AAYUDHA: A Tool for Automatic Segmentation and Labelling of Continuous Tamil Speech

Laxmi Sree B. R.
Faculty in Computer Science
Dr. G R Damodaran College of Science
Coimbatore - 641 014 India

Suguna M.
Faculty in Computer Science
Dr. G R Damodaran College of Science
Coimbatore - 641 014 India

ABSTRACT
Speech! An effective way of communication between human is now becoming an alternative way to communicate between human and machine. This alternative way is now-a-days used in many real time systems for faster, easier and comfortable response and communication. Speech segmentation and labelling are the process that lay as a key to decide the accuracy of several speech related research. A tool 'AAYUDHA' is proposed that enables automatic segmentation and labelling of continuous speech in Tamil. Two different segmentation algorithms, one based on Fast Fourier Transform (FFT) feature set and 2D filtering and other based on Discrete Wavelet Transform (DWT) feature set and its energy variation in different sub-bands are implemented. The segmentation accuracy of those algorithms is analyzed. Further the segmented speech is labelled using a baseline Hidden Markov Model (HMM) based acoustic model. A speech corpus named 'KAZHANGIYAM' is created which includes the recorded Tamil speech of various speakers. The database also includes the information of manually segmented data of those speech data. This speech corpus is used to analyze the accuracy of the algorithms used in the proposed tool. This tool concentrates on the phonetic level segmentation of Tamil speech. The tool shows an acceptable segmentation and labelling accuracy.

General Terms
Speech segmentation, labeling, algorithm, speech corpus, segmentation tool.

Keywords
FFT, DWT, automatic segmentation, labelling, Tamil speech.

1. INTRODUCTION
Speech segmentation is the process in which a speech signal is divided into smaller units by identifying the boundaries between words, syllables or phonemes in the recorded waveforms of spoken natural languages[1][2]. The difficulty of this problem is compounded by the phenomenon of co-articulation of speech sounds. The classical solution to this problem is manual segmentation or automatic segmentation of speech. The manual segmentation can be done using the tools like Praat[3], ESPS, WaveSurfer[4]. Manual segmentation of speech into phonemes is a highly time-consuming, tiresome and difficult process. This calls for automating the segmentation process. Proposed here is a tool for the automatic segmentation of continuous speech in Tamil for phoneme level. The proposed tool will be seen as a friendly tool to segment Tamil speech comparatively providing results as existing alignment tool, namely HTK [5], and Praat plug-in tool for phonetic segmentation, respectively, which facilitates the alignment process possible by novel users. The whole process starts from a sound file to its orthographic (or phonetic) transcription within a text file or in a convenient TextGrid format.

The rest of the paper is organized as follows: Section 2 reviews the literature of speech segmentation, section 3 gives an overview on the proposed tool 'AAYUDHA', section 4 describes about the speech corpus 'KAZHANGIYAM', section 5 discusses about the various algorithms used in the tool, section 6 discusses the experimental results and section 7 is the conclusion and future enhancement.

2. LITERATURE SURVEY
G Lakshmi Sarada et al [6] proposed a group delay based algorithm for automatic segmentation of continuous speech signal into syllable-like units for Indian languages and an Unsupervised and Incremental Training method for grouping of similar syllable segments. During training process, isolated style HMM models are generated for each of the clusters and the speech signal is segmented into syllable-like units during testing which are then tested against the HMMs obtained during training and results in a syllable recognition performance of 42•6% and 39•94% for Tamil and Telugu. The proposed feature extraction technique that uses the features extracted from multiple frame sizes and frame rates results in a recognition performance of 48•7% and 45•36%, for Tamil and Telugu respectively. The performance of segmentation followed by labelling is superior to that of a flat start syllable recognizer (27•8% and 28•8%for Tamil and Telugu respectively).

[7] Presented the evaluation of different feature extraction techniques for continuous speech data. In speech recognition system, extraction of human voice feature is the most important process. Pitch and formants are the significant features of any human voice. Feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), Zero Crossings with Peak Amplitudes (ZCPA) are studied in the literature. Principle Component Analysis (PCA) is used to enhance the results of feature extraction techniques such as ZCPA.

