ISSN: 0976-9102 (ONLINE)
DOI: 10.21917/ijivp.2015.0152

WORD BASED TAMIL SPEECH RECOGNITION USING TEMPORAL FEATURE BASED SEGMENTATION

A. Akila[1] and E. Chandra[2]
[1]*D.J. Academy for Managerial Excellence, India*
E-mail: akila.ganesh.a@gmail.com
[2]*Department of Computer Science, Bharathiar University, India*
E-mail: crcspeech@gmail.com

Abstract

Speech recognition system requires segmentation of speech waveform into fundamental acoustic units. Segmentation is a process of decomposing the speech signal into smaller units. Speech segmentation could be done using wavelet, fuzzy methods, Artificial Neural Networks and Hidden Markov Model. Speech segmentation is a process of breaking continuous stream of sound into some basic units like words, phonemes or syllable that could be recognized. Segmentation could be used to distinguish different types of audio signals from large amount of audio data, often referred as audio classification. The speech segmentation can be divided into two categories based on whether the algorithm uses previous knowledge of data to process the speech. The categories are blind segmentation and aided segmentation.

The major issues with the connected speech recognition algorithms were the vocabulary size will be larger with variation in the combination of words in the connected speech and the complexity of the algorithm is more to find the best match for the given test pattern. To overcome these issues, the connected speech has to be segmented into words using the attributes of speech. A methodology using the temporal feature Short Term Energy was proposed and compared with an existing algorithm called Dynamic Thresholding segmentation algorithm which uses spectrogram image of the connected speech for segmentation.

Keywords:

Short Term Energy, Missed Detection Percentage, Deviation Percentage, Dynamic Thresholding Segmentation, Temporal Feature based Segmentation

1. INTRODUCTION

Identifying the boundaries between words in a connected speech is the process of segmentation in connected speech. The term applies both to the mental processes used by humans, and to artificial processes of natural language processing. Manual segmentation (hand labeling) and Automatic Segmentation (AS) are the two ways of segmentation.

Segmentation of speech usually performed manually by phoneticians, but it is extremely time consuming and costly [1]. When large amount of segmented and labeled speech data is required, manual segmentation would be a tiresome job. Due to human variability of visual and perceptual capabilities, there will be always disagreement among skilled human labeling experts in their results of labeling the same waveform. In order to speed up the task and to avoid the disagreement, researchers were trying to develop automatic labeling methods. The detection of boundary of the words using an algorithm by the machine is known as automatic segmentation [2].

1.1 RELATED STUDY

The Tamil language has 247 characters [3] which constitutes of 12 vowels, 18 consonants, 216 vowel consonants and 1 special character. The language also has 6 grantha letters derived from Sanskrit. Tamil is a strong grammatical language which has large set of rules. The consonants could be categorized as stop, fricatives, nasal, lateral and glides. The categories and the Tamil consonants of each category are shown in Table.1.

Table.1. Category of Tamil Consonants and Grantha Characters

Sl. No.	**Consonants Category**		**Tamil Characters**
1	Stop/Plosives		க், ச், ட், த், ப்
2	Nasal		ந், ம், ன், ண், ங், ஞ்
3	Lateral	Alveolar	ல்
		Retroflex	ள்
4	Fricatives	Flap	ர்
		Trill	ற்
		Retroflex Frictionless Constituent	ழ்
5	Glides		ய், வ்
6	Grantha letters		ஜ், ஷ், ஸ், ஹ், ஸ்ரீ, க்ஷ்

Some of the consonants shown in Table.1 cannot appear as the first lexeme of any word in Tamil. The nasal consonants ன், ண், ங் cannot be first lexeme of a word [4]. Similarly the fricatives ற், ழ் and lateral ள் cannot appear at the initial positions of a word. The vowel consonants derived from these consonants also would not appear as the first letter of any Tamil word. The nasals ந், ங், ஞ் cannot come as the last letter of a word in Tamil language [5] which was shown in Table.2. All the Tamil words follow these grammatical rules. The segmentation of Tamil connected words could be done using acoustic cues which uses these grammatical rules and the attributes which are discussed in the next section. The algorithms that use any of these cues for segmentation are aided segmentation algorithms.

