# Automatic Segmentation of Hindi Speech into Syllable-Like Units

Ruchika Kumari[1]
Department of ECE
Maharaja Surajmal Institute of Technology
GGSIPU,
New Delhi, India
Research Scholar
Indira Gandhi Delhi Technical University for Women
New Delhi, India

Amita Dev[2]
Vice-Chancellor
Indira Gandhi Delhi Technical University for Women
New Delhi, India

Ashwani Kumar[3]
Department of ECE
Indira Gandhi Delhi Technical University for Women
New Delhi, India

***Abstract*—To develop the high-quality Text-to-Speech (TTS) system, appropriate segmentation of continuous speech into the syllabic units placed an important role. The research work has been implemented for automatic syllable based speech segmentation technique for continuous speech for the Hindi language. The experiments were conducted by using the energy convex hull approach for clean, continuous speech for Hindi. In this method, the Savitzky-Golay filter was applied on the short term energy (STE) signal to increase the signal to noise ratio (SNR), followed by applying the median filter to preserve the boundaries, hence smoothing the energy curve. Also, the Hamming sliding-window was applied twice on speech signal to get the more accurate depth of convex hull valleys. Further, the algorithm was tested on 50 unique utterances chosen from the travel domain. The accuracy of the proposed algorithm has been calculated and obtains that 76.07% syllables have time-error less than 30 ms with manual segmentation reference. The performance of the proposed algorithm is also analyzed and gives better-segmented accuracy as compared to the existing group delay segmentation technique for fricatives or nasal sounds. The syllable base segmented database is suitable for the speech technology system for Hindi in the travel domain.**

***Keywords*—Database; short term energy; convex hull; speech segmentation; syllable***

## I. Introduction

Speech is considered as quasi-periodic signal since the characteristic of the signal changes over time. Segmentation is the process of splitting the speech signal into several parts. Speech can be segmented into various units, such as words, syllables, and phones. TTS is the ability of a machine to convert the given text in a language to spoken speech.

The accurate segmentation and label play a vital role in developing the TTS. The speech synthesis system makes use of various speech and language technology. It is being used to enhance human-machine interactions such as in mobile communication, screen reader, remote access to online information. The various application of speech synthesis includes talking aids, health care, banks, travel and tourism, visual and speech impairment, etc. Building a TTS for any language requires a corpus, which is a labor-intensive and time-consuming task. The research aim is to develop and analyze continuous speech segmentation as syllable like units for the Hindi language. Hindi is one of the official languages of India. It is a primary communication language for a large number of Indian populations and in other parts of the world. Most of the research has been done in other languages, such as European, English, Mandarin, Arabic, etc. However, less work has been done in the Hindi language due to a lack of standard database and pronunciation rule. As Hindi is syllable-centric in nature, the syllable is considered as an appropriate segment to a label. Several advance works have been reported to the phoneme level segmentation technique but still lacking on syllable base level.

The objective of the paper is to propose a time-domain automatic segmentation technique based on STE and convex hull approach for the Hindi language. Moreover, applied Savitzky-Golay filter [13] and median filter to get smoother energy curve and also apply Hamming sliding-window twice on STE to get a smoother curve and more profound valleys to make it easy to set the threshold boundary. The performance of resultant syllable units is calculated in terms of time duration, which is compared with the existing group delay and manual segmentation techniques.

The remaining paper is organized as follows: Section II describes the literature review. Section III describes the methods and procedures. Section IV explains the information about acoustic-phonetic features in Hindi. Section V describes the energy convex hull algorithm approach. Section VI gives experimentation based on the proposed algorithm. In Section VII, the result and time error analysis are discussed. Section VIII gives a subjective evaluation. Section IX describes the conclusion of the paper.

