Reprinted from "Journal of the Computer Society of India", Vol. 7 No. 1, December 1976

Computer Recognition of Telugu Vowel Sounds

D Dutta Majumder
AK Dutta
SK Pal
ISI, Calcutta

The paper presents a method for automatic recognition of vowel speech sounds using a three dimensional weighted discriminant function. Experiments were conducted with Telugu (one of the major Indian languages) speech sounds, both long and short, and all the processing of speech data was carried out on an electronic digital computer Honeywell 400. Different sets of training samples, viz., 5%, 10%, 20%, 40%, 60%, 80%, and 100% were randomly drawn from each pattern class with which reference events and weighting coefficients corresponding to a training set were evaluated. Standard deviation and percentage variation of features are used as weighting coefficients. Final assignment of the unknown pattern was done on the basis of the maximum computed value of the discriminant function. Pattern classes with the first two maximum values of discriminant functions were also noted to find out the importance of linguistic knowledge in vowel identification. It is also found that if the sample size is sufficient to characterize the class, the recognition score is not affected by increasing the sample size. The recognition score was found to be about 82% and an increment in recognition by a factor of 15% is caused by incorporating a second choice from the classifier.

Introduction

In the field of Communication Science and Computer Technology, the general problem of designing an automatic machine for pattern recognition in general and speech recognition in particular is one of the most fascinating aspects for achieving faster and better communication between man and machine. Pattern recognition can be viewed as a two-fold task consisting of the extraction of significant characterising features which determine the invariant and common properties of a set of samples followed by the classification technique based on the measurement of these characteristic features. The present paper is a part of the work being carried out in the field of Automatic Speech Recognition (ASR) (Dutta Majumder *et al*, 1968 (a), 1968 (b), 1969, 1970, 1975, 1976) by computers at the Electronics and Communication Sciences Laboratory of the Indian Statistical Institute.

Recognition of speech patterns is a very complex problem involving multilevel decision processes (Dulley *et al*, 1958, Sakai *et al*, 1963, and Suzuki *et al*, 1967). In order to achieve clarity of understanding the whole problem of recognition has been broken up into various parts, e.g., recognition vowels, consonants, recognition of different phonemes in isolation and in connected speech. There are several researchers (Reddy 1966, Shaffer *et al*, 1970, Broad, 1972, Paliwal *et al*, 1976) dealing with the problem of speech recognition using time, frequency or time-frequency (spectograph) domain analysis. Because of the remarkable inter-repetition stability (Broad, 1972) and close relation of the formants to the phonetic concepts of segmentation and equivalence, the formants are found to have potential application in speech and speaker recognition. The use of a computer in speech recognition was first made by Forgie *et al*, 1959, who recognised ten vowels with 93% accuracy using a single speaker. Suzuki *et al*, 1967 reported successful recognition score using first three formants of Japanese vowels. In a recent work of Paliwal *et al*, 1976, the machine was found to render correct decision in 80 to 93% English vowels for one speaker, with the first three formants and their amplitudes as recognition features. The size of the sample space in these experiments was small enough containing utterances of isolated vowels.

In the present paper a scheme for recognition of vowels in various consonantal contexts and experimental results for different speakers have been presented. The vowel qualities are considerably influenced in connected speech by the adjoining consonants. Recognition criteria considered are some interrelated functions of the first three formant frequencies. The data consists of about 900 utterances of 10 Telugu vowels occurring in the first nuclear position of a selected commonly used vocabulary of Telugu words. The spectra for these words were made on a Kay Sonagraph Model 7029 A. From these spectrograms the acoustic data were obtained manually. The formant frequencies for the first nuclear vowels were taken at the central position of the steady state. Where the steady state was not observable because of the extreme shortness of the vowel, the congruence of the off-glide and on-glide was taken as the nuclear position of the vowel.

Electronics and Communication Sciences Department, Indian Statistical Institute, 203, Barrackpore Trunk Road, Calcutta 700 035

The discriminant function used for classification is Euclidean distance (Sebestyen, 1962) in the measurement space from the representative points of each class, the weighting coefficients for the coordinates used are function of "percentage" and "standard deviations" of the corresponding features for two different experiments.

