DYNAMIC SPECTRAL SHAPE FEATURES FOR SPEAKER-INDEPENDENT AUTOMATIC RECOGNITION OF STOP CONSONANTS

Stephen A. Zahorian and Zaki B. Nossair

Department of Electrical and Computer Engineering
Old Dominion University
Norfolk, Virginia 23529

ABSTRACT

In this study several acoustic feature sets and automatic classifiers were investigated to determine a combination of features and classifier which would enable accurate bottom-up speaker and vowel independent automatic recognition of initial stop consonants in English. The features evaluated included a form of cepstral coefficients and formants, each computed both for one static frame and as spectral trajectories over various segments of the speech signal. The classifiers investigated included Bayesian Maximum Likelihood (BML), artificial neural network (NN), and K Nearest Neighbor (KNN) classifiers. The most accurate results, over 93% of the six stops correctly identified with a speaker-independent classifier, were obtained with the BML classifier using cepstral coefficient trajectories as a 20-dimensional feature vector. These results for stop recognition are higher than any results previously reported for a data base of similar diversity.

INTRODUCTION

Many workers in the field of automatic speech recognition have attempted to devise signal processing schemes for extracting features for automatic identification of stops [7, 8, and 9]. To date none of the speaker-independent automatic recognition schemes have performed nearly as well as human listeners in identifying stops. Lamel et al. [4] 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. The objectives of our study were to use automatic classification experiments to help define a set of acoustic features which encode information sufficient to reliably distinguish the initial stop consonants for speaker-independent ASR applications. We explored in detail features based on overall spectral shape versus features based on spectral peaks (formants) and compared features based on one static spectrum sampled at the release burst versus features based on dynamic spectra.

(Paper presented at ICASSP-90, April 3-6, 1990, Albuquerque, NM)

DATA BASE

Ninety-nine CVC isolated tokens were recorded for each of 30 native English talkers (12-bit 16 kHz A-D). Ten of these talkers were adult males, ten were adult females, and ten were children between the ages of 7 and 11. Eighty-four of these syllables began with one of the six stop consonants /b,p,d,t,g,k/, the focus of this study, and the other 15 syllables began with one of the four consonants /h,l,m,w/ which were needed for other experiments conducted with the vowel portions of the syllables. The vowel in each syllable was one of the eleven vowels /aa,iy,uw,er,ih,ae,eh,ao,ah,uh,ow/ and the final consonant was one of the 9 consonants /b,d,g,k,t,p,v,s,h/. Each initial stop was paired with at least one instance of each of the eleven vowels, i.e. each initial stop was spoken in eleven vowel contexts. The acoustic regions of all speech stimuli were manually labeled (burst onset, beginning of vowel transition, and beginning of steady-state vowel) for use with the automatic classification routines.

SPEECH PARAMETERS

The two feature sets investigated were formants and a form of cepstral coefficients. Since the cepstral coefficients were computed somewhat differently than the usual method, 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.

Formants

Formants were computed for the initial stops in a multi-stage process as follows. The speech signal was first digitally lowpass filtered at 3.8 kHz with a 49th order FIR linear-phase lowpass filter and resampled at 8 kHz. The speech signal was then high-frequency preemphasized with transfer function ($1-.75\ z^{-1}$). 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 5 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, 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 were used to represent that stimulus. In addition to 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.

Cepstral coefficients

The cepstral coefficients, i.e., the DCTC's, were computed as follows. First, the speech signal was high-frequency preemphasized with transfer function ($1-.95\ z^{-1}$). The speech signal was then windowed using a Hamming window. Depending on the length of the window, either a 256 or 512 FFT was computed for each speech frame. 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 then defined as the a_n's in the equation

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

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,

$$P(n) = \sum_{k=1}^{N} C_k \cos((k-1)\,\pi\, n/(L-1)), \qquad (2)$$

where P(n), $1 \le n \le L$, 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, in general the best results were obtained with N = 3.

