Recognition of Unaspirated Plosives—A Statistical Approach

A. K. DATTA, N. R. GANGULI, AND S. RAY

Abstract—In this paper the results of a study of the computer recognition of unaspirated plosives in commonly used polysyllabic words uttered by three different informants are presented. The onglide transitions of the first two formants and their durations have been found to be an effective set of features for the recognition of unaspirated plosives. The rates of transition of these two formants as a feature set have been found to be significantly inferior to the features mentioned earlier. The maximum likelihood method, under the assumption of a normal distribution for the feature set, provides an adequate tool for classification. The assumption of both intergroup and intragroup independence of the features reduces recognition scores. A prior knowledge of target vowels is found necessary for attaining reasonable efficiency. A prior knowledge of voicing manner improves classification efficiency to some extent. The physiological factors responsible for the variation of the recognition score for the various plosives are discussed. For labials and velars the recognition score is very high, nearly 90 percent. An attempt to correlate the dynamics of tongue-body motion with the variations in recognition scores has been made. Back vowels as targets have been found to give improved classification of the preceding consonants. A comparison of the result of machine recognition with those of published results on perception tests has been included. The results are found to be of the same order.

I. Introduction

AUTOMATIC speech recognition (ASR) occupies probably the most important role in the field of intelligence communication between man and machine. The ultimate goal of ASR is an automata which can extract the full intelligence content of speech and interpret the message contained in it for decision making and information transferring purposes. This complex problem has been investigated in depth in its various aspects, namely, the general classificatory and decision making aspects, as well as the particular problems of recognition of vowels, consonants, and other phonemes, both in isolation and in connected speech [1]-[15]. The variability of the features associated with human speech has prompted the use of statistical techniques in ASR particularly when the assumption of distribution functions does not pose a problem. The maximum likelihood method has been shown to be a potentially effective tool for such purposes [9]-[15].

Fundamentally, general ASR systems aim at the recognition of speech without any inherent limitations of vocabulary. In such a general approach it endeavors, at the primary level, identification of the smallest units of speech (phonemes) from the speech signal solely on the basis of the acoustic, phonetic and prosodic properties of these units. The reconstruction of semantic units from those phonemes using linguistic constraints is a task for the higher level ASR. This approach may be clearly distinguished from the specific ASR approaches which aim at the recognition of words or such other larger units over the field of a limited vocabulary. These approaches though highly restrictive are lucrative, exhibiting fairly good efficiency for very specific purposes. The efficiency, however, falls sharply as the size of vocabulary increases [8]. The present work has the approach of a general primary ASR system and intends to throw some light on the area of the machine recognition of unaspirated plosive consonants. These consonants have been taken from CV combinations, occurring in polysyllabic commonly used words. The present status of automatic vowel recognition is reasonably satisfactory [1]-[5]; the main emphasis here has been to investigate the machine recognition of unaspirated plosives on the assumption that the target vowels are known. A comparison of recognition scores when the target vowels are known with those when target vowels are unknown is also included. The selection of necessary acoustic, phonetic and prosodic features, their measurements and the nature of variation of these, have been considered in some detail along with the various classificatory methods employed.

The eight unaspirated plosive consonants of Telugu, a major Indian language, studied in the CV context of the ten major vowels are /k/, /g/, /ṭ/, /ḍ/, /t/, /d/, /p/, and /b/. The results of the classification of these consonants in various contexts and manners and their differential behaviors have been discussed with reference to the physiological factors involved.

II. Feature Selection and Measurement

The plosives are perceived and distinguished by their manner and place of articulation. The differentiation between various manners of articulation is a segmentation problem which has not been taken up here. The determination of the place of articulation of a plosive from the speech spectra is a difficult task. The acoustic cues for these sounds are supposed to be in the burst spectra, the aperiodic, as well as the vocalic transitions [16]-[26]. The principal source of information regarding acoustic cues for the identification of the place of articulation of plosives has been through experiments on perception. Various acoustic properties associated with these sounds are influenced by the following vowels. However, recent studies by Stevens *et al.* [23] suggest that cues for the place of articulation, which remain invariant under the influence of the fol-

Manuscript received September 15, 1978; revised February 15, 1979 and August 30, 1979.

