

A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications

G. Saha¹, Sandipan Chakroborty², Suman Senapati³

Department of Electronics and Electrical Communication Engineering

Indian Institute of Technology, Kharagpur, Kharagpur-721 302, India

Email: {gsaha@ece.iitkgp.ernet.in¹, sandipan@ece.iitkgp.ernet.in², speech_ece@rediffmail.com³}

Abstract

Pre-processing of Speech Signal serves various purposes in any speech processing application. It includes Noise Removal, Endpoint Detection, Pre-emphasis, Framing, Windowing, Echo Canceling etc. Out of these, silence/unvoiced portion removal along with endpoint detection is the fundamental step for applications like Speech and Speaker Recognition. The proposed method uses Probability Density Function (PDF) of the background noise and a Linear Pattern Classifier for classification of Voiced part of a speech from silence/unvoiced part. The work shows better end point detection as well as silence removal than conventional Zero Crossing Rate (ZCR) and Short Time Energy (STE) function methods.

1. Introduction

Pre-Processing of Speech Signal is very crucial in the applications where silence or background noise is completely undesirable. Applications like Speech and Speaker Recognition [1] needs efficient feature extraction techniques from speech signal where most of the voiced part contains Speech or Speaker specific attributes. Endpoint Detection [2],[3] as well as silence removal are well known techniques adopted for many years for this and also for dimensionality reduction in speech that facilitates the system to be computationally more efficient. This type of classification of speech into voiced or silence/unvoiced [4] sounds finds other applications mainly in Fundamental Frequency Estimation, Formant Extraction or Syllable Marking, Stop Consonant Identification and End Point Detection for isolated utterances.

There are several ways of classifying (labeling) events in speech. It is accepted convention to use a three-state representation in which states are (i) silence (S), where no speech is produced; (ii) unvoiced (U), in which the vocal cords [5] are not vibrating, so the resulting speech waveform is aperiodic or random in nature and (iii) voiced (V), in which the vocal chords are tensed and therefore vibrate periodically when air flows from the lungs, so the resulting waveform is quasi-periodic [6]. It should be clear that the segmentation of the waveform into well-defined regions of silence, unvoiced, signals is not exact; it is often difficult to distinguish a weak, unvoiced sound (like /f/ or /th/) from silence, or weak voiced sound (like /v/ or /m/) from unvoiced sounds or even silence.

However, it is usually not critical to segment the signal to a precision much less than several milliseconds; hence, small errors in boundary locations usually have no consequence for most applications. Since for most of the practical cases the unvoiced part has low energy content and thus silence (background noise) and unvoiced part is classified together as silence/unvoiced and is distinguished from voiced part.

Two widely accepted methods namely Short Time Energy (STE) [6],[7] and Zeros Crossing Rate (ZCR) [6],[7] have been used for a long time for silence removal. But they have their own limitation regarding setting thresholds as an ad hoc basis. STE uses the fact that energy in voiced sample is greater than silence/unvoiced sample. However, it is not specific about how much greater it needs to be for proper classification and varies case to case. On the other hand ZCR has a demarcation rule specifying that if the ZCR of a portion speech exceeds 50 then this portion will be labeled as unvoiced or background noise whereas any segment showing ZCR at about 12 is considered to be the voiced one. One attempt [8] was made by taking these two methods together and results reported only 65% accuracy with respect to manually labeled speech sample.

In this paper, we detect silence/unvoiced part from the speech sample using uni-dimensional Mahalanobis Distance [9] function which itself is a Linear Pattern Classifier [9],[10]. Our algorithm uses statistical properties of background noise as well as physiological aspect of speech production and does not assume any ad hoc threshold. We also show the algorithm's performance using the measure of correctness taking manually labeled speech as a reference. The experiments are done on two kinds of speeches which are a running text read from a paragraph and a combination lock number. The result shows better classification for the proposed method in both the cases when compared against conventional silence/unvoiced detection methods. We assume that background noise present in the utterances are Gaussian [11] in nature, however a speech signal may also be contaminated with different types of noise [12]. In such cases the corresponding properties of the noise distribution function are to be used for detection purpose.