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

$$D_i(x) = (x-\mathbf{x}_i)^T \mathbf{R}_i^{-1} (x-\mathbf{x}_i) + \ln|\mathbf{R}_i| - 2 \ln P(G_i), \qquad (3)$$

$$1 \le i \le M,$$

is minimized. In Eq. (3) $\mathbf{x}_i$ is the centroid for category i, $\mathbf{R}_i$ is the covariance matrix for category i, and $P(G_i)$ is the a priori probability for category 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 multi-variate Gaussian [2].

In addition to the BML classifier experiments were also performed with a two-layer feedforward perceptron-like neural network (NN) classifier [5] and a K nearest-neighbor (KNN) classifier [2]. These alternate classifiers were used to verify that the relative rankings of feature sets determined via the BML classifier were also obtained with alternate classifiers, and were not merely specific to one particular classifier.

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

EXPERIMENTS

Classification experiments based on burst spectra

Our first series of automatic classification experiments was conducted to optimize the DCTC computations 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. Experiments were conducted to evaluate various nonlinear amplitude scales, nonlinear frequency scales, frequency ranges, and the number of DCTC's used as features, in terms of their effect on automatic classification accuracy of initial stops. Based on these experiments a log amplitude scaling, a bilinear frequency warping [6] with a coefficient of .5, and a frequency range of 200 to 6000 Hz, were selected.

DCT coefficients 2-10, computed as outlined above, were used as 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 + amplitudes) for each of three conditions: (1), all six stops (6S); (2), voiced stops only (3V); and (3), unvoiced stops only (3U). For the case of all six stops both the place and voicing features must be distinguished whereas for conditions 2 and 3 only place of articulation must be determined. Figure 1 summarizes the training and test results for these conditions and these parameter sets, all based on the BML classifier.

The results given in Fig. 1 indicate that overall 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 reliably distinguish all six stops, with the highest test recognition rate of only 64% for this condition. For the cases of voiced stops or unvoiced stops considered individually, 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 far 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 overall 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.

Classification experiments based on dynamic spectra

The experiments reported in the previous section indicated that overall spectral shape shows promise for cueing stop consonant identity, but that the global spectral shape derived from a single frame is insufficient. Thus a number of experiments were conducted to identify features from several frames of speech data beginning at the burst onset. One series of automatic classification experiments was conducted to determine the approximate time 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 a classification experiment was performed for each of these intervals. The series was repeated for DCTC's and for formants + amplitudes. In all cases and for each time interval, the parameters for each frame were encoded with a three-term cosine basis vector expansion over time. A BML classifier was used. The results of these tests indicated that test results for both the DCTC's and formants + amplitudes, increase as the time interval used for feature extraction increases up to 60 ms. However, the results were consistently lower for the formant trajectories versus the DCTC trajectories.

Figure 2 summarizes the automatic classification results obtained with DCTC trajectories, formant trajectories, or formant and amplitude trajectories, for the 60 ms interval beginning with the burst onset. The figure shows that for each condition, the highest recognition rates are obtained with DCTC's, followed by formants + amplitudes, followed by formants alone.

Additional tests were performed to investigate 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. 2 were repeated with identical signal processing, except that the features were timed to begin with the initial transition rather than the burst onset. The results of this experiment are given in Fig. 3. Comparing the results given in Fig. 3 with the results given in Fig. 2, for every condition and for each feature set, we can see that the identification of initial stops significantly decreases when the features are extracted from a time interval 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 time interval 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. Therefore, these results indicate that the burst section is essential for reliable identification of initial stops.

