Synthesis-based recognition of continuous speech

K. K. Paliwal and P. V. S. Rao

Speech and Digital Systems Group, Tata Institute of Fundamental Research, Homi Bhabha Road, Bombay 400005, India

(Received 13 June 1979; accepted for publication 1 December 1981)

An acoustic phonemic recognition system for continuous speech is presented. The system utilizes both steady-state and transition segments of the speech signal to achieve recognition. The information contained in formant transitions is utilized by the system by using a synthesis-based recognition approach. It is shown that this improves the performance of the system considerably. Recognition of continuous speech is accomplished here in three stages: segmentation, steady-state recognition, and synthesis-based recognition. The system has been tried out on 40 test utterances, each 3–4 s in duration, spoken by a single male speaker and the following results are obtained: 5.4% missed segment error, 8.3% extra segment error, 52.3% correct recognition using only steady-state segments, and 62.0% correct recognition using both steady-state and transition segments.

PACS numbers: 43.70.Sc

INTRODUCTION

There are in existence a number of acoustic phonemic recognition systems for speech which were developed either independently (Reddy, 1967; Niederjohn and Thomas, 1973; Hess, 1976; Paliwal and Rao, 1977) or as part of so-called speech understanding systems (Reddy *et al.*, 1973; Schwatz and Makhoul, 1975; Weinstein *et al.*, 1975; Goldberg, 1975). These systems use primarily the acoustic information contained in the steady-state segments of continuous speech for recognition of phonemes. Since there is considerable acoustic variability in continuous speech arising from different interphonemic contexts and speaking rates, steady-state information alone is not enough for the phonemic transcription of continuous speech. Also, it is well established that the acoustic information contained in transition segments plays an important role in human perception (Delattre *et al.*, 1955; Lindblom and Studdert-Kennedy, 1967). Hence, it would seem that it would be a good strategy for machine recognition systems to try to utilize this transitional information to the maximum possible extent.

A simple and straightforward procedure for utilizing the transitional information for speech recognition would be to store sample transition segments in some parameterized form for all possible phoneme pairs (assuming that context effects between immediate neighbors only are important). One could then classify the input transition segment by comparing it, after time normalization, with all stored transition segments using some suitable distance measure. Such an approach, as used by Dixon and Silverman (1977), would require prohibitively large memory. A more economic and flexible approach to utilize this transitional information for recognition has been proposed by Thosar and Rao (1971, 1976). Their scheme utilizes interphoneme contextual information contained in formant transitions and employs internal trial synthesis and feedback comparison as a means for recognition. This synthesis-based recognition scheme has been tried out on vowel–stop–vowel utterances and encouraging results are reported. Cook (1976) and Klatt (1974) have also used a synthesis-based strategy for word verification.

The work presented in this paper is an extension to Thosar and Rao's work. Their synthesis-based recognition scheme with some modifications has been used here to recognize continuous speech. The system, shown in the form of a block diagram in Fig. 1, accomplishes speech recognition in three stages: segmentation, steady-state recognition, and synthesis-based recognition. The recognition system accepts unconstrained continuous speech spoken in an ordinary office-room environment. The system as implemented here is trained for a single male speaker.

For the purpose of training and testing the recognition

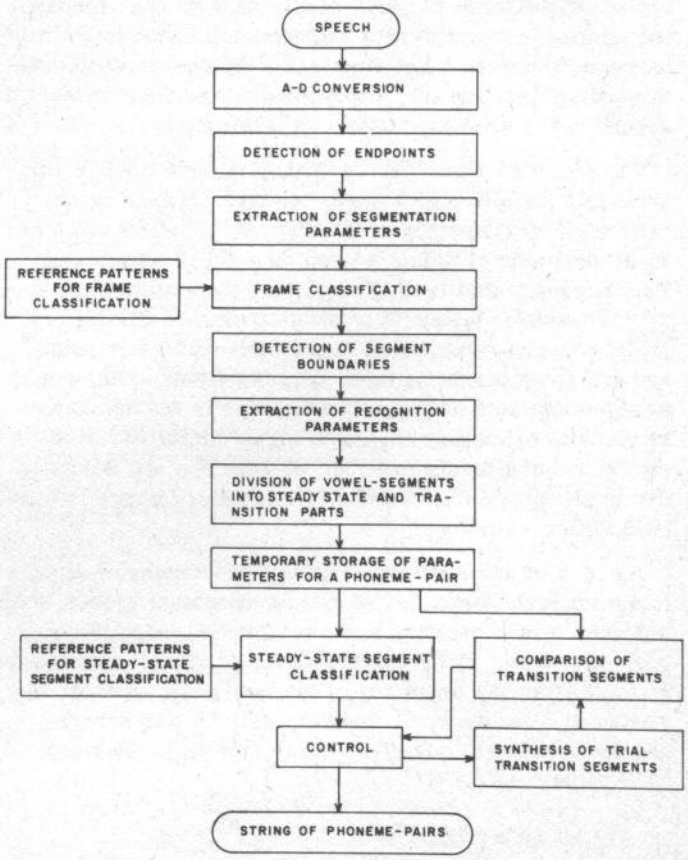

FIG. 1. Block diagram of the recognition system.

