World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:1, No:8, 2007

Recognition of Isolated Speech Signals using Simplified Statistical Parameters

Abhijit Mitra, Bhargav Kumar Mitra and Biswajoy Chatterjee

Abstract— We present a novel scheme to recognize isolated speech signals using certain statistical parameters derived from those signals. The determination of the statistical estimates is based on extracted signal information rather than the original signal information in order to reduce the computational complexity. Subtle details of these estimates, after extracting the speech signal from ambience noise, are first exploited to segregate the polysyllabic words from the monosyllabic ones. Precise recognition of each distinct word is then carried out by analyzing the histogram, obtained from these information.

Keywords— Isolated speech signals, Block overlapping technique, Positive peaks, Histogram analysis.

I. Introduction

VOICE signals, considered as sequences of discrete states, are often regarded as stationary stochastic signals within any fixed state for convenient mathematical treatment. Such an approach, however, ignores the time-varying characteristics of speech signals and therefore, accurate recognition of speech signals is seldom achieved in practice. Nevertheless, many works have been reported in the literature [1]-[2] that exploit the stationary property of the speech signals for easier computational purpose. New models have also been developed that duly take into consideration the dynamics of speech, e.g., Hidden Dynamic Model [3] or Dynamic HMM Model [4], as well as time sequential specifics, e.g., Trend-HMM Models with discriminative training [5]-[6]. These models, however, increase the computational complexity for deploying stack decoding or restoring methods to describe the dynamic sequence. Certain other speech recognition techniques [7]-[8] have also dealt with the dynamics of speech taking into consideration the noise effects.

We present here a novel approach for proper detection of certain primitive isolated speech signals, mainly confined within monosyllabic and disyllabic words, which takes into account dynamic property of the speech yet reducing the computational load due to its simplicity. The major computational advantage associated with it stems from the fact that it is based on extracted signal parameters rather than the original samples of the speech waveform. In this method, the speech signal is first extracted from pre- and post-ambience noise by using a block-overlapping technique. Next, with the help of a moving average parameter, the words are differentiated based on the number of syllables, i.e., monosyllabic and disyllabic cases. A positive peak concept is then introduced and analyzing the histogram of such positive peaks, each of the words is recognized precisely. It is shown in the results of the proposed scheme that gradually 70% of the permanent memory space is set free at the later stages which makes the memory utilization of such an approach more convenient. Certain possible extensions of the approach are also discussed in the conclusion.

A. Mitra is with the Department of Electronics and Communication Engineering, Indian Institute of Technology (IIT) Guwahati, North Guwahati - 781039, India (e-mail: a.mitra@iitg.ernet.in).

B. K. Mitra is with the Department of Electronics and Communication Engineering, Institute of Engineering & Management (IEM), Kolkata, India (e-mail: bkmitra09@vsnl.net).

B. Chatterjee is with the Department of Computer Science and Engineering, Institute of Engineering & Management (IEM), Kolkata, India (e-mail: biswajoy@iemcal.com).

II. Speech Data Bank

The task of developing a voice recognition system was carried out as a part of implementing a voice enabled on-line examination system, where a speaker pronounces the following words: BEGIN, YES, NO and CANCEL to start, answer or terminate various questions respectively. To utter any of these words, the speaker was given two seconds of time so as to store almost 27,000 discrete samples of speech signal and the captured signal of this duration was stored as .wav file in the memory of the computer. Fig. 1 shows a typical example of YES, NO, CANCEL and BEGIN waveforms.

Note that, among the above mentioned four words, YES and NO belong to monosyllabic category while the two other belong to polysyllabic case.

The speech data bank consists of all of these four words of 40 different male speakers, which were taken in the morning, afternoon and nighttime. Physical illness was also taken into account. Two different environments were considered - windows off and climatizer off/on, in order to include the ambience noise effect.

III. The Proposed Scheme

The simple yet robust technique deployed here to recognize constrained speech signals is based on three operational modules. Phase wise, these are discussed below. Note that the entire scheme has been executed using the MATLAB software, taking all the data formats as 32 bits including sign.

A. Extraction of the isolated speech signal trait from a low noise environment

Because of the non-deterministic nature of the exact position of the speech signals within the specified 27,000 samples, the captured signal trait is corrupted by both (1) pre- and post-ambience noise and (2) ambience noise superimposed on the signal itself.

