ISSN:0975-9646

Grapheme Segmentation of Tamil Speech Signals using Excitation Information with MFCC and LPCC Features

Geetha K[#1], Dr. R. Vadivel[*2]

[#]Department of Computer Science, D.J. Academy of Managerial Excellence
Coimbatore, India

[2]Assistant Professor
Department of Information Technology, Bharathiar University

Abstract— **The major components of automatic Speech Recognition(ASR)are the pronunciation dictionary, language models, acoustic model and decoder. The Pronunciation dictionaries define the mapping between the words and basic sounds of a language and thus play a vital role in speech recognition systems. Construction of the pronunciation dictionary is expensive and time consuming since it requires the knowledge of the target language. Since phoneme is context dependent, a higher unit than phoneme is considered with the aim to develop a sophisticated tool for speech to text application. The proposed segmentation algorithm is tested on the continuous speech read by 4 native speakers. Energy and spectral centroid features are used to remove the silent portion and vowel onset point is used as the anchor point to find the beginning of the vowel. The proposed method is analysed with various time tolerances and the results are presented.**

Keywords— **Speech Segmentation, Grapheme Segmentation, MFCC, LPCC.**

I. INTRODUCTION

Modern Automatic Speech Recognition(ASR) systems use phoneme as the sub word unit in larger vocabulary systems in non phonetic languages like English. For languages with a reasonable grapheme-to-phoneme relation, the grapheme-based modeling is a fast and efficient method that avoids the labor and cost intensive manual generation of pronunciation dictionaries[1]. Kanthak and Ney[2] proposed Grapheme-Based Speech Recognition (GBSR) for the language Dutch, Italy, German and English.

Reya et al. [3] concluding in their research that graphemes are perceptual reading units so that it can be considered the minimal ‘functional bridges’ in the mapping between orthography and phonology. Galescu and Allen [4] present a statistical model for language independent bi-directional conversion between spelling and pronunciation, based on joint grapheme/phoneme units extracted from automatically aligned data. Killer [5] has tried building grapheme based speech recognition as a way to build large vocabulary speech recognition systems.

Tamil is phonetic language, follows unique letter to sound rules and it is straight forward in the sense that it has close relation between what has been written to what would be pronounced. One of the unique feature of Tamil is that it does not have distinct letters for voiced and unvoiced plosives, although both are present in the spoken language as allophones[6]. The pronunciation duration of Tamil linguistic units is proposed in Table I which is used as a cue to find the boundary between VOPs in the proposed method.

Table I
Pronunciation duration of Tamil linguistic units.

Linguistic unit	Time length of pronunciation in milliseconds
Consonants	125
Short Vowel	250
Consonants + Short Vowel	
Long Vowel	500
Consonants + Long Vowel	
Diphthong	375
Consonants + Diphthong	

This paper is organized as follows: In Section 2 the review of literature. Section 3 describes speech segmentation and its mathematical representation. Section 4 describes about the data used for the experiment, endpoint detection method to segment speech data into individual words and the feature used for the experiment Section 5 presents proposed framework. Section 6 presents experimental results and comparative analysis of the proposed method with manual segmentation. Section 7 provides the conclusion and the scope for future work.

II. LITERATURE REVIEW

Kanthak et.al [2] proposed decision tree based grapheme acoustic sub-word units with phonetic questions. They have reduced the manual effort in question generation by automating it and presented experimental results on four corpora with different languages: Dutch ARISE corpus, the Italian EUTRANS EVAL00 evaluation corpus, the German VERBMOBIL ’00 development corpus and the English North American Business ’94 20k and 64k development corpora. They had shown that for the Dutch, German and Italian corpora, the presented approach works surprisingly well and increases the word error rate by not more than 2% relative. But on the English NAB task the error rate is about 20% higher compared to experiments using a pronunciation lexicon.

Borislava Mimer et.al.[1] investigated enhanced decision tree clustering scheme for context dependent modeling for grapheme based speech recognition on GlobalPhone database. They trained grapheme based speech recognizers using Janus Recognition Tool kit in two languages: English and German. They did the experiment by extracting MFCC features of 6 neighbors followed by dimensionality reduction techniques and Feature space adaptation. They showed that the word error rate is reduced 9.3% and 4.1% in English and German respectively and suggested flexible parameter tying is a successful scheme for grapheme-based speech recognition.

