Dynamic spectral shape features as acoustic correlates for initial stop consonants

Zaki B. Nossair and Stephen A. Zahorian
Department of Electrical and Computer Engineering, Old Dominion University, Norfolk, Virginia 23529

(Received 8 February 1990; accepted for publication 12 February 1991)

A comprehensive investigation of two acoustic feature sets for English stop consonants spoken in syllable initial position was conducted to determine the relative invariance of the features that cue place and voicing. The features evaluated were overall spectral shape, encoded as the cosine transform coefficients of the nonlinearly scaled amplitude spectrum, and formants. In addition, features were computed both for the static case, i.e., from one 25-ms frame starting at the burst, and for the dynamic case, i.e., as parameter trajectories over several frames of speech data. All features were evaluated with speaker-independent automatic classification experiments using the data from 15 speakers to train the classifier and the data from 15 different speakers for testing. The primary conclusions from these experiments, as measured via automatic recognition rates, are as follows: (1) spectral shape features are superior to both formants, and formants plus amplitudes; (2) features extracted from the dynamic spectrum are superior to features extracted from the static spectrum; and (3) features extracted from the speech signal beginning with the burst onset are superior to features extracted from the speech signal beginning with the vowel transition. Dynamic features extracted from the smoothed spectra over a 60-ms interval timed to begin with the burst onset appear to account for the primary vowel context effects. Automatic recognition results for the 6 stops (93.7%) based on 20 features was better than the rates obtained with human listeners for a 50-ms segment (89.9%) and only slightly worse than the rates obtained by human listeners for a 100-ms interval (96.6%). Thus the basic conclusion from our work is that dynamic spectral shape features are acoustically invariant cues for both place and voicing in initial stop consonants.

PACS numbers: 43.72.Ar, 43.71.Es, 43.72.Ne

INTRODUCTION

The identification of invariant acoustic correlates for stop consonants remains one of the most challenging problems in acoustic-phonetic research. In numerous studies, speech scientists have investigated basic questions such as the relative importance of the burst and the transition in providing stop information, whether the information is primarily context invariant or whether it depends on the adjacent vowel, and whether the cues are primarily static (i.e., dependent on features extracted from one speech frame sampled at the beginning of the signal) or dynamic (i.e., dependent on features extracted from several frames of the speech signal). Clear-cut answers have been difficult to find. In early work, Fischer-Jorgensen (1954) showed that the effectiveness of burst cues depends on both the stop and the vowel. For example, the burst of /b/ and /g/ appears to signal the stop identity before the vowels /i/ and /u/ but not /a/, whereas the /d/ burst signals the stop before /i/ but not before /a/ and /u/. Experiments with synthetic speech at Haskins Laboratory indicated that the formant transitions that follow the release burst encode the primary acoustic cues for place of articulation in initial stop consonants (Delattre *et al.*, 1955; Liberman *et al.*, 1954; Liberman *et al.*, 1967). However, since the formant transition patterns vary greatly depending on the vowel, these cues presumably are highly dependent on the vowel. Other studies, using tape splicing experiments, have shown that the burst section of the waveform contains sufficient cues for stop consonants and that these cues are primarily independent of the following vowel (Cole and Scott, 1974a, 1974b). In contrast to the claims of Cole and Scott, Dorman *et al.* (1977), who also used tape splicing experiments, claimed that burst and formant transitions are functionally equivalent (complement each other) context-dependent cues to stop consonants. Schouten and Pols (1983), who conducted a perceptual study on Dutch stops similar to the study by Dorman *et al.*, emphasized that the initial burst carries more information than the vocalic transition, for any vowel context, for both unvoiced and voiced stops.

More recent studies with English stops claim that the global shape of the spectrum sampled over the first 20–50 ms of the speech waveform contains acoustically invariant cues to place of articulation in stop consonants, across vowel context and talkers (Stevens and Blumstein, 1978, 1981; Blumstein and Stevens, 1979, 1980; Kewley-Port, 1983; Kewley-Port *et al.*, 1983; Kewley-Port and Luce, 1984). However, while these investigators have agreed that the cues are encoded in the global shape of the spectrum and are invariant across vowel context and talkers, they have disagreed upon the form of these cues. Stevens and Blumstein initially argued that the cues are encoded in the global shape of one static spectrum computed from one 25.6-ms speech frame sampled at the release burst of each stimulus (Stevens and Blumstein, 1978; Blumstein and Stevens, 1980). Using templates based on the features proposed by Stevens and Blum-

0001-4966/91/062978-14$00.80

stein, a panel of human viewers was able to correctly label place of articulation for 85% of voiced and voiceless stops (Blumstein and Stevens, 1979). However, in other more recent studies, (Lahiri and Blumstein, 1981; Ohde and Stevens, 1984) these views have been replaced by a description of the acoustic correlates that is more dynamic in nature. Kewley-Port (1983) has claimed that invariant cues to place of articulation lie in the dynamic changes in spectral energy over time observed in the first 50 ms beginning with the release burst. Using templates based on her features, human viewers were able to correctly identify place of articulation for approximately 88% of initial voiced stops (Kewley-Port, 1983; Kewley-Port and Luce, 1984). In contrast, however, Suomi (1985), using automatic classification experiments for initial stops, claimed that the cues for stops are heavily dependent on the vowel context, thus making reliable stop consonant recognition impossible unless the vowel is known.

Many workers in automatic speech recognition have attempted to devise signal processing schemes for extracting features for automatic identification of stops. Searle *et al.* (1979) extracted features over a 100-ms interval from the output of a psychophysically motivated filter bank and were able to classify 77% of the six initial stops correctly from these features. Tanaka (1981) used formant transitions (trajectories of the first three formants and their relative amplitudes) over the first 50 ms of the speech waveform and reported automatic recognition rates of 81% for voiced stops and 84% for voiceless stops. In studies completed over the last 10 years, automatic classifiers for initial stops have generally been able to achieve about 90% speaker-independent recognition for stops (Yoder and Jamieson, 1987; Rossen *et al.*, 1988) and as high as 98% recognition for the speaker-dependent case (Waibel *et al.*, 1989). To date, none of the speaker-independent automatic recognition schemes match the performance of human listeners in identifying stops. Lamel *et al.* (1987) found that listeners can identify about 97% of initial stops correctly and 85% of mid and final stops correctly, even with consonants extracted from continuous speech, from a wide variety of talkers.

A. Objectives

The objectives of the present study were to extend the results of the previously mentioned investigations in several ways to develop in more detail a set of acoustic features that encode sufficient information to distinguish the initial stop consonants. The feature sets were investigated primarily through automatic recognition experiments. We explored in detail features based on overall spectral shape versus features based on spectral peaks (formants). We compared features based on one spectrum sampled at the release burst (the static spectrum) versus features based on the temporal history of several frames of spectra (dynamic spectra). We also compared dynamic spectral features timed to begin with the burst versus dynamic spectral features timed to begin with the vowel transition after voicing onset. The use of automatic classification experiments in this study allowed us to fine tune the definition of the acoustic features for initial stops using a large set of talkers and utterances over a wide range of conditions that produce variability in the acoustic-phonetic properties of speech. Additionally, in contrast to most previous studies for which acoustic correlates were only defined for place of articulation in initial stops, the features defined in this study cue both place of articulation and voicing.

Another main objective of this work, directed specifically to automatic recognition of initial stops, was to define a small set of "informationally rich" features, well suited for use in automatic classification. Part of the motivation for this objective is the practical desire to be able to adequately train an automatic classifier. That is, a basic principle of pattern recognition theory is that the amount of data required to train a classifier increases exponentially with the number of features (Duda and Hart, 1973). An even more fundamental motivation is to be able to extract enough acoustic-phonetic information to enable reasonably accurate phoneme recognition without the extensive use of high-level speech knowledge. Thus an additional goal of this work was to identify features and signal processing techniques that can improve the performance of phoneme-based automatic speech recognition systems.

In the remainder of this paper, we describe our database, explain our feature extraction and automatic classification algorithms, give experimental procedures and results, and finally give a general discussion and conclusions based on these experiments.

I. DATABASE

A. Tokens

Naturally produced CVC-isolated tokens were used in all experiments of this study. Eighty-four CVC tokens were recorded for each of 30 native English talkers. Ten of these talkers were adult males, ten were adult females, and ten were children between the ages of 7 and 11. The six initial stops were /b/, /d/, /g/, /p/, /t/, and /k/. The vowel in each syllable was one of the 11 vowels /a,i,u,æ,ɝ,ɪ,ɛ,ɔ,ʊ,U,o/ and the final consonant was one of the 8 consonants /b,d,g,k,t,p,v,s/. The final consonants were chosen to maximize the number of meaningful words and also because of requirements of other experiments conducted with the same data base. About 2/3 of the tokens were meaningful words and about 1/3 were nonsense syllables. Each initial stop was paired with at least one instance of each of the 11 vowels, i.e., each initial stop was spoken in 11 vowel contexts. The total number of tokens (84) was greater than 66 (6 consonants × 11 vowels) because of other experiments (not reported in this paper) performed with the vowel and final stop portions of the tokens.

