85.3	
ɛə	39.1	66.7	**	55.2	92.5	**
i	87.5	88.1		81.2	97.2	**
ɔ	92.6	92.0		80.1	86.9	
u	88.2	89.3		80.9	79.8	
oʊ	42.3	84.6	**	54.9	82.3	**
aɪ	55.1	84.7	**	70.5	79.5	
aʊ	42.9	81.7	**	63.4	88.0	**
eɪ	66.5	95.5	**	63.6	94.3	**
ɔɪ	79.5	97.7	**	38.1	96.6	**
All	73.2	88.6	**	71.9	88.6	**

concatenated conditions is mainly due to the improved performance of diphthongs in the latter condition.

Table 2 shows a vowel-by-vowel analysis of the scores from the two spaces. All diphthongs performed significantly better in the concatenated than the midpoint condition, but only four out of 12 monophthongs (/æ ɜ a ɛə/) had significantly better scores on the concatenated condition.

Neural Network. 88.6% of vowels were correctly classified on the concatenated network compared with 71.8% correct on the midpoint network: These results are significantly different [t (7) = −7.4, $p < 0.002$]. A vowel-by-vowel analysis (Table 2) shows that all diphthongs except /aɪ/, and six out of 12 monophthongs, have significantly higher scores on the concatenated network.

A similar two-way ANOVA was carried out on these results showing again main effects of condition [F (1,7) = 20.9, $p < 0.0001$] and type [F (1,7) = 7.9, $p < 0.01$] and a significant interaction between condition and type [F (1,7) = 5.0, $p < 0.05$]. Again this interaction suggests that the improvement between the midpoint and concatenated conditions is mainly due to the improved performance of diphthongs in the concatenated condition.

Discussion

On the measure of total number of correctly classified vowels, Experiment 1 has shown that scores are higher from the concatenated than the midpoint space. However, an

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

examination of the results in terms of the separate vowel types shows that most diphthongs, but only a small number of monophthongs, benefit from the inclusion of the spectral slices close to the vowel margins.

These results lend only limited support to the dynamic theory and are in fact consistent with predictions of the target theory: Diphthongs perform better on multiple spectral slices because they have multiple targets distributed in time, but classifications of monophthongs, which have a single target close to the vowel midpoint, are generally no better from three spectral slices than from one.

Nevertheless, an explanation is needed for the small number of monophthongs which performed better in the concatenated space. Under one interpretation, their improved performance might be attributed to their inherent dynamic structure which aids phonetic identification. Another interpretation might attribute the better performance of these monophthongs entirely to the improvement in the performance of diphthongs: That is, since diphthongs are more clearly delineated in the concatenated space, monophthongs are less likely to be confused with them.

If these monophthongs are inherently dynamic (first interpretation), then better scores should be obtained in a concatenated space in which training and testing are carried out *on monophthongs only*. If, on the other hand, the presence of diphthongs accounts for the improvement in the classification of these monophthongs (second interpretation), then concatenated and midpoint scores should be the same when training and testing are carried out on monophthongs only. The next experiment sought to adjudicate between these hypotheses.

EXPERIMENT 2

Method

The methodology was exactly parallel to that of Experiment 1. The only difference was that training and testing were carried out on the 12 monophthongs only. For this experiment, the numbers of tokens in the three talker groups were: Group A, 1076; Group B, 535; Group C, 1072.

Results

As before, the results are based on pooling the classification scores from the two separate runs (run 1: training on talker set A, testing on talker set C; run 2: training on talker set C, testing on talker set A).

Gaussian classification. As in the first experiment, the comparisons between the midpoint and concatenated spaces were made on the peak scoring dimensions, as determined from the evaluation set. For the first run, peak scores were obtained on 8 (midpoint) and 15 (concatenated) dimensions; for the second run, peak scores were obtained on 7 (midpoint) and 19 (concatenated) dimensions.

