International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:11, 2008

On Preprocessing of Speech Signals

Ayaz Keerio, Bhargav Kumar Mitra, Philip Birch, Rupert Young, and Chris Chatwin

***Abstract*—**Preprocessing of speech signals is considered a crucial step in the development of a robust and efficient speech or speaker recognition system. In this paper, we present some popular statistical outlier-detection based strategies to segregate the silence/unvoiced part of the speech signal from the voiced portion. The proposed methods are based on the utilization of the 3σ edit rule, and the Hampel Identifier which are compared with the conventional techniques: (i) short-time energy (STE) based methods, and (ii) distribution based methods. The results obtained after applying the proposed strategies on some test voice signals are encouraging.

***Keywords*—**STE based methods, Mahalanobis distance, 3σ edit rule, Hampel Identifier.

I. Introduction

PRE-PROCESSING of speech signals, i.e. segregating the voiced region from the silence/unvoiced portion of the captured signal is usually advocated as a crucial step in the development of a reliable speech or speaker recognition system. This is because most of the speech or speaker specific attributes are present in the voiced part of the speech signals [1]; moreover, extraction of the voiced part of the speech signal by marking and/or removing the silence and unvoiced region leads to substantial reduction in computational complexity at later stages [2], [1]. Other applications of classifying speech signals into silence/unvoiced region and voiced region, as described in [1], are: Fundamental Frequency Estimation, Formant Extraction or Syllable Marking, Stop Consonant Identification, and End Point Detection for isolated speech signals.

One of the accepted ways of labeling a speech signal is the three state representation: (i) Silence region (S) where no speech is produced, (ii) Unvoiced region (U), where the resulting waveform is aperiodic or random in nature as the vocal chords do not vibrate, and (iii) Voiced region (V) where the resulting waveform is quasi-periodic as the vocal chords are tensed and hence vibrate periodically [3], [1]. It should be made clear that the segmentation of the speech signal in the aforementioned regions is not very rigid; however, it has been noted that small errors in the boundary locations seldom have any significant effect in most of the applications [3]. Mention should also be made of the fact that as the energy in the unvoiced portion of the speech signal is usually low, this region is usually clubbed together with the silence region and segregated from the voiced part [1].

The short-time energy based methods to segregate the silence/unvoiced region from the voiced portion are in general fast; however, these methods are stymied by the fact that the thresholds needed to implement them are chosen on an *ad hoc* basis. This means that the recognition system has to be retuned every time there is some change in the ambience [2]. It has been noted that when the method is applied with another popular technique to segregate the different portions of the speech signal, namely the zero cross detection rate (ZCR) method, the success achieved is around 65% [1]. The second category of methods, i.e. distribution based methods, heavily rely on the distribution of the first few thousand samples of the signal which are assumed to be part of the noise region [1]. The methods completely fail if the noise is not well described by the chosen distribution function. The outlier based detection strategies, on the other hand, can be applied on a constrained database where one class of the data is ideally a small fraction of the other class [4]. Here, also, the 3σ edit rule depends heavily on the distribution of the entire data; however, as will be discussed below, the strategy based on Hampel Identifier can be successfully applied to segregate the silence/unvoiced region from the voiced portion. The motivation behind trying the statistical outlier-detection based strategies stems from the fact that if the microphone is kept on for a few seconds before and after the utterance of the word(s), then the samples of the voiced portion can be treated as outliers and the samples of the silence/unvoiced region as inliers.

The paper has been organized as follows: Section II talks about the short time energy (STE) based methods. Section III elaborates one of the distribution based methods, and Section IV broaches the issue of how various popular outlier-detection strategies can be deployed to meet our desired goal. The obtained results are discussed in section V, and finally conclusions are drawn in section VI.

II. Short-Time Energy (STE) based Methods

STE based methods of speech signal segregation utilizes the fact that energy in the voiced region is greater than in the silence/unvoiced region [5], [6], [2], [1]. So, a small window

A. Keerio is with Industrial Informatics and Manufacturing Systems, Department of Engineering and Design, University of Sussex, UK, Falmer, Brighton BN1 9QT (phone: +44-1273-872642; fax: +44-1273-678399; e-mail: a.keerio@sussex.ac.uk, author for correspondence).

B. K. Mitra is with Industrial Informatics and Manufacturing Systems, Department of Engineering and Design, University of Sussex, Falmer, Brighton BN1 9QT (e-mail: b.k.mitra@sussex.ac.uk).