B. Recording conditions and signal preprocessing

All recording sessions were held in a sound-attenuated room. The typical sound level of speech sounds was approximately 36 dB above the background noise level in the room. An automated recording procedure was used for recording each talker. An Electro-Voice RE10 dynamic cardioid microphone was used in all recording sessions. A level-activated trigger with a 310-ms pretrigger buffer insured that the entire signal was captured. The experimenter asked each talker to repeat any syllable that the experimenter judged to be mispronounced. Each speech waveform was low-pass filtered at 7.5 kHz with a 6th-order Butterworth analog filter before sampling at 16 kHz with a 12-bit A/D converter. All speech files were digitally high-pass filtered at 240 Hz with a 62nd-order linear phase FIR filter to remove low-level, low-frequency noise in the signal.

C. Segmentation

The acoustic regions of all speech files were manually labeled using an interactive computer waveform editor. The stimuli were segmented into a maximum of ten acoustic regions. The acoustic regions pertinent to initial stops were defined as follows.

(1) Prevoicing (PV), a portion of periodic or voiced signal that occurs just before the consonantal release. This acoustic segment occurs in about 10% of the /b/, /d/, and /g/ syllables.

(2) Initial burst (IB), a portion of frication noise occurring just after the consonantal release.

(3) Initial aspiration (IA), a portion of aspirative noise that occurs after the initial burst. This aspirative segment usually appears for unvoiced stops but very rarely for voiced stops.

(4) Initial transition (IT), a periodic waveform that begins at the first voicing pulse and ends at the start of the steady-state vowel.

(5) Steady-state vowel (SV), a section of periodic steady-state waveform.

The mean and standard deviation for the voice onset time (VOT), the IT interval, and the SV interval are given in Table I. The voice onset time (VOT) was defined as the time from the onset of IB to the beginning of the first voicing pulse (the beginning of IT). Note that the duration of IB was not used in later experiments since it was often difficult to locate the precise boundary between the burst and the aspiration for the unvoiced stops. Also note that the defined starting point for IT, the beginning of voicing onset for both voiced and unvoiced stops, is not the same as that used in some data reported in the literature. Our definition for IT was chosen because the onset of voicing is reasonably clearly defined. All segmentation points were defined through visual and auditory inspection of the relevant parts of the waveform. As an additional aid, the spectral derivative (Furui, 1986a) was also displayed along with each acoustic waveform to help define the boundaries between the different acoustic segments. To minimize transients, all segmentation points were chosen to coincide with zero crossings in the speech waveform.[1] After labeling the entire data base and excluding the misrecorded or mispronounced files, the number of usable stimuli for the six initial stop consonants was 2481 (out of 84×30 = 2520).

TABLE I. Mean and standard deviation (s.d.) of the length (in ms) of three acoustic segments for stop consonants spoken in syllable initial position.

			A. Voiced stops			
	/b/		/d/		/g/	
Segment	Mean	s.d.	Mean	s.d.	Mean	s.d.
VOT	11.3	4.8	16.0	5.9	25.0	11.0
IT	31.2	11.4	34.0	12.3	38.8	16.8
SV	163.5	80.9	165.0	85.0	151.4	62.1
			B. Unvoiced stops			
	/p/		/t/		/k/	
Segment	Mean	s.d.	Mean	s.d.	Mean	s.d.
VOT	78.4	29.7	87.5	30.1	95.6	27.6
IT	32.2	10.0	35.3	12.5	36.2	12.7
SV	126.5	64.4	150.0	77.8	157.6	79.2

II. SPEECH PARAMETERS AND CLASSIFICATION METHODS

A. Speech parameters

In this section, we explain the signal processing techniques used in computing the two features sets investigated in this study. These two feature sets are formants and a form of cepstral coefficients. Since the cepstral coefficients were computed as a cosine transform of the nonlinearly scaled magnitude spectra, and were computed somewhat differently than the usual method for computing cepstral coefficients, we refer to them as discrete cosine transform coefficients (DCTC's). The formants encode the peaks in the spectrum and are traditionally considered to be the primary acoustic cues to phoneme identity. The DCTC's encode the smoothed overall shape of the spectrum. Thus these two parameter sets represent two different points of view regarding the most important acoustic-phonetic features.

1. Formants

Formants were computed for the initial stops 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 pre-emphasized with transfer function $(1-0.75 z^{-1})$. The signal was windowed with a 25-ms Hanning window and a 10th-order LP model was computed. The roots of the LP polynomial were computed in order to determine up to five formant candidates (frequency, amplitude, and bandwidth) for each frame. Formant candidates were obtained for 25 frames (5-ms frame spacing), beginning at the burst for the voiced stops and were computed for 50 frames (5-ms frame spacing) for the unvoiced stops. Finally, a formant tracking routine (similar to McCandless, 1974), which makes use of the continuity property and the bandwidth limitation of formants, was used to track the formants from the last frame (i.e., a vowel region) back to the burst. The resulting formant values in the initial region of each stimulus represent that stimulus. Besides the formants (F1, F2, and F3), the log of the formant amplitudes (A1, A2, and A3) and the formant bandwidths (B1, B2, and B3) were also computed and used as parameters.

The performance of the formant tracking routine was verified by comparison of the computed formant trajectories for 400 stimuli with manually tracked formants for the same stimuli. All trajectories obtained using the formant tracking routine for these 400 stimuli matched the manually obtained trajectories. However, for both the manual and automatic case it was not always possible to determine continuous formant trajectories from the burst onset through the vowel. Rather, at least for some of the stimuli, it appeared that one set of spectral peaks began in the burst and eventually disappeared while another set appeared with the onset of the vowel transition. For many tokens, particularly labial and alveolar stops, the formants are simply not well defined in the burst and aspiration segments. Spectral zeros in the burst spectra introduce additional problems with formant tracking based on an all-pole model. Formants also may change rapidly during these segments. Thus the 25-ms analysis window, selected to provide good tracking during intervals of slowly varying formants, may prevent good tracking in the initial portions of the waveform. Despite these limitations, the automatic routine appeared to track the formants as continuously as was possible with manual labeling of the formants from LP spectral peaks.

2. Cepstral coefficients

The cepstral coefficients, i.e., the DCTC's, were computed over the original frequency range (0–8 kHz) as follows. First, the speech signal was high-frequency pre-emphasized with transfer function $(1\text{–}0.95\,z^{-1})$.[2] The speech signal was then windowed using a Hamming window. Depending on the length of the window, either a 256- or 512-point FFT was computed for each speech frame. The magnitude spectrum of each speech frame was computed from the complex-valued output of each FFT. Let $H(f)$ denote the magnitude spectrum of a speech frame, $H'(f)$ a nonlinearly amplitude scaled version of $H(f)$, $H'(f')$ a nonlinearly warped version of $H'(f)$, and let $[H'(f')]$ be a portion of $H'(f')$ over a selected frequency range. The DCT coefficients are defined as the $a_n's$ in the equation

$$[H'(f')] = \sum_{n=1}^{n=N} a_n \cos[(n-1)\cdot\pi\cdot f']. \qquad (1)$$

Several pilot experiments were conducted to evaluate various nonlinear amplitude scales, nonlinear frequency scales, and frequency ranges, in terms of their effect on automatic classification accuracy of initial stop consonants. Based on these experiments a log amplitude scaling was selected. Bilinear frequency warping (Oppenheim and Johnson, 1972) with a coefficient of 0.5 was also selected for the primary experiments. That is,

$$f' = f + \frac{1}{\pi}\tan^{-1}\left\{\frac{0.5\sin(2\cdot\pi\cdot f)}{1-0.5\cos(2\cdot\pi\cdot f)}\right\}. \qquad (2)$$

The DCT coefficients were computed over a frequency range of 200–6000 Hz. Thus the DCT coefficients are traditional cepstral coefficients except for added flexibility in frequency range selection, frequency scaling, and amplitude scaling. Note that a_1, which we call DCTC1, is a measure of the average level of the spectrum; a_2, which we call DCTC2, is a measure of the spectral tilt, and so on. Also note that a smoothed spectrum can be computed from the DCTC's, with the degree of smoothing dependent on the number of DCTC's used to reconstruct the spectrum.

To illustrate the differences between LP smoothing of spectra and DCTC smoothing of spectra, Fig. 1 depicts the original FFT magnitude spectrum, a 14th-order LP model spectrum, and the same spectrum computed from ten DCTC's. These spectra were computed from a 25.6-ms speech frame sampled at the burst onset of the initial stop of the word "dot" for a female talker. Because of the length of the time window, this "burst" spectra is heavily influenced by the following /a/. These spectra are plotted on a log amplitude scale, using a frequency scale with a bilinear warping coefficient $= 0.5$, for a frequency range from 200 Hz–6 kHz. Although both the LP and DCTC spectra are greatly smoothed with respect to the FFT spectra, the LP spectrum gives much more emphasis to the spectral peaks than does the DCTC spectrum. In addition, since the DCTC smoothing occurs after the frequency warping and the log amplitude scaling, whereas the LP smoothing was performed on linear amplitude and frequency scales and then simply displayed on the log amplitude scale and warped frequency scale, the smoothing is quite different for the two cases. The 14th-order LP spectrum is compared with the 10th-order DCTC spectrum, since the LP spectrum was originally computed over the full-frequency range, DC to 8 kHz, versus the 200- to 6000-Hz range used for the DCTC computations.

