Spectral-shape features versus formants as acoustic correlates for vowels

Stephen A. Zahorian and Amir Jalali Jagharghi[a)]
Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, Virginia 23529

(Received 27 September 1991; revised 2 December 1992; accepted 14 June 1993)

The first three formants, i.e., the first three spectral prominences of the short-time magnitude spectra, have been the most commonly used acoustic cues for vowels ever since the work of Peterson and Barney [J. Acoust. Soc. Am. **24**, 175–184 (1952)]. However, spectral shape features, which encode the global smoothed spectrum, provide a more complete spectral description, and therefore might be even better acoustic correlates for vowels. In this study automatic vowel classification experiments were used to compare formants and spectral-shape features for monopthongal vowels spoken in the context of isolated CVC words, under a variety of conditions. The roles of static and time-varying information for vowel discrimination were also compared. Spectral shape was encoded using the coefficients in a cosine expansion of the nonlinearly scaled magnitude spectrum. Under almost all conditions investigated, in the absence of fundamental frequency ($F0$) information, automatic vowel classification based on spectral-shape features was superior to that based on formants. If $F0$ was used as an additional feature, vowel classification based on spectral shape features was still superior to that based on formants, but the differences between the two feature sets were reduced. It was also found that the error pattern of perceptual confusions was more closely correlated with errors in automatic classification obtained from spectral-shape features than with classification errors from formants. Therefore it is concluded that spectral-shape features are a more complete set of acoustic correlates for vowel identity than are formants. In comparing static and time-varying features, static features were the most important for vowel discrimination, but feature trajectories were valuable secondary sources of information.

PACS numbers: 43.72.Ar, 43.70.Fq, 43.71.Es

INTRODUCTION

The search for acoustically invariant cues to phones remains one of the most challenging and fundamental problems in speech science (Pisoni, 1985). Researchers ranging from speech scientists to automatic speech recognition engineers have toiled long and diligently to define acoustic cues that are important to perception and that also improve the performance of automatic speech recognition systems. For the case of vowels, there is, according to "simple target" theory (Strange, 1989a), a unique vocal tract configuration ("target") for the production of each vowel. Therefore, according to the theory, the primary acoustic correlates necessary to distinguish vowels from one another can be extracted from the static spectral characteristics of vowels near the central regions of their acoustic waveforms.

Ever since the classic paper by Peterson and Barney (1952), the first three formants, i.e., the first three spectral prominences, have been regarded as the primary source of this spectral information. The first two formants, in particular, are considered to be the most important perceptually (Fant, 1960), with formant three playing a supporting role. Peterson and Barney plotted vowels in a formant-one/formant-two space and showed that, to a large degree, phonologically similar vowels (as judged by listeners) cluster in this space while phonologically dissimilar vowels are more separated. Over the past 40 years, numerous papers have been presented and published that examine various aspects of formant representations of vowels. In many studies (e.g., Syrdal and Gopal, 1986; Miller; 1989; Nearey, 1989), models of vowel perception have been developed using formants as the primary acoustic correlates of vowel perception. These studies show that vowel position in a formant space is highly correlated with vowel perception. Despite this widespread use of formants in representing speech spectra, particularly for vowels, it is not clear that the formants play a fundamental role in speech perception. On the one hand, in experiments with the auditory nerves of cats (Sachs and Young, 1979; Young and Sachs, 1979; Delgutte, 1984; and Delgutte and Kiang, 1984), a possible mechanism was shown for accurately encoding the formants of steady-state vowels at the auditory nerve level. Nevertheless, inconsistencies remain. For example, in vowel identification tasks using vowels synthesized from the Peterson and Barney formant data, human listeners typically achieved considerably lower identification rates than in listening tests with the original speech stimuli (Hillenbrand and McMahon, 1987). It is also possible to synthesize vowels with identical values for $F1$ and $F2'$ (thought to be a better perceptual indicator than $F2$ alone) but which are identified as different vowels (Bladon, 1983). Vowel perception for fixed $F1$ and $F2$ values

[a)]Present address: ViGyan, Inc., 30 Research Drive, Hampton, VA 23666.

 0001-4966/93/94(4)/1966/17/$6.00

also depends on *F0*, fundamental frequency of voicing; therefore some investigators have modified their formant models to include the *F0* interaction. Besides perceptual studies, automatic recognition of vowels based on formants with sophisticated pattern recognition schemes is never quite as accurate as recognition rates obtained by human listeners (e.g., Hillenbrand and Gayvert, 1987).

Bladon (1982) made several arguments against a formant representation of speech and favored a representation based on gross spectral shape. First, he made the argument that changes in formant frequencies, or any mechanisms that presumably enhance spectral peaks, also change the spectral shape. The primary objections, however, to a formant-based representation of speech were based on "*reduction, determinacy,* and *perceptual adequacy.*" The reduction objection stems from the observation that a formant representation is an incomplete spectral description. The determinacy objection stems from the great difficulty in determining the locations of formants for many conditions. The perceptual adequacy objection was raised on the grounds that perceptual distance for vowels with widely spaced spectral peaks are poorly predicted by a formant representation of the spectra but can be predicted by a spectral-shape model.

Global spectral shape features have been investigated as acoustic correlates for vowels. For example, Pols and his colleagues (Plomp *et al.*, 1967; Pols *et al.*, 1969; Klein *et al.*, 1970) completed an extensive series of experiments using a principal-components spectral-shape representation of vowel spectra. They demonstrated that a plot of vowel data in a rotated principal-components-one versus principal-components-two parameter space resembles the Peterson–Barney vowel data plotted in a formant space. This group also determined that vowels could be automatically classified as accurately from a principal-components representation as from a formant representation.

Besides spectral-shape representations based on low-ordered terms in a series basis vector expansion, several investigators examined two-formant models of vowel spectra (e.g., Carlson *et al.*, 1975; Chistovich *et al.*, 1979; Bladon, 1983; Beddor and Hawkins, 1984; and Chistovich, 1985, for a review and tutorial of work in this area). In the typical experimental paradigm, subjects matched a many-formant stimulus with a two-formant stimulus, usually by adjusting the second formant—or matched a given two-formant stimulus with a single-formant stimulus. The results of these experiments often suggested a match to the "center of gravity" of the target vowel formants, thus implying an averaging over frequency. This spectral integration takes place only when the distance between formants is less than a critical distance of 3–3.5 bark on the frequency scale. Chistovich (1985) also showed that for closely spaced formants, both formant amplitudes and formant frequencies affect perceived vowel quality in a manner consistent with a spectral averaging process. For widely spaced formants, spectral averaging does not appear to take place. The spectral averaging over small frequency ranges and apparent lack of spectral averaging over the entire spectrum is consistent with the hypothesis that overall spectral shape is crucial to vowel identity, provided spectral shape is characterized appropriately.

Another important issue in the search for acoustically invariant cues to vowel perception is the relative importance of static versus temporal cues. There is ample evidence showing that the static spectral properties of vowels are not always sufficient cues for perception, and that some time-varying information contained in the interval surrounding the vowel "center" is also required. For example, Fairbanks and Grubb (1961) examined the static spectral characteristics of vowels by presenting nine isolated vowels produced by phonetically trained speakers to experienced listeners. The overall identification rate was 74%, significantly lower than the 94.4% rate obtained by Peterson and Barney (1952), where vowels were presented in an /hVd/ context and, thus, presumably contained more temporal information.

Stevens and House (1963) have also shown that the acoustic properties of phonologically equivalent vowels vary greatly because of coarticulation with adjacent phonemes. In continuous speech, "target" states are seldom reached (Lindblom, 1963; Stevens and House, 1963). This "target undershoot" or "vowel reduction" problem can be compensated for if listeners make use of the direction and rate of change of formants to identify vowels (Lindblom and Studdert-Kennedy, 1967). The effects of "target undershoot" can also be compensated for by pattern recognition procedures (Broad, 1976; Kuwabara, 1985). Many recent perceptual studies (Strange *et al.*, 1976; Gottfried and Strange, 1980; Strange *et al.*, 1983; Williams, 1986; Di Benedetto, 1989a,b; Strange, 1989b) support the hypothesis that both static cues and "gestures," representing the time history of spectra, are essential for reliable perception of vowels.

Returning to the discussion of the main thesis of this paper, despite at least some evidence to the contrary, among speech scientists "steady-state" formants remain the acoustic features of choice for vowels. In this study, we used sophisticated signal processing techniques combined with automatic pattern classification to investigate in detail two sets of spectral features, global spectral-shape parameters versus spectral peaks, as acoustic cues for automatic classification of vowels. We hypothesized that overall global spectral shape provides a more complete spectral description than do three formants and therefore classification based on spectral-shape features should be superior to that based on three formants. The formants are important and efficient acoustic cues because they constrain global spectral shape to a large degree. We also examined the role of static and gestural acoustic features in classifying vowels and examined the extent of contextual influence on acoustic features. Finally, we compared error patterns in the perception of vowels by human listeners with error patterns resulting from automatic classification experiments for the two feature sets.

In this study, for comparison with formants, the discrete cosine transform coefficients (DCTCs) of the nonlinearly scaled spectrum were used as the spectral-shape parameters. Except for small differences, the DCTCs were

computed the same as are the cepstral coefficients commonly used in speech processing. Static features, which represent the vowel for one instant in time, were computed from the middle of the quasi-steady-state portion of each vowel segment. Gestural features, which represent the trajectory of the vowel spectra over a short segment, were computed as the coefficients in a series basis vector expansion over the selected segment. Various segments of the speech signal were investigated for relevant vowel information. Many automatic classification experiments were conducted with a large database to refine the features under investigation for a wide range of conditions.

The present study is a second paper devoted primarily to comparing formants and global spectral shape features for speech recognition, following a paper investigating initial stop consonants (Nossair and Zahorian, 1991). The format of this paper is the same as that of the previous paper to facilitate more convenient comparisons. The database, basic signal processing techniques, and general objectives are the same for both papers. The primary difference between the two papers is, of course, that the present paper focuses on vowels, rather than initial stops. Additionally, however, all the signal processing and classification techniques were refined to enhance the feature computations and to improve the statistical reliability of results. Another related paper focused on speaker-normalization issues for the case of vowels, but again using the same database and same underlying signal processing methods, is Zahorian and Jagharghi (1991).

