Acoustic evidence for dynamic formant trajectories in Australian English vowels

Catherine I. Watson[a] and Jonathan Harrington
Speech Hearing and Language Research Center, Macquarie University, Sydney 2109, Australia

(Received 19 September 1997; revised 18 September 1998; accepted 5 March 1999)

The extent to which it is necessary to model the dynamic behavior of vowel formants to enable vowel separation has been the subject of debate in recent years. To investigate this issue, a study has been made on the vowels of 132 Australian English speakers (male and female). The degree of vowel separation from the formant values at the target was contrasted to that from modeling the formant contour with discrete cosine transform coefficients. The findings are that, although it is necessary to model the formant contour to separate out the diphthongs, the formant values at the target, plus vowel duration are sufficient to separate out the monophthongs. However, further analysis revealed that there are formant contour differences which benefit the within-class separation of the tense/lax monophthong pairs. © *1999 Acoustical Society of America.* [S0001-4966(99)05706-9]

PACS numbers: 43.72.Ar, 43.70.Fq, 43.70.Hs [JH]

INTRODUCTION

In the last 50 years, many different kinds of investigations have established the first two formant center frequencies as the main determiners of vowel quality. Research into articulatory-to-acoustic modeling (Stevens, Kasowski, and Fant, 1953; Fant, 1960), studies of acoustic phonetic cues (Peterson and Barney, 1952; Ladefoged, 1967), as well as various different kinds of speech perception experiments (Delattre *et al.*, 1952; Klein, Plomp, and Pols, 1970) have all demonstrated the strong correlation between the first two formant frequencies and decreasing phonetic height and backness, respectively, thereby establishing a quadrilateral-like shape when vowel tokens are plotted in the $-F2/-F1$ plane (Essner, 1947; Joos, 1948). The formant frequencies are usually extracted at the acoustic vowel target (Lindblom, 1963; Lehiste and Peterson, 1961) which is presumed to be the section of the vowel that is least influenced by phonetic context effects: the vowel target is therefore closest in quality to the same phonetic vowel in a citation-form production in a context that is largely uninfluenced by flanking consonants.

Although the effectiveness of the first two, or possibly first three, formant frequencies in vowel separation is indisputable, it is also recognized that temporal information provides many kinds of cues to vowel quality. Some of these are very well-known. For example, the tense and lax vowels in English are acoustically long and short in duration, respectively, resulting in minimal pairs such as ''heed''/''hid'' and, in Australian English, ''dark''/''duck,'' in which the durational cues are as important as those due to formant differences. Equally, most varieties of English include diphthongs that have an early first target compared with that of monophthongs, as well as spectral transitions which either attain or (more commonly) point towards a second target (Bladon, 1985; Fox, 1983; Gay, 1970; Gottfried, Miller, and Meyer, 1993). The time at which the target occurs (relative to the vowel onset and offset) may provide another source of temporal information: some vowels may have long or short vowel onglides or offglides resulting in a considerable temporal displacement of the vowel target from the temporal midpoint (e.g., for American English, see Lehiste and Peterson, 1961; Huang, 1986; Strange, 1989b; for Australian English, see Bernard, 1981; Cox, 1996; Harrington, Fletcher, and Beckman, in press; Harrington, Cox, and Evans, 1997).

Since vowels are distinguished not only in height and backness as cued principally by formants at the vowel target, but also by various temporal features as represented by some of the time-varying cues of the kind described above, it is perhaps not surprising that a number of acoustic studies (Harrington and Cassidy, 1994; Hillenbrand *et al.*, 1995; Huang, 1992; Nearey and Assmann, 1986; Zahorian and Jagharghi, 1993) as well as various perception experiments in which spliced sections of syllables were presented to listeners for identification (e.g., Benguerel and McFadden, 1989; Jenkins, Strange, and Miranda, 1994; Strange *et al.*, 1976; Strange, Jenkins, and Johnson, 1983; Strange, 1989a), have found that vowels are suboptimally differentiated if the acoustic information is entirely based on a static spectral slice either as formants, or some other kind of parameterization at the vowel target.

Thus, acoustic classification studies in Harrington and Cassidy (1994), Hillenbrand *et al.* (1995), Huang (1992), and Zahorian and Jagharghi (1993) have all found that vowels of different quality are more effectively separated when the acoustic parameters are based on spectral information extracted at multiple time points, rather than just at the vowel target. As discussed in Harrington and Cassidy (1994), at least part of the reason for this finding is that diphthongs in particular, which are characterized by a trajectory between two targets (Bladon, 1985; Holbrook and Fairbanks, 1962), remain largely undifferentiated from those monophthongs to which their first targets are closest in quality in a single-target space. To a certain extent, the reasoning in Harrington and Cassidy (1994) is consistent with a number of studies by

[a]Electronic mail: watson@srsuna.shlrc.mq.edu.au

Nearey (Andruski and Nearey, 1992; Nearey and Assmann, 1986) which have shown that, for Canadian English at least, many vowels which might have been assumed to be monophthongal in fact exhibit a quasidiphthongal quality. Therefore, the spectral change which is observed in some of these vowels cannot just be attributed to the influence of phonetic context, but is instead a systematic property of the vowel itself, in much the same way that the movement towards a low $F1$ and high $F2$ approaching the vowel offset is an inherent feature of the diphthong /aɪ/ in Australian English, Southern British English, and no doubt many other varieties of English (Wells, 1982).

The numerous speech-perception experiments by Strange, Jenkins, and colleagues (see, e.g., Jenkins *et al.*, 1994 and Strange, 1989b for reviews of this literature) are founded on a very different interpretation of the way in which time-varying spectral information benefits vowel discrimination. Their experiments are usually based on presenting listeners with various kinds of edited speech stimuli for identification. In some of these, the syllables contain only the initial and final transitions and exclude the relatively steady-state part of the vowel; in others, listeners are asked to identify the vowel from syllables without transitions, i.e., containing only the section around the target. Most of these studies show that listeners identify vowels from the targetless (transition-only) syllables either with a very low error rate, or as well as from the original unmodified syllables (e.g., Benguerel and McFadden, 1989; Fox, 1989; Jenkins, Strange, and Edman, 1983; Nearey, 1989; Parker and Diehl, 1984; Rakerd and Verbrugge, 1987; Verbrugge and Rakerd, 1986), and there is also recent evidence (Hillenbrand and Gayvert, 1993) to show that when vowels were synthesized from the steady-state values in Peterson and Barney (1952), they were not accurately identified by listeners. At one level—which is perhaps not in dispute—Strange, Jenkins, and colleagues (henceforth SJ) conclude that there must be considerable information contained in the vowel transitions (perhaps as much, if not more than at the target) for vowel identification to be possible despite the fact that the section of the vowel which is usually assumed to contain the most important information—i.e., the vowel target—has been discarded. At another level, which is certainly more controversial, SJ attribute listeners' identifications of vowels from targetless syllables to the theory that vowel perception is made with reference to dynamic articulatory information which is predominantly encoded in the syllable's transitions (e.g., Fowler, 1987, and Strange, 1987). This idea is, in turn, based on the earlier action theory framework of speech production as developed by Fowler and colleagues (Fowler, 1980, 1983, 1986; Fowler and Rosenblum, 1989), and on the more recent task-dynamic model of speech production (Browman and Goldstein, 1992; Saltzman and Munhall, 1989) which SJ believe to be ''on the right track'' from the point of view of our understanding of how listeners decode segments phonetically from the acoustic speech signal (Jenkins *et al.*, 1994).