Alternate classifiers

The results of a comparison of the BML classifier with the KNN and NN classifiers mentioned previously are given in Fig. 4. The KNN was implemented for K = 1, since this gave the best results. The NN classifier was implemented as a two-layer feedforward perceptron structure with 20 hidden nodes. The tests were performed with spectral trajectories extracted from the 60 ms interval beginning with the burst. For each classifier, the figure shows that the best recognition rates were obtained with DCTC parameters followed by formants + amplitudes followed by formants alone. In general the BML classifier results in the highest recognition rates. For the case of the formants, however, the NN classifier performs better than the BML classifier, undoubtedly because the formants violate the multivariate Gaussian assumption of the BML classifier. Nevertheless, in summary, the highest overall recognition rates were achieved with DCTC trajectories and the BML classifier.

CONCLUSIONS

In this study, we compared two different feature sets, 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 in a speaker-independent manner with over 93% accuracy based on dynamic spectral shape features spanning a time interval of approximately 60 ms beginning with the release of the burst. The error rate increases dramatically if features are extracted from a single 25 ms frame, if formants and amplitudes are used as parameters rather than DCTC's, or if the burst is missing from the interval sampled. The error rate for the test speakers also increased 22% if the DCTC coefficients were computed without frequency warping and over the full frequency range. These results were obtained with broad range of speakers (men, women, and children) and with a large number of vowel contexts (11). We determined that a 20-dimensional feature vector, representing the smoothed DCTC trajectories was sufficient to account for both vowel context and speaker effects. Thus the dynamic properties of smoothed spectral shape convey a great deal of information about both place of articulation and the voicing features for initial stop consonants. The recognition results obtained in our study are higher than any results reported in the literature for speaker and vowel independent initial stop recognition. We believe the techniques used in our work could be applied to slightly longer signal segments and a much larger data base to obtain even higher automatic recognition rates for initial stops.

ACKNOWLEDGMENTS

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

REFERENCES

[1] 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.

[2] Duda, R. O. and Hart, P. E. (1973). Pattern Classification and scene analysis, (John Wiley & Sons, New York).

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

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

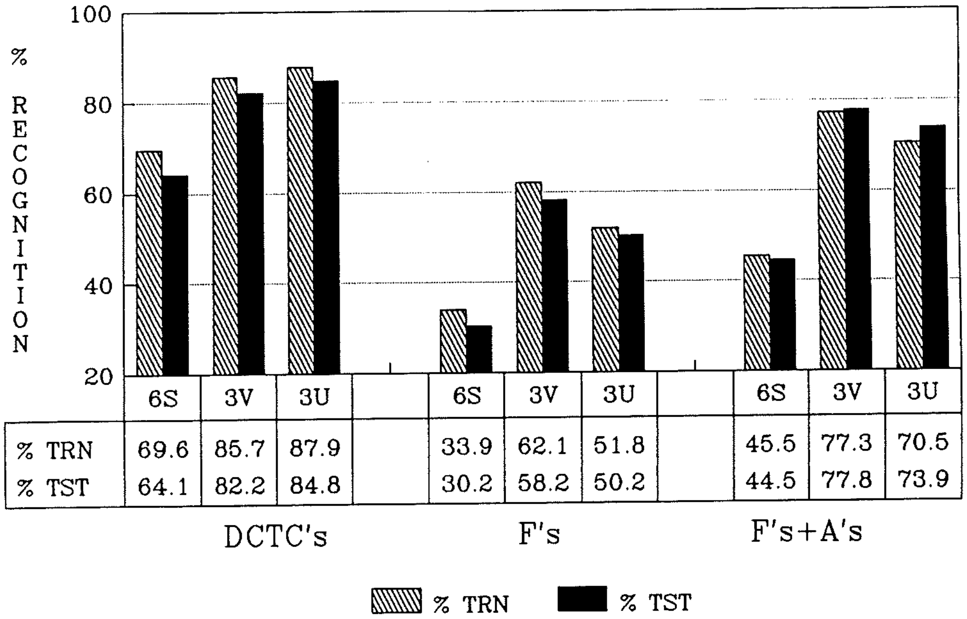
[5] Lippmann, R. P. (1987). "An introduction to computing with neural nets," IEEE ASSP magazine, April, 4-22.