Extraction of Features and Recognition Criteria for Vowel Identification Scheme

Pattern Recognition is considered as a process of decision making in which a new input is decided to be a member of a particular group by the comparison of its attributes with those of previously known members of that class. To obtain knowledge about the features space of the Telugu vowel patterns (/ə/, /a: /,/i ˫/,/ i: / , / u˥ /, /u: / , / e ˫ / , /e:/ , /o ˥ /, and/o:/), they were selected in a CNC (consonant nucleus consonant) combination as the vowel qualities are considerably influenced in connected speech by the adjoining consonants. All these speech samples were recorded by five adult male speakers on an AKAI tape recorder. From these five informants three were chosen on the basis of a listening experiment based on the opinion of 10 listeners. The frequency analysis was done on a Kay Sonagraph model 7029 A. The first three formant frequencies were taken at the steady state of the first nuclear vowel position. Out of 871 samples the third formant for 384 could not be clearly obtained. For these vowels the average value of the third formant for that vowel corresponding to the particular speaker was injected. Thus all the samples with the measured value of the three features F_1, F_2, and F_3 could be thought to constitute a three dimensional feature space, each dimension representing a property of the event.

The significant information available about the event thus could be expressed as a three-dimensional feature vector

$$F = \begin{bmatrix} F_1 \\ F_2 \\ F_3 \end{bmatrix}$$

the coordinates of which would have numerical values indicating the amount of each property of the event. Each event with a set of all measured features will therefore, correspond to a single point in the feature space Ω_F.

The ensemble of points, for the purpose of recognition, is assumed to be isolated, statistically independent from one another and to lie in regions separable by surfaces in the feature space Ω_F. The points which represent a set of non-identical events of a common group are also considered to be close to each other, as measured by some method of measuring distances. All the features of a particular sample point represented by the different coordinate directions are not equally significant in defining the characteristics of a class to which like events belong. it is reasonable to assume that the feature with larger variation is less characteristic in nature. Therefore in measuring closeness or similarity, lower weight is to be given to features having large variations. In the present experiment, the features with increasing variance have been weighted with decreasing values of a "feature weighting" coefficient, W_n, n=1,2,3 and percentage variation and variance of the formants as weighting coefficients were studied.

Of the ten Telugu vowels, long and short categories, viz., /i ˫ , i:/ , /u ˥ , u:/ , /e ˫ , e:/ and /o ˥ , o:/ were found to differ from one another mainly in duration (Dutta Majumder *et al.*, 1974) but phonetically they are not distinctively different. With this background, the number of pattern classes to be recognised were transformed to six namely, /ə/,/a:/,/i/,/u/,/e/and /o/ which are phonetically different from one another. Therefore, the above designed feature space Ω_F could be viewed to be constituted by various pattern classes C_1, C_2, C_{3s} and C_{3l}, C_{4s} and C_{4l}, C_{5l}, and C_{6s} and C_{6l} where subscripts *s* and *l* stand for shorter and longer categories. Through classification, this 3-dimensional feature space is to be divided into such regions which contain vowels differing only in phonetic features. Some of these regions would contain two subregions, one for the short vowels another for the longer ones. The next task before classification is, obviously, the selection of reference vectors or "prototypes"d enoting the representative point of each class. Of the entire events belonging to a category one may choose a set of events called a "training set" from which prototype events can be chosen as having the mean value of the feature measurements for each coordinate direction. The weighting coefficients for the specified class could be developed for these training sets. Though longer and shorter types of a particular vowel are treated to be in the same group, they are given individual reference vectors computed over their respective training sets.

The last and final task of a pattern recognition system is classification by a suitable classifier, whose function is to examine a maximum similarity between the reference vectors of a class and a new input vector. For the vowel classes / i / , / u / , / e /and/o/, the "closeness" is measured separately for both the shorter and the longer subgroups, and input pattern is assigned to a pattern class which is associated with maximum similarity for either of the subgroups. A suitable classifier based on a "maximum discriminant function" criterion is described in the subsequent section.

Classificatory Method Used

Let us consider an N-dimensional feature vector space Ω_F in which $C_1, C_2, \ldots, C_j, \ldots, C_m$ are the m possible pattern classes to be recognised with their corresponding m reference vectors $R_1, R_2, \ldots, R_j, \ldots, R_m$ and

let

$$F = \begin{bmatrix} F_1 \\ F_2 \\ \vdots \\ F_n \\ \vdots \\ F_N \end{bmatrix} \quad (1)$$

be designated as the feature vector, the numerical value of F_n representing the nth property of the event. Then the problem of classification is to assign each point in the vector space represented by F to a proper class—a mapping from feature space to decision space.

