AUTOMATIC IDENTIFICATION OF SILENCE, UNVOICED AND VOICED CHUNKS IN SPEECH

Poonam Sharma[1] and Abha Kiran Rajpoot[2]

[1,2]Assistant Professor, Computer Science Engineering Department
Sharda University, Greater Noida
poonamsharma.2289@gmail.com
akrajpoot@gmail.com

ABSTRACT

The objective of this work is to automatically segment the speech signal into silence, voiced and unvoiced regions which are very beneficial in increasing the accuracy and performance of recognition systems. Proposed algorithm is based on three important characteristics of speech signal namely Zero Crossing Rate, Short Time Energy and Fundamental Frequency. The performance of the proposed algorithm is evaluated using the data collected from four different speakers and an overall accuracy of 96.61 % is achieved.

KEYWORDS

Zero Crossing Rate, Short Time Energy, Fundamental Frequency, Cepstrum.

1. INTRODUCTION

Speech can be considered as a time varying signal whose features changes very frequently when it is being taken for a large time. Classification of the speech signal into regions of silence, voiced and unvoiced can increase the recognition rate and improve the overall performance of the recognition systems. . In silence state of speech no sound is being produced so the energy and the amplitude of the signal is very low. This is important to identify silence region. Once identified than that part of the speech signal can be ignored for further recognition process. In unvoiced stage of speech, vocal cords do not vibrate so the resulting speech is random in nature like the sounds of whisper or aspiration. Finally in the voiced excitation of speech vocal cords are tensed and vibrate periodically. Voiced excitation for the speech sound will result in a pulse train called as fundamental frequency. Voiced excitation is used when articulating vowels and some of the consonants [1].

In recent years considerable efforts has been spent by researchers in solving the problem of classifying speech into silence/voiced/unvoiced parts. Various pattern recognition based [2] and statistical and non statistical techniques has been applied for deciding whether the given segment of a speech signal should be classified as voiced speech or unvoiced speech or silence [3 and 4]. Various other methods based on feed forward networks have also been developed [5, 6 and 7].

David C. Wyld (Eds) : ICCSEA, SPPR, CSIA, WimoA - 2013
pp. 87–96, 2013. © CS & IT-CSCP 2013 DOI : 10.5121/csit.2013.3509

The method we used in this work is a simple and fast approach and can overcome the problem of classifying the speech into silence/voiced/unvoiced. In section 2 we discuss various features and the facts observed from them. In section 3 the proposed work and algorithm is described and in section 4 results are being discussed.

2. Features And Observed Facts

2.1 Zero Crossing Rate

It is a measure of number of times in a given time interval or frame that the amplitude of speech signal passes through a value of zero. The rate at which zero crossing occurs is a simple measure of the frequency content of the signal. This feature is very useful for analysis and segmentation of the speech signal. Voiced speech usually shows a low zero crossing count and generally in quite conditions the zero-crossing count for silence is expected to be lower than for unvoiced speech, but larger or comparable to that for voiced speech [8]. For this application rate at which zero crossing occurs was calculated by taking a window of 20 ms and for the voiced region of the speech signal it was observed that zero crossing rate was always less than 0.1 and for the silence region it was between 0.1 and 0.3. For the unvoiced region ZCR was observed to be greater than 0.3 as shown in Figure 1.

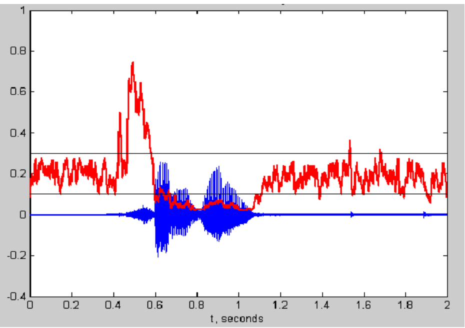

Fig. 1. Zero crossing rate plotted over signal for word "samajhdaar".

