Evaluation of various features of Gujarati continuous numerals speech signal used for segmentation

[1] Bharat C. Patel & [2]Apurva A. Desai
[1]Asst. Professorand I/c Principal,[2]Professor and Head
[1]Smt. Tanuben& Dr. Manubhai Trivedi college of information Science, Surat, India
2Department of Computer Science, Veer Narmad South Gujarat University, Surat, India

Received: April 01, 2019 Accepted: May 12, 2019

ABSTRACT: *Speech is a common medium used for communication by human being. In modern era of information technology, speech is also used for communication with machine. To communicate with machine through speech, it is required to analyze speech signal and segment it into basic units such as words, phonemes or syllables. For large vocabulary and continuous speech recognition, the sub word-unit-based approach is a feasible option than the whole word-unit-based approach. For preparing a large inventory of sub word units, an automatic segmentation is preferable to manual segmentation as it significantly reduces the work associated with the generation of templates and gives more reliable results. This paper provides a guideline for the readers, working in the field of automatic speech recognition. An overview of various features used for automatic segmentation based on the various research papers is also presented and reviewed in the paper. In this paper we discuss some features for automatically segmentation of speech signal into phonetic units and also explain the result obtained by different features of continuous speech signal of Gujarati numerals.*

Key Words: *Speech Segmentation, Time-domain Features, Frequency-domain features.*

I. Introduction

In the field of computer science, human-computer interaction is possible with the help of one of the technologyknown as speech recognition. Automatic Speech Recognition (ASR) is allows a computer to recognize the words spoken by a person through telephone, microphone or other devices. In other words, automatic speech recognition is a process that converts analogue signal into a sequence of words in any regional language as perthe algorithm implemented as a computer program. A speech signal may be discrete or continuous. A discrete speech signal consists of a phonetic utterance whereas a continuous speech signal consists of two parts: one carries the speech information known as verbal part of speech signal, and the second carries the silent or noise, without any verbal information known as non-verbal part of speech signal, that are in between the utterances. The verbal part of speech can be further divided into two main types: Voiced and Unvoiced speech in the literature(AlaaEhabSakran et al, 2017).

When an air flow from the lungs; the vocal cords are tensed and vibrated periodically as a resultit generates quasi-periodic speech waveform known as voiced speech signal. The silent part of speech signal contains no speech; unvoiced speech signal is produced when vocal cords are not vibrated so that resulting speech waveform is aperiodic or random in nature (Rabiner, L. R. et al., 1993 and Bharat C. Patel et al., 2015).

Speech segmentation plays an important role in the development of a syllable-centric automatic speech recognition system. The purpose of segmentation is to convert speech signal into a phonetic unit. Speech segmentation is a process of decomposing the speech signals into acoustic units serially. In the last few years, a variety of segmentation algorithms have been proposed and shown to work on specific tasks. However, a robust and general-purpose solution for real-time ASR is not there. Speech segmentation is useful in various applications like speech recognition, speech synthesis, speech enhancement, speech corpus collection, speaker verification and in the research field of natural language processing.Basically, there are two approaches to segment speech signal that is manual and automatic segmentation.

Most of the procedures have followed one of the two basic approaches to the problem. The first approach is to utilize the explicit information that is known as a priori, such as the correct phonetic transcription of the utterance. The incoming speech signal is then segmented using reference templates corresponding to the phonetic events. The second approach does not require any explicit information, but utilizes only the acoustical information that is contained within the speech signal to be segmented, such as the amount of spectral change from one speech frame to the next.

This Paper is organized as follows: Section II describes the literature survey on feature used by authors for automatic speech segmentation; Section III describes the features of automatic speech segmentation; Section IV describes the segmentation results obtained by different features speech signal and last section represents the conclusion of the paper.

II. Literature Survey

Over the past years, several procedures for automatic segmentation of speech signal have been proposed in the literature. Speech can be efficiently segmented into its basic units which are words, phonemes and syllables using automatic segmentation.

V. Kamakshi Prasad et. al. (2004) proposed a novel approach for segmenting the speech signal into syllable-like units. They proposed a group delay based approach to processing the short-term energy for determining segment boundaries. The performance of this technique is tested on both continuous speech utterances and connected digit sequences. It is shown that the segmentation performance is quite satisfactory. The error in segment boundary is less than equal to 20% of syllable duration for 70% of the syllables. In addition to true segments, an overall 5% insertions and deletions have also been observed.