## II. Literature Review

The accurate segmentation of speech is an essential factor in creating a high quality of TTS. Zhao and O'Shaughnessy [1] implemented algorithms of the convex hull in speech segmentation. Similarly, Ling and colleagues [2] used speech

segmentation to cleft palate speech of the Mandarin language using a convex hull. They initially extracted syllables from the speech utterances and classified as "quasi-unvoiced" or "quasi-voiced" and estimated the segmentation accuracy, which came out to be high. K. Prasad et al. [3] and Hema A Murthy [4] have performed an algorithm based on short-term energy and group delay processing of the magnitude spectrum for determining segmented syllable boundaries for the Indian languages and TIMIT database. Panda and Nayak [5] carried out successful automated speech segmentation of Hindi, Bengali, and Odia languages using vowel offset point identification technique along with Zero Crossing Rate (ZCR) segmentation method with the manual segmentation approach. Similarly, Stan et al. [6] used an ALISA tool to segment sentence-level alignment of speech with imperfect transcripts. This method helped in the creation of a new speech corpora. This method found that utilizing the speech segmentation tools and transcribing speech data is reduced. Hamza Frihia and Halima Bahi [7] reported the Hidden Markov Model (HMM) and support vector machine (SVM) model to generate the phoneme-based speech segmentation for the Arabic language for application of speech recognition. Sandrine Brognaux and Thomas Drugman [8] presented the HMM algorithm speech segmentation on the phone level for English, French, or under-rescore Language. Jon Ander G´omez and Marcos Calvo [9] shown the segmentation technique with a combination of HMM and DTW (Dynamic Time Wrapping) to achieved phone boundaries on the Albayzin and TIMIT database. Asaf Rendel et al. [10] shown that the HMM-GMM modeling technique is applied to the TIMIT corpus to get phoneme speech segmentation, and SVM is used to refine the obtained phone boundaries. The accuracy of the above modeling technique is 96%. Fréjus A. A. Laleye [11] published the algorithm based on STE & Zero crossing rate (ZCR) and perform the machining phase using the set of Fuzzy rules to get the syllable and phone boundaries on Fongbe language spoken in Benin, Tago, and Nigeria. Balyan et al. [12] built a medium-sized database for passenger rail information systems for the Hindi language in the phoneme level using HMM. The database consists of 630 utterances with 12674 words to facilitate the researcher in TTS and automatic speech recognition (ASR). Arum Boby et al. [21] presented the speech segmentation for Indian language consider as a phone level by using deep neural network (DNN) and convolutional neural network (CNN) framework. Md. Mijanur Rahman and Md Al-Amin Bhuiyan have created the database on time and frequency domain approach on word level and achieve a segmentation accuracy rate of 96.25 for Bangla Language [22]. Yahia Hasan Jazyah [23] has reported the segmentation of audio data such as human speech in both English and Arabic languages by using Dynamic Windows and Thresholds. The algorithm achieved a segmentation accuracy rate up to 91.6% in average for English and 89.0% for the Arabic language.

## III. Methods and Procedures

The following steps are carried out to design a Speech corpus.

- Selection of text sentences from news domains
- Recording of the selected text
- Syllabification of the speech signal

### A. Selection of Sentences

The selection of the 150 sentences has been manually selected from various sources relevant to Metro travel information announcements in Delhi Rail for building the speech synthesis system. Adequate care has been taken to include all types of the required information so that the recording has enough occurrence of each type of Hindi sound [14].

### B. Recording of Speech Corpus

The steps followed for recording the speech wav files were as follows:

- Professional male speaker voice has been recorded to maintain constant pitch and prevent stress phenomenon in noise and echo-free studio.
- The speaker has clear pronunciation and no articulacy defect.
- The sampling frequency was set to 16 kHz store in 16-bit PCM with Mono mode type.
- The speaker is required to read each text sentence, and the recorded sample was saved as wav files.

## IV. Acoustic- Phonetic Features in Hindi

The acoustic-phonetic of Hindi differs from the European languages. Hindi is mostly phonetic in nature, i.e., there is one to one correspondence between written symbols and the spoken sentences. Hindi phonemes can be divided into vowels and Consonants. The Hindi alphabet consists of 10 pure vowels (/ə/, /ɑ/,/i/, /I/, /u/, /U/, /æ//e/,/o/, /ɔ:/) including two diphthongs namely; /æ/ and /ɔ:/.All these vowels have their nasalized form also. Creaky and whispered vowels are rarely used [15]. The Hindi consonants consist of 4 semivowels, 4 fricatives, and 25 stop consonants (including 5 nasals). The stop consonants are ordered systematically in the Hindi language, and this order may suggest ideas for developing a recognition/synthesis system [17, 18]. Classification of Hindi consonants and vowels are presented in Table I.