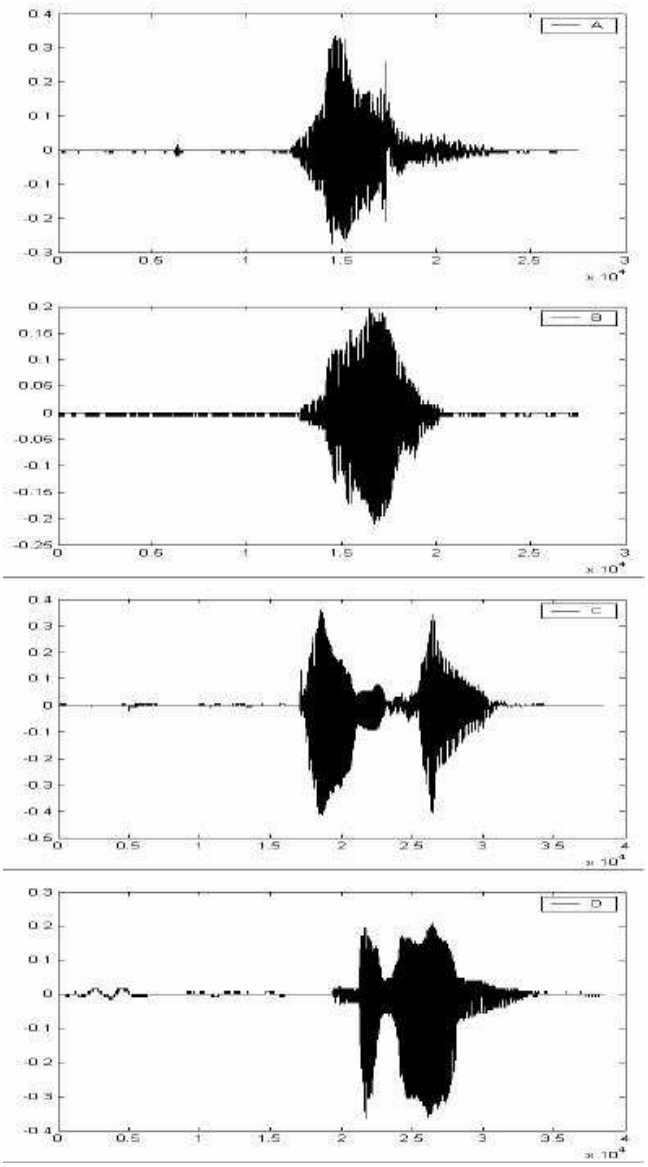

Fig. 1. Typical examples of captured speech waveforms of (A) YES (B) NO (C) CANCEL and (D) BEGIN.

The purpose of this module is to extract the actual constrained speech signal trait by neglecting the pre- and post-ambience noise. To serve this purpose, a *Block Overlapping Technique* has been used. In this technique, the captured signal is divided into several overlapped blocks, where the size of each block and overlap between two successive blocks have two different fixed values. The entire method is based on the observation that the sum of the absolute values of the sample-magnitudes for a moderate-length block is much less for the block containing samples of the spurious signal or ambience noise than that containing samples of the actual signal. Thus, after noting the experimental data, a threshold value was determined such that if the aforementioned parameter for a block is less than the threshold value then that block is rejected, otherwise accepted. The overlap between successive blocks was maintained so that even if we reject a particular block containing samples of the actual signal, the loss would be restricted to a value L where:

$$0 < L < N - N_1[= M] \tag{1}$$

with N being the number of samples in a block, $N_1(< N)$ denoting the number of overlapped samples between two adjacent blocks, and M being the number of non-overlapped samples for the same two adjacent blocks mentioned above. Moreover, it should also be noted if the first accepted block contains some spurious samples then also the number of such samples will be restricted by the value L as given in equation (1). If it is required to calculate the overall accuracy, then such losses at the front and tail end need to be considered and the total loss L_T is given by the following expression:

$$0 < L_T < 2(N - N_1)[= 2M]$$

The selection of the values of N and M was based on the following cost function:

$$f(M, N) = \left(\frac{n - N}{M} + 1\right) N \tag{2}$$

where n is the number of samples captured in 2 seconds (=27,000). Further investigation on extreme point detection using equation (2) revealed all the points with $M = N$ as saddle points. However, to meet the requirements and to optimize the computation we studied a statistics based on different values of N, M and the corresponding increments or decrements in the computational operations and accuracy of uncorrelated sample values. From such statistics it has been noted that $N = 600$ and $M = 200$ would give optimized results considering both the number of operations and % of overlapped samples per block $[= (1 - M/N) \times 100\%]$; and the value of the number of samples lost in these cases can be restricted to $(2 \times 199/27,000) \times 100\% = 0.15\%$ of that of the original signal. An estimate of such micro-operations vs. overlapped samples (%) is given in Table 1.