This paper is organized as follows. In section 2 we describe the theoretical background. Section 3 presents the algorithm along with a short discussion regarding computational complexity and defining the measure of correctness. The results are presented in section 4 and section 5 describes the principal conclusion.

2. Theoretical Background

2.1 Speech Signal and its Basic Properties

The speech signal [13] is a slowly time varying signal [14] in the sense, that, when examined over a sufficiently short period of time (between 5 and 100 msec), its characteristics are fairly stationary; however, over long periods of time (on the order of 1/5 seconds or more) the signal characteristics change to reflect the different speech sounds being spoken. Usually first 200 msec or more (1600 samples if the sampling rate is 8000 samples/sec) of a speech recording corresponds to silence (or background noise) because the speaker takes some time to read when recording starts. Figure 1 illustrates the fact.

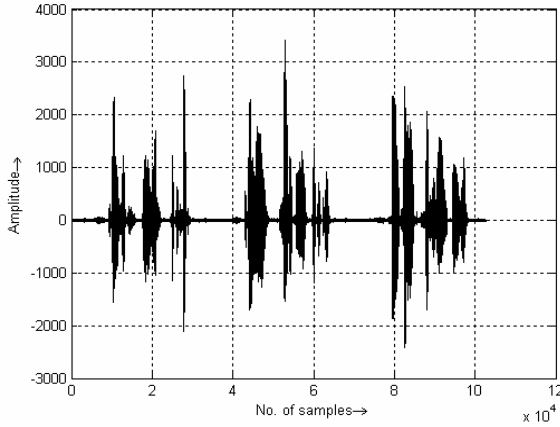


Fig. 1. Diagram of a typical Speech Signal

2.2 Gaussian or Normal Distribution

One of the most important results of the probability theory is the Central Limit Theorem [9], which states that, under various conditions, the distribution for the sum of d independent random variables approaches a particular limiting form known as the normal distribution. As such, the normal or Gaussian probability density function is very important, both for theoretical and practical reasons. In one dimension, it is defined by:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (1)$$

The normal density is traditionally described as ‘bell-shaped curve’; it is completely determined by the numerical values for two parameters, the mean μ and the variance σ^2 . This is often emphasized by writing $p(x) \sim N(\mu, \sigma^2)$, which is read as “ x is distributed normally with mean μ and variance σ^2 ”. The distribution is symmetrical about the mean, the peak occurring at $x=\mu$ and the width of the ‘bell’ is proportional to the standard deviation σ . Normally distributed data points tend to cluster about the mean. Numerically, the probabilities obey

$$\Pr[|x - \mu| \leq \sigma] \approx 0.68 \quad (2)$$

$$\Pr[|x - \mu| \leq 2\sigma] \approx 0.95 \quad (3)$$

$$\Pr[|x - \mu| \leq 3\sigma] \approx 0.997 \quad (4)$$

as shown in Fig. 2 given below :

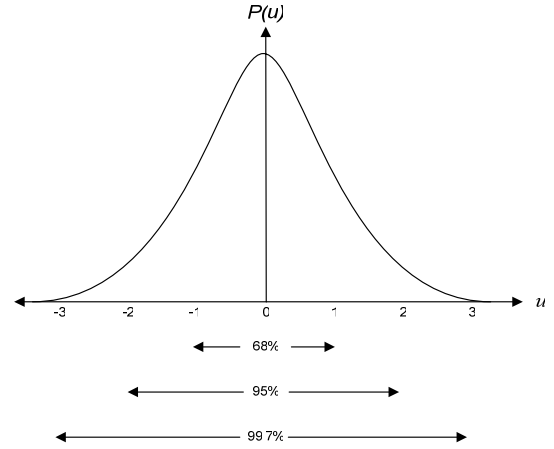


Fig. 2. A one-dimensional Gaussian distribution, $p(u) \sim N(0,1)$, has 68% of its probability mass in the range $|u| \leq 1$, 95% in the range of $|u| \leq 2$, and 99.7% in the range of $|u| \leq 3$.