The properties of the group delay functions are reviewed [8]. The information in speech signals that is derived from short-time Fourier analysis is represented in terms of features in conventional methods. The features extracted from the magnitude of the Fourier transform (FT) are measured during analysis and the significance of FT phase is highlighted for representing information in speech signal. The group delay function is implemented to extract the information in the short-time FT phase. The characteristics of the vocal-tract system are captured using modified group delay function. Applications of group delay functions such as segmentation of speech into syllable boundaries are discussed in detail.

Bartosz Ziółko [9] proposed a method for speech signals segmentation based on the time-frequency analysis. This approach works on the basis of discrete wavelet transform resulting in power spectrum and derivative information that are used to locate the boundaries of phonemes. To validate the efficiency of segmentation method, a corpus of male speaker in Polish language is implemented using a statistical classification method.

Automatic phoneme recognition techniques from spoken speech are explored [10]. The objective of this method is to extract as much information about phoneme from as long temporal context as possible and this is carried out by the Hidden Markov Model / Artificial Neural Network (HMM/ANN) hybrid system. The proposed approach involves two phases. In the first phase, the Temporal Pattern (TRAP) system is implemented and compared to other systems based on conventional feature extraction techniques. In the second phase, a new Split Temporal Context (STC) system is proposed that reduces the complexity. The performance of the system is further improved by using three-state phoneme modeling and phonotactic language model and reaches 21.48 % phoneme error rate when applied on the TIMIT database and databases with noise.

Piero COSI [11] presented an interactive Segmentation and Labelling Automatic Module (SLAM) for Windows-based Personal Computers. The proposed system is planned to work with big amount of speech material of scientists which is highly time-consuming task. The system is implemented using Microsoft C++ and Windows 3.1 SDK software, which is based on the Multi-Level Segmentation theory, that runs preferably on Intel 386/486-based personal computers running DOS 5.00 or higher and equipped with VGA and SuperVGA boards.

The segmentation of the acoustic signal into syllabic units is an essential stage in the development of a syllable-centric ASR system. T.Nagarajan [12] presented a group delay based approach with minimum phase to segment spontaneous speech into syllable-like units. Three different minimum phase signals are derived from the short term energy functions of three sub-bands of speech signals. This proposed method carried out experiments on Switchboard and OGI-MLTS corpus resulting in 40 m.sec. of time for finding out the error in segmentation for 85% of the syllable segments.

3. AAYUDHA – TOOL DESCRIPTION

The tool has player, recorder, segmentation and labelling functionalities in it. The following are the key options available in the tool. This used uses the algorithm similar to [13] and [14] to perform the segmentation task. The following Figure 1 shows the home screen of tool.

3.1 Player Tab

Player is used to view the speech file chosen for segmentation. In the player, one can select single file or multiple file for the pre-view. Player has options like play, pause, stop, forward, backward and volume. The File→ Open option is used to open sound files.

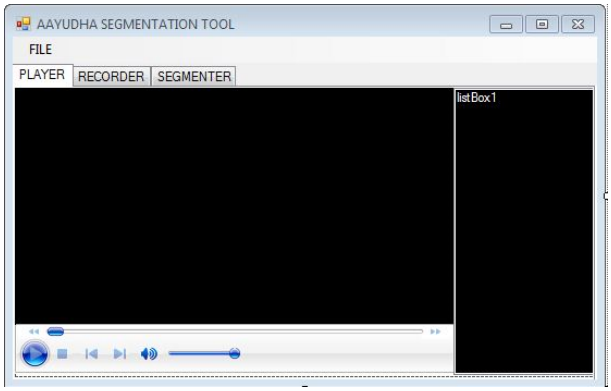

Figure 1 : Home screen of the tool

3.2 Recorder Tab

If one clicks the recorder tab it will show CLICK HERE FOR RECORD (see Figure 2). Clicking that will navigate the user to another window for recording the speech. The recording window have options to start recording, stop recording, play, pause, stop, save, delete and a small panel to display the details about the recorded speech .The panel display the details like date and length of the recorded speech. The save option is available, which allows the user to select the file location and enter a name for the file to save. It saves the file in .WAV format.