D.S.Shete and S. B. Patil(2014)proposed two methods to separate the voiced-unvoiced parts of speech from a speech signal. These are zero crossing rate (ZCR) and energy. They evaluated the results by dividing the speech sample into some segments and used the zero crossing rate and energy calculations to separate the voiced and unvoiced parts of speech. The results suggest that zero crossing rates are low for voiced part and high for unvoiced part where as the energy is high for voiced part and low for unvoiced part. Therefore, these methods are proved more effective in separation of voiced and unvoiced speech.

Bharat C. Patel and Apurva A. Desai (2015)presented a model which uses speech features such as Short-Term Energy (STE), Zero-Crossing Rate (ZCR) and peaks, and developed an algorithm to segment continuous Gujarati speech signal into word or sub-words. The experiments are carried out on continuous Gujarati speech signal and obtained results are presented.

TorbjomSvendsent and Frank K. Soong (1987) discussed some methods for automatically segmenting speech into phonetic units. They described three different approaches, one based on template matching, one based on detecting the spectral changes that occur at the boundaries between phonetic units and one based on a constrained-clustering vector quantization approach. An evaluation of the performance of the automatic segmentation methods is given. The methods were tested experimentally in a speaker independent manner and the results were compared with manual segmentation using coincidence rate as a performance measure. The template matching approach resulted in a higher coincidence rate (92%) than the constrained clustering method (76%). An algorithm combining the two methods did not improve the coincidence rate.

Anupriya Sharma and Amanpreet Kaur (2013) described the process of automatic segmentation of speech using group delay technique. This includes segmentation of continuous Punjabi speech into syllable like units by using the high resolution properties of group delay. This group delay function is found to be a better representative of the STE function for syllable boundary detection.

Nipa Chowdhuryet. al. (2009) presented a new word separation algorithm for Continuous Bangla Speech Recognition. The algorithm is developed by considering prosodic feature with energy to separate Bangla speech into words. The result shows that energy and pitch information is important parameter for word separation. They used features like energy, zero crossing rate and pitch to separate words and result shows that 98% word boundaries are correctly detected.

Md. Mijanur Rahman and Md. Al-Amin Bhuiyan(2013)presented several dynamic thresholding approaches for segmenting continuous Bangla speech sentences into words/sub-words. They have proposed three efficient methods for speech segmentation: two of them are usually used in pattern classification (i.e., k-means and FCM clustering) and one of them is used in image segmentation (i.e., Otsu's thresholding method). They also used new approaches blocking black area and boundary detection techniques to properly detect word boundaries in continuous speech and label the entire speech sentence into a sequence of words/sub-words. K-Means and FCM clustering methods produce better segmentation results than that of Otsu's Method. The proposed system achieved the average segmentation accuracy of 94% approximately.

AgnelWaghelaet. al.(2014) discussed the implementation of an algorithm which automatically detects the silence, voiced and unvoiced parts of a speech signal, which can drastically improve the accuracy of a speech recognition system. The algorithm is based on three important characteristics of a speech Signal – Zero Crossing Rate, Short Time Energy and Fundamental Frequency.

T. Nagarajanet. al. (2003)presented a minimum phase group delay based approach to segment spontaneous speech into syllable like units. Here, three different minimum phase signals are derived from the short term energy functions of three sub-bands of speech signals, as if it were a magnitude spectrum. The experiments are carried out on Switchboard and OGI-MLTS corpus and the error in segmentation is found to be utmost 40msec for 85% of the syllable segments.

Md. Mijanur Rahman and Md. Al-Amin Bhuiyan(2012)presented simple and novel feature extraction approaches for segmenting continuous Bangla speech sentences into words/sub-words. These methods are based on two simple speech features, namely the time-domain features and the frequency-domain features. The time-domain features, such as short-time signal energy, short-time average zero crossing rate and the frequency-domain features, such as spectral centroid and spectral flux features are extracted in this research work. After the feature sequences are extracted, a simple dynamic thresholding criterion is applied in order to detect the word boundaries and label the entire speech sentence into a sequence of words/sub-words. The proposed automatic speech segmentation system achieved segmentation accuracy of 96%.

Yin Win Chit and Dr.Renu(2017) proposed the time-domain features and frequency-domain features based on fuzzy knowledge for continuous speech segmentation task via a nonlinear speech analysis. They used time-domain features such as short-time Energy and Zero-crossing Rate and frequency-domain feature such as Spectral Centroid for generating the significant segments. Fuzzy Logic technique will be used not only to fuzzify the calculated features into three complementary sets namely: low, middle, high but also to perform a matching phase using a set of fuzzy rules. The output of the Fuzzy Logic is phonemes, syllables and disyllables of Myanmar Language. The result of the system will recognize the continuous words of input speech.