There are, however, at least two major unresolved issues that arise from the body of research on vowel dynamics carried out by SJ in the last 20 years. First, although SJ relate the results of their perception experiments to the existence of articulatory-dynamic features that are coded in the acoustic signal and that are primary in listeners' identification of vowels, the exact nature of these cues remains elusive: as Jenkins *et al.* (1994) state, the acoustic phonetic attributes of a vowel may be represented by a multitude of targets, rather than two, three, or four targets—which essentially implies that it is the whole dynamic spectral trajectory, rather than any number of static targets, which cues the vowel. Second—and this is closely related to the first point—it has not been established how much of the dynamic spectral change is relevant to vowel identification and what proportion is attributable to other factors: as Jenkins *et al.* (1994) acknowledge, numerous studies over the last 50 years have shown that a good deal of spectral change is caused by context effects that might *complicate*, rather than enhance, the relationship between the acoustic signal and the vowel class.

The research methodology for the present investigation is closely based on Harrington and Cassidy (1994) and is similar to other investigations (e.g., Huang, 1986, and Zahorian and Jagharghi, 1993) which seek to determine the extent to which the dynamic aspects of the vowel's acoustic signal contribute to its identification beyond the ''static'' information which is available at the vowel target. Essentially, the aim of this research is to begin to find answers to the following question: if two classification algorithms are applied to the same body of acoustic vowel data, such that the first is based on vowel-target information alone and the second on both vowel-target and dynamic information throughout the vowel's time course, to what extent are vowels more effectively separated in the second space compared with the first? This is the approach taken in Harrington and Cassidy (1994), but in contrast to that study, the present one is based on a larger group of talkers (62 male, 70 female) and a different dynamic classification algorithm. In Harrington and Cassidy (1994), the contribution of dynamic information to vowel separation beyond that encoded at the target was estimated by comparing classification scores obtained from three time slices in the vowel with those from a single time slice at the midpoint using either a Gaussian model or a time-delay (recurrent) neural network to classify the vowels. The general conclusion from that study was that, since only diphthongs but not monophthongs were more effectively separated in the classifications from three time slices compared with one, the results were more consistent with a target theory of vowel distinction than the dynamic theory proposed by SJ. Part of the aims of the present study is to reassess this interpretation by classifying a large number of vowels using an algorithm that represents the dynamic behavior of the formants with a discrete cosine series. The advantage of this approach over that taken in Harrington and Cassidy (1994) is twofold: first, it does not require the extraction of spectral information at a number of arbitrary time points; second, the parameters from the discrete cosine series are considerably less correlated with each other than those obtained by merely sampling the formant tracks. The second aim of the paper is to try to define more precisely the nature of the dynamic information that characterizes some (or all) vowels when classifications from the dynamic algorithm show superior scores compared with single-target classifications. Finally, we consider the ad-

equacy of the discrete cosine series as an algorithm for representing dynamic changes in the vowel's spectral shape.

I. EXPERIMENT 1

There are two aims in this experiment. The first is to assess the adequacy of classifying vowels (diphthongs and monophthongs) from target information plus total vowel duration alone. This assessment is made by comparing the results of a classification from the target-only space with a dynamic space formed from the discrete cosine transformation (DCT) applied to the formant trajectories. If there is information for vowel separation beyond that which is represented at the vowel target, then the separation between the vowels should be greater in DCT space.

The second part is an investigation into whether modeling the formant contour might benefit the within-class separation of the monophthongs. Once again, we will compare the results of a classification from a target-only space to a dynamic space formed by modeling the formant trajectories with DCT coefficients. To neutralize as far as possible the obvious contribution from acoustic vowel duration differences, this parameter is excluded in the classifications.

A. Method

1. Talkers

The vowels that were analyzed were taken from the isolated word materials collected under the Australian National Database of Spoken Language (ANDOSL) (Millar, Harrington, and Vonwiller, in press; Millar *et al.*, 1994; Vonwiller *et al.*, 1995). In selecting the talkers for ANDOSL, an attempt was made to cover three age ranges (18–30, 31–45, 46+ yrs) and also to select Australian English talkers from the main accent types that form a continuum from *cultivated* to *broad*, where cultivated bears the closest resemblance to British English received pronunciation (RP) and broad is identified as the most characteristically ''Australian'' accent and shares some characteristics with London Cockney English (Cochrane, 1989; Horvath, 1985). *General* Australian is the third recognized major accent type and falls between these two categories on this continuum (Blair, 1993; Horvath, 1985). This accent variation is determined primarily by socioeconomic factors, and there is scarcely any regional accent variation in Australia. (See Cox, 1996; Bernard, 1967; and Harrington *et al.*, 1997 for acoustic phonetic analyses of Australian English vowels and Harrington and Cassidy, 1994 for a comparison with RP vowels.) The talkers were also balanced for gender. A total of 266 talkers were recorded as part of the ANDOSL project, of which 140 subjects were nonaccented native speakers of Australian English, born in Australia, and with the exception of three subjects, of Anglo/Celtic origin (Vonwiller *et al.*, 1995). Six of these talkers did not complete the recording sessions for the list of single-word utterances, leaving 134 subjects who participated in the isolated word task. Two further subjects were removed from consideration because exceptionally, they produced rhotic vowels (Australian is nonrhotic). The present study is concerned with the isolated-word productions from the remaining 132 talkers.

The differences between the three accent types, although perceptible, are minor compared with the extent of accent variation found in American English. The acoustic phonetic studies by Bernard (1970), Cox (1998), Harrington *et al.* (1997), and Watson, Harrington, and Evans (1998) of Australian English vowels all show that the differences in vowel quality are confined primarily to the duration of the onglide in the tense high vowels (most extensive for ''broad''), a raised $F2$ in broad /u/ (as in WHO'D) which may be due either to less lip-rounding and/or more tongue-fronting, and various first target differences in the rising diphthongs /aɪ aʊ eɪ/ (as in HIGH, HOW, and HAY, respectively). However, even these accent differences are generally not sufficient to substantially increase the overlap between different vowel categories. This was substantiated by a pilot study in which we carried out many of the classifications to be reported in this paper on the data from the male general talkers alone: the classification scores never showed any significantly greater overlap between the vowel categories than from the same classification applied to the male talkers' vowels pooled across the three accent categories. We repeated this study for the female talkers and got exactly the same results.

2. Materials

As described in Millar *et al.* (in press), and Vonwiller *et al.* (1995), the ANDOSL talkers read citation-form productions of 25 different words. For the present paper, we selected the words from an /hVd/ or /hV/ context. In addition, we also selected HOIST and TOUR words to include the /ɔɪ ʊə/ diphthong nuclei, neither of which occurred in the /hV/ or /hVd/ context. This gave us 19 words in total.