If the feature vector F representing an input is a member of the class C_k, denoted as $F \in C_k$, the "Discriminant function", $D_k(F)$, associated with the class C_k, $k = 1, 2, \ldots, m$, must then possess the largest value (Fu, 1968). That is, a classificatory decision would be as follows :

$$\text{Decide } F \in C_k, \text{ if } D_k(F) > D_j(F) \quad (2)$$

in which k and j may have any integer value $1, 2, \ldots,$ m and $k \neq j$. Then obviously, the decision boundary in the N-dimensional feature space Ω_F, between regions associated with classes C_k and C_j respectively would be governed by the expression

$$D_k(F) - D_j(F) = 0 \quad (3)$$

with $k \neq j$, $k, j = 1, 2, \ldots, m$.

Let the reference vector in R_j associated with pattern class C_j be denoted as

$$R_j^{(l)} \in R_j \quad , l = 1, 2, \ldots, P_j, \quad (4)$$

where P_j is the number of reference vectors in the set R_j. In the present case, $P_j = 2$ (shorter and longer subgroups) for the vowel classes /i/, /u/, /e/ and /o/ and $P_j = 1$ for the vowel classes/ *d* / and/ a : /.

Defining the distance between an input pattern F and a prototype point R_j as the smallest of the distances between F and each reference vector $R_j^{(l)}$ in R_j, it can be expressed as,

$$d(F,R_j) = \underset{l=1,2,\ldots,P_j}{\text{Min}} |F - R_j^{(l)}|, j = 1, 2, \ldots, m. \quad (5)$$

or

$$d(F,R_j) = \underset{l=1,2,\ldots,P_j}{\text{Min}} \left[\sum_n (F - R_{jn})^2\right]^{\frac{1}{2}}, \quad j = 1, 2, \ldots, m, \quad (6)$$

where subscript n indicates the component value of the vector along the nth co-ordinate direction. Introducing the "weighting coefficients" $W_1^{(l)}, W_2^{(l)}, \ldots, W_j^{(l)}, \ldots, W_m^{(l)}$ in which $W_j^{(l)}$ is associated with reference vector $R_j^{(l)}$, equation (6) can be rewritten as

$$d(F,R_j) = \underset{l=1,2,\ldots,P_j}{\text{Min}} \left[\sum_n \left\{W_{jn}^{(l)} (F_n - R_{jn}^{(l)})\right\}^2\right]^{\frac{1}{2}}, \quad (7)$$

where $j = 1, 2, \ldots, m$, $W_{jn}^{(l)} = \rho / \sigma_{jn}^{(l)}$, ρ is an arbitrary constant and $|W_{jn}^{(l)}| \leq 1$. $\sigma_{jn}^{(l)}$ being considered to be either "standard deviation" or "coefficient of variation" of the measured features corresponding to prototype point $R_{jn}^{(l)}$ could be computed over the members of any size sample of a pattern class. $\sigma_{jn}^{(l)}$ when defined as standard deviation of the features along the nth coordinate direction in class $C_j^{(l)}$, $W_{jn}^{(l)}$ can have the form

$$\left[W_{jn}^{(l)}\right]^2 = \frac{1}{\sigma_{jn}^{(l)2}} = \frac{1}{\frac{1}{M}\sum_{i=1}^{M} (F_{ijn}^{(l)} - R_{jn}^{(l)})^2}$$

for $\rho = 1$

$$\text{or, } \left[W_{jn}^{(l)}\right]^2 = \frac{1}{\frac{1}{M}\sum_{i=1}^{M} \left[F_{ijn}^{(l)}\right]^2 - \left[R_{jn}^{(l)}\right]^2} \quad (8)$$

If, instead of standard deviation coefficient of variation is used, then,

$$W_{jn}^{(l)} = \frac{R_{jn}^{(l)}}{(F_{jn}^{(l)})_{max} - (F_{jn}^{(l)})_{min}}, \quad (9)$$

$$\text{where } R_{jn}^{(l)} = \frac{1}{M}\sum_{i=1}^{M} F_{ijn}^{(l)}, \quad j = 1, 2, \ldots, m, \quad (10)$$

and $(F_{jn}^{(l)})_{max}$ and $(F_{jn}^{(l)})_{min}$ denote the maximum and minimum value respectively for nth co-ordinate in class $C_j^{(l)}$ and M is the total number of features $F_{ijn}^{(l)}$, representing nth components of a specific set of training samples associated with pattern class $C_j^{(l)}$.