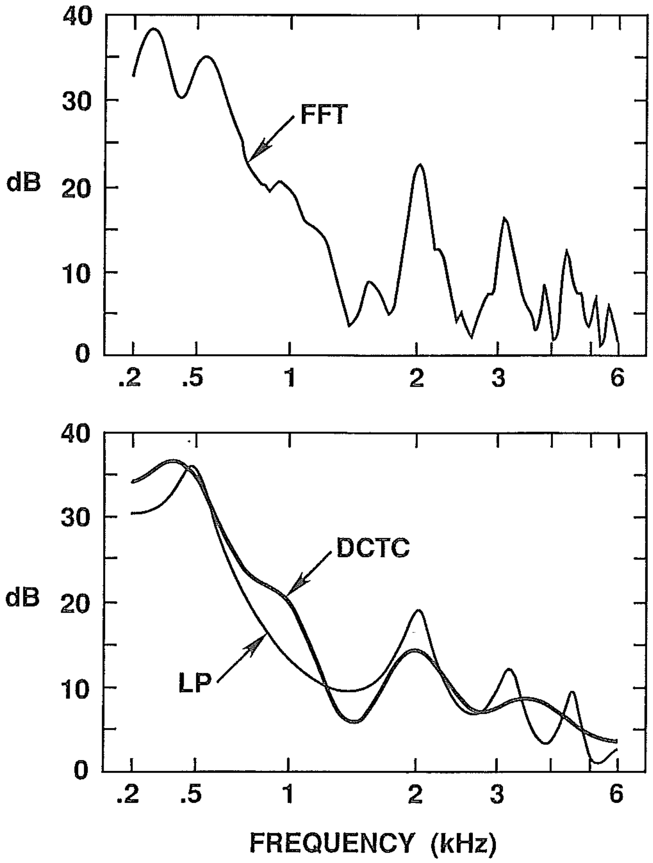

FIG. 1. The spectrum of the burst onset for /d/ in "dot" from a female speaker computed with three methods: the FFT spectrum, a 10th-order DCT spectrum, and a 14th-order LP spectrum.

B. Features for dynamic spectra

Speech features were also computed for each of several speech frames, in order to evaluate automatic recognition accuracy for the case of dynamic spectra. Several methods were first investigated for sampling the spectra and for combining the parameters of several frames. These methods were evaluated in terms of automatic stop consonant recognition accuracy. The best approach found was to sample the speech spectra with frames equally spaced starting at the burst. 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 cosine basis-vector expansion. That is,

$$P(n) = \sum_{k=1}^{N} C_k \cos \frac{(k-1)\cdot\pi\cdot n}{(L-1)}, \qquad (3)$$

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. Although several values for N were investigated, the best results were usually obtained with $N = 3$.

Thus the coefficients C_1, C_2, and C_3 in Eq. (3) encode the smoothed trajectory of a speech parameter. Here, C_1 is the average value of a parameter, C_2 is a measure of the tilt over time of a parameter, and C_3 provides additional detail of a parameter trajectory. These coefficients were the features for the automatic classifier. Thus, with this approach, time-smoothed dynamic parameters are the classification features. To illustrate the effect of this smoothing, Fig. 2 shows the original trajectories of two formants (F1 and F2) after tracking and the same trajectories after each was smoothed with a three-term cosine expansion. These trajectories correspond to the first 120 ms of the word "dot" for a female talker.

The choice of the cosine basis vectors for the time expansion was motivated by pilot experiments comparing cosine basis vectors, Legendre polynomial basis vectors, and least-squares polynomial curve fitting. 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,[3] and thus the cosine basis vector expansion was selected for the primary experiments.

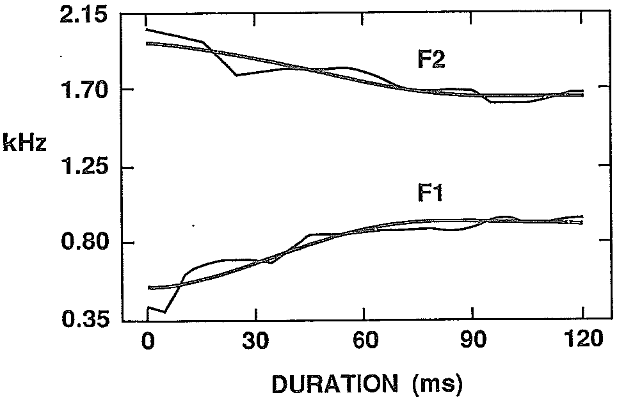

FIG. 2. Formant trajectories for $F1$ and $F2$ before and after smoothing with a three-term cosine series expansion. These two trajectories are for the first 120 ms of the word "dot" for a female talker.

C. Classifiers

All feature sets for the stops were evaluated in terms of their effects on automatic classification accuracy with a Bayesian maximum likelihood classifier (BML). That is, each stimulus was classified according to the category for which the distance

$$D_i(\mathbf{x}) = (\mathbf{x} - \mathbf{x}_i)^T \mathbf{R}_i^{-1} (\mathbf{x} - \mathbf{x}_i) + ln|\mathbf{R}_i| - 2\, ln\mathbf{P}(G_i), \quad 1 \leqslant i \leqslant M, \qquad (4)$$

is minimized. In Eq. (4), $\mathbf{x}$ 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 $\mathbf{P}(G_i)$ 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).

In all the automatic classification experiments reported in this paper the speakers used for training the classifier were different from those used for testing the classifier. More specifically, 15 speakers, 5 adult males, 5 adult females, and 5 children, were used to train the classifier and the other 15 speakers of our data base, 5 adult males, 5 adult females, and 5 children, were used for evaluation. Thus all comparisons of features sets are derived from **speaker-independent** automatic recognition experiments.

III. EXPERIMENTS

A. Listening experiment

In addition to the automatic classification experiments, we also conducted a listening experiment. The objectives of this experiment were: (1) to evaluate our data base; (2) to obtain an estimate of the interval of the speech signal required by human listeners to identify the initial stops; (3) to use the identification rates obtained by human listeners as a control for the results obtained by automatic classification experiments; and (4) to determine the extent to which the presence of the entire vowel signal helps listeners to identify the initial stop.

1. Method

a. Stimulus specification. The experiment was conducted with the data from 9 of the 30 talkers—3 adult males, 3 adult females, and 3 children. These nine talkers were selected as follows. An automatic classification experiment for the 6 initial stops was first conducted based on the data of each of the 30 talkers. The results of this automatic classification experiment were then used for ranking the ten adult males, the ten adult females, and the ten children. For each of the three groups, we chose the highest ranking, lowest ranking, and middle ranking talker for use in the listening experiment. This method for selecting talkers was chosen to allow a subsequent comparison of the talker rankings by automatic classification versus talker rankings by listeners.

b. Subjects. Five female students at Old Dominion University served as subjects for the listening experiment. Subjects were contacted through school bulletin board ads and were paid for their participation. Subjects were phonetically naive and had no known history of a hearing or speech disorder at the time of the experiment.

c. Procedure. The first step in our procedure was to insure that all stimuli of the nine talkers had approximately equal amplitude when presented to the subject. To do this, each stimulus was normalized so that each complete CVC syllable had the same maximum short-time absolute value. That is, the average absolute value was computed for each 25-ms segment of each token. Each token was then scaled so that the maximum of the average absolute values for each token would be the same.[4] The stimuli of each talker were then randomly ordered and were automatically presented to each subject individually in a sound-treated room through headphones (Koss-Pro/4×PLUS). For each listening condition the nine talkers were also presented to the subject in random order, but grouped together in blocks of adult males, adult females, and children. Each subject attended an initial 1/2-h training session in which the experimental procedure was explained and 40 CVC syllables from a female talker outside the testing set were presented. In this training session the subject entered computer-keyboard responses for the initial consonant, the vowel, and the final consonant. The subject was informed, during this training session, of incorrect responses and was allowed to listen again to that stimulus. After this training session, the subjects were presented with stimuli for the first listening condition. No feedback was given after the initial training session. The listeners could listen to each stimulus as many times as they desired but had to make a forced choice response among one of the six stops. This experiment was performed for six initial stop-segment conditions, summarized as follows: (1) the entire CVC syllable (CVC); (2) the beginning of the burst up to the end of the following vowel (IB-SV); (3) the beginning of the burst up to the end of the transition to the following vowel (IB-IT); (4) a 50-ms segment beginning at the burst onset (IB-50); (5) a 100-ms segment beginning at the burst onset (IB-100); and (6) the beginning of the initial transition up to the end of the following vowel (IT-SV) (for /b,d,g/ only).