A natural measure of the distance from x to the mean is the distance $|x-\mu|$ measured in units of standard deviation which can be analytically expressed as:

$$r = \frac{|x-\mu|}{\sigma} \quad (5)$$

and defined as ‘Mahalanobis Distance’ from x to μ (In the one-dimensional case, this is sometimes called z-score). Thus for instance the probability is 0.95 that the Mahalanobis distance from x to μ will be less than 2. If a random variable x is modified by (a) subtracting its mean and (b) dividing by its standard deviation, it is said to be standardized. Clearly, a standardized normal random variable $r=(x-\mu)/\sigma$ has zero mean and unit standard deviation—that is,

$$p(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \quad (6)$$

which can be written as $p(u) \sim N(0,1)$.

3. Method

3.1 The Algorithm

The algorithm described below is divided into two parts. First part assigns label to the samples by using a statistical properties of background noise while the second part smoothens the labeling by the physiological aspects from the speech production process. The Algorithm two passes over speech samples. In Pass I (Step 1 to 3) we use statistical property of background noise to make a sample

as voiced or silence/unvoiced. In Pass II (Step 4 and 5) we use physiological aspects of speech production for smoothening and reduction of probabilistic errors in statistical marking of Pass I.

Step 1: Calculate the mean and standard deviation of the first 1600 samples of the given utterance. If μ and σ are the mean and the standard deviation respectively then analytically we can write,

$$\mu = \frac{1}{1600} \sum_{i=1}^{1600} x(i) \quad (7)$$

$$\sigma = \sqrt{\frac{1}{1600} \sum_{i=1}^{1600} (x(i) - \mu)^2} \quad (8)$$

Note that background noise is characterized by this μ and σ .

Step 2: Go from 1st sample to the last sample of the speech recording. In each sample check whether one-dimensional Mahalanobis distance function i.e. $|x - \mu|/\sigma$ greater than 3 or not. Analytically,

$$\text{If, } \frac{|x - \mu|}{\sigma} > 3 \quad (9)$$

the sample is to be treated as voiced sample otherwise it is an silence/unvoiced.

Note that the threshold reject the samples upto 99.7% as per given by equation no. 4 in a Gaussian Distribution thus accepting only the voiced samples.

Step 3: Mark the voiced sample as 1 and unvoiced sample as 0. Divide the whole speech signal into 10 ms non-overlapping windows. Now the complete speech is represented by only zeros and ones.

Step 4: Consider there are M no. of zeros and N number of ones in a window. If $M \geq N$ then convert each of ones to zeros and vice versa. This method adopted here keeping in mind that a speech production system consisting of vocal chord, tongue, vocal tract etc. cannot change abruptly in a short period of time window taken here as 10 ms.

Step 5: Collect the voiced part only according to the labeled '1' samples from the windowed array and dump it in a new array. Retrieve the voiced part of the original speech signal from labeled 1 samples.

The algorithm is illustrated in the flow chart given in fig.3. Note that in the proposed method after μ and σ calculation it requires one division and condition checking per sample in Pass I and in Pass II one condition checking per 80 samples (10 ms). In ZCR method sign of each sample is checked and number of reversal in a window (80 samples, 10 ms) is calculated. Then the no. of sign reversal is checked if within certain range for voice part classification purpose. In STE method each sample sequence energy is calculated and then summed over 80

samples (10 ms) window. Then a condition checking is done on this sum to classify it as voiced part or not. Thus the proposed method is computationally comparable to