Thus the corresponding weighted discriminant function $D_j(F)$ with respect to class C_j is essentially

$$D_j(F) = \frac{1}{d(F, R_j)} = \underset{l=1,\ldots,P_j}{\text{Max}} \left[\sum_{n=1}^{N} \left\{ W_{jn}^{(l)} \cdot (F_n - R_{jn}^{(l)}) \right\}^2 \right]^{-\frac{1}{2}}$$

where $j = 1, 2, \ldots, m$. The classification model for the pattern classes of Telugu vowels could be described as :

Decide the input F in jth class, i. e., $F \in C_j$ if,

$$\underset{l=1,2,\ldots,p_j}{\text{Max}} \left[\sum_{n=1}^{N} \left\{ W_{jn}^{(l)} (F_n - R_{jn}^{(l)}) \right\}^2 \right]^{-\frac{1}{2}}$$

$$> \underset{l=1,2,\ldots,p_k}{\text{Max}} \left[\sum_{n=1}^{N} \left\{ W_{kn}^{(l)} (F_n - R_{kn}^{(l)}) \right\}^2 \right]^{-\frac{1}{2}} \quad (11)$$

for $k, j = 1, 2, \ldots, m$ and $k \neq j$.

A vocabulary consisting of Telugu words was selected so as to encompass as many CN and NC combinations as possible with an emphasis on the use of commonly used words. These were recorded by five adult male speakers on an AKAI tape recorder inside a big auditorium. On the basis of a listening experiment by 10 listeners, only 871 samples of three speakers were selected. The Spectrographic analyses of these utterances were done on a Kay Sonagraph Model 7029A. The analyses were carried out in the normal mode and the band 80Hz to 8kHz with wide band filters having bandwidth 300Hz was chosen.

Formant frequencies F_1, F_2, and F_3 were obtained manually at the steady state of the vowels. Wherever, due to the extreme shortness of the vowels, steady states were not observed, the measurements were made at the point of congruence of the off-glide and on-glide. The samples which did not depict a prominent third formant were allowed to have an injected average third formant $(F_3)_{av}$. computed over all members of that class of vowels for the particular speaker. The number of samples which fall in this category is 384.

In the third step, different sets of training samples, viz., 5%, 10%, 15%, 20%, 40%, 60%, 80% and 100% were randomly drawn from each pattern class with which reference events $R_{jn}^{(l)}$ (equation 10) and weighting coefficients $W_{jn}^{(l)}$(equation 8) corresponding to a training set were evaluated. Since a set of 5% events in a class encompasses only 4 to 6 samples, training sets lower than it were not taken. The purpose of selecting various sample sizes was to study the effect of the size of training set on overall recognition score. In a few cases, where the standard deviation of the magnitudes of a coordinate in a training set was zero, the corresponding weighting coefficient $W_{jn}^{(l)}$ was set at unity. Though it does not satisfy equation (7), it is still logical in the sense that, an attribute occurring with identical magnitude in all members of a training set is an "all important" feature of the set and hence its contribution in the discriminant function need not be reduced.

Finally, all the ten Telugu vowels having phonetic symbol /ɖ/,/a:/,/i ↦ /,/i/,/i:/,/u ↦ /,/u:/,/e ↦ /,/e:/,/o ↦ /and/o:/, for the purpose of classification,were reduced to the six classes which are phonetically quite different from one another. Longer and shorter vowels were grouped together as they differ significantly only in duration. To assign a proper class to an unknown, its discriminant function Dj(F) using formula (11), corresponding to all the classes were measured. It was then decided to belong to a class associated with maximum value of Dj (F). It is also reasonable to note the pattern classes associated with the first two maximum values from which a correction can be attempted whenever a supervisory programme indicates a wrong classification. The function of the supervisory scheme may be such that the correctness of a class corresponding to first choice was first examined on the basis of syntactic or semantic knowledge of the CNC word. If it did agree, the selection would correspond to first choice, otherwise, the second choice of the classifier was taken to be final.

In the second experiment, the percentage variation of the features were measured and corresponding weighting coefficients (equation 9) were used to compute Dj (F). The third and last experiment encompassed the measurment of Dj (F) without deforming the feature space (with $W_{jn}^{(l)}|_{l=1}$) and classification scores obtained from these various approaches are described in the next section.