Based on the total number of vowels correctly classified, classifications were no more accurate in the concatenated than in the midpoint space. In the concatenated space, 90.2% of vowels are correctly classified, and the corresponding score in the midpoint space was 92.4%: These scores are not significantly different [$t\ (7) = 1.82$, $p > 0.05$]. A subsequent vowel-by-vowel analysis (Table 3) showed that there were no significant differences in the performance of the individual monophthongs in the two conditions,

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

TABLE 3

Scores for Gaussian (with and without noise) and neural network classification methods, monophthongs only. Stars (**) mark vowels where the two sets of scores differ significantly ($p < 0.05$)

	GAUSSIAN			GAUSSIAN WITH NOISE		NETWORK		
	Single	Concat		Single	Concat	Single	Concat	
æ	91.1	93.8		85.4	83.3	89.1	91.7	
ɛ	96.6	96.6		77.8	77.3	96.6	98.3	
ɪ	97.9	94.3		87.5	91.7	94.8	91.7	
ɒ	88.1	91.5		71.0	80.1	79.0	83.0	
ʊ	98.3	94.3		96.6	93.2	93.2	94.9	
ʌ	89.1	77.6		83.3	76.4	78.7	94.8	**
ɜ	85.6	91.2		60.2	67.4	83.4	80.1	
a	97.7	98.3		87.6	79.1	94.9	98.9	
ɛə	85.1	63.8	**	54.6	44.8	76.4	91.4	
i	93.2	90.3		72.2	78.4	93.2	98.3	
ɔ	99.4	93.8		88.1	80.1	93.2	94.9	
u	86.5	95.5		85.4	78.7	90.4	88.8	
All	92.4	90.2		79.2	77.7	88.6	92.2	

with the exception of /ɛə/ which performed significantly better in the midpoint than the concatenated condition.

The total correct scores reported above for the midpoint and concatenated conditions are close to 100%, and in order to ensure that the differences between the conditions are not obscured by ceiling effects, the same classifications were carried out a second time, but with added noise. Specifically, each of the dimensions of the training and testing data were multiplied by a random value between 0 and 1 prior to classification. The same number of dimensions were used in the classifications-with-noise as in the no-noise classifications. Total correct scores were 77.7% correct for the concatenated condition, and 79.2% correct for the midpoint condition (i.e., a reduction of 12.5% and 13.2% compared with the no-noise midpoint and concatenated conditions respectively). These total correct classification scores are once again not significantly different [t (7) =1.07, $p > 0.1$]. A subsequent vowel-by-vowel analysis showed no significant differences between the midpoint and concatenated noise conditions. Table 3 compares the no-noise and noise conditions.

Neural network. Using the neural network, 88.6% of vowels were correctly classified from the midpoint space compared with 92.2% from the concatenated space, a nonsignificant difference [t (7) = −1.53, $p > 0.1$]. A vowel-by-vowel analysis (Table 3) showed no significant difference between the two cases except for /ʌ/ which scored significantly better on the concatenated space [t (7) = −3.04, $p < 0.05$].

Discussion

The results from Experiment 2 favor the interpretation that diphthongs are responsible for the improved scores of some monophthongs when classifications are made from

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

multiple spectral slices. With the exception of / ʌ /, none of the vowels performed better in the concatenated than in the midpoint conditions when training and testing were carried out on monophthongs only. With regard to / ʌ /, the evidence is equivocal for two reasons: Firstly, because the Gaussian and neural network classifications produced different results; secondly, because an examination of the two separate training and testing runs of the neural network experiment showed that significantly better scores were obtained in the concatenated condition for / ʌ / only on one run (training on group C, testing on group A), but not on the other (training on group A, testing on group C).