I. DATABASE

The database for the experiments was obtained by recording 99 CVC syllables produced in isolation by each of 30 speakers. This database is also described in both Nossair and Zahorian (1991) and Zahorian and Jagharghi (1991). Summarizing briefly, ten of the speakers were men (M), ten were women (W), and ten were children (C) between the ages of 7 and 11 (five male, five female). The distribution of dialect regions for these speakers was Southern (15), Mid Atlantic (7), Northern (6), and New England (2). However, many of the speakers had moved extensively during childhood and adolescence, thus making it difficult to label the speech of these individuals with regional dialects. The Southern speakers were primarily from urban areas of Virginia. The CVC syllable list contained approximately 9 instances of each of the 11 vowels /iy,ih,eh,ae,ah,aa,ao,ow,uh,uw,er/.[1] The initial consonant was one of /b,d,g,p,t,k,hh,l,w/. The final consonant was one of /b,d,g,p,t,k,v,s/. The speech signals were low-pass filtered at 7.5 kHz and sampled at 16 kHz with a 12-bit analog-to-digital converter. The total number of usable vowel stimuli was 2922 (out of 99*30 = 2970).

In comparison with the Peterson and Barney vowel database (1952), our database is larger (2970 tokens versus 1520) but is derived from fewer speakers (30 vs 76) from a more restricted geographical region. The primary difference, however, is the much greater consonantal context, with every vowel paired with at least one instance of /b,d,g,p,t,k/ in both initial and final position, as opposed to the fixed /hVd/ context for the Peterson and Barney vowel data. Thus, our database also enables an investigation of context effects. In any event, the Peterson and Barney vowel data was not suitable for the present experiments since only the formant data, rather than the actual speech signals, have been preserved.

The acoustic regions of all speech files were manually labeled through visual and auditory inspection of the waveform, with the aid of an interactive computer waveform editor. In addition to the acoustic speech waveform, the spectral derivative (Furui, 1986) was also displayed to help define the boundaries between the acoustic segments. All segmentation points were selected to coincide with voicing-pulse zero crossings in the speech waveform. The acoustic regions most relevant for vowels are the following: (a) initial transition (IT), a periodic waveform that begins at the first voicing pulse and ends at the start of the steady-state vowel; (b) steady-state vowel (SV), a section of quasiperiodic steady-state waveform—the SV region was defined as that portion of the vowel with a high almost constant amplitude and a low value of the spectral derivative;[2] and (c) final transition (FT), the voiced transition region from the end of the SV region to the beginning of the final burst.

II. SPEECH PARAMETERS AND CLASSIFICATION METHODS

A. Speech parameters

In this section we explain the signal processing techniques used in computing the two feature sets investigated in this study. These two feature sets are formants and discrete cosine transform coefficients (DCTCs). Fundamental frequency $F0$ was also used as an additional feature for some experiments. Although we have used the same features in previously reported studies, we have refined the signal processing for feature extraction, as described below.

1. Formants

Formants were computed for the vowels in a multistage process as follows. The speech signal was first digitally low-pass filtered at 3.8 kHz with a 49th-order FIR linear-phase low-pass filter and resampled at 8 kHz. The speech signal was then high-frequency preemphasized with transfer function ($1-0.75\ z^{-1}$). The signal was windowed with a 50-ms Hanning window and a tenth-order LP model was computed. The roots of the LP polynomial were computed to determine up to five formant candidates (frequency, amplitude, and bandwidth) for each frame.[3] Formant candidates were obtained for each frame over the interval of interest. The frame spacing was variable, depending on the token, since a fixed total number of frames was processed for each interval. However, the spacing was typically between 2 and 6 ms.

The "actual" formants were selected from the formant candidates using a dynamic programming approach as described by Talkin (1987). This approach can be summarized as follows. Dynamic programming (Sakoe and Chiba, 1978) selects the lowest-cost[4] path among the set of

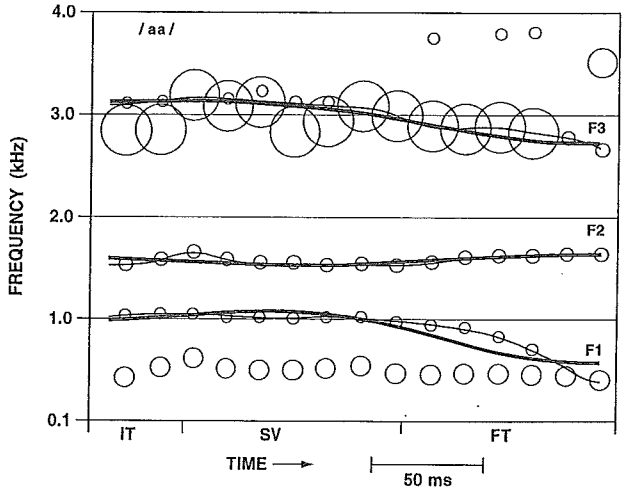

FIG. 1. The LP poles, original formant trajectories (light lines), and smoothed formant trajectories (heavy lines) for $F1$, $F2$, and $F3$ for /aa/ from a female speaker (extracted from the syllable "TOG") for the IT–FT interval. Each circle is centered at an LP pole location with radius proportional to the pole bandwidth. The formant trajectories were smoothed with a three-term cosine expansion over time.

formant candidates over the tracking interval. The cost of each path is the sum of "local" costs and "transition" costs encountered for that path. Local costs depend on the deviation of formant candidates from expected formant values and the bandwidth of the candidates. Transition costs model the constraint that formants generally change smoothly from frame to frame. Formants are also constrained such that $F3 > F2 > F1$. Dynamic programming is then used to find the minimum cost path through the candidates. The equations used to compute the local costs and transition costs, along with empirically determined constants, are given in the Appendix.[5]

The formant seed values were chosen independently for each vowel and for each speaker type (M, W, C) as follows. Initial seed values were obtained from the published Peterson and Barney data. These values were then used to track formants, and the average values were recomputed for each vowel and for each speaker type, and used as the updated seed values. The entire process was repeated several times (approximately five) until the average values no longer changed. Thus, the algorithm was not fully automatic, since both the vowel and speaker type were inputs to the procedure.

The performance of the formant tracking routine was verified by visual inspection of the computed formant trajectories for a large percentage of the stimuli. Figure 1 depicts formant candidates and formant trajectories (light lines) as computed by the tracking algorithm for a typical token. The heavier lines are smoothed formant trajectories, used for classification experiments, as described in a later section of this paper. The average formant frequencies are given in Table I for each vowel and each speaker category. These values were computed over the center frame of the SV interval.

2. DCT coefficients

Global spectral shape was represented as the coefficients in a discrete cosine transform expansion of a selected frequency range of the magnitude spectra (chosen as a portion of the original 0- to 8- kHz range), after nonlinear amplitude and frequency scaling. Consistent with our previous work, we refer to these coefficients as DCTCs rather than cepstral coefficients to emphasize their interpretation as the coefficients in a cosine expansion of the magnitude spectrum. With our notation, DCTC1, the coefficient of the constant term, is a measure of the average level of the spectrum; DCTC2, the coefficient of a half-cycle of a cosine, is a measure of the spectral tilt; DCTC3, the coefficient of a full cycle of a cosine, is a measure of spectral compactness; higher-ordered DCTCs provide additional spectral resolution. Also note that a smoothed spectrum can be computed from the DCTCs, with the degree of smoothing dependent on the number of DCTCs used to reconstruct the spectrum.

Figure 2 depicts examples of FFT spectra, tenth-order LP spectra, and spectra reconstructed from 11 DCTCs, for the vowels /iy/ and /aa/ from a male speaker. Note that although both the DCTCs and the LP model smooth the high-resolution spectra, the LP spectra provides more resolution for spectral peaks whereas the DCTC smoothing models peaks and valleys with equal resolution. Another difference is that the DCTC spectra provide more resolution at low frequencies than at high frequencies as a result

TABLE I. Average formant frequencies (in hertz) for 11 vowels for each of three speaker categories.

	F1			*F2*			*F3*		
Vowel	M	W	C	M	W	C	M	W	C
/iy/	272	338	313	2209	2837	2705	2971	3456	3517
/ih/	410	486	564	1859	2284	2615	2600	3093	3521
/eh/	550	745	875	1740	2123	2436	2535	3041	3526
/ae/	656	922	1116	1748	2089	2345	2483	2981	3409
/ah/	596	793	862	1289	1599	1627	2400	2872	3470
/aa/	749	981	1125	1192	1440	1590	2501	2847	3440
/ao/	637	822	906	1004	1176	1327	2557	2860	3432
/ow/	456	532	660	1176	1419	1645	2307	2789	3320
/uh/	439	528	573	1234	1437	1558	2349	2848	3400
/uw/	324	400	400	1396	1617	1806	2352	2766	3267
/er/	445	542	614	1286	1532	1596	1656	1992	2140

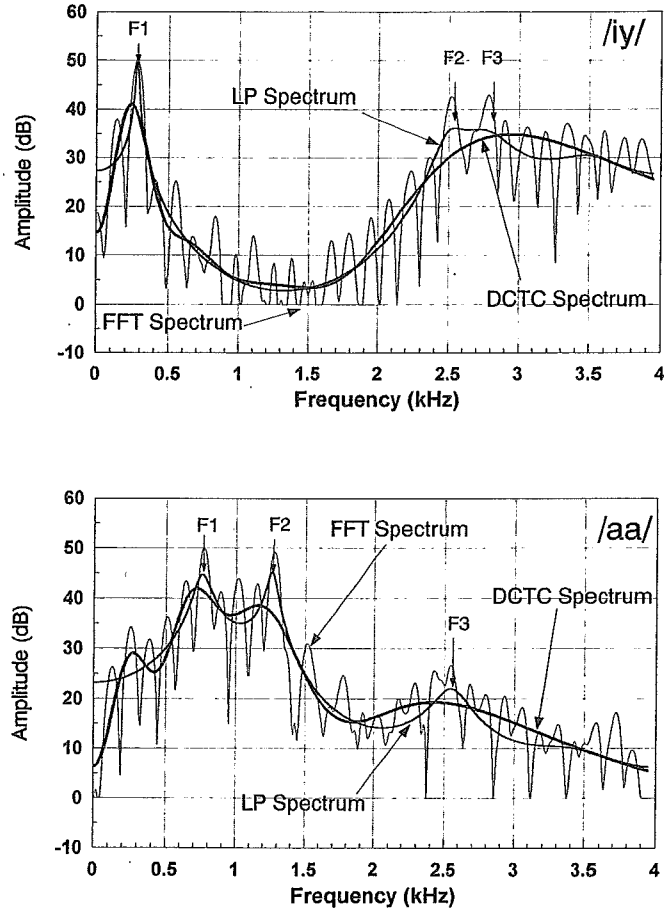

FIG. 2. FFT, DCTC, and LP spectral plots for /iy/ and /aa/ from a male speaker.

of bark warping, as opposed to uniform frequency resolution for the LP case. Further discussion of the spectral plots depicted in Fig. 2 is deferred until the results of automatic classification experiments are presented.