The authors are with the Electronics and Communication Sciences Laboratory, Indian Statistical Institute, Calcutta, India.

lowing vowels, are to be found more in the spectrum of the onset of stimuli than in the continuing transitional dynamics of the speech wave. Though the formant transitions are not considered to be primary cues for the identification of articulatory position, results [23], [24] indicate that the aperiodic and vocalic transitions combined form an effective cue for this purpose once the target vowel is fixed. The unambiguous and accurate determination of the poles and zeros, the important characteristics of the place of articulation, in the burst spectra is extremely difficult because of the weak intensity and short duration (of the order of 5 to 10 ms) of the burst. On the other hand, a formant tracker can trace the transitory movements of the formants, particularly vocalic transitions of the first three formants. The present experiment proposes to study how effectively the onglide transitions of the first two formants may be used for automatic recognition of plosives according to their place of articulation. The basic features which have been considered here to represent the transition adequately for this purpose are the amount of transitions $[\Delta F]$ and the duration of transitions $[\Delta t]$.

The acoustic characteristics depicting the articulatory situation at the time of plosion would be best represented if the formant values were known at that time. The hypothesis that the transitional data can provide cues for the determination of the articulatory position of the plosives requires that these should adequately reflect the cavity characteristics closest to the time of plosion. It is, therefore, necessary that the initial values of the formants be obtained at the instant of release of the consonantal obstruction. Unfortunately, exact determination of the formant position just after the release is unclear and unambiguous for unaspirated plosives. The transition from plosion to the steady state of target vowels contains an initial aperiodic portion where the vibration of the vocal chords has not yet commenced. The energy content in this state is very weak and defies easy detection of the spectral structures. In the spectrograms also, this part is visible occasionally only when there is a slight aspiration. It is, therefore, necessary to resort to extrapolation of the obtainable formant-transition data. The tongue-body movement from after release to the time of the attainment of a reasonably steady state is very complex [25], [27]. The initial movement is very fast, which merges into a movement of larger time constant in the later stage. In the absence of dependable data on the dynamics of the tongue-body motion, a simple linear extrapolation has been used here to trace the formant transition back to the time of release (Fig. 1). Such measurements have been found to tally with the spectrographic data where traces of the aperiodic transition have been discernible.

The frequencies of the formants at the steady state as measured from the base line to the central line of the formant bands, where they are parallel to the base line, are subtracted from the values of the respective formants extrapolated to the time of the end of burst spectra. These values give the amount of formant transitions ΔF_1 and ΔF_2 (Fig. 1). The scale used for this measurement is derived from the callibrated tone of 500 Hz recorded on each and every spectrogram. The duration of transition Δt is measured from the end of burst spectra to the point where both the first two formants have attained a

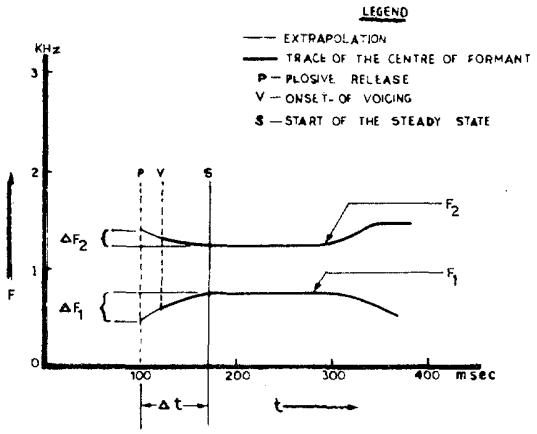


Fig. 1. Extrapolation of formants.

reasonably steady state. The time duration has been measured with a scale formed from the average of two time marker recordings, one at the beginning and one at the end of every group of 50 spectrograms. Throughout the recordings, no significant variation of these scales has been noticed.

Altogether, four different feature combinations have been tried from the three basic transitional features, namely, ΔF_1, ΔF_2 and Δt, the derived features like rates of transition $\Delta F_1/\Delta t$ and $\Delta F_2/\Delta t$, and the inverse of the duration of transition $1/\Delta t$. The feature combinations tested are $[\Delta F_1, \Delta F_2]$, $[\Delta F_1, \Delta F_2, \Delta t]$, $[\Delta F_1, \Delta F_2, 1/\Delta t]$, and $[\Delta F_1/\Delta t, \Delta F_2/\Delta t]$.