P. M. Birch is with Industrial Informatics and Manufacturing Systems, Department of Engineering and Design, University of Sussex, Falmer, Brighton BN1 9QT (e-mail: p.m.birch@sussex.ac.uk).

R. C. D. Young is with Industrial Informatics and Manufacturing Systems, Department of Engineering and Design, University of Sussex, Falmer, Brighton BN1 9QT (e-mail: r.c.d.young@sussex.ac.uk).

C. R. Chatwin is with Industrial Informatics and Manufacturing Systems, Department of Engineering and Design, University of Sussex, Falmer, Brighton BN1 9QT (e-mail: c.r.chatwin@sussex.ac.uk).

is taken and the energy of the window calculated; if the total energy of the window is more than the chosen threshold, then samples of the window are retrieved, otherwise dropped. We conducted two different experiments; first one with a non-overlapping moving window of fixed size, and another with an overlapping moving window of fixed size. The overlapping moving window technique has been used to prevent valuable information loss [7]. It can be shown that if the all the samples of the captured data are divided into several over-lapping blocks of size T with the distance between the start of two blocks as N, then the total loss $(\mathrm{L_T})$ can be restricted to [7]:

$$0 < \mathrm{L_T} < 2(\mathrm{T} - \mathrm{T_1}) = 2\mathrm{N} \tag{1}$$

where in (1), $\mathrm{T_1}$ is the number of overlapped samples between two successive positions of the fixed sized window.

Note that in the above expression, the values of N and T should be chosen very carefully in order to maximize the desired system accuracy as well as to minimize the operational counts. In our case, the selection of values has been based on the optimization of the following approximate cost function:

$$\mathrm{f(N,T)} = \left(\frac{\lambda - \mathrm{T}}{\mathrm{N}} + 1\right)\mathrm{T}$$

$$\cong \left(\frac{\lambda - \mathrm{T}}{\mathrm{N}}\right)\mathrm{T} \tag{2}$$

where in (2), λ is the total number of samples in the raw data.

It is evident that such a function would always yield the value $\mathrm{f(N,T)} = \lambda$ for points with the condition $\mathrm{N} = \mathrm{T}$; these points are found to be saddle points. We, however, have emphasized having overlapping blocks so that the total loss can be minimized as per the system requirement. From a system accuracy point-of-view, the ideal value of N should be 1 and therefore, $\mathrm{N} = \mathrm{T}$ is not a feasible solution. On the other hand, decreasing the value of N translates into the increasing the number of total unit operations; we have used the following values for our experiments $\mathrm{T} = 4410$ (200msec window, sampling rate 22050 samples/second) and $\mathrm{N} = 2205$.

Taking an average value of λ, say 308265, the total unit operations for one successful execution of the algorithm were 606335 approximately, and the total loss was restricted to $\left(\frac{2\mathrm{N}}{\lambda} \times 100\right) = 1.43\%$ of the entire data.

III. Distribution based Methods

A. Distribution based Methods

Here, we have to assume the nature of the background noise, and then use relevant distribution functions for the segregation purpose [8], [1]. If we assume ambience noise present in the captured signal is Gaussian in nature, then the uni-dimensional Mahalanobis distance function, which itself is a Linear Pattern Classifier [9], [10], [1] can be used to extract the voiced part from the signal.

In this case, at first a 200msec window is chosen to find out the parameters of the Gaussian distribution; the time duration of the window is chosen considering the fact that the speaker will take more than 200msec to initiate speaking after recording starts [1].

Note that the Gaussian distribution in one dimension is defined as:

$$g(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{(x-\mu)^2}{2\sigma^2}} \tag{3}$$

where in (3), μ is the mean and σ is the standard deviation of the distribution.

It can be calculated that the probabilities obey:

$$\Pr[|x - \mu| \le \sigma] \approx 0.68$$
$$\Pr[|x - \mu| \le 2\sigma] \approx 0.95$$
$$\Pr[|x - \mu| \le 3\sigma] \approx 0.997$$

The Mahalanobis distance (r) from x to μ is defined as:

$$r = \frac{|x - \mu|}{\sigma} \tag{4}$$

From the calculations of the probabilities it is evident that there is a probability of 99.7% that the distance, *r*, will be less than 3.