TABLE I. Description of Hindi Phoneme

| Shorts Vowels | | | Long vowels | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| अ | इ | उ | ए | ओ | आ | ई | ऊ | ऐ | औ |
| Unvoiced | | | | | Voiced | | | | Nasals |
| Unaspirated | | Aspirated | | | Unaspirated | | | Aspirated | |
| क | | ख | | | ग | | | घ | **ण** |
| च | | छ | | | ज | | | झ | **ङ** |
| ट | | ठ | | | ड | | | ढ | **ञ** |
| त | | थ | | | द | | | ध | **न** |
| प | | फ | | | ब | | | भ | **म** |
| | | | | | | | | | |
| Semi Vowel | | | | | Fricatives | | | | |
| य | र | ल | व | | श | | ष | स | ह |

## V. Syllable base Segmentation Algorithm

The syllables are identified from the speech database. The fundamental of the database is multiple forms of the unit phoneme, syllable, and words. In the Hindi language, the syllable types are *CV*, CVC, VC, V, CCV, and CCVC [14, 16]. The database distribution of syllables is mentioned below in Table II.

The | syllable likes boundary identification is performed by using an energy convex hull approach. The steps are as follows:

- Let's x(t) is the represented continuous speech signal, and $x[n]$ be digitized speech signal.
- Determine the Short-term energy (STE) by applying the overlapped Hamming window (N= 400).

$$Q(n) = \sum_{m=-\infty}^{\infty} [X(m)]^2 w(n-m)$$

$$E(n) = 10 * logQ(n)$$

$$W(n) = .54 - .46cos\left(\frac{2\pi n}{N-1}\right); 0 \leq n < N$$

- Apply the Savitzky-Golay smoothing filter and Median filter to reduce the noise and preserve boundaries.
- Estimate the initial syllable decision threshold for initial syllable detection.
- Apply the Hamming window for refining the boundaries of syllable- like units.

$$D(n) = 10 * \log[Q(n) + 1]$$

$$p(n) = \sum_{m=-\infty}^{\infty} D(m)w(n-m)$$

- Reset the threshold on $p(n)$ to obtain correct syllable boundary

The block diagram in Fig. 1 shows the steps involved to obtain of syllable – like segmented speech.

TABLE II. Distribution of Various Syllable in Hindi

| Syllables | Relative Frequency (%age) |
|---|---|
| CV | 69.69 |
| CVC | 22.00 |
| VC | 2.78 |
| V | 3.60 |
| CVCC | 1.18 |
| CCVC | 0.89 |
| CCV | 0.48 |

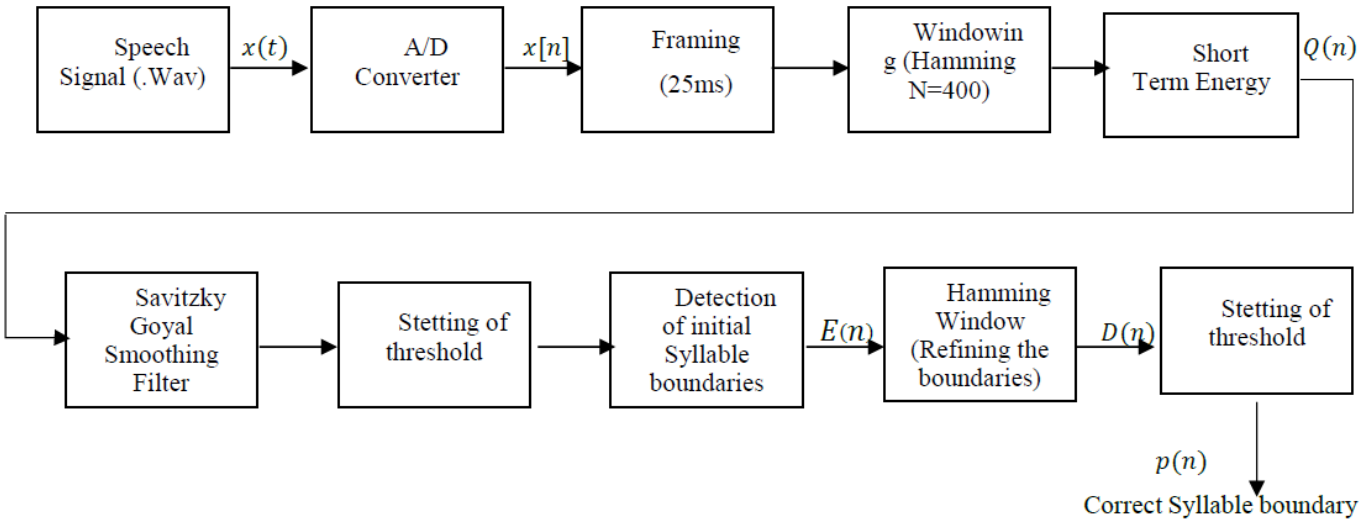


Fig. 1. Block Diagram Showing Steps Taken in Finding Syllable Boundaries.