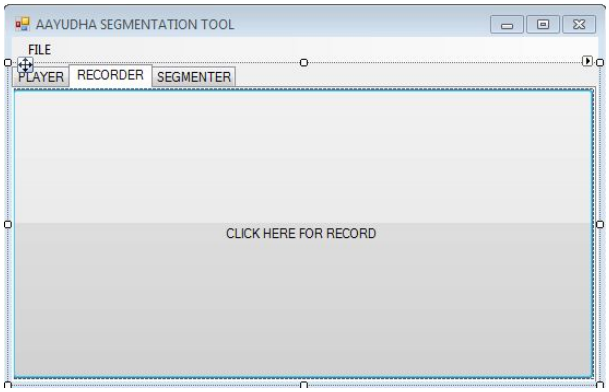

Figure 2 : Recorder tab in the tool

3.3 Segmenter Tab

This tab (see Figure 3) of the tool allows the user to segment and label the selected speech file. This tab provides option to

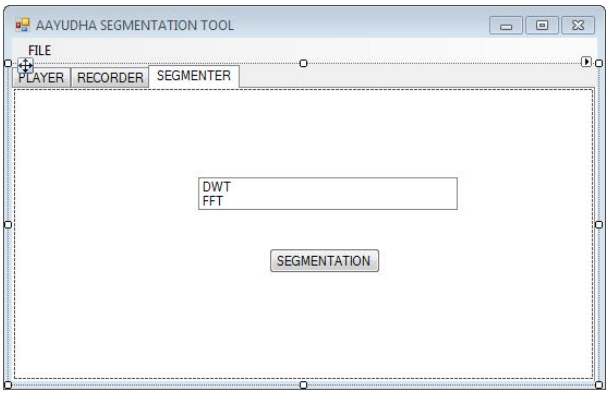

Figure 3 : Segmentation tab showing the choice of two different algorithms

select the segmentation algorithms that are implemented. The algorithms are discussed in the following sections 5.1 and 5.2.

4. SPEECH CORPUS

Any research is need of database to validate itself. Due to lack of benchmark speech corpus in Tamil, a speech corpus named 'Kazhangiyam' Tamil Corpus is created. Tamil speech is directly dependent on the grapheme; it is a grapheme based language. Tamil speech comprises of 35 different phonemes which are listed in Table 1.

Totally, forty five Tamil sentences that deliberately cover all the phonemes are taken into consideration. The speech was collected from thirty nine speakers whose mother tongue is Tamil. The speakers were scattered in the age group from 18 to 40. Sample sentences are listed in Appendix I. The collected speech is manually segmented into world level, syllable level and phonetic level segmentations to enable the validation of proposed tool. The manual segmentation was carried out using Praat (see Figure 4). Thus databases of word collection, syllable collection and phoneme collection have been created.

Table 1 : Phonemes of Tamil Language

அ	a	ஞ்	nj
ஆ	aa	ட்	t/d
இ	i	ண்	nn
ஈ	ii	த்	th/dh
உ	u	ந்	N
ஊ	uu	ப்	p/b
எ	e	ம்	m
ஏ	ee	ய்	ei
ஐ	ai	ர்	r
ஒ	o	ல்	l
ஓ	oo	வ்	v
ஔ	au	ழ்	zh
க்	k/g	ள்	L
ங்	ng	ற்	rr
ச்	ch/s	ன்	n