But sometime for silence region zero crossing rate was coming out to be less than 0.1 due to some constant background sound as shown in the Figure 2.

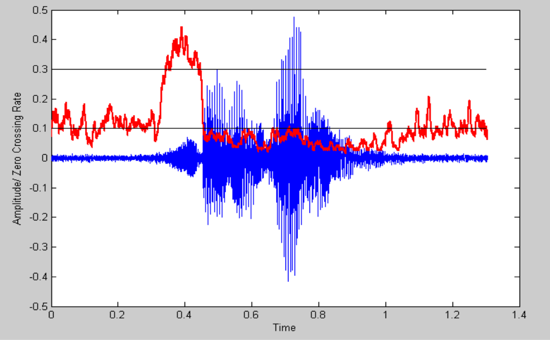

Fig. 2. Signal and ZCR for word "shalgam" showing ZCR less than 0.1 in silence region

Also unvoiced region was having sometimes less zero crossing rate. So it was concluded that zero crossing rate alone cannot be used for classification.

2.2 Short Time Energy

It provides a representation that reflects the amplitude variations in speech signal. In this method hamming window of average size *i.e.* 50ms was taken and short time energy was calculated because large size of the window does not reflect the variations in amplitude and small size does not give a smooth function. For classification purpose threshold value was calculated dynamically using the following process (http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals):-

- Compute the histogram of the feature sequence's values.
- Detect the histogram's local maxima.

Let M1 and M2 be the positions of the first and second local maxima respectively. The threshold value is computed using the following equation :-

$$T = \frac{W.M_1 + M_2}{W + 1}$$

The energy of the voiced sound is much higher than the energy of silence and the energy of unvoiced sounds is lower than for voiced sounds, but often higher than for silence [21]. In this method for voiced region the short time energy was always found to be greater than dynamically calculated threshold value as shown in the Figure 3.

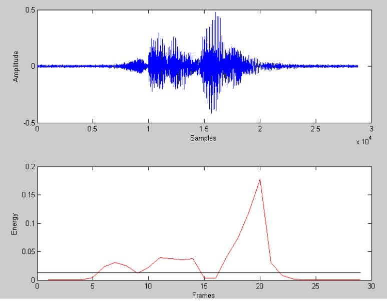

Fig. 3. Signal and STE of "kabutar".

As a result a combination of Short Time Energy and ZCR could be used for the classification purpose.

2.3 Fundamental Frequency

Fundamental frequency also known as pitch is usually the lowest frequency component, or partial, which relates well to most of the other partials. For this method to work cepstrum approach was used and the fundamental frequency was calculated for frames of 40 ms.

It is the quality of pitch that it raises when something is spoken voiced and then falls [19]. So for the unvoiced and silence region fundamental frequency is always zero as shown in Figure 3.This quality of fundamental frequency was used to overcome all the problems that were coming from only taking the ZCR and STE into consideration for the classification purpose

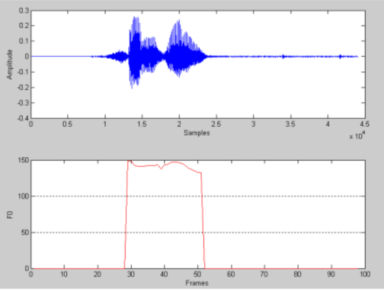

Fig. 4. Signal and F0 of “shalgam” showing zero value of F0 in unvoiced and silence region.

3. METHOD

After analyzing the results from different features calculated the algorithm was designed for the identification of silence, unvoiced and voiced chunks in speech signal. This algorithm was then implemented in MATLAB 2011a.