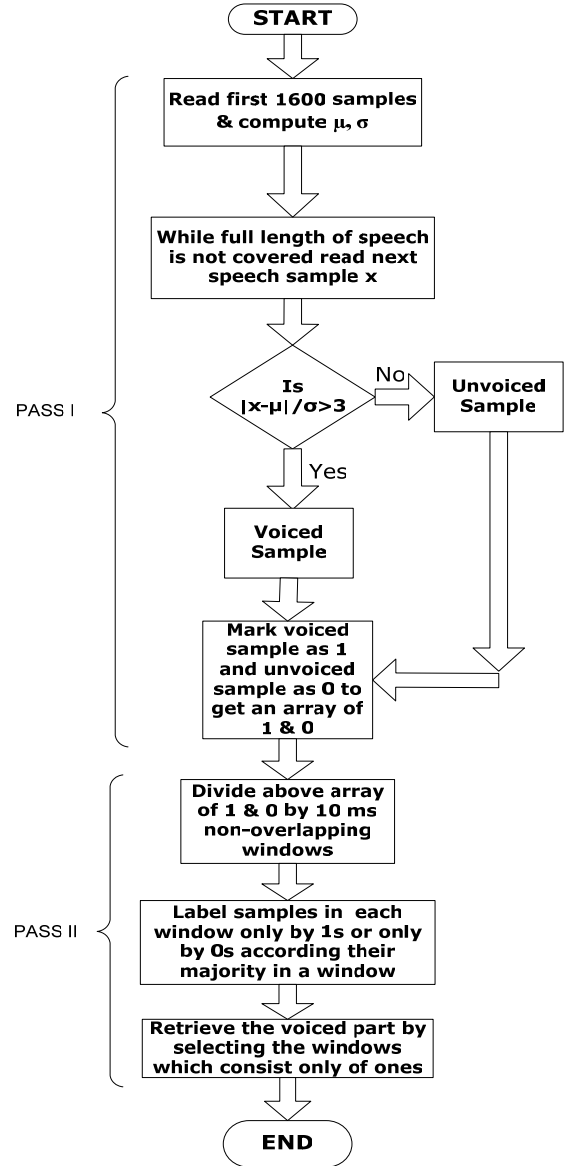


Fig. 3. Flow Chart of the algorithm

conventional STE & ZCR based silence/unvoiced detection method and can be used for real time analysis. However, in the result section we show the proposed method is superior performance wise for silence-voice classification. Note that in STE and ZCR method threshold is calculated after few trials or adhoc basis whereas, proposed method uniquely defines threshold first instance.

3.2 Percentage of Correctness

Percentage of correctness regarding extraction of voiced sample from a speech signal is defined as follows:

$$\% \text{ of correctness} = 100 - \frac{|N_{\text{manual}} - N_{\text{algorithm}}|}{N_{\text{manual}}} \times 100 \quad (11)$$

Where, N_{manual} is the no. of voiced samples from manually labeled speech, $N_{\text{algorithm}}$ is the no. of voiced samples from a specified algorithm.

4. Results

Two experiments are conducted here. In the first experiment, a combination lock number ('26-81-57-29-94-52-35-79-89') from YOHO database is taken while in the second one a running text is read from a paragraph for about 20 sec duration is considered as a speech sample. The second speech is recorded keeping fan, air condition and computers on. For both utterances three algorithms 1) STE 2) ZCR with STE together (because ZCR when used showed poor performance) and 3) Proposed Method are used and output waveforms are presented as follows. Figure 4 and 8 show two original speech samples for two different utterances. Figure 5, 6 and 7 are the results of the combinational lock no. and fig. 9, 10 and 11 are the results of the running text for STE, ZCR-STE & proposed method respectively. Table 1 summarizes results showing percentage of correctness in detection of all three algorithms for both the phrases. Note that the result shows that all the algorithms perform better for YOHO data (combinational lock no.) which is relatively noise free than the speech (running text) collected from noisy environment for the second experiment. Proposed method performs well in both the experiments than conventional ZCR & STE method.

Table 1: Performance Index of the Algorithms using percentage of correctness criteria

Phrases	STE	ZCR-STE	Proposed Method
Combination lock number	77.9531%	70.3720%	83.5565%
Running Text	50.8391%	50.1231%	59.7181%