B. Application of the Mahalanobis Distance Method to Extract the Voiced Portion

A window of duration 200msec is chosen (4410 samples, sampling rate being 22,050 samples/sec) to calculate the mean and the standard deviation of the distribution. The duration of the window is chosen based on the assumption that the speaker will take more than 200ms of time to initiate the process of speaking; in other words, this means the chosen window contains ambience noise samples. Now the Mahalanobis distance is calculated for each sample after the window. If the distance calculated is more than 3 then the sample is restored as a voice sample, otherwise dropped.

IV. Outlier-Detection based Strategies

Data points, ξ_k, in a data set, ξ, that do not agree with our expectations based on the bulk of the data are termed as outliers [11], [12]. The popular automatic outlier-detection approaches depend on two estimates: (1) an estimate of a nominal reference value for the data set, and (2) a scatter estimate of the data. Based on these estimators, outliers can be detected based on the following criterion:

$$|\xi_k - \xi_{kr}| > \alpha\gamma \quad \Rightarrow \xi_k = \xi_{ko}, \qquad \forall \xi_k \in \xi \tag{5}$$

where in (5), ξ_{kr} is the nominal reference value of the dataset, α is the threshold parameter, γ the scatter estimate, and ξ_{ko} an outlier.

A. The '3σ edit rule'

The outlier-detection strategy based on the '3σ edit rule' considers the mean of the data values of the data set as the nominal reference value and the corresponding standard deviation as an estimate of the scatter:

$$\xi_{kr} = \xi_{mean} = \frac{1}{N}\sum_{k=1}^{N}\xi_k \tag{6}$$

where in (6), N is the total number of observations in the data set.

$$\gamma = \left[\frac{1}{N-1}\sum_{k=1}^{N}\left(\xi_k - \xi_{mean}\right)^2\right]^{1/2} \tag{7}$$

Note that if the distribution is assumed to be approximately normal, then the probability of getting a data value greater than three times the standard deviation of the data $(\alpha = 3)$, added to the mean, is around 0.3% [14]. The technique suffers from the fact that both the mean and the standard deviation of the data are very much outlier sensitive [12]. Moreover, it should be kept in mind that the strategy is based on the fact that the underlying distribution is approximately Gaussian [4].

B. Strategy Based on Hampel Identifier

In this case, the outlier resistant median (breakpoint value of 50%) and the median absolute deviation from the median (MAD) scale estimates replace the outlier sensitive mean and standard deviation estimates respectively. The median of a data sequence is obtained as follows [13], [4]:

1. The observations are ranked according to their magnitude.

2. If N is odd, the median is taken as the value of the $\left[\frac{(N+1)}{2}\right]^{th}$ ranked observation; otherwise if N is even, the median is taken as the mean of the $\left(\frac{N}{2}\right)^{th}$ and $\left[\left(\frac{N}{2}\right)+1\right]^{th}$ ranked observations.

The MAD scale estimate is defined as:

$$\gamma = MAD_{se} = 1.4826 \times median\left\{\left|\xi_k - \xi_{median}\right|\right\} \tag{8}$$

where in (8), 'the factor 1.4826 was chosen so that the expected value of γ is equal to the standard deviation for normally distributed data' [12].

The strategy, although quite often very effective in practice [12], completely fails if more than 50% of the observations are of the same value, then the scale estimate is equal to 0, i.e. every data value greater than the median would then be considered as an outlier.

Note, for such cases the boxplot outlier-detection strategy can also be used by replacing the mean with the median and the standard deviation with the interquartile deviation [14].

C. Application of Outlier-Detection Strategies

The microphone is kept on for a few seconds before and after the utterance of the speech sample by the speaker. This ensures that the silence/unvoiced region forms the major class, and the voiced samples the minor class. The two aforementioned strategies are then applied to demarcate one class from the other, i.e. to segregate the voiced portion from the noise part of the captured raw data.

V. Results

The different strategies have been tested and compared based on the utterance of (i) a three two-digit combination lock number phrase i.e. 35-72-41 (pronounced as thirty-five/ seventy-two/ forty-one) from the YOHO speech database, maintained by the Linguistic Data Consortium at http://www.ldc.upenn.edu/ and (ii) a running text "*Acoustic signal processing extracts the desired information from speech signals*" read from a paragraph. Fig. 1 and Fig. 2 below show two typical speech waveforms of the two experiments.