Table.2. Tamil Graphemes that cannot be present in word beginning and word ending

Sl. No.	**Characters that cannot be present in word beginning**		**Characters that cannot be present in word ending**	
1	Nasal	ன், ண், ங்	Nasal	ந், ங், ஞ்
2	Fricatives	ற், ழ்		
3	Lateral	ள்		

Connected speech segmentation is a process to identify the words in the speech using the knowledge of the language used in the recognition system. Identification of the words in the absence of systematic cues to word boundaries which differ for every language is the segmentation problem [11].

The segmentation of connected speech into isolated word has been performed using statistical approaches like Brandt's Generalized Likelihood Ratio (GLR) and Divergence Algorithm [12]. In the segmentation algorithms, segmentation points were identified by finding the discontinuities in the speech signal. The word boundaries of connected words can be identified using the phonological rules. The phonological rules were formed for each phoneme categories by identifying the frequency of occurrence of phonemes in the word beginning and in the word ending [13].

The prosodic information like pitch and duration in a speech could be used to segment the connected speech into isolated words [11]. Eventhough word segmentation could be done using these procedures, the segmentation errors like false boundary identification, omission of boundary, insertion of boundary were possible [12].

The Prosodic cues like rhythmic properties, phrase boundary cues, statistical regularities, Transitional Probabilities (TP) between syllables or diphones [6] were used by people for segmentation [7]. Even stress could be used to locate the word edges. Duration, pitch and intensity were used to differentiate the stressed and unstressed syllables. The co-occurrence statistics which defines the similarity occurrence of syllables or phones in certain order could be used to perform the segmentation [8] [9].

The intonation and speech style variability has been used in segmentation to improve the performance of the connected speech recognition system. The rise in pitch at the beginning of each accentual phrase known as Accentual Phrase Rise (APR) could be used for identifying the word boundary as accentual phrases might be present at the beginning of a word [10]. These methodologies use the cues for segmentation. The proposed segmentation algorithm uses the blind segmentation method.

2. MAJOR ATTRIBUTES FOR CONNECTED SPEECH SEGMENTATION

2.1 INTONATION

Intonation or the melody of speech plays a central role in human communication. It is one of many elements in the speech signal that needs to be decoded to translate speech sound into meaning, but its role in speech understanding is crucial, since it can give us immediate cues to the start of a new word or phrase in the speech stream, and to the meaning of utterances. As a consequence, when the intonation is wrong, communication often breaks down.

At the acoustic level, intonation carries information about the physical features of the variability of fundamental frequency, statistical distributions, short and long disturbances, relation to segmental features on physical level.

The presence of intonational boundary could be found using the feature vectors which consist of acoustic information like pitch, duration and intensity [14]. The boundary could be manipulated using the decreased standard deviation of pitch.

2.2 SPEECH STYLE VARIABILITY

Speakers were being capable of changing the way they speak, when necessary for successful communication. The speaker can vary between speaking rapidly or slowly, or between quietly and loudly in a way that is appropriate to the communication situation. The duration of the speech may vary with the different speech styles like read speech, connected speech and spontaneous speech [15].

2.3 PHONOLOGICAL PROBABILITY

Phonological probability refers to the frequency with which a phonological segment and a sequence of phonological segments occur on a given position in a word. The knowledge of phonotactic constraints derived from the participant's own language may be recruited to segment the speech. Phonological probability influences the production of the spoken language. The tool Phonotactic Probability Calculator calculates the probability of each phoneme in the word. The word which has length up to 17 phonemes could be given as input word to the tool [16]. The position specific probability of each segment, the position specific biphone probability of each biphone, sum of all phoneme probability and sum of all biphone probability were given as output for each word given to the calculator tool. With these probability values, the highest possible phoneme can be identified to manipulate the word boundary.