 0001-4966/82/041016-09$00.80

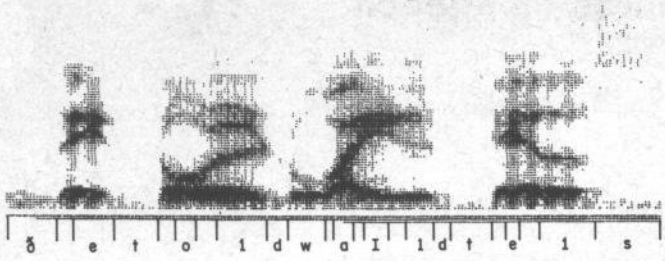

FIG. 2. Digital spectrogram of the speech utterance "They told wild tales" with manual segmentation and labeling.

system, it is necessary to have correct segmentation and labeling of speech utterances. This is done here manually using the digital spectrograms of speech utterances plotted by the computer on a line printer. Figure 2 shows the digital spectrogram of a speech utterance "They told wild tales." In this figure, the boundaries of steady-state and transition segments are marked manually and the steady-state segments are manually labeled.

I. DATA ACQUISITION AND END-POINT DETECTION

All the speech utterances used in the present study are recorded in an ordinary office room using an Akai GX–365 tape recorder and an Altec–Lansing 681A dynamic microphone. The corpus chosen for evaluating the system is a phonetically balanced set of 40 English sentences.[1] These are spoken twice by a single male speaker whose mother tongue is Marathi (an Indian language spoken in the western part of the country). The recording is done in two different sessions separated by an interval of one week. The first set of 40 spoken sentences is used for training the recognition system, while the second set is used for testing the system.

The recorded signal is fed through a high-fidelity variable-gain amplifier to a fixed, passive lowpass filter (with a cutoff frequency of 8000 Hz and an attenuation of 36 dB per octave) which is used as a de-aliasing filter. The filtered signal is digitized using a sampling frequency of 20 kHz by means of a 12-bit analog-to-digital (A/D) converter connected to the CDC–160A Computer and stored on magnetic tape. For digitization, the quantization step size of the A/D converter is set manually by visually examining the audio signal amplitude of all the recorded utterances on an oscilloscope and adjusting the amplifier gain so that the full quantizer range is utilized without clipping.

In order to avoid the unnecessary processing of silence intervals which precede and follow the actual speech utterance, it is necessary to determine the endpoints of each utterance. For this, the digitized signal is scanned forward from the beginning of the recording interval and backward from the end. Endpoint detection is accomplished by using an adaptive energy threshold (Rabiner and Sambur, 1975).

II. SEGMENTATION

For dividing the continuous speech signal into a sequence of phoneme sized acoustic segments, a recognition-then-segmentation approach is employed. In this approach, the speech signal is analyzed at the rate of 100 frames per second and each frame is independently classified into a phonemic class. Segmentation is achieved by noting a change in frame classification. Reddy *et al.* (1973) and Dixon and Silverman (1976) classified every frame into 40 phonemic classes making a straightforward implementation of a classification algorithm. This, however, is quite expensive because the distance measure has to be calculated for each of the 40 phonemic classes. In the present effort, these phonemes are grouped into five broad categories: (1) vowel (VO: /i/, /ɪ/, /e/, /æ/, /ʌ/, /ɑ/, /ɔ/, /o/, /U/, /u/); (2) vowel-like (VL: /m/, /n/, /ŋ/, /j/, /r/, /l/, /w/); (3) voiced stop (VS: /b/, /ð/, /d/, /g/); (4) unvoiced stop [US: /p/, /t/, /k/, silence (denoted by SI)]; and (5) fricative (FR: /f/, /θ/, /h/, /s/, /ʃ/, /z/).[2] This grouping of phonemes into fewer categories not only saves computation time but also reduces frame classification errors, thus leading to a smaller number of extra phonetic boundaries detected by the segmenter in the speech signal.

The following six parameters are selected for frame classification: total energy, voice frequency energy (80 to 300 Hz), low-frequency energy (300 to 1000 Hz), mid-frequency energy (1000 to 3200 Hz), high-frequency energy (3200 to 7000 Hz), and zero crossing rate. These parameters are found to provide adequate discrimination between the classes (Paliwal and Rao, 1977; Paliwal, 1978). The parameters total energy and zero crossing rate are measured for every frame directly from the speech waveform, using a rectangular window of 10 ms duration. The other four parameters are measured from the power spectrum. In order to compute the power spectrum, the speech waveform is weighted by a 12.2-ms Hamming window and then subjected to a 256-point discrete Fourier transformation using a radix-2 fast Fourier transform (FFT) algorithm.

The six-dimensional pattern so obtained for each frame is classified into one of the five classes using a pattern classifier which needs: (1) a definition of a distance measure and (2) reference patterns for each of the five classes in a six-dimensional parametric space. In the present study, a weighted Euclidean distance measure is used and its value d_i for the ith class is given by

$$d_i^2 = \sum_{j=1}^{N} [w_{ij}(X_j - m_{ij})]^2, \tag{1}$$

where N is the dimensionality of the space (here it is 6), m_{ij} and w_{ij} are jth components of the mean vector and the weight vector, respectively, of the ith class and X_j is the jth component of the input pattern to be classified. The class-conditional weights w_{ij} associated with different parameters are taken here to be proportional to the standard deviations σ_{ij} of the respective parameters (Sebestyen, 1962). The input pattern X is classified into category i if

$$d_i < d_j, \tag{2}$$

for all $j \neq i$.