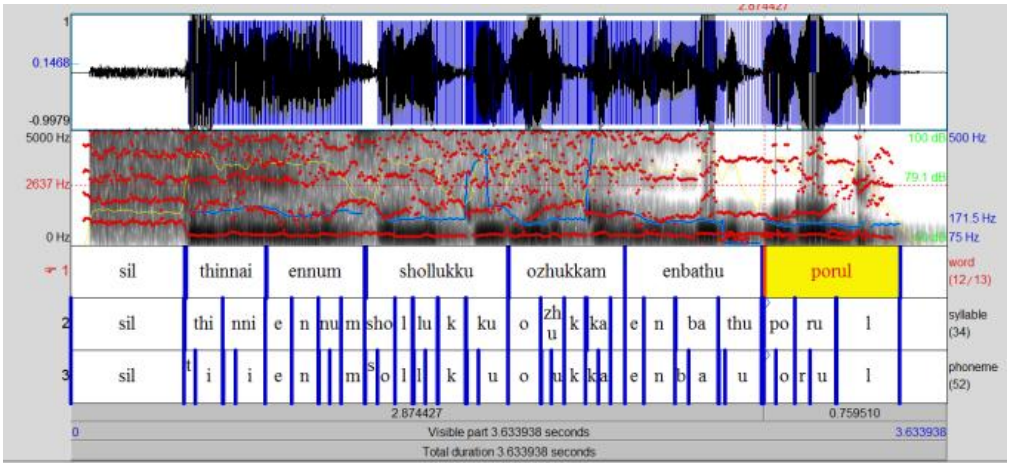

Figure 4 : Hand segmentation using Praat

5. SEGMENTATION AND LABELLING OF SPEECH

5.1 DWT based segmentation algorithm (DWTS)

Input: Raw speech
Output: Phonetic level segmented speech

1. Pre-emphasize the raw speech data.
2. Extract the DWT features of the speech signal. Limit the levels of decomposition using the entropy measure. Higher entropy of data supports further decomposition. The feature set is represented as, $D = \{D_1, D_2, D_3, \ldots\ldots, D_n\}$, where D_j is the j^{th} feature vector, with n feature vectors, and d_{ij} is the i^{th} feature of j^{th} vector.
3. Calculate the power (pow_n) of each subband for each frame from the DWT feature vectors of the speech as follows, $pow_n(i) = \frac{\beta \sum_{j=n-l+1}^{n+l} d_{ij}}{2l}$, where $2l$ is the length of the window.
4. Calculate the change in power between frames as follows, $f_n = \max(pow_{i+1}) - \max(pow_i)$, and smoothen f.
5. Identify the peaks in f by adding a distance constraint between peaks, which gives the segmentation points for the phonetic level segmentation.

5.2 FFT based segmentation algorithm (FFTS)

Input: Raw speech
Output: Phonetic level segmented speech

1. Pre-emphasize the raw speech.
2. A hamming window passed on the filtered speech and speech frames are created.

3. The FFT features of speech frames are extracted and represented as F, where each element f_{ij} in F is the j^{th} feature of i^{th} feature vector.

4. Normalize each frame fi of F by using the mean value σ_i of F, and apply the hyperbolic tangent mapping to compress the coefficients of the feature vectors to simulate human hearing's non-linear sensitivity. $N = \tanh(\alpha . F)$, where α=0.45.

5. A cross correlation matrix C of N is calculated and a custom 2D triangle filter comprising of two triangles in the bottom and a square on its top is applied on the diagonal of the correlation matrix C and the energy at each location is recorded in P, where $p_i = \sum_{i=i-\frac{d}{2}}^{i=i+\frac{d}{2}} \sum_{j=i-\frac{d}{2}}^{j=i+\frac{d}{2}} C_{ij} T_{ij}$, where T is the upper triangular matrix used to represent the filter. The filter design is shown in the Figure 5.

6. The pits in P denotes the segment location, which is further constrained based on the distance between adjacent segment points.

