Asian Journal of Information Technology 15 (19): 3770-3779, 2016
ISSN: 1682-3915

A Novel Entropy Based Algorithm to Remove Silence from Speech and Classifying the Residue as Voiced/unvoiced Regions

[1]R. Johny Elton, [1]P. Vasuki, [2]J. Mohanalin
[1]Department of Electronics and Communication Engineering, K.L.N College of Information Technology, Tamil Nadu, India
[2]Department of Electrical and Electronics Engineering, Lourdes Matha College of Science and Technology, Kerala, India

Abstract: For any speech synthesis, voiced portion of speech plays a crucial role. Major researchers have focussed on the most sophisticated statistical approaches whereas least importance was given to the time-domain or frequency domain approaches citing their limitations. The issues of statistical approaches are dealt with by adding new features making them more complex rather than resolving complexity. So we propose an algorithm which uses one feature namely sample entropy to classify speech signal. In our proposed algorithm, silence removal is achieved by fuzzy entropy and sample entropy is used to classify the residual speech signal as voiced or unvoiced regions. The performance of the proposed algorithm is analysed using TIMIT database. The proposal outperforms the existing approaches with a 94.98 % accuracy of information during silence removal from speech signals and the classification rate is analysed using Receiver Operating Characteristics (ROC) which yields an accuracy of 92.78 %.

Key words: Voiced, unvoiced, fuzzy entropy, sample entropy, India

INTRODUCTION

In speech analysis, classification of speech signal into VOICED/UNVOICED (V/UV) provides the groundwork for acoustic segmentation. This classification demands, classifying the regions of speech based on the vibration of the vocal cords (glottal activity). Speech, in general, can be classified as Silence (S), Voiced (V) and Unvoiced (UV) as in Fig. 1. Regions of speech where vocal cords tend to vibrate due to the excitation of air from the lungs is referred to as voiced speech and regions where there are no such vibrations but presence of noise like turbulence are referred to as unvoiced speech. Silence region is identified as absence of speech (no excitation) which has only background noises. The classification of speech signal into Voiced and Unvoiced (V/UV) regions plays a crucial role in major speech processing applications like speech signal modelling (Yin *et al*., 2009), language identification, pitch detection (Rouat *et al*., 1997), etc. Also other applications like automatic speech recognition.

(ASR) (Strik and Cucchiarini, 1999; Bosch, 2003), speech enhancement (Krishnamoorthy and Prasanna, 2011; Paulikas and Navakauskas, 2005; Karthikeyen *et al*., 2013), diagnosing voice disorders (Hariharan *et al*., 2013; Maier *et al*., 2009), emotion recognition (Bosch, 2003), etc. heavily rely on estimating pitch frequency (Yegnanarayana and Murty, 2009) which deserves locating voiced regions of speech (Paulikas and Navakauskas, 2005; Ercelebi, 2003).

Any V/UV/S classification falls into three categories such as time-domain, frequency-domain and statistical approaches. In time-domain approach, V/UV/S classification could be made using a single parameter derived from the speech signal such as rms energy or Zero-Crossing Rate (ZCR). But if the nature of the recorded speech lacks high environment fidelity, the accuracy of the method will be limited because relying on a single parameter usually overlaps between categories (Qi and Hunt,, 1993; Atal and Rabiner, 1976). Moreover, large analysis frames are to be considered for V/UV classification which is connected to the determination of pitch (Childers *et al*., 1989; Childers and Lee, 1991) and for UV/S, it becomes even more complex because often unvoiced regions are treated as silence which affects the efficiency of recognition of speech. Similarly for frequency-domain approaches (Arifianto, 2007), the significance of voiced sounds is measured by one or more acoustic features such as energy, periodicity and short-term correlation. Some more parameters that

Corresponding Author: R. Johny Elton, Department of Electronics and Communication Engineering, K.L.N College of Information Technology, Tamil Nadu, India

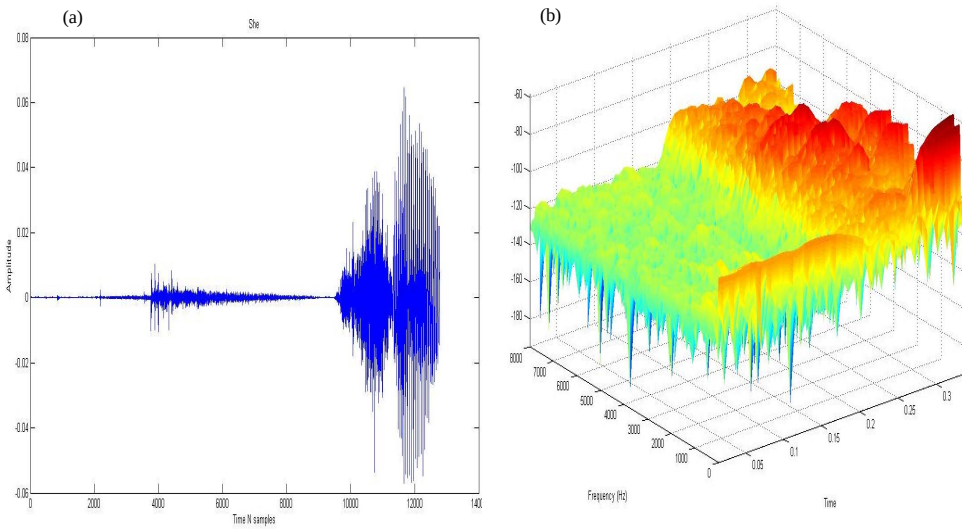

Fig. 1: Speech signal for the utterance 'She' with silence(s), unvoiced (uv) and voiced (v) regions (Left) and Spectrogram of the speech utterance 'She' showing three sound regions (Right)