The reference patterns (mean vector and standard de-

TABLE I. Confusion matrix for frame classification. (Frame classification accuracy = 83.9%.)

Intended \ Classified	VO	VL	VS	US	FR
VO	79.9%	19.0%	0.6%	0.1%	0.4%
VL	17.2%	71.9%	10.6%	0.1%	0.2%
VS	0.0%	3.8%	90.7%	4.8%	0.7%
US	0.0%	0.0%	3.2%	96.3%	0.5%
FR	0.3%	5.0%	1.9%	7.6%	85.2%

viation vector) are computed for each class from data in the training set as follows:

$$m_{ij} = \frac{1}{N_i} \sum_{k=1}^{N_i} X_{ijk} \tag{3}$$

and

$$\sigma_{ij}^2 = \frac{1}{N_i} \sum_{k=1}^{N_i} (X_{ijk} - m_{ij})^2, \tag{4}$$

where N_i is the number of preclassified training patterns in the ith class and X_{ijk} the jth component of the kth training pattern of the ith class. The performance of the frame classifier on the test set data is shown in Table I in the form of a confusion matrix. The classifier achieves a frame classification accuracy of 83.9%. Most of the confusions that occur are between VO and VL classes.

Once frame classification is completed for an entire utterance, segmentation of this utterance is carried out in two steps. In the first step, assuming the minimum duration of each phonemic segment to be 30 ms, isolated 10-ms frames bearing labels which are different from those of both their neighbors are taken to have been wrongly classified and are corrected. In the second step, contiguous frames with identical labels are grouped together to form individual acoustic segments and segment boundaries are inserted where the labels of two adjacent frames differ. In cases where two adjacent phonemes belong to the same category (as /b/ and /d/ in the word "robbed," /ɑ/ and /ʊ/ in "house," /r/ and /l/ in "girl"), segment boundaries are missed at this stage. Such composite segments are detected on the basis of their relatively long durations, divided into two equal segments and treated independently.

TABLE III. Confusion matrix for segment classification. (Segment classification accuracy = 89.9%.)

Intended \ Classified	VO	VL	VS	US	FR
VO	86.5%	13.0%	0.5%	0.0%	0.0%
VL	7.1%	87.2%	5.7%	0.0%	0.0%
VS	0.0%	1.1%	96.6%	2.2%	0.0%
US	0.0%	0.0%	3.1%	96.9%	0.0%
FR	0.0%	1.7%	0.9%	5.2%	92.2%

The performance of the segmentation system is evaluated by comparing the machine obtained segment boundaries with those obtained manually from digital spectrograms and is judged on the basis of two types of errors: the missing segment error and the extra segment error. When the system is used to segment the 40 test utterances which have 971 manually derived segments, 52 segment boundaries are missed and 81 segment boundaries are wrongly inserted by the system. The segmentation results thus show 5.4% missed segment error and 8.3% extra segment error. These results, summarized in Table II, are comparable to those reported earlier in the literature (Goldberg, 1975; Schwartz and Makhoul, 1975; Baker, 1975; Dixon and Silverman, 1976, 1977).

The performance of the system for classification of segments into the five broad categories is given in the form of a confusion matrix in Table III. Segment classification accuracy is found here to be 89.9%. Most of the classification errors are due to confusion between VO and VL categories. This is comparable to the segment classification accuracy of 88.6% reported by Dixon and Silverman (1976) at the phoneme-class level (for seven phoneme classes: silence, voiced stop, nasal, aspiration, fricative, glide, vowel-like).

III. STEADY-STATE RECOGNITION

The steady-state recognition system labels the acoustic segments on the basis of their steady-state proper-

TABLE II. Comparison of our results on segmentation with those reported by others.

Reference	Missed segments	Extra segments	Other remarks
Goldberg (1975)	3.7%	27.6%	Tested on 1085 segments in 40 sentences spoken by a single speaker.
Schwartz and Makhoul (1975)	5.1%	5.4%	Tested on 473 segments in 15 sentences spoken by five male speakers.
Baker (1975)	9.3%	17.6%	Tested on 216 segments in five sentences spoken by four male and one female speakers.
Dixon and Silverman (1976)	6.9%	10.5%	Tested on 6175 segments in 8.5 min of speech, spoken by a single speaker.
Dixon and Silverman (1977)	6.19%	6.07%	Tested on 1507 segments in 2 min of speech, spoken by a single speaker.
Ours	5.4%	8.3%	Tested on 971 segments in 40 sentences spoken by a single speaker.

TABLE IV. Confusion matrix for phonemic recognition of continuous speech using steady-state segments. (Phoneme recognition accuracy = 52.3%.)

