A syllable-based isolated word recognizer for Tamil handling OOV words

P. G. Deivapalan, Hema. A. Murthy
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai, India 600036
Email: {deivapalan,hema}@lantana.tenet.res.in

***Abstract*—Conventional large vocabulary isolated word recognition (LVIWR) systems use a sub-word based continuous speech recognizer to recognise words. In this paper, we approach the problem of isolated word recognizer (IWR) somewhat differently. The word is segmented into subword units, namely, syllable like units using a group delay based approach [1]. During training, syllable models are trained. During testing, the speech signal is segmented into syllable like units. The syllable is recognized using the trained syllable models. A modified *edit distance* based approach is used to identify the word.**

I. Introduction

Many applications such as Airline reservations, data and information retrieval systems require LVIWR. In large vocabulary system, sub-word based approach is a more viable alternative than word based approach inorder to overcome the problem of training data size, as different words can share common segments [2], [3]. In case of Indian languages like Tamil, syllable is considered to be acoustically stable subword unit [4]. Also experiments in [5]–[7] have shown that syllable is a better representative unit for speech recognition. Conventional syllable based IWR tries to align syllable models to form a word. The WER of such systems is high when compared to word based IWR. An improvement in WER could be possible, had those systems use boundary information of the syllables. In this paper, we attempt to use boundary information and propose a new approach for syllable-based IWR. In this approach, the given word utterance is segmented into syllable-like units and each units are recognized using individually trained syllable models. The obtained syllable sequence is then used to output the recognized word.

A segment and recognize based approach [8] has been used for continuous speech recognition (CSR). In [8], the system achieves atmost 30% reduction in WER when compared with conventional CSR approach. In this paper, we use the proposed approach for LVIWR. The main difference between our approach and the one described in [8] is that, we use a simple *edit distance* based matching instead of computationally expensive language model to output the recognized word. The main contribution of this method is that it handles out-of-vocabulary (OOV) word.

Inorder to obtain syllable-like units, group delay based segmentation [1] algorithm has been used. It has been shown in [1] that, group delay based segmentation provides better segment boundaries when compared with short term energy function used elsewhere in [9] for a continuous speech utterance. In this paper, we analyse consistency of group delay segmentation algorithm for isolated speech utterance under various cases.

The proposed syllable based IWR consists of segmentation unit, recognition unit, and word output unit. The given test utterance is automatically segmented into syllable-like units, and these units are then recognized against isolated syllable models producing a sequence of syllables. Finally, the recognized word is obtained from syllable sequence using a simple *edit distance* [10] based matching. Apart from improved word accuracy, the proposed system provides the following advantages: (i) The number of models remain manageable even if the vocabulary size is increased. (ii) In case of OOV words, the system outputs syllable sequence.

The rest of this paper is organized as follows. In Section II, an overview of the proposed syllable based IWR is explained. Section III gives the experimental setup. Section IV briefly outlines the group delay based segmentation and its usefulness in extracting boundary information for isolated word utterance. Section VI describes the procedure for obtaining recognized word from syllable sequence using *edit distance* based approach. In Section VII, the performance of the proposed system is discussed. Section VIII discusses the experiments and results on confusable words. Section IX discusses the experimental results on OOV words. Section X concludes this paper with future direction.

II. System description

The system consists of two parts: a) the training part and b) the testing part. During training, syllable boundaries are manually marked. Different examples of a syllable are used to build isolated style syllable models. Testing is made up of three phases: (i) segmentation of speech utterance into syllable units, (ii) decoding the units using isolated syllable models, and (iii) generating word output from syllable sequence. Fig. 1 shows the functional block diagram of the testing part of the system. The solid lines in Fig. 1 refers to the path taken by a typical word recognition system that uses isolated word models, while the dotted lines refer to the path taken by the proposed system which uses syllable models. Each of the stages are explained briefly in this section.