Since scores from monophthongs are in general no better from three, compared with a single, spectral slice, the evidence suggests that spectral change does not benefit monophthong identification. An alternative interpretation might be that the spectral change for monophthongs between the three spectral slices is negligible: if classifications are being made from three nearly identical spectral slices, it follows that three-slice classification will be no better than one. In order to test this second hypothesis, two Euclidean distance calculations were made in the 16-dimensional, normalised Bark space, for each vowel token: From the spectral slice at the 20% time-point to the midpoint spectral slice; and from the midpoint spectral slice to the spectral slice at the 80% time point. These two Euclidean distance calculations were summed to give a spectral change measure for each of the 3781 vowel tokens. The results of this exercise showed that, while the spectral change is certainly less for monophthongs than diphthongs, it is far from negligible: Indeed, for some monophthongs the extent of spectral change is similar to that of diphthongs (Figure 1). In order to qualify the interpretation that spectral change in monophthongs does not benefit their identification, a comparison was made between the average magnitude of spectral change for each of the 12 monophthong classes and the differences between their scores on the three-slice and the single-slice classifications (in the monophthong-only classifications of Experiment 2): If spectral change *per se* is beneficial to monophthongal identification, then those monophthongs which showed the greatest improvement on the three-slice compared with the single-slice classifications should also have the greatest magnitude of spectral change. The correlations between magnitude of spectral change and score differences (three-slice *vs.* single slice) are very close to zero for both the Gaussian ($r = 0.05$) and neural network ($r = 0.10$) classifications. Since a greater magnitude of spectral change does not imply superior classification scores on three, compared with a single, spectral slice, the evidence further suggests that spectral change does not benefit monophthong identification.

Taken together, Experiments 1 and 2 have shown that diphthongs behave separately as a class from monophthongs, and also that there is very limited evidence to suggest that monophthongs are inherently dynamic.

However, it would be premature to reject the considerable perceptual and acoustic evidence in favor of vowels as dynamic for the reason that the concatenated condition does not adequately represent vowels as a *succession* of spectral slices in time. Specifically, by treating the spectral data from the three time points as separate dimensions, the concatenated condition does not encode the fact that the spectral slices follow each other in a particular temporal order.

In an attempt to address this issue, a further experiment was carried out using a neural network (henceforth the *recurrent* network) with an architecture that is sensitive to the temporal order of the spectral slices in the speech signal.

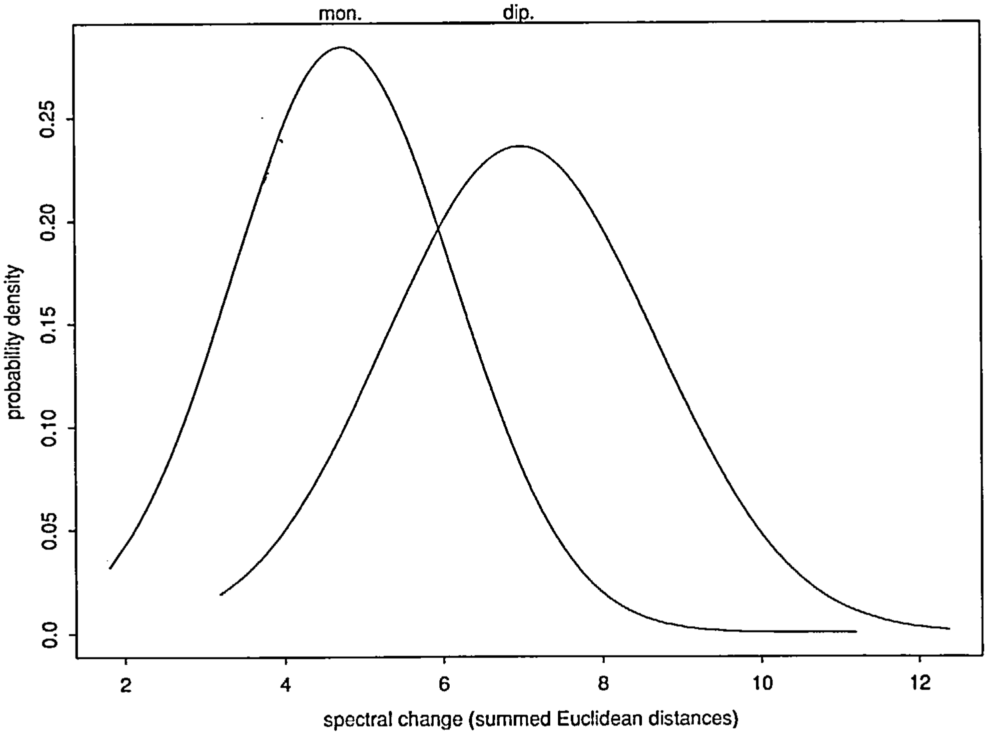

Fig. 1. The distribution of an index of spectral change (summed Euclidean distance between three successive time slices) for the monophthongs and diphthongs used in the study.