III. Features of automatic speech segmentation

Automatic speech recognition requires evaluate the features of speech signal to identify important part of the audio segment. By the nature of production, a continuous speech signal contains voiced, unvoiced and silence regions. In order to separate voiced and unvoiced part from continuous speech signal, it requires to track the amplitude or spectral changes in the signal by using short-time energy or spectral features and to detect the segment boundaries at the locations where amplitude or spectral changes exceed a minimum threshold level. This can be possible through two types of features: Temporal features also known as time-domain features and spectral features which are also known as frequency-domain features.

The temporal features are simple to extract and have easy physical interpretation. The energy of signal, zero crossing rate, maximum amplitude, minimum energy is time domain features and they are widely used for speech segment extraction (Md. Mijanur Rahmanand Md. Al-Amin Bhuiyan, 2012).The spectral features are obtained by converting the time based signal into the frequency domain using the Fourier Transform. The fundamental frequency, frequency components, spectral centroid, spectral flux, spectral density, spectral roll-off is frequency based features and they can be used to identify the notes, pitch, rhythm, and melody.

3.1. Time-Domain Signal Features

The energy associated with speech is time varying in nature. Hence, in automatic processing of speech it is necessary to know that how the energy is varying with time and to be more specific, how energy associated with short term region of speech signal. The most used time-domain features are short-term average zero-crossing rate, short-time energy, Average Magnitude Difference Function (AMDF), Short time Auto correlation and Pitch Period Extraction using Short time Autocorrelation.

3.1.1. Short-term Zero-Crossings Rate

In the context of discrete-time signals, a zero crossing is occurred if successive samples have different algebraic signs. It indicates that the rate at which zero crossings occur is a measure of the frequency content of a signal. Zero Crossing Rate (ZCR) is compute number of times in a given time interval that the amplitude of the speech signals passes through a value of zero.

The short-term zero-crossing rate is defined using Eq1.as:

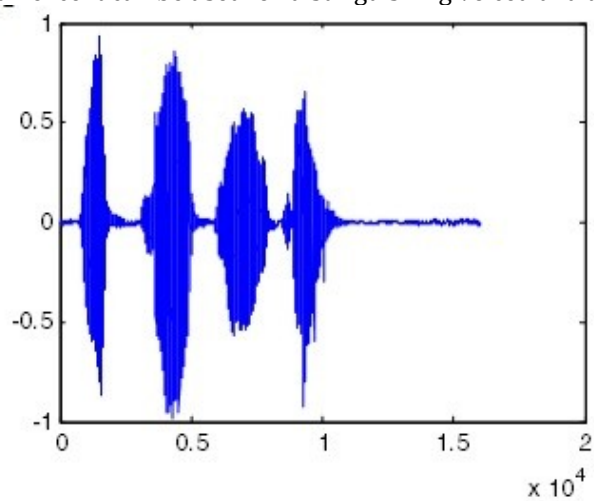
$$Z_x = \sum_m |sgn[s(m)] - sgm[s(m-1)]|w(n-m) \qquad (1)$$

Where sgn (s(m))= 1 if s(m) ≥ 0
= -1 if s(n)<0

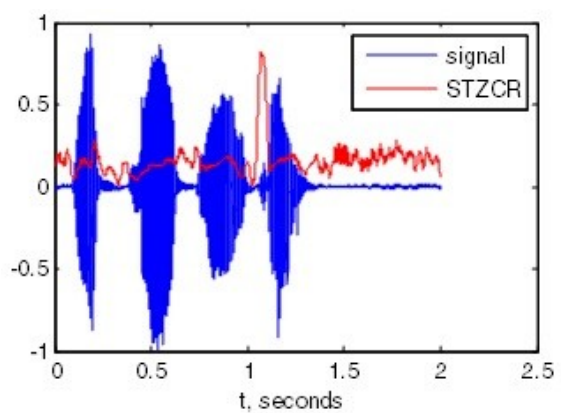
Where s(m) is the speech signal and w(m) is the window. ZCR provides information about the number of zero-crossings present in a given frame. Obviously, if the number of zero crossings is more in a given frame

indicate that the signal is changing rapidly and may contain high frequency information. On the other hand, if the number of zero crossing is less in a given frame indicate that the signal is changing slowly and may contain low frequency information. Thus ZCR gives indirect information about the frequency content of the signal. A reasonable generalization is that, if the ZCR is high, it is Unvoiced and if ZCR is low, it is Voiced speech.