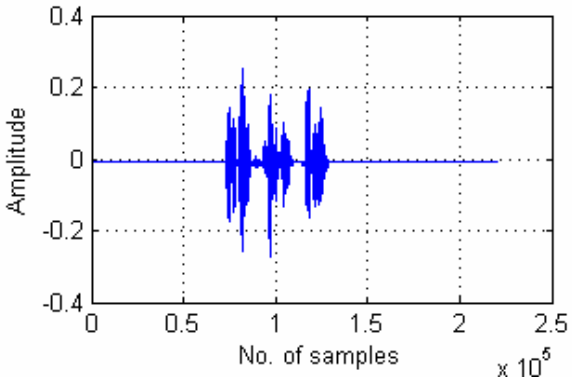

Fig. 1 Typical speech waveform of the three two digit combinational lock number experiment (35-72-41)

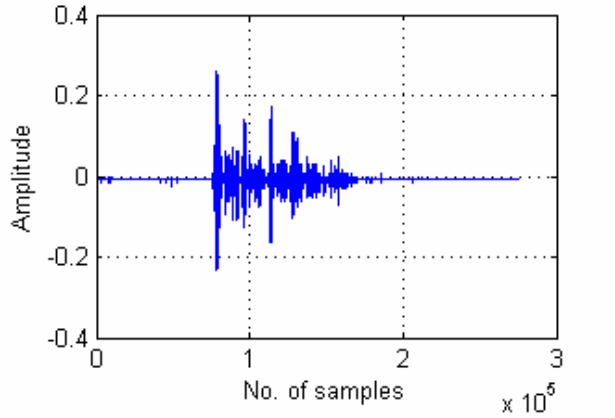

Fig. 2 Typical speech waveform of the running text experiment: "*Acoustic signal processing extracts the desired information from speech signals.*"

A. Speech Data Bank

The voice samples for the first experiment are obtained from the YOHO database. The voice samples for the second experiment were obtained from 20 male and 20 female speakers. Note that the speech signals were captured with an omni-directional microphone, allowing a maximum mouth to microphone distance of 12 inches, while the minimum distance was allowed as 2 inches.

Also note that the speech signals were recorded in the laboratory keeping the laboratory climatizer on but doors and windows closed.

B. Performance Criterion

A single 'figure-of-merit', percentage distortion *d* is defined [1] to analyse the performance of the method:

$$d = \frac{|Vman - Vmeth|}{Vman} \times 100\% \qquad (9)$$

where, in (9) $Vman$ is the count of manually labeled voiced samples, and $Vmeth$ is the count of the voiced samples obtained as the output of an applied method.

Table I and II respectively summarize the results by showing the percentage distortion for male and female voice samples that occurred after applying each strategy for the two experiments using equation (9). The STE based strategies i.e. (i) the non-overlapping moving window (of fixed size) and (ii) the overlapping moving window (of fixed size) show encouraging results for both the experiments, however their results for combination lock number experiment are better. It is obvious from the obtained results that the results of the overlapping moving window (of fixed size) based strategy are better than the non-overlapping moving window (of fixed size) based strategy. Note that the *ad hoc* thresholds required to implement STE based strategies need to be selected carefully. In this work the *ad hoc* thresholds were set manually on a trail and error basis. Fig. 3 and Fig. 4 below, show the typical results of the non-overlapping moving window (of fixed size) based strategy. Fig. 5 and Fig. 6 show the typical results of the overlapping moving window (of fixed size) based strategy.

TABLE I
THE SUMMARY OF AVERAGE PERCENTAGE DISTORTION FOR MALE VOICE SAMPLES

Strategy	Combination lock number	Running text
Non-overlapping moving window	4.92%	11.91%
Overlapping moving window	3.66%	8.40%
Mahalanobis Distance	20.91%	9.71%
3σ edit rule	28.17%	30.43%
Hampel Identifier	12.52%	9.68%

TABLE II
THE SUMMARY OF AVERAGE PERCENTAGE DISTORTION FOR FEMALE VOICE SAMPLES

Strategy	Combination lock number	Running text
Non-overlapping moving window	7.80%	10.69%
Overlapping moving window	5.06%	7.03%
Mahalanobis Distance	17.03%	18.90%
3σ edit rule	41.53%	45.20%
Hampel Identifier	16.89%	17.91%