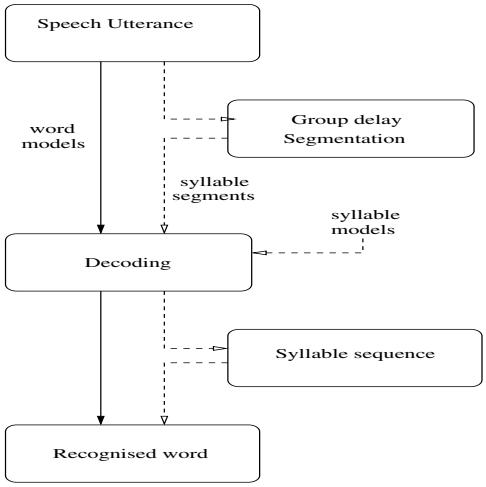


Fig. 1. Functional block diagram of the testing part of syllable-based IWR.

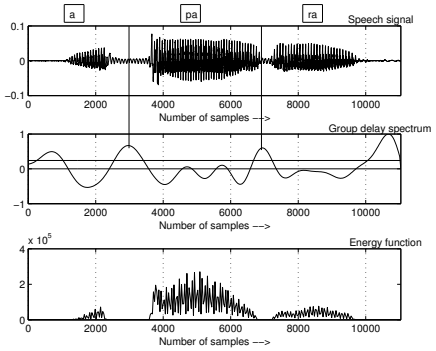


Fig. 2. Group delay based segmentation of utterance *a-pa-ra*. Peak threshold value of 0.25 is used.

- **Segmentation unit** - Group delay based segmentation algorithm [1] is used to obtain syllable units from the isolated speech utterance (see Fig. 2) by detecting peaks in the group delay spectrum. The solid vertical lines correspond to the syllable boundaries.
- **Decoding unit** - The obtained syllable units are Viterbi decoded [11] using isolated syllable models. A sequence of syllables is formed corresponding to the segmented units.
- **Word output unit** - The syllable sequence is then compared against dictionary which consist of syllable sequence of all the words in the vocabulary. A distance score (*edit distance* between recognized sequence and original sequence from dictionary) is obtained for all such pairs. The sequence that maximise the score is considered as the recognized word. If the minimum distance score is very high, the word is assumed to be OOV and the syllable sequence is output.

III. Experimental Setup

The database consists of 400 Tamil words which are used in the context of sports and cricket. The training dataset contains 40 speakers (20 male and 20 female), while the testing dataset contains 10 speakers (5 male and 5 female). The isolated speech utterances were recorded at 16 bits with sampling frequency of 16000 Hz. The 42 (13 MFCC, 13 velocity, 13 acceleration and 3 energy) dimensional features were used in these experiments. Selected binaries of HTK3 [12] toolkit were also used in these experiments.

The following experiments have been carried out:

- Overall performance of the system on test dataset.
- Performance of the system on most confusable words.
- Performance of the system in case of OOV words.

IV. Group delay based segmentation

In [13], it has been shown that minimum phase group delay function derived from the short term energy function can be used for segmentation of a speech utterance into syllable like units. Fig. 2 demonstrates the segmentation of a word utterance *apAra* (using the technique), which is segmented into three syllables *a, pA, ra*. The syllable boundaries are marked by the solid vertical line. The location of the peaks in the minimum phase group delay function approximately corresponds to syllable boundaries.

The desirable property of any segmentation algorithm is in producing consistent segmental boundaries. The performance of the group delay segmentation algorithm is evaluated for the following cases: (i) short duration words and long duration words , (ii) short duration example and long duration example for the same word. Figures 3 and 4 demonstrates the perfomance of the algorithm for case (i) and case (ii) respectively.

A. Segmentation of a short duration word and long duration word

Consider a short duration word, *ani* which consists of two syllables *a-ni* and a long duration word, *thikumukAdinArgal* which consists of seven syllables *thik-ku-muk-kA-di-nAr-gal*. The average word duration for *ani* and *thikumukAdinArgal* are 365.9 ms and 772.5 ms respectively. The segmentation is performed on the words. The results are shown in the Fig. 3. It demonstrates that the syllable boundaries are identified accurately in both cases. In order to suppress spurious boundaries, a peak threshold value of 0.25 (found experimentally) is used.