In case of speech, the nature of signal changes with time over few milliseconds. For example, from initial voiced to unvoiced and back to voice and so on. ZCR is computed frame by frame manner and the typical frame size of 10-30 milliseconds with half the frame size as shift is used. Figure 1 shows a speech signal of the continuous Gujarati numerals "1234" and in figure 2 computation of short term ZCR shown in red color. As it can be observed that in case of unvoiced sounds, the ZCR value is significantly high compared to the region of voiced sounds and hence it can be used for distinguishing voiced and unvoiced regions.



[Figure: 1 Speech wave of continuous Gujarati numerals'1234']



[Figure: 2 Short-term Zero Crossing Rate of continuous Gujarati numerals '1234']

3.1.2. Short Term Energy

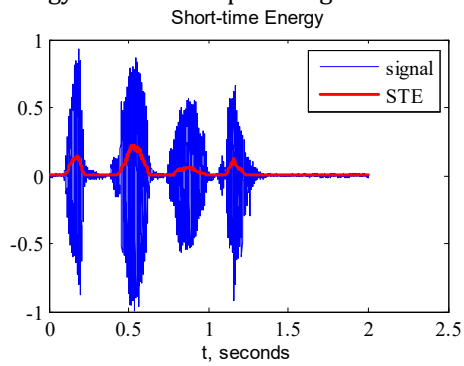
Like energy of speech, the amplitude of the speech signal also varies with time. In general, the amplitude of unvoiced speech segment is much lower than the amplitude of voiced segment. The Short term energy can be calculating using Eq 2:

$$E_n = \sum_m [s(m)w(n-m)]^2 \qquad (2)$$

Where s(m) is a short time speech segment obtained by passing the speech signal x(n) through window w(n).

The major implication of this is that it provides a foundation for differentiate voiced speech from unvoiced speech. However, one difficulty with Short term energy is that it is very sensitive to the large signal amplitude levels, because it's a square function.

To computer the short term energy, the signal is divided into frames. The length of frame size is in the range of 10 to 30 milliseconds. The typical value for the frame size is about 20 milliseconds. On the other hand, we get much smoothed version of energy for larger frame sizes and cannot find time varying nature of short term energy. Figure 3 shows the energy contour for speech signal taken for study.



[Figure: 3 Short term energy of continuous Gujarati numerals '1234']

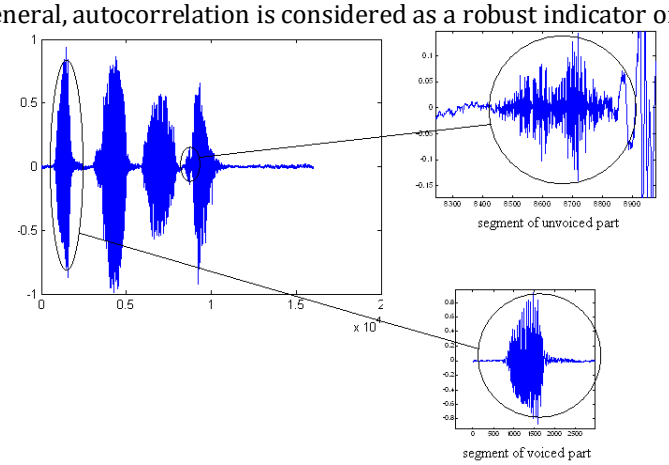
3.1.3. Short time Auto correlation function:

To find the similarity between two sequences of speech signal can be computed using the cross correlation tool of signal processing. That means cross correlation refers to the case of having two different sequences for correlation whereas the autocorrelation refers to the case of having only one sequence for correlation. The main significanceof autocorrelation is to examine how similar the signal characteristics with respect to time. We can achieve thisby providing different time lag for the sequence and computing with the given sequence as reference. The autocorrelation is a very useful in case of speech processing. However due to the non-stationary nature of speech signal, a short term version of the autocorrelation is needed. Short-time autocorrelation is defined using Eq 3 as follow.