All of the word tokens were labeled phonetically at the Speech Hearing and Language Research Center, Macquarie University, using the procedures described below and in Croot, Fletcher, and Harrington (1992). Any words that were incorrectly produced (e.g., /hæd/ for HARD) were removed from consideration; we also rejected all TOUR words which had been produced with a monophthongal /ɔ/ nucleus (as in HOARD). The final distribution of the words used in this study, together with their (phonological) subcategorisations *tense monophthong*, *lax monophthong*, *rising diphthong*, and *falling diphthong* used in this paper, are shown in Table I.

3. Recording, digitization, labeling

The subjects were all recorded in an anechoic environment at the National Acoustics Laboratories, Sydney. The material was recorded in a single session, and, for the isolated word lists, was presented to the subjects on a computer screen one word at a time to avoid list intonation (see Millar *et al.*, in press, for further details). The speech data was digitized at 20 000 Hz with a 16-bit resolution and the first three formant center frequencies and their bandwidths were automatically tracked in ESPS/Waves (the settings were 12th-order linear predictive coding analysis, cosine window, 49-ms frame size, and 5-ms frame shift).

All automatically tracked formants were checked for accuracy, and hand corrections were made when considered necessary. Formant tracking errors were especially common

TABLE I. The list of vowels used in the study, the words from which they were extracted, and the number of vowel tokens from the female speakers and male speakers, respectively.

Tense monophthongs				Lax monophthongs			
		Number of tokens				Number of tokens	
Word	Phoneme	Female	Male	Word	Phoneme	Female	Male
HEED	i	70	62	HID	I	70	62
WHO'D	u	69	62	HOOD	ʊ	70	62
HOARD	ɔ	70	61	HOD	ɒ	69	62
HARD	a	68	62	HUD	ʌ	70	62
HEARD	ɜ	68	62	HEAD	ɛ	69	62
				HAD	æ	70	62

Rising diphthongs				Falling diphthongs			
		Number of tokens				Number of tokens	
Word	Phoneme	Female	Male	Word	Phoneme	Female	Male
HAY	eɪ	70	62	HEAR	ɪə	70	60
HOE	oʊ	70	62	HAIR	ɛə	70	61
HOIST	ɔɪ	70	62	TOUR	ʊə	63	48
HIDE	aɪ	70	62				
HOW	aʊ	70	62				

in vowels which have $F1$ and $F2$ close together (i.e., back-rounded vowels such as HOARD) but also in some high-front vowels when $F2$ and $F3$ merged. Occasionally, a formant that was very low in amplitude might not have been tracked (e.g., $F3$ in high-back vowels) and for a few very high-pitched voices, the fundamental was sometimes mistracked as $F1$. Predictably, many more errors occurred for the female data than for the male data. Approximately one-third of all the vowels required some hand correction, and this was done by redrawing the formant tracks by hand using either the Waves or EMU (Harrington and Cassidy, in press) tools, occasionally after recalculating the spectrogram with different fast Fourier transform (FFT) sizes to allow a closer examination of some of the harmonics in relation to the tracked formants.

The acoustic onset of the vowel was marked at the onset of voicing as shown by strong vertical striations in the spectrogram, and by the onset of periodicity in the waveform. The acoustic offset of the vowel in the /hVd/ context was marked at the closure of the [d] corresponding to a cessation of regular periodicity for the vowel and/or a substantial decrease in the amplitude of the waveform. The acoustic vowel target was marked as a single time point between the acoustic onset and offset according to the criteria (see, also, Harrington *et al.*, 1997). For high- and mid-front vowels, the target was marked where $F2$ reached a peak; for mid- and high-back vowels, the target was marked where $F2$ reached a trough; for open vowels, the target was marked at the $F1$ peak. When there was no evidence for a target based on formant movement (this could happen especially in central vowels), then other acoustic criteria were used such as the time at which the amplitude reached its maximum value. If there was no acoustic evidence of any kind for a target—which implies neither formant nor amplitude change between the acoustic onset and offset of the vowel—the target was marked at the vowel's temporal midpoint. In rising diphthongs, two targets were marked using the same sets of criteria as for the monophthongs; however, only the first target was used in this study. For the three falling diphthongs, only the first target could be reliably marked at the $F2$ maximum (HERE, HAIR), or the $F2$ minimum (TOUR).

Figure 1 shows the averaged trajectories for the first three formants for each vowel class from the male data. The tokens were time aligned at the vowel target (first target for diphthongs) and averaged separately in each vowel class.

4. The discrete cosine transformation

Similar to Zahorian and Jagharghi (1993), we represented the time-varying nature of the formant trajectories by modeling the trajectories with coefficients of the discrete cosine transform (DCT). The cosine basis function we used to model the trajectories was

$$C(m)\cos(\theta), \tag{1}$$

where $C(m)$ is the amplitude of the mth cosine and θ is related to the number of sample points in a trajectory. The amplitudes of the cosines, $C(m)$, are the output of a DCT. There are several ways the DCT can be expressed (Rao and Yip, 1990): the form chosen was

$$C(m)=\frac{2}{N}k_m\sum_{n=0}^{N-1}x(n)\cos\left(\frac{(2n+1)(m-1)\pi}{2N}\right)$$
$$m=1,...,N, \tag{2}$$

where $x(n)$, $n-0,...,N-1$ is the trajectory of the feature being modeled, N is the number of points in the trajectory, $C(m)$ denotes the mth DCT coefficient, and k_m is $1/\sqrt{2}$ when $m=1$, and is 1 when $m\neq 1$.

Figure 2 shows the second formant frequency trajectories for /i/ (from HEED) and /a/ (from HARD) and their decomposition into the first three cosines of the basis function [see Eq. (1)]. The first cosine is a straight line (the dc offset) at a value proportional to the mean of the original