Each subject completed these six conditions in approximately five 1-h sessions over a 2-week period. All subject responses were scored using an automatic scoring program.

TABLE II. Summary of listening results for identification of initial stops for a variety of conditions. All results were obtained by averaging over five listeners and nine talkers. All results are for six stops, except for condition IT-SV, which was for the three voiced stops only.

No.	Condition	% identification
1	CVC	98.2
2	IB-SV	97.6
3	IB-IT	95.7
4	IB-50	89.9
5	IB-100	96.6
6	IT-SV (for the three voiced stops only)	74.5

2. Results

Table II gives the average percent identification, averaged over five listeners and nine talkers, for the six initial stop consonants for each of the six listening conditions. As Table II shows, the average percent correct for the six initial stops, based on listening to the entire syllable was 98.2%, thus showing that nearly all the stop tokens were identifiable by all the listeners. Comparison of conditions 2 and 3 of Table II indicates that there is approximately a 2% percent improvement in the identification rate of initial stops based on the burst through the end of the steady-state vowel versus the burst through the end of the initial transition. Inspection of the confusion matrices (Tables III and IV) for these two conditions indicates that almost all the improvement in condition 2 over condition 3 was due to improvements in the identification of voiced stops. This change in error pattern seems reasonable since the interval consisting of the burst plus initial transition is generally much shorter for voiced stops than for unvoiced stops. Conditions 4 and 5 of Table II (confusion matrices given in Tables V and VI) show that listeners were able to identify stops with about 90% and 97% accuracy, respectively, from 50 and 100 ms of the speech waveform timed to begin with the burst. Again, the recognition rate improved much more for the voiced stops than for the unvoiced stops—the voiced stop results changed by 9.7% vs 3.5% for the unvoiced stops. For all four confusion matrices given, the identification rates for the unvoiced stops are somewhat higher than for the voiced stops. There is also, except for /g/–/k/ pair and the 100-ms interval, a greater tendency to confuse voiced stops with unvoiced stops with the same place of articulation than vice versa.

Condition 6, based only on the voiced stops, shows that the identification rate of initial stops decreases significantly if the burst section is removed (96.8% for voiced stops for the IB-SV interval versus 74.5% for the IT-SV interval). A two-tailed *t* test indicated that a difference of 1.2% was significant at the 95% confidence level. This large difference in identification rates thus suggests that the burst section is essential for identifying initial stops, even for voiced stops.

The results of these experiments are in general agreement with other similar experiments reported in the literature. For example, the 97% rate for condition 5 is very similar to the rate obtained by Lamel *et al.* (1987) with stop stimuli extracted from continuous speech. The general result

TABLE III. Confusion matrix resulting from listening experiment for identification of six initial stops based on the IB-SV interval.[a]

	/b/	/d/	/g/	/p/	/t/	/k/
/b/	96.3	0.5	0.6	2.4	0.2	0.0
/d/	0.9	95.1	3.4	0.0	0.6	0.0
/g/	0.0	0.9	99.0	0.0	0.0	0.2
/p/	0.0	0.0	0.0	99.7	0.0	0.3
/t/	0.0	0.0	0.0	0.0	97.0	3.0
/k/	0.0	0.0	0.0	0.9	0.4	98.7

TABLE IV. Confusion matrix resulting from listening experiment for identification of six initial stops based on the IB-IT interval.

	/b/	/d/	/g/	/p/	/t/	/k/
/b/	93.7	0.3	1.3	4.7	0.0	0.0
/d/	2.2	89.2	4.1	0.0	4.1	0.4
/g/	0.3	0.5	96.0	0.0	0.0	3.1
/p/	0.5	0.0	0.0	99.4	0.2	0.0
/t/	0.0	0.0	0.0	0.0	97.1	2.9
/k/	0.0	0.0	0.1	0.7	0.4	98.7

that listeners can accurately identify initial stops from a short interval timed to begin at the start of the burst also agrees with the results of several studies (for example, Tekieli and Cullinan, 1979; Kewley-Port *et al.*, 1983), although the details of the exact length of time required do not agree. Additional discussion of the results of this experiment, and comparison to the automatic classification results, is deferred until the results of the automatic classification experiments are presented.

B. Classification experiments based on static burst spectra

The objective of the first series of automatic classification experiments was to optimize features for identification of initial stop consonants based on the burst spectrum computed from one 25.6-ms speech frame sampled at the burst onset of each stimulus. This speech frame was sampled from the stimulus using a half-Hamming window. A 25.6-ms half-Hamming window is rectangular over the first 12.8 ms and the "second half" of a 25.6-ms Hamming window for the last 12.8 ms. This window was timed to begin with the burst onset. A pilot experiment indicated that slightly higher (about 1%) automatic classification rates were obtained with a half-Hamming window versus the full Hamming window. This is also the same window used for this application in previous work (Blumstein and Stevens, 1979). Our objectives were to determine to what extent the burst spectrum carries information for initial stops, for both place of articulation and voicing features, and to contrast the success of formants versus DCTC's in representing this information.

For the case of DCT coefficients several experiments were conducted to optimize the signal processing involved in the spectral shape representation prior to comparison of the DCTC representation of the burst spectrum with the formant representation. Variables investigated included the frame length, several forms of nonlinear amplitude and frequency scales, frequency range, and the number of DCT coefficients required. Since most of these effects were small, we only give a brief summary of the results, except for the experiment to investigate the required number of DCT coefficients. In particular, the best frame size was found to be 25.6 ms (vs 20 and 30 ms), the best amplitude scaling was log (versus linear or power function), the best frequency warping was bilinear with a coefficient of 0.5 (versus other coefficients for bilinear scaling and mel and sine warping), and the best frequency range was 200–6000 Hz (versus other ranges selected from 0–8000 Hz). In all cases "best" was defined as highest automatic recognition results for 6 stops with data from the 15 test speakers, over the conditions investigated.

TABLE V. Confusion matrix resulting from listening experiment for identification of six initial stops based on a 50-ms interval timed to begin with the burst.

	/b/	/d/	/g/	/p/	/t/	/k/
/b/	90.2	0.0	2.6	7.3	0.0	0.0
/d/	1.9	84.1	6.9	0.0	7.1	0.0
/g/	0.2	0.5	84.8	1.2	0.2	13.1
/p/	3.2	0.3	0.3	95.0	0.2	1.1
/t/	0.0	0.6	1.1	0.9	95.5	2.0
/k/	0.0	0.0	5.5	2.4	2.0	90.1

TABLE VI. Confusion matrix resulting from listening experiment for identification of six initial stops based on a 100-ms interval timed to begin with the burst.

	/b/	/d/	/g/	/p/	/t/	/k/
/b/	97.3	0.2	0.8	1.8	0.0	0.0
/d/	1.5	93.5	3.9	0.0	0.9	0.2
/g/	0.5	0.9	97.6	0.0	0.0	1.0
/p/	1.1	0.0	0.0	98.0	0.0	0.9
/t/	0.0	0.3	0.3	0.2	97.4	1.7
/k/	0.0	0.1	3.3	1.0	0.0	95.5

The most important optimization experiment was the one conducted to determine the number of DCT coefficients that should be used to represent each spectrum for classification. This test was performed using the parameter values listed above for frame length, frequency range, etc. Sixteen DCT coefficients were computed from the spectrum of one frame sampled at the burst onset of each stimulus in the database. Classification rates were obtained as a function of the number of DCTC's. Figure 3 depicts the results. Note that as the value on the abscissa increases, the level of spectral detail available to the classifier increases. Figure 3 shows that the recognition rate for test data improves very little if DCTC's are added after DCTC7. The test rate even decreases slightly if DCTC's with indices higher than 10 are used. Note that DCTC1, corresponding roughly to overall signal level, was not used for the results plotted in Fig. 3 because its use decreased the test recognition rate. Therefore these results imply that the burst spectrum can be encoded as a relatively smooth spectrum for use in automatic classification of stops.

Thus DCT coefficients 2–10, computed as outlined above, are an encoding of the smoothed spectral shape of the burst spectrum. The first three formants and their log amplitudes were also computed for the burst spectrum. Automatic classification experiments were then conducted for three parameter sets (DCTC's; formants; formants plus amplitudes) for each of three conditions: (1) voiced stops only (3V); (2) unvoiced stops only (3U); and (3) all six stops (6S). For conditions 1 and 2 place of articulation must be determined whereas for all six stops, both the place and voicing features must be distinguished. Figure 4 summarizes the training and

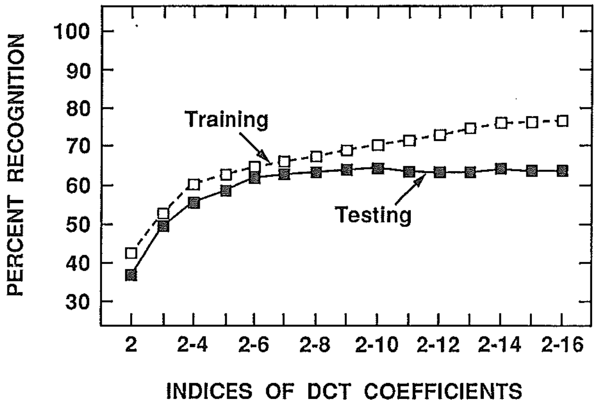

FIG. 3. Automatic classification rates for the six initial stops as a function of the number of DCT coefficients for one static spectrum.

test results for these conditions and these parameter sets, all based on the BML classifier.