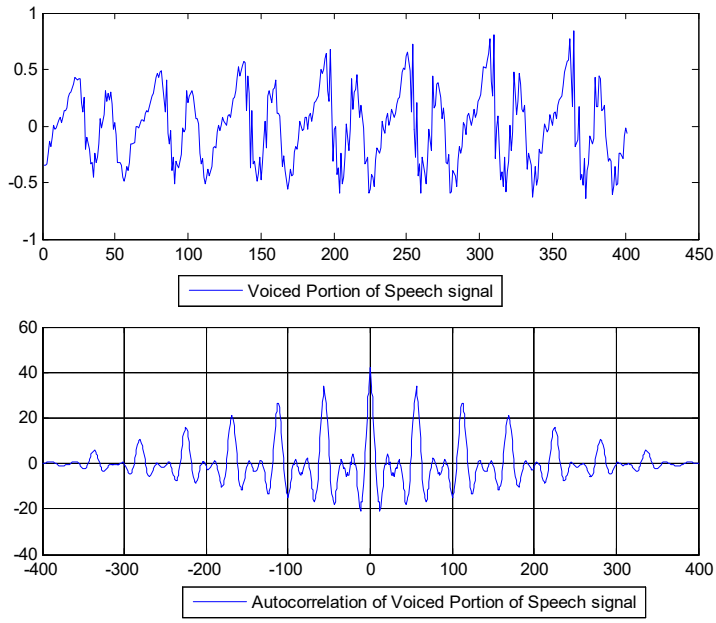
$$R_n(k) = \sum_{m=0}^{m=N-1-k} \left[x(n+m)w'(m)\right]\left[x(n+m+k)w'(k+m)\right] \qquad (3)$$

Where, $R_n(k)$ is a short time autocorrelation at sample n in the signal x. W is a window and it is actually the autocorrelation of a windowed speech segment. The short-time autocorrelation tells the periodicity in a signal. To reveal periodicity of signal, we have taken voiced and unvoiced segment of continuously spoken Gujarati numerals '1234'. As shown in figure 4 that samples 8400:8800 is taken as unvoiced segment and samples 1000:1400 is taken as voiced segment of continuous spoken Gujarati numerals '1234'.

Here, we notice in figure 5 that how the autocorrelation of the voiced speech segment retains the periodicity. On the other hand in figure 6, we see that the autocorrelation of the unvoiced speech segment looks more like noise. Figure 5 and Figure 6 show segments of voiced and unvoiced speech with their corresponding autocorrelation sequences respectively. As we can see that the nature of short term autocorrelation sequence is primarily different for voiced and unvoiced segments of speech. Hence information from the autocorrelation sequence can be used for make distinction between voiced and unvoiced segments.In general, autocorrelation is considered as a robust indicator of periodicity.

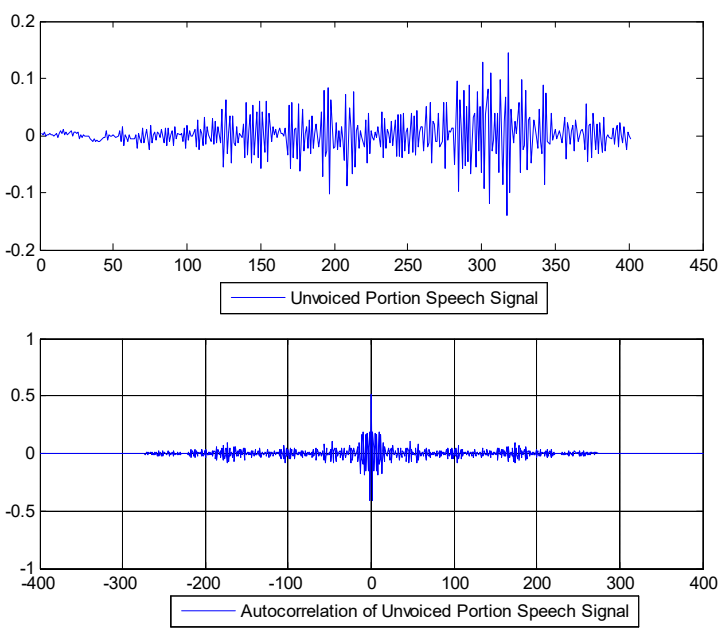


[Figure 4: Segments of voiced and unvoiced part of continuous Gujarati numerals '1234']



[Figure 5**:** voiced segments of speech and Autocorrelation sequence]

The experimental result shows that the autocorrelation of voiced speech should give strong peak at the periodic value and no such peak in case of unvoiced speech. Therefore, the autocorrelation of speech has become a standard approach for enhancing pitch. Autocorrelation function is very useful tool in speech processing and used to identify the similarities of speech characteristics with respect to time.



[Figure 6**: Un**voiced segments of speech and Autocorrelation sequence]

3.1.4. Average Magnitude Difference Function (AMDF)

The concept of AMDF is very close to auto correlation function except that it estimates the distance instead of similarity between a frames. For discrete signalsAMDF function is given by the Eq 4 as follow:

$$M_n = \sum_m |s(m)|w(n-m) \qquad (4)$$

For the average magnitude computation, the dynamic range (ration between maximum to minimum) is approximately the square root of the dynamic range for the standard energy computation. Thus the differences in level between voiced and unvoiced regions are not pronounced as for the short time energy.
There are many variations of AMDF found in the literature such as High Resolution AMDF (HRAMDF), Circular AMDF (CAMDF), and so on. Ghulam Muhammad (2011) presented a paper on Extended Average Magnitude Difference Function (EAMDF). In his work the EAMDF spread over previous and next frames along with current frame, and thereby possesses greater smoothing power. An experiment shows the efficient noise robustness of the proposed method both in white and color (restaurant) noise. This method can significantly contribute to speech/music discrimination, voice-enabled security, among others.