In Indian Languages, Consonant-Vowel (CV) segment occurs frequently in a word. Since CV unit has vital role as basic unit of speech production, CV units can be considered as sub-word units[7]. Research works are done in stop consonant vowel units in Hindi[8][9], CV units in Telugu[10], and Malayalam[11].

Tamil Language has distribution of phonemes with restricted rules such as

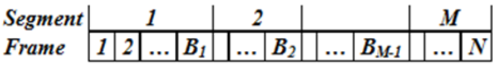
- Has restricted number of consonant clusters.
- Initial consonant clusters are not acceptable.
- All words ends with either a vowel or a sonorant.
- No word ends in an obstruent except borrowed words etc.[12].

III. Speech Segmentation

Segmentation of speech signal is a fundamental task in Speech Recognition System. Most of the speech recognition systems in non phonetic languages use phoneme as the basic unit for modeling. Since phonemes are context dependent and follows tedious process to find the boundary, higher levels of speech units than phoneme, such as grapheme, syllable are tried in many researches.

A. *Mathematical Formulation of Segmentation*

The problem of speech segmentation is described in Fig.1. Let a speech utterance of S samples represented by $\{X_i\}_{i=1}^{N}$, is the first derivative MFCC parameter vector sequence of N speech frames, where X_i is p dimensional parameter vector at frame 'i'. The segmentation problem is to find M consecutive segments in the N frame sequence. Let the boundaries of the segment be denoted by the sequence of integers$\{B_i\}_{i=1}^{M}$. The ith segment stats at frame B_{i-1} +1 and ends at frame B_i; where B_0=0 and B_M=N.



Segment	*1*				*2*				*M*	
Frame	*1*	*2*	…	B_1	…	B_2	…	B_{M-1}	…	*N*

Figure 1. Segmentation of $\{X_i\}_{i=1}^{N}$, into M segments

IV. Data, Pre Processing and Feature Extraction

A. *Data*

Frequently used Tamil words uttered by 4 native speakers are recorded with the help of a unidirectional microphone and considered as data set. Data are recorded using a recording tool audacity in a normal room. The sampling rate used for recording is 16 kHz. The description about the data used is also given in Table II.

TABLE II
Specification used in Creating Dataset

Description	*Feature*
Language	Tamil
Speech type	Continuous speech
Recording Conditions	Room Environment
Number of Speaker	4
Gender	2: Male 2 Female
Age group	25-30
Region	Native

B. *Pre Processing*

The silence removal is one of the dimensionality reduction techniques in speech signal processing. An algorithm for salience removal for the proposed method is same as that proposed in[13]. The acoustic features signal energy and the spectral centroid are used for every frame of 50ms. The Computation of Energy is given in eqn. (1) and Spectral centroid is given in eqn. (2).

$$\text{Signal Energy}= A = \frac{1}{N}\sum_{n=1}^{N}|x_i(n)|^2 \qquad (1)$$

$$\text{Spectral Centroid } C_i = \frac{\sum_{k=1}^{N}(k+1)x_i(k)}{\sum_{k=1}^{N}x_i(k)}x_i(k), \qquad (2)$$

Both energy and spectral centroid will be low in the silent region of the speech signal. The results obtained after pre-processing stage using the above algorithm is represented in the fig 2. In Figure 2(a) the short time energy of the original and the filtered signal is plotted and in Figure 2(b) the spectral centroid of the original and the filtered signal is plotted and in Figure 2(c) the original wave form is shown.

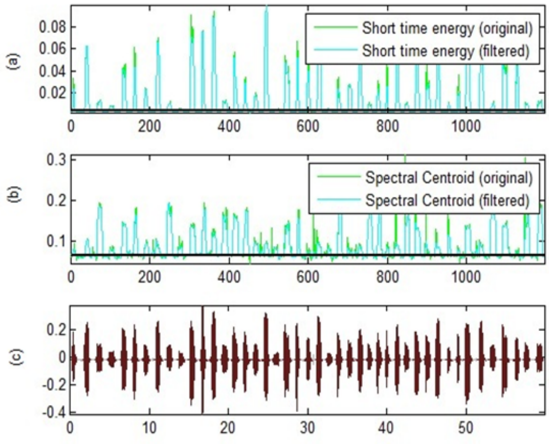


Figure 2 (a)Short term Energy of the Speech Signal (b) Spectral Centroid of the signal (c) Waveform of speech signal