3.1 Flow Chart

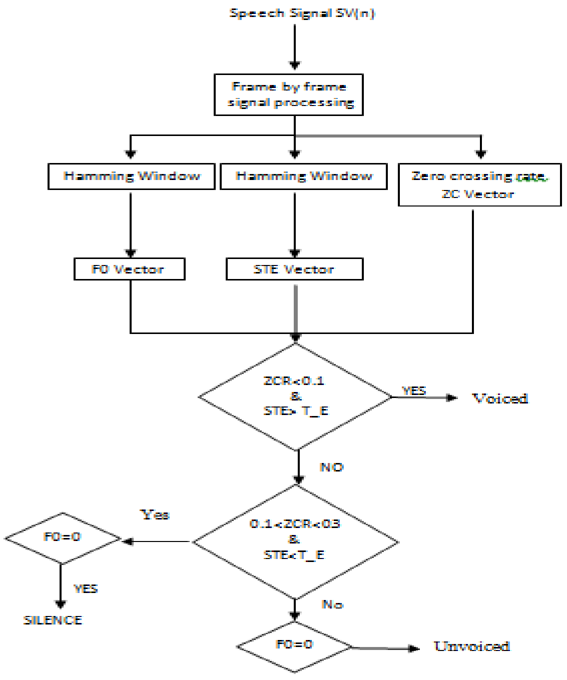

Fig. 5.

4. RESULTS

After applying the algorithm discussed the output was in the form of a matrix and was of the same length as that of the length of the speech signal. The simplest output of the algorithm is shown in Fig. 6.

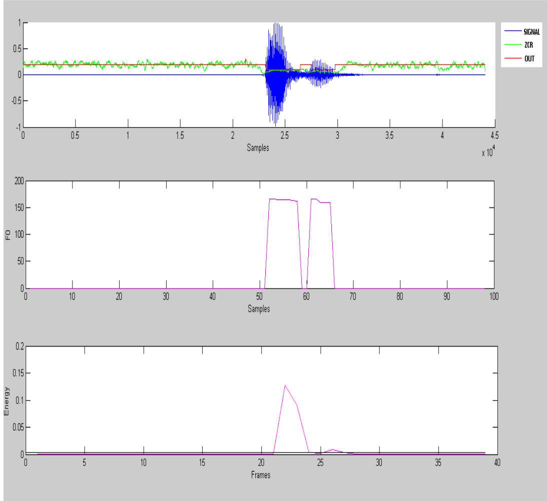

Fig. 6. Output of algorithm for word “bahar”.

The above example is identifying all the parts almost correct and with very high degree of accuracy. But there were some cases when the algorithm was not able to correctly identify very few samples of speech. For example in Figure 5 word “shor” was spoken by a female speaker and it can be clearly seen that some of the unvoiced region (were fricative /sh/ was spoken) came into the category of silence and some of the silence region (in the end).

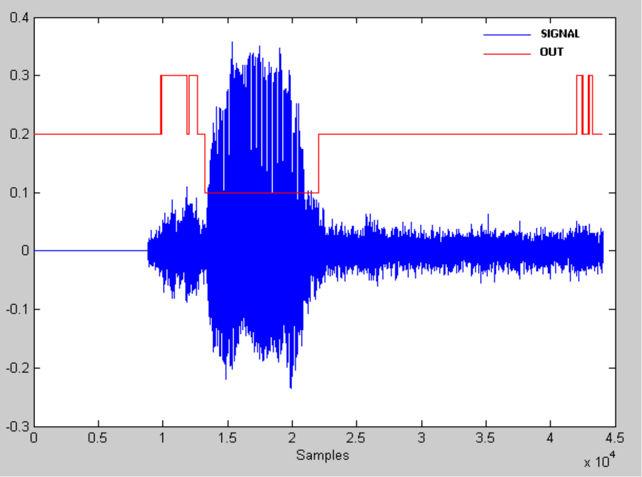

Fig. 7. Output of algorithm for word “shor” spoken by female speaker.