The DCTCs were computed using the same processing described in Nossair and Zahorian (1991) except for slight differences as follows. The signal was windowed with a 25-ms Hamming window. No high-frequency preemphasis was used.[6] The frequency scale was warped using bark frequency scaling (Zwicker, 1961; Syrdal and Gopal, 1986) rather than bilinear frequency warping (Oppenheim and Johnson, 1972) as used in previous studies. Note, however, that bilinear warping is very similar to bark warping if the bilinear warping coefficient is between 0.5 and 0.6. The log amplitude scaling of the magnitude spectrum was modified slightly such that the scaling is logarithmic for the uppermost 50 dB of the range for each frame with hard limiting (a "floor") at -50 dB relative to the peak spectral value for each frame. This method, determined empirically,[7] eliminates large negative spikes in the log that would occur for very low-energy spectral valleys. Figure 3 shows DCTCs 3–6 (light lines) for the same token and same segment as was used to depict the formant tracks in Fig. 1. Note that both the formants and DCTCs are relatively constant in the SV segment and vary more in the IT and FT segments.

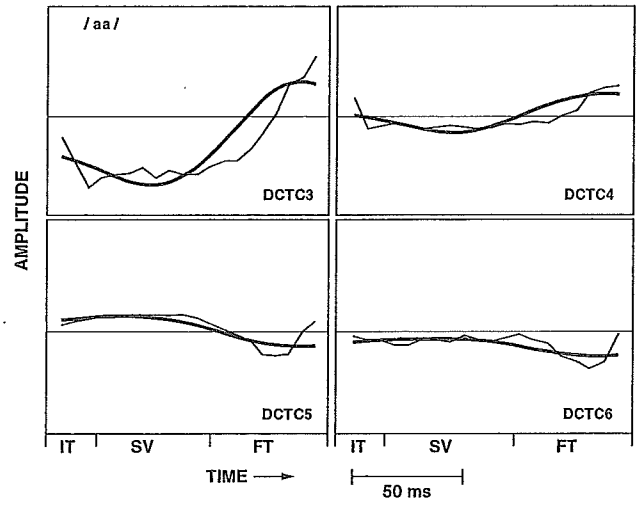

FIG. 3. DCTC trajectories (light lines) and smoothed trajectories (heavy lines) for DCTCs 3–6 for /aa/ from a female speaker (extracted from TOG) for the IT–FT interval. The DCTC trajectories were smoothed with a three-term cosine expansion over time.

3. Fundamental frequency

Fundamental frequency ($F0$) was computed using a form of the SIFT fundamental frequency algorithm (Markel, 1972). That is, the LP residual was computed for a window of speech (50 ms for males, 40 ms for females and children) in the steady-state portion of each vowel with a 12th-order LP inverse filter. The $F0$ values were computed from peaks in the autocorrelation of the residual after low-pass filtering at 1 kHz. Then, $F0$ was smoothed with median smoothing over a seven-frame window. The details of the signal processing for the $F0$ extraction, including the LP window lengths, were developed and investigated in previous studies (Zahorian and Gordy, 1983; Effer, 1985). For the static case, $F0$ was computed for 15 frames over the SV segment, and the resultant $F0$ for the center frame was used for experiments. For the case of time-varying features, $F0$ values were computed for each frame of the vowel.

B. Features for time-varying spectra

Speech features were also computed for each of several speech frames, to evaluate automatic recognition accuracy for the case of time-varying spectra. Several methods were investigated for sampling the spectra and for combining the parameters of several frames. The best approach found, insofar as automatic vowel classification is concerned, was to sample the speech spectra with frames equally spaced over the desired interval. The value of each parameter for each frame (i.e., a vector with a length equal to the number of frames) was then expanded using a discrete cosine series (DCS) expansion. That is,

$$P(n)=\sum_{k=1}^{N} C_k \cos\frac{(k-1)\pi(n-0.5)}{L}, \qquad (1)$$

where $P(n)$, $1\leqslant n\leqslant L$, is the parameter value for frame n, L is the total number of frames, N is the number of cosine

coefficients used to encode P, and the C_k are the cosine coefficients. Thus the coefficients C_k, $1 \leqslant k \leqslant N$, in Eq. (1) encode the smoothed trajectory of a speech parameter. C_1 is the average value of a parameter, C_2 is a measure of the tilt over time of a parameter, and higher-number terms encode additional details of a parameter trajectory. The C_k, which we call DCS coefficients, were then used as the features for classification. To illustrate the effect of this smoothing, Fig. 1 shows the smoothed trajectories for formants (heavy line, obtained with a three-term DCS expansion) as well as the original trajectories. Figure 3 also depicts both the original and smoothed DCTC trajectories.

C. Classifiers

All feature sets for vowels were evaluated using automatic classifiers as described in this section. The primary classifier used for experiments was a Bayesian maximum likelihood classifier (BML). That is, each stimulus was classified according to the category for which the distance

$$D_i(\mathbf{x}) = (\mathbf{x}-\mathbf{x}_i)^T \mathbf{R}_i^{-1} (\mathbf{x}-\mathbf{x}_i) + \ln|\mathbf{R}_i| - 2 \ln P(G_i), \quad 1 \leqslant i \leqslant M, \tag{2}$$

is minimized. In Eq. (2), $\mathbf{x}$ is the feature vector, $\mathbf{x}_i$ is the centroid for category G_i, $\mathbf{R}_i$ is the covariance matrix for category G_i, and $P(G_i)$ is the *a priori* probability for category G_i. Thus each category is characterized according to the centroid of all the training data in that category and the covariance matrix of the training data for that category. This classifier is optimum if the feature vector components are multivariate Gaussian (Duda and Hart, 1973). One variation of this method is to compute a single pooled covariance matrix over all categories. For this case the distance measure is referred to as Mahalanobis (MAH). Another variation is to assume that the covariance for each category is an identity matrix and to assume that the *a priori* probabilities are equal for all categories. For this case the distance measure is Euclidean distance (EUC).

Besides the traditional classifier described above, we also used an artificial neural network (ANN) for classification. The ANN used was fully interconnected with one hidden layer and sigmoid nonlinearities (Lippmann, 1987). The ANNs were trained with back propagation to recognize vowels using the features under investigation. Pilot experiments were used to determine the "best" number of hidden nodes. Based on the pilot testing 35 hidden nodes were selected for the reported experiments.[8]

In all the automatic classification experiments reported in this paper (except for a few special cases noted), the speakers used for training the classifier were different from those used for testing the classifier. To maximize the amount of training data (a fundamental problem with statistical classifiers) and to insure that all speakers were weighted equally, the following "round-robin" procedures were used. Specifically, for the BML, MAH, and EUC classifiers, 29 speakers were used for training and 1 for testing; the overall procedure was then repeated 30 times and test results averaged. For the case of the ANN classifiers, the database was partitioned into 24 speakers for training, 3 for evaluation, and 3 for testing. The networks were trained on the training data, until performance was maximized on the evaluation data. At this level of training, performance was tested on the remaining three speakers. This process was repeated ten times and test results averaged. This data management method was chosen for the ANN classifier to help insure the optimal amount of network training. Thus all comparisons of feature sets are derived from *speaker-independent* automatic recognition experiments.

III. EXPERIMENTS

A. Listening experiment

Besides the automatic classification experiments, we also conducted a listening experiment. The objectives of the experiment, conducted with natural speech, were to (1) evaluate our database; (2) obtain an estimate of the relative importance of various acoustic segments in supplying vowel information to human listeners; and (3) use the identification rates obtained by human listeners as a control for the results obtained by automatic classification experiments.

This experiment was conducted using the same methods as described in Nossair and Zahorian (1991). Summarizing briefly, the experiment was conducted with the data from 9 of the 30 talkers—3 adult males, 3 adult females, and 3 children. These speakers were chosen based on performance from an automatic vowel classification experiment, with the goal of choosing a representative set of speakers. Within each group of three, one speaker was chosen with relatively high automatic vowel classification rates, one with low automatic vowel classification rates, and one with average vowel classification rates. The dialect regions of these speakers were Southern (5), Mid Atlantic (3), and New England (1). Five paid normally hearing female students at Old Dominion University served as subjects. Each subject attended an initial half-hour training session in which the experimental procedure was explained and 40 CVC syllables from a female talker outside the testing set were presented. Feedback was given to subjects during the training session only. For each experimental condition, the listeners could listen to each stimulus as many times as they desired but had to make a forced choice response among the 11 vowels. The five listening conditions were (i) the entire CVC syllable (CVC), (ii) the beginning of the initial transition up to the end of the final transition (IT–FT), (iii) the beginning of the burst through the end of the steady-vowel segment (IB–SV),

TABLE II. Average vowel identification rate for each listening condition.

Condition	Identification rate (%)
CVC	91.3
IT–FT	88.7
IB–SV	91.0
SV–END	88.9
SV	85.4

TABLE III. Confusion matrix for the CVC listening condition.

	/iy/	/ih/	/eh/	/ae/	/ah/	/aa/	/ao/	/ow/	/uh/	/uw/	/er/
/iy/	100.0										
/ih/		99.1	0.9								
/eh/		1.5	96.2	2.3							
/ae/			3.3	82.3		14.5					
/ah/				0.3	93.4	0.3	0.8		4.8	0.5	
/aa/					5.2	73.6	19.5	1.6			
/ao/					0.9	15.8	82.9	0.2		0.2	
/ow/								99.6		0.4	
/uh/					11.5			0.7	83.3	4.4	
/uw/					0.7			0.7	4.5	94.0	
/er/					0.3				0.3		99.4

(iv) the beginning of the steady vowel through the end of the token (SV–END), and (v) the steady-vowel segment (SV).

Each subject completed these five conditions, in the order listed, in approximately five 1-hr sessions over a 2-wk period. Tokens were blocked by talker, with the order of tokens within each block and the order of blocks separately randomized for each listener. The conditions were arranged in order of expected difficulty to maximize the experience of the listeners before testing with the more difficult stimuli. Table II gives the average percent identification, averaged over five listeners and nine talkers, for the 11 vowels for each of the five listening conditions. As Table II shows, the average percent correct for the 11 vowels, based on listening to the entire syllable, was 91.3%, thus showing that there was significant ambiguity in vowel identity for human listeners, even if the entire word was available. Comparison of the other conditions in Table II shows that IB–SV is nearly equivalent to listening to the entire CVC, the IT–FT and SV–END conditions result in nearly identical average results, which are lower than for the best conditions, and there is a further drop in performance for the SV case.