parameters that can be considered are autocorrelation coefficient at the first lag, the first coefficient of a p^{th}-order Linear Prediction (LP) analysis, long-term normalized autocorrelation peak strength (in the range 2-15 ms), normalized LP error, normalized low-frequency energy, cepstral peak strength, harmonic measure from the instantaneous frequency amplitude spectrum (Atal and Rabiner, 1976; Erkelens and Broersen, 1998).

Traditionally, V/UV/S classification is mainly used for the determination of periodicity of speech signal. But, it is not always that the vocal fold vibrations yield periodic signal, because in real speech even the voiced region contains some random like aperiodic components (Klatt and Klatt, 1990; Holmberg *et al*., 1994). This is noticed obviously in case of voiced fricative, (e.g., /v/, /z/), in breathy vowels (Hermes, 1991) or in cases where speech is produced with weak phonemes. This aperiodic component can also be identified in normal vowels due to turbulence of air around the instant of glottal closure (Naylor *et al*., 2007) which gives rise to aspiration noise (Pinto *et al*., 1989; Qi and Hunt, 1991). This aperiodic component helps in characterizing voice quality attributes such as breathiness or roughness. Breathiness can be noted due to glottal air leakage and due to turbulence noise during phonation whereas; roughness can be noted because of the low-frequency noise component (D'Alessandro *et al*., 1993). Moreover, inclusion of aperiodic component in voiced excitation helps producing natural sounding synthetic speech (Yegnanarayana *et al*., 1998; Christophe, 1998). Also synthetic speech with desired voices is possible if detailed characterization of the source is available. So whenever there is a failure in detection of periodicity in the voiced speech, V/UV/S ends up in error classification. V/UV/S classification relies predominantly on one more factor, namely threshold. Setting up threshold differs for individual parameters and its decision can be combined in a hierarchical fashion. Usually, manual threshold is applied based on the experimental study but it often limits the performance of the classification.

So researchers focussed on the other sophisticated approaches, namely the statistical approach to overcome these problems, where the role of threshold is completely neglected because the classification rate relies mainly on the input parameters chosen and on the efficiency of the classifier. So pattern recognition approach (Atal and Rabiner, 1976; Siegel Bessey, 1982) was followed by combining one or more features instead of relying on single-feature where it requires more training on the set of features chosen. Qi and Hunt 1993) proposed hybrid features based V/UV/S classification by including new features and selecting optimal set of features, but the network training demands much longer time. Mostly these statistical approaches rely on Artificial Neural Networks (ANN), Gaussian Mixture Model (GMM) or Hidden Markov Models (HMM). These approaches do not rely on the knowledge of speech production mechanism and invariably the performance evaluation does not rely on the classification of V/UV/S regions of speech.

In this study, we propose entropy based approaches that would overcome the limitations identified with the existing approaches. As per information theory, entropy (Shannon, 1948) is used to retrieve the information in a

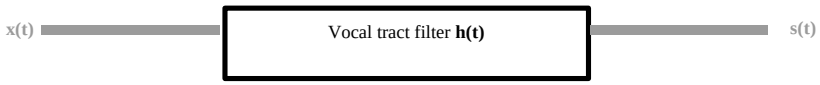

Fig 2: Simple model for speech production

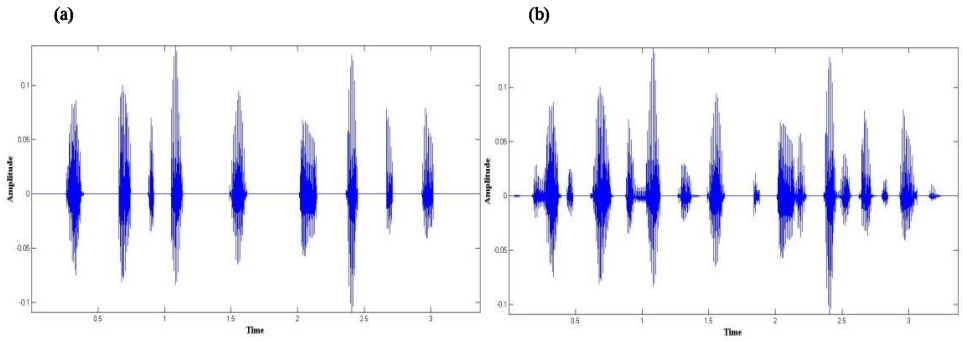

Fig. 3: Impact of threshold for the speech utterance by male speaker when using feature a. Energy and b. ZCR

signal. When this is applied over (Kosko, 1986) fuzzy domain (fuzzy entropy), it helps in suppressing background noises (silence) (Qaimkhani and Hossain, 2008), thereby retrieving a residual speech inclusion of both V and UV. This idea is more suited when cases of compression of speech signals are required during transmissions. For V/UV classification, we use sample entropy (Richman and Moorman, 2000) as feature because of its usefulness in measuring non-linear time series data and its efficiency is evaluated by the selection of optimal threshold. We use sample entropy to identify the regions where vocal cords that vibrate and those that do not, which are labelled as V and UV, respectively.