As statistical techniques with parametric representation have been used for classification, an examination of the nature of the distribution of the basic features is in order. The distribution of formant frequencies for the steady state of vowels has been reported to be normal [28]. Since the amount of transition is obtained from the difference of two values of the formant at two different instances, and as the physiological mechanism remains basically the same, its distribution also may reasonably be assumed normal. As the data for each CV combination are not sufficiently large to conduct a rigorous normality test, an indirect approach has been developed to test the normality of Δt. Descriptive measures of skewness $[\beta_1]$ and kurtosis $[\beta_2]$ [29] have been calculated for each of the features ΔF_1, ΔF_2, and Δt. These values have been compared with 0 and 3, the corresponding measures for a normal variable. The features are then ranked separately on the basis of β_1 and β_2 values, in each case the feature with the lowest difference being given the lowest rank. The two sets of ranks for β_1 and β_2 are then added, and combined ranks varying from 1 to 3 are then given to the features on the basis of their rank totals. In the case of the same rank total for more than one feature, they have been differentiated by looking into the individual differences of their $[\beta_1, \beta_2]$—values from $[0, 3]$. For each feature, the ranks obtained with different CV combinations have been averaged. They are presented in Table I.

As is evident from this table, the average rank of Δt is less than or at most equal to the larger of those of ΔF_1 and ΔF_2. Therefore, as ΔF_1 and ΔF_2 have already been taken to be normal, Δt can also be so considered.

TABLE I
AVERAGE RANKS OF DIFFERENT FEATURES

Feature	Plosive Type		
	Unvoiced	Voiced	Pooled
ΔF_1	1.82	1.95	2.04
ΔF_2	2.09	2.05	2.13
Δt	2.09	2.00	1.83

III. METHODS OF CLASSIFICATION

Let $x' = (x_1, x_2, \cdots, x_n)$ be an n-dimensional feature vector. Let x be multivariate normal with parameters μ_k and Σ_k, $k = 1, 2, \cdots, m$ where m is the number of groups. In the general maximum likelihood method of classification, under the assumption of equal *a priori* probability of the groups, x is assigned to that group k for which L_k defined by (1) is maximum [30].

$$L_k = -\tfrac{1}{2}\log_e|\Sigma_k| - \tfrac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k). \quad (1)$$

Altogether, three different classification methods have been used. In method 1 the above mentioned formula has been used without any restriction on the interdependence of the features. In method 2 the variations of the features are assumed to be group independent and therefore the dispersion matrices have been taken to be equal. Here L_k reduces to

$$L_k = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda^{ij}x_i\mu_{kj} - \tfrac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda^{ij}\mu_{ki}\mu_{kj} \quad (2)$$

where $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_m = \Sigma$, say, and $\Sigma^{-1} = (\lambda^{ij})$.

In the third method the features are assumed to be intragroup independent. The subregions containing different groups are also assumed to occupy equal volume in the feature space. In other words, the two assumptions made are 1) Σ_k is diagonal and 2) $|\Sigma_k|$ is constant. Here the problem reduces to assigning x to the group k, provided D_k defined by (3) is minimum.

$$D_k = \sum_{i=1}^{n}(x_i - \mu_{ki})^2/\sigma_{ki}^2 \quad (3)$$

where σ_{ki}^2 is the variance of the ith feature in the kth group. This method is known as the minimum-weighted distance method of classification [31].

IV. EXPERIMENTAL PROCEDURE

The present study has been conducted with eight unaspirated plosives namely /k/, /g/, /ṭ/, /ḍ/, /t/, /d/, /p/, and /b/. They have been selected in the CV combinations with ten major vowels both long and short. These vowels are /ə/, /a/, /e/, /e:/, /o/, /o:/, /u/, /u:/, /i/, /i:/. The coarticulation effect of distant vowels on consonants has been observed to be quite significant [32]. The CV combinations, therefore, have been taken from commonly used polysyllabic Telugu words. The six hundred words uttered by three male informants constitute the whole speech sample. The recordings of these words, spoken by ten native educated informants, have been made in an empty auditorium $12m \times 30m \times 6m$ size on TDK tapes with an AKAI 1710 tape recorder. The informants belonged to the age group of thirty to thirty-five years. The tapes were replayed before an audience of 12 native students of postgraduate and undergraduate classes. The three informants were chosen on the basis of highest recognition scores. The spectrograms have been made with a Kay Sonagraph having Sona-Marker facility. The selected bandwidth of the system has been 80 Hz to 8 KHz, with the filter bandwidth of 300 Hz. The formant transitions and the duration of transitions have been measured manually from the spectrograms.