The hypothesis that was tested was as follows. If spectral information, as it evolves in time, is crucial to vowel identification, then we would predict that scores should be higher from the time-delay recurrent network than from the concatenated network.

EXPERIMENT 3

Method

The recurrent network that was used in this Experiment incorporates directly the temporal nature of the signal into its architecture with the use of time-delayed links between units. *Temporal Flow* networks (Watrous 1988, 1990) were developed specifically to solve the problem of the representation of time in neural networks for speech processing. Watrous (1990) showed that these networks are capable of a series of phonetic discrimination tasks including place and manner of articulation, voicing and formant trajectories. Two strategies can be used to encode temporal aspects of the signal with these networks. Firstly, delay links from one layer to the next allow the later layer to see the first over a given temporal window and allow "the formation of general filters which can be used for feature detection and pattern matching" (Watrous, 1990, Section II). Secondly, delayed links within the same layer, called recurrent links, "can be used to condition unit responses by temporal context, in

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

order to generate and decode sequential events" (*op. cit.*).

Since the performance of a network can be artificially affected by the degrees of freedom of the network, the number of time-delay links in the hidden layer of the recurrent network was chosen such that the total number of links of the recurrent network matched, as far as possible, that of the concatenated network: That is, each hidden unit sees activation from the current input units plus the hidden units (including itself) one time-slice ago. Each output unit (one per vowel as before) is then connected to every hidden unit with links of zero delay. This network has 1080 links (links between 17 input units and the 20 hidden units, between the 20 hidden units and themselves, and between the 20 hidden units and 17 output units), which is similar to the number of links in the concatenated network.

Training and testing were carried out on the same sets of data as described in the Method section of Experiment 1 using the same critical band values as input to the networks. The networks were again trained until performance on an open test began to degrade.

For the recurrent condition, the input to the network consisted of three separate spectral slices (20%, 50%, 80% time points) plus total vowel duration (17 dimensions per time slice) presented in sequence together with the appropriate output for each vowel. In this case, the unit corresponding to the correct vowel was turned on for the duration of the vowel while all other units were turned off.

Results

86.5% of vowels were correctly identified on the recurrent network compared with 88.6% on the concatenated network. These results are not significantly different [$t = 1.57$, $df = 7$, $p > 0.1$]. A vowel-by-vowel analysis showed that there were no significant differences between the two conditions.

Given this result, the hypothesis that the temporal order of the three spectral slices is important to vowel identification is not supported.

GENERAL DISCUSSION

The series of experiments that has been carried out in this study has sought to investigate the extent to which vowels are dynamic under two different interpretations of 'dynamic'. The first of these is closely related to the concept of the compound target theory as discussed in Nearey and Assmann (1986) and more recently Andruski and Nearey (1992), which suggests that some monophthongs exhibit a consistent formant movement which is not attributable to context effects. The results of the studies by Huang (1992) and Zahorian and Jagharghi (1993), both of which show superior classifications from multiple spectral slices compared with a single slice, are consistent with the position that transitional information contributes to vowel identification. The second position is more closely related to the various dynamic theories of speech production (e.g. Fowler, 1986, Saltzman and Munhall, 1989) which imply that articulatory movement, and therefore the changing acoustic signal as it unfolds in time, provides listeners with the prime cues to vowel identification.

Under the first interpretation (that vowels are consistently characterized by inherent spectral change), we would expect classifications from multiple spectral slices to be

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

better than those from a single spectral slice taken from the spectral midpoint. Our results have shown that diphthongs clearly benefit from the spectral information provided by additional spectral slices, but that monophthongs do not. When training and testing are carried out only on monophthongs, classifications are no better from multiple spectral slices than from a single spectral slice at the midpoint. These results are entirely consistent with a target theory of speech perception: Diphthongs perform better in classifications from multiple spectral slices because they have at least two targets, but since monophthongs are cued by a single target close to the midpoint, the additional information in the transitions makes no difference to their recognition scores.