The results given in Fig. 4 indicate that global spectral shape features are much better for identifying the stops than are the values of formants in the burst interval. As expected, none of the features are sufficient to distinguish all six stops reliably, with the highest test recognition rate of only 64% for this condition. Considering the cases of voiced stops or unvoiced stops separately, place of articulation can be identified with over 82% accuracy based on spectral shape versus approximately 50% based on the formant values or 73% based on formants and their amplitudes. For all conditions, spectral shape is much more effective for classifying the stops than formants alone. The improvement in recognition accuracy in adding the formant amplitudes to the formant frequencies, which thus adds information about global spectral shape, also lends support to the hypothesis that the shape of the spectrum carries the most information. Note, however, that the addition of bandwidths to the formant + amplitude parameter set did not improve the recognition rate. In any case, even the 82% and 84% rates obtained for voiced and unvoiced consonants, respectively, are far less than the rates possible by human listeners, leading to the conclusion that although the spectral shape of the burst onset carries information about place of articulation, this information is incomplete, at least insofar as automatic classification of stops is concerned. Although a perceptual experiment was not conducted specifically for speech stimuli consisting of a 25.6-ms frame, the perceptual results obtained from the 50-ms and 100-ms segments (conditions 4 and 5 from Table II) imply that the burst spectrum is also unlikely to contain complete information for perception.

To gain more insight into the role of spectral shape in cueing place of articulation, plots were made of the average spectra at burst onset for each of the six stop categories, as obtained from a DCTC 1–10 representation. These plots are shown in Fig. 5. These plots are averages of all tokens of each stimulus in our data base (i.e., approximately 400 tokens for each case). Because of this averaging over 30 speakers and 11 vowel contexts, these spectra are much more smoothed than would be a typical spectral plot computed from a single token. Also note that the amplitude scale is logarithmic and the frequency scale is for bilinear frequency warping. These average spectra vary considerably as place of articulation changes, but are very similar for the same place of articulation (i.e., /b,p/ form a pair, /d,t/ form a pair, and /g,k/ form a pair). Thus these plots are consistent with our classification results that suggest place of articulation can be distinguished reasonably well from the burst spectra, but the voicing feature cannot be distinguished from the burst spectra.

C. Classification experiments based on dynamic spectra

The experiments reported in the previous section indicated that global spectral shape shows promise for cueing stop consonant identity, but that the global spectral shape derived from a single frame is insufficient. Thus several experiments were conducted to identify features from several frames of speech data beginning at the burst onset. The ob-

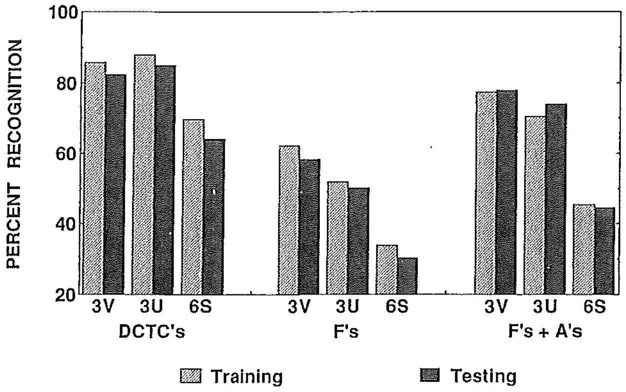

FIG. 4. Summary of automatic recognition results from one static spectrum for voiced stops (3V), unvoiced stops (3U), and all six stops (6S) for each of three parameter sets.

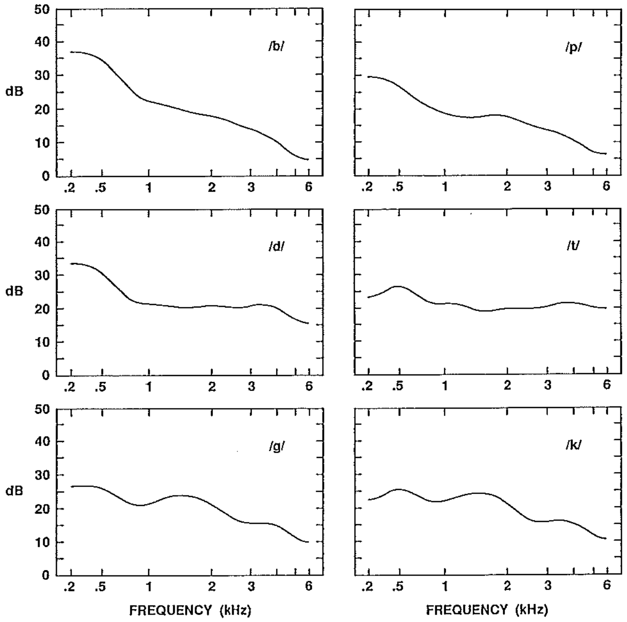

FIG. 5. The average spectra at burst onset for each of the six initial stops as obtained from a DCTC 1–10 representation, averaged over 11 vowel contexts and 30 speakers.

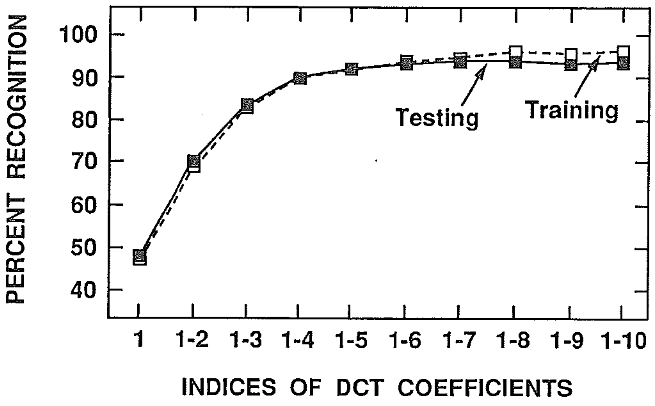

FIG. 6. Automatic classification accuracy for the six initial stops as a function of the number of DCT coefficients for dynamic spectra spanning the first 60 ms of each stimulus.

jective of the first experiment was to examine once again the effects of the level of spectral detail on classification accuracy for the six stops, for the case of a dynamic spectral shape representation. To investigate this issue, 10 DCT coefficients were computed for each of 11 frames spanning the first 60 ms of each stimulus, with a frame size of 15 ms and frame spacing of 5 ms. Each of these 10 DCT coefficients was then expanded with a 3-term cosine expansion over those 11 frames. The three coefficients of the cosine expansion of each DCT coefficient were used as the feature vector for the classifier to represent that DCT coefficient. The automatic classification accuracy was evaluated for the six stops as a function of the number of DCT coefficients used as input for the classifier. Figure 6 shows the results. The recognition rate increases until DCTC7 is added, but then levels off, or even decreases slightly. These results imply that the first seven DCT coefficients are sufficient for representing the spectrum of each speech frame. The implication is that the highly smoothed global shape of the spectrum encodes most of the information required for stop identification. For the case of dynamic spectral shape, the time trajectory of DCTC1, representative of signal level, does improve the recognition rate by about 1.3%. The discrepancy between the effect of DCTC1 for the static and dynamic features implies that the temporal changes in DCTC1, not its average level, contain the stop information. Thus the results depicted in Fig. 6, which are similar to those given in Fig. 3 for the static spectrum, imply that DCT coefficients 1–7 are sufficient for encoding each frame of the time-varying spectrum for use in automatic classification of initial stops.

Additional testing investigated the effects of window length and frame spacing on recognition accuracy. Neither window length nor frame spacing were found to affect this accuracy significantly. However, of the values tested (frame spacings of 2.5, 5, and 10 ms; window lengths of 10, 15, and 20 ms) a frame spacing of 5 ms and a frame length of 15 ms gave the best results (by about 1% over the worst case) and were thus used in later experiments.

Another series of automatic classification experiments was conducted to determine the approximate interval, measured from the beginning of the burst of each stimulus, over which the dynamic features should be extracted to classify initial stops. In these tests, dynamic features were extracted from 20-, 30-, 40-, 50-, 60-, 75-, and 90-ms intervals and classification results were computed for each of these intervals. Each classification test was performed for the three voiced stops, for the three unvoiced stops, and for the six stops combined. The series was repeated for DCTC's and for formants plus amplitudes.[5] In all cases and for each time interval, the parameters were encoded with a three-term cosine basis vector expansion over time. Figure 7 depicts the results of these tests.