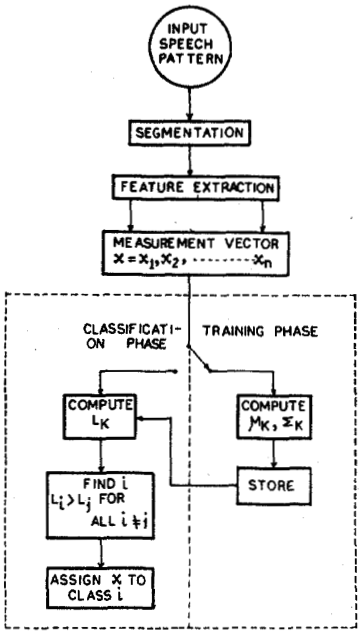


Fig. 2. Block diagram for plosive recognition.

Fig. 2 represents the present scheme in the context of a general ASR system. Segmentation, feature extraction, and formation of the measurement vector have been explained earlier. The classifier has two separate phases. In a nonadaptive system the input vector is switched to the training phase initially. After an adequate number of training samples are fed, the input vector is channeled to the classification phase. In the training phase the estimated values of the mean representative vector and the dispersion matrix for each class are computed and stored in permanent store. In the classification phase, for each input vector x, the discriminant scores L_k are first computed with the help of stored values of μ_k and Σ_k for all the m classes. The class number i, for which the score is a maximum, is then obtained through usual sorting procedures and the input vector is then assigned to that class.

V. DISCUSSION OF THE RESULTS

On the same basic transitional data, twelve classification experiments have been conducted with four feature combinations under each of the three methods of classification. Again, for each experiment, the data were grouped into a total of twenty-one subgroups, seven subgroups under each of the three main groups. The three main groups are unvoiced plosives [U], voiced plosives [V], and the unvoiced and voiced

TABLE II
AVERAGE RECOGNITION SCORES FOR DIFFERENT FEATURE COMBINATIONS BY DIFFERENT METHODS WITH TARGET VOWEL KNOWN

Feature Combination	Method 1 Maximum Likelihood Method [Diff. Disp. Matrices]	Method 2 Maximum Likelihood Method [Equal. Disp. Matrices]	Method 3 Minimum Weighted Distance Method
$\Delta F_1, \Delta F_2, \Delta t$	74.9	68.9	69.0
$\Delta F_1, \Delta F_2, 1/\Delta t$	73.9	68.5	68.9
$\Delta F_1, \Delta F_2$	67.2	61.6	63.8
$\Delta F_1/\Delta t, \Delta F_2/\Delta t$	67.0	62.6	63.6

mixed together [M]. The subgroups are one for each of six target vowels and one for all vowels taken together. The groups have been so formed that an assessment of the effect of prior knowledge of the voicing manner and the target vowels on the efficiency of recognition can be made.

Table II presents a summary of all these recognition results. It immediately reveals that the maximum likelihood method without any restrictive assumption using the feature set $[\Delta F_1, \Delta F_2, \Delta t]$ gives the best recognition results. Both the assumptions of intragroup independence and intergroup variational independence of the features lower the recognition score by about 6 percent. The last two classification procedures exhibit almost equal performance. The method of minimum distance, however, is the computationally simplest method. It may be of some interest where cost consideration is an important factor and one can afford this relatively small amount of loss in performance.

The performance of the feature set $[\Delta F_1, \Delta F_2, 1/\Delta t]$ is only marginally inferior to the set $[\Delta F_1, \Delta F_2, \Delta t]$, whereas those of the other two feature sets are significantly worse. This leads to the conclusion that all the three basic transitional parameters, namely, the amount and the duration of the transitions of formants, are important cues for the classification of the unaspirated plosives. The basic features in the original form represent these plosives better than the derived features, at least for the classificatory methods employed here. In the succeeding paragraphs the results will be discussed only in terms of method 1 and the feature set B, $[\Delta F_1, \Delta F_2, \Delta t]$, for which the recognition scores have been consistently higher.

Fig. 3 shows that the recognition score improves significantly if one knows the succeeding vowels beforehand. In fact, for method 1 and feature set B, the improvement is of the order of 60 percent. This strongly suggests that the transitional data are useful only with *a priori* knowledge of the target vowels. Further, the prior knowledge of voicing manner improves the classification score by about 10 percent.