2.4 ACOUSTIC CUES

The boundary could be identified using the acoustic cues got from various features extracted from the speech. The acoustic cues like signal energy, fundamental frequency, periodicity, extracted formant contours and spectral characteristics could be used to manipulate the boundary of a word [17]. The signal intensity whose slope peaks near the boundaries, pitch analysis in the frequency range 75 Hz to 600 Hz to identify the voiced regions and thereby identifying the boundary were also some of the acoustic cues used for boundary identification [17].

2.5 RHYTHMIC CHARACTERISTIC

The rhythm of speech is a subjective impression which is presumably derived from acoustic properties. The rhythmic measures like percentage of vocalic intervals, standard deviation of vocalic intervals and variability index of syllable could be used for segmentation [18].

3. SEGMENTATION USING DYNAMIC THRESHOLDING

The connected word could be segmented using the methods like K-Mean clustering, Fuzzy C-Means Clustering and Ostu Thresholding method. The Ostu method has been used in image segmentation. The approaches like blocking black area and boundary detection were used to detect the boundaries of the connected speech. The Dynamic Thresholding Segmentation (DTS) method was implemented for continuous speech [19]. The continuous speech was segmented into words as words have well defined acoustic representation. The method was implemented for connected speech in this paper work.

3.1 SPEECH SPECTROGRAM

The spectrogram of the given speech was formed using the Fast Fourier Transform (FFT) of the signal [20]. The boundary of each word in the connected speech was found using the spectrogram image stored. Initially the discrete Fourier transform [21] of the sample data of the signal was found. The FFT was found using the Eq.(1).

$$X = \sum_{n=1}^{N} x * e^{\frac{-j2\pi(n-1)}{N}} \qquad (1)$$

where, N is length of sample data, X is the FFT of the given signal and x is the sample data of the signal.

The spectrogram would be stored as a Tagged Image File Format (TIFF) image. The image was used to further manipulation in the Dynamic Thresholding Segmentation method.

3.2 FORMATION OF GRAYSCALE INTENSITY IMAGE

The image was read using the MATLAB function imread and the array containing the image data has been returned.

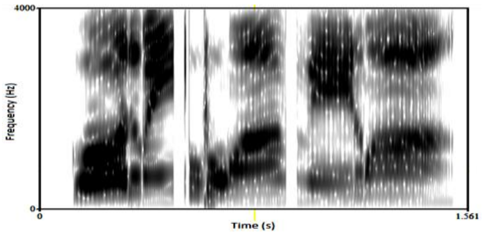

Fig.1. Gray scale Intensity Image of a spectrogram

The array of image data represents the Red Green Blue (RGB) image. The RGB image has to be converted to gray scale image by eliminating hue and saturation information while retaining the luminance. The rgb2gray function in MATLAB has been used to convert true color image to gray scale intensity image as shown in the Fig.1.

3.3 MANIPULATION OF THRESHOLD VALUE

The Threshold value for converting the grayscale image to binary image has been manipulated using the K-Means Clustering algorithm [22]. The average value of the K cluster centroids was assigned as the threshold.

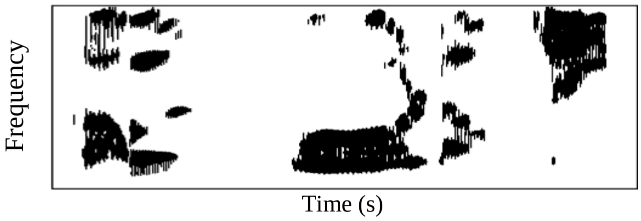

Fig.2. Thresholded Spectrogram Image of the speech signal

The K-Means clustering is a method of vector quantization where the similar data were grouped into K clusters. The clusters were formed using the minimum mean from the cluster centroids. The threshold value was used to convert the grayscale image into binary image [19]. The binary image is shown in Fig.2.