MATERIALS AND METHODS

The speech samples required for the validation process are collected from the TIMIT database consisting of both short and long utterances. The size of speech utterances ranges from 3-5 sec. Speech samples are collected from TIMIT database (Dekrom *et al*., 1993). Speech utterances are selected randomly from the database, each uttering 2 sentences each from 8 different dialect regions. The speakers include both male and female each contributing around 300 and 180 respectively in count. The selection of the samples are chosen in such a way that a person with a low pitched voice, persons changing their locality more often, persons with good accenture to speak, persons adapting new accents, persons with a cold and on the road to recovery, voice disorders, etc.

Problem statement and objectives: In general, speech s(t) is formed as a sequence of sounds which has low frequency components (<2 kHz) and high frequency (between 2 and 8 kHz) components. This can be mathematically represented as excitation from the source x(t), which is the input to the vocal tract acting like a filter h(t). The convolution of the source x(t) and filter h(t) is given by Eq. 1:

$$s(t) = x(t) * h(t) \quad (1)$$

Where x(t) can be excitation signal from the glottal pulse generator (periodic) or random noise like pulse sequences. A simple speech production model is shown in Fig 2. The objectives are set as follows:

- To overcome the weakness encountered in existing single-feature approach used for V-UV classification
- To propose a new algorithm using Fuzzy entropy as a feature to remove silence from the speech signal
- To propose a new algorithm using Sample entropy to classify voiced and unvoiced regions of the speech signal

Silence removal and Voiced/Unvoiced features

Energy: For any given speech signal, the energy of the signal is reflected on the amplitude variations of the speech signal. Usually, the energy tends to be higher in voiced regions than that of the unvoiced regions shown in Fig 3a. Energy of the signal is computed by Eq. 2:

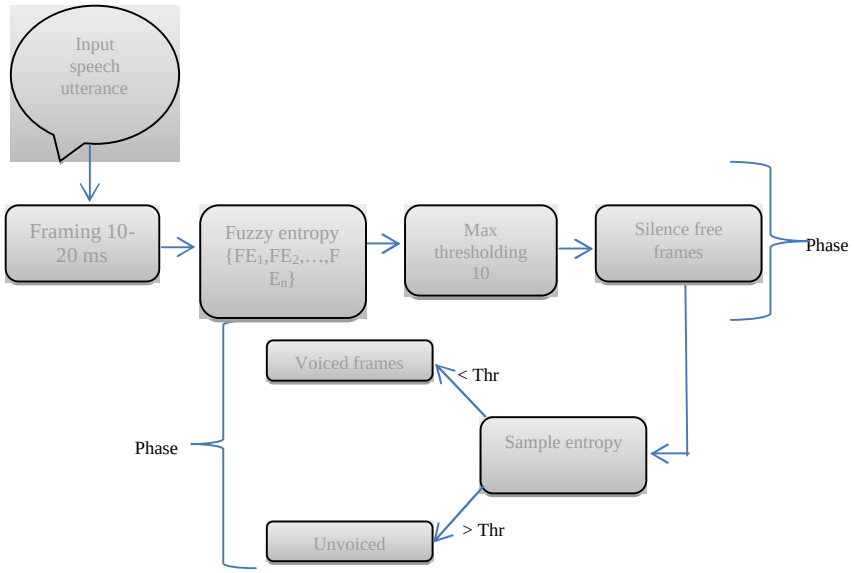

Fig. 4: Block diagram of the proposed algorithm

$$E_i = \sum_{i=1}^{n} s_i^2 \quad (2)$$

Zero Crossing Rate (ZCR): ZCR is defined as the rate at which the speech signal crosses zero and the unvoiced speech has higher ZCR than voiced regions because most of the energy of the unvoiced speech is located around high frequencies as shown in Fig 3b. This is given by Eq. 3:

$$ZCR_i = \frac{1}{n}\sum_{i=1}^{n} |\text{sign}[(s_i)] - \text{sign}[(s_{i-1})]| \quad (3)$$