The algorithm was applied to all the 15 words spoken in Hindi by four persons (3 male and 1 female). Accuracy of the algorithm was calculated by checking how many samples in the spoken word were identified correctly compared to the manual classification of the voiced, unvoiced and silence region in the word and then dividing them by total no of samples. The accuracy of the algorithm for four different speakers taking all the 15 words spoken3 times is shown in Table 1, 2, 3 and 4.The overall accuracy of the algorithm is shown in Table 5.

Table 1:Accuracy of first speaker (male)

Word Spoken	Accuracy (Spoken 1st time)	Accuracy (Spoken 2nd time)	Accuracy (Spoken 3rd time)	Average accuracy
"ajay"	99.73	99.71	99.2	99.54
"kabutar"	95.7	98.6	97.5	97.26
"shalgam"	92.6	90.9	98.9	94.13
"ghar"	94.6	94.7	93.6	94.3
"bahar"	91.4	89.85	90.8	90.68
"aag"	95	94.3	96.6	95.3
"aam"	98.4	96.1	97.27	97.25
"dhaga"	96.8	98.18	93.75	96.24
"gadi"	98.9	99.3	95.4	97.86
"ghadi"	98.43	96.45	96.78	97.22
"ghas"	97	99.5	99.18	98.56
"hawa"	97.2	98.13	97.8	97.71
"kal"	98.7	99.2	99.1	99
"mitti"	95.4	94.13	92.1	93.87
"shor"	97.89	98.54	98.2	98.21
Average accuracy for speaker1				96.47

Table 2:Accuracy of second speaker(male)

Word Spoken	Accuracy (Spoken 1st time)	Accuracy (Spoken 2nd time)	Accuracy (Spoken 3rd time)	Average accuracy
"ajay"	93.9	94.3	94.2	94.13
"kabutar"	96.4	97.2	92.3	95.3
"shalgam"	95	96.2	93.6	94.93

“ghar”	92.4	86.4	95.7	91.5
“bahar”	98.4	89.9	92.3	93.53
“aag”	98.7	98.34	99.18	98.74
“aam”	98.2	98.56	96.7	97.82
“dhaga”	96.7	98.4	93.24	96.11
“gadi”	99.52	94.9	90.45	94.96
“ghadi”	96.7	95.62	96.9	96.41
“ghas”	94.42	88.2	92.3	91.64
“hawa”	96.5	97.8	98.1	97.47
“kal”	98.2	99.13	98.46	98.6
“mitti”	96.4	92.61	97.28	95.43
“shor”	92.1	86.8	88.9	89.27
Averageaccuracy forspeaker2				95.06

Table 3: Accuracyof third speaker (female)

Word Spoken	**Accuracy (Spoken 1st time)**	**Accuracy (Spoken 2nd time)**	**Accuracy (Spoken 3rd time)**	**Average accuracy**
“ajay”	99.5	98.8	98.8	99.03
“kabutar”	96.5	98.18	97.7	97.46
“shalgam”	97.75	96.59	99.5	97.95
“ghar”	97.5	97.73	98.18	97.80
“bahar”	99.2	98.6	98.8	98.87
“aag”	98.4	98.12	97.24	97.92
“aam”	99.1	98.62	99.32	99.01
“dhaga”	99.47	97.67	99.23	98.79
“gadi”	97.16	98.7	98.6	98.15
“ghadi”	98.68	99.13	99.2	99

“ghas”	97.25	97.8	97.34	97.46
“hawa ”	98.7	95.67	97.54	97.3
“kal”	98.74	99.56	98.9	99.07
“mitti”	98.3	98.57	98.18	98.35
“shor”	99.6	99.23	97.34	98.72
Averageaccuracy forspeaker3				98.33

Table 4:Accuracy of fourth speaker(male)