## VI. Experimentation

The experiment is done on the word and sentence level of medium size database consisting of 150 sentences of the duration of approx. 45 mins spoken by a single male speaker and obtained 1175 syllables units.

The 50 sentences of a syllable are processed manually by using PRAAT [19] speech analysis to check the performance of the proposed techniques. Fig. 2 shows the manually segmented output of the input wav file "Yahhan line do ke liye badle". This input wav file consists of 9 syllable units.

### A. Initial Boundary Detection

On STE Q (n), the Savitzky–Golay [12] filter is applied for signal smoothing, and the SNR ratio is improved. Further, the median filter is used to preserve the boundaries and the smoothing energy curve.

To detect the initial boundary, a threshold is required to be estimated in the short term energy curve. To get the threshold in training set in the average STE of utterance was calculated.

However, the threshold can't be set to this value. For example, in Fig. 3, the utterance contained five possible syllable boundaries points A to E when the energy threshold was set to the average STE curve of a speech signal. The threshold value is -17 dB. If the threshold was kept higher than -17 dB, more valley points might be obtained, which are incorrect. If the threshold were kept lower, then the valley points E and C would be removed. The threshold value was reset from -17 dB to -32 dB to obtain the correct boundaries based on the above observation. After experimentation with a Hindi training set, it was seen that the threshold value between -28 dB to -38 dB gives more accurate segmentation boundaries.

### B. Convex Hull Boundary Detection Analysis

In this approach, a sliding Hamming window is applied on Q(n) shown in Fig. 4 to obtain P(n). It is seen, the STE curve is smoother deeper valley is obtained, which makes it easy to set convex- hull threshold value.

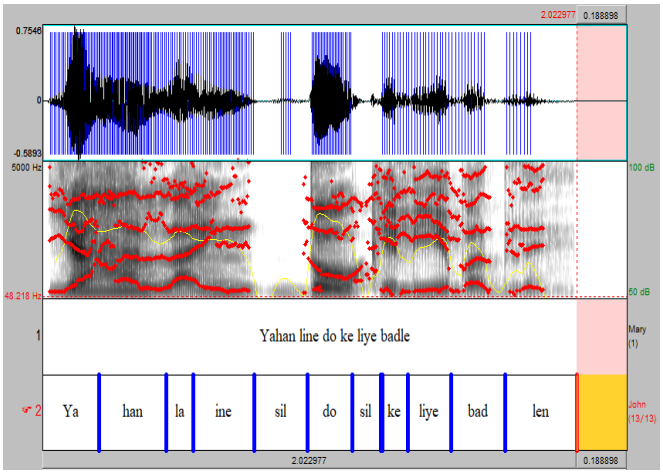


Fig. 2. Manual Segmentation of Continuous Speech at the Syllable Level.

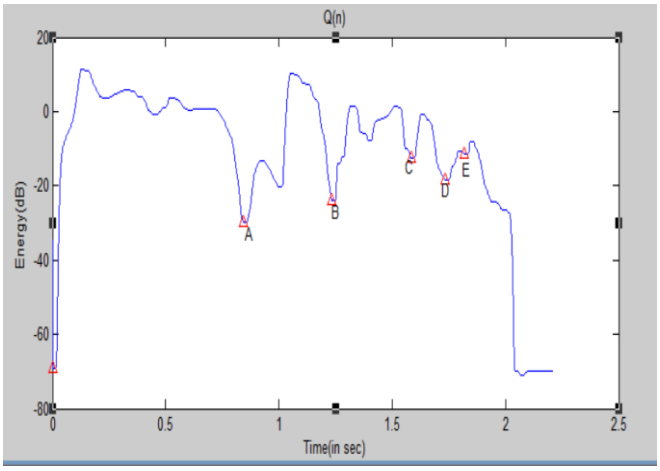


Fig. 3. STE Curve Syllable Points in an utterance.