Spoken phoneme	Recognized phoneme																															
	ʌ	ɑ	ɪ	i	ʊ	u	e	æ	o	ɔ	m	n	ŋ	j	r	l	w	b	ð	d	g	p	t	k	SI	f	θ	h	s	ʃ	z	Total
ʌ	30	21	0	3	3	0	0	7	10	2	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78
ɑ	1	36	0	0	1	0	0	7	0	4	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52
ɪ	0	1	18	20	0	0	2	4	8	0	0	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58
i	0	0	2	25	1	1	0	1	1	0	3	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39
ʊ	0	0	1	0	1	2	0	0	6	0	0	1	3	0	0	0	10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	25
u	0	0	0	0	0	4	0	0	2	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
e	1	0	10	0	1	0	10	8	4	0	0	1	0	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40
æ	0	12	3	1	0	0	0	30	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48
o	2	0	0	0	1	0	0	0	20	0	0	0	0	0	0	0	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	31
ɔ	1	21	0	0	0	0	0	2	1	13	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39
m	0	0	0	0	0	0	0	0	0	0	28	0	5	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34
n	0	0	0	1	0	0	0	0	0	0	10	39	15	2	2	0	3	0	0	2	1	0	0	0	0	0	0	0	0	0	0	75
ŋ	0	0	0	0	0	0	0	0	0	0	1	0	6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
j	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
r	1	0	0	0	0	0	0	0	1	0	0	1	8	0	18	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	32
l	2	0	1	3	1	0	0	1	3	0	1	1	5	1	1	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	23
w	0	0	0	0	0	0	0	0	0	0	0	0	8	1	1	0	18	4	0	1	1	0	0	0	0	0	0	0	0	0	0	34
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	14	2	0	0	0	0	0	1	0	0	0	0	0	0	17
ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	16	2	9	0	0	1	0	0	0	0	0	0	0	36
d	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	9	0	4	13	0	0	0	0	0	0	0	0	0	0	27
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	6	0	0	0	0	0	0	0	0	0	0	9
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	3	3	0	0	0	0	0	0	13
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	7	4	12	7	0	0	0	0	0	0	31
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	8	2	0	0	0	0	0	0	11
SI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	10	4	9	81	0	0	0	0	0	0	108
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	0	8	0	1	0	0	0	14
θ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	3
h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	5	0	0	1	8
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	42	1	5	50
ʃ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	1	19
z	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	4	2	8	21
Total	38	91	35	53	9	7	12	60	57	19	43	49	55	13	28	3	50	39	20	12	35	27	8	37	94	9	6	8	46	21	16	1000

ties into 31 phoneme classes. For this, nine parameters are used. The first six parameters are the same as those used for the frame classification process described earlier. The remaining three parameters are the frequencies of the first three formants for vocalic sounds and energies in the ranges 3000 to 4200, 4200 to 5800, and 5800 to 7000 Hz for nonvocalic sounds. These energies are computed directly from the power spectrum evaluated already. Formant frequencies can be extracted from the power spectrum either by using an analysis-by-synthesis procedure (Bell *et al.*, 1961) or by means of a peak-picking procedure. Since the analysis-by-synthesis procedure is iterative in nature and needs too much computation time, the peak-picking procedure is used for extracting formants.

For peak-picking, the power spectrum has to be smoothed. This has been done by using the selective linear prediction technique (Makhoul, 1975) which gives a better smoothed spectrum for formant extraction than that obtained by using the cepstrum smoothing method (Schafer and Rabiner, 1970). A portion of the power spectrum from 0 to 3350 Hz is selected for smoothing and the ten predictor coefficients are computed. The smoothed spectrum in the selected range is obtained by computing a 128-point discrete Fourier transform of these predictor coefficients by using a pruned decimation-in-time FFT algorithm (Skinner, 1976). The peaks in the range 0 to 3200 Hz are picked and form the raw data from which the formants can be extracted by using an algorithm suggested by Markel (1973). The formant trajectories so obtained are smoothed using a nonlinear 3-point median filter (Rabiner *et al.*, 1975) and a zero phase 3-point Hanning filter (coefficients $\frac{1}{4}$, $\frac{1}{2}$, $\frac{1}{4}$).

The segmentation system described in the preceding section partitions the continuous speech signal into a sequence of phoneme-sized acoustic segments with each segment classified as one of the five broad categories. The acoustic segments not belonging to the vowel category (i.e., consonant segments) are assumed not to contain any transitions and treated here, for recognition purposes, as steady-state segments. The acoustic segments belonging to the vowel category are assumed to include steady-state and transition segments. These are further segmented on the basis of transitions of the second formant which, among the formants, has the maximum variability and carries the highest functional load. This is accomplished in two steps. First, only the middle 20% of the segment is considered as the steady-state segment and the average value of the second formant frequency is computed over this segment. The steady-state segment is then extended in both directions until the absolute deviation of the second formant frequency from its average value exceeds 5% of the average. This algorithm works in most cases. However, where the transition in the second formant is very small, the entire vowel segment gets detected as the steady-state segment. In such cases, 40% of the segment in the middle is considered as the steady-state segment and the remaining segments on both sides are labeled as transitions.

The values of the nine parameters are averaged over each of the steady-state segments. The nine-dimensional vector so obtained is then classified into one of the 31 phoneme classes, using a weighted Euclidean distance classifier given by Eqs. (1) and (2), by matching it with only those phonemes which belong to the broad category into which the segment has been already classified by the segmentation system, thus reducing the time taken by the classifier. The reference patterns (mean and standard deviation vectors) needed by the weighted Euclidean distance classifier are computed from the manually labeled steady-state segments of the 40 training utterances by using Eqs. (3) and (4).