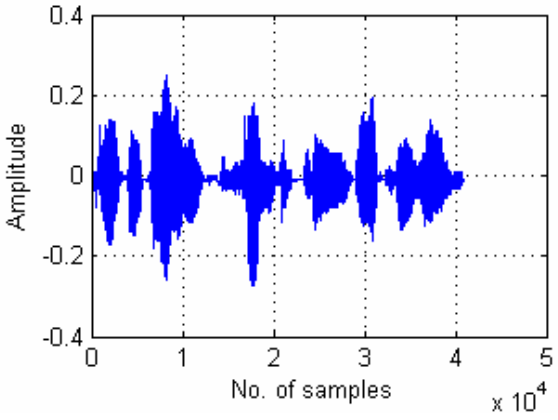

Fig. 3 Typical output of the non-overlapping moving window (of fixed size) based strategy for the combination lock number experiment

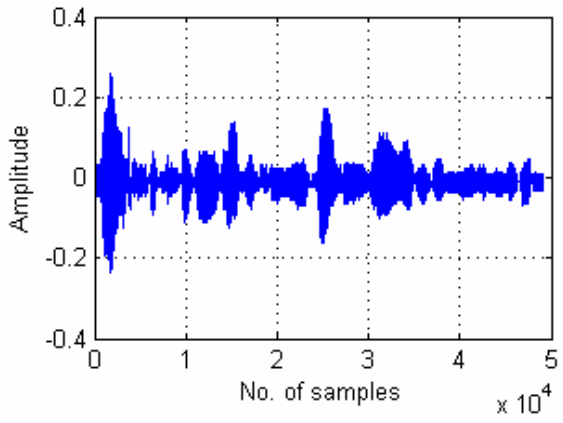

Fig. 4 Typical output of the non-overlapping moving window (of fixed size) based strategy for the running text experiment

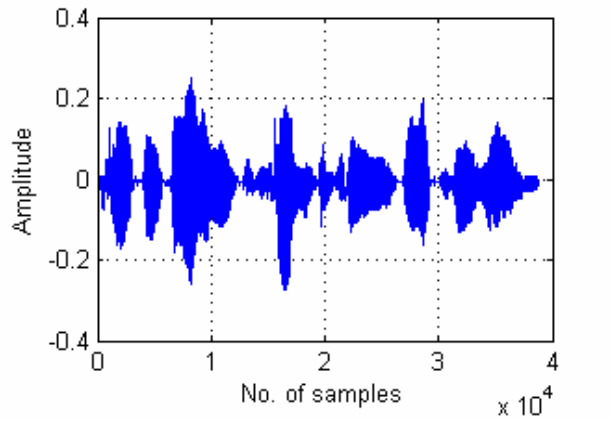

Fig. 5 Typical output of the overlapping moving window (of fixed size) for the combination lock number experiment

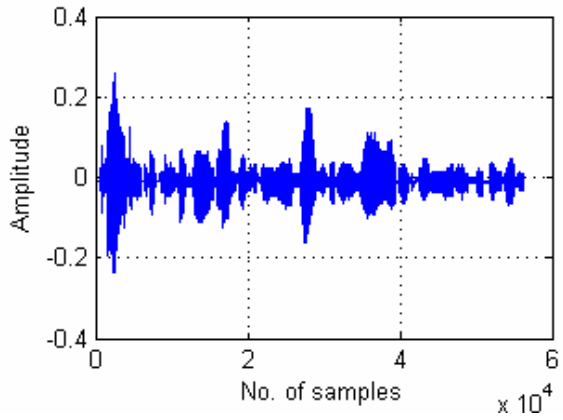

Fig. 6 Typical output of the overlapping moving window (of fixed size) for the running text experiment

The distribution based 'Mahalanobis distance' method depends on the nature of the noise distribution. If the distribution is not Gaussian in nature it fails completely. For the running text experiment the Mahalanobis distance method showed moderately good results. However it failed completely for some of the samples tested for the combination lock number experiment as shown in Fig. 7, and Table I (the high percentage distortion 20.91%). Fig. 8 below show the typical result of the Mahalanobis Distance based method for running text experiment.