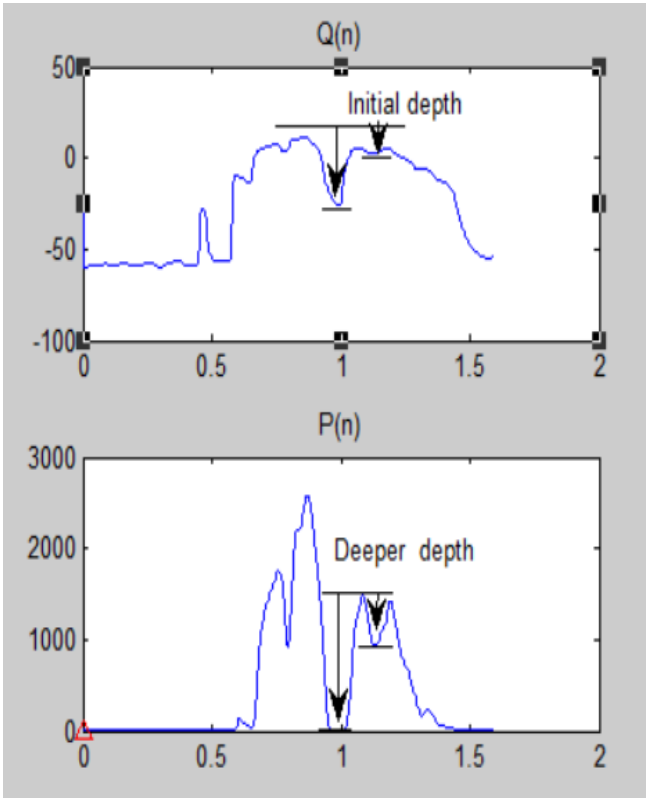


Fig. 4. Comparison of the Valley of the Energy Curve and Convex Hull Curve.

Fig. 5 shows the output of the segmentation algorithm for the input speech utterance "यहाँ लाइन दो के लिए बदले" ("yahan line do ke liye badlen"). It is seen that the input speech signal is segmented into three initial syllable units "यहाँ लाइन", "दो" and "के लिए बदले" ("yahan line", "do" and "ke liye badlen"). On the application of the convex hull approach, the speech is re-segmented into nine syllables units. "य", "हाँ", "ला", "इन", "दो", "के", "लिए", "बद" and "ले" ("ya", "han", "la", "ine", "do", " ke", "liye", "bad" and "len"). The same process has been applied for 50 utterances and obtained 402 syllables like units.

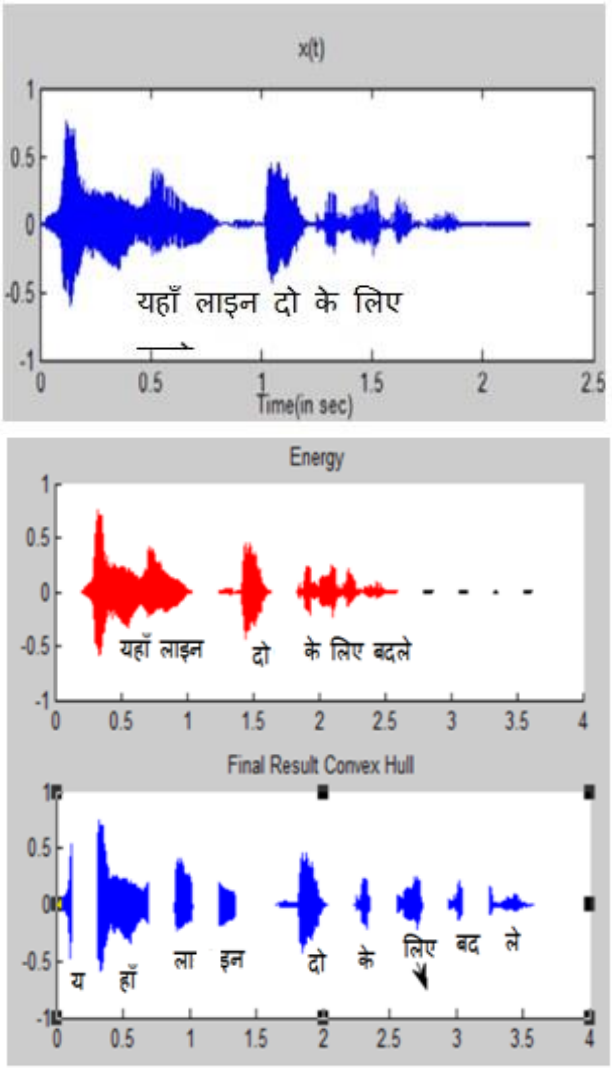


Fig. 5. The Waveform of Input Speech x (t) and Segmented Output Syllable units of STE and Convex Hull.