B. Segmentation of a short duration example and long duration example of the same word

Consider a short duration example and a long duration example of the word, *pAkistAnudan* which consists of five syllables *pA-kis-tAn-nu-dan*. The duration of these examples is 847.5 msec and 1.1025 seconds. The segmentation procedure is performed on these examples and the results are shown in the Fig. 4. It demonstrates that the syllable boundaries are identified accurately in case of long duration example whereas it missed a syllable boundary in case of short duration example. This is shown as dotted lines in Fig. 4. But interestingly, although a boundary is missed, the location of the boundaries

TABLE I
PERFORMANCE (IN %) OF GROUP DELAY BASED SEGMENTATION ON THE TEST DATASET.

Syllable count	13938
No of insertions	916
No of deletions	738
Performance (in %)	87.97

TABLE II
SYLLABLE PERFORMANCE (IN %) OF THE SYSTEM ON THE TEST DATASET.

Insertion Errors (in %)	7.2
Deletion Errors (in %)	6.0
Substitution Errors (in %)	40.6
Accuracy (in %)	61.5

of other syllables are not disturbed. A simple duration analysis could be used to improve the segmentation performance as indicated in [5].

V. DECODING SYLLABLE SEGMENTS

The syllable like segments are decoded using isolated syllable models. The total number of syllable models is 238, whereas the total number of word models is 400. The syllable performance of the system is shown in the Table. II. For example, let us consider three syllable segments s_1, s_2, and, s_3 obtained from the segmentation algorithm as shown in Fig. 2. Viterbi decoding is used to obtain the most likely syllable.

Syllable sequence a, pA, ra corresponding to s_1, s_2, s_3 is obtained. The proposed system achieves a syllable accuracy of 61.5% (see Table. II). Table. II reveals that there is more substitution errors than insertion errors or deletion errors. While analysing the substitution errors, we found the list of confusable syllable pairs and a partial list is shown in Table III. Observe that the confusable syllables in each pairs are rather close.

VI. WORD OUTPUT FROM SYLLABLE SEQUENCE

Edit distance is the most popular distance measure for comparing two strings. It is computed as the minimum number of basic operations required to convert one string into another. The basic operations are (i) inserting a character, (ii) deleting a character, and (iii) substituting a character. We have modified this approach, considering syllable as a basic unit rather than characters. The *edit distance* is computed in two stages. In the first stage, the basic operations are performed on syllables, while in the second stage, the basic operations are performed on characters within the syllable. Any non-zero cost in second stage results in an unit cost for first stage computation. Fig. 5 demonstrates the computation of *edit distance* between test syllable sequence and syllable sequence of a particular word in the dictionary.

The score is given as follows:

$$score = \min_i ||seq_{test} - seq_i|| \quad (1)$$

$$word = \arg\min_i ||seq_{test} - seq_i|| \quad (2)$$

where $i = 1, \ldots, V$ and V is total number of words in the vocabulary.

TABLE III
PARTIAL LIST OF CONFUSION PAIRS WITH GREATER THAN 1 OCCURENCES.

(tha, thu)	(ka, tha)	(ti, thi)	(ku, thu)
(tu, thu)	(ka, ra)	(path, pa)	(il, vil)
(ka, kan)	(a, pa)	(ka, kal)	(kal, thal)
(thir, thirk)	(thu, thum)	(thi, thir)	(kum, vum)
(nai, thai)	(thil, til)	(tha, than)	(vil, vi)
(na, nar)	(il, thil)	(tai, thai)	(thil, thi)

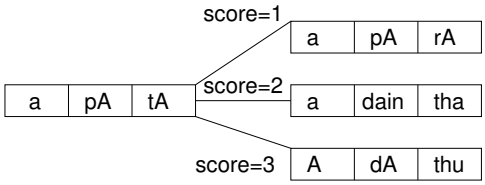


Fig. 5. Edit distance based word searching algorithm.