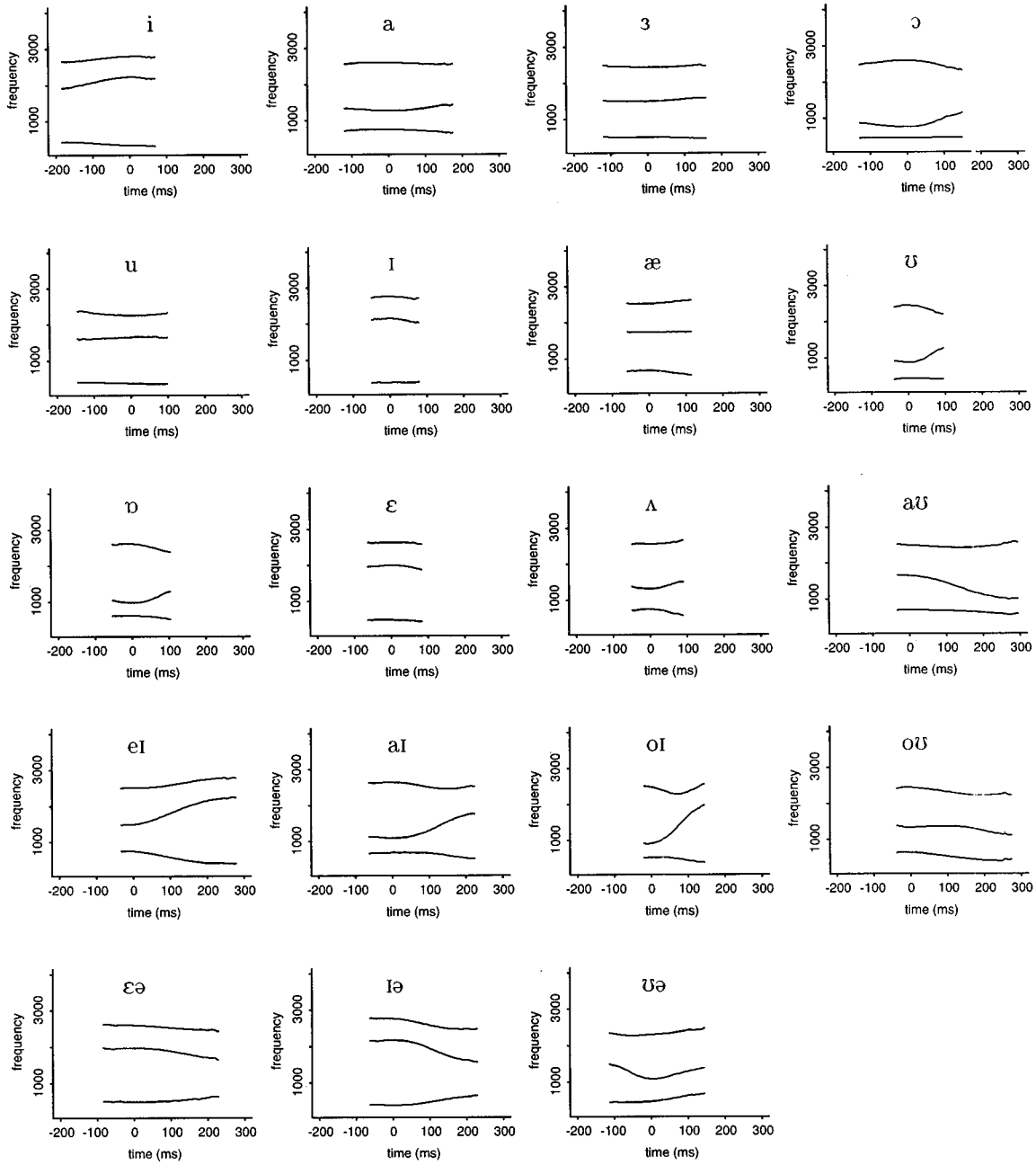

FIG. 1. The averaged formant trajectories for the first three formants for each vowel class from the male data, aligned at the target for the monophthongs and the first target for the diphthongs. The averaged formant trajectories were truncated at the left/right edges at the average durations from the vowel target to the vowel onset/offset. For each vowel, the target is positioned at 0 ms.

formant trajectory calculated between the formant onset and offset. The second basis function is a half-cycle cosine wave which models both the direction and the magnitude of tilt of the formant trajectory's tilt. The third basis function, which is a whole-cycle cosine wave, is a measure of the trajectory's curvature. Consider, for example, the $F2$ trajectory of a high-front vowel in a flanking labial context. In this case, the curvature would be extreme because the trajectory has to span the divergent low $F2$ labial loci and the high $F2$ target. By contrast, a mid-front vowel in a flanking alveolar context would have a much lower value on this parameter because the consonant loci and $F2$ target values are very similar (resulting in an almost straight-line trajectory).

We performed an initial pilot study to decide which DCT coefficients we should use to encode the formant trajectories. We considered the female and male data separately. The results of this study (Table II) give the percentage of vowels correctly identified in vowel classification experiments when the first three formant trajectories were encoded with different combinations of the first three coefficients of the DCT. From these results, we concluded that the first and second DCT coefficients played significant roles in separating the vowels, whereas the third coefficient did not. The total number of vowels correctly classified using the first two coefficients to encode the formant trajectories was significantly greater than using the first and third coefficients, the second and third coefficients, and the first coefficient only ($\alpha \ll 0.002$; the significance levels were Bonferroni-corrected

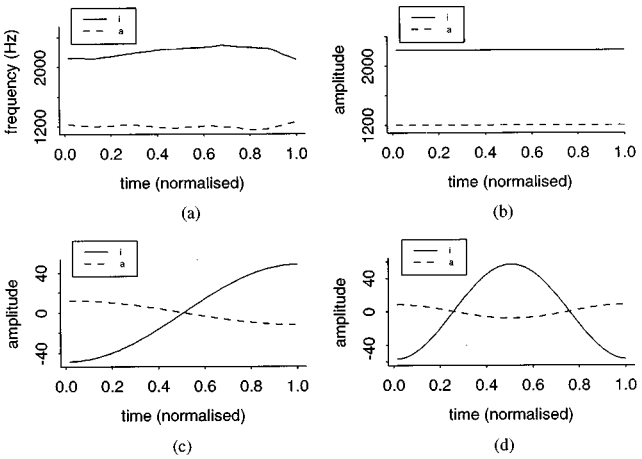

FIG. 2. The F_2 trajectory of an /i/ token and an /a/ token and their representation from the first three elements in the cosine basis function.

to reduce the probability of a type I error due to multiple testing). Further, there was no significant different in the vowel classification scores from the first two DCT coefficients and the first three DCT coefficients. We repeated the test for the monophthong vowels only and got exactly the same findings. For these reasons, we decided to represent each formant trajectory with the first two discrete cosine transformation (DCT) coefficients (a six-parameter space). This is in contrast to earlier studies (e.g., Zahorian and Jagharghi, 1993) where the first three DCT coefficients were used.

Further support for using just the first two DCT coefficients to model the Australian English formant trajectories can be seen in Fig. 1. The trajectories are not complex shapes, and the most distinguishing features are the mean values of the formants and the slope of the formant trajectories. The means cannot by themselves function to distinguish some tense from lax vowels (e.g., /a/ from /ʌ/), nor monophthongs from diphthongs that have similar formant values at the first target (e.g., /a/ from /aI/). Neither is the discriminating power of the slope (the second DCT coefficient) sufficient by itself because most monophthongs have similar slopes but different means.

5. Classification procedure

All results reported in this paper were obtained using a Gaussian classification technique. In Gaussian classification, the centroid and covariance matrix of the training set are estimated for each vowel class. Tokens from the test set are then classified based on their Bayesian distances to each of the class centroids.

A ''round-robin'' procedure was used to train and test the classifier. In this procedure, all the tokens for a single speaker were used as a test set and the remaining speakers' tokens formed the training set. This was repeated for all the speakers in turn, and the overall classification score was the summation of the individual results.

In the experiment, vowel-classification scores from two sets of parameters were compared with each other. In the first part of the experiment, the first set was formed from a four-dimensional space that included $F1-F3$ at the vowel target and acoustic vowel duration. The second set was formed from a seven-dimensional space that included the first two DCT coefficients of $F1-F3$ and acoustic vowel duration.

The results of the target and DCT analyses were compared using a *t*-test on data broken down on a per-speaker basis. On a vowel-by-vowel basis, all the significance levels were Bonferroni-corrected to reduce the probability of a type I error due to multiple testing. For the complete vowel set (19 in total), only the variables with $p<0.003$ are significantly different at an experiment-wide significance level of $p<0.05$. For the monophthong set (11 vowels in total), only the variables with $p<0.005$ are significantly different at an experiment-wide significance level of $p<0.05$. The female and male data were compared separately to avoid any confounding influences due to vowel normalization effects. In general, when a significant difference between two methods was identified for the female data, it was also identified in the male data (or vice versa); therefore, significant differences reported below are applicable to both the male and female data unless stated otherwise.