3.4 IDENTIFICATION OF VOICED AND SILENCE AREA

The voiced and silence regions of the speech were identified by blocking the black area of the threshold spectrogram image. The Intensity value of thresholded spectrogram image was represented as an array. The column wise intensity values were used to block the black area in the image [19]. The sum of each column intensity value was found and if the value was below half of the number of rows in the array, then all the values of that column were replaced by all white value (1). Similarly if the value was above half of the number of rows in the array, then all the values of that column were replaced by all black (0) value. The blocked image of the speech signal is shown in Fig.3. The black region specifies the voiced part and the white region specifies the silence part of the connected speech signal.

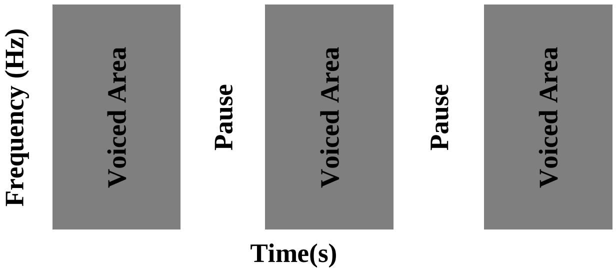

Fig.3. Image after blocking the black area of the speech signal

3.5 WORD BOUNDARY COMPUTATION

The left and right boundary of each word in the connected speech was found using the blocked image. The first column with the black area after the column with white area was considered as the left boundary of the word and the column with the black area before the white area was considered as the right boundary of the word.

4. PROPOSED METHODOLOGY – WORD BASED TAMIL SPEECH RECOGNITION USING TEMPORAL FEATURE BASED SEGMENTATION

The speech input would be a connected speech which was initially segmented into words. The segmentation was done using the Short Term Energy (STE). The signal was segmented into frames and STE [22] of each frame was computed using the Eq.(2). The Frame with Short Term Energy less than a threshold value was found and that frame was the boundary of each word in the connected word speech.

$$E_{sqr} = \sum_{n=1}^{N} s(n)^2 \qquad (2)$$

where, N is the length of the sampled signal, $s(n)$ is the amplitude of the speech signal.

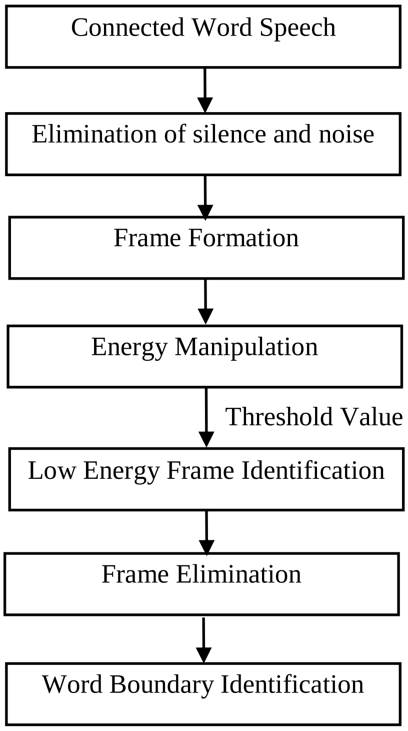
Connected Word Speech
Elimination of silence and noise
Frame Formation
Energy Manipulation
Threshold Value
Low Energy Frame Identification
Frame Elimination
Word Boundary Identification

Fig.4. Temporal Feature based Segmentation of Connected Speech

The voice activity detection was done to identify the end points. The signal was divided into n frames of size m. The frames whose energy is less than a threshold value were identified. The consecutive frames which have the short term energy less than threshold value were discarded and only the first frame of the consecutive frame was considered as the boundary. This process was used to split the connected speech into words. The Fig.4 shows the process of segmentation of connected speech into words in the Temporal Feature based Segmentation (TFS).

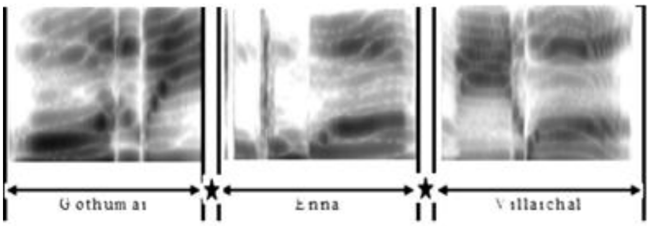

Fig.5. Spectrogram Representation of a Connected Speech

The Fig.5 represents the wideband spectrogram representation of a connected speech signal which constitutes of three words. Between each word, there exists a pause which was indicated by ‘P’ in the Fig.5. It could be clearly noticed from the figure that the spectrogram was free when the pause occurs in the signal. The Fig.6 represents the spectral representation of the connected speech of the same signal represented in the Fig.5. In the figure, the signal has zero amplitude which represents the pause in the speech signal.