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

[7] 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.

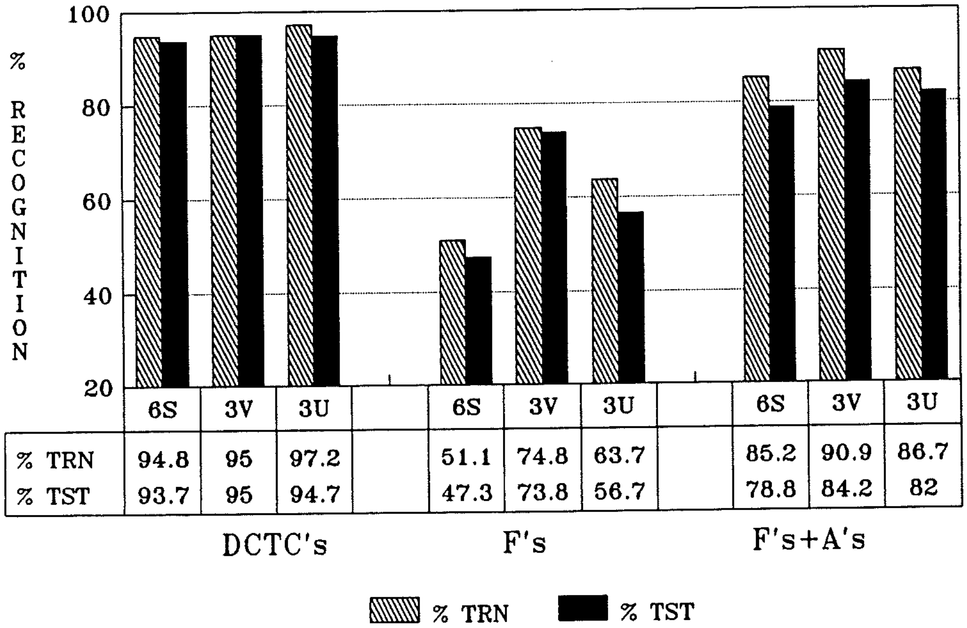
[8] Tanaka, K. (1981). "A parametric representation and clustering method for phoneme recognition--application to stops in a CV environment," IEEE Trans. Acoust., Speech, Signal Process. 29, 1117-1127.

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



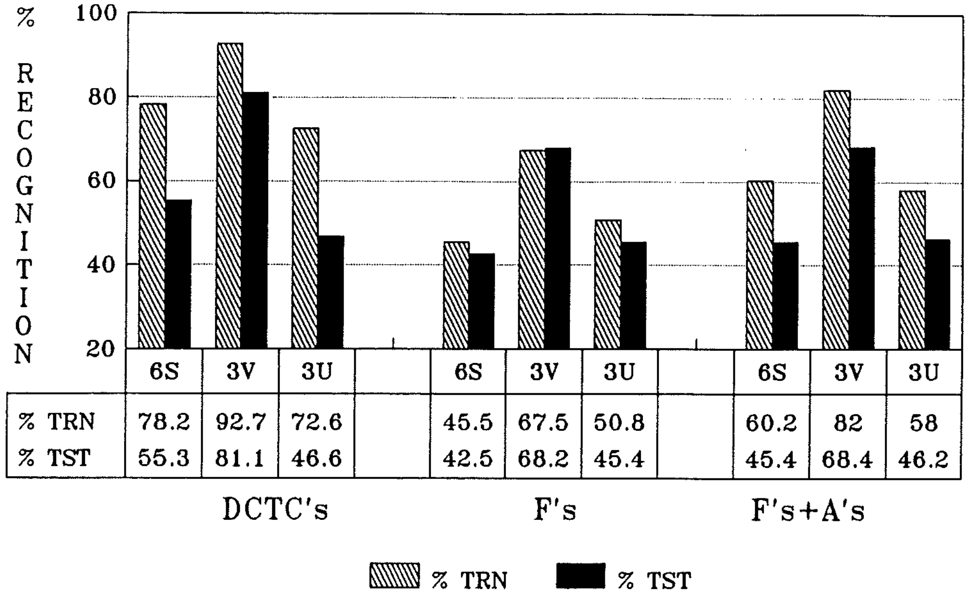
	6S	3V	3U		6S	3V	3U		6S	3V	3U
% TRN	69.6	85.7	87.9		33.9	62.1	51.8		45.5	77.3	70.5
% TST	64.1	82.2	84.8		30.2	58.2	50.2		44.5	77.8	73.9