3.2. Frequency-Domain Signal Features

The time domain features such as energy, zero crossing rate and autocorrelation can be computed using the short term time domain analysis. The different frequency or spectral components in the speech signal are not directly visible in the time domain. Using Fourier transform representation,we can convert time domain representation into frequency domain representation. The conventional Fourier representation of signal processing is inadequate to provide information about the time varying nature of spectral information present in speech. Hence, the need for short term version of Fourier transform more commonly known as Short term Fourier Transform (STFT). Frequency-domain analysis is extensively used in the area of communications, geology, remote sensing, and image processing. The time-domain analysis demonstrates how a signal changes over time whereas frequency-domain analysis give you an idea about how the signal's energy is distributed over a range of frequencies. A frequency-domain representation also includes information on the phase shift that must be applied to each frequency component in order to recover the original time signal with a combination of all the individual frequency components. Commonly used frequency-domain features are spectral centroidand spectral fluxfeature sequences that used discrete Fourier transform.

3.2.1. Spectral Centroid

In digital signal processing,the spectral centroid is a measure used to characterize a spectrum. It is usuallyrelated with the determination of the brightness of a sound. This measure is obtained by evaluating the “center of gravity” using the Fourier transform’s frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes as shown in Eq 5 follow:

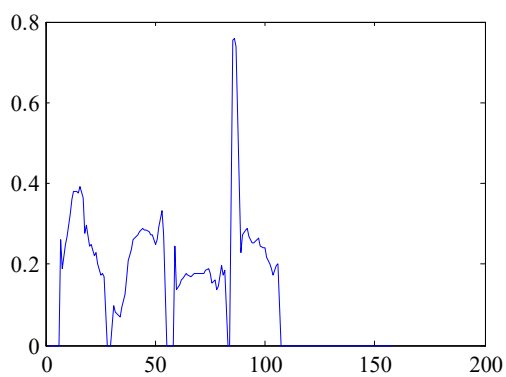
$$centroid = \frac{\sum_{n=0}^{N-1} f(n)x(n)}{\sum_{n=0}^{N-1} x(n)} \qquad (5)$$

Where x(n) is the magnitude of bin number, f(n) is the center frequency of that bin in Discrete Fourier Transform (DFT) spectrum. The DFT is given by the Eq 6 and can be computed efficiently using a fast Fourier transform (FFT) algorithm (Cooley et. al., 1965).

Type equation here.

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-j 2\pi k \frac{n}{N}} \; ; k = 0, \ldots, N-1 \qquad (6)$$

The "center of gravity" of the spectrum iscomputed using Spectral Centroid. This feature is a measure of the spectral position, with high values corresponding to “brighter” sounds (T Giannakopoulos, 2009) as shown in Figure-7.

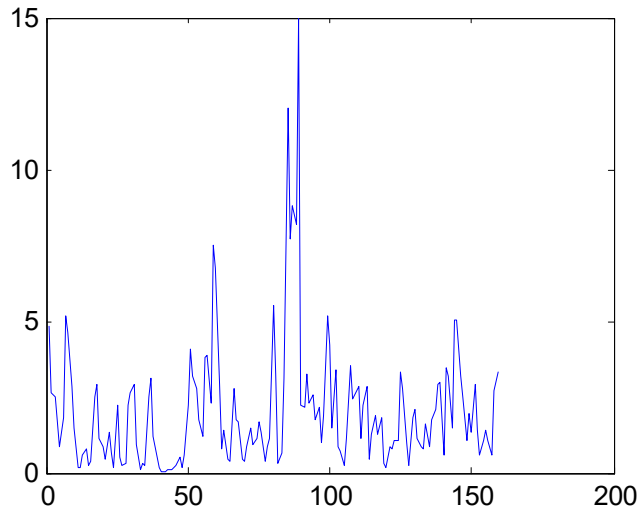


[Figure: 7 Spectral centroid of continuous Gujarati numerals ‘1234’]

It is not sure if the spectral centroid would be useful for classifying different genres of music but will show some spectral components of the music, which are mixed sounds.