Below examples are shown in Table III to obtain as syllable boundaries units for a few input wav files.

TABLE III. EXAMPLES ILLUSTRATING SYLLABLES SEGMENTS

| Input Wav file | Obtained Syllable output |
|---|---|
| सफदरजंग | सफ् दर् जंग् |
| सेवा में नहीं | से वा मे न ही |
| कृपया दरवाजो से दूर हट कर खड़े हो | कृप् या दर् वा जो से दूर् हट् कर् ख डे हो |
| लाजपत नगर | लाज् पत् न गर् |

## VII. RESULT

The performance of the segmentation algorithm is analyzed on a set of 50 test samples. Time error analysis is calculated to test the accuracy of the segmented syllable-likes unit for each syllable. The research also includes silence occurrence in the sentence as discuss:

$$\text{Time\_Error} = |duration\ of\ manual\ segment\ boundary - duration\ of\ the\ automatic\ segment\ boundary|$$

Table IV shows the result of the segmented output and the calculated error rate of the proposed algorithm & existing group delay technique [20]. The error rate obtained in the energy convex hull algorithm performs better as it has a lower value.

Experiments performed in Fig. 6 demonstrate in the graph that the energy convex hull segmentation technique achieves better results that are closer to the outcome achieved by manual segmentation. But, the group delay based method shows a high degree of variation in syllable durations compared to the energy convex hull approach.

The same process has applied a set of words and sentences to find overall performance segmented syllable like units of continuous speech by using proposed and group delay segmentation techniques.

The performance results are shown in Table V and found that the group delay-based algorithm approach shows an accuracy rate of 63.05%. The proposed algorithm energy convex hull approach achieves an accuracy rate of 76.12% of segmented speech in less than 30 ms.

In the proposed algorithm, the final segmentation result is obtained after applying the double sliding widow along with the reset of the threshold value. After analysis, it is observed that if the threshold is set between 2200-2800 for Hindi speech, it gives an accurate syllable boundary. During the experiment, it was found that the duration of time error was higher for fricative and nasal sound, but it provided better results as compared to group-delay segmentation. The threshold value for fricative sound {e.g., shakur basti (शाकुर बस्ती), safdarjung (सफदरजंग), udghoshnaa (उदघोषना), Station (स्टेशन), Shalimar (शालीमार), etc.} is set at approx. 2600 to 2700 as these sounds are high energy signals. For nasal sound (e.g., mangolpuri (मंगगोलपुरी), nagar (नगर), anand (आनंद), nirmal (निर्मल), etc.) the threshold is set at approx. 2300 to 2400.

TABLE IV. DURATION OF SEGMENTED OUTPUT BY USING MANUAL SEGMENTATION (PRAAT TOOL), GROUP DELAY ALGORITHM AND ENERGY CONVEX HULL ALGORITHM

| Obtained syllable units | Duration of manual segmentation (sec) | Duration of group delay algorithm (sec) | Error rate (msec) | Duration of energy convex hull algorithm (sec) | Error rate (msec) |
|---|---|---|---|---|---|
| य (***ya***) | 0.19 | 0.12 | 44.00 | 0.15 | 21.05 |
| हाँ (***han***) | 0.18 | 0.20 | 11.54 | 0.18 | 2.25 |
| ला (***la***) | 0.21 | 0.243 | 16.48 | 0.22 | 4.35 |
| इन (***ine***) | 0.33 | 0.35 | 5.06 | 0.34 | 1.20 |
| दो (***do***) | 0.34 | 0.30 | 10.19 | 0.32 | 3.57 |
| के (***ke***) | 0.11 | 0.12 | 6.25 | 0.16 | 45.45 |
| लिए (***liye***) | 0.18 | 0.20 | 10.53 | 0.19 | 5.56 |
| बद (***bad***) | 0.21 | 0.15 | 39.52 | 0.17 | 21.23 |
| ले (***len***) | 0.16 | 0.15 | 4.35 | 0.16 | 1.26 |