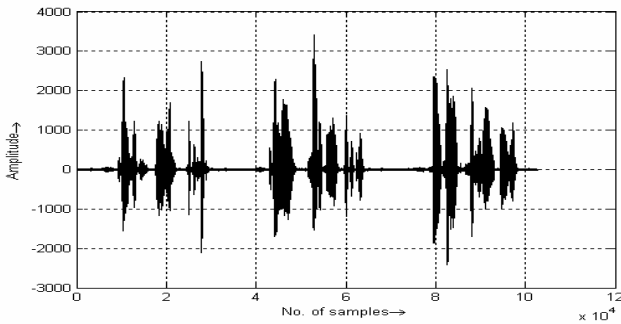


Fig. 4. Original speech signal for combination lock Number

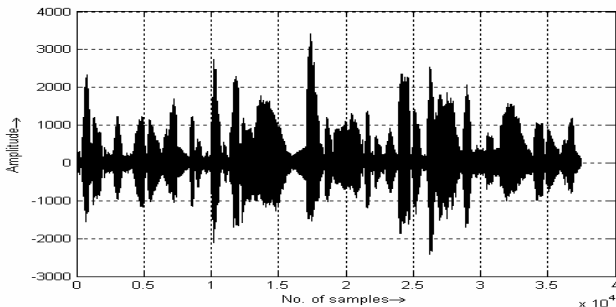


Fig. 5. Output of STE method for combination lock number

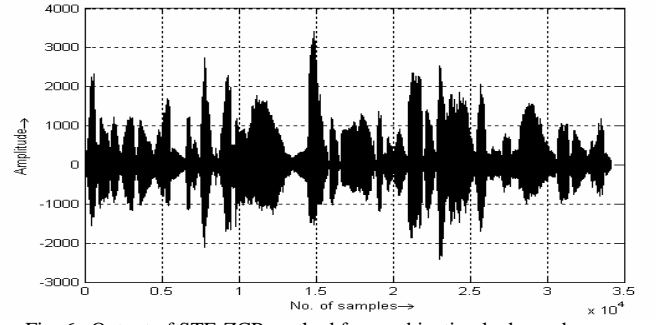


Fig. 6. Output of STE-ZCR method for combination lock number

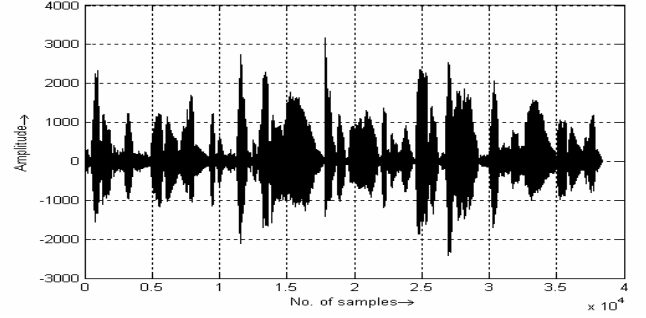


Fig. 7. Output of Proposed Method for combination lock number

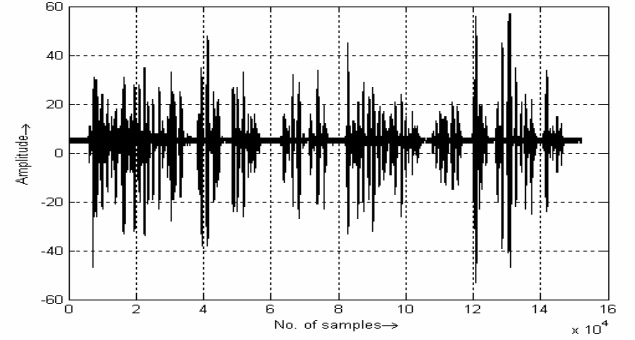


Fig. 8. Original speech signal for running text

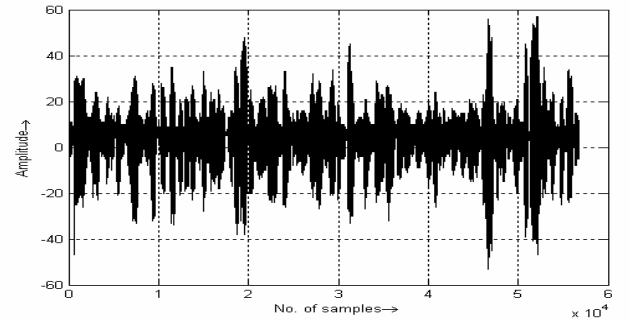


Fig. 9. Output of STE method for running text

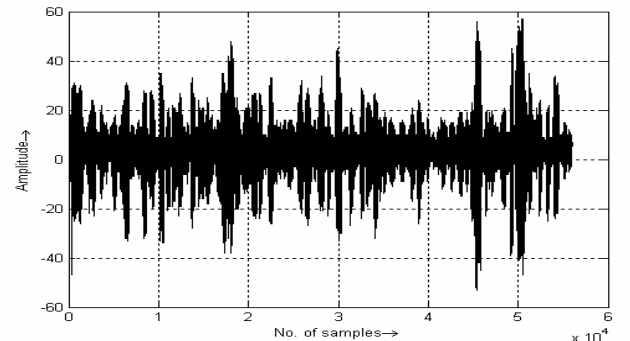


Fig. 10. Output of STE-ZCR method for running text

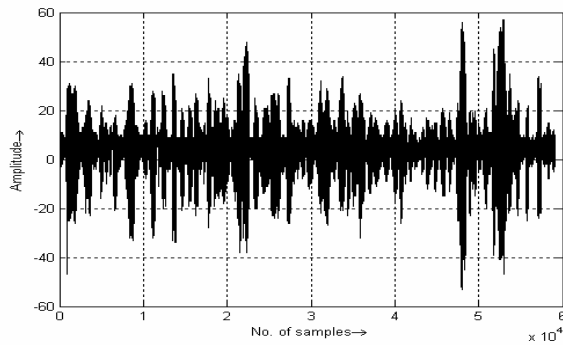


Fig. 11. Output of Proposed Method for running text

5. Conclusion

A new silence end-point detection technique for Speech/Speaker Recognition is presented. The method uses statistical properties of background noise and also the physiological aspects of speech production process. The method assumes the noise to be white Gaussian. However, for other types of noises [12] a similar approach of characterization of noise through probabilistic model can be used. The threshold used in this method is uniquely specified and require no trial and error or adhocism. It is shown to be computationally efficient for real time applications and it performs better than conventional methods for speech samples collected from noisy as well as noise free environment.

6. Acknowledgement

The work is partly supported by Indian Space Research Organization (ISRO), Government of India.

References

- [1] J. P. Campbell, Jr., "Speaker Recognition: A Tutorial", Proceedings of The IEEE, Vol.85, No.9, pp.1437-1462, Sept.1997.
- [2] Koji Kitayama, Masataka Goto, Katunobu Itou and Tetsunori Kobayashi, "Speech Starter: Noise-Robust Endpoint Detection by Using Filled Pauses", Eurospeech 2003, Geneva, pp. 1237-1240.