Results

Recognition scores for the individual vowel classes are shown in Tables I and II in which the "Weighting coefficient" is the reciprocal of the standard deviation of features calculated over the entire set of samples. The performance of the machine in recognising vowels is explained through confusion matrices for different weighting coefficients. A numeral in a cell denotes the number

of instances for which same decision was made by the classifier and diagonal elements thus represent the number of utterances correctly recognised. The recognition score with the reciprocal of standard deviation as the weighting coefficient (Tables I and II) was found to be maximum (≈82%). The scores for different classes were also found to be larger (with the exception of only one vowel class /a:/) with this coefficient than those without any weighting coefficient (Table IV). This is expected as the fixation of proper weights ensures correct representation of the importance of different features in classification. Interestingly, the weights based on percentage variation were found to decrease the recognition score (Table III). It was revealed on investigation that these weights tended to attach more importance to the second formant frequency whereas the weights based on standard deviation attached importance in the order of 1st, 2nd and 3rd formant frequencies. The significant decrease (about 10% on overall score) tends to indicate that for vowel classification the 1st formant is more important than the 2nd formant.

Tables I and II reveal that the recognition score for the set having an injected average 3rd formant frequency is somewhat larger than that for the set having actual 3rd formants. It has been generally accepted that the 3rd formant frequency is not a very important cue for recognition. The individual variation of the 3rd formant for the the second set probably created some confusion thus lowering the total recognition score.

Investigation of the scope of correction of errors by providing a second choice has been carried out. Lower numerals in diagonal cells of Tables I and II represent the number of errors corrected. It is seen that the system can provide an improvement of recognition by 15% by incorporating second choice under the control of a supervisory scheme using extra acoustical features as discussed earlier.

Results of the investigation on the optimum size of training set of samples is graphed in Fig. 1, where the percentage of classification is seen to be almost constant from and above 15% of training samples. Such a set contains about 12 to 16 samples for each class. Therefore, it could be stated that the variation of recognition rate with the training set of events becomes insignificant after a size sufficiently

TABLE I
CONFUSION MATRIX FOR VOWEL RECOGNITION WHEN STANDARD DEVIATION OF FORMANTS IS CONSIDERED AS WEIGHTING COEFFICIENT

ACTUAL CLASS

CLASSIFIED AS	i	e	ɑ	a:	o	u
i	89 (15)	26				
e	16	119 (42)	2		1	
ɑ		11	20 (4)	10	3	1
a:			2	38 (7)		
o		7	2	1	74 (10)	1
u			1		9	54 (1)

TABLE II
CONFUSION MATRIX FOR VOWEL RECOGNITION WHEN STANDARD DEVIATION OF FORMANTS IS CONSIDERED AS WEIGHTING COEFFICIENT (SAMPLES WITH INJECTED $(F_3)_{av}$)

ACTUAL CLASS

CLASSIFIED AS	i	e	ɑ	a:	o	u
i	66 (1)					
e	1	35 (8)	7			
ɑ		4	32 (7)	7	4	
a:			6	32 (1)		
o		5		1	64 (26)	5
u					25	90 (3)

TABLE III
CONFUSION MATRIX FOR VOWEL RECOGNITION WHEN PERCENTAGE VARIATION OF FORMANTS IS CONSIDERED AS WEIGHTING COEFFICIENT

ACTUAL CLASS

CLASSIFIED AS	i	e	ɑ	a:	o	u
i	86	41				
e	18	105	5		3	
ɑ		13	16	5		
a:	1		3	32		
o		1	1	11	53	5
u		3	2	1	11	50

TABLE IV
CONFUSION MATRIX FOR VOWEL RECOGNITION WITHOUT WEIGHTING COEFFICIENT

ACTUAL CLASS

CLASSIFIED AS	i	e	ɑ	a:	o	u
i	93	40				
e	10	91	3		1	
ɑ		25	14	3		
a:	2	1	7	40	2	
o		6	2	6	63	4
u			1		21	51

large to provide a good representative and weighting coefficients which characterise the classes.