TABLE IV
PERFORMANCE OF THE SYSTEM. PSB-IWR: PROPOSED SYLLABLE-BASED IWR, CSB-IWR: CONVENTIONAL SYLLABLE-BASED IWR.

Method	Basic unit	Accuracy (in %)	Time (in CPU sec)	Total No of units
IWR	Word	97	1.103	400
PSB-IWR	Syllable	96.37	1.071	238
CSB-IWR	Syllable	77	1.05	238

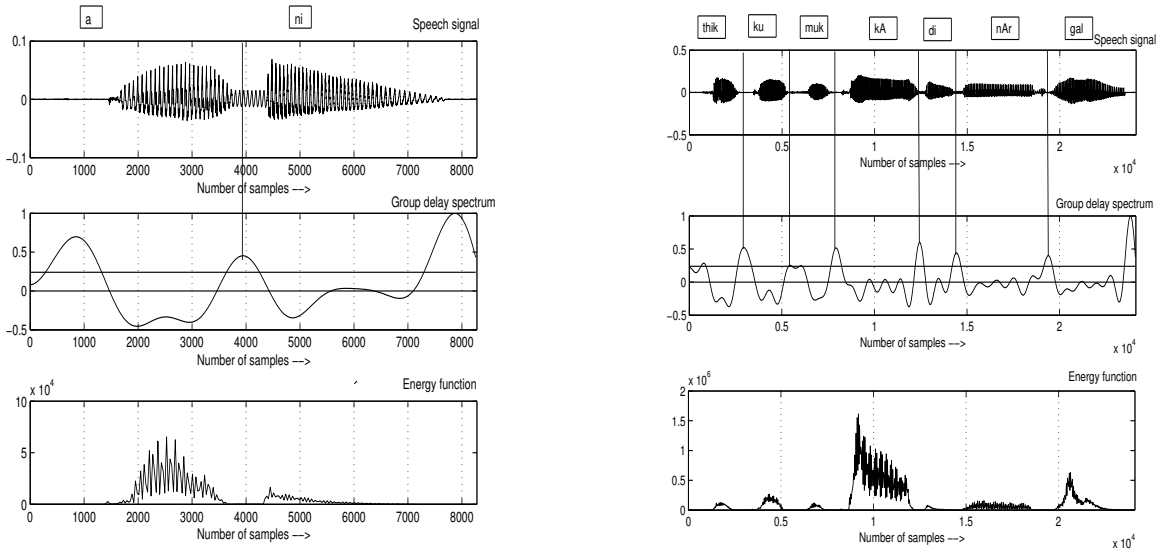
VII. RESULTS AND DISCUSSION

Table. IV shows the performance of the system. We have compared our segmental approach with non-segmented approach which directly use word models. The baseline performance is obtained with word models and the accuracy for the test dataset was found to be 97%. Using segmental approach, we obtained an accuracy of 96.37% for the test dataset. The recognition time of the system in the case of word models is slightly higher than in the case of syllable models since the number of units is less in case of syllables. On the other hand, conventional syllable-based IWR perform rather poorly.

As mentioned in Table. III, the confusability among syllables need to be addressed. A preliminary evaluation on a small set of confusable words has been carried out and is discussed in Section VIII.