Fuzzy Entropy (FE)-Silence removal (Phase I): Let S^n_1 through S^n_{N-n+1} be the vector sequences of the speech, for a given frame size n, for an N speech sample time series {s(i): 1 = i = N} which is given by Eq. 4:

$$S_i^n = \{s_i, s_{i+1}, \ldots, s_{i+N-n}\}, \quad i = 1, 2, \ldots, N - n + 1 \quad (4)$$

then, by Shannon entropy (Shannon, 1948), $p(s_i)$ is the probability of the sample s_i^n given in Eq. 5:

$$H(S_i^n) = -\sum_{i=1}^{n} p(s_i)\log_2 p(s_i) \quad (5)$$

This Shannon entropy applied over fuzzy sets is called Fuzzy entropy (Kosko, 1986). Let μs_i^n be the membership function which is given by Eq. 6:

$$\mu_{S_i^n} = e^{-\frac{(s_i - c)^2}{2\sigma^2}} \quad (6)$$

Where:

c = Is the center

σ = Is the standard deviation computed over m

The entropy over this fuzzy set is given by Eq. 7:

$$H(\mu_{S_i^n}) = -p(\mu_{s_i})\log_2 p(\mu_{s_i}) \quad (7)$$

Sample Entropy(SampEn)-Voiced/Unvoiced (Phase II): Sample entropy (Richman and Moorman, 2000) is an alteration of approximate entropy (ApEn) which is used in the measure of complexity of the time series data, but its properties differ in two ways: it excludes self-matching and it avoids template-wise approach. For the computation of SampEn, the given time series of the speech signal s_i: $1 \le i \le N$ is to be embedded m-dimensional space with delay. The vector construction is given by $x_m^j = \left[s^{j+k}\right]_{k=0}^{m-1}$ where j = 1,..., N -m+1. The probability $B_m(r)$ that two sequences match for m points is computed by counting the average number of vector pairs (in the embedded space), for which the absolute distance between their corresponding scalar elements $|s_i, s_k+k| < r$) is lower than the tolerance r. Similarly, $A_m(r)$ is defined for an embedding dimension of m+1. Sample entropy of each frame can be computed using Eq. 8:

$$\text{Samp En}(m,r) = -\ln\frac{A_m(r)}{B_m(r)} \quad (8)$$

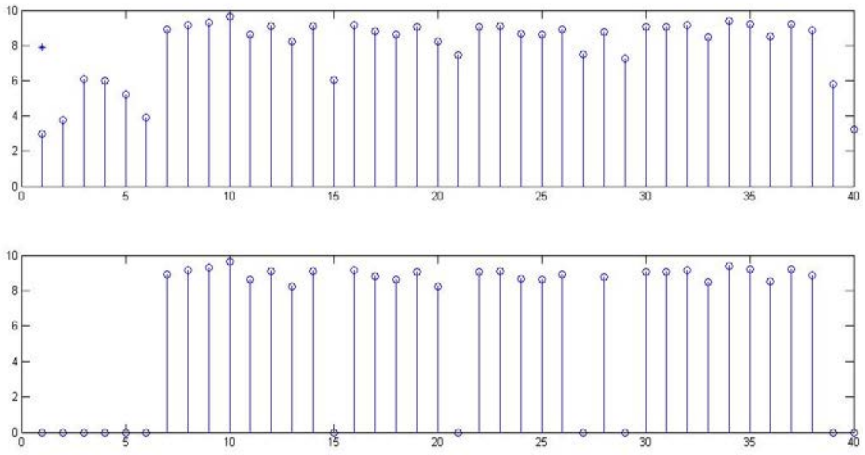

Fig. 5: a) Fuzzy entropy features for frames of size 10 ms; b) Impact of threshold ({FEn >th}) on feature set using Fuzzy entropy to remove silence region frames from the given input speech signal

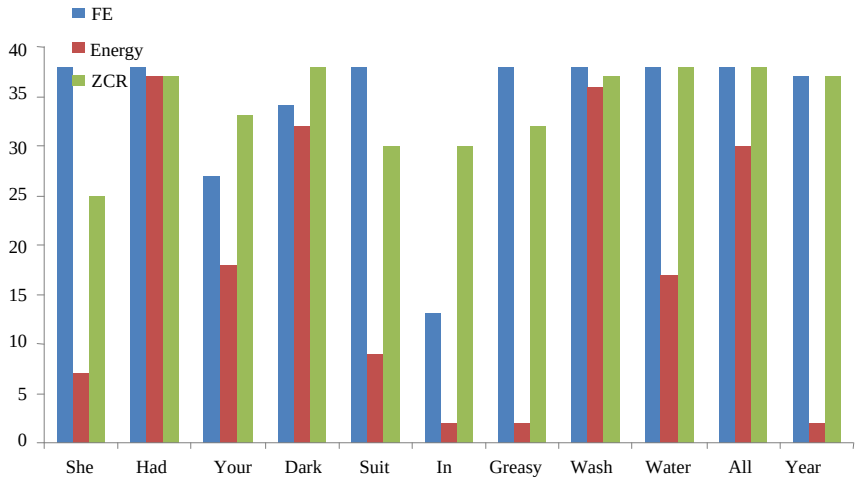

Fig 6: Speech utterances of both male and female speakers uttering the sentence "She had your dark suit in greasy wash water all year" after removing silence using FE, Energy and ZCR

Sample entropy features are obtained using SampEn by specifying a length of m points and a tolerance window r, which is set to be 2 and 0.2 σ, respectively.