Figure 7 shows that test results for both the DCTC's and

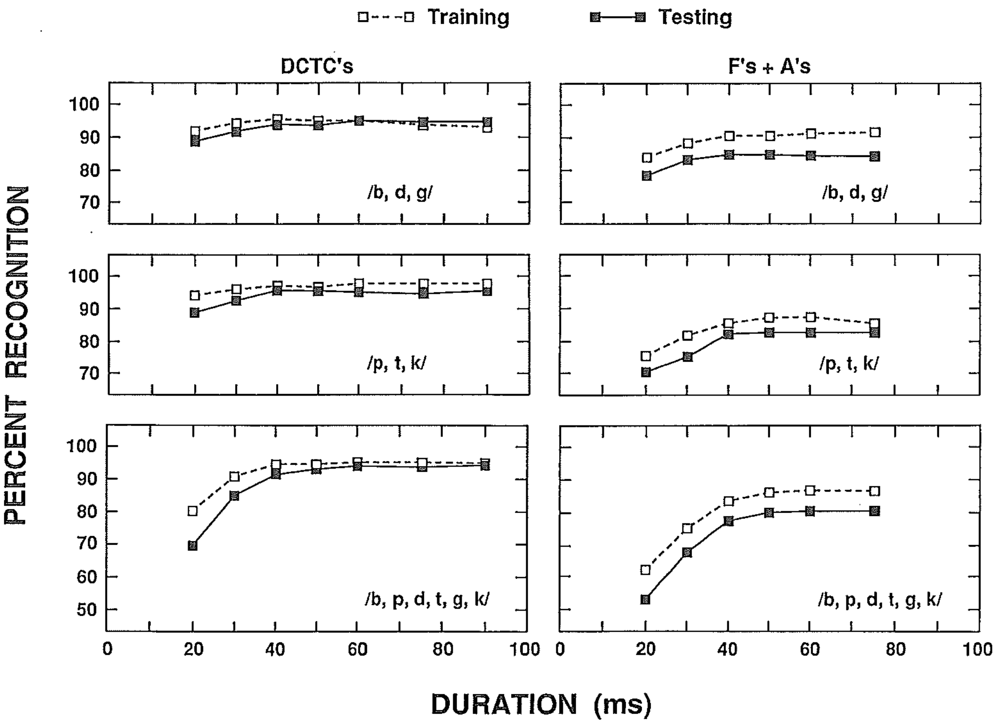

FIG. 7. Automatic classification accuracy for the three voiced stops, for the three unvoiced stops, and for all six initial stops as a function of the time interval used to extract the dynamic features. The dynamic features are the smoothed DCT coefficient trajectories for the left-hand panels and formant and amplitude trajectories for the right-hand panels.

formants plus amplitudes increase as the interval used for feature extraction increases. However, the results are consistently lower for the formant trajectories versus the DCTC trajectories. The recognition rates increase for time intervals up to 60 ms, for the case of the six stops and the three voiced stops. However, for unvoiced stops test results increase only until the interval reaches a length of 40 ms. Although the differences in the recognition rates at these two intervals for the voiced and unvoiced stops were not significant at the 99% confidence level, the indication that voiced stops require a longer time interval for reliable automatic identification than do the unvoiced stops is consistent with other sources of evidence. For example, the listeners required a longer portion of the signal to identify voiced stops than unvoiced stops. The listeners identified initial voiced stops with 86.4% accuracy from the first 50 ms of each stimulus and with 96.1% accuracy from the first 100 ms. In contrast, the listeners could identify 93.5% of the unvoiced stops correctly from a 50-ms interval and 97.0% of the unvoiced stops from a 100-ms interval. These results imply that cues for the unvoiced stops are contained in a shorter time interval than for the voiced stops. These results are also in agreement with perceptual results reported by other investigators. For example, Tekieli and Cullinan (1979) reported that the average signal interval required by listeners to identify initial voiced stops is 34 ms versus 17 ms for unvoiced stops.[6]

Figure 8, in bar graph form, summarizes the automatic classification results obtained with DCTC trajectories, formant trajectories, and formant plus amplitude trajectories, for the 60-ms interval beginning with the burst onset. The figure shows that for each condition, the highest recognition rates correspond to DCTC's, followed by formants + amplitudes, followed by formants alone.

Additional tests investigated the role of the initial transition interval, without the burst, in supplying cues for initial stops. Therefore all the classification tests used for the results shown in Fig. 8 were repeated with identical signal processing, except that the features were timed to begin with the onset of voicing in the initial transition rather than the burst onset. The results of this experiment are given in bar graph form in Fig. 9. Comparing the results given in Fig. 9 with the results given in Fig. 8, for every condition and for each feature set, the identification of initial stops significantly decreases when the features are extracted from a signal timed to begin at the start of the initial transition rather than at the start of the burst. For example, the recognition rate of the six stops using DCT coefficients extracted from a signal starting at the first voicing pulse is 55.3% compared to 93.7% if the burst is included. Even for the case of the three voiced stops, and with DCT coefficients as the parameter set, the recognition rate drops to 81% if the burst is not included versus 95% if the burst is included. These results show that the vowel transition regions alone do not contain sufficient acoustic cues to identify initial stops reliably. These results are also in agreement with the perceptual results obtained from our listening experiment. The listeners reported 74.5% identification for initial voiced stops for stimuli that started from the first voicing pulse and ended with the end of the vowel. Therefore, these results imply that the burst section is essential for reliable identification of initial stops. Another point to be noted from Fig. 9 is that identification of unvoiced stops based on the vowel transition is only slightly above chance (46% vs 33% for chance), indicating that the vocalic transitions, at least those occurring after the onset of voicing, carry almost no information regarding the identity of unvoiced stops.

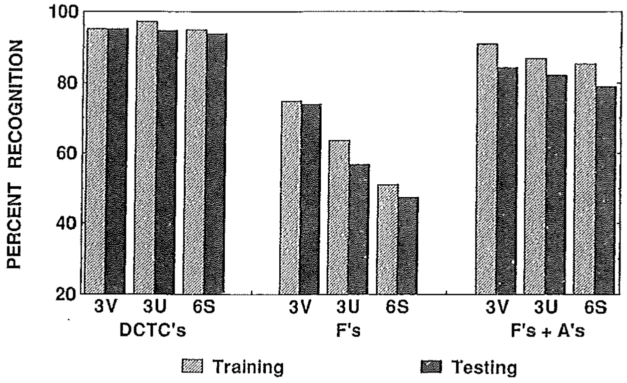

FIG. 8. Summary of automatic recognition results obtained from dynamic spectra, timed to begin with the onset of the burst.

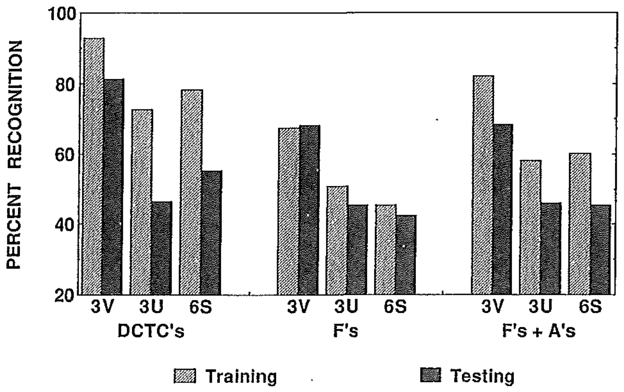

FIG. 9. Summary of automatic recognition results obtained from dynamic spectra, timed to begin with the beginning of the vowel transition.

In summary, the best automatic classification results of initial stops are obtained from the smoothed spectral shape features (DCT coefficient trajectories) extracted from an interval approximately 60 ms in duration and beginning with the burst onset. As Fig. 8 shows, the recognition results based on formants are very low compared with results obtained from DCT coefficients for every condition tested. Although the feature set consisting of formants plus amplitudes is much more effective than formants alone, the DCTC's are much better still. Thus the experiments of this section suggest that changes in the global shape of the spectrum over a 60-ms interval beginning at the burst are very effective for cueing both the place of articulation and the voicing feature in initial stop consonants.