- [3] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition", in Proc. ICASSP2002, vol. 4, 2002, pp. 3808-3811.
- [4] A. Martin, D. Charlet, and L. Mauuary, "Robust speech / non-speech detection using LDA applied to MFCC", in Proc. ICASSP2001, vol. 1, 2001, pp. 237-240.
- [5] K. Ishizaka and J.L Flanagan, "Synthesis of voiced Sounds from a Two-Mass Model of the Vocal Chords," Bell System Technical J., 50(6): 1233-1268, July-Aug., 1972.
- [6] Atal, B.; Rabiner, L., "A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition" Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on , Volume: 24 , Issue: 3 , Jun 1976, Pages: 201 - 212.
- [7] D. G. Childers, M. Hand, J. M. Larar, " Silent and Voiced/Unvoiced/ Mixed Excitation(Four-Way),

Classification of Speech", IEEE Transaction on ASSP, Vol-37, No-11, pp. 1771-74, Nov 1989.

- [8] Mark Greenwood and Andrew KInghorn, "SUVing: Automatic Silence/Unvoiced/Voiced Classification of Speech", Presented at the university of Sheffield.
- [9] Richard. O. Duda, Peter E. Hart, David G. Strok, "Pattern Classification", A Wiley-interscience publication, John Wiley & Sons, Inc, Second Edition, 2001.
- [10] Sarma, V.; Venugopal, D., "Studies on pattern recognition approach to voiced-unvoiced-silence classification", Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '78. , Volume: 3, Apr 1978, Pages: 1 - 4.
- [11] L. R. Rabiner and R.W. Schafer, "Digital Processing of Speech Signals", First Edition, Chapter-4, Pearson Education, Prentice-Hall.
- [12] http://cslu.ece.ogi.edu/nsl/data/SpEAR_technical.html.
- [13] J. L. Flanagan, Speech Analysis, Synthesis, and Perception, 2nd ed., Springer-Verlag, New York, 1972.
- [14] L. R. Rabiner and B. H. Juang, "Fundamentals of speech recognition," 1st Indian Reprint, Pearson Education.