Fig. 4 presents the recognition scores of the various plosives in different contexts. It may be noted that for both the extreme points of articulation, namely velar and labial, the recognition scores are significantly higher than the other two points. In fact, for voiced labials and velars, the average recognition score has reached the 90 percent level. Fully correct classification has been observed for 10 out of 48 CV combinations. High recognition scores for bilabials may be due to the lesser degree of involvement of the tongue in the articulation of these plosives. The tongue, therefore, has a larger time to settle for the target vowels. Fig. 5, drawn from the results of cineradiographic studies of the movements of different tongue points by Houde [27], illustrates the much lesser degree of complication in the tongue motions for labials in comparison with those for velars. It has been suggested that the main source of variation of formant frequencies is due to the inaccuracy of the positioning of the tongue [28]. The transitions of formants are, therefore, expected to be more stable for the bilabials than for the other plosives. This probably explains the high classification scores for the labials. Ta-

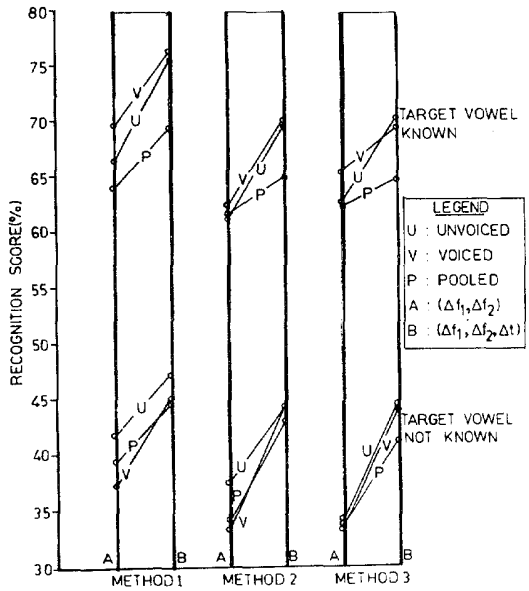


Fig. 3. Comparative performances of different methods and feature combinations.

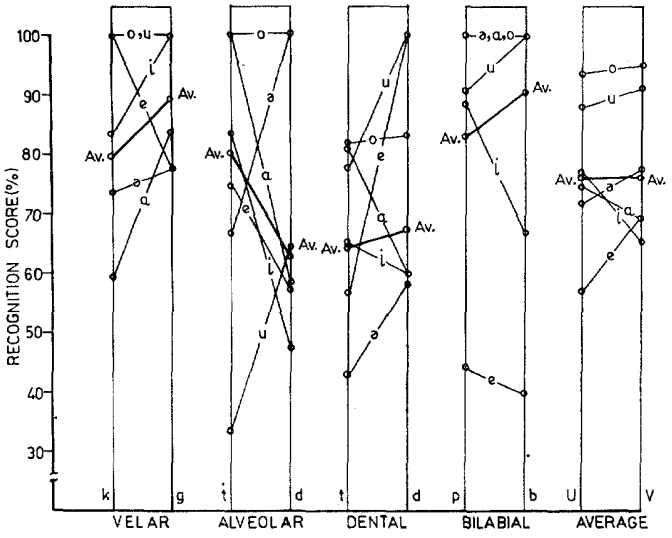


Fig. 4. Recognition scores of plosives for different target vowels.

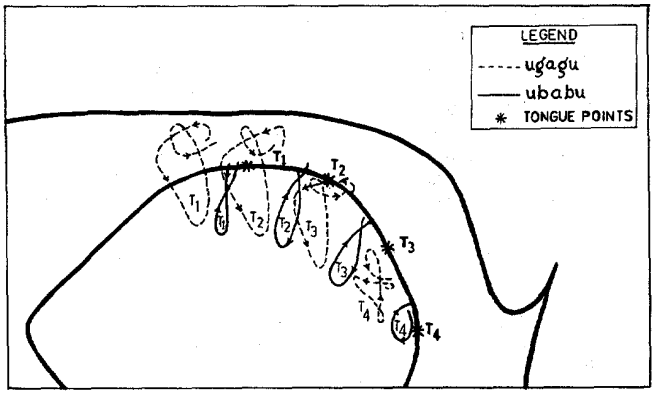


Fig. 5. Articulatory diagram of tongue-body motion.

TABLE III
AVERAGE MOVEMENTS IN mm OF DIFFERENT TONGUE POINTS FOR SPECIFIED TRANSITIONS

Transition	Tongue Points				Recognition Score [%]
	T_1	T_2	T_3	T_4	
gi	2.5	3.5	3.5	3.5	100
ga	10.0	11.5	12.5	8.0	85
gu	2.5	2.5	1.5	1.0	100

ble III presents average total displacements of different tongue points (results obtained from the same source [27]) for /gi/, /ga/, and /gu/ utterances and the corresponding recognition scores. The negative correlation of the displacement with recognition scores further substantiates the point.