A separate implication of this first set of results is that monophthongs and diphthongs are two distinct acoustic phonetic categories. This result is inconsistent with the dynamic theory of vowels as developed by Strange and her colleagues, not only because all vowels are presumed to be dynamic in this theory, but also because the concept of a 'vowel target', which is crucial to defining the distinction between monophthongs and diphthongs (one target *vs.* two), is presumed to be irrelevant, or of secondary importance. Whatever their significance for the dynamic theory of vowels, the results from these experiments suggest that future perceptual experiments which seek to determine the relevant saliency of transitions and targets for vowel identification, should be undertaken separately for monophthong and diphthong categories.

Under the second interpretation of 'dynamic' (that the temporal development of the acoustic signal is crucial to vowel identification), we would expect that vowel recognition should benefit from a knowledge of the temporal order of spectral information taken at different time points in the acoustic signal. In order to test this hypothesis, we compared the scores from two kinds of networks. In the first, vowels were classified by concatenating spectral slices, which provides no direct information about the temporal order in which they occur. In the second, vowels were classified from a network in which the dynamically changing features of the vowel are preserved. Since a comparison between these two networks showed nonsignificant differences, we can conclude that the temporal order of the spectral slices is not important for distinguishing between Australian English vowels.

In summary, our results are more consistent with a target theory of vowels. Diphthongs are clearly dynamic because the relevant phonetic information is distributed across the vowel. However, the fact that this phonetic information occurs in a particular temporal order seems to be irrelevant to the identification of the vowels in this study. The results of this study provide little evidence that monophthongs are dynamic, since spectral information in the transitions does not generally improve recognition scores.

The conclusion that favors the target theory of vowel specification rests on the caveat that only Australian English vowels have been tested, and that different results may well be produced for vowels in other accents. We also stress that these networks have only tested two of the many possible interpretations of dynamic, and that vowels may well be cued by other kinds of dynamic information that we have not investigated. Nevertheless, the results do suggest that the theory that vowels are inherently dynamic warrants further investigation.

(Received November 22, 1993; accepted July 13, 1994)

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

REFERENCES

ANDRUSKI, J.E., and NEAREY, T.M. (1992). On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables. *Journal of the Acoustical Society of America*, **91**, 390 – 410.

ASSMANN, P.F., NEAREY, T.M., and HOGAN, J.T. (1982). Vowel identification: orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America*, **71**, 975 – 989.

BENGUEREL, A-P., and McFADDEN, T.U. (1989). The effect of coarticulation on the role of transitions in vowel perception. *Phonetica*, **46**, 80 – 96.

BERNARD, J. (1970). Towards the acoustic specification of Australian English. *Zeitschrift für Phonetik*, **23**, 113 – 128.

CHAN, L., and CHEUNG, Y. (1986). Analysis and recognition of isolated Putonghua vowels by Karhunen-Loève transformation techniques. *Speech Communication*, **5**, 299 – 330.

CLARK, J. E., (1981). A low-level speech synthesis by rule system. *Journal of Phonetics*, **9**, 451 – 476.

CROOT, K., FLETCHER, J., and HARRINGTON, J. (1992). Levels of segmentation and labelling in the Australian national database of spoken language. In J. Pittam (Ed.), *Proceedings of the 4th International Conference on Speech Science and Technology* (pp. 86 – 90). Canberra: Australian Speech Science and Technology Association.

DIEHL, R.L., McCUSKER, S.B., and CHAPMAN, L.S. (1981). Perceiving vowels in isolation and in consonantal context. *Journal of the Acoustical Society of America*, **69**, 239 – 248.

FOWLER, C.A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, **8**, 113 – 133.

FOWLER, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, **14**, 3 – 28.

FOX, R. (1989). Dynamic information in the identification and discrimination of vowels. *Phonetica*, **46**, 97 – 116.

GOTTFRIED, T.L., and STRANGE, W. (1980). Identification of coarticulated vowels. *Journal of the Acoustical Society of America*, **68**, 1626 – 1635.

HENTON, C.G. (1983). Changes in the vowels of received pronunciation. *Journal of Phonetics*, **11**, 353 – 371.