B. Segregation of the words based on number of syllables

Unlike the waveform of the monosyllabic words, the waveform of each polysyllabic word is seen to contain a distinct trough after the utterance of each syllable, if observed carefully. To implement this observational finding, a new parameter ρ_i has been conceived, defined as follows:

$$\rho_i = \frac{\sum_{j=1}^{N} |x_{i,j}|}{max_i\{\sum_{j=1}^{N} |x_{i,j}|\}} \tag{3}$$

where $x_{i,j}$ denotes the jth sample value of the i^{th} block, $max_i\{\sum |x_{i,j}|\}$ is the maximum value of the absolute sum of all the samples considering all the i^{th} blocks, where $i = \{1, 2, ..., \beta\}$, β is the number of blocks of size N available after trimming the signal and N denotes the number of samples per block (600 in our case). Bar plots of ρ_i vs. i for different signals are shown in Fig. 2. Next, with the help of this *moving average parameter* ρ_i, we define another threshold value ρ as follows:

$$\rho = \frac{1}{\beta\sqrt{2}} \sum_{i=1}^{\beta} \rho_i \tag{4}$$

TABLE I
NUMBER OF MICRO-OPERATIONS VERSUS % OF OVERLAPPED SAMPLES PER BLOCK, FOR DIFFERENT VALUES OF M AND N.

N	M	No. of micro-operations	Overlapped samples per block (%)
100	1	2690100	99
	50	53900	50
200	1	5360200	99.5
	50	107400	75
	100	53800	50
400	1	10640400	99.75
	100	106800	75
	200	53600	50
	300	35600	25
600	1	15840600	99.83
	100	159000	83.33
	200	79800	66.67
	300	53400	50
	400	40200	33.33
800	1	20960800	99.875
	100	210400	87.5000
	200	105600	75
	300	70400	62.5
	400	52800	50
	600	35200	25

which serves as the decision parameter of the above explained technique. The decision about a word, whether that is monosyllabic or polysyllabic, is taken according to the following algorithm: *Due to the increasing nature of the moving average curves at the initial values, the first few sample values of* ρ_i *are discarded and then all the next values are stored in the memory till* $\beta/2^{th}$ *index and if it is seen that* $\rho_i < \rho$ *for any five consecutive index values within this range, the word is decided as* ***polysyllabic****, otherwise the decision is taken in favor of* ***monosyllabic*** *case.*
After detecting a word according to the number of syllables, the next task left for us is to recognize each of them precisely under each category. This is discussed next.

C. Recognition of each distinct word

In a stochastic process like speech generation, it has frequently been observed that if frequency of amplitudes of the signal is taken then the histogram takes a definite shape reflecting the uniqueness in the pattern of the waveform. In order to simplify the above procedure, we considered here the positive peaks of the waveform. Any positive peak (say P_n) implies a particular peak which follows the following relationship:

$$P_{n-1} < P_n > P_{n+1} \tag{5}$$

where P_n , P_{n-1} and P_{n+1} are the n^{th}, $(n-1)^{th}$ and $(n+1)^{th}$ sample values respectively and each one of them is greater than zero.
Now the frequency of occurrence of each such positive peak samples are found out and the normalized frequency values are plotted against the sample values. A straight line on the average value (0.5) of each such normalized histogram is drawn next and the decision for each distinct word is taken according to the number of cuts on the histogram w.r.t. this line as follows:
If the histogram plot crosses this average line by a number $>= 3$*, the decision is taken as NO (for a monosyllabic word) or BEGIN (for a disyllabic word), otherwise it is taken in favor of YES or CANCEL respectively*.
Typical examples of such histograms along with the number of cuts for all the four signals of our context are shown in Fig. 3, which clearly depict the validity of our decision criteria. It has been observed that employing such a simple technique yields more than 98% accuracy in different ambiences, which is discussed in the next section.