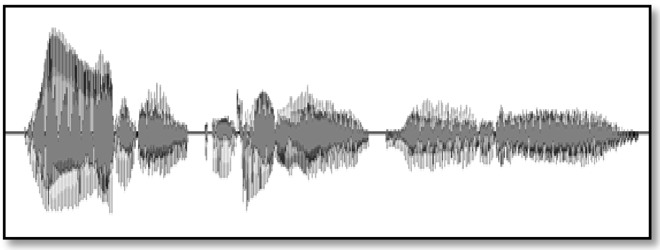

Fig.6. Spectral Waveform Representation of a Connected Speech

4.1 TFS PROCEDURE

Step 1: The speech was preprocessed using Endpoint Detection Algorithm using Mahalanobis Distance [23].

Step 2: The Short Term Energy (STE) for each frame using the Eq.(2).

Step 3: The frame with STE less than the threshold value was manipulated and the frame number of those frames were stored in a vector.

Step 4: The first frame of the consecutive frames was considered as the end boundary of the word.

Step 5: The next frame to the last frame of the consecutive frame was considered as the beginning of the next word.

Step 6: Repeat step 4 and step 5 till end of the frame of the speech signal.

4.2 WORD BOUNDARY IDENTIFICATION

4.2.1 Frame Formation:

The speech signal received from the user has been preprocessed for removing the silence signal at the left and right of the speech signal. Then the signal has been formed as frames of size 80 sample data. The frame size was chosen as 80 sample data because the speech was recorded in the speech rate 8 KHz. The number of frames was computed by dividing the length of the speech signal divided by the length of a single frame [24]. The last frame might be a unique frame of different size less than or equal to 80 sample data. The Short Term Energy (STE) of each frame was manipulated in the next step.

4.2.2 Energy Manipulation:

Short-term energy is the principal and most natural feature that has been used for segmentation of speech. Energy is a measure of how much signal there is at any one time. The short term energy (STE) is the energy of the speech signal at a particular instant of time. It differentiates between voiced, unvoiced and silence part of the speech. The short term energy is high for voiced speech, low for unvoiced speech and zero for silence. The threshold value could be set to segregate the speech signal. The values less than the threshold value were counted as zeros. The boundary of the word could be found using variation of the STE above the threshold value to STE below the threshold value. The short term energy was found using the Eq.(2).

4.2.3 Low Energy Frame Identification:

The frame which has short term energy less than a threshold value has been considered to manipulate the boundary of each word in the connected speech. The frame number of the frames whose STE was less than the threshold value was stored as a vector. The threshold value was computed manually using the STE manipulated with MATLAB. The vector of frames was used in the next step to eliminate the excess frames.

4.2.4 Frame Elimination:

The vector of frames derived from the previous step has been used to manipulate the word boundary in this step. The consecutive frames whose STE was less than the threshold were taken into consideration for identifying the word boundary. When the number of consecutive frames was more than one, then these consecutive frames represent the pause between the words in the connected speech. The first frame was treated as the boundary of the word. The frame next to the last consecutive frame was treated as the beginning of the next word. The process was repeated till the end of the vector which constitutes the frame numbers of the frames whose STE was less than the threshold.

4.3 COMPARISON OF DTS AND TFS ALGORITHM

The Dynamic Thresholding Segmentation algorithm uses complex process as the algorithm converts the spectrogram image to the color spectrogram and then into gray scale image. The Temporal Feature based Segmentation uses the short term energy for segmentation which does not require any additional process of conversion.