3.2.2. Spectral flux

Spectral flux is measured the changing of power spectrum of the signal(as shown in Figure 8) and it is calculated by comparing the power spectrum for one frame against the power spectrum from the previous frame which is also known as the Euclidean distance between the two normalized spectra. It is most important feature to separate the music from speech signal.The spectral flux can be used to determine the timbre of an audio signal, or in onset detection (Bello J P et. al., 2005) among other things.



[Figure: 8 spectral flux of continuous Gujarati numerals '1234']

Spectral flux is defined as squared difference between two normalized magnitude of successive spectral distribution and it represent the successive signal frames. It is calculated by the Eq 7 as:

$$SF_i = \sum_{k=1}^{N/2} (|X_i(k)| - |X_i(k-1)|)^2 \quad (7)$$

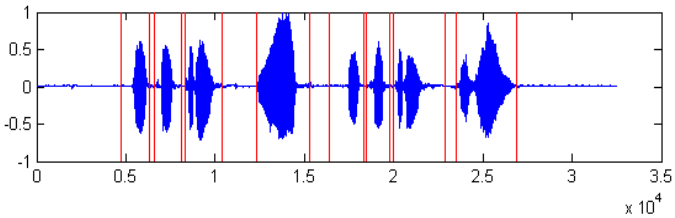
Where X (k) is the DFT coefficient of ithshort term frame with length N. Spectral flux is used to find out the tone of audio signal.

Spectral flux is also called Spectral Variation. Spectral flux is described as the variation value of spectrum between the adjacent two frames in a short-time analyze window.

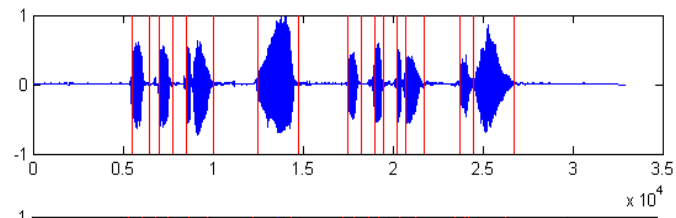
IV. Experimental results

The recording was carried out in the environment which was not truly noiseless. Speech signal are recorded in .wav form using headphone, sampling frequency of the signal is 8000Hz and recording time duration is 4 to 5 seconds. We have collected 50 continuousally spoken Gujarati numerals for experimental purpose. All the data collected in natural environment so it may have external noise like fan, backgound music, vaccume cleaner, phoen ringing and so on. Moreover, the recording instruments were also produced a little noise in some cases. We have taken time domain features such as STE and ZCR and frequency domain feature as spectral centroid to segment speech signal of continuous Gujarati numerals.

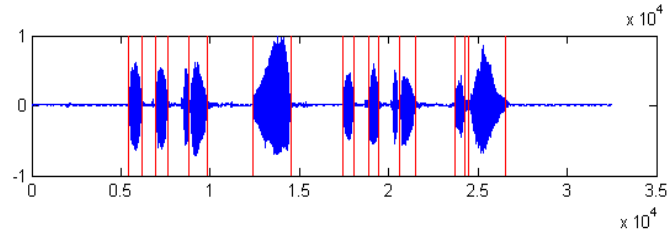
Figure: 9(A), 9(B) and 9(C) shows the segmentation result of continuous Gujarati numerals 3131000 (એકત્રીસ લાખ એકત્રીસ હજાર) using spectral centroid, zero-crossing rate and short-term energy respectively. It is clear that using spectral centroid feature (figure:9(A)) the word હજાર is segmented as whole word whereas using Zero-crossing rate and short-Term energy features it can be broken down into two sub-word as હજ + ાર(figure: 9(B) and 9(C)).



[Figure 9(A): segmentation of continuous wave of '3131000.wav' using Spectral Centroid]



[Figure 9(B): segmentation of continuous wave of '3131000.wav' using Zero-Crossing Rate]



[Figure 9(C): segmentation of continuous wave of '3131000.wav' using Short-Term Energy]

There are total number of 230 utterances presence in 50 continuous Gujarati numerals. The performance of the segmentation of these continuous numerals are shown in Table 1.

Table 1: Segmentation Result of speech signal using different features

Sr. No.	Features	No. of utterances successfully segmented	Percentage of correct segments
1	Zero-Crossing Rate	216	93.91%
2	Short-Term Energy	218	94.78%
3	Spectral Centroid	212	92.17%
4	ZCR + STE	224	97.39%

As we can see that using individual features, we get good segmentation result for coninuous Gujarati numerals but success rate is very much increases using hybride features that is zero-crossing rate and short term energy features. We also observed that, some of the continuous Gujarati numerals are not segmented correctly. This is happen because, there might be presence of noise or not spoken in correct manner at the time data collection. If we use sound proof environment and spoken in correctly manner at the time of data collection, would increased segmentation result. Alternatively, we can improve the segmentation result by using more hybridge features of speech signal at the time of processing.