Figure 10 depicts smoothed spectra over a time-frequency plane for each of the six stop consonants. The plots shown were obtained by first smoothing the spectra with DCTC's one through seven in frequency and then smoothing each of

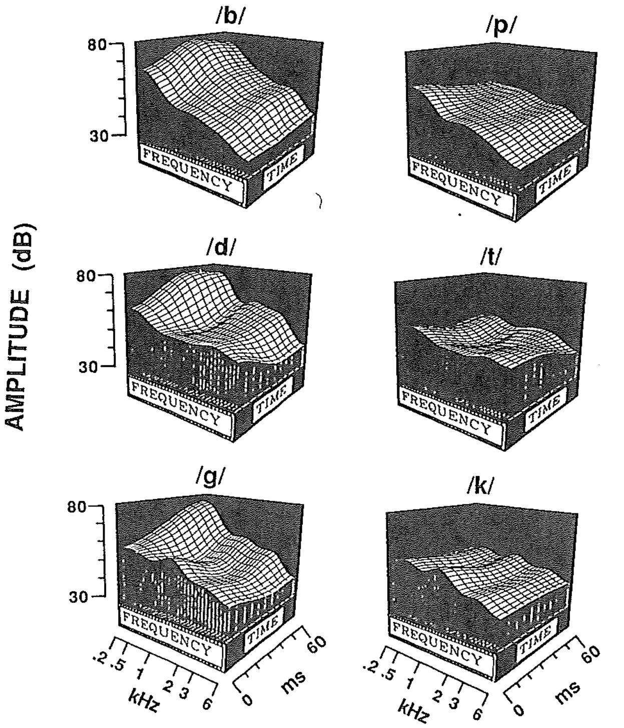

FIG. 10. Smoothed spectra over a time-frequency plane for the six initial stops.

these DCTC's with a three-term cosine expansion over time. Additional smoothing occurs, as for the burst spectra depicted in Fig. 5, because of averaging over all 30 speakers and all 11 vowel contexts. The plots span a frequency range of 200–6000 Hz and a time interval of 0–60 ms beginning with the burst onset. Inspection of these plots reveals systematic differences among the six stop categories. The voiced consonants are clearly differentiated from the unvoiced consonants in that the low-frequency energy for the voiced stops increases with time whereas the low-frequency energy for the unvoiced stops remains constant or decreases. In fact, we found that the addition of VOT as an extra feature parameter did not improve classification accuracy. The bilabials are characterized by a smooth surface in the time-frequency plane with a negative spectral tilt, particularly for the low-time frames. The alveolars are characterized by a nearly flat onset spectra and decreasing high-frequency energy as time increases. The velars are characterized by three broad spectral peaks, with the midfrequency one the most prominent. Note that these spectral features are consistent with the dynamic features described by Kewley-Port (1983).

IV. DISCUSSION

Our experiments show that the dynamic spectral shape features (DCT coefficient trajectories) extracted from the first 60 ms beginning with the burst are very effective for encoding consonantal information. Using these dynamic spectral shape features we were able to identify initial stops nearly as accurately with an automatic classifier as could human listeners. The automatic classification rate of 93.7% derived from these dynamic spectral shape features was higher than the perceptual rate for the 50-ms condition (89.9%) and lower than the perceptual rate for the 100-ms condition (96.6%). These automatic classification results were obtained using a 20-dimensional feature vector, corresponding to the coefficients of a 3-term cosine expansion over time for DCTC's one through six plus the coefficients of a 2-term cosine expansion for DCTC7.[7] Presumably the automatic classification rates did not improve in our experiments for time intervals greater than 60 ms since the amount of training data was insufficient to determine reliable statistical estimates for the feature spaces required to encode these longer intervals.

We performed a t test for the recognition results obtained for the six initial stops using the data of the 15 talkers used for testing. This test indicated that the 95% confidence interval for our recognition accuracy results is $+ - 2.1\%$. This test was conducted only for the case of dynamic spectral shape features extracted from the first 60 ms of each stimulus. However, similar results would be expected for other cases. The results of this t test imply that if our experiments were repeated with other test talkers, the recognition results would generally be within 2.1% of the results obtained using the 15 test talkers used in our study. Therefore test classification accuracy differences of 3% or more are statistically significant.

In an attempt to derive a lower dimensionality feature space, we also used linear discriminant analysis to reduce our 20 features to five maximally discriminating features. The automatic classification rate based on two discriminant scores was 71.6%, and the rate based on five discriminant scores was 86%. Thus these low dimensionality spaces are not sufficient for reliable identification of initial stops. We also examined the relative importance of the 20 features through a feature ranking algorithm similar to the one described by Cheung (1978). Table VII lists the ten features ranked highest by this feature ranking program, in terms of their contribution to recognition rate on test data [refer to Eq. (3) for notation]. The table shows that the three most discriminating features for the 60-ms interval are the average value of DCTC2 (the spectral tilt), the temporal slope of DCTC3 (a measure of changes in spectral compactness), and the temporal slope of the spectral tilt (DCTC2). Eight

TABLE VII. Ten highest ranking features selected by a feature ranking algorithm as contributing the most to automatic stop identification for test talkers. These 10 features were selected from 20 dynamic spectral shape features. The test recognition rate based on all 20 features was 93.7%.

Index	% testing	% increase	Features
1	43.1	43.1	C_1 of DCTC2
2	57.4	14.3	C_2 of DCTC3
3	67.9	10.5	C_2 of DCTC2
4	75.5	7.6	C_1 of DCTC3
5	79.2	3.7	C_2 of DCTC4
6	82.8	3.6	C_1 of DCTC4
7	85.5	2.7	C_2 of DCTC3
8	86.7	1.2	C_1 of DCTC6
9	88.2	1.5	C_1 of DCTC1
10	89.9	1.7	C_1 of DCTC5

out of the 10 top-ranked features, from which 89.9% identification of test data was obtained, consist of the first and second time cosine basis vector coefficients used to encode the spectral trajectories of the first five DCTC's.

All of our experiments suggest that dynamic spectral shape contains sufficient cues for reliable automatic classification of initial stops. Since we did not perform a perceptual test with synthetic stimuli, with conflicting cues indicated by dynamic spectral shape versus those supplied with alternative features (such as formants), we did not directly test the perceptual significance of dynamic spectral shape. Comparisons of stimuli locations in our 20-dimensional feature space with stimuli locations in a 2-dimensional perceptual space obtained from multidimensional scaling of the listening experiment data were also not feasible because of the great disparities in the dimensionality of the two spaces. However, automatic classification results based on dynamic spectral shape features and the BML classifier are generally consistent with perceptual results. For example, as mentioned previously, the identification rates for both human listeners and our automatic classifier are much better if the burst plus vocalic transition is used versus the vocalic transition only. The identification rates for both listeners and the automatic classifier increase as the signal interval increases, up to a certain point. Both automatic and perceptual identification also require longer portions of the speech signal for voiced stops versus unvoiced stops.

To further test the hypothesis that automatic classification based on spectral shape trajectories closely parallels human perception of initial stops, we compared the perceptual rankings of the nine talkers used in the listening experiment with the automatic ranking for the same nine talkers. The perceptual rate was obtained by averaging the identification rate of all listeners for each talker for the condition in which listeners were presented with the first 50 ms of each stimulus. The automatic rate was obtained by training the classifier using 29 speakers and testing the classifier with data of the speaker in question. Table VIII gives the automatic identification rate and the perceptual identification rate for each of the nine speakers. The table shows that the ranking based on automatic classification matches the ranking from the listeners for the female and male speakers. For the children, the automatic rankings and perceptual rankings do not match for two of the three speakers. This mismatch for the case of children might be because of greater variability in speech productions for children relative to that of adults.

TABLE VIII. Automatic recognition rates versus the perceptual identification rates for the nine talkers used in the listening experiment. The automatic recognition rates were obtained using the coefficients of a three-term cosine expansion over time for each of DCTC's one through seven. DCTC's were extracted from the first 60 ms of each stimulus. The perceptual identification rates listed are for the IT-50 (50-ms) condition. The first three talkers listed are adult females, the second group of three are adult males, and the final three are children.

Talker I.D.	Perceptual rate	Automatic rate
F04	93.5	96.4
F02	93.0	92.6
F09	90.4	90.5
M06	90.9	95.1
M10	89.7	90.4
M05	83.8	88.1
C04	96.2	93.8
C02	89.7	78.2
C10	82.2	87.0

V. CONCLUSIONS

In this study, several aspects of acoustic cues for initial stop consonants were investigated. We compared global spectral shape features (DCTC's) versus formants, as acoustic cues for initial stops. We also compared static versus dynamic features and investigated the role of initial transitions as cues for initial stops. Our experiments indicate that the six initial stops can be automatically classified, independently of both vowel context and talker, with over 93% accuracy based on dynamic spectral shape features spanning a signal interval of approximately 60 ms beginning with the release of the burst. In contrast, features extracted from the initial transitions beginning at the onset of voicing, leaving off the burst, cannot reliably distinguish initial stops. The static spectral shape at the burst onset is sufficient to distinguish place of articulation with approximately 82% accuracy. Formant trajectories can be used to distinguish place of articulation for only about 73% of initial voiced stops and 56% of initial unvoiced stops. The relatively poor performance of formants is not surprising given all the difficulties with formant tracking in the burst and aspiration regions of the stop waveforms. However, even for the case of voiced stops with the burst removed, where the reliability of the formant tracking is presumably highest, the DCTC classification results are superior to the formant results by a wide margin (82% vs 69%). The primary conclusion from our experimental work is that the dynamic properties of spectral shape convey a great deal of information about both place of articulation and the voicing features for initial stop consonants. Although considerable additional effort could be devoted to more sophisticated formant tracking algorithms for use with stop waveforms, our classification results with dynamic global spectral shape imply that formants are not really required.