TABLE II. The percentage of vowels correctly identified in vowel-classification experiments when the first three formant trajectories were encoded with different combinations of the first three coefficients of the DCT.

	All vowels		Monophthongs	
Parameter set	Female	Male	Female	Male
$C(1)$	53.3	52.7	72.3	75.5
$C(1),C(2)$	84.0	86.2	81.8	85.5
$C(1),C(3)$	68.6	68.0	73.1	76.8
$C(2),C(3)$	65.2	65.1	59.4	57.1
$C(1),C(2),C(3)$	84.2	85.6	81.1	85.4

B. Results

Table III shows the results of classifications from formants at the vowel target (henceforth, *target analysis*) and vowel duration and those from the DCT coefficients (henceforth, DCT *analysis*) and vowel duration. The table shows that the total number of correctly classified vowels was significantly higher in the DCT analysis than in the target analysis. On a vowel-by-vowel basis, it is predominantly the diphthongs that had a significantly greater separation in the DCT analysis than in the target analysis. This suggests that rather than all vowels being dynamic, it is the diphthongs that are best described using a dynamic method of analysis.

Some support for this view is provided by the results in Table IV, in which the same two classification experiments were carried out again, but this time by training and testing only on the (so-called) monophthongs (see Table I). In this case, the results show first very high classification scores for both the target analysis and the DCT analysis, and, in clear contrast to the experiment in which training and testing was carried out on all 19 vowel types, no significant differences between the two analyses on a vowel-by-vowel basis.

When total vowel duration was excluded as a parameter in the classification, there was still no significant difference in the classification scores for many of the monophthongs in

TABLE III. The classification scores (percent correctly classified) from the target analysis (*left column*) and from the DCT analysis (*right column*) for the male and female talkers after training and testing on the first three formant frequencies and total vowel duration. All the *p* values are given and significant differences are shown by **. Note on a vowel-by-vowel basis, the Bonferroni correction was applied ($\alpha<0.003$).

Female				Male			
	Target & duration	DCT coef. & duration	*p*		Target & duration	DCT coef. & duration	*p*
æ	87.1	91.4	0.08	æ	95.2	93.5	0.6
ɛ	68.1	56.5	0.03	ɛ	91.9	85.5	0.1
ɪ	82.9	84.3	0.7	ɪ	85.5	95.2	0.01
ɒ	79.7	89.9	0.007	ɒ	80.6	98.4	0.002**
ʊ	82.9	94.3	0.01	ʊ	91.8	96.7	0.2
ʌ	95.7	90.0	0.1	ʌ	98.4	96.8	0.3
ɜ	88.2	92.6	0.3	ɜ	88.7	95.2	0.1
a	60.3	95.6	<0.003**	a	75.8	98.4	<0.003**
i	72.9	95.7	<0.003**	i	69.4	100.0	<0.003**
ɔ	97.1	98.6	0.3	ɔ	96.7	95.1	0.6
u	84.1	88.4	0.2	u	87.1	96.8	0.01
eɪ	40.0	98.6	<0.003**	eɪ	56.5	100.0	<0.003**
ɔɪ	48.6	100.0	<0.003**	ɔɪ	64.5	98.4	<0.003**
aɪ	66.7	100.0	<0.003**	aɪ	82.3	91.9	0.08
aʊ	31.4	95.7	<0.003**	aʊ	41.9	96.8	<0.003**
oʊ	70.0	94.3	<0.003**	oʊ	37.1	90.3	<0.003**
ɪə	34.3	78.6	<0.003**	ɪə	60.0	78.3	0.02
ɛə	80.0	70.0	0.05	ɛə	85.2	88.5	0.5
ʊə	90.5	95.2	0.2	ʊə	89.6	95.8	0.3
all	71.5	90.0	<0.05**	all	77.7	94.3	<0.05**

the target analysis compared with the DCT analysis. However, as Table V shows, four vowels that form two tense/lax pairs in Australian English (/i,ɪ/ and /a,ʌ/) were more effectively separated in the DCT analysis, which suggests that there may be some differences in the shapes of the formant trajectories that contribute to their separation.

C. Discussion

The object of the first part of the experiment has been to assess how effectively vowels were separated when they are classified from a ''static'' section at the vowel target (the target analysis) and from ''dynamic'' information which was represented by discrete cosine coefficients (the DCT analysis). Consistent with Harrington and Cassidy (1994), the results of the classifications suggest that only those vowels that have traditionally been labeled ''diphthongs'' benefited significantly from the additional information which is encoded in the dynamic analysis. If those vowels that are traditionally labeled ''monophthongs'' were inherently dynamic, we would expect higher classification scores in the DCT analysis than when training and testing are carried out on these vowels alone. In summary, the results given in Tables III and IV are so far entirely compatible with a target theory of vowels: the information for monophthong identification is encoded at the vowel target, while diphthong identification requires information from more than one static spectral section.

Table V, however, suggests it would be premature to conclude that the formant trajectory shapes of monophthongs contain no useful information for the separation of /i,I/ and

TABLE IV. The classification scores (percent correctly classified) from the target analysis (*left column*) and from the DCT analysis (*right column*) for the male and female talkers following training and testing on the first three formant frequencies and total vowel duration of the monophthongs only. All the *p* values are given and significant differences are shown by **. Note on a vowel-by-vowel basis, the Bonferroni correction was applied ($\alpha<0.005$).

Female				Male			
	Target & duration	DCT coef. & duration	*p*		Target & duration	DCT coef. & duration	*p*
æ	91.4	92.9	0.3	æ	96.8	93.5	0.2
ɛ	58.1	56.5	0.03	ɛ	91.9	87.1	0.2
ɪ	82.9	84.3	0.7	ɪ	85.5	95.2	0.01
ɒ	91.3	89.9	0.3	ɒ	96.8	100	0.2
ʊ	95.7	94.3	0.6	ʊ	95.1	96.7	0.6
ʌ	97.1	91.4	0.04	ʌ	98.4	98.4	1.0
ɜ	92.6	92.6	1.0	ɜ	91.9	95.2	0.3
a	97.1	98.5	0.3	a	98.4	98.4	1.0
i	92.9	97.1	0.2	i	96.4	100.0	0.2
ɔ	98.6	98.6	1.0	ɔ	98.4	98.4	1.0
u	87.0	88.4	0.7	u	90.3	96.9	0.04
all	90.4	89.5	0.3	all	94.6	96.3	0.03**

TABLE V. The classification scores (percent correctly classified) from the target analysis (*left column*) and from the DCT analysis (*right column*) for the male and female talkers following training and testing on the first three formants of the monophthongs excluding total vowel duration. All the *p* values have been given and significant differences are shown by **. Note on a vowel-by-vowel basis, the Bonferroni correction was applied (α<0.005).