V. Conclusion

In this paper, we discussed different time domain features such as short-time energy, zero crossing rate, autocorrelation methods etc and frequency domain features such as Spectral Centroid and Spectral flux methods and the use these features in signal processing. Moreover, the researchers new in the field of signal processing can easily understand the importance of these features. This paper provides basic idea to the researchers about the detection of voiced and unvoiced/silent part from continuous speech signal using the features discussed above.

From the experimental result, we can hypothesize that the segmentation of continuous speech signal using time domain features, like ZCR and STE, we obtained good result compared to frequency domain features. The performance of segmentation was evaluated on 230 utterances and it was found that about 97.39% of the syllables boundaries were identified successfully using zero-crossing and short term energy hybrid feature. In future, we can use either other feature or fusion features of speech signal for segmentation purpose.

REFERENCES

1. AlaaEhabSakran ,SherifMahdyAbdou , Salah Eldeen Hamid and Mohsen Rashwan2017. A Review: Automatic Speech Segmentation. International Journal of Computer Science and Mobile Computing, 6(4): 308 – 315.
2. Rabiner, L. R. and B. H. Juang1993. Fundamentals of Speech Recognition. Englewood Cliffs, NJ: Prentice-Hall.

3. V. Kamakshi Prasad, T. Nagarajan and Hema A. Murthy 2004. Automatic segmentation of continuous speech using minimum phase group delay functions. Speech Communication,429 – 446.
4. D.S.Shete and S.B. Patil2014. Zero crossing rate and Energy of the Speech Signal of Devanagari Script.IOSR Journal of VLSI and Signal Processing (IOSR-JVSP), 4(1): 01-05.
5. Bharat C. Patel and Apurva A. Desai 2015. Segmentation of Gujarati words from Continuous spoken Gujarati speech signal.VNSGU Journal of Science and Technology, 4(1): 106 – 112.
6. TorbjomSvendsent and Frank K. Soong1987. On the Automatic Segmentation of Speech Signals.IEEE International Conference on Acoustics, Speech and Signal Processing.
7. Anupriya Sharma and Amanpreet Kaur2013. Automatic Segmentation of Punjabi Speech Signal using Group Delay. Global Journal of Computer Science and Technology, 13(12): 7-10.
8. Nipa Chowdhury, Md. AbdusSattar and Arup KantiBishwas2009. Separating Words from Continuous Bangla Speech. Global Journal of Computer Science and Technology 172-175.
9. Md. Mijanur Rahman and Md. Al-Amin Bhuiyan2013. Dynamic Thresholding on Speech Segmebtation. International Journal of Research in Engineering and Technology,2(9): 404-41.
10. AgnelWaghela, Rohan Reddy, ShivangiRai, Aditya Pawar and NamrataGharat2014. SUV Detection Algorithm for Speech Signals. International Journal of Advanced Research in Computer Science and Software Engineering, 4(4): 958-963.
11. T. Nagarajan, Hema A. Murthy and Rajesh M. Hegdem2003. Segmentation of speech into syllable-like units. EUROSPEECH, 2893-2896.
12. Md. Mijanur Rahman and Md. Al-Amin Bhuiyan2012. Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches.International Journal of Advanced Computer Science and Applications, 3(11):131-138.
13. Yin Win Chit and Dr. Renu2017. Fuzzy Logic Based Segmentation for Myanmar Continuous Speech Recognition System. American Scientific Research Journal for Engineering, Technology, and Sciences (ASRJETS), 31(1): 183-190.
14. Ghulam Muhammad2011. Extended Average Magnitude Difference Function Based Pitch Detection.The International Arab Journal of Information Technology, 8(2): 197-203.
15. Cooley, James W. and Tukey, John W. 1965.An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation: Journal Review, 19: 297–301.
16. T Giannakopoulos2009. Study and application of acoustic information for the detection of harmful content and fusion with visual information. Ph.D. dissertation, Dept. of Informatics and Telecommunications, University of Athens, Greece.
17. Bello J P, Daudet L, Abdallah S, Duxbury C, Davies M, and Sandler MB2005. A Tutorial on Onset Detection in Music Signals. IEEE Transactions on Speech and Audio Processing, 13(5): 1035–1047.