The steady-state recognition system has been tried out on the test set of 40 sentences. Table IV presents the results obtained in the form of a confusion matrix.[3] The system achieves phonemic transcription with 52.3% accuracy. The recognition score increases to 72.6% if the first two choices are considered and to 82.0% if the first three choices are included. These results compare favorably with those reported in the literature. Goldberg (1975), for instance, has reported (for 29 phoneme classes) a recognition accuracy of 28.7% in the first choice, 44.4% in the first two choices, and 54.6% in the first three choices. Dixon and Silverman (1977) have achieved a phoneme recognition accuracy of 57.9% using only the steady-state properties of the acoustic segments.

IV. SYNTHESIS-BASED RECOGNITION

The synthesis-based recognition system uses both steady-state and transition segments for recognizing continuous speech. The recognition procedure considers a transition segment and the two steady-state segments on either side of it at a time. The steady-state segments are tentatively recognized by the steady-state recognizer and the M best phonemic choices are retained along with their distance measures. The control component passes the two lists of tentatively recognized phonemes, each containing M elements, to a synthesizer which in turn generates trial patterns representing the transition segments for all the M^2 combinations possible for the phoneme pair. These synthesized transition segments are compared with the input transition segment stored in the temporary store using a weighted Euclidean distance measure and the results of this comparison are transferred to the control component. The control component then combines the distances for the two steady-state segments and the transition segment and outputs a list of most likely phoneme pairs along with their confidence measures.

Here, two points should be noted. First, the steady-state recognizer not only contributes to the decision measure used for final recognition, but also limits the number of transition segments required to be synthesized by the internal synthesizer. If the value of M is 3 and the total number of phonemes in the vocabulary is 30, this reduction in number is by a factor of 100. Secondly, since all the nine trial patterns generated internally are compared with the input pattern, the system is capable of providing the recognition output in the form of a list of most probable phoneme pairs along with their good-

ness measures. This leaves scope for a final choice to be made in each case using higher level linguistic information.

Formant transitions are among the most important and extensively studied context effects and known to play a significant role in the human perception of stops (Delattre *et al.*, 1955) and vowels (Lindblom and Studdert-Kennedy, 1967). So the first three formant frequencies are the only parameters used here to represent speech during the transition segments and only consonant–vowel (CV) and vowel–consonant (VC) types of transition segments are used in the recognition process. Holmes *et al.* (1964) and Rao and Thosar (1974) have studied speech synthesis-by-rule using linear interpolation of formants during transitions and found the resulting synthetic speech to be highly intelligible. So for synthesizing the transition segments, the time variation of each of the three formants is assumed here to be linear.

In order to generate the trial transition segments for each phoneme pair, the synthesizer needs information about the duration of the transition segment, the formant-transition slopes and the formant frequencies at the vowel end of the transition segment. The duration of the synthesized transition segment is adjusted to be the same as that of the input transition segment. The values of the formant transition slopes are taken from a prestored table. (The procedure for computing the prestored values of the formant transition slopes for all the consonants occurring in all the vowel contexts is described in the footnote.[4]) The formant frequencies at the vowel end of the transition segment are adjusted to be the same as those computed from the input speech signal for the end frame of the steady-state vowel segment (i.e., the frame immediately next to the transition segment). This ensures that variations in vowel formants from utterance to utterance do not interfere in the recognition process.

The procedure for synthesizing the trial transition segments is illustrated in Fig. 3. The figure shows a hypothetical spectrographic display (showing a formant trajectory) of a CV utterance. Two vertical lines (at $t=t_1$ and $t=t_2$) divide the CV utterance into three segments: steady-state consonant segment ($t<t_1$), transition segment ($t_1 \leq t < t_2$), and steady-state vowel segment ($t \geq t_2$). The computed formant frequencies for the input speech signal are shown here by closed circles. Assume that the M best phonemic choices given by the steady-state recognizer are $\{C_i,\ i=1,2,\ \ldots, M\}$ for the steady-state consonant segment and $\{V_i,\ i=1,2,\ \ldots, M\}$ for the steady-state vowel segment. The synthesizer then has to synthesize M^2 trial transition segments for M^2 combinations $\{(C_i, V_j),\ i=1,2,\ldots,M, j=1,2,\ldots,M\}$. This is done by using the following linear equation

$$f_{ij}(t)=s_{ij}(t-t_2)+f_{\text{ter}},\quad t_1 \leq t < t_2, \tag{5}$$

where $f_{ij}(t)$ is the formant frequency of the synthesized transition segment for the combination (C_i, V_j) at time t, s_{ij} the formant-transition slope taken from the prestored table for the combination (C_i, V_j), and f_{ter} the terminal formant frequency at the vowel end of the transition segment to be synthesized and is taken to be equal to the formant frequency value computed from the input speech signal for the first frame (i.e., at $t=t_2$) of the steady-state vowel segment. The synthesized transition segments for the M^2 combinations are shown in the figure by the dashed lines.