Female				Male			
	Target	DCT coef.	*p*		Target	DCT coef.	*p*
æ	90.0	90.0	1.0	æ	90.3	88.7	0.6
ɛ	52.2	58.0	0.3	ɛ	88.7	80.6	0.1
ɪ	55.7	82.9	<0.005**	ɪ	59.7	90.3	<0.005**
ɒ	91.3	94.2	0.3	ɒ	91.9	90.3	0.6
ʊ	72.9	74.3	0.8	ʊ	83.6	86.9	0.4
ʌ	62.9	80.0	0.02	ʌ	66.1	64.5	0.8
ɜ	88.2	89.7	0.7	ɜ	91.9	96.8	0.2
a	45.6	67.6	<0.005**	a	40.3	77.4	<0.005**
i	68.6	92.9	<0.005**	i	66.1	93.5	<0.005**
ɔ	82.9	87.1	0.4	ɔ	85.2	88.5	0.5
u	76.8	82.6	0.5	u	85.2	90.3	0.3
all	71.8	81.8	<0.05**	all	77.2	86.7	<0.05**

/a,ʌ/. Many researchers have shown that there are clear time-varying differences in the formants of monophthong tense–lax pairs which may contribute to their separation beyond total vowel-duration differences. It is also known that Australian tense high vowels can be produced with delayed targets that may serve to distinguish them from their lax counterparts.

Figure 3, which shows averaged time-normalized formant trajectories for these vowel pairs for both male and female talkers, suggests that these tense/lax vowel pairs may differ in the time at which the target occurs relative to the vowel onset and offset. Also shown in Fig. 3 are formant trajectories for /ɔ/ and /ʊ/ because, although Table V shows no significant differences for these vowels in the two types of classification, a closer examination of the confusion matrices pointed to fewer confusions between these vowel pairs in the DCT analysis than in the target analysis. For all three vowel pairs shown in Fig. 3, the relative time at which the vowel target occurs is delayed for the long tense vowels /i,a,ɔ/ compared with the corresponding short vowels that are similar in quality. This suggests that the time of the target is different for these tense/lax pairs: consequently, the shape of the formant contours may be different even though the respective onset, target, and offset formant values for the tense/lax pairs are quite similar. Therefore, the time of the target provides some contributory information to the distinction between such vowels, which results in a more effective separation between them in the DCT analysis compared with the target analysis.

Further evidence to support this view is shown in Table

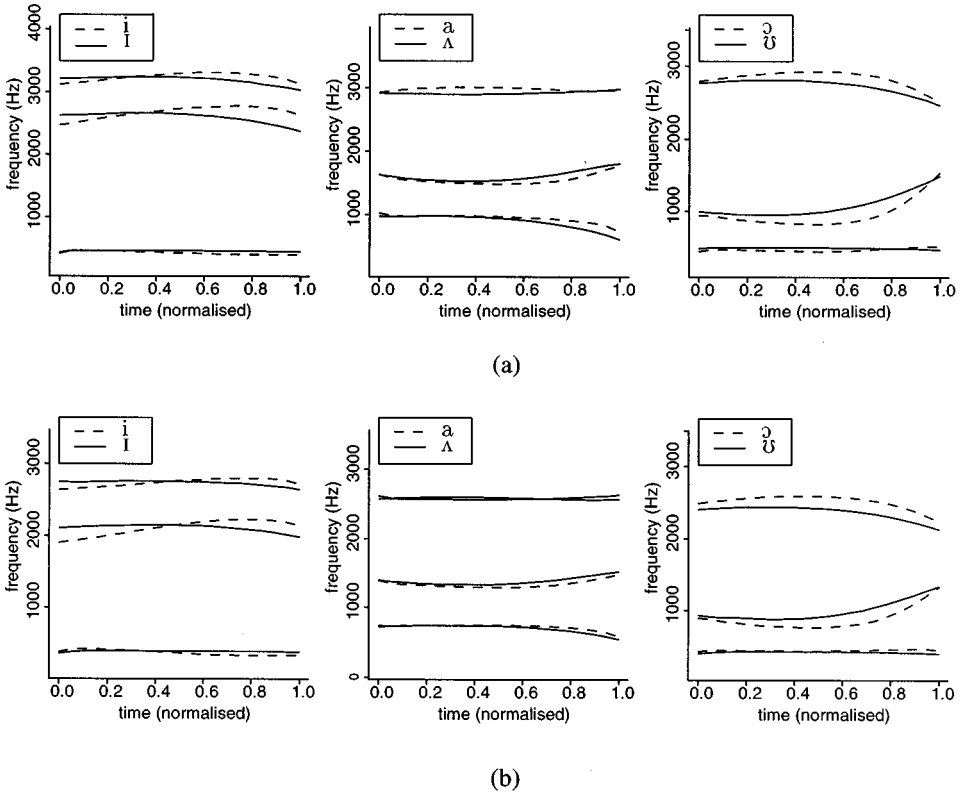

FIG. 3. Averaged time-normalized tracks of the first, second, and third formants for the tense/lax pairs /i,ɪ/, /a,ʌ/, and /ɔ,ʊ/ from (a) the female data, and (b) the male data.

TABLE VI. The mean and standard deviation for the target position (the ratio of the time between the vowel onset and the target, to the total vowel duration) for the tense/lax pairs from the female and male data.

	Female			Male	
	Mean	Standard deviation		Mean	Standard deviation
i	0.690	0.129	i	0.736	0.087
ɪ	0.357	0.107	ɪ	0.401	0.108
a	0.424	0.124	a	0.421	0.123
ʌ	0.343	0.088	ʌ	0.383	0.101
ɔ	0.484	0.087	ɔ	0.466	0.107
ʊ	0.312	0.077	ʊ	0.316	0.065

VI, which lists the mean and standard deviation of the time at which the target occurs relative to the acoustic vowel boundaries (therefore, a mean value of 0.5 would indicate the target occurs, on average, at the acoustic vowel midpoint). These data show that the lax vowels /ɪ,ʌ,ʊ/ had earlier target times (i.e., shorter onglides) than the tense vowels /i,a,ɔ/, respectively. These differences are significant for all the tense/lax pairs in the female data, and the /i,ɪ/ and /ɔ,ʊ/ tense/lax pairs in the male data ($\alpha \ll 0.003$, the significance level was Bonferroni-corrected).

II. SUMMARY

The experiments in this study of a large corpus of Australian English vowels produced by 132 male and female talkers have shown that vowels differ in the extent and nature of dynamic information which contributes to their identification. First, many vowels, which are traditionally labeled monophthongal were as adequately classified from a static section at the vowel target as from time-varying formant information that was represented by the discrete cosine transformation on the formant frequencies. Second, Australian English vowels that are traditionally labeled diphthongal were more accurately identified in the DCT space that encodes time-varying formant information than from a single static section at the target. As discussed in Harrington and Cassidy (1994), this is because these vowels are defined by at least two targets which are inadequately represented in a single-target space. Thirdly, some tense vowels traditionally labeled as monophthongs had proportionally delayed targets compared with those of their lax counterparts.