Confusion matrices are given for the CVC, IT–FT, and SV conditions in Tables III–V, respectively. The confusion matrices for the IB–SV and SV–END conditions are omitted since the confusion patterns are similar to those for the CVC and IT–FT conditions, respectively. Each row of each matrix represents the vowel intended by the speaker and the columns are the responses of the listeners to each spoken vowel. The vowels /iy,ih,uw,er,ah/ were easily identified for all conditions. Most confusions were for vowels that are separated only one step in vowel height (/ih,eh/, /eh,ae/, /uh,uw/, /ow,uh/) or one step in the frontness (/uh,ah/, /er,ow/, /aa,ao/, /ae,aa/). The percentage of vowel errors due to vowel height for the CVC, IT–FT, and SV conditions, respectively, were 18%, 20%, and 27% while the corresponding numbers due to changes in frontness were 32%, 61%, and 50%. There were relatively few errors for the tense/lax vowel pairs /iy,ih/ and /ow,ao/, whereas the tense/lax vowels /uw,uh/ were somewhat confused in all listening conditions. The neutral vowel /ah/ was confused with the largest number of other vowels. There were few errors for vowels widely spaced on the traditional vowel triangle. The confusion between vowel pairs was fairly symmetric except for the vowel pairs /ah,uh/ and /ae,aa/. The main increases in vowel identification accuracy between the SV and CVC conditions were for the vowels /ih,eh,ao,ow,uh/, implying that time-varying and/or contextual information is needed to resolve these vowels.

Additional discussion of the results of the listening experiment, and comparison to automatic classification results, is deferred until the results of the automatic classification experiments are presented.

B. Automatic classification based on static spectra

1. Optimization experiments

For both the formants and DCTCs several optimization experiments were conducted to enhance the signal

TABLE IV. Confusion matrix for the IT–FT listening condition.

	/iy/	/ih/	/eh/	/ae/	/ah/	/aa/	/ao/	/ow/	/uh/	/uw/	/er/
/iy/	98.2	0.7	0.9								
/ih/	0.3	96.0	2.0	0.3				1.1		0.3	
/eh/	0.3	3.8	92.2	3.0	0.5				0.3		
/ae/			9.0	84.3	0.3	6.3	0.3				
/ah/		0.3	0.8		93.2	1.0	0.5	0.5	3.5		0.3
/aa/		0.2			5.2	76.8	16.1	0.5			
/ao/					0.4	24.2	74.2	0.7	0.2		0.2
/ow/								98.4		1.6	
/uh/					25.9			1.5	70.0	2.6	
/uw/					2.1			1.2	3.3	93.3	
/er/					0.6						99.4

TABLE V. Confusion matrix for the SV listening condition.

	/iy/	/ih/	/eh/	/ae/	/ah/	/aa/	/ao/	/ow/	/uh/	/uw/	/er/
/iy/	98.8	0.2	0.5							0.5	
/ih/	0.3	91.7	5.1	0.6	0.6				0.6	1.1	
/eh/	0.5	5.6	83.8	7.6	2.0	0.3					0.3
/ae/			9.5	81.8		8.3	0.5				
/ah/		0.3	0.8		89.1	2.0	0.3	2.8	4.6		0.3
/aa/				1.6	7.0	73.6	17.7				
/ao/					2.2	24.2	73.6				
/ow/		0.2	0.7		3.8			84.0	6.1	4.3	0.9
/uh/			0.4		24.4		0.4	2.2	67.8	4.8	
/uw/					0.7			0.5	2.4	96.4	
/er/			0.3		1.2				0.3		98.6

processing prior to the primary experiments. All refinements in signal processing were evaluated in terms of automatic vowel recognition results for the 11 vowels used in this study. All constants listed in the equations in the Appendix for the formant tracking were determined from pilot experiments. The constants listed gave "good" results, as determined both from automatic identification experiments and graphical inspection of resultant formant tracks. The window length (50 ms), low-pass filter, and high-frequency preemphasis constant (0.75) were also selected from pilot tests. In addition, one series of tests was performed to compare the differences between linear and bark scaling of formants, since some researchers have obtained higher automatic recognition results with bark scaling of formants as opposed to linear scaling (Syrdal and Gopal, 1986). The results of our experiment, in terms of automatic recognition results for four classifiers, are given in Fig. 4. For the simplest classifiers, i.e., EUC and MAH, vowel identification was significantly higher based on bark formants as compared to linear formants. However, for the more sophisticated classifiers, i.e., BML and ANN, the results were nearly identical for the two frequency scales. Apparently the BML and ANN classifiers are able to form complex decision regions that compensate for the lack of nonlinear scaling of the original features. Nevertheless, since, overall the results based on bark scaling were better than those derived from linear scaling of the formants, bark scaling was used for all later experiments. For the case of DCT coefficients, the primary optimization experiments were to evaluate (1) the number of DCT coefficients needed, (2) the required frequency range, and (3) various forms of nonlinear amplitude and frequency scales. Results of selected experiments are given in Figs. 5 and 6.

Figure 5 depicts the results of the experiment to determine the number of DCT coefficients that should be used to represent each spectrum for classification. This test was performed using a frequency range of 75 to 5500 Hz, log amplitude scaling, and bark frequency warping. Sixteen DCT coefficients were computed from the spectrum of one frame sampled at the center frame of each SV segment. Classification rates were obtained as a function of the number of DCTCs. Training and test results, obtained with the BML classifier, are shown for all speakers. In all cases, consecutively numbered DCTCs beginning with DCTC2 were used. Thus, as the value on the abscissa increases, the level of spectral detail available to the classifier increases. The difference between the training and test recognition rate increases as the number of features increases. The test recognition rate for all speakers peaks at 77.0% using DCTCs 2–11 (or ten total) and then decreases slightly if DCTCs with indices higher than 11 are used. DCTC1, corresponding roughly to overall signal level, was not used for the results plotted in Fig. 5 because its use did not

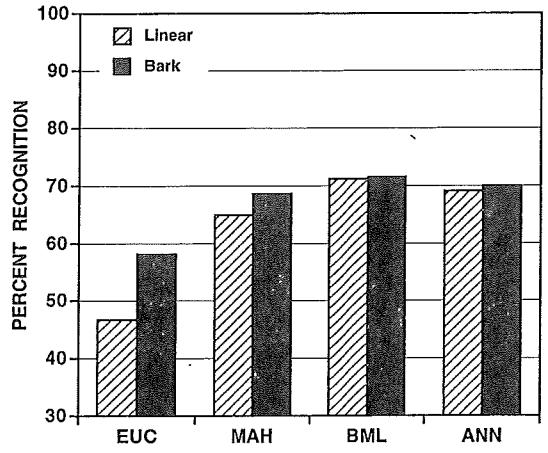

FIG. 4. Comparison of linear versus bark formant scaling of formants with four classifiers for 11 vowels.

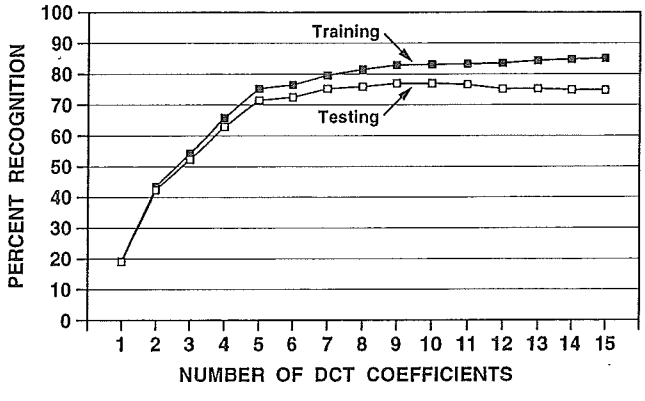

FIG. 5. Automatic recognition of 11 vowels versus the number of DCTCs used. For each case, consecutive DCTCs beginning with DCTC2 were used.

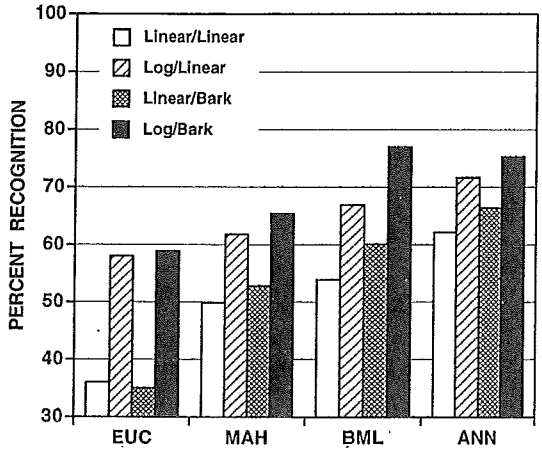

FIG. 6. Automatic recognition of 11 vowels from DCTCs 2–11, using linear and log amplitude scaling, linear and bark frequency warping, and four type of classifiers. The DCTCs were computed over a frequency range of 75 to 5500 Hz.

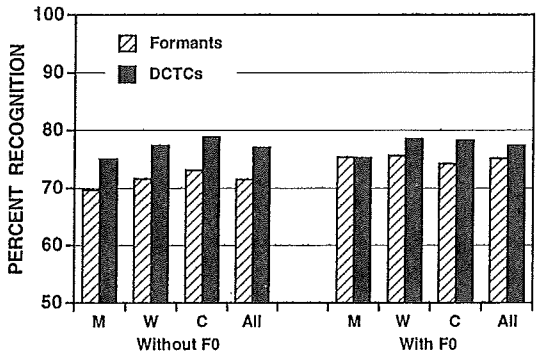

FIG. 7. Automatic recognition of 11 vowels from formants, DCTCs, formants+$F0$, and DCTCs+$F0$, as computed from one static spectrum.

improve the recognition rate. Therefore these results imply that the vowel steady-state spectrum can be encoded as a relatively smooth spectrum for use in automatic classification of vowels.

Several classification experiments were conducted to examine the effect of bandwidth used to compute DCTCs on classification accuracy. The "optimum" range, as determined from these tests, of 75 to 5500 Hz was used in later experiments. As expected, with high-frequency components removed from the signal, identification degraded the most for children and the least for men, whereas removal of low-frequency components was most detrimental for the men. In general, the drop in performance was gradual as bandwidth was reduced. For example, there was a 9.1% degradation in classification rates if a range of 0.3 to 3.0 kHz (approximately telephone bandwidth) was used.