Proposed algorithm: V/UV decision region is a very important and challenging part in our proposed algorithm. V regions have higher energy and occupy low frequency range (LFR) and the UV regions fill the high frequency range (HFR). The principle behind the classification problem depends on setting up of proper thresholds to the features selected that would distinguish the characteristics of V and UV regions. Therefore, selection of optimal thresholds to distinguish the regions of interest becomes crucial. The architecture of the proposed algorithm is shown in Fig. 5. The steps involved in the algorithm are as follows:

- Step 1; The speech signal is divided into frames of size 10-20 msec
- Step 2; Feature vectors using FE for phase I is calculated for each frame which is given by FE_1, FE_2, ..., FE_n calculated using the formula in Eq. 7
- Step 3: Threshold using averaging is computed over the feature vector extracted using FE and its impact on the FE feature vectors is shown in Fig 5. Maximization of FE, truncates silence region of the speech signal preserving both V and UV regions shown in Fig. 6
- Step 4; In phase II to the selected frames, sample

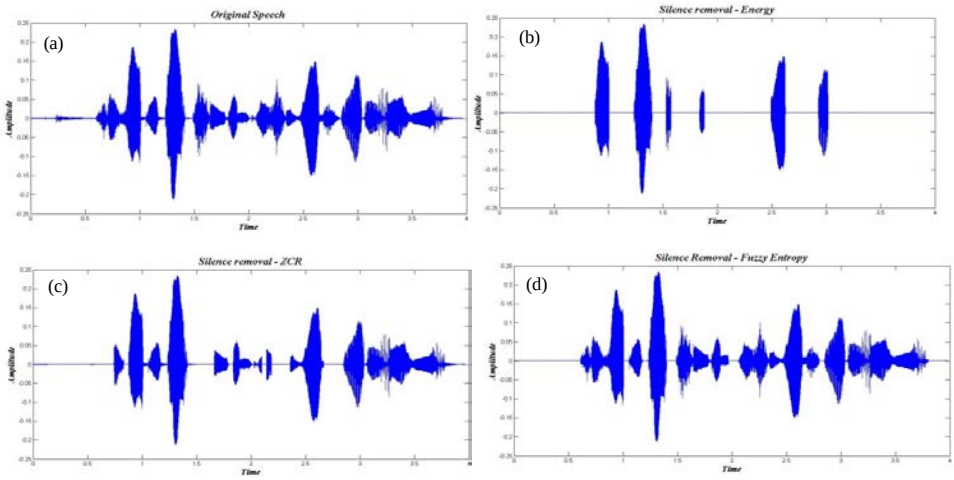

Fig 7: Silence removal from the speech signal with a frame size of 10 msecs and thresholding by averaging. a. Input speech signal b. Output when using Energy as a single feature c. Output when using ZCR as a single feature d. Output when using Fuzzy entropy (Proposed) as a single feature

entropy is calculated using Eq. 8. Threshold is calculated for the SE vectors $\{SE_1, SE_2, \ldots, SE_n\}$ and the classification is made. Classes C_1 and C_2 contain voiced and unvoiced regions, respectively

RESULTS AND DISCUSSION

Phase I: In this phase, silence removal is to be deployed and a sample output of the same is shown in Fig. 7a to d. Differentiating silence and unvoiced region is difficult when considering a single feature based analysis. Entropy based classifications are solid for two class problems, here; speech (V and UV) and non-speech (silence) are the two classes. V regions occupy LFR which has high energy and UV around HFR with relatively low energy as compared to V regions. However, UV from silence discrimination is more significant, because existing energy based silence removal measures energy level based on V region i.e., energy is high in V regions and low in UV regions but it's hard to differentiate with one feature and moreover large frames of samples are required for the analysis. So applying threshold over energy frames, only portion of speech with high energy content is preserved and major HFR components are removed along with the silence observed in Fig 7b. On the other hand, ZCR for UV is high and V is low, but when removing silence (background noise) some UV regions are also treated as background noises and are removed which affects some edges contained in the speech signal observed in Fig 7 c. But the proposed FE, simplifies this by the optimal threshold selected by segregating the LFR and HFR (V and UV regions) from the background noises (silence) which is noticed from Fig 7d. This helps in preserving HF edge information available in the speech signal, because formant frequencies are available in HFR. A special case is analysed by adding white Gaussian noise to TIMIT database speech utterances of Signal-Noise Ratio (SNR) ranging from 0 to 30 dB. The efficiency of the phonemes extracted during silence removal is given by Eq. (9),

$$\%\text{of Accuracy} = 100 - \frac{|N_{original} - N_{processed}|}{N_{original}} \quad (9)$$

Where:

$N_{original}$ = Refers to the number of phoneme occurrences before silence removal and

$N_{processed}$ = To the phoneme occurrences after removing silence by the algorithm

The efficiency of the feature is calculated by adding white Gaussian noise to the input speech signal and its SNR and accuracy of phonemes after silence removal by FE, ZCR and Energy are compared and is shown in Fig. 8. Silence removal in speech signal for male speaker uttering the sentence, "Don't ask me to carry an oily rag like that" by energy, ZCR and fuzzy entropy is shown in Fig. 9 and silence removal of the same at SNR = 20 dB is shown in Fig. 10.