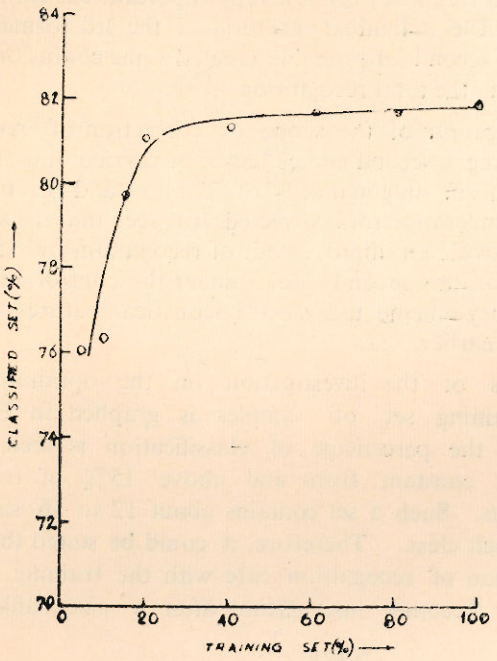

Fig. 1. Effect of training sample size on recognition score

Conclusion

The classification technique described here, is more general and simpler in comparison to the earlier experiments (Suzuki *et al*, 1967, Dutta Majumder *et al* 1975, 1976, Paliwal *et al*, 1976) comprising straightforward distance measurements of a point in a 3-dimensional feature vector space from a reference point with a suitable "weighting coefficient." Recognition score (≈82%) of Telugu vowels could be thought to be invariant with any training set, containing enough sample points to describe the properties of a pattern class. Confusion in machine recognition of a vowel is seen to be limited only to neighbouring classes constituting a vowel triangle. This is in agreement with that of the other experiments (Dutta Majumder *et al*, 1975, Paliwal *et al*, 1976). Again, it could be concluded that, the formant frequencies (F_1,F_2, and F_3) of a vowel are not the only characterising features besides, a linguistic knowledge of vowels, to have correct classification, has been taken into account.

A rechecking on 5% samples revealed that, the error in manual extraction of the formant frequencies lies within a range of ± 20 Hz. The formants were obtained at the steady state of vowel. So far as vowel recognition is concerned, this is not supposed to differ much from that obtained from dynamic characteristic, via digital processing. The prototype points chosen in this method are nothing but the average values of the coordinates corresponding to samples in specific training set, and the reference points so obtained are not distributed throughout the class, but rather are close to one another and describe properties of only a part of the class. An improved process of evaluating prototype events that is expected to ensure better classification is in progress and will be reported in due course.

Appendix

Summary of the paper " Studies on Acoustic Phonetic Features of Telugu Speech Sounds," by Dutta Mujumder *et al*, 1974.

The paper deals with the acoustic phonetic features of Telugu speech sounds as obtained from the spectrographic analysis of 362 words. Results are summerised below :

a) The variation of the first three formant frequencies have been statistically tested and found to follow the normal distribution. It was found that the spread of first formant is from 325 Hz for vowel [i:] to 710 Hz for [a:] and that for second formant is from 923 Hz for vowel [u:] to 2260 Hz [i:] and that for third formant is from 2400 Hz for [a:] to 2757 Hz for [i:].

b) The shorter vowels have a tendency towards centralisation in the high tongue position than their corresponding longer counterparts. A comparison of these data with the LEA data of Fant (1969) indicates that the accuracy of positioning of the tongue is the main reason for the larger variation of the formant frequencies for shorter categories.

c) For front vowels, low tongue position shows clearly larger variation than the high tongue position. For back vowels, the tongue position is not discriminatory in this respect. For high tongue position, the short vowels show larger variations in formant frequencies for back vowels than front vowels. The case is reversed for long low vowels. The high, long, and low, short vowels do not show such discrimination with respect to variations in formant frequencies for the back front portion of the tongue.

d) The speaker to speaker variation for short vowels is significantly larger than those for long vowels. For both long and short front vowels, the duration of open vowels are somewhat larger than that of close vowels.

e) The velar [h] fricative consonant is characterised by two consonantal formants, one in the low region 300 - 1400 Hz and second in 2200 - 4000 Hz range. The palatal [ʃ] and dental [s] fricatives represent only a single concentration. Concentration for [ʃ]

extends to a much higher frequency than that of dental [s]. The dental fricative is very distinct in its appearence in the spectrogram, the velar being the weakest.

f) The burst energy spectrum of the voiceless unaspirated fricative [t ʃ] bears close resemblance to that of the unvoiced dental stops [t]. Voiced affricates [dz], [t ʃ h] and [dzh] have burst spectrum similar to that of voiced velar stops [g], [kh] and [g] respectively and the frictional energy spectrum have striking similarity with [s], [ʃ] and [ʃ] respectively.