Figure 6 depicts vowel classification results for various combinations of amplitude and frequency scales with four types of classifiers. All experiments were performed using DCTCs 2–11, computed over a frequency range of 75 to 5500 Hz, and other conditions as mentioned previously. The results clearly show that both nonlinear frequency and amplitude scaling are required to achieve high vowel classification rates. DCTCs computed using a log amplitude scale and bark frequency scale resulted in approximately 25% higher vowel classification accuracy than for DCTCs computed using linear amplitude and frequency scales. Therefore these scalings were used for additional experiments.

Since the results depicted in both Fig. 4 (formants) and Fig. 6 (DCTCs) are higher for the BML classifier than for any of the other three classifiers tested, these other three classifiers were not used for additional experiments reported in this paper.

2. Comparison of formant and DCTC results

In order to more thoroughly compare formants and DCTCs as features for representing the static vowel spectrum for use with automatic vowel classification, a summary of automatic classification experiments is depicted in Fig. 7. The figure depicts results using formants and DCTCs, with and without $F0$, as obtained with the BML classifier for four classes of speakers—men (M), women (W), children (C), and all speakers (A). With spectral features alone, vowel classification based on smoothed spectral shape is superior to that obtained with three formants. With the addition of $F0$, the two feature sets are more nearly equivalent. The addition of $F0$ increased recognition rates by 3.6% for formants but only 0.3% for DCTCs. The overall (i.e., all-speaker case) highest rate for DCTCs was 77.0% without $F0$ and 77.3% with $F0$ added. The corresponding rates obtained with formants were 71.5% and 75.1%. A correlated one-tailed t test showed that the DCTC results (all speakers) were higher than the formant results at the 99% level of confidence if $F0$ is not used ($t=3.32$). With $F0$ added as a feature, the DCTC results were higher than the formant results at the 95% confidence level ($t=1.77$).

As another comparison of formants and DCTCs, a feature evaluation algorithm similar to that described by Cheung (1978) was used to rank the top five features of various feature sets. In our implementation of this procedure, the BML classifier first determines M features that individually contribute most to classification accuracy. The algorithm then adds features that, in combination with any of the M highest-ranking features already selected, result in the highest classification rate. The procedure is iterated, finding M "best" feature sets, each with the desired number of total features. This procedure was used with three initial feature sets: (1) formants, amplitudes, bandwidths, and $F0$; (2) DCTCs and $F0$; and (3) formants, DCTCs, and $F0$. For each feature set, the algorithm was programmed to determine the best four subsets with five features each. We also note that these computations were derived from training data only (all data from all speakers), since the difference between training and test results is generally small for low-dimensionality feature spaces and since the round-robin testing described previously was not feasible with this algorithm.

The results are given in Table VI. For the formant feature set, one set of five features was significantly better

TABLE VI. The "best" five static features for vowel classification, selected from various original feature sets. The classification rates given are cumulative.

	Formants+amps+BWs+$F0$					DCTCs+$F0$					Formants+DCTCs+$F0$ Set 1					Formants+DCTCs+$F0$ Set 2				
Feature	$F1$	$F2$	$F3$	$F0$	$A1$	DCTC3	DCTC5	DCTC6	DCTC2	DCTC4	$F1$	$F2$	$F3$	$F0$	DCTC7	DCTC3	$F2$	$F3$	$F1$	$F0$
% classified	28.3	60.5	73.0	77.8	78.4	34.6	48.2	60.8	69.5	75.3	28.3	60.5	73.0	77.8	78.9	34.6	52.3	65.3	73.7	78.9

than any other set of five features. The first formant frequency contributed most individually to vowel identity. However, the addition of $F2$ improved vowel classification by an amount even greater than that obtained with $F1$ alone. Clearly $F1$ and $F2$ are the most important features, with $F3$ and $F0$ contributing to vowel discrimination, but to a lesser extent. The formant amplitudes and bandwidths contributed almost no information, with $A1$, the most important of these, adding only 0.6% to vowel classification. Therefore formant amplitudes and bandwidths were not used in additional experiments.

The results for the DCTC feature set, again with a single solution, showed that the most important features are DCTC3, DCTC5, and DCTC6, in that order. The most important single feature, DCTC3, has more vowel discriminating power than either $F1$ or $F2$ singly. However, the top three DCTC features are significantly poorer than three formants (60.8% vs 73.0% of vowels correctly classified). If five features were used, the DCTC feature set was more nearly comparable to the formant feature set (75.3% vs 78.4%). For the third feature set, consisting of both formants and DCTCs, the feature selection algorithm determined two "best" sets of five features, both similar in overall performance. Set 1 consists of formants, selected in the same order as for the formant feature set, with DCTC7 completing the set of the top five features. Set 2 begins with DCTC3, and then incorporates the formants and $F0$. $F0$ was selected in either the fourth or fifth position for all cases except for the DCTC+$F0$ feature set.

3. Discussion

The experiments reported in this section showed that if a large number of features are used (ten or more), the DCTCs are superior to formants for automatic vowel classification. However, no set of three DCTCs contains as much discrimination power as do the three formant frequencies. If $F0$ is an additional feature, the two spectral feature sets are more nearly equivalent for vowel classification.

More insight into the differences between the two feature sets can be gained by inspection of spectral plots such as those shown in Fig. 2. For the /iy/ token, both the DCTCs and formants indicate the general regions of spectral energy. However, the broad energy peak from approximately 2.5 to 4 kHz is better represented by the DCTC spectrum than by the second and third formants at 2.55 and 2.8 kHz. The /aa/ token has substantial energy throughout the low-frequency range of approximately 200 to 1500 Hz, with spectral peaks at 250, 800, and 1300 Hz. The peaks at 800 and 1300 Hz represent $F1$ and $F2$, respectively, for the formant model, but the peak at 250 Hz is ignored. Although this low-frequency energy peak may be due to glottal source characteristics rather than the vocal tract, the peak nevertheless is presumably important to perception of the /aa/.

In summary, three formants, even with their bandwidths and amplitudes included, appear to be insufficient to encode all the important properties of natural speech spectra. In contrast to three formants, ten DCTCs simply provide a much more complete spectral description. Since much of the information missing in formants is related to the voicing source, the addition of $F0$ to formants provides a substantial increase in classification accuracy. In contrast, $F0$ is less important with DCTC features, since the dominant effects of the vocal source are already accounted for.

With all of the feature sets used, automatic classification of 11 vowels obtained from the static spectrum was still significantly worse than that obtained by human listeners for 11 vowels, for any of the listening conditions. As mentioned previously, several recent studies have also noted the importance of time-varying spectral properties for vowel perception (Strange, 1989a,b; Nearey, 1989; Di Benedetto, 1989a,b). Therefore, additional experiments, reported in the following section, were conducted to evaluate the role of time-varying information in enhancing classification accuracy.

C. Classification experiments based on time-varying spectra

1. Optimization experiments

For both DCTCs and formants as features, and over several acoustic intervals, pilot experiments were conducted to compare cosine basis vectors, Legendre polynomial basis vectors, and least-squares polynomial curve fitting for encoding parameter trajectories. Unlike either of the polynomial curve fitting methods, the cosine basis vectors restrict the smoothed curve to a slope of zero at both the beginning and end of the interval, thus potentially preventing good matches to rapidly varying features at the start or end of an interval. However, in the pilot tests, the cosine basis vector features resulted in slightly higher recognition rates (although not statistically significant) than the rates obtained with either the Legendre polynomial basis vectors or polynomial curve fitting, and thus the cosine basis vector expansion was selected for the primary experiments.

Pilot experiments were also used to determine which DCTC coefficients to use and the number of cosine coeffi-

cients to use in the trajectory expansions. From these experiments, DCSs 1–3 were used for each formant, DCS2 was used for DCTC1 and DCTC2, DCSs 1–3 were used for each of DCTCs 3–7, and DCSs 1 and 2 were used for DCTC8. Note that DCTCs 9–11 were not used as features for time-varying spectra. Similarly DCSs 1 and 2 were used to encode $F0$ trajectories. Thus the number of features used to encode trajectory information was 9 for formants, 11 for formants+$F0$, 19 for DCTCs, and 21 for DCTCs +$F0$.

For the case of DCTCs, experiments were also conducted to investigate the effects of bandwidth on classification accuracy. All tests were performed over the IT–FT interval. In general, the results of these tests were quite similar to those obtained for the static spectra. The primary differences were that the classification rates were higher and that the classification rate degraded less rapidly as bandwidth was decreased for the time-varying spectra relative to the static case. For example, for "telephone bandwidth" (0.3–3.0 kHz), results for the time-varying features degraded 5.9% relative to results obtained from the optimum bandwidth of 75 to 5500 Hz (as compared to a 9.1% drop for the static case). Thus, apparently temporal cues compensate to some extent for the lack of bandwidth.

2. Vowel discrimination versus acoustic region

Experiments were conducted to evaluate the vowel information contained in multiple frames of six different acoustic regions as follows. Note that the center of the first frame was always at the beginning of the acoustic region and the center of the last frame was always at the end of the acoustic region.

(1) *Region IT.* Spectral feature sets were computed for ten equally spaced frames covering the entire IT region.

(2) *Region SV.* Spectral feature sets were computed for ten equally spaced frames covering the entire SV region.

(3) *Region FT.* Spectral feature sets were computed for ten equally spaced frames covering the entire FT region.

(4) *Region IT–FT.* Spectral feature sets were computed for 25 equally spaced frames covering the entire region from the beginning of IT to the end of FT.

(5) *Region IT+170.* Spectral feature sets were computed for 17 frames with a frame spacing of 10 ms starting at the beginning of IT. Note that for some short vowels, the end of the region may have been beyond the end of the final consonant.

(6) *Region SV-20.* Spectral feature sets were computed for ten frames with a frame spacing of 10 ms starting at 20 ms before the start of the SV segment. Thus, for some voiced consonants, the beginning of the region may have been prior to start of the initial transition.

Note that because of the frame selection method, the acoustic region of conditions (1)–(4) were time normalized, whereas the acoustic regions for conditions (5) and (6) were not. As described previously, a DCS expansion was computed for each feature over the entire region of interest. A summary of the results for both formants and DCTCs for 11 vowels, obtained with the BML classifier, is depicted in Fig. 8. For comparison, the corresponding result obtained from static features is also depicted.

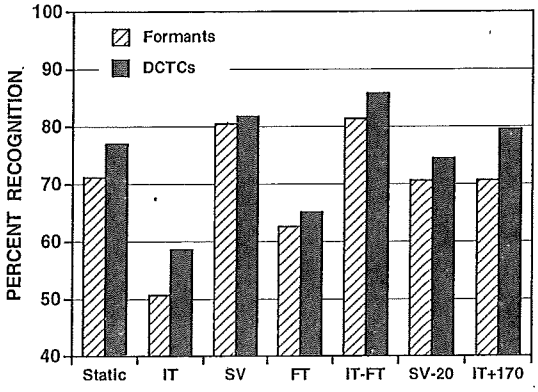

FIG. 8. Automatic recognition of 11 vowels from formant and DCTC trajectories, computed over various acoustic regions as noted. Results are based on the maximum likelihood classifier (BML).