5. EXPERIMENTAL RESULT

The evaluation of the proposed model was manipulated using the accuracy.

5.1 SPEECH CORPUS

To evaluate the performance of the proposed methodologies, two kinds of agriculture database Agrodat1 and Agrodat2 were constructed. The first database Agrodat1 was a Tamil connected speech database and the other was Agrodat2 which is a syllables database of 200 distinct Tamil syllables for agriculture. Both the databases were recorded with a single input channel microphone in a quiet environment.

The Agrodat1 consists of 200 connected speech sentences with a maximum of 3 words each was uttered 100 times by four speakers (2 male and 2 female). Each sentence was recorded with an average length of 3 seconds. 100 utterances of 200 syllables from the same four speakers constructed the training data and 10 utterances of 100 syllables randomly selected from the 200 syllables were used for evaluation. In the database Agrodat2, 20 utterances of each syllable were recorded with variation in speech rate. Each syllable was modeled by a five state left to right HMM and four mixtures Gaussian were used for representing the observation distribution in each state. Speech signal was sampled at the rate of 8 kHz and segmented into 25 ms frame with 15ms overlap at every 10 ms. Each frame was parameterized by 39 dimensional feature vectors consisting of 3 energy coefficients, 12 mel frequency cesptral coefficients and their first and second derivatives.

5.2 ACCURACY

The accuracy could be measured using the measures like Word Error Rate (WER), Correctness Percentage, False Detection Percentage, Match Error Rate (MER) etc. The accuracy of segmentation algorithms could be measured using the Missed Detection Percentage (MDP) and False Detection Percentage (FDP) [25]. The MDP and the deviation from the manual method were the measures used in this work to measure the accuracy. The accuracy of the procedure was manipulated by finding the boundary frame number of each segment using the proposed model and the segmentation boundary found manually. The deviation was calculated by the difference of the manual method and the proposed segmentation procedure. The Missing Detection Percentage is measured by finding the percentage of the number of frames missed from boundary detection divided by the actual number of frames which is given in Eq.(3) [25].

$$MDP = \frac{Missed\ Detection}{Actual\ points} \times 100 \tag{3}$$

Table.3. Boundary of each word in the Connected Speech manipulated using DTS and TFS

	Manual			DTS			TFS		
	B1	B2	B3	B1	B2	B3	B1	B2	B3
CT1	28	70	113	28	70	113	28	70	113
CT2	36	70	101	36	70	101	36	70	101
CT3	38	85	135	38	85	135	38	85	135
CT4	46	83	140	46	83	140	46	83	140
CT5	55	106	145	55	106	145	55	106	145
CT6	53	85	130	53	85	130	53	85	130
CT7	38	92	126	38	92	126	38	92	126
CT8	45	90	134	45	90	134	45	90	134