The alveolar and dental plosives show consistently lower recognition scores. A reference to the confusion matrices (Appendix) indicates that the confusions amongst these two plosives themselves are mainly responsible for this. In fact, of the total error of 26 percent for unvoiced plosives in these two groups, 18 percent is due to the confusion amongst them. For voiced plosives, the respective figures are 35 percent and 20 percent. The closeness of these two articulatory positions is possibly the main reason for this confusion.

Each segment inside a vertical bar represents the plosive pair of a particular articulatory position in combination with a particular target vowel (Fig. 4). The slope of a segment reflects the relative status of the voiced over unvoiced counterpart of the plosive pair with respect to the recognition score. The preponderance of positive slopes for all groups except alveolar indicates that voiced plosives are generally better classified. However, the alveolar group exhibits a reverse trend. The plosives are generally better classified for the target vowels /o/, /u/, and /ə/. The differences in the classification scores between voiced and unvoiced plosives with respect to these vowels are also small. It may be of interest to note that plosive recognition is, on an average, poorer when the targets are front vowels. The back and central vowels provide very good targets in this respect.

Much attention has been given towards automatic recognition of words with a limited vocabulary. Reddy [8] has given

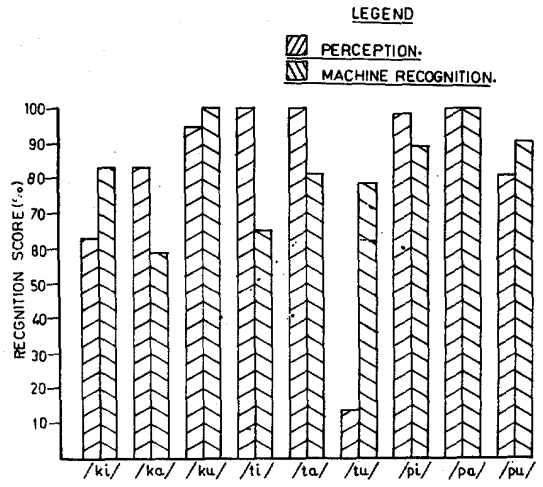


Fig. 6. Comparison of machine recognition and human perception.

a useful review of work in this field. Recent experiments by Rabiner [33] with 54 selected words have achieved recognition score of 85 percent. Unfortunately, machine recognition of consonants has failed to attract as much attention. It would, therefore, be difficult to find a frame of reference for the evaluation of the present results in a specific context. However, Sakai *et al.* [1] reported about 70 percent recognition for consonants, in general. Pal *et al.* [10] have reported some experiments using fuzzy algorithms where recognition scores vary from 60 percent to 85 percent for plosives. The present experiment produces an overall recognition score of 75 percent, starting from 62 percent for /t/ to 90 percent for /b/ and /g/.

These results of machine recognition compare well with human perception. Experiments with segmented and gated speech [24] reveal that the highest recognition scores were obtained when aperiodic plus vocalic transitions of the CV syllables were presented to the listeners. Fig. 6 represents a comparison of the results of this perception test [24] and those of present machine recognition. The overall performance of the machine is marginally better than that of native listeners. Perception tests conducted by Stevens *et al.* [23], with voiced synthetic plosives /b/, /d/, and /g/ in conjunction with vowels /a/, /i/, and /u/, provided an average score of 81 percent. A corresponding figure for these plosives with the vowels /ə/, /a/, /o/, /i/, /u/, and /e/ in the present experiment is about 82 percent.

VI. CONCLUSION

The feature set $[\Delta F_1, \Delta F_2, \Delta t]$ has been found to provide adequate cues for the classification of unaspirated plosive sounds. The assumption of a normal distribution for these features has also been found reasonable. The features display significant dependence both inside the groups and between groups. The use of transitions of higher formants is likely to improve recognition scores, though these are more difficult to trace [21], [25]. A prior knowledge of the target vowel is necessary and prior knowledge of the voicing manner is cer-

tainly helpful for the automatic recognition of plosives. In the present state of ASR this information can be satisfactorily obtained without much difficulty.