From our point of view acoustic correlates for initial stops can be viewed as either context independent or dependent, depending on terminology. The cues are context dependent in the sense that consonant cues are coarticulated with the vowel, even in the burst section of the waveform.[8] Therefore, low-dimensionality feature spaces are inadequate to represent these cues, if the vowel context is allowed to change. On the other hand, the cues are context independent in that there is no need to explicitly identify the vowel to recognize the consonant. That is, the consonants can be classified from high-dimensionality feature spaces that encode sufficient consonant and vowel information in an integrated fashion to allow consonant identification. We also found no evidence to indicate that the identification of onset of voicing, or determination of voice onset time, improves the automatic classification of stops. As noted previously, the addition of VOT to dynamic spectral shape features did not improve automatic classification of stops. The burst onset appears to be the critical timing point in the signal. Features extracted from the smoothed spectral shape over a fixed signal interval beginning with the burst are sufficient to enable reliable identification of the initial stops for an automatic classifier. Human listeners are also able to identify initial stops reliably from a short segment of the signal timed to begin with the burst.

We can at present make no claim that the particular dynamic global shape features used in our automatic classification experiments are the same or even closely parallel features used in perception. However, at least roughly speaking, the automatic classification results obtained from the dynamic spectral shape features are similar to perception results. The results of our study therefore support the conjecture that the features used for perception are derived from dynamic and highly smoothed spectral shape, as noted in previous studies (Kewley-Port, 1983). Such features might be more readily identifiable if the front-end spectral processing more closely approximated that performed by the human auditory system.

ACKNOWLEDGMENTS

This work was supported by Grant No. IRI-8702649 from the National Science Foundation.

[1] Although segmentation at zero crossings does not eliminate the possibility of spurious sound generation, no additional signal processing, such as tapered endpoints, was used for the segments.

[2] The careful reader will notice that the high-frequency preemphasis is different for the formant processing versus the DCTC processing. The preemphasis values used were selected from values used in the literature for other similar experiments (Lee, 1989; Talkin, 1987), and also from our own pilot experiments for the two types of processing. The basic reason for the difference in the pre-emphasis coefficient for the two cases is the difference in the sampling rates (8 vs 16 kHz). Therefore, a larger value of the pre-emphasis coefficient is required for the DCTC case in order that the effective analog pre-emphasis roughly match over the more important low-frequency range.

[3] In most experiments, the recognition results obtained with Legendre polynomials were identical to those obtained with least-squares polynomial curve fitting. The polynomial methods were equivalent to those described by Furui (1981, 1986b) for computing delta cepstrum coefficients for automatic speech recognition.

[4] Although amplitude equalization is not, of course, equivalent to loudness equalization, this gain normalization did prevent dramatic changes in loudness among the stimuli. In any case, subjects were not required to evaluate loudness—the amplitude normalization was performed because some of the speakers (primarily the children) varied considerably in loudness from token to token.

[5] Tests were also made with formants plus amplitudes plus bandwidths. However, since the classification results were no better than for formants plus amplitudes, the results of these experiments are not given.

[6] From our experiments, we could only speculate concerning the apparent discrepancy in signal interval required for voiced stop versus unvoiced stop identification. One conjecture was that voiced stops are more vowel dependent, even in the burst, than are unvoiced stops, and thus require a longer interval to include sufficient cues for stop identification. However, our experimental evidence did not support this hypothesis. In an automatic vowel identification experiment based on the burst (using a 25-ms Hamming window centered at the burst onset) 26.8% of vowels could be identified for the case of voiced stops versus 28.2% for the case of unvoiced stops (versus 9.1% for chance for 11 vowels).

[7] The third term of the cosine expansion for DCTC7 was not used because classification results degrade slightly if this term is used. The feature ranking experiment (results in Table VII) also showed that the C_3 terms for DCTC's 4–7 affected classification rates very little.

[8] The experimental results mentioned in footnote 6 show that the burst contains some vowel information. Visual inspection of many samples of the burst spectra also indicated that the vowel strongly influences the spectra.

[9] For all confusion matrices, the rows are stimuli intended by the talker. Columns are stimuli identified by the listeners.

Blumstein, S. E., and Stevens, K. N. (1980). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. 67, 648–662.

Blumstein, S. E., and Stevens, K. N. (1979). "Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. 66, 1001–1007.

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

Cole, R. A., and Scott, B. (1974a). "The phantom in the phoneme: Invariant cues for stop consonants," Percept. Psychophys. 15, 101–107.

Cole, R. A., and Scott, B. (1974b). "Toward a theory of speech perception," Psychol. Rev. 81, 348–374.

Delattre, P., Liberman, A. M., and Cooper, F. S. (1955). "Acoustic loci and transitional cues for consonants," J. Acoust. Soc. Am. 27, 769–773.

Dorman, M. F., Studdert-Kennedy, M., and Raphael, L. J. (1977). "Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues," Percept. Psychol. 22, 109–122.

Duda, R. O., and Hart, P. E. (1973). *Pattern classification and scene analysis* (Wiley, New York).

Fischer-Jorgensen, E. (1954). "Acoustic analysis of stop consonants," Miscellanea Phonetica 2, 42–49.

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

Furui, S. (1986b). "Speaker-independent isolated word recognition using dynamic features of speech spectrum," IEEE Trans. Acoust. Speech Signal Proc. 34, 52–59.

Furui, S. (1981). "Cepstral analysis techniques for automatic speaker verification," IEEE Trans. Acoust. Speech Signal Proc. 29, 254–272.

Kewley-Port, D., and Luce, P. A. (1984). "Time-varying features of initial stop consonants in auditory running spectra: A first report," Percept. Psychol. 35, 353–360.

Kewley-Port, D. (1983). "Time-varying features as correlates of place of articulation in stop consonants," J. Acoust. Soc. Am. 73, 322–335.

Kewley-Port, D., Pisoni, D. B., and Studdert-Kennedy, M. (1983). "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoust. Soc. Am. 73, 1779–1793.

Lahiri, A., and Blumstein, S. E. (1981). "A reconsideration of acoustic invariance for place of articulation in stop consonants: Evidence from cross-language studies," J. Acoust. Soc. Am. Suppl. 1 70, S39.

Lamel, L. F. (1987). "Identification of stop consonants from continuous speech in limited context," J. Acoust. Soc. Am. Suppl. 1 82, S80.

Lee, K. F, and Hon, H. W. (1989). "Speaker-independent phone recognition using Hidden Markov Models," IEEE Trans. ASSP-37, 1641–1648.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). "Perception of the speech code," Psychol. Rev. 74, 431–461.

Liberman, A. M., Delattre, P. C., Cooper, F. S., and Gerstman, L. J. (1954). "The role of consonant-vowel transitions in the perception of the stop and nasal consonants," Psychol. Monogr.: Gen. Appl. 68, 1–13.

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

Ohde, R. N., and Stevens, K. N. (1984). "Revisiting stop-consonant perception for two-formant stimuli," J. Acoust. Soc. Am. Suppl. 1, 75, S66.

Oppenheim, A. V., and Johnson, D. H. (1972). "Discrete representation of signals," Proc. IEEE 60(6), 681–691.

Rossen, M. L., Niles, L. T., Tajchman, G. N., Bush, M. A., Anderson, J. A., and Blumstein, S. E. (1988). "A connectionist model for consonant-vowel syllable recognition," ICASSP-88, 59–62.

Schouten, M. E. H., and Pols, L. C. W. (1983). "Perception of plosive consonants—The relative contributions of bursts and vocalic transitions," in *Sound Structures: Studies for Antonie Cohen*, edited by M. P. R. van den Broecke, V. J. J. P. van Heuven, and W. Zonneveld (Foris, Dordrecht, The Netherlands), pp. 227–243.

Searle, C. L., Jacobson, J. Z., and Raymond, S. G. (1979). "Stop consonant discrimination based on human audition," J. Acoust. Soc. Am. 5, 799–809.

Stevens, K. N., and Blumstein, S. E. (1978). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. 64, 1358–1368.

Stevens, K. N., and Blumstein, S. E. (1981). "The search for invariant acoustic correlates of phonetic features," in *Perspectives on the Study of*

Speech, edited by P. D. Eimas and J. Miller (Erlbaum, Hillsdale, N.J.), pp. 1–38.

Suomi, K. (**1985**). "The vowel-dependence of gross spectral cues to place of articulation of stop consonants in CV syllables," J. Phon. **13**, 267–285.

Talkin, D. (**1987**). "Speech formant trajectory estimation using dynamic programming with modulated transition costs," J. Acoust. Soc. Am. Suppl. 1 **82**, S55.

Tanaka, K. (**1981**). "A parametric representation and clustering method for phoneme recognition—Application to stops in a CV environment," IEEE Trans. Acoust. Speech, Signal Proc. **29**, 1117–1127.

Tekieli, M. E., and Cullinan, W. L. (**1979**). "The perception of temporally segmented vowels and consonant-vowel syllables," J. Speech Hear. Res. **22**, 103–121.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., and Lang, K. (**1989**). "Phoneme recognition using time delay neural networks," IEEE Trans. **ASSP 37**, 328–339.

Yoder, K. S., and Jamieson, L. H. (**1987**). "Speaker-independent recognition of stop consonants," **ICASSP-87**, 864–867.