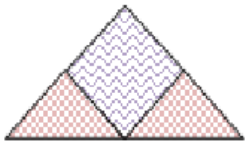

Figure 5 : 2D filter design

5.3 Labelling using Hidden Markov Model

Once the speech has been successfully segmented, it has to be labelled. The labelling process is done using the acoustic model built using Hidden Markov Model (HMM). The DWT features extracted for the segmentation process is used as an input to the labelling process. To build the acoustic model, the training samples of different phonemes are represented as P_1, P_2, P_3....P_N, where N is the total number of phonemes, including one for silence. The HMMs for phonemes is built with 3 states, which depicts the initial, middle and final states of phonemes. Thus, the 3 states along with their $mean_k$, cov_k *(covariance), and* t_{ij} *(transition probability) a*re initialized. Viterbri algorithm is used to train each training set, so that each frame in the training data is aligned to some state of the HMM. The updated $mean_k$, cov_k, *and* t_{ij} are calculated at the end of each iteration. The generating probability (g) of individual training data is summed up using the updated HMM values. The iteration is repeated until the difference between the successive g is less than the predefined threshold (τ).The segmented test speech is applied to the labelling module, which then classifies each segment of the speech to the HMM class which has the maximum generating probability.

6. EXPERIMENTAL RESULTS AND DISCUSSION

The speech instances of different speakers were recorded for the database in a controlled environment. The speakers were allowed to preview the content they are reading. The front end design of the tool is done using VB.NET. The algorithms are implemented using Matlab and then integrated with VB.NET.

The results of DWT based segmentation algorithm and FFT based segmentation algorithm are compared (see Table 2) for a speech sample "திணை என்னும் சொல்லுக்கு ஒழுக்கம் என்பது பொருள் ". The sample speech consists of 37 boundary points out of which the DWTS and FFTS identified 27 and 30 points correctly respectively. The accuracy of the algorithms was tested using the performance

Table 2 : Result of the algorithms DWTS and FFTS on a sample speech "திணை என்னும் சொல்லுக்கு ஒழுக்கம் என்பது பொருள்" consisting of 37 boundary points

Algorithm	No. of Correctly identified boundary Points	No. Incorrectly identified boundary points	No. of Missed boundary points
DWT Segmentation	27	12	10
FFT Segmentation	30	9	7

Metrics namely, precision, recall and F-measures. The precision, recall an F-measures are calculated as follows:

$$Presicion = \frac{No.of\ correct\ segment\ points\ that\ were\ identified}{Total\ no.of\ segment\ points\ that\ were\ identified}$$

$$Recall = \frac{No.of\ correct\ segment\ points\ that\ were\ identified}{Total\ no.of\ actual\ segment\ points}$$

$$F-measure = 2.\frac{Precision.Recall}{Precision + Recall}$$

The Table 3 shows the performance of the two algorithms DWTS, and FFTS using precision, recall and F-measure (see Figure 6). The table shows that the segmentation accuracy is better for DWTS when compared to FFTS for the dataset 'Kazhangiyam'. The accuracy of FFTS is seen to be better than that of DWTS.

Table 3: Comparison of Precision, recall and F-measure

Algorithm	Precision	Recall	F-measure
DWT Segmentation	0.75	0.75	0.75
FFT Segmentation	0.81	0.83	0.82

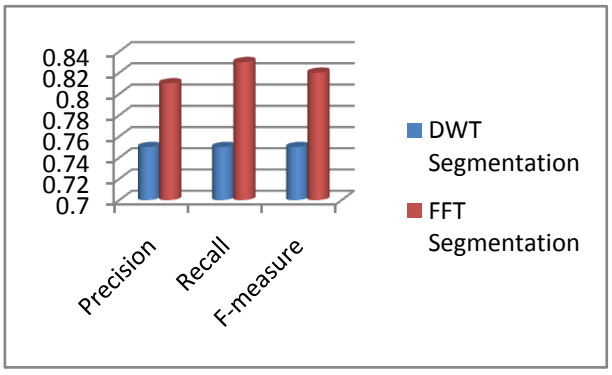

Figure 6 : Chart showing the precision, recall and F-measure of the algorithms DWTS and FFTS

The segmented speech obtained as a result of segmentation process was further sent as input to the labelling process. From the manually segmented speech, for each phoneme, various instances obtained were used to train and test the acoustic model. For training each model, 80% of the speech database is used while 20% are used for testing. The model turned out with a good labelling accuracy of 87.7%.