For some researchers, a dynamic theory of vowel identification and perception has been advocated which is presumed to be in opposition to a theory of classifying vowels from targets (a target classification is often assumed to be static). For example, this opposition between dynamic and static target is clearly stated by Hillenbrand *et al.* (1995), who comments that ''static spectral targets are neither necessary nor sufficient for accurate vowel recognition.'' Although we agree that time-varying information is often very important in vowel identification, we do not believe that this need imply that targets are irrelevant. As the results from our paper show, the target times of vowels can vary, and for tense/lax pairs this can be measured by different rates of spectral change. This variation in the target time is one of the ways in which vowels are dynamic. Another is that some vowels have two targets. And yet another is that there are durational differences between tense/lax vowels. These are all important sources of dynamic variation, but none of them by themselves, or together, implies that the concept of a target is not relevant to vowel identification (see, also, Harrington and Cassidy, 1994). In summary, although there is clear evidence for the presence of dynamic information that benefits vowel identification, the results of this study are entirely compatible with a target theory of vowel identification.

TABLE VII. The classification results for separating monophthongs and diphthongs using DCT analysis and vowel duration, and DCT analysis only. Significant differences are shown by ** ($\alpha < 0.01$).

	Female		
	DCT coef. & duration	DCT coef.	*p*
Monophthongs	98.3	98.3	1.0
Diphthongs	88.8	84.8	<0.01**
All	94.3	92.6	<0.01**
	Male		
	DCT coef. & duration	DCT coef.	*p*
Monophthongs	97.5	97.5	1.0
Diphthongs	87.1	83.7	<0.01**
All	93.2	91.8	<0.01**

This study has also shown that the first two coefficients of the discrete cosine transformation, which model the mean and the slope of the formant trajectory, not only effectively distinguish between the citation-form vowels produced by multiple talkers in a very similar consonantal context (mostly /hVd/ and /hV/), but also encode the dynamic information in diphthongs and in monophthong tense/lax pairs that differ in the relative timing of the vowel target. Another advantage in using a DCT analysis on the formants is that vowels can be effectively separated without the need to mark explicitly one or more vowel targets, which can be complicated when a vowel appears either not to have a steady-state section, or if the formants reach a minima or maxima at different times.

One final application for the DCT analysis is in monophthong/diphthong distinction. There was a good separation between the monophthongs and diphthongs in the DCT analysis in experiment 1 (in contrast to the target analysis). Exploiting this knowledge, we relabeled all the vowels as either monophthong or diphthong and repeated the DCT analysis as outlined in experiment 1. It can be seen from Table VII (the classification scores from DCT analysis including vowel duration) that most of the vowels were correctly identified; that is, either as monophthong or diphthong. We repeated the above analysis without including vowel duration as a parameter in order to be certain that the monophthong/diphthong distinction was primarily due to differences in the shapes of the formant contours. These results are also given in Table VII, where it can be seen that, although the total number of vowels and diphthongs correctly classified were significantly less when vowel duration was not included with the DCT analysis, all the classification rates were still very high.

Finally, it must be emphasized that, although the present corpus included all the phonemic vowels of Australian English from a large number of male and female talkers, these were all restricted to a very similar consonantal context (mostly /hVd/ and /hV/), and it is certain that other forms of dynamic vowel changes would be introduced by considering a wider range of contexts beyond those considered here. However, this need not imply that it is necessary to abandon the model that the salient information for vowel identification can be represented by one or two targets which may be variably timed relative to the vowel onset and offset. We envisage, for example, that the additional complexity introduced by variable coarticulatory influences could be modeled using a second-order differential equation that relates consonantal loci, rate of formant change, and position of the vowel target (e.g., Moon and Lindblom, 1994), together with a phase specification for timing the separate transitions towards, and away from, a vowel target, as suggested by much of the task-dynamic literature (e.g., Browman and Goldstein, 1992; Saltzman and Munhall, 1989). These issues, as well as the effectiveness of the DCT coefficients in separating vowels in multiple contexts in a similarly large acoustic speech corpus, are currently being investigated.

ACKNOWLEDGMENTS

The authors thank Steve Cassidy, Felicity Cox, Zoe Evans, William Thorpe, and in particular Sallyanne Palethorpe, James Hillenbrand, and Terry Nearey for their help and suggestions. This research was supported by an Australian Research Council large grant.

Andruski, J. E., and Nearey, T. M. (**1992**). ''On the sufficiency of compound target specification of isolated vowels in /bVb/ syllables,'' J. Acoust. Soc. Am. **91**, 390–410.

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. R. L. (**1967**). ''Some measurements of some sounds of Australian English.'' Unpublished doctoral dissertation, Sydney University.

Bernard, J. R. L. (**1970**). ''Towards the acoustic specification of Australian English,'' Zeitschrift fur Phonetik **2/3**, 113–128.

Bernard, J. R. L. (**1981**). ''Australian pronunciation,'' in *The Macquarie Dictionary*, edited by A. Delbridge (Macquarie Library, Sydney), pp. 18–27.

Bladon, R. A. W. (**1985**). ''Diphthongs: a case study of dynamic auditory processing,'' Speech Commun. **4**, 145–154.

Blair, D. (**1993**). ''Australian English and Australian national identity,'' in *The Languages of Australia* (Australian Academy of the Humanities, Canberra), Vol. 14, pp. 62–70.

Browman, C. P., and Goldstein, L. (**1992**). ''Articulatory phonology: an overview,'' Phonetica **49**, 155–180.

Cochrane, G. R. (**1989**). ''Origins and development of the Australian accent,'' in *Australian English: The Language of a New Society*, edited by D. Blair and P. Collins (University of Queensland Press, St. Lucia), pp. 176–186.

Cox, F. M. (**1996**). ''An acoustic study of vowel variation in Australian English.'' Unpublished doctoral dissertation, Macquarie University.

Cox, F. M. (**1998**). ''The Bernard data revisited,'' Aust. J. Ling. **18**(1), 29–55.

Croot, K., Fletcher, J., and Harrington, J. (**1992**). ''Levels of segmentation and labelling in the Australian national database of spoken language,'' in *Proceedings of the 4th International Conference on Speech Science and Technology*, edited by J. Pittam (Australian Speech Science & Technology Association, Brisbane), pp. 86–90.

Delattre, P., Liberman, A. M., Cooper, F. S., and Gerstman, F. J. (**1952**). ''An experimental study of the acoustic determinants of vowel color: observations on one- and two-formant vowels synthesised from spectrographic patterns,'' Word **8**, 195–210.

Essner, C. (**1947**). ''Recherche sur la structure des voyelles orales,'' Arch. Néerlandaises Phonétique Exp. **20**, 40–77.

Fant, G. (**1960**). *The Acoustic Theory of Speech Production* (Mouton, The Hague).

Fowler, C. A. (**1980**). ''Coarticulation and theories of extrinsic timing,'' J. Phonetics **8**, 113–133.

Fowler, C. A. (**1983**). ''Converging sources of evidence on spoken and perceived rhythms in speech: cyclic productions of vowels in monosyllabic stress feet,'' J. Exp. Psychol. **112**, 386–412.

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

Fowler, C. A. (**1987**). ''Perceivers as realists, talkers too: commentary on papers by Strange, Diehl *et al.*, and Rakerd and Verbrugge,'' J. Mem. Lang. **26**, 574–587.