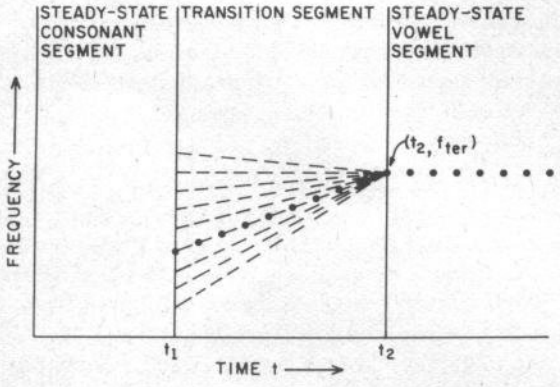

FIG. 3. Illustration of how to synthesize the transition segments. Formant values computed from the input speech signal are shown by closed circles. Dashed lines show synthesized transition segments.

For comparing the synthesized trial transition segments with the input transition segment, the following distance measure is defined

$$d_i^2=\sum_{j=1}^{N}\left(\sum_{k=1}^{3}[W_k(F_{jk}-f_{jk}^i)]^2\right), \tag{6}$$

where d_i is the distance of ith phoneme pair, N is the duration of the transition segment, W_k is the weight associated to kth formant, F_{jk} and f_{jk}^i are the frequencies of the kth formant at the jth point of transition for input transition segment, and the ith trial transition segment, respectively. The weights W_1, W_2, and W_3 associated with the first three formants are fixed here to 0.42, 0.5, and 0.08, respectively.

For final recognition using both steady-state and transition segments, the distances for the two steady-state segments and the transition segment are combined into a distance d given by

$$d=d_{\text{SI}}+W\cdot d_{\text{TR}}+d_{\text{SI}}, \tag{7}$$

where d_{SI} and d_{SI} are the distances for initial and final steady-state segments and d_{TR} is the distance for the transition segment. Here, both steady-state distances are given equal weights, but the transition distance is weighted by a factor W with respect to the steady-state distances. The weight W is fixed here to a value that gives the best recognition results on the training set. Before the steady-state and transition distances are used in Eq. (7), they are normalized as follows: if M is the number of choices given by the recognizer and d_i' represents the distance associated with the ith choice, then the normalized distance d_i associated with ith choice is given by

$$d_i=(M\cdot d_i')\Big/\sum_{j=1}^{M}d_j'. \tag{8}$$

The synthesis-based recognition system gives at its output a sequence of recognized phoneme pairs. Thus a

TABLE V. Confusion matrix for phonemic recognition of continous speech using both steady-state and transition segments. (Phoneme recognition accuracy = 62.0%.)

Spoken phoneme	Recognized phoneme: ʌ	ɑ	I	i	ʊ	u	e	æ	o	ɔ	m	n	ŋ	j	r	l	w	b	ð	d	g	p	t	k	SI	f	θ	h	s	ʃ	z	Total
ʌ	40	19	0	3	2	0	0	6	5	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	78
ɑ	1	40	0	0	1	0	0	5	0	2	0	1	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	52
I	0	1	33	11	0	0	2	4	2	0	0	2	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	58
i	0	0	2	26	0	1	0	1	1	0	3	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39
U	0	0	0	0	4	2	0	0	4	0	0	1	3	0	0	0	10	0	0	0	1	0	0	0	0	0	0	0	0	0	0	25
u	0	0	0	0	0	4	0	0	2	0	0	0	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13
e	1	0	6	0	1	0	16	8	2	0	0	1	0	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40
æ	0	4	2	1	0	0	0	39	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	48
o	2	0	0	0	1	0	0	0	20	0	0	0	0	0	0	0	7	0	1	0	0	0	0	0	0	0	0	0	0	0	0	31
ɔ	1	13	0	0	0	0	0	2	1	21	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	39
m	0	0	0	0	0	0	0	0	0	0	29	0	4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	34
n	0	0	0	1	0	0	0	0	0	0	7	47	13	2	2	0	0	0	0	2	1	0	0	0	0	0	0	0	0	0	0	75
ŋ	0	0	0	0	0	0	0	0	0	0	1	0	6	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8
j	0	0	0	0	0	0	0	0	1	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
r	1	0	0	0	0	0	0	0	1	0	0	1	5	0	21	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	32
l	2	0	1	3	1	0	0	1	3	0	1	1	5	1	1	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	23
w	0	0	0	0	0	0	0	0	0	0	0	0	6	1	1	0	20	4	0	1	1	0	0	0	0	0	0	0	0	0	0	34
b	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	15	1	0	0	0	0	0	1	0	0	0	0	0	0	17
ð	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4	23	2	6	0	0	1	0	0	0	0	0	0	0	36
d	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	9	0	8	9	0	0	0	0	0	0	0	0	0	0	27
g	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	1	6	0	0	0	0	0	0	0	0	0	0	9
p	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	8	0	2	3	0	0	0	0	0	0	13
t	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	12	9	6	0	0	0	0	0	0	31
k	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	8	2	0	0	0	0	0	0	11
SI	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	9	4	7	84	0	0	0	0	0	0	108
f	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	0	8	0	1	0	0	0	14
θ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	1	3
h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	6	0	0	1	8
s	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	43	0	5	50
ʃ	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	18	1	19
z	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	3	2	3	2	9	21
Total	48	77	44	45	10	7	18	66	42	24	41	57	47	13	31	3	49	36	26	16	28	23	16	31	96	9	5	9	46	20	17	1000

phoneme B in the phoneme sequence ABC is recognized as B1 using the context effects between phonemes A and B and as B2 using the context effects between B and C. For evaluating the performance of the recognition system, the phoneme B is considered to be recognized as B1 or B2 depending on the distance measures associated with them.