Word Spoken	**Accuracy (Spoken 1st time)**	**Accuracy (Spoken 2nd time)**	**Accuracy (Spoken 3rd time)**	**Average accuracy**
“ajay”	97.29	98.64	98.8	98.24
“kabutar”	96.4	98.59	97.27	97.42
“shalgam”	95.5	97.02	95.4	95.97
“ghar”	99.45	97.25	98.6	98.43
“bahar”	96.75	98.4	98.4	97.85
“aag”	96.85	97.50	93.40	95.92
“aam”	97.08	97.42	96.05	96.85
“dhaga”	94.35	96.63	98.25	96.38
“Gadi”	94.88	94.44	94.82	94.71
“Ghadi”	95.36	97.46	95.60	96.14
“Ghas”	97.69	95.05	93.03	95.26
“Hawa”	98.02	99.24	97.09	98.12
“Kal”	94.74	97.24	94.53	95.5
“Mitti”	94.64	96.42	98.76	96.61
“Shor”	93	96.78	96.5	95.43
Averageaccuracy forspeaker4				96.58

Table 5: Overall accuracy of algorithm

Speaker	**Average accuracy**	**Overall accuracy**
1	96.47	**96.61**
2	95.06	
3	98.33	
4	96.58	

5. CONCLUSION

This algorithm is efficient in solving the problem of identifying the unvoiced, voiced and silence chunks in speech. Three fundamental features namely: ZCR, STE and F0 are used in the algorithm for the classification purpose and an accuracy of 96.61 % is achieved. The errors in the system are mainly in the starting and the ending of the word due to little noise or lower energy during the starting and ending of the word.

REFERENCES

[1] Rabiner, L. and Juang, B. H., 1993. Fundamental of Speech Recognition. PTR Prentice-Hall, New Jersey.

[2] Atal, B. S. and Rabiner, L. R., 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to Speech Recognition. IEEE transactions on acoustic, speech, and signal processing, vol. 24, no. 3, pp. 201-212.

[3] Chung, M., Kushner, W. M. and Damoulakis, J. N., 1985. Word Boundary Detection and Speech Recognition of Noisy Speech by Means of Iterative Noise Cancellation Techniques. IEEE International Conference on ICASSP. Vol. 10, pp. 1838.

[4] Cox, V. B. and Timothy, L. V., 1980. Nonparametric rank order statistics applied to robust voiced-unvoiced-silence classification. IEEE transactions on acoustic, speech, and signal processing, vol. 28, issue. 5, pp. 550-561.

[5] Ghiselli-Crippa, T. and El-Jaroudi, A., 1991. A fast neural net training algorithm and its application to voiced-unvoiced-silence classification of speech. IEEE transactions on acoustic, speech, and signal processing, vol. 1, pp. 141-144

[6] Gupta, R, 2006. Speech Recognition for Hindi, M. Tech Thesis, Dept. Computer Science and Eng. IIT Bombay, Bombay.

[7] Qi, Y. and Hunt, B. R., 1993. Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and a Network Classifier. IEEE transactions on speech and audio processing, vol. 1, no. 2, pp. 250-255.

[8] Atal, B. S. and Rabiner, L. R., 1976. A pattern recognition approach to voiced-unvoiced-silence classification with applications to Speech Recognition. IEEE transactions on acoustic, speech, and signal processing, vol. 24, no. 3, pp. 201-212.

[9] (http://www.mathworks.com/matlabcentral/fileexchange/28826-silence-removal-in-speech-signals):-

[10] Sarma, V. V. S. and Venugopal, D., 1978. Studies on pattern recognition approach to voiced-unvoiced-silence classification. IEEE International Conference on ICASSP, Vol. 3, pp. 1-4.

[11] Zhao, X., O‟Shaughnessy, D. and Minh-Quang, N., 2007. A Processing Method for Pitch Smoothing Based on Autocorrelation and Cepstral F0 Detection Approaches. International Symposium on Signals, Systems and Electronics, pp. 59-62.

[12] Raman Rao, G. V. and Srichand, J., 1996. Word boundary detection using pitch variations. Forth international conference of spoken language, vol. 2, pp. 813-816.