For both formants and DCTCs, the "best" acoustic region found was IT–FT, followed by SV. For these intervals, the results were substantially higher than the results for the static case for both formants and DCTCs, showing that feature trajectories do supply additional vowel information. For example, for the SV region, the DCTC results improved by 4.8% and the formant results improved by 9.3%. The results improved even more for the IT–FT interval. Inspection of the confusion matrices for the static versus the trajectory features indicated the largest reduction in errors for /ow/ and /uh/, particularly for confusions between these two vowels. Other major confusions were also reduced to a lesser extent, with the notable exception of the /aa/ /ao/ confusion.

For every condition tested, the DCTC rates were higher than the formant rates. All differences between formants and DCTC results were significant at the 99% confidence level. The results showed that each of the transition segments, IT and FT, contain moderate (58.6% and 65.1%, respectively, for DCTCs) vowel information. Note that the features extracted from fixed length intervals were not as effective as features extracted from the "best" time-normalized intervals. Since the features for the time-normalized IT–FT interval resulted in the highest classification accuracy, these features were used for additional testing.

3. Comparison of formant and DCTC results

Another series of vowel classification tests was conducted to compare the formant and DCTC trajectory features for the IT–FT interval. Results are summarized in Fig. 9 to illustrate several points: (1) formants versus DCTCs; (2) speaker type effects; and (3) the use of $F0$ trajectory as an additional feature. Several conclusions can be drawn from inspection of this figure. First, the DCTC trajectories are superior to formant trajectories, in the absence of $F0$ information. With $F0$ included, the differences

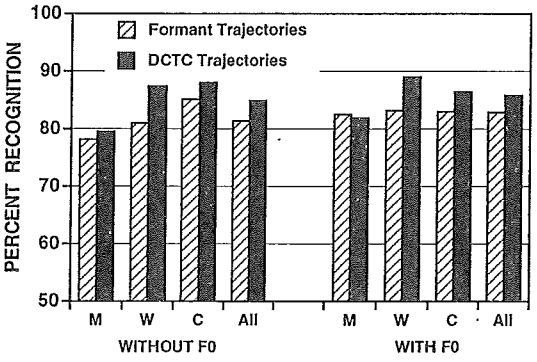

FIG. 9. Automatic recognition of 11 vowels from formant and DCTC trajectories, with and without an *F*0 trajectory, computed over the IT–FT interval.

between the two feature sets are reduced somewhat. A correlated one-tailed t test showed that differences of approximately 2.0% were significant at the 95% confidence level (or approximately 3% at the 99% confidence level). The difference between the two feature sets was 3.5% without *F*0 and 2.9% with *F*0 included. Therefore, the DCTC results were significantly higher than the formant results (t scores of 2.92 and 2.40 for the two cases). In comparing speaker types, the highest recognition rates are obtained for either the children or women, with lower results for the males (0.5% to 8.5%).[9] The use of the *F*0 trajectory improves the classification rates for all speakers in all cases, but more so for formants than for DCTCs.

The feature evaluation algorithm mentioned previously was also used to rank the top five features for three initial sets of features: (1) formants and *F*0, (2) DCTCs and *F*0, and (3) formants, DCTCs, and *F*0. The procedure used was identical to that described for the static case, except for the features used. The results are given in Table VII. For the formant feature set, the average values of the formants and *F*0 were found to be the most important features. The "tilt" of *F*2 (DCS2 of *F*2) was the most important trajectory feature. For the DCTCs the top five features were the average values of DCTCs 3–6 and 8. Thus, the results were very similar to those obtained for the static case, except that DCTC2 was no longer selected and DCTC8 was substituted. For the combined feature set, the average formant values and the average value of DCTC3 were found to be most useful for vowel discrimination.

In summary the results of the feature evaluation showed that the average value terms are much more important than are the terms that represent changes over time. However, in other tests, with the feature evaluation algorithm programmed to select the best ten features, slope terms were chosen in positions 6–10. Thus feature trajectories play a supporting role for vowel discrimination. Also note that the best five-dimensional space is better for formants than DCTCs (83.6% vs 76.6% vowel classification rates). If the "full" feature sets were used, the DCTC classification rates were significantly higher (99% confidence level) than corresponding formant results, but the magnitudes of these differences were generally small.

D. The role of duration

For both static features and features that represent trajectories over time-normalized intervals, all duration cues were missing. To check the possibility that duration information is important for vowel classification, we also used the lengths of IT, SV, and FT as additional features. Classification experiments were performed for five cases: (1) duration features only, (2) static formants augmented with duration features, (3), static DCTCs augmented with duration features, (4) formant trajectories augmented with duration features, and (5) DCTC trajectories augmented with duration features. For the duration features alone, the classification rate was 15.6% (vs 9.1% for chance). For each of the other four cases, the classification rates increased by less than 1%. None of the increases were statistically significant at the 95% level of confidence. Thus the duration cues appear to play a small role in vowel discrimination. Duration cues might, however, be more significant for vowels in naturally spoken continuous speech.

E. Comparison with listening results

The experimental results reported in the preceding sections showed that automatic vowel classification is generally superior using DCTC features versus formants, particularly in the absence of *F*0 information. However, those results do not show whether vowel perception is more closely linked to overall spectral shape or spectral peaks. We have obtained some previous evidence to show that the perception of phonologically similar vowels, synthesized with conflicting cues to vowel identity in terms of spectral shape and formants, more closely follows spectral shape cues (Jagharghi, 1990; Jagharghi and Zahorian, 1990). However, in this section we address this point by examining correlations between listening results for vowel perception and automatic classification results.

In particular the confusions obtained from listening experiments (i.e., Tables IV and V with the diagonal re-

TABLE VII. The "best" five temporal features for vowel classification, selected from various original feature sets. Each feature is labeled with feature index followed by the index of the coefficient in the DCS expansion. Thus, for example, DCTC12 is the second DCS coefficient of DCTC1.

											Formants+*F*0+DCTCs									
	Formants+*F*0					DCTCs+*F*0					Set 1					Set 2				
Feature	*F*11	*F*21	*F*31	*F*01	*F*22	DCTC31	DCTC51	DCTC61	DCTC41	DCTC81	*F*11	*F*21	*F*31	*F*01	DCTC12	DCTC31	*F*21	*F*31	*F*01	*F*11
% classified	28.3	58.9	73.1	79.8	83.6	35.8	51.1	63.4	71.7	76.6	28.3	58.9	73.1	79.8	84.1	35.8	53.0	65.9	74.2	80.5

TABLE VIII. Correlation coefficients between confusions from automatic recognition experiments with confusions from listening experiments.

	DCTCs		Formants	
	Static	Trajectories	Static	Trajectories
Without $F0$	0.48	0.73	0.32	0.62
With $F0$	0.48	0.74	0.40	0.67

moved) were correlated with confusions obtained from automatic classification experiments. The SV listening results (Table V) were correlated with automatic results from a single static spectrum while the IT–FT listening results (Table IV) were correlated with the automatic results obtained with the features extracted from the IT–FT interval.

The correlations obtained from comparing classification results with the perceptual results, for formants and DCTCs as features, are summarized in Table VIII. Note that in computing these correlation coefficients, the automatic classification results were obtained from the same nine speakers as were used in the listening tests. Confusion matrices obtained from all 30 speakers for two of the feature sets are also given in Tables IX and X. The confusion patterns for perceptual experiments and automatic classification experiments were positively correlated, with correlation coefficients ranging from 0.32 to 0.74. For all cases, the correlations were higher for the case of DCTCs than for formants. The correlations were also always higher for those cases for which automatic classification results were more accurate.

F. Context effects

Stevens and House (1963) argued that consonantal environment greatly affects the acoustic properties of phonologically equivalent vowels in a CVC context. To explore this issue with our database and with four feature sets (static formants+$F0$, static DCTCs+$F0$, formant and $F0$ trajectories, and DCTC and $F0$ trajectories), we computed vowel error profiles from classification experiments for various initial and final consonants and depict the results in Fig. 10. Results are shown only for the six English stops +/hh/ in syllable initial position and only for the six stops in syllable final position, since our database paired only these consonants with all 11 vowels.

Figure 10 shows that there are some systematic differences in error patterns depending on the consonantal context. The error patterns are similar with either formants or DCTCs as features. However there are differences in the error distributions between static and time-varying features and even larger differences between consonants appearing in initial or final position. For example, the fewest vowel classification errors were made with either /b/ or /d/ in syllable initial position if classification was based on a single frame of features. If temporal features were used for vowel classification, error rates were lowest with /b/, /d/, /g/, and /hh/. Thus, the center of the vowel is apparently least coarticulated if preceded by the voiced stops /d/ or /b/. However, coarticulation due to initial /g/ or /hh/ is readily accounted for if feature trajectories are used for classification. Similarly, coarticulation is apparently the highest for an initial /t/, as demonstrated by the relatively higher vowel error rates for this condition. For final consonants the lowest error rates were for vowels from syllables ending in /b/ whereas the highest error rates were for vowels from syllables ending in the velar stops /g/ or /k/.

IV. DISCUSSION AND CONCLUSIONS

Several issues regarding vowel identification were investigated in this study. The main issue was the role of global spectral shape parameters (DCTCs) versus spectral peaks ($F1$, $F2$, and $F3$) as acoustic cues to vowel identity. This issue was investigated using automatic classification experiments for both static and time-varying spectral features. We also examined the relative importance of various acoustic regions to vowel identification, the role of additional features such as $F0$ and duration, context effects, and the link between automatic classification results and human perception of vowels. The principal conclusions with regard to the investigations in this study are the following.

(1) Both global-shape features and formants are adequate, but redundant, information-bearing parameters for vowels. Generally results based on the two feature sets were more similar than different. The advantage of formants is that a large amount of information is contained in

TABLE IX. Confusion matrix from automatic classification using formant and $F0$ trajectories as features.