Appendix

Confusion Matrices for the Feature Set $[\Delta F_1, \Delta F_2, \Delta t]$ –Target Vowel Known

Unvoiced

	Classified as /k/	/ṭ/	/t/	/p/	Total No. of Observations
Actual Plosive					
/k/	71	5	15	0	89
/ṭ/	2	33	5	1	41
/t/	8	19	60	6	93
/p/	1	6	5	60	72
Total					295

Voiced

	Classified as /g/	/ḍ/	/d/	/b/	Total No. of Observations
Actual Plosive					
/g/	69	3	3	2	77
/ḍ/	4	49	19	6	78
/d/	9	13	56	5	83
/b/	0	4	1	50	55
Total					293

Unvoiced and Voiced Pooled

	Classified as Velar	Alveolar	Dental	Bilabial	Total No. of Observations
Actual Plosive					
Velar	129	12	22	3	166
Alveolar	7	85	16	11	119
Dental	23	53	89	11	176
Bilabial	1	13	7	106	127
Total					588

Acknowledgment

The authors express their thanks to Prof. D. Dutta Majumder, Head of the Electronics and Communication Sciences Laboratory, Indian Statistical Institute, Calcutta, India, for his valuable suggestions, to the Andhra Association of Calcutta, India, for voice recording, and to S. J. Gupta for all typing work. The authors also would like to thank the reviewers for their constructive comments.

References

[1] T. Sakai and S. Doshita, "The automatic speech recognition system for conversational sounds," *IEEE Trans. Electron. Comput.*, vol. EC-12, pp. 835–846, Dec. 1963.

[2] D. Dutta Majumder, A. K. Datta, and S. K. Pal, "Computer recognition of Telugu vowel sounds," *J. Comput. Soc. India*, vol. 7, pp. 14–20, 1976.

[3] J. W. Forgie and C. D. Forgie, "Results obtained from a vowel recognition computer program," *J. Acoust. Soc. Amer.*, vol. 31, pp. 1480–1489, 1959.

[4] S. K. Pal and D. Dutta Majumder, "Fuzzy sets and decision making approaches in vowel and speaker recognition," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-7, pp. 625–629, July 1977.

[5] H. Suzuki, H. Kasuya, and K. Kido, "The acoustic parameters for vowel recognition without distinction of speakers," in *Proc. Conf. Speech Commun. and Processing*, Massachusetts Inst. Tech., Cambridge, 1967, pp. 92–96, (reprints).

[6] R. W. Schaffer and L. R. Rabiner, "Systems for automatic formant analysis of voiced speech," *J. Acoust. Soc. Amer.*, vol. 47, pp. 634–648, 1970.

[7] D. J. Broad, "Formants in automatic speech recognition," *Int. J. Man-Machine Studies*, vol. 4, p. 411, 1972.

[8] D. J. Reddy, "Speech recognition by machine–A review," *Proc. IEEE*, vol. 64, pp. 501–531, Apr. 1976.

[9] A. K. Datta, N. R. Ganguli, and S. Ray, "Computer recognition of plosive speech sounds," in *Proc. Comput. Soc. India*, Feb. 1978, pp. 122–134.

[10] D. Dutta Majumder and S. K. Pal, "On automatic plosive identification using fuzziness in property sets," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-8, pp. 302–308, 1978.

[11] A. K. Datta, N. R. Ganguli, and S. Ray, "Computer recognition of consonantal speech sounds," in *Proc. 4th Int. Joint Conf. Pattern Recognition*, Kyoto, Japan, Nov. 1978, Session-B8, pp. 1047–1049.

[12] —, "Transition–A cue for identification of plosives," *J. Acoust. Soc. India*, vol. VI, pp. 124–131, 1978.

[13] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-24, pp. 550–558, 1976.

[14] F. Jelinek, "Continuous speech recognition by statistical methods," *Proc. IEEE*, vol. 64, pp. 532–556, 1979.

[15] F. Itakura and S. Saito, "Analysis synthesis telephone based on the maximum likelihood ratio method," in *Proc. 6th Int. Cong. Acoust.*, Y. Kohasi, Ed., Aug. 1968, paper C-5-5, pp. C17–C20.

[16] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *J. Acoust. Soc. Amer.*, vol. 27, pp. 769–773, 1955.

[17] M. Halle, G. W. Hughes, and J. P. A. Radley, "Acoustic properties of stop consonants," *J. Acoust. Soc. Amer.*, vol. 29, pp. 107–116, 1957.

[18] F. S. Cooper, P. C. Delattre, A. M. Liberman, J. M. Borst, and L. J. Gerstman, "Some studies on the perception of synthetic speech sounds," *J. Acoust. Soc. Amer.*, vol. 24, pp. 597–606, 1952.