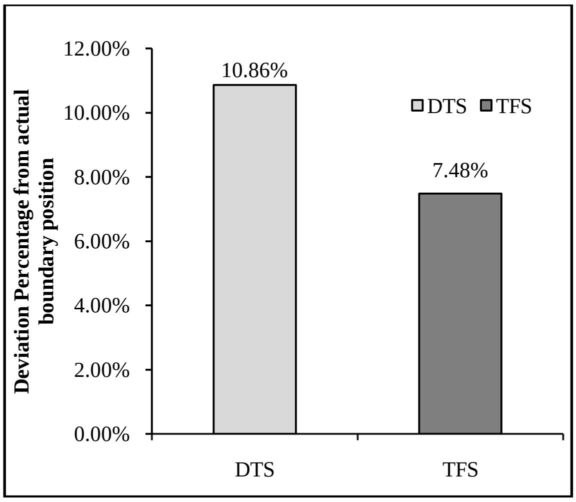

Fig.7. Boundary Frame Deviation of DTS and TFS Algorithms from actual boundary

Table.4. Missed Detection Percentage of the DTS and TFS Algorithms

Algorithm	Missed Detection Percentage
DTS	23.82%
TFS	14.94%

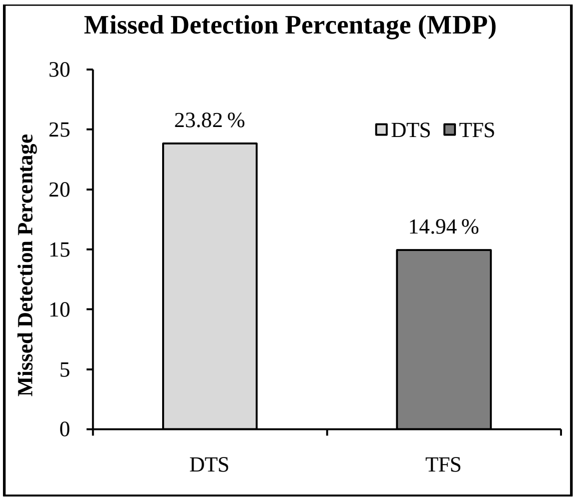

Fig.8. Graphical representation of Missed Detection Percentage of the DTS and TFS Algorithms

The boundary of each word in the connected speech was found using the existing method Dynamic Thresholding segmentation and the proposed method Temporal Feature based segmentation. The performance accuracy was computed by identifying the deviation from the boundary position found manually from the connected speech signals with the help of audacity tool which is shown in Table.3. Three word connected words was segmented into three isolated words using the DTS and TFS algorithms. The boundary of the three isolated words was represented using B1, B2 and B3 respectively. The speech signal was represented as collection of frames in the algorithms DTS and TFS. It could be noticed from the Fig.7 that the deviation was less in TFS compared to the DTS algorithm. The deviation was less by 3.38% in the proposed TFS algorithm. The missed detection percentage of the proposed TFS algorithm was lesser than the DTS algorithm as shown in Fig.8 and Table.4. The MDP in Temporal Feature Segmentation was 8.9% less than the existing Dynamic Thresholding Segmentation Algorithm.

6. CONCLUSION

The connected speech could be segmented into words using many attributes like pitch, duration, rhythmic cues. The proposed method has used the short term energy to perform segmentation. The performance of the proposed methodology and the Dynamic Thresholding Segmentation was measured using the accuracy. The accuracy was measured in terms of the variation of the boundary position from the actual boundary position. As the speech signal is divided into frames, the boundary was measured in terms of variation in the boundary frame number. The accuracy of boundary identification was increased in the proposed methodology by using short term energy which is a temporal feature.

REFERENCES

[1] T. Jayasankar, R. Thangarajan and J. Arputha Vijaya Selvi, "Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis", *International Journal of Computer Applications*, Vol. 25, No. 1, pp. 31-36, 2011.

[2] Mohammed. A. Al-Manie, Mohammed. I. Alkanhal and Mansour. M. Al-Ghamdi, "Arabic Speech Segmentation: Automatic verses Manual Method and Zero Crossing Measurements", *Indian Journal of Science and Technology*, Vol. 3, No. 12, pp. 1134-1138, 2010.

[3] Akila A. Ganesh and Chandra Ravichandran, "Syllable Based Continuous Speech Recognizer with Varied Length Maximum Likelihood Character Segmentation", *International Conference on Advances in Computing, Communications and Informatics*, pp. 935-940, 2013.

[4] Hanitha Gnanathesigar, "Tamil Speech Recognition using Semi Continuous Models", *International Journal of Scientific and Research Publications*, Vol. 2, No. 6, pp. 1-5, 2012.

[5] South Asia Language Resource Center, www.southasia.sas.upenn.edu/ tamil/index.html

[6] Alan Langus, Erika Marchetto, Ricardo Augusto Hoffmann Bion and Marina Nespor, "Can Prosody be used to Discover Hierarchical Structure in Continuous Speech?", *Journal of Memory and Language*, Vol. 66, No. 1, pp. 285-306, 2012.

[7] Mikhail Ordin and Marina Nespor, "Transition Probabilities and Different Levels of Prominence in Segmentation", *Language Learning*, Vol. 63, No. 4, pp. 800-834, 2013.