C. Feature Extraction

The speech signal is decomposed into a sequence of overlapping frames. The frame size of 25 ms and 10ms frame shift were used for the segmentation approach considered. The input speech data are pre emphasized with co-efficient of 0.97 using a first order digital filter. The samples are weighted by a Hamming window for avoiding spectral distortions. The resulting windowed frames are used to extract both features MFCC and LPCC

1) MFCC Feature Extraction

The resulting windowed frame obtained from pre processing is used to extract the 12 Mel Frequency Cepstral Coefficients(MFCC) which excludes the zero order coefficient that represent the total energy and this vector is used to find the boundary. The short-time Fourier transform analysis performed after windowing to compute the magnitude spectrum. It is followed by filter bank design with triangular filters uniformly spaced on the mel scale between 300 Hz to 3400 Hz as lower and upper frequency limits. The filter bank is applied to the magnitude spectrum values to produce Filter Bank Energies (FBEs) 20 per frame. Log-compressed FBEs are then de-correlated using the Discrete Cosine Transform (DCT) to produce cepstral coefficients. The co-efficients are rescaled to have similar magnitudes achieved through liftering with L = 22. The steps involved in the MFCC feature extraction are shown in Fig. 3.

2) LPCC Feature Extraction

The resulting windowed frame obtained from pre processing is used in auto correlation analysis, in which each frame of the windowed signal is auto correlated and provides p + 1 auto correlations for each frame, where p is the order of LPC analysis. Normally, 8 to 16 are used for values for p. In LPC analysis, each frame of p + 1 auto correlations are converted into LPC parameter set in which might be the LPC coefficients, the reflection coefficients, the log area ratio coefficient, the cepstral coefficients or any desired transformation of the above sets. The formal method for converting from autocorrelation coefficient to an LPC parameter set is known as Durbin's method. Then the LPC parameter set is converted into LP Cepstral Coefficients(LPCC)[14] and the process is clearly depicted in Fig.4.

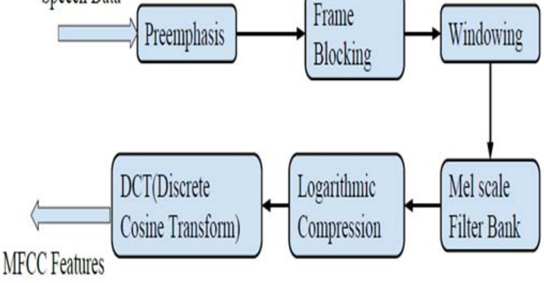


Figure 3. MFCC Features Extraction

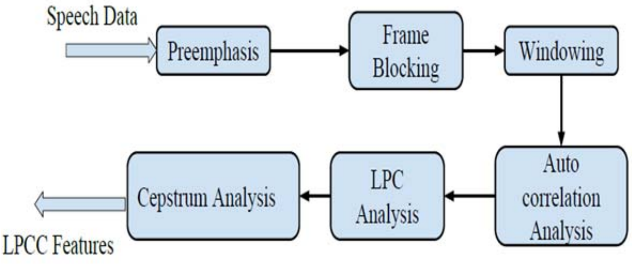


Figure 4. LPCC Features Extraction

V. PROPOSED METHOD

The outline of the proposed method is presented in figure 5. Initially, words are extracted using silence removal algorithm. The acoustic characteristics of the each word are analyzed by performing short-time analysis of speech signals. Each word is processed as a sequence of frames of size 20msec with a shift of 10 msec. For each frame 13 LPCC features and 13 MFCC features are extracted and each extraction method is described in the previous section. Number of vowel onset points is found. Distance between VOPs and the number of frames with the more spectral changes within the VOPs define the number of boundaries NB. If consonant cluster exits NB is set as 2 other wise set as 1. Actual boundaries are identified by finding the frame in which more spectral transition[15] occurs which is estimated using the eqn. 3 proposed in [16].

$$\Delta \boldsymbol{x_m}(\boldsymbol{i}) = \frac{\sum_{k=-K}^{K} \boldsymbol{k h_k x_m}(\boldsymbol{i}+\boldsymbol{k})}{\sum_{k=-K}^{K} \boldsymbol{h_k}} \qquad (3)$$

where h_k is a symmetric window of length (2K+1).