	/iy/	/ih/	/eh/	/ae/	/ah/	/aa/	/ao/	/ow/	/uh/	/uw/	/er/
/iy/	95.9	1.4						0.3		2.4	
/ih/	0.8	85.7	6.3	0.4					1.3	5.5	
/eh/		7.9	70.2	16.2	2.3			0.4	1.5	0.4	1.1
/ae/			13.4	81.7	3.4	1.1					0.4
/ah/			1.9	3.7	79.4	7.1	1.9	2.2	3.4		0.4
/aa/				0.7	4.1	80.6	14.6				
/ao/			0.3		2.4	24.3	70.8	0.3	1.7		
/ow/		0.3	1.3		2.0			86.6	4.0	5.7	
/uh/			1.1		10.0		1.1	5.6	77.8	3.9	0.6
/uw/	1.0	5.8						5.4	0.3	87.1	0.3
/er/			0.4	2.1	0.3				2.1	0.4	94.9

TABLE X. Confusion matrix from automatic classification using DCTC and $F0$ trajectories as features.

	/iy/	/ih/	/eh/	/ae/	/ah/	/aa/	/ao/	/ow/	/uh/	/uw/	/er/
/iy/	98.3	0.7							0.3	0.7	
/ih/	0.8	87.3	6.8						1.7	3.4	
/eh/		7.9	79.2	9.1	1.5	0.4	0.4		0.8		0.8
/ae/		1.1	12.7	84.3	0.7	1.1					
/ah/		0.4	2.6	1.5	80.8	8.7	2.6	1.1	1.5		0.8
/aa/				1.4	4.1	78.2	14.6	0.3	0.7		0.7
/ao/			0.3		2.4	22.2	74.3		0.7		
/ow/		0.3	1.7		1.3			94.0	0.3	2.3	
/uh/		1.1	2.2		8.3		0.6		84.4	2.8	0.6
/uw/	1.7	4.4						2.4	2.0	89.1	0.3
/er/			0.9	1.3	3.0	0.4			1.3		93.2

a few features. The advantage of DCTCs is that classification accuracy is even better than with formants, provided enough DCTCs (ten or more) are used. Thus, although spectral peaks contain most of the spectral information required for vowel discrimination, the overall smoothed spectral shape provides an even more complete description. We found no evidence that the precise location of spectral peaks is needed for vowel identification, given that the peaks are blurred in the highly smoothed DCTC spectra. However, as previously noted in the literature (Broad and Clermont, 1989; Zahorian *et al.*, 1992), approximate formant frequencies can be computed from the spectral shape features.

To the extent that automatic classification of vowels from either formants or DCTCs is different, the DCTC features are superior. The DCTC results were also more highly correlated with listener responses. In line with Bladon's (1982) "reduction" argument, we believe the reason is simply that the DCTCs provide a more complete spectral description than do the formants. Although the formants "work well" to characterize the spectra of most vowels, there are still many instances of spectra for which a three-formant model is inadequate. The plots depicted in Fig. 2 illustrate examples for which formants can be readily identified, but yet for which some prominent spectral features are poorly represented. Attempts to use more than three formants for vowels are likely to be very difficult, since, for many stimuli, there are no more than three apparent formants.[10]

Our results predict that stimuli with two "close" narrow-band spectral peaks will be perceived almost identically to a stimulus with one broader spectral peak that gives the same contribution to overall spectral shape. Conversely, spectral peaks that are separated by a distance large enough to make a significant contribution to smoothed spectral shape cannot be ignored. Perceptual experiments would be required to test these predictions. It could be argued that the formant results are unreliable because of potential errors in the tracking algorithm. However, it must be emphasized that considerable effort and signal processing sophistication were used to maximize the accuracy of the formant tracking. The tracking algorithm was not fully automatic since the algorithm made use of the "expected" formant frequencies for each token as anchor points for computations, thus "biasing" the algorithm to err on the side of improving formant-based vowel classification. Also, graphical inspection of the LP spectra and corresponding formant tracks (such as shown in Fig. 1) for a large portion of the database revealed almost no apparent errors. As yet additional evidence, classification experiments based on formant data from three speaker categories resulted in higher recognition rates for children and women than for men. If the formant tracking were error prone, the results for the children should be lower than for the men, since formant tracking is more difficult for children than for adult male speakers. Thus all these points strongly support the validity of the formant tracking algorithm used in this study.

(2) The experimental results imply that, for monopthongal vowels spoken in an isolated-word CVC context,

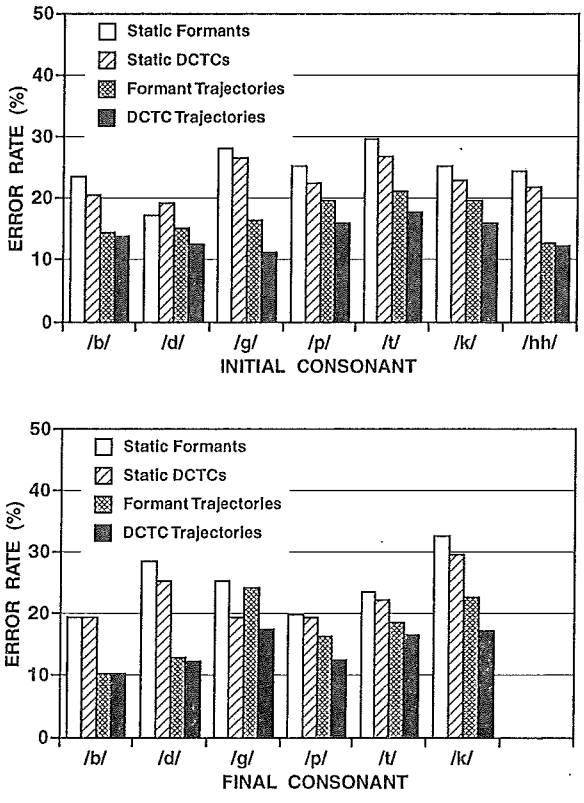

FIG. 10. Vowel error rates as a function of initial and final consonantal context of CVC syllables.

static spectral cues are more important than temporal cues. This contrasts with stop consonants, for which temporal cues are very important (Nossair and Zahorian, 1991). Even for vowels, however, feature trajectories are important secondary sources of information.

(3) The correlations between vowel confusions using automatic classifiers and those resulting from perceptual experiments were high, with correlation coefficients as large as 0.74. We therefore make the argument that automatic classification experiments can be used as a valuable tool to help understand human speech perception. On the average, perceptual confusions were more similar to confusions resulting from classification based on spectral-shape features than to those resulting from classification based on formants, thus supporting the claim that overall spectral shape is more important to perception than are the precise locations of spectral peaks.

(4) All acoustic regions in the vicinity of the steady-state vowel carry acoustic cues to vowel identity. However, of these regions, the steady-vowel is the most important, followed by the final transition, followed by the initial transition. This conclusion is in general agreement with the results of Strange (1989b), which showed that listeners can identify vowels from more than one acoustic region.

(5) Besides spectral features, fundamental frequency of voicing can be used as an additional feature to improve speaker-independent vowel classification, particularly for formants as primary features. For DCTC features, the addition of $F0$ contributes less to recognition accuracy, presumably because the DCTCs already contain most of the information encoded by $F0$. Duration cues, at least for the database used in this study, appear to have a minor role for vowel discrimination.

Note that even the best automatic classification rates reported in this paper were considerably lower than the rates obtained by human listeners (85.9% vs 91.3%). However, speaker normalization of spectral features (Zahorian and Jagharghi, 1991) can be used to increase classification rates to the same level as those achieved by human listeners. Additionally, because of the very high-dimensionality spaces required to encode time-varying features, it is possible that our database was not large enough to adequately train the classifiers for these features. The classification rates for training data with time-varying features were typically 4% to 6% higher than the test classification rates.

The main experimental result of this study, namely, that vowels can be automatically classified with a high degree of accuracy from acoustic information, supports the theory of acoustic-phonetic invariance for vowels. The results obtained in this study are more statistically reliable and comprehensive than those obtained from any previous study. First, the size of the database used was larger than that in most previous studies and included more consonantal contexts. The two principal parameter sets were also much more thoroughly investigated under many more conditions. The results of this research give insight into possible mechanisms for human decoding of speech. The results also have potential application in the field of automatic speech recognition (ASR). Specifically, the feature extraction process used in this study can be used to improve the acoustic preprocessing and phonetic analysis components of ASR systems, commonly called the front end.

ACKNOWLEDGMENTS

This work was supported by Grant Nos. IRI-8702649 and BCS-9010334 from the National Science Foundation.

APPENDIX: EQUATIONS FOR COMPUTING LOCAL AND TRANSITION COSTS FOR FORMANT TRACKING

The local cost LC of selecting formant candidate q in frame i as formant k is given by

$$\mathrm{LC}(q,i)=0.0007\ \mathrm{BandW}(q,i)+5.0\left(\frac{|\mathrm{FmtS}(k)-\mathrm{FmtCand}(q,i)|}{\mathrm{FmtS}(k)+\mathrm{FmtCand}(q,i)}\right),$$

where BandW is the bandwith in hertz of the formant candidate, FmtS is the expected value of the formant (the "seed" value), and FmtCand is the frequency in hertz of the formant candidate.

The transition cost TC of selecting formant candidate p in frame $i-1$ and formant candidate q in frame i is given by

$$\mathrm{TC}(p,q,i)=5.0\left[\left(\frac{\mathrm{FmtCand}(q,i)-\mathrm{FmtCand}(p,i-1)}{\mathrm{FmtCand}(q,i)+\mathrm{FmtCand}(p,i-1)}\right)^2+1.5\ \mathrm{DEV}(p,q,i)\right].$$

The term DEV was designed to give large costs to abrupt transitions using

$$\mathrm{DEV}(p,q,i)=|\mathrm{SLC}(p,q,i)-[\mathrm{SLP}(p,i-1)+\mathrm{SLN}(q,i)]/2|,$$

where

$$\mathrm{SLC}(p,q,i)=\mathrm{FmtCand}(q,i)-\mathrm{FmtCand}(p,i-1)$$

is equal to the slope of the track between candidates p and q in frames $i-1$ and i, respectively; and

$$\mathrm{SLP}(p,i-1)=\mathrm{FmtCand}(p,i-l)-\mathrm{FmtCand}'(p,i-2)$$

is equal to the slope of most likely track from frame $i-2$ to frame $i-1$, given that the track selects candidate p in frame $i-1$; and

$$\mathrm{SLN}(q,i)=\mathrm{FmtCand}''(q,i+1)-\mathrm{FmtCand}(q,i)$$

is equal to the slope of most likely track from frame i to frame $i+1$, given that the track selects candidate q in frame i.

FmtCand′ $(p,i-2)$ is the formant candidate in frame $i-2$ closest to candidate p in frame $i-1$ and FmtCand″ $(q,i+1)$ is the formant candidate in frame $i+1$ closest to candidate q in frame i. Thus, DEV is large if the local slope of the track deviates substantially from the most likely

track slope immediately preceding and following this segment of the track. The DEV term helps to insure continuity of formant tracks.