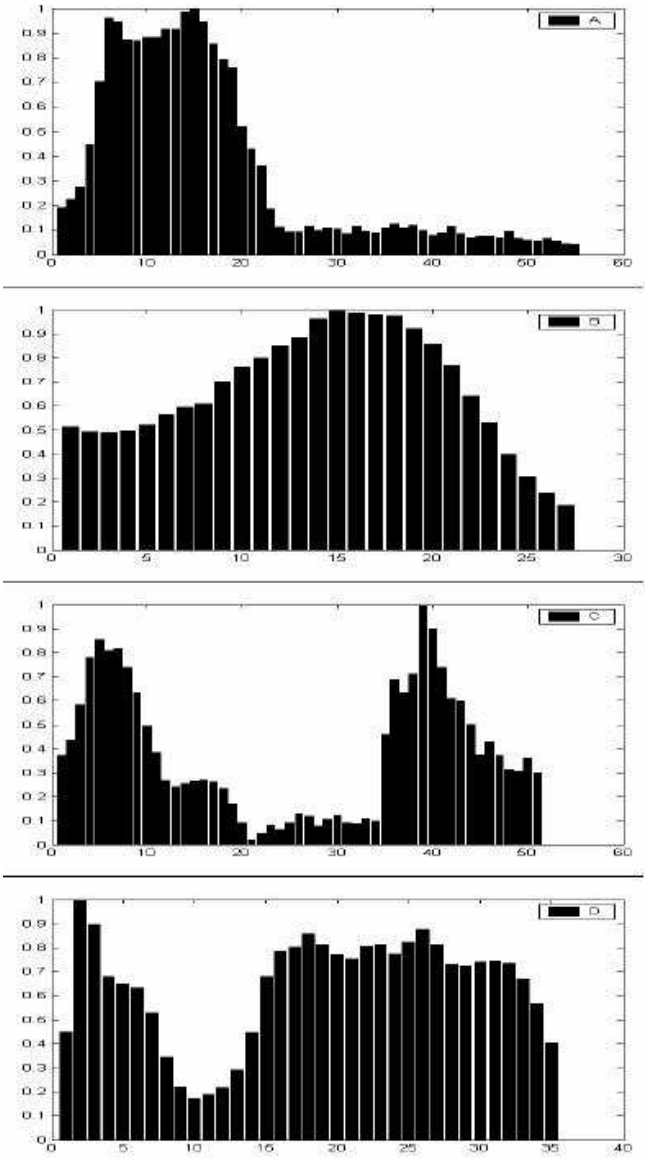

Fig. 2. Normalized bar plots of ρ_i (in Y axis) versus i (in X axis) for (A) YES (B) NO (C) CANCEL and (D) BEGIN waveforms.

IV. RESULTS AND DISCUSSIONS

The process was tested by 40 different male speakers with each of them speaking each word 10 times in 2 different environments, hence with a total of 3200 samples. For both environments, all windows were kept closed but for the first

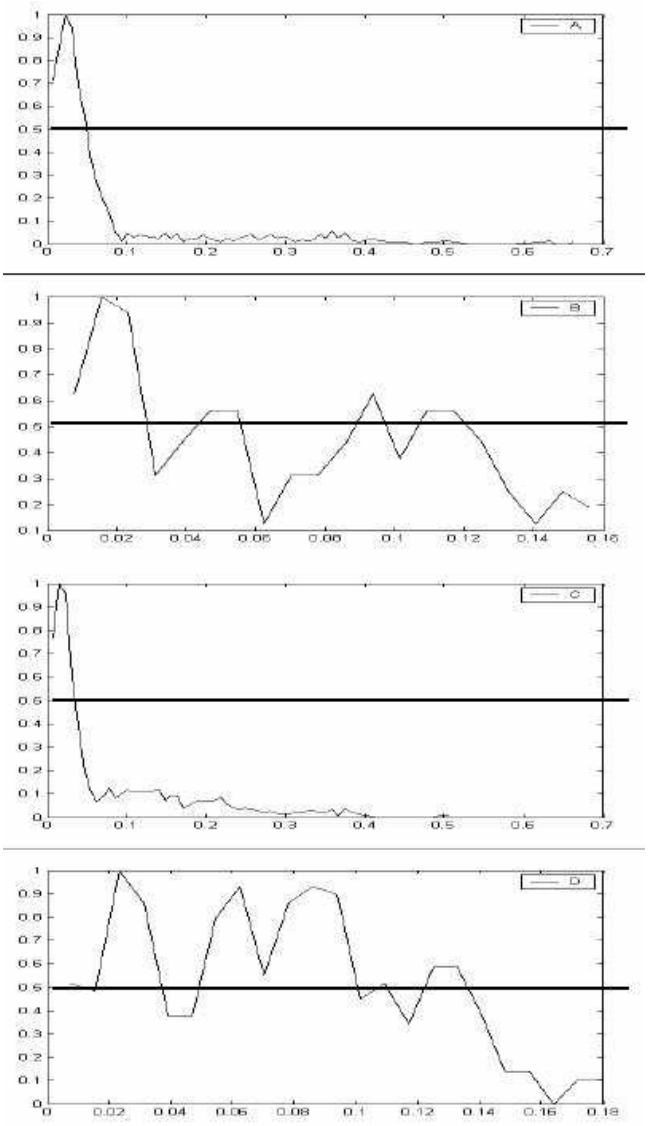

Fig. 3. Histogram plots of positive peaks and the corresponding number of cuts w.r.t. the average value 0.5 for (A) YES (B) NO (C) CANCEL and (D) BEGIN waveforms.