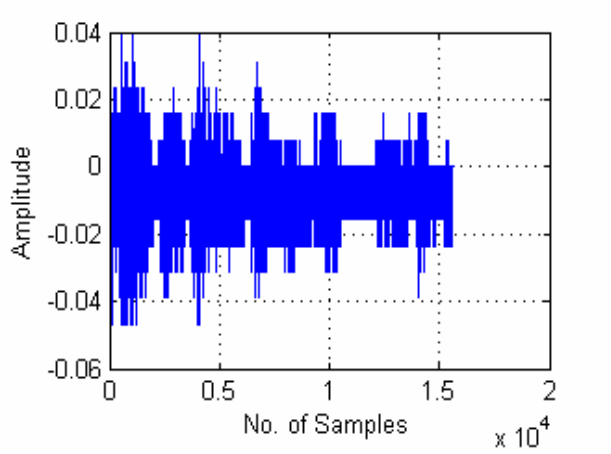

Fig. 7 Output of the Mahalanobis Distance based method for one of the Combination lock number experiments where it failed completely

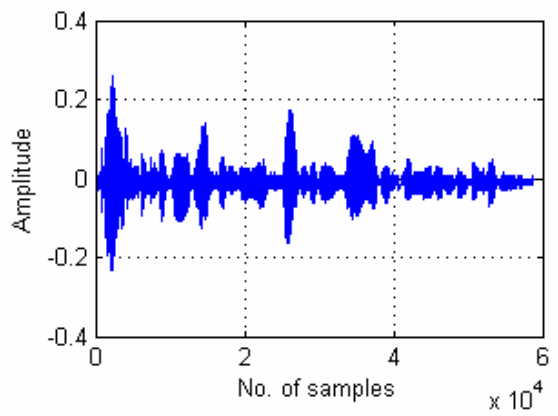

Fig. 8 Typical output of Mahalanobis Distance based method for the running text experiment

One of the outlier-detection based strategies the '3 σ edit rule' fails completely for both the experiments showing a very high percentage distortion (as shown in Table I and Table II). Fig. 9 and Fig. 10 below show the typical results of the '3 σ edit rule' based method.

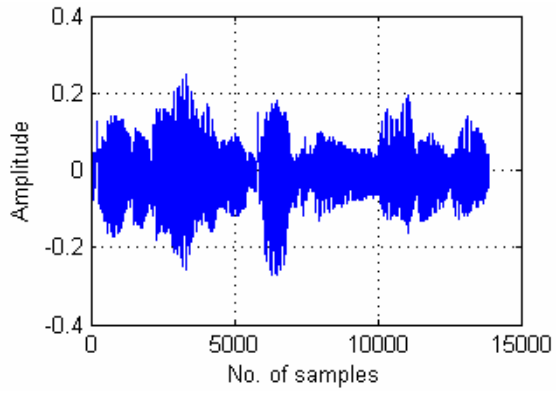

Fig. 9 Typical output of the '3 σ edit rule' based method for the combination lock number experiment

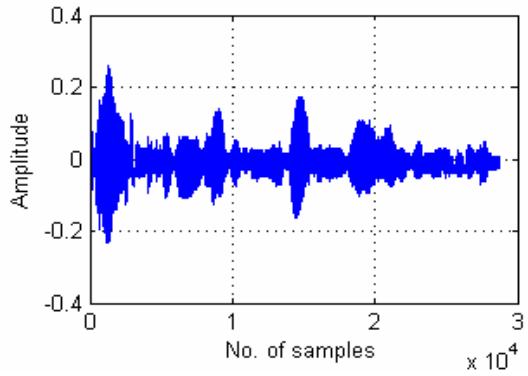

Fig. 10 Typical output of the '3 σ edit rule' based method for the running text experiment

However, the outlier-detection based strategy the Hampel Identifier show good results for both the experiments. Thus it is observed from the obtained results that the Hampel Identifier based method can be used to segregate the voiced and unvoiced/silence portions of the speech signals. Note that the success rate of this strategy depends on the fact that one class of data should ideally be a small fraction of the other class. Fig. 11 and Fig. 12 below show the typical results of the Hampel Identifier based method.

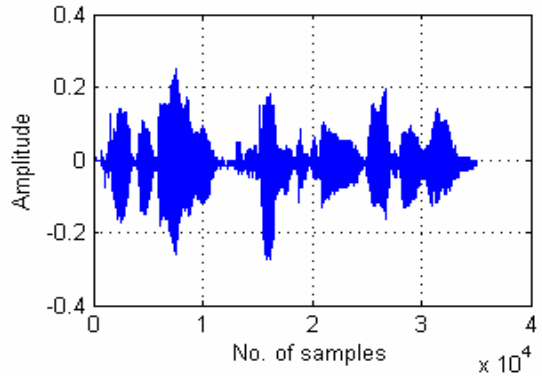

Fig. 11 Typical output of the Hampel Identifier based method for the combination lock number experiment