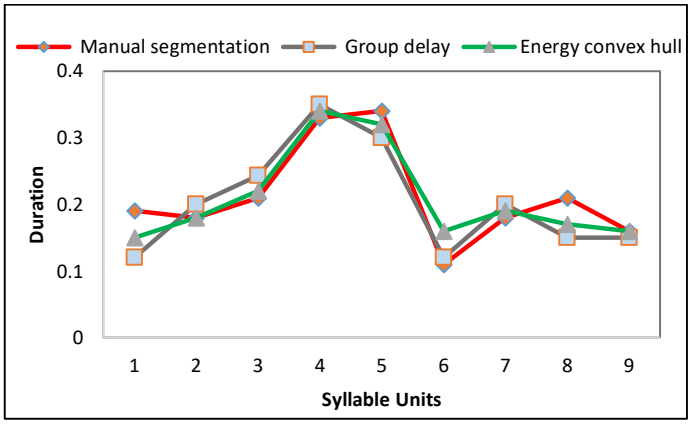


Fig. 6. Duration of Syllable units Obtained by manual Analysis and Segmentation Algorithm.

TABLE V. TIME ERROR ANALYSIS OF OVERALL SEGMENTATION CONTINUOUS SPEECH

| Algorithm | Time–error (msec) | ≤ 30 | 31-40 | 41-50 | > 50 | Total no. of segments |
|---|---|---|---|---|---|---|
| Purposed algorithm | Number of segments | 306 | 18 | 16 | 62 | 402 |
| | Performance (in %age) | 76.12 | 4.47 | 3.98 | 15.42 | |
| Group Delay | Number of segments | 275 | 51 | 28 | 82 | |
| | Performance (in %age) | 63.07 | 11.69 | 6.40 | 18.08 | |

## VIII. SUBJECTIVE EVALUATION

Accuracy is an essential factor in measuring the performance of segmented speech. In this work, five subjects were considering for perception evaluation of segmented speech. Subjects were asked to access the accuracy on a 5 points scale (1-Unsatisfactory, 2-Poor, 3-Fair, 4-Good, and 5-Excellent) for each of the segmented sentences. The test is carried out for the segmented sentences generated by group delay and energy convex hull approach. The mean opinion score (MOS) is calculated for the accuracy of segmented speech. Table VI shows that the segmented accuracy rate is improved in the convex hull approach.

TABLE VI. MEAN OPINION SCORE FOR THE QUALITY OF SEGMENTED CONTINUOUS SPEECH

| **Algorithm** | **No of Test samples** | **Accuracy rate** |
|---|---|---|
| Energy Convex hull | 50 | 4.18 |
| Group Delay | 50 | 4.02 |

## IX. CONCLUSION

In this paper, the energy convex hull algorithm is proposed for segmenting the speech signal into syllable-like units for improving the segmentation performance. The algorithm is applied to speech corpus, and segmented syllabic units are obtained. The algorithm calculated the time duration of each syllable unit and obtained a time error rate about manual segmentation syllable units to validate the accuracy of the proposed algorithm. After a comprehensive analysis, it is found that the segmented boundary errors are ≤ 30 ms for 76.07% of the total syllables. The performance of the algorithm gives an accurate result as compared to the existing group delay segmentation technique. Hence the proposed algorithm is highly useful to create syllable like speech units as it takes a few milliseconds to obtain syllabic units over manually labelling process of speech segmentation, which is a very time-consuming and strenuous task.

This algorithm may also be extended over large databases for building the high quality of TTS by the researcher for the limited and unlimited domain. Further, the research may be extended to reduce errors by applying various optimization techniques - machine learning (DNN, CNN, or hybrid models) and fuzzy-based algorithms.

## REFERENCES

[1] X. Zhao, and D. O'Shaughnessy, "A New hybrid approach for automatic speech signal segmentation using silence signal detection, energy convex hull, and spectral variation," IEEE International Conference, pp. 000145-000148, 2008.

[2] J. Li, and F. Shen, "Automatic segmentation of Chinese Mandarin speech into syllable-like," Asian Language Processing (IALP) 2015 International Conference, pp. 57-60, 2015.