HUANG, C.B. (1992). Modelling human vowel identification using aspects of formant trajectory and context. In Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 43 – 61). Amsterdam: IOS Press.

JENKINS, J.J. (1987). A selective history of issues in vowel perception. *Journal of Memory and Language*, **26**, 542 – 549.

KLEIN, W., PLOMP, R., and POLS, L. (1970). Vowel spectra, vowel spaces, and vowel identification. *Journal of the Acoustical Society of America*, **48**, 999 – 1009.

MACCHI, M.J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. *Journal of the Acoustical Society of America*, **68**, 1636 – 1642.

MANNELL, R., and CLARK, J.E. (1987). Text to speech rule and dictionary development. *Speech Communication*, **6**, 317 – 324.

MITCHELL, A.G., and DELBRIDGE, A. (1965). *The Speech of Australian Adolescents*. Sydney: Angus and Robertson.

NEAREY, T.M., and ASSMANN, P.F. (1986). Modeling the role of inherent spectral change in vowel identification. *Journal of the Acoustical Society of America*, **80**, 1297 – 1308.

O'SHAUGHNESSY, D. (1987). *Speech Communication*. Reading, MA: Addison-Wesley.

PARKER, E.M., and DIEHL, R.L. (1984). Identifying vowels in CVC syllables: effects of inserting silence and noise. *Perception & Psychophysics*, **36**, 369 – 380.

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

RAKERD, B., and VERBRUGGE, R.R. (1987). Evidence that the dynamic information for vowels is talker independent in form. *Journal of Memory and Language*, **26**, 558 – 563.

RAKERD, B., VERBRUGGE, R.R., and SHANKWEILER, D.P. (1984). Monitoring for vowels in isolation and in a consonantal context. *Journal of the Acoustical Society of America*, **76**, 27 – 31.

SALTZMAN, E., and MUNHALL, K. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, **1**, 333 – 382.

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

STRANGE, W. (1989a). Evolving theories of vowel perception. *Journal of the Acoustical Society of America*, **85**, 2081 – 2087.

STRANGE, W. (1989b). Dynamic specification of coarticulated vowels spoken in sentence context. *Journal of the Acoustical Society of America*, **85**, 2135 – 2153.

STRANGE, W., JENKINS, J.J., and JOHNSON, T.L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, **74**, 695 – 705.

STRANGE, W., VERBRUGGE, R.R., SHANKWEILER, D.P., and EDMAN, T.R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, **60**, 213 – 224.

VERBRUGGE, R.R., and RAKERD, B. (1986). Evidence of talker independent information for vowels. *Language and Speech*, **29**, 39 – 55.

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**, 1753 – 1772.

ZAHORIAN, S.A., and JAGHARGHI, A.J. (1993). Spectral shape features versus formants as acoustic correlates for vowels. *Journal of the Acoustical Society of America*, **94**, 1966 – 1982.

ZWICKER, E. (1961). Subdivision of the audible frequency range into critical bands (Frequenzgruppen). *Journal of the Acoustical Society of America*, **33**, 248.

APPENDIX: AUSTRALIAN ENGLISH

It is generally considered that there are three main accent groups of Australian English which are, for the most part, unaffected by regional variation. The accents are *Cultivated, General, and Broad Australian*. It is also usual to consider these three groups as points on a continuum extending from Cultivated, which bears the strongest resemblance to British English Received Pronunciation (RP), to Broad, which shares some similarities with London Cockney English (although there are also many differences). These three accent groups do not take account of the many migrant-Australian accents which are spoken by roughly 20 – 25% of the population, nor of Aboriginal-Australian accents.

From a *phonemic* point of view, there is more or less systemic equivalence between the vowel systems of most Australian accents and British English RP. Additionally, both Australian English (AE) and British English RP are non-rhotic.