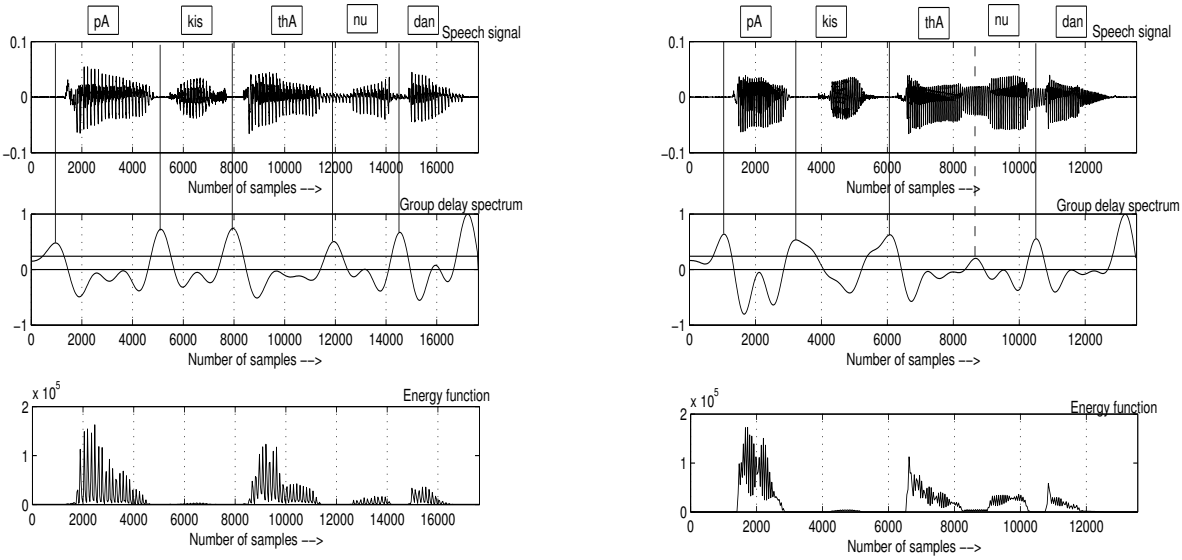
VIII. EXPERIMENTS ON CONFUSABLE SET OF WORDS

The chosen dataset contains five confusable word pairs such that given two words in a pair differ atmost by one



(a) Case I (b) Case II

Fig. 3. Effect of segmenting short duration word, *a-ni* and long duration word *thik-ku-muk-kA-di-nAr-gal*. Case I and Case II is from the same speaker. Peak threshold value of 0.25 is used.



(a) Case I (b) Case II

Fig. 4. Effect of segmenting a longer duration example and shorter duration example of the same word, *pA-kis-thA-nu-dan* from different speakers. The duration of the example in Case I and Case II are 17640 samples (1.1025 sec) and 13560 samples (847.5 ms) respectively. Peak threshold value of 0.25 is used.

TABLE V
PERFORMANCE OF THE SYSTEM ON CONFUSABLE WORDS.

Total No of examples	Syllable accuracy (in %)	WER (in %)
100	87	1

syllable e.g., *a-di-thAr* and *a-di-thAl* differ only in the last syllable. For such words, there is a possibility that it could be misrecognised. In order to study the performance of the system for confusable words, an experiment is carried out taking ten confusable words. The words are as follows *a-di-thAl*, *a-di-thAr*, *a-dain-tha*, *a-dain-tha-thu*, *A-du-va-dhu*, *A-du-va-thAl*, *a-ni-kE*, *a-ni-ku*, *A-tath-thai*, and *A-tath-thil*. Each word has ten examples.

As shown in the Table V, the confusable words are recognised with a syllable accuracy of 87%. From the syllable accuracy (which is quite high), it can be concluded that syllable boundaries are quite accurately determined since the acoustic model is able to distinguish between words that differ only in a single syllable.

IX. EXPERIMENTS ON OOV WORDS

The primary benefit of the system lies in its success in handling the OOV words. In this experiment, another database (B) which consists of 100 Tamil words that have not been seen. This database contains 10 speakers (5 male and 5 female). If minimum distance score between recognised syllable sequence and words in the dictionary contain more than 39% of syllable as misclassified, then the spoken word is considered as an OOV word (since overall syllable accuracy is 61%). The syllable sequence output is shown in Table. VII and system performance is shown in Table. VI. It was found that, only 65% of the syllables that occur in OOV words present in the training dataset. Hence a very low syllable performance has been reported. But recognized syllables match actual syllables in either vowel part or consonant part. It has also been shown that the readability of the OOV word is quite high.

TABLE VI
PERFORMANCE OF THE SYSTEM FOR OOV WORDS IN DATABASE (B).