[8] Franco Lancia, "Word co-occurrence and theory of meaning", pp. 1-39, 2005, Available at: http://www.soc.ucsb.edu/faculty/mohr/classes/soc4/summer_08/pages/Resources/Readings/TheoryofMeaning.pdf

[9] Mehryar Mohri, Pedro Moreno and Eugene Weinstein, "Discriminative Topic Segmentation of Text and Speech", *13th International Conference on Artificial Intelligence and Statistics*, Vol. 9, pp. 533-540, 2010.

[10] Natasha Warner, Lynnika Butler and Takayuki Arai, "Intonation as a Speech Segmentation Cue: Effects of Speech Style", *9th Conference on Laboratory Phonology*, pp. 37-42, 2004.

[11] Ansgar. D. Endress and Marc. D. Hauser, "Word segmentation with Universal Prosodic Cues", *Cognitive Psychology*, Vol. 61, No. 2, pp. 177-199, 2010.

[12] M. S. Salam, Dzulkifli Mohamad and S. H. Salleh, "Segmentation of Malay Syllables in Connected Digit Speech using Statistical Approach", *International Journal of Computer Science and Security*, Vol. 2, No. 1, pp. 23-33, 2008.

[13] Daniela Braga, Diamantino Freitas, Luis Coelho, Antonio Moura and Maria Joao Barros, "On the Identification of Word-Boundaries using Phonological Rules for Speech Recognition and Labeling", *Forum Acusticum*, pp. 2675-2680, 2005.

[14] Sameer. R. Maskey, Andrew Rosenberg and Julia Hirschberg, "Intonational Phrases for Speech Summarization", *Proceedings of the 9th Annual Conference*

of the International Speech Communication Association, pp. 2430-2433, 2008.

[15] Pavel. A. Skrelin, "Segment Features in Different Speech Styles", *9th Conference Speech and Computer*, 2004.

[16] Michael. S. Vitevitch and Paul. A. Luce, "A Web-based Interface to Calculate Phonotactic Probability for Words and Nonwords in English", *Behavior Research Methods, Instruments & Computers*, Vol. 36, No. 3, pp. 481-487, 2004.

[17] D. R. Van Niekerk and E. Barnard, "Acoustic Cues Identifying Phonetic Transitions for Speech Segmentation", *Council of Scientific and Industrial Research*, pp. 55-59, 2008.

[18] Anastassia Loukina, Greg Kochanski, Chilin Shih, Elinor Keane and Ian Watson, "Rhythm Measures with Language-Independent Segmentation", *INTERSPEECH*, pp. 1531-1534, 2009.

[19] Md Mijanur Rahman and Md Al-Amin Bhuiyan, "Dynamic Thresholding on Speech Segmentation", *International Journal of Research in Engineering and Technology*, Vol. 2, No. 9, pp. 404-411, 2013.

[20] Miael Nilsson and Marcus EJnarsson, "Speech Recognition using Hidden Markov Model", *Department of Telecommunications and speech Processing, Biekinge Institute of Technology*, pp. 17-39, 2002.

[21] Sadaoki Furui, "*Digital Speech Processing, Synthesis and Recognition*", Second Edition, Marcel Dekker: New York, 2001.

[22] Lawrence Rabiner and Biing-Hwang Juang, "*Fundamentals of Speech Recognition*", Prentice-Hall, 1993.

[23] Matthias Wolfel and Hazim Kemal Ekenel, "Feature Weighted Mahalanobis Distance: Improved Robustness for Gaussian Classifiers", *13th European Signal Processing Conference*, pp. 1-4, 2005.

[24] Kuldeep Kumar, R. K. Aggarwal and Ankita Jain, "A Hindi Speech Recognition System for Connected Words using HTK", *International Journal of Computational Systems Engineering*, Vol. 1, No. 1, pp. 25-32, 2012.

[25] Behrouz Abdolali and Hossein Sameti, "A Novel Method for Speech Segmentation Based on Speakers Characteristics", *Signal & Image Processing: An International Journal*, Vol. 3, No. 2, pp. 65-78, 2012.