The system has been tried out on a test set of 40 utterances. Using both steady-state and transition segments, a recognition score of 62.0% is obtained in the first choice, 72.0% in the first two choices, and 78.5% in the first three choices. The recognition results are shown in Table V in the form of a confusion matrix.

V. DISCUSSION

As mentioned earlier, the steady-state recognition system recognizes 52.3% of the phonemes in the first choice. The incorporation of transition segments in the recognition system using the synthesis-based recognition approach improves the phoneme recognition score from 52.3% to 62.0%.[5] This improvement in recognition score is significant at the 99% confidence level (Wilks, 1949, p. 218) and points to the usefulness of transition segments in phonemic recognition of continuous speech. These results compare favorably with those reported by Dixon and Silverman (1977) who obtained phoneme recognition scores of 57.9% using only steady-state segments and 63.3% using both steady-state and transition segments. The Dixon and Silverman system, it may be noted, stores transition segments (sampled at seven time instants) for all possible combinations of phoneme pairs and uses the whole power spectrum to represent these transition segments. The present system, on the other hand, uses only CV and VC types of linearly synthesized transition segments for matching with the input transition segment and only the first three formants to represent these transition segments.[6]

A major advantage of the present synthesis-based recognition approach is that the acoustic variability in the input speech signal which occurs from utterance to utterance due to different articulation speeds and other factors is compensated here by the adaptive nature of the synthesized transition segments (i.e., by taking the formant frequencies at the vowel end of the synthesized transition segment to be the same as those computed from the input speech signal for the first frame of the steady-state vowel segment). This makes the recognition system less prone to errors. Such compensation for the acoustic variability cannot easily be provided for in the stored transition segment approach (Dixon and Silverman, 1977). Another advantage of synthesis-based recognition approach is that it requires comparatively less memory than the stored transition segment approach. The computational effort in synthesizing the linear transition segments is very little; this is comparable to the computations involved in time normalization of stored transition segments in the other approach.

VI. CONCLUSION

In this paper, a phonemic recognition system for continuous speech has been described. The system utilizes both steady-state and transition segments of the speech signal for recognition. The acoustic information contained in the transition segments is incorporated into the system using a synthesis-based recognition approach. It has been shown that utilization of interphonemic contextual effects contained in formant transitions improves the performance of the steady-state recognizer considerably.

Though the recognition system as implemented here utilizes context effects between immediate neighbors only, it can, in principle, also incorporate context effects between nonadjacent phonemes. Since the synthesis-based recognition approach offers an economic method of making use of transition segments, it is comparatively easy to train the system for a new speaker. The present system depends heavily on the correct estimation of formants within the transition segments. In order to follow fast transitions faithfully, the formant extraction method should be able to estimate the formant frequencies correctly for small analysis intervals. A pitch-asynchronous formant extraction method which yields correct estimates of formant frequencies for such small analysis intervals of voiced speech is under development.

ACKNOWLEDGMENT

The authors are thankful to the referees for their constructive comments.

[1]The 40 English sentences used in the present investigation are taken from the lists 49, 50, 71, and 72 of the report "IEEE recommended practice for speech quality measurement," IEEE Trans. Audio Electroacoust. AU-17, 225–246 (1969).

[2]In the present paper, phonemes /v/ (as in "vote") and /ʒ/ (as in "azure") are not considered because these do not occur in Indian languages. Most Indian speakers of English therefore substitute these by the phonemes /w/ (as in "will" and /z/ (as in "zoo"), respectively. Also, the phoneme /ð/ (as in "then") is articulated as a stop and does not show any fricativelike noise. This is therefore treated as a voiced stop.

[3]While reporting the recognition results on the basis of steady-state information, some of the recognition systems reported in the literature (Reddy, 1967; Dixon and Silverman, 1977) group all the voiced stops into one class and all the unvoiced stops into another class. This is done due to the availability of very little steady-state information to discriminate these stops individually. In the present paper, we report the steady-state recognition results for these stops individually (as shown in Table IV) because of two reasons. First, we do not want to discard any acoustic information which may be present in the steady-state segments due to the stop burst. Second, even if this information is very little, this would be reflected by the distance measure associated with each phonemic choice given by the steady-state recognizer.

[4]A question that arises is whether the data in the training set itself can be used for deriving the formant-transition slopes for synthesizing the trial transition segments. The training set used being too small to have in it all the consonants in all the vowel contexts, the synthesizer would then not be able to generate trial transition segments for all the M^2 combinations of phoneme pairs. It would be artificially forced to trim down the list of M^2 combinations of phoneme pairs. This would bias the recognition process, pushing recognition scores upwards if the texts for the training and test sets are the same and causing the performance to deteriorate unduly if they

are different. This problem is solved here by recording once, all possible CV syllables in a carrier sentence "Shall I speak (CV) tomorrow?" and computing the values of formant-transition slopes from them. (For the phoneme /ŋ/, VC syllables are used for computing the formant-transition slopes.)