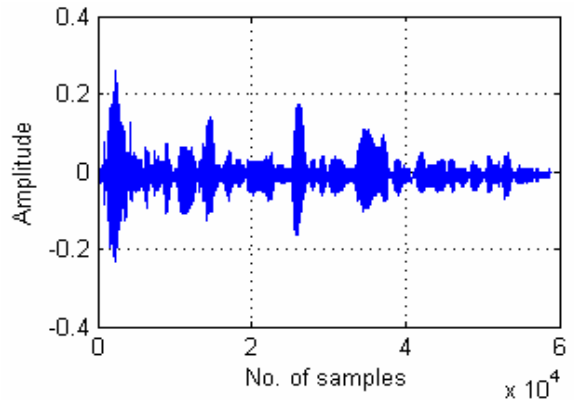

Fig. 12 Typical output of the Hampel Identifier based method for the running text experiment

VI. Conclusion

Three different approaches are presented for the segregation of voiced and unvoiced/silence portions of the speech signals using the statistical properties of the ambient noise. It is observed from the results that the STE based strategies have shown moderately good results and can be used for the purpose of segregation of voiced and unvoiced/silence portions of the speech signals. These strategies are also generally efficient computationally; however these strategies depend heavily on the selection of a suitable *ad hoc* threshold, which is manually set on a trial and error basis. The distribution based strategies depend on prior knowledge about the noise statistics. The Mahalanobis distance method fails for one of the experiment because the noise distribution was not Gaussian. The outlier-detection based strategies, the '3 σ edit rule', fails completely for both experiments, while the Hampel Identifier based strategy gives better results and can be used for the segregation of the voiced and unvoiced/silence portions of the speech signals; moreover it makes the overall preprocessing method inherently automatic. Thus it can be concluded that the STE based strategies and the Hampel Identifier based strategy give better results for the segregation of voiced and unvoiced/silence portions of the speech signals.

References

[1] Saha. G., Chakroborty. S., and Senapati. S, "A new Silence Removal and End Point Detection Algorithm for Speech and Speaker Recognition Applications", in Proc. of Eleventh National Conference on Communications (NCC), IIT-Kharagpur, India, January 28-30, 2005, pp. 291-295.

[2] Mitra. A., Chatterjee. B., Mitra. B. K., "Identification of Primitive Speech Signals using TMS320C54X DSP Processor", in Proc. of Eleventh National Conference on Communications (NCC), IIT- Kharagpur, India, January 28-30, 2005, pp. 286-290.

[3] Rabiner. L. R., and Juang. B. H., "Fundamentals of Speech Recognition", AT&T, 1993, Prentice-Hall, Inc.

[4] Mitra. B. K., Young. R., Chatwin. C., "On shadow elimination after moving region segmentation based on different threshold selection strategies", Optics and Lasers in Engineering, vol. 45, no. 11, pp. 1088-1093, July 2007.

[5] 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], vol 24, 3, June 1976, pp. 201-212.
[6] Childers. D. G., Hand. M., Larar. M. J., "Silent and Voiced/Unvoiced/Mixed Excitation (Four Way), Classification of Speech", IEEE Trans. On ASSP, vol 37, 11, Nov 1989, pp1771-74.
[7] Mitra. A, Mitra. B. K., Chatterjee. B., "Recognition of Isolated Speech Signals using Simplified Statistical Parameters", Proceedings of World Academy of Science, Engineering and Technology, vol.8, pp. 151-154, October 2005.
[8] Rabiner. L. R., Schafer. R. W., "Digital Processing of Speech Signals", First Edition, Prentice-Hall.
[9] Duda R. O., Hart. P. E, Strok. D. G., "Pattern Classification", Second Edition, John Wiley and Sons Inc., 2001.
[10] Sarma. V., Venugopal. D., "Studies on pattern recognition approach to voiced-unvoiced-silence classification", IEEE international conference on ICASSP, 78, 3, April 1978, pp. 1-4.
[11] Hawkins D.M, "Identification of Outliers", Great Britain, Chapman and Hall, 1980.
[12] Pearson R.K, "Outliers in process modeling and Identification", IEEE transactions. Consol Syst. Technologies 2001. lo (1) pp55-63.
[13] Halin G.Z, Shafiro S.S, "Statistical Models in Engineering", USA: Wiley, 1967.
[14] Pearson R.K. "Mining imperfect data dealing with contamination and incomplete records", Philadelphia: SIAM; 2005.