Phase II: In this phase, V/UV classification is experimented from the output obtained from silence removal phase. Sample entropy is computed over the

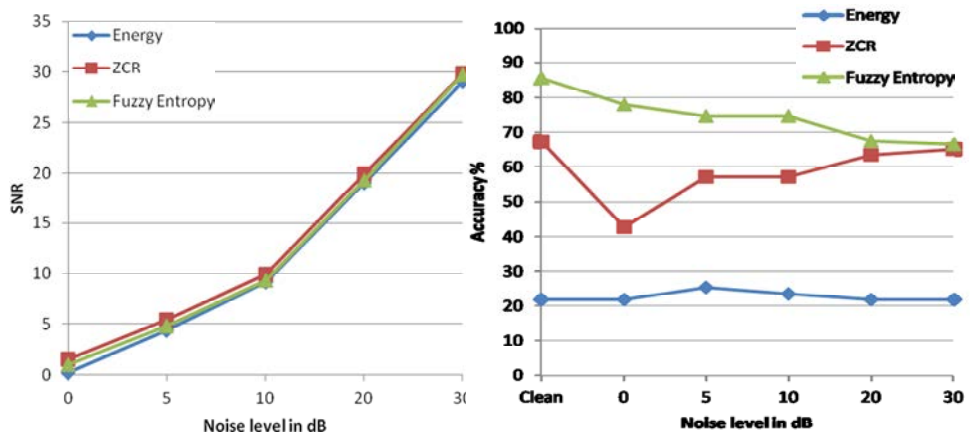

Fig 8: Silence removal under noisy conditions with noise level from 0 to 30 dB and a. its gain for Energy, ZCR and Fuzzy Entropy b. Accuracy in % for Energy, ZCR and Fuzzy Entropy

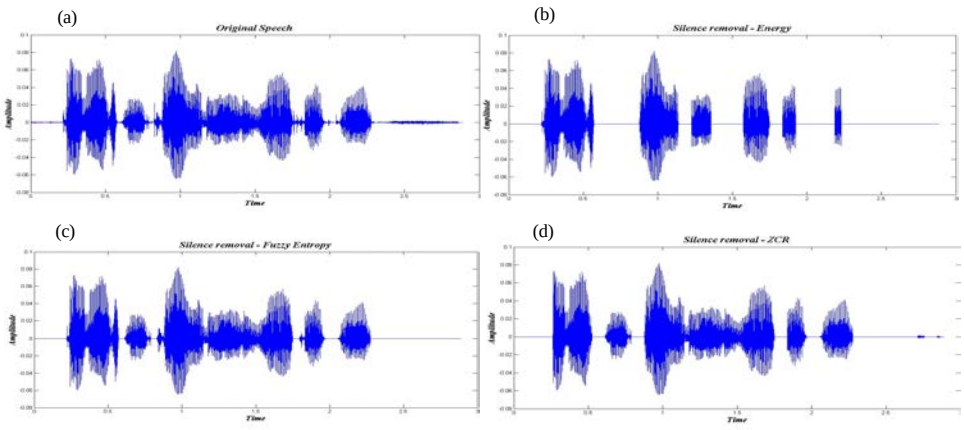

Fig 9: Silence removal from the speech signal uttered by a male speaker thresholded by averaging: a) Input speech signal; b) Output when using energy; c) Output when using ZCR d) Output when using fuzzy entropy (proposed)

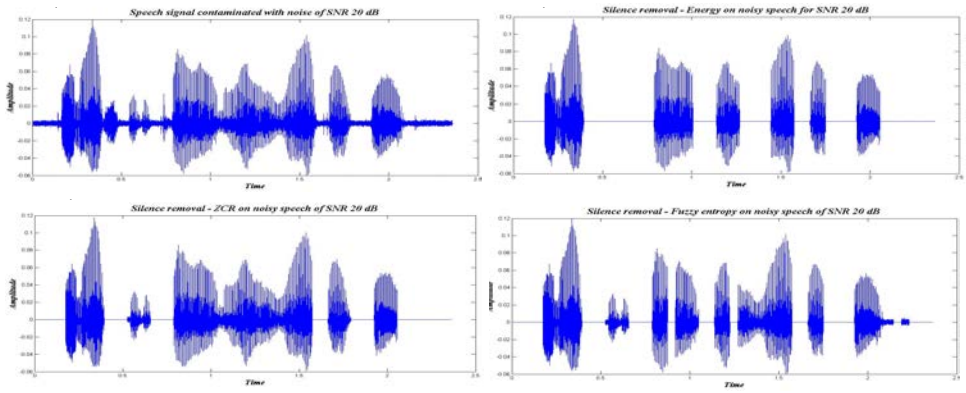

Fig 10: Silence removal from the speech signal added with white Gaussian noise of SNR 20 dB: a) Input speech signal; b) Output when using energy; c) Output when using ZCR; d) Output when using Fuzzy entropy (Proposed)

frames selected from the previous phase. Here, the classification mainly depends on the optimal threshold value being selected. Threshold is computed using entropy (Shannon, 1948) given by the formula in Eq. 10:

$$H = \sum_{i=1}^{k} p_i \log_2\left(\frac{1}{p_i}\right) \quad (10)$$

where, p_i is the probability of SampEn vectors and $\sum_{i=1}^{k} pi = 1$. The efficiency of the classification into classes C_1 and C_2 is computed by the maximization of entropy (Beenamol *et al*., 2012) technique which yields an optimal threshold, T_{VUV}, given by Eq. 11:

$$T_{VUV} = \text{argmax}\{H(C_1) + H(C_2)\} \quad (11)$$

The performance of the classification process is evaluated by Receiver Operating Characteristics (ROC) curve. For the given V/UV classifier, and a frame, there are four possible outcomes. If the frame encountered is voiced and it is classified as voiced, it is termed as true voiced; instead if it is classified as unvoiced, it is termed as false voiced. If the frame encountered is unvoiced it is classified as unvoiced frame, then it is termed as true unvoiced; but, if it is classified as voiced, it is termed as false unvoiced. Figure 11 shows the confusion matrix for the V/UV problem and basic manipulations can be computed from it. The true voiced rate of the classifier can be estimated by Eq. 12:

$$\text{tv rate} = \frac{\text{Voiced correctly classified}(\text{TV})}{\text{Total Voiced}(\text{V})} \quad (12)$$

and false unvoiced rate is estimated as in Eq. 13:

$$\text{fuv rate} = \frac{\text{Unvoiced incorrectly classified}(\text{FUV})}{\text{Total Unvoiced}(\text{UV})} \quad (13)$$

and accuracy for V/UV classification is given by Eq. 14,

$$\text{Accuracy}_{VUV} = \frac{\text{TV} + \text{TUV}}{\text{V} + \text{UV}} \quad (14)$$

The impact of threshold for the V/UV classification problem is shown in Fig. 12. Shannon Entropy (SE) based thresholding proved to be more significant in the process of classification. Here, low voiced frames are also detected. Voiced regions constituting both periodic and aperiodic components (Yegnanarayana *et al*., 1998; D'Alessandro *et al*., 1998) are also retrieved. The issue faced is that, when a frame detected is with both V and UV regions, the frame is considered as V frame for

True Voiced	**False Unvoiced**
False Voiced	**True Unvoiced**

Fig 11: Confusion matrix for V/UV classification

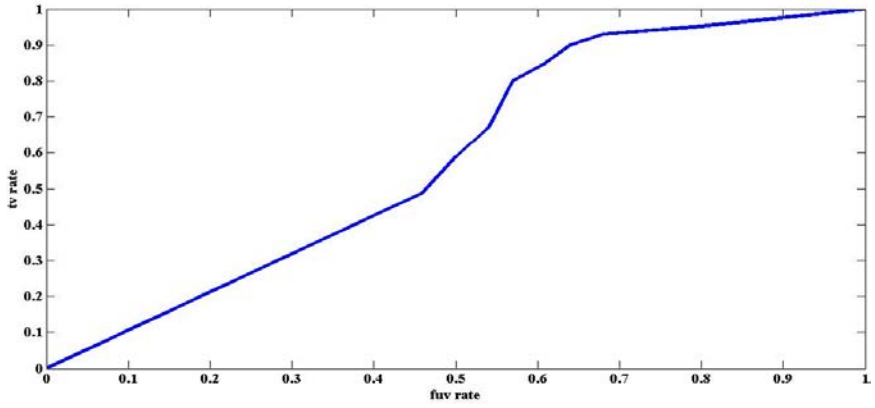

Fig. 12: ROC curve obtained by diminishing the threshold parameter by a factor of 5-50 %

Table 1: Performance of Voiced/Unvoiced Classification under noisy conditions

White	
SNR	Accuracy (%)
Clean	92.76
30 dB	90.86
20 dB	87.14
10 dB	83.51
5 dB	78.78
0 dB	40.37

manipulation purposes. The effectiveness of the feature vector using sample entropy is tested under noisy conditions varying the SNR values for 0-30dB and its accuracy of classification is computed which is given in Table 1.

CONCLUSION

A novel algorithm for the V/UV classification based on time domain approach using sample entropy has been proposed in this article. Moreover, removing background noise (silence) from the speech signal using fuzzy entropy has also been proposed. Features chosen for the process of classification of speech signal into V/UV regions and silence removal seems to be efficient with reasonable accuracy. The research significantly reveals some difficulties related in the setting up of thresholds that can be overcome in the case of silence removal. Averaging threshold works better in noise-free speech signals, but under noisy conditions the threshold selection seems irrelevant. Robustness of sample entropy as a feature for V/UV classification proves to be more reliable even under noisy environments. The proposed algorithm has provided significantly better results than existing methods in terms of V/UV classification accuracy even under noisy environments. One of the main advantages of the proposed algorithm in comparison to the existing methods for the V/UV classification is that selection of optimal threshold based on Shannon entropy over the proposed feature eases the classification of V/UV regions. Moreover, the proposed algorithm does not require knowledge of pitch frequency or GCIs of the speech signal in advance.

REFERENCES

Arifianto, D., 2007. Dual parameters for voiced-unvoiced speech signal determination. Proceedings of the 2007 IEEE International Conference on Acoustics Speech and Signal Processing-ICASSP'07, April 15-20, 2007, IEEE, Honolulu, Hawaii, ISBN: 1-4244-0727-3, pp: 749-752.