[1]The DARPABET phonetic notation has been used throughout this paper. The correspondence between DARPABET and IPA notation for the vowels used in this study is (iy,i), (ih,I), (eh,ɛ), (ae,æ), (ah,ʌ), (aa,ɑ), (ao,ɔ), (ow,o), (uh,U), (uw,u), (er,ɝ).

[2]Although attempts were made to be as objective as possible in the manual labeling, the decisions did involve subjectivity on the part of the experimenters. No fixed values for "high" and "low" were used.

[3]Formant bandwidths were computed using the equation $B = -(f_s/\pi) \ln |z_f|$, where f_s is the sampling frequency in hertz and z_f is the LP pole for the formant (Markel and Gray, 1976). Formant amplitudes were computed by evaluating the LP spectrum at the formant frequencies.

[4]Costs are numerically evaluated penalties computed for each possible set of formant tracks. The penalties are larger as tracks deviate more from expected behavior.

[5]The algorithm used for the formant data in this paper differs from the McCandless algorithm (1974) used to track formants in our previously reported results. The present algorithm was selected over the McCandless approach both from visual inspection of formant tracks (fewer apparent errors) and from results of vowel classification experiments (approximately 1.5% higher vowel classification rates).

[6]Window types and durations were chosen for each type of analysis according to best performance in automatic recognition experiments—thus, the differences in windows for the two types of analyses. Also the high-frequency preemphasis was found to be beneficial for formant analysis but not for DCTC analysis. However, these differences in processing resulted in only small differences in classification rates.

[7]The level of −50 dB was determined by adjusting the threshold in 5-dB steps from −20 dB to −100 dB. The −50-dB level resulted in approximately 2% higher recognition rates than the −100-dB level, as used in our previous work. This does not appear to be an artifact of our particular database, since similar results were also obtained with the DARPA/TIMIT database.

[8]However, results degrade only on the order of 1% with 25, 50, or 100 hidden nodes.

[9]A possible reason for the lower performance of the classifier with male vowels is the number of /ao/ /aa/ confusions was larger for the male speakers than for females or children, presumably because of dialect variations. Another factor might be that if the female vowel spectra are more similar on the average to the spectra of children than to male speakers, the statistical classifiers would form better representations for the female and children data than for the male data.

[10]In a pilot experiment, we used the neural network classifier with all five formant candidates (frequencies, amplitudes, and bandwidths) for vowel classification. However the results were about 10% lower than using the three tracked formants.

Beddor, P. S., and Hawkins, S. (**1984**). "The 'center of gravity' and perceived vowel height," J. Acoust. Soc. Am. Suppl. 1 **75**, S86.

Bladon, R. A. V. (**1982**). "Arguments against formants in the auditory representation of speech," from *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granstrom (Elsevier, Amsterdam), pp. 95–102.

Bladon, R. A. V. (**1983**). "Two-formant models of vowel perception: shortcomings and enhancements," Speech Commun. **2**, 305–313.

Broad, D. (**1976**). "Toward defining acoustic phonetic equivalence for vowels," Phonetics **33**, 401–424.

Broad, D. J., and Clermont, F. (**1989**). "Formant estimation by linear transformation of the LPC cepstrum," J. Acoust. Soc. Am. **86**, 2013–2017.

Carlson, R., Fant, C. G. M., and Granstrom, B. (**1975**). "Two-formant models, pitch and vowel perception," in *Auditory Analysis and Perception of Speech*, edited by C. G. M. Fant and M. A. A. Tatham (Academic, New York), pp. 55–82.

Cheung, R. S. (**1978**). "Feature selection via dynamic programming for text independent speaker recognition," IEEE Trans. **ASSP-26**, 397–403.

Chistovich, L. A. (**1985**). "Central auditory processing of peripheral vowel spectra," J. Acoust. Soc. Am. **77**, 789–805.

Chistovich, L. A., Sheikin, R. L., and Lublinskaja, V. V. (**1979**). "'Centres of gravity' and spectral peaks as the determinants of vowel quality," in *Frontiers of Speech Communication Research*, edited by B. Lindblom and S. Ohman (Academic, London), pp. 143-157.

Delgutte, B. (**1984**). "Speech coding in the auditory nerve: II. processing schemes for vowel-like sounds," J. Acoust. Soc. Am. **75**, 879–886.

Delgutte, B., and Kiang, N. Y. S. (**1984**). "Speech coding in the auditory nerve: I. Vowel-like sounds," J. Acoust. Soc. Am. **75**, 866–878.

Di Benedetto, M.-G. (**1989a**). "Vowel representation: Some observations on temporal and spectral properties of the first formant frequency," J. Acoust. Soc. Am. **86**, 55–66.

Di Benedetto, M.-G. (**1989b**). "Frequency and time variations of the first formant: Properties relevant to the perception of vowel height," J. Acoust. Soc. Am. **86**, 67–77.

Duda, R. O., and Hart, P. E. (**1973**). *Pattern Analysis and Scene Classification* (Wiley, New York).

Effer, E. A. (**1985**). "An investigation to improve linear predictive vocoder pulse excitation models," Master's thesis, Old Dominion University.

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

Fairbanks, G., and Grubb, P. (**1961**). "A psychophysical investigation of vowel formants," J. Speech Hear. Res. **4**, 203–221.

Furui, S. (**1986**). "On the role of spectral transition for speech perception," J. Acoust. Soc. Am. **80**, 1016–1025.

Gottfried, T. L., and Strange, W. (**1980**). "Identification of coarticulated vowels," J. Acoust. Soc. Am. **68**, 1626–1635.

Hillenbrand, J., and Gayvert, R. T. (**1987**). "Speaker-independent vowel classification based on fundamental frequency and formant frequency," J. Acoust. Soc. Am. **81**, S93.

Hillenbrand, J., and McMahon, B. J. (**1987**). "The role of static spectral properties in vowel identification," J. Acoust. Soc. Am. **82**, S37.

Jagharghi, A. J. (**1990**). "A comparative study of spectral peaks versus global spectral shape as invariant acoustic cues for vowels," Ph.D. dissertation, Old Dominion University, Aug. 1990.

Jagharghi, A. J., and Zahorian, S. A. (**1990**). "Vowel perception: Spectral shape versus formants," J. Acoust. Soc. Am. **87**, S159.

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

Kuwabara, H. (**1985**). "An approach to normalization of coarticulation for vowels in connected speech," J. Acous. Soc. Am. **77**, 686–694.

Lindblom, B. (**1963**). "Spectrographic study of vowel reduction," J. Acoust. Soc. Am. **35**, 1773–1778.

Lindblom, B., and Studdert-Kennedy, M. (**1967**). "On the role of formant transitions in vowel recognition," J. Acoust. Soc. Am. **42**, 830–843.

Lippmann, R. P. (**1987**). "An introduction to computing with neural nets," IEEE ASSP Mag. April, 4–22.

Markel, J. D. (**1972**). "The Sift algorithm for fundamental frequency estimation," IEEE Trans. Audio Electroacoust. **20**, 367–377.

Markel, J. D., and Gray, A. H. (**1976**). *Linear Prediction of Speech* (Springer-Verlag, New York).

McCandless, S. S. (**1974**). "An algorithm for automatic formant extraction using linear prediction spectra," IEEE Trans. **ASSP-22**, 135–141.

Miller, J. D. (**1989**). "Auditory-perceptual representation of the vowel," J. Acoust. Soc. Am. **85**, 2114–2134.

Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.

Nossair, Z. B., and Zahorian, S. A. (**1991**). "Dynamic spectral shape features as acoustic correlates for initial stop consonants," J. Acoust. Soc. Am. **89**, 2978–2991.

Oppenheim, A. V., and Johnson, D. H. (**1972**). "Discrete representation of signals," Proc. IEEE **60**, 681–691.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of the vowels," J. Acoust. Soc. Am. **24**, 175–184.

Pisoni, D. B. (**1985**). "Speech perception: Some new directions in research and theory," J. Acoust. Soc. Am. **78**, 381–388.

Plomp, R. Pols, L. C. W., and van de Geer, J. P. (**1967**). "Dimensional analysis of vowel spectra," J. Acoust. Soc. Am. **41**, 707–712.

Pols, L. C. W., van der Kamp, L. J. Th., and Plomp, R. (**1969**). "Perceptual and physical space of vowel sounds," J. Acoust. Soc. Am. **46**, 458–467.

Sachs, M. B., and Young, E. D. (**1979**). "Encoding of steady-state vowels

in the auditory nerve: Representation in terms of discharge rate," J. Acoust. Soc. Am. **66**, 470–479.
Sakoe, H., and Chiba, S. (**1978**). "Dynamic programming algorithm optimization for spoken word recognition," IEEE Trans. **ASSP-26**, 43–49.
Stevens, K. N., and House, A. S. (**1963**). "Perturbation of vowel articulations by consonantal context: An acoustic study," J. Speech Hear. Res. **6**, 111–128.
Strange, W. (**1989a**). "Evolving theories of vowel perception," J. Acoust. Soc. Am. **85**, 2081–2087.
Strange, W. (**1989b**). "Dynamic specification of coarticulated vowels spoken in sentence context," J. Acoust. Soc. Am. **85**, 2135–2153.
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.
Syrdal, A. K., and Gopal, H. S. (**1986**). "A perceptual model of vowel recognition based on the auditory representation of American English vowels," J. Acoust. Soc. Am. **79**, 1086–1100.
Talkin, D. (**1987**). "Speech formant trajectory estimation using dynamic programming with modulated transition costs," J. Acoust. Soc. Am. **82**, S55.
Williams, D. R. (**1986**). "Role of dynamic information in the perception of coarticulated vowels," Ph.D. dissertation, University of Connecticut.
Young, E. D., and Sachs, M. B. (**1979**). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," J. Acoust. Soc. Am. **66**, 1381–1403.
Zahorian, S. A., and Gordy, P. E. (**1983**). "Finite impulse response (FIR) filters for speech analysis and synthesis, International Conference on Acoustics, Speech, and Signal Processing **83**, 808–811.
Zahorian, S. A., and Jagharghi, A. J. (**1991**). "Speaker normalization of static and dynamic vowel spectral features," J. Acoust. Soc. Am. **90**, 67–75.
Zahorian, S. A., Kelkar, S., and Livingston, D. (**1992**). "Formant estimation from cepstral coefficients using a feedforward memoryless neural network," *Proceedings of IJCNN 92* (IEEE, Piscataway, NJ), Vol. 4, pp. 673–678.
Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (*Frequenzgruppen*)," J. Acoust. Soc. Am. **33**, 248.