[19] A. M. Liberman, P. C. Delattre, F. S. Cooper, and L. Gerstman, "The role of consonant–Vowel transitions in the perception of the stop and nasal consonants," *The Psychological Monographs: General and Applied*, vol. 68, pp. 1–13, 1964.

[20] A. M. Liberman, P. C. Delattre, and F. S. Cooper, "The role of selected stimulus variable in the perception of unvoiced stop consonants," *Amer. J. Psych.*, vol. 65, pp. 497–516, 1952.

[21] A. M. Liberman, "Some results of research on speech perception," *J. Acoust. Soc. Amer.*, vol. 29, pp. 117–123, 1957.

[22] K. N. Stevens and S. E. Blumstein, "Quantal aspects of consonant production and perception–A study of retroflex consonants," *J. Phonet.*, vol. 3, pp. 215–234, 1975.

[23] —, "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Amer.*, vol. 64, pp. 1358–1368, 1978.

[24] C. LaRiviers, H. Wintz and E. Heriman, "Vocalic transitions in the perception of voiceless initial stops," *J. Acoust. Soc. Amer.*, vol. 27, pp. 470–475, 1975.

[25] G. Fant, "Stops in CV syllabels," Speech Communication Laboratory, Royal Inst. Tech., Stockholm, Sweden, Tech. Rep. STL-QPSR4, 1969.

[26] R. A. Cole and B. Schott, "Toward a theory of speech perception," *Psychol. Rev.*, vol. 81, pp. 348–374, 1974.

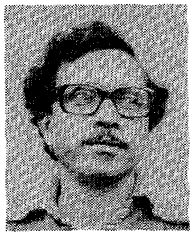
[27] R. A. Houde, Speech Communcation Research Laboratory, CA, SCRL Monograph No. 2, 1968.

[28] D. Dutta Majumder, A. K. Datta, and N. R. Ganguli, "Some studies on acoustic phonetic features of Telugu vowels," *Acustica*, vol. 41, pp. 55–64, 1978.

[29] G. Udny Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*. London, England: Griffin, 1953, p. 159.

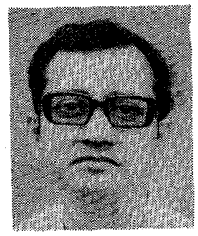
[30] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*. New York: Wiley, 1958, p. 147.

[31] G. S. Sebestyen, *Decision-Making Process in Pattern Recognition*. New York: Macmillan, 1962, p. 18.
[32] S. E. G. Ohman, "Coarticulation in VCV utterances–Spectrographic measurements," *J. Acoust. Soc. Amer.*, vol. 39, pp. 151–168, 1966.
[33] L. R. Rabiner, "On creating reference templates for speaker–independent recognition of isolated words," *IEEE Trans. Acoust., Speech., Signal Processing*, vol. ASSP-26, pp. 34–42, 1978.



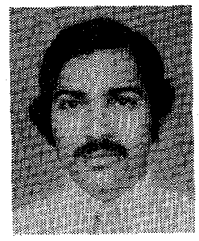
A. K. Datta was born in 1935. He graduated with honors in physics from Calcutta University, Calcutta, India, in 1955, and received the M.Sc. degree in pure mathematics in 1963 as an external noncollegiate student.

Since 1955 he has been with the Electronics and Communication Sciences Laboratory, Indian Statistical Institute, Calcutta, India, where he worked in the field of accounting machines, computer memory, and computer hardware before taking up pattern recognition. His present research activities include speech acoustics, speech pattern recognition, handwritten character recognition and robotics.



N. R. Ganguli was born on September 1, 1939. He received the Engineer's degree in telecommunication engineering from Jadavpur University, India, in 1961.

From 1962–1968, he was engaged in the development of computers in the Indian Statistical Institute and Jadavpur University Joint Computer Project. Since 1969 he has been with the Electronics and Communication Sciences Laboratory, Indian Statistical Institute, Calcutta, India. His current research interests are in the areas of speech analysis, synthesis, and recognition.



S. Ray was born on January 2, 1953. He received the M.Stat. degree with specialization in computer science from the Indian Statistical Institute, Calcutta, India, in 1972.

From 1973–1976, he worked as a Programmer in the National Institute of Rural Development, Hyderabad, India, and the Indian Oil Corporation, Calcutta, India. He is presently with the Electronics and Communication Sciences Laboratory, Indian Statistical Institute, Calcutta, as a Programmer. His research interests include computer-oriented statistical methods of pattern recognition.