Figure 1. Summary of automatic recognition results from one static spectrum for various conditions. These include six stops, unvoiced stops, and voiced stops, for each of three parameter sets.



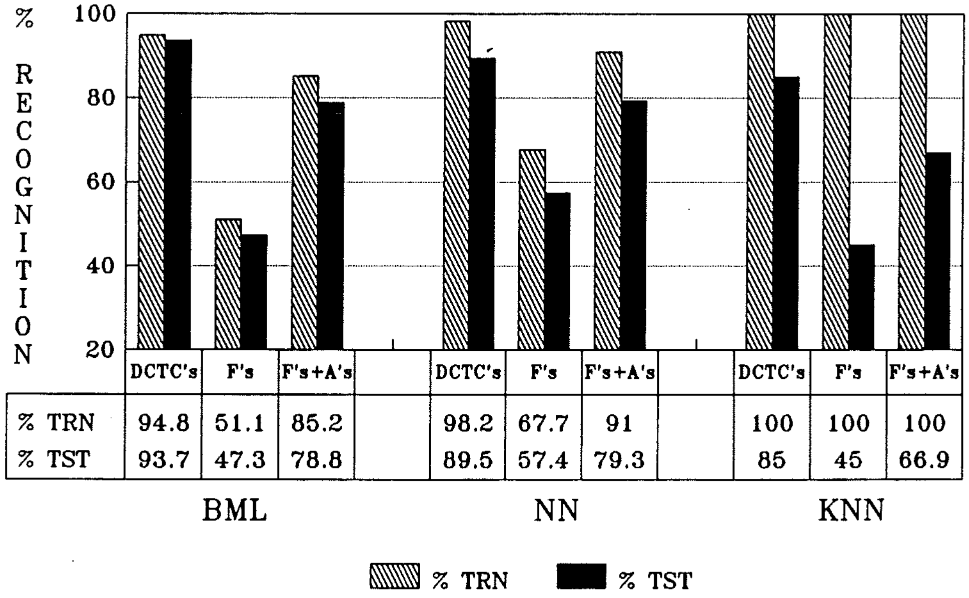
	6S	3V	3U		6S	3V	3U		6S	3V	3U
% TRN	94.8	95	97.2		51.1	74.8	63.7		85.2	90.9	86.7
% TST	93.7	95	94.7		47.3	73.8	56.7		78.8	84.2	82

Figure 2. Summary of automatic recognition results obtained from dynamic spectra for various conditions as noted.



	DCTC's			F's			F's+A's		
	6S	3V	3U	6S	3V	3U	6S	3V	3U
% TRN	78.2	92.7	72.6	45.5	67.5	50.8	60.2	82	58
% TST	55.3	81.1	46.6	42.5	68.2	45.4	45.4	68.4	46.2

Figure 3. Summary of automatic recognition results obtained from dynamic spectra, timed to begin with the beginning of the vowel transition.



	BML			NN			KNN		
	DCTC's	F's	F's+A's	DCTC's	F's	F's+A's	DCTC's	F's	F's+A's
% TRN	94.8	51.1	85.2	98.2	67.7	91	100	100	100
% TST	93.7	47.3	78.8	89.5	57.4	79.3	85	45	66.9

Figure 4. Automatic classification rates as a function of classifier type.