[5]If the sample (consisting of the data in the test set) is assumed to be drawn from a binomial distribution, the 95% confidence limits for the recognition score can be calculated from a formula given by Wilks (1949, p. 200). These limits are 52.3% ±3.1% using the steady-state segments and 62.0% ± 3.0% using both steady-state and transition segments.

[6]The usefulness of nonlinear time functions for synthesizing the transition segments is under investigation. Preliminary results indicate that using nonlinear transitions does not improve the recognition performance appreciably.

Baker, J. M. (1975). "A new-time domain analysis of human speech and other complex waveforms," Ph. D. dissertation, Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA.

Bell, C. G., Fujisaki, H., Heinz, J. M., Stevens, K. N., and House, A. S. (1961). "Reduction of speech spectra by analysis by synthesis technique," J. Acoust. Soc. Am. 33, 1725–1736.

Cook, C. (1976). "Word verification in a speech understanding system," 1976 IEEE Int. Conf. Rec. Acoust., Speech, Signal Process., 553–556.

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

Dixon, N. R., and Silverman, H. F. (1976). "A general language-operated decision implementation system (GLODIS): Its application to continuous speech segmentation," IEEE Trans. Acoust. Speech Signal Process. ASSP-24, 137–162.

Dixon, N. R., and Silverman, H. F. (1977). "The 1976 modular acoustic processor (MAP)," IEEE Trans. Acoust. Speech Signal Process. ASSP-25, 367–378.

Goldberg, H. G. (1975). "Segmentation and labeling: A comparative performance evaluation," Ph. D. dissertation, Computer Science Department, Carnegie-Mellon University, Pittsburgh, PA.

Hess, W. J. (1976). "A pitch-synchronous digital feature extraction system for phonemic recognition of speech," IEEE Trans. Acoust. Speech Signal Process. ASSP-24, 14–25.

Holmes, J. N., Mattingly, I. G., and Shearme, J. N. (1964). "Speech synthesis by rule," Lang. Speech 7, 127–143.

Klatt, D. H. (1974). "Word verification in a speech understanding system," in *Speech Recognition*, edited by D. R. Reddy (Academic, New York), pp. 321–341.

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

Makhoul, J. (1975). "Spectral linear prediction: Properties and applications," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 283–296.

Markel, J. D. (1973). "Application of a digital inverse filter for automatic formant and F_0 analysis," IEEE Trans. Audio Electroacoust. AU-21, 154–160.

Niederjohn, R. J., and Thomas, I. B. (1973). "Computer recognition of the continuant phonemes in connected English speech," IEEE Trans. Audio Electroacoust. AU-21, 526–535.

Paliwal, K. K., and Rao, P. V. S. (1977). "Acoustic phonetic recognition of continuous speech," 9th Int. Cong. Acoust., Madrid, Spain, Paper I–40.

Paliwal, K. K. (1978). "Computer recognition of continuous speech," Ph. D. thesis, University of Bombay.

Rabiner, L. R., and Sambur, M. R. (1975). "An algorithm for determining the endpoints of isolated utterances," Bell Syst. Tech. J. 54, 297–315.

Rabiner, L. R., Sambur, M. R., and Schmidt, C. E. (1975). "Applications of a nonlinear smoothing algorithm to speech processing," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 552–557.

Rao, P. V. S., and Thosar, R. B. (1974). "A programming system for studies in speech synthesis," IEEE Trans. Acoust. Speech Signal Process. ASSP-22, 217–225.

Reddy, D. R. (1967). "Computer recognition of connected speech," J. Acoust. Soc. Am. 42, 329–347.

Reddy, D. R., Erman, L. D., and Neely, R. B. (1973). "A model and a system for machine recognition of speech," IEEE Trans. Audio Electroacoust. AU-21, 229–238.

Schafer, R. W., and Rabiner, L. R. (1970). "System for automatic formant analysis of voiced speech," J. Acoust. Soc. Am. 47, 634–648.

Schwartz, R., and Makhoul, J. (1975). "Where the phonemes are: Dealing with ambiguity in acoustic-phonetic recognition," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 50–53.

Sebestyen, G. S. (1962). *Decision-Making Processes in Pattern Recognition* (MacMillan, New York), pp. 17–23.

Skinner, D. P. (1976). "Pruning the decimation in-time FFT algorithm," IEEE Trans. Acoust. Speech Signal Process. ASSP-24, 193–194.

Thosar, R. B., and Rao, P. V. S. (1971). "A software system for speech recognition," Proc. 7th Int. Cong. Acoust., Budapest, Hungary, Paper 19c–3.

Thosar, R. B., and Rao, P. V. S. (1976). "An approach towards a synthesis-based speech recognition system," IEEE Trans. Acoust. Speech Signal Process. ASSP-24, 194–196.

Weinstein, C. J., McCandless, S. S., Mondshin, L. F., and Zue, V. W. (1975). "A system for acoustic-phonetic analysis of continuous speech," IEEE Trans. Acoust. Speech Signal Process. ASSP-23, 54–67.

Wilks, S. S. (1949). *Elementary Statistical Analysis* (Princeton U. P., Princeton, NJ).