case, climatizer was kept off while for the other, it was kept on for including the ambience noise effect. For the first case, a success rate of 100% was achieved and for the second case the success rate was 98.25%. Table 2 shows the number of average cuts and therefore the accuracy of the proposed scheme for 10 different speakers. Another important aspect as revealed by the experimental statistics is the gradual release of the memory. Note that after trimming the signal, around 54% of the memory can be reallocated and in the final phase around 72% of the still captured memory space can be released. As time taken for completion of the first operation is around 60% of the total time, so after that time 54% of memory can be reallocated and after 90% of the total time required, 92% of the still used memory can be freed. Hence, memory management can also be considered as very efficient in the proposed technique.

TABLE II
NUMBER OF AVERAGE CUTS IN THE HISTOGRAMS OF POSITIVE PEAKS OF CORRESPONDING 10 DIFFERENT SPEAKERS (X IS THE SPEAKER SERIAL NUMBER, A1 DENOTES THE AMBIENCE WHERE WINDOWS AND CLIMATIZER ARE OFF AND A2 REFERS TO THE AMBIENCE WHERE WINDOWS ARE OFF BUT CLIMATIZER IS ON).

X	Yes		No		Cancel		Begin	
	A1	A2	A1	A2	A1	A2	A1	A2
1	1	1.2	5.6	5.4	1	1	8.1	7.4
2	1	1.1	6.1	5.9	1	1	7.2	7.2
3	1.1	1.1	4.8	4.3	1	1.1	6.7	6.9
4	1	1.3	4.8	5	1	1.1	7.5	6.9
5	1.2	1.2	6	6.1	1	1	7.3	7.4
6	1.1	1.1	5.5	5.4	1.1	1	8	6.7
7	1	1.1	5.4	5.4	1	1	7.2	7.9
8	1	1.3	5.2	5.3	1	1.2	5.6	5.4
9	1.1	1.1	5.4	5.1	1	1	6.5	6.5
10	1	1	4.6	3.9	1	1.1	7.1	6.2

V. CONCLUSIONS

In this paper, an effective approach is given to recognize certain primitive isolated speech signals by determining some statistical estimates. The scheme presented here takes into account the dynamic property of the speech and reduces the computational complexity as well. The speech recognition system, discussed here, has been carried out as a part of realizing a voice-enabled on-line examination system with only two options. Developing the same recognition system for multiple-choice questions with an option for changing the answer would be an excellent future topic of research in the analogous lines. Also, implementation of a voice-enabled username password system, which combines both the speech and speaker recognition for male and female separately, would be a good alternative for future study. Active investigation on the same topic is now being carried out by the same authors.

ACKNOWLEDGEMENT

The authors gratefully acknowledge the insightful suggestions of Dr. Goutam Saha of IIT-Kharagpur in several technical discussions with Dr. Abhijit Mitra during the initial development of this work.

REFERENCES

[1] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals*. Prentice- Hall, Englewood Cliffs, NJ, 1978.
[2] J. L. Flanagan, "Speech Coding", *IEEE Trans. on Communications*, vol. COM-27, April 1979. pp. 710-737.
[3] J. Picone *et. al.*, "Initial Evaluation of Hidden Dynamic Models on Conversational Speech", in *Proc. IEEE ICASSP*, Phoenix, Arizona, USA, May 1999.
[4] F. Chen and E. Chang, "A New Dynamic HMM Model for Speech Recognition", in *Proc. EUROSPEECH 2001*, Scandinavia, 2001.
[5] D. Sun, L. Deng and C. Wu, "State-dependent Time Warping in the Trended Hidden Markov Model", *Signal Processing*, vol. 39, no. 1, 1994. pp. 263-275.
[6] L. Deng, "Speech Recognition using Autosegmental Representation of Phonological Units with Interface to the Trended HMM", *Speech Communication*, vol. 23, 1997. pp. 211-222.
[7] A. Agarwal and Y. M Cheng, "Two-stage Mel Warped Wiener Filter for Robust Speech Recognition", in *Proc. ASRU*, December 12-15, 1999.
[8] L. Deng et. al., "Large-Vocabulary Speech Recognition under Adverse Acoustic Environments", in *Proc. ICSLP*, vol.3, pp. 806-809, 2000.