[3] K. Prasad, T. Nagarajan, and H. A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay function," Speech Communication Vol. 42, pp. 1883-6, 2004.

[4] H. A Murthy, and B. Yegnanarayana, "Group delay functions and its applications in speech technology," Indian Academy of Sciences, pp. 745–782, 2011.

[5] S. P. Panda, and A. K. Nayak, "Automatic speech segmentation in syllable centric speech recognition," International Journal of speech technology, pp. 9-18, 2015.

[6] A. Stan, Y. Mamiya, J. Yamagishi, P. Bell, O. Watts, R. A. J. Clark, and S. King, "ALISA: An automatic lightly supervised speech segmentation and alignment tool," Computer Speech & Language, 35, pp. 116 – 133, 2016.

[7] H. Frihia, and H. Bahi, "HMM/SVM segmentation and labelling of Arabic speech for speech recognition application," International Journal of Speech Technology 20(3), pp-563-573, 2017.

[8] S. Brognaux, and T. Drugman, "HMM-based speech segmentation: Improvements of fully automatic approaches," IEEE/ ACM Transactions on Audio, Speech, and Language Processing, 24(1), pp. 5–15, 2016.

[9] J. A. G´omez, and M. Calvo, "Improvements on Automatic Speech Segmentation at the Phonetic Level," Springer-Verlag Berlin Heidelberg, pp. 556-564, 2011.

[10] A. Rendel, A. Sorin, R. Hoory and, A. Breen, "Towards Automatic Phonetic Segmentation for TTS," International Conference on Acoustics, Speech and Signal Processing, pp. 4533-4536, 2012.

[11] F. A. A. Laleye, E. C. Ezin, and C. Motamed, "Fuzzy-based algorithm for Fongbe continuous speech segmentation," Pattern Analysis and Application 20, pp. 855–864, 2017.

[12] A. Balyan, S. S. Agrawal, and A. Dev, "Automatic phonetic segmentation of Hindi speech using hidden Markov model," AI & Soc Springer-Verlag London Limited, pp. 543–549, 2012.

[13] A. Savitky and M. J. E. Goyal, "Soothing and differentiation of data by simplified least square procedure," Anal Chem., vol. 36, no.9, pp. 1627-1639, 1964.

[14] A. Balyan, S. S. Agrawal, A. Dev, "Building Syllable dominated Speech corpora for Metro Rail Information system," International Conference of O-COCOSDA-2008, pp. 135-140, Kyoto, Japan, Nov 25-27, 2008.

[15] K. Arora, Sunita, K. Verma, and S. S. Agrawal, "Automatic extraction of phonetically rich sentences from large Text corpus of Indian Languages," In INTERSPEECH, pp. 2885-2888, 2004.

[16] S. S. Agrawal, “Emotions in Hindi speech-analysis, perception and recognition," International Conference on Speech Database and Assessments (Oriental COCOSDA), pp.7-13, 2011.

[17] B. Aarti1, and S. K. Kopparapu, “Spoken Indian language identification: a review of features and databases,” Indian Academy of Sciences, pp. 1-14, 2018.

[18] S. Bhatt, A. Dev, and A. Jain, “Confusion analysis in phoneme based speech recognition in Hindi,” Journal of Ambient Intelligence and Humanized Computing, DOI:10.1007/s12652-020-01703-x, 2020.

[19] P. Boersma, and D. Weenik, Praat: A System for Doing Phonetics by Computer. http://www.praat.org/ , 2001.

[20] A. Balyan, A. Dev, R. Kumari, and S. S. Agrawal, “Labelling of Hindi Speech,” IETE Journal of Research, vol 62, issue 2, pp. 146-153, 2015.

[21] A. Baby, J. J. Prakash, and H. A. Murthy, “ A Hybrid approach to Neural Networks Based Speech Segmentation,” International Symposium Frontiers of Research on Speech and Music, 15-16, National Institute of Technology (NIT) Rourkela, 2017.

[22] Md. M. Rahman, and Md. A. Bhuiyan, “Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches,” International Journal of Advanced Computer Science and Applications, Vol. 3, No. 11, pp. 131-138, 2012.

[23] Y. H. Jazyah, “Speech segmentation using dynamic window and thresholds or Arabic and English Language,” Journal of Computer Science, Volume 14, Issue 4, pp. 485-490, 2018.