The salient *phonetic* differences between the AE and RP vowels are partly

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

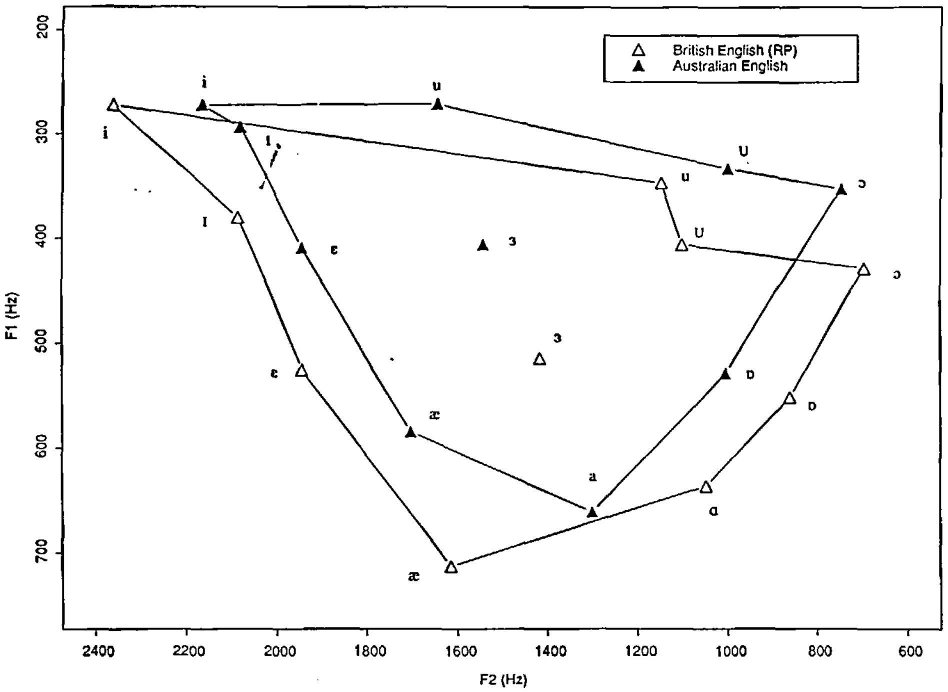

Fig. 2. A comparison of the F1/F2 plane for British English and Australian English.

illustrated by a plot of the vowels in the formant plane in Figure 2. Ten vowels are represented per accent. For the Australian accent, each vowel is an average of 110 tokens taken from /CVd/ citation-form words (C = /b d g p t k f s t h/) produced by the same five male speakers who participated in this study (thus 22 tokens per speaker). A basic form of speaker normalization was applied prior to averaging, in which a speaker's centroid values (mean of F1 and mean of F2) were subtracted from each of the same speaker's F1 and F2 values. The normalized F1 and F2 values were then rescaled by adding the mean of F1 and the mean of F2 across all five speakers to the normalized values. The RP vowels are taken from Henton (1983). These are based on 10 male speakers' productions of /hVd/ citation-form words.

The plot illustrates some of the principal differences between these two accent groups that have been previously noted (Bernard, 1970). Firstly, the front AE vowels /ɪɛ/ are closer than their RP counterparts. Secondly, some of the RP back vowels, notably /u/ (*who'd*) and /ɑ/ (*hard*, transcribed as /a/ in AE), have fronted counterparts in AE. Thirdly, some of the RP back vowels, in particular /ɔ/ (saw) and /U/ (hood) have raised equivalents in AE.

Concerning the categorization of vowels as monophthongs and diphthongs, both AE and RP have rising diphthongs /eɪ oʊ aɪ aʊ ɔɪ /. Some AE talkers also have three centering diphthongs /ɪə ɜə ʊə/ (*here, there, cure*), although there is a strong tendency, particularly in younger AE talkers for these to be produced as monophthongs (the /ʊə/

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013

phoneme is now replaced by /ɔ/ for most Australian talkers). The remaining vowels are presumed to be monophthongal in AE, with the possible exception of /i u/, which in some AE accents, particularly Broad Australian, can sometimes be produced with a long onglide from schwa (Mitchell and Delbridge, 1965). In the Macquarie University text-to-speech system of Australian English (Clark, 1981; Mannell and Clark, 1987), all the rising and centering diphthongs are currently specified by two targets; the monophthongs are all specified by a single target with the exception of /i/ which can be realized with an optional long onglide.

Downloaded from las.sagepub.com at Bayerische Staatsbibliothek on February 25, 2013