Atal, B. and L. Rabiner, 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition. IEEE. Trans. Acoust. Speech Signal Process., 24: 201-212.

Beenamol, M., S. Prabavathy and J.M ohanalin, 2012. Wavelet based seismic signal de-noising using Shannon and Tsallis entropy. Comput. Math. Applic., 64: 3580-3593.

Bosch, L.T., 2003. Emotions speech and the ASR framework. Speech Commun., 40: 213-225.

Childers, D.G. and C.K. Lee, 1991. Vocal quality factors: Analysis synthesis and perception. J. Acoustical Soc. Am., 90: 2394-2410.

Childers, D.G., M. Hahn and J.N. Larar, 1989. Silent and voiced-unvoiced-mixed excitation (four-way) classification of speech. IEEE. Trans. Acoust. Speech Signal Process., 37: 1771-1774.

D'Alessandro, C., V. Darsinos and B. Yegnanarayana, 1998. Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. IEEE. Trans. Speech Audio Process., 6: 12-23.

DeKrom, G., 1993. A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. J. Speech Lang. Hearing Res., 36: 254-266.

Ercelebi, E., 2003. Second generation wavelet transform-based pitch period estimation and voiced-unvoiced decision for speech signals. Appl. Acoust., 64: 25-41.

Erkelens, J.S. and P.M. Broersen, 1998. LPC interpolation by approximation of the sample autocorrelation function. IEEE. Trans. Speech Audio Process., 6: 569-573.

Hariharan, M., C.Y. Fook, R. Sindhu, A.H. Adom and S. Yaacob, 2013. Objective evaluation of speech dysfluencies using wavelet packet transform with sample entropy. Digital Signal Process., 23: 952-959.

Hermes, D.J., 1991. Synthesis of breathy vowels: Some research methods. Speech Commun., 10: 497-502.

Holmberg, E.B., J.S. Perkell, R.E. Hillman and C. Gress, 1994. Individual variation in measures of voice. Phonetica, 51: 30-37.

Klatt, D.H. and L.C. Klatt, 1990. Analysis synthesis and perception of voice quality variations among female and male talkers. J. Acoust. Soc. Am., 87: 820-857.

Kosko, B., 1986. Fuzzy entropy and conditioning. Inf. Sci., 40: 165-174.

Krishnamoorthy, P. and S.M. Prasanna, 2011. Enhancement of noisy speech by temporal and spectral processing. Speech Commun., 53: 154-174.

Maier, A., T. Haderlein, U. Eysholdt, F. Rosanowski and A. Batliner *et al*., 2009. PEAKS-A system for the automatic evaluation of voice and speech disorders. Speech Commun., 51: 425-437.

Naylor, P.A., A. Kounoudes, J. Gudnason and

M. Brookes, 2007. Estimation of glottal closure instants in voiced speech using the DYPSA algorithm. IEEE. Trans. Audio Speech Lang. Process., 15: 34-43.

Paulikas, S. and D. Navakauskas, 2005. Restoration of voiced speech signals preserving prosodic features. Speech Commun., 47: 457-468.

Pinto, N.B., D.G. Childers and A.L. Lalwani, 1989. Formant speech synthesis: Improving production quality. IEEE. Trans. Acoustics Speech Signal Process., 37: 1870-1887.

Qaimkhani, I.A. and E. Hossain, 2008. Efficient silence suppression and call admission control through contention-free medium access for VoIP in WiFi networks. IEEE. Commun. Mag., 46: 90-99.

Qi, Y. B.R. Hunt, 1993. Voiced-unvoiced-silence classifications of speech using hybrid features and a network classifier. IEEE. Trans. Speech Audio Process., 1: 250-255.

Richman, J.S. and J.R. Moorman, 2000. Physiological time-series analysis using approximate entropy and sample entropy. Am. J. Physiol. Heart Circulatory Physiol., 278: H2039-H2049.

Rouat, J., Y.C. Liu and D. Morissette, 1997. A pitch determination and voiced-unvoiced decision algorithm for noisy speech. Speech Commun., 21: 191-207.

Shannon, C.E., 1948. A mathematical theory of communications. Bell Syst. Tech. J., 27: 379-423.

Siegel, L. and A. Bessey, 1982. Voiced-unvoiced-mixed excitation classification of speech. IEEE. Trans. Acoust. Speech Signal Process., 30: 451-460.

Strik, H. and C. Cucchiarini, 1999. Modeling pronunciation variation for ASR: A survey of the literature. Speech Commun., 29: 225-246.

Yegnanarayana, B. and K.S.R. Murty, 2009. Event-based instantaneous fundamental frequency estimation from speech signals. IEEE. Trans. Audio Speech Lang. Process., 17: 614-624.

Yegnanarayana, B., D.C. Alessandro and V. Darsinos, 1998. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. IEEE. Trans. Speech Audio Process., 6: 1-11.

Yin, B., E. Ambikairajah and F. Chen, 2009. Voiced-unvoiced pattern-based duration modeling for language identification. Proceedings of the 2009 IEEE International Conference on Acoustics Speech and Signal Processing, April 19-24, 2009, IEEE, Taipei, Taiwan, ISBN: 978-1-4244-2353-8, pp: 4341-4344