7. CONCLUSION AND FUTURE ENHANCEMENT

The tool AAYUDHA has been implemented with the functionalities to play, record, segment and label the speech. It has shown acceptable results for segmentation and labelling. It was supported with two segmentation algorithms namely, DWT based segmentation and FFT based segmentation, in which FFT based segmentation shows better results. Labelling of segmented speech is implemented using a baseline HMM based acoustic model. The labelling can further be improved by applying more featured data representation of phonemes in the states of HMM. Neural network, graph based approaches are considered as future enhancements to improve the performance of the tool. The size of the speech corpus will be increased, which also enables to increase the accuracy of segmentation and labelling of the tool.

8. ACKNOWLEDGEMENTS

This work is funded by UGC, New Delhi, India.

9. REFERENCES

[1] Ranjani, H. G. (2008). Explicit Segmentation Of Speech For Indian Languages (Doctoral dissertation, Indian Institute of Science Bangalore-560 012 India).

[2] Elminir, H. K., ElSoud, M. A., & El-Maged, L. A. (2012). Evaluation of different feature extraction techniques for continuous speech recognition. *International Journal of Science and Technology*, 2(10).

[3] Boersma, P., Weenink, D., "Praat: doing phonetics by computer", http://www.praat.org, accessed in Mar 2010.

[4] Sjölander, Kåre, and Jonas Beskow, "Wavesurfer-an open source speech tool." *Interspeech*. 2000.

[5] Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., ... & Valtchev, V. (1997). *The HTK book* (Vol. 2). Cambridge: Entropic Cambridge Research Laboratory.

[6] Sarada, G. L., Lakshmi, A., Murthy, H. A., & Nagarajan, T. (2009). Automatic transcription of continuous speech into syllable-like units for Indian languages. *Sadhana*, *34*(2), 221-233.

[7] Elminir, H. K., ElSoud, M. A., & El-Maged, L. A. (2012). Evaluation of different feature extraction techniques for continuous speech recognition.*International Journal of Science and Technology*, *2*(10).

[8] Murthy, H. A., & Yegnanarayana, B. (2011). Group delay functions and its applications in speech technology. *Sadhana*, *36*(5), 745-782.

[9] Schwarz, P. (2009). Phoneme recognition based on long temporal context.

[10] Ziolko, B., Manandhar, S., Wilson, R. C., & Ziolko, M. (2006, September). Wavelet method of speech segmentation. In *Signal Processing Conference, 2006 14th European* (pp. 1-5). IEEE.

[11] Cosi, P. (1993). SLAM: Segmentation and labelling automatic module. In*Third European Conference on Speech Communication and Technology*.

[12] Nagarajan, T., Murthy, H. A., & Hegde, R. M. (2003). Segmentation of speech into syllable-like units. *Energy*, *1*(1.5), 2.

[13] Okko Rasanen, Unto Laine and Toomas Altosaar (2011). Blind Segmentation of Speech Using Non-Linear Filtering Methods, Speech Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-996-7, InTech, DOI: 10.5772/16433.

[14] Ziolko, B., Manandhar, S., Wilson, R. C., & Ziolko, M. (2006, September). Wavelet method of speech segmentation. In Signal Processing Conference, 2006 14th European (pp. 1-5). IEEE.

10. APPENDIX I

- மோகினியாட்டம் நாட்டியக் கலைகளில் ஒன்று.
- யுவராஜ் உண்மையைக் கூறினான்.
- ஒருவரின் யூகம் எப்போதும் சரியாக இருக்காது.
- ரவி மிகவும் வீரமானவன்.
- முல்லைக்குத் தேர் ஈந்தவன் பாரி.
- ரேவதியின் வீட்டில் ரோஜா செடி உள்ளது.
- ரௌத்திரம் பழகு என்பது பாரதியின் கூற்று.
- வைகை நதி மதுரையில் பாய்கிறது.
- தேனீ நமக்குத் தேனை உணவாகத் தருகிறது.
- பொன்னூஞ்சல் மிகவும் விலை மதிப்பானது.
- நம்முடைய எண்ணம் எப்போதும் உயர்ந்ததாக இருக்க வேண்டும்.