Step 1: Read speech Signal.
Step 2: Extract words using silence removal algorithm
Step 3: For Each word W_i
Step 4: Derive Perceptual and Production features
Step 5 : Find the Number of VOPs
Step 6: For each consecutive pair of VOPs:
Step 7: Find the number of boundaries NB based on duration and number of spectral peak between VOPs.
Step8: Find the frame as boundary where more spectral transition occurs
Step 9: Repeat the steps 6 to step 8 for all remaining VOP frames
Step 10: Repeat the steps 4 to 8 for all words
Step11: Evaluate the performance with manual segmentation.

Figure. 5. Procedure for Proposed Grapheme Segmentation Method

Vowel onset Point is a point at which the consonant region ends and vowel region begin in a CV utterance. Utterances of CV units consist of different speech production events like closure, burst, aspiration, transition and vowel. All CV units have a distinct VOP in their production, which is the significant property useful in CV unit segmentation.

Different methods are found in the literature to find the VOP by utilizing various features and their combinations[17][18]. The combined evidence plot of source excitation, spectral peak and modulation spectrum has been used to find VOP in the proposed method[19]. Steps involved in finding the vowel onset point are presented and VOP detected for a sample Tamil word is shown in the figure 7.

Step 1: Preemphasize input speech signal with α=0.97

Step 2: Compute the LP residual using frame size of 20ms, frame shift of 10 ms and 10th order LP analysis.

Step 3: Find the Hilbert envelope of the LP residual.

Step 4: Convolute Hilbert envelope with a modulated Gabor window of size 800

Step 5: Select peaks preceded with negative regions called valleys

Step 6: Eliminate Local Peaks which are closer to prominent peaks with distance less than 50ms to 70 ms

Step 7: Select prominent peaks which are the vowel onset points

Figure 6. Procedure to find the VOP

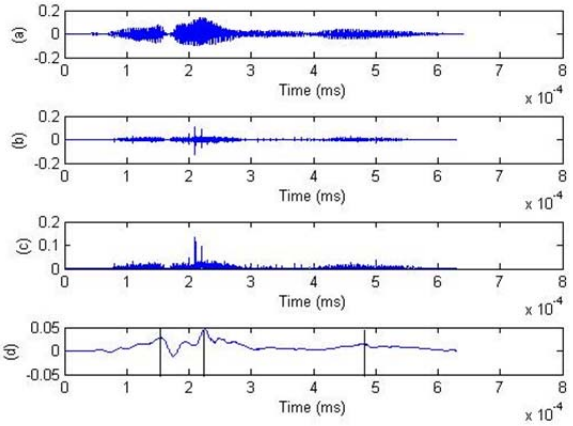


Figure 7 (a) Speech signal of sample Tamil word (b) LP residual (c) Hilbert envelope of LP residual, (d) VOP evidence plot with hypothesized VOP

VI. Experimental Results

Performance of the proposed method is calculated by finding the relation between the number grapheme boundaries detected correctly by the algorithm to the actual number of grapheme boundaries presented in the speech corpus as hand labelled boundaries. The match is between the detected grapheme boundary and the hand labelled grapheme boundary is computed and analysed by considering the tolerance value ranging from range 10 ms to 30ms. Match percentage is computed by Number of Grapheme boundaries detected by the algorithm to the hand labelled boundaries given by eqn. 4 and the results are presented in Table III.

$$\text{Match percentage } (\%M) = \frac{\text{NGBD}}{\text{NGBDHM}} \qquad (4)$$

Where

NGBBD: Number of Grapheme Boundary Detected

NGBDHM: Number of Grapheme Computed by Hand Labeled Method

Table III

Performance of Grapheme Segmentation

Speaker	Duration in Seconds	No. of Words	No. of Graphemes	Feature	Percentage of Match with tolerance(%M)		
					10ms	20ms	30ms
S1	8	5	13	MFCC	68.97	75.86	96.55
				LPCC	55.17	62.07	89.66
S2	15	10	33	MFCC	67.86	80.36	91.07
				LPCC	58.93	73.21	85.71
S3	29	20	70	MFCC	65.45	74.55	89.09
				LPCC	60.00	70.91	83.64
S4	64	50	173	MFCC	76.92	80.59	92.31
				LPCC	73.63	77.66	89.74
Average					62.73	72.83	89.29

VII. Conclusions

Segmentation of connected Tamil speech signal into grapheme has many in applications like speech to text editors, speech recognition in mobiles and coding. In the proposed work, a novel Tamil grapheme speech segmentation task has been carried out to identify the boundaries of graphemes in the given speech. The proposed method uses the VOPs as anchor points to identify the position of beginning of the vowel unit, the duration of graphemes and the number of spectral changes between the VOPs are used as cue to find the boundaries of graphemes. The performance of the proposed method is evaluated using a range of frame tolerance to find the percentage of match with handmade segmentation. The proposed method provides reasonable results for segmentation of graphemes. In future the work can be extended with enhanced alignment technique to increase the percentage of match with actual boundaries. To avoid the problem identifying the boundary between consonant clusters, syllable based sub word unit may also be proposed.