Models	Total no of models	syllable performance (in %)
syllable	238	32.20

X. CONCLUSION AND FUTURE WORK

In this paper, we have proposed and built a syllable-based IWR for Tamil using segment and recognize based approach. It performs better than conventional syllable-based IWR. The readability of the syllable sequence, outputed in case of OOV words, is quite high. A typical application of this syllable-based IWR would be in embedded speech recognition, where heavy duty language models that disambiguate words can be avoided.

TABLE VII
A PARTIAL LIST OF OOV WORDS AND THEIR SYLLABLE SEQUENCE.

Word	Original syllable sequence	Recognised syllable sequence
etiyathE	e-ti-ya-thE	e-ti-yi-thE
kunamataitha	ku-nA-ma-tai-tha	ku-na-ma-thai-ka
kuvithanar	ku-vi-tha-nar	ku-vi-ta-nar
thotarthu	tho-tar-thu	tho-nar-thu
ullathu	ul-la-thu	pul-la-thu
vativam	va-ti-vam	a-ti-vam
varai	va-rai	va-rai
vEkamAna	vE-ka-mA-na	vE-kam-na
vivaram	vi-var-ram	i-var-ram

As large vocabulary isolated word dataset is not available for Tamil, we could only test our system on much smaller dataset. Preparing a large vocabulary database for Indian languages still is a major concern. The bottleneck is not in collecting database but labeling the collected database. In a step towards that, we have built an automatic tool to segment and label the database in terms of syllable [14].

REFERENCES

[1] T. Nagarajan and Hema A. Murthy, "Group delay based segmentation of spontaneous speech into syllable-like units," in *ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, pp. 115-118, Tokyo, April 2003.

[2] L. Shi-wook, T. kazuyo, and I. Yoxhiaki, "Multilayer subword units for open-vocabulary spoken document retrieval," in *8th International Conference on Spoken Language Processing, INTERSPEECH-2004*, pp. 1553-1556, Korea, October 2004.

[3] T. Svendsen, K. K. Palliwal, E. Harborg, and P. O. Husoy, "An improved sub-word based speech recognizer," in *IEEE Conference on Acoustic, Speech, and Signal Processing, ICASSP-89*, vol. 1, pp. 108-111, Glasgow, May 1989.

[4] South Asia Language Resource Center [Online]. Available http://www.southasia.upenn.edu/tamil/tamilsound.html

[5] Lakshmi. A, Hema. A. Murthy, "A syllable based continuous speech recognizer for Tamil", in *IEEE Internaltional Conference Speech and Language Processing, INTERSPEECH*, pp. 1878- 1881, Pittsburgh, Pennsylvannia, September 2006.

[6] Osamu Fujimura, "Syllable as a unit of speech recognition", in *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 23, no. 1, pp. 82-87, February 1975.

[7] Gauvian J. L, "A syllable-based isolated word recognition experiment", *IEEE International Conf. on ICASSP'86*, vol. 11, pp. 57-60, Apr 1986.

[8] Lakshmi, "A syllable-based continuous speech recognizer for Tamil", M. S. Thesis, Indian Institute of Technology Madras, Chennai, July 2007.

[9] A.G. Adami and H. Hermansky, "Segmentation of speech for speaker and language recognition," in *Proc. of the Eurospeech 2003*, pp. 841-844, Geneva, Switzerland, 2003.

[10] Richard O. Duda, Peter E. Hart, and David G. Stork, "Pattern Classification", pp. 413-421, 2nd Edition., A Wiley-Interscience Publication, NY, 2000.

[11] G. D. Fornay, "The Viterbi algorithm," in *Proc. IEEE*, vol. 61, no.3, pp. 268-278, March 1973.

[12] HTK Hidden Markov Model Toolkit [Online]. Available: http://htk.eng.cam.ac.uk

[13] T. Nagarajan, V. Kamakshi Prasad, and Hema A. Murthy, "The minimum phase signal derived from the magnitude spectrum and its application to speech segmentation", in *Sixth Biennial Conference of Signal Processing and Communications*, July 2001.

[14] P. G. Deivapalan, Mukund Jha, Rakesh Guttikonda, and Hema. A. Murthy, "DONLabel: An Automatic Labeling Tool for Indian Languages", (preprint).