g) The concentration of acoustic energy for [r] is nearer to [l]. There are two different [l] in Telegu language. One is alveolar [l] and the other is palatal [l̇]. The formant like energy concentration is more prominent and steady for [l] than that of [l̇].

h) The plosives show an occlusion period ranging from 50 msec to 180 msec and a burst period of 3 msec to 5 msec. For aspirated plosives the period of aspiration ranges from 30 msec to 75 msec. The velar consonants have the largest period of aspiration and the dentals have the lowest aspiration period. The total duration of affricates varies from 100 msec to 175 msec. The fricative consonants have the largest duration (120 msec to 152 msec) of energy concentration among all the consonants. The palatal fricatives have larger duration than the dental ones. [l] has larger duration than [l̇].

Acknowledgement :

The authors wish to acknowledge with thanks the helpful co-operation from Shri N. R. Ganguli, Shri B. Mukherjee, Shri S. C. Kundu and Shri S. K. Seal in completing the work and to Professor C. R. Rao, FRS, Director of the Indian Statistical Institute for his kind interest. Thanks are also due to the Council of Scientific and Industrial Research for providing financial assistance to Shri S. K. Pal, one of the authors, in the form of a research fellowship.

References :

Broad DJ, 1972, Formants in Automatic Speech Recognition, *Int. J. Man-machine Studies*, 4, 411.

Dulley H, Balashek S, 1958, Automatic Recognition of Phonetic Patterns in Speech, *JASA*, 30, 721-732.

Dutta Majumder D, Dutta AK, 1968a, A Model for Spoken Word Recognition, *Proc. Automazione E Strumentazione*, Milan, Italy, 250-259.

Dutta Majumder D, Dutta AK, 1968b, Some Studies on Automatic Speech Coding and Recognition Procedure, *Ind. J. Phys.*, XLII, 7, 425-443.

Dutta Majumder D, Dutta AK, 1969, An Analyser-Coder for Machine Recognition of Human Speech, *JIETE*, 15, 4, 233-243.

Dutta Majumder D, Dutta AK, 1970, Some Studies on the Suitability of Associative Memory in Pattern Recognition System, *Proc. Automazione E Strumentazione*, Milan, Italy, 203-218.

Dutta Majumder D, Dutta AK, Ganguly NR, Mukherjee B, Sarkar R, 1974, Studies of Acoustic Phonetic Feature of Telugu Speech Sounds, *ECSL Series on Phonological Studies, Mahalanobis Memorial Issue.*

Dutta Majumder D, Dutta AK, Pal SK, 1975, Vowel Identification Using Piecewise Separation Technique, *JASA* (Communicated).

Dutta Majumder D, Dutta AK, Pal SK, 1976, The Concept of Fuzzy Sets and its Application in Pattern Recognition Problems, *Proc. 11th Annual Convention of the Computer Society of India*, 20-23rd Jan, Hyderabad, SD 02.

Fant G, 1969, *Speech Analysis*, STL - QPSR, 4.

Fu KS, 1968, *Sequential Methods in Pattern Recognition and Machine Learning*, Academic Press, London.

Forgie JW, Forgie CD, 1959, Results obtained from a Vowel Recognition Computer Program, *JASA*, 31, 1480-1489.

Paliwal KK, Rao PVS, 1976, Computer Recognition of Isolated Phonemes, *Proc. 11th Annual Convention of the Computer Society of India*, 20-23rd Jan, Hyderabad, SR 04.

Reddy DR, 1966, Segmentation of Speech Sounds, *JASA*, 40, 307-312.

Shaffer RW, Rabiner LR, 1970, Systems for Automatic Formant Analysis of Voiced Speech, *JASA*, 47, 634-648.

Sebestyen GS, 1972, *Decision Making Processes in Pattern Recognition*, The Macmillan Co. New York.

Sakai T, Doshita S, 1963, The Automatic Speech Recognition System for Conversaticnal Sounds, *IEEE Trans. Electronic Computers*, EC-12, 5, 835-846.

Suzuki H, Kasuya H, Kido K, 1967, The Acoustic Parameters for Vowel Recognition without Distinction of Speakers, *Conf. on Speech Comm. and Processing*, MIT, 92-96 (Conf. reprints).