Fowler, C. A., and Rosenblum, L. D. (**1989**). ''The perception of phonetic gestures,'' Haskins Lab. Status Rep. Speech Res. **99/100**, 102–117.

Fox, R. (**1983**). ''Perceptual structure of monophthongs and diphthongs in English,'' Lang. Speech **26**, 21–49.

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

Gay, T. (**1970**). ''A perceptual study of American English diphthongs,'' Lang. Speech **13**, 65–88.

Gottfried, M., Miller, J. D., and Meyer, D. J. (**1993**). ''Three approaches to the classification of American English diphthongs,'' J. Phonetics **21**, 205–229.

Harrington, J., and Cassidy, S. (**1994**). ''Dynamic and target theories of vowel classification: Evidence from monophthongs and diphthongs in Australian English,'' Lang. Speech **37**(4), 357–373.

Harrington, J., and Cassidy, S. (1999). *Techniques in Speech Acoustics* (Kluwer Academic, Dordrecht).

Harrington, J., Cox, F., and Evans, Z. (**1997**). ''An acoustic analysis of cultivated, general, and broad Australian English speech,'' Aust. J. Ling. **17**, 155–184.

Harrington, J., Fletcher, J., and Beckman, M. E. (in press). ''Manner and place conflicts in the articulation of accent in Australian English,'' in *Papers in Laboratory Phonology 5*, edited by M. Broe (Cambridge University Press, Cambridge).

Hillenbrand, J., and Gayvert, R. T. (**1993**). ''Vowel classification based on fundamental frequency and formant frequencies,'' J. Speech Hear. Res. **36**, 647–700.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). ''Acoustic characteristics of American English vowels,'' J. Acoust. Soc. Am. **5**, 3099–3111.

Holbrook, A., and Fairbanks, G. (**1962**). ''Diphthong formants and their movements,'' J. Speech Hear. Res. **5**, 38–58.

Horvath, B. M. (**1985**). *Variation in Australian English: The Sociolects of Sydney* (Cambridge University Press, Cambridge).

Huang, C. B. (**1986**). ''The effect of formant trajectory and spectral shape on the tense/lax distinction in American vowels,'' in *Proceedings IEEE International Conference on Acoustics, Speech and Signal Processing* (Institute of Electrical & Electronic Engineers, Tokyo), pp. 893–896.

Huang, C. B. (**1992**). ''Modelling human vowel identification using aspects of format trajectory and context,'' in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (IOS, Amsterdam), pp. 43–61.

Jenkins, J., Strange, W., and Edman, T. (**1983**). ''Identification of vowels in 'vowelless' syllables,'' Percept. Psychophys. **34**, 441–450.

Jenkins, J. J., Strange, W., and Miranda, S. (**1994**). ''Vowel identification in mixed-speaker silent-center syllables,'' J. Acoust. Soc. Am. **95**, 1030–1043.

Joos, M. (**1948**). ''Acoustic phonetics,'' Language **24**, 1–136.

Klein, W., Plomp, R., and Pols, L. C. W. (**1970**). ''Vowel spectra, vowel spaces, and vowel identification,'' J. Acoust. Soc. Am. **48**, 999–1009.

Ladefoged, P. (**1967**). *Three Areas of Experimental Phonetics* (Oxford University Press, Oxford).

Lehiste, I., and Peterson, G. (**1961**). ''Some basic considerations in the analysis of intonation,'' J. Acoust. Soc. Am. **33**, 419–425.

Lindblom, B. (**1963**). ''Spectrographic study of vowel reduction,'' J. Acoust. Soc. Am. **35**, 1773–1781.

Millar, J., Harrington, J., and Vonwiller, J. (**1997**). ''Spoken language data

resources for Australian speech technology,'' J. Electr. Electron. Eng., Aust.
Millar, J., Vonwiller, J., Harrington, J., and Dermody, P. (**1994**). ''The Australian National Database of Spoken Language,'' in *Proceedings of the International Conference on Acoustics Speech and Signal Processing* (Institute of Electrical & Electronic Engineers, Adelaide), Vol. 2, pp. 67–100.
Moon, S.-J., and Lindblom, B. (**1994**). ''Interaction between duration, context, and speaking style in English stressed vowels,'' J. Acoust. Soc. Am. **96**, 40–55.
Nearey, T. M. (**1989**). ''Static, dynamic, and relational properties in vowel perception,'' J. Acoust. Soc. Am. **85**, 2088–2113.
Nearey, T. M., and Assmann, P. F. (**1986**). ''Modelling the role of inherent spectral change in vowel identification,'' J. Acoust. Soc. Am. **80**, 1297–1308.
Parker, E. M., and Diehl, R. L. (**1984**). ''Identifying vowels in CVC syllables: effects of inserting silence and noise,'' Percept. Psychophys. **36**, 369–380.
Peterson, G., and Barney, H. (**1952**). ''Control methods used in a study of vowels,'' J. Acoust. Soc. Am. **24**, 175–184.
Rakerd, B., and Verbrugge, R. R. (**1987**). ''Evidence that the dynamic information for vowels is talker independent in form,'' J. Mem. Lang. **26**, 558–563.
Rao, K. R., and Yip, P. (**1990**). *Discrete Cosine Transform: Algorithms, Advantages, Applications* (Academic, New York).
Saltzman, E., and Munhall, K. (**1989**). ''A dynamical approach to gestural patterning in speech production,'' Ecological Psychol. **1**, 333–382.
Stevens, K. N., Kasowski, S., and Fant, G. (**1953**). ''An electrical analog of the vocal tract,'' J. Acoust. Soc. Am. **25**, 734–742.
Strange, W. (**1987**). ''Information for vowels in formant transitions,'' J. Mem. Lang. **26**, 550–557.
Strange, W. (**1989a**). ''Dynamic specification of coarticulated vowels spoken in sentence context,'' J. Acoust. Soc. Am. **85**, 2135–2153.
Strange, W. (**1989b**). ''Evolving theories of vowel perception,'' J. Acoust. Soc. Am. **85**, 2081–2087.
Strange, W., Jenkins, J. J., and Johnson, T. L. (**1983**). ''Dynamic specification of coarticulated vowels,'' J. Acoust. Soc. Am. **74**, 695–705.
Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). ''Consonant environment specifies vowel identity,'' J. Acoust. Soc. Am. **60**, 213–224.
Verbrugge, R. R., and Rakerd, B. (**1986**). ''Evidence of talker-independent information for vowels,'' Lang. Speech **29**, 39–55.
Vonwiller, J., Rogers, I., Cleirigh, C., and Lewis, W. (**1995**). ''Speaker and material selection for the Australian National Database of Spoken Language,'' J. Quant. Ling. **3**, 177–211.
Watson, C., Harrington, J., and Evans, Z. (**1998**). ''An acoustic comparison between New Zealand and Australian English vowels,'' Aust. J. Ling. **18**(2), 185–207.
Wells, J. C. (**1982**). *Accents of English: Beyond the British Isles* (Cambridge University Press, Cambridge).
Zahorian, S., and Jagharghi, A. (**1993**). ''Spectral-shape features versus formants as acoustic correlates for vowels,'' J. Acoust. Soc. Am. **94**, 1966–1982.