REFERENCES

[1] Mimer, B., Stüker, S., & Schultz, T. (2004). Flexible decision trees for grapheme based speech recognition. In Proceedings of the 15th Conference Elektronische Sprachsignalverarbeitung (ESSV).

[2] Kanthak, S., & Ney, H. (2002, May). Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition. In ICASSP (Vol. 2, pp. 845-848).

[3] reya

[4] L. Galescu and J. Allen. Bi-directional conversion between graphemes and phonemes using a joint n-gram model. In Proceedings of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, 2001.

[5] Killer, M., Stüker, S., & Schultz, T. (2003, September). Grapheme based speech recognition. In INTERSPEECH.

[6] Thangarajan R.; Natarajan A M.; Selvam M. (2009) Syllable modeling in continuous speech recognition for Tamil language. International Journal of Speech Technology, pp. 47–57.

[7] C.C. Sekhar, S.M. Santhosh and B. Yegnanarayana, "A Modular Approach for Recognition of Isolated Stop-Consonant-Vowel (SCV) Utterances in Indian Languages", Journal of the Acoustic Society of India, Vol. 23, No. 1, pp. 28-35,1995.

[8] G. Lakshmi Sarada, A. Lakshmi, Hema A. Murthy and T. Nagarajan, "Automatic Transcription of Continuous Speech into Syllable-like Units for Indian Languages", Sadhana, Vol. 34, No. 2, pp. 221-233, 2009.

[9] Anil Kumar Vuppala1, K. Sreenivasa Rao and Saswat Chakrabarti, "Improved Consonant-Vowel Recognition for Low Bit-rate Coded Speech", International Journal of Adaptive Control and Signal Processing, Vol. 26, No. 4, pp. 333-349, 2011.

[10] T.M. Thasleema and N.K. Narayanan, "Wavelet Transform based Consonant-Vowel (CV) Classification using SVMs", Proceedings of the 19th International Conference on Neural Information Processing, pp. 250-257, 2012.

[11] Prasanna S R M.; Reddy B V S.; Krishnamoorthy P. (2009) Vowel onset point detection using source, spectral peaks, and modulation spectrum energies. IEEE Transactions on Audio, Speech, and Language Processing. 17(4), 556–565.

[12] http://www.languageinindia.com/dec2001/nramaswami.html

[13] Theodorous G. A method for silence removal and segmentation of speech signals. Computational Intelligence Laboratory (CIL), Institute of Informatics and telecommunications, NSCR Demokritos; Greece. 2009.

[14] Lawrence Rabiner and Biing-Hwang Juang.; (1993). Fundamentals of Speech Recognition. Prentice Hall. 102-108.

[15] SaiJayram A K.;, Ramasubramanian V.; Sreenivas T V. (2002) Robust Parameters for Automatic Segmentation of Speech. Proc. of ICASSP, 1-513-1-516.

[16] Dusan S.; Rabiner L.(2006). On the relation between maximum spectral transition positions and phone boundaries. Proceedings of 9th International Conference on Spoken Language Processing, 645-48.

[17] Vuppala, A.K. , Rao, K.S. , Chakrabarti, S. " Improved vowel onset point detection using epoch intervals", AEU –Int. J. Electron. Commun. , 2012a ,66 (8), 697–700 .

[18] Vuppala, A.K. , Yadav, J. , Chakrabarti, S. , Rao, K.S. , "Vowel onset point detection for low bit rate coded speech", IEEE Trans. Audio Speech, Lang. Process. 2012b , 20 (6), 1894–1903.

[19] Prasanna, S. R. M., Reddy, B. V. S., and Krishnamoorthy, P., "Vowel onset point detection using source, spectral peaks, and modulation spectrum energies", IEEE Transactions on Audio, Speech, and Language Processing, 2009, 17(4), 556–565.