Robust Syllable Segmentation and its Application to Syllable-centric Continuous Speech Recognition

Rajesh Janakiraman, J. Chaitanya Kumar, Hema A. Murthy
Department of Computer Science and Engineering
Indian Institute of Technology Madras
Chennai, India - 600036
Email: rajesh,chaitanya,hema@lantana.tenet.res.in

***Abstract*—The focus of this paper is two-fold: (a) to develop a knowledge-based robust syllable segmentation algorithm and (b) to establish the importance of accurate segmentation in both the training and testing phases of a speech recognition system.**

A robust segmentation algorithm for segmenting the speech signal into syllables is first developed. This uses a non-statistical technique that is based on group delay(GD) segmentation and Vowel Onset point(VOP) detection. The transcription corresponding to the utterance is syllabified using rules. This produces an annotation for the train data. The annotated train data is then used to train a syllable-based speech recognition system. The test signal is also segmented using the proposed algorithm. This segmentation information is then incorporated into the linguistic search space to reduce both computational complexity and word error rate(WER). WER's of 4.4% and 21.2% are reported on the TIMIT and NTIMIT databases respectively.

I. Introduction

Syllables have long been regarded as robust units of speech perception and recognition[1], [2]. Automatic segmentation and labeling of speech at the phonetic level is not very accurate while syllable boundaries are more precise and well defined. In a speech recognition framework, although the syllable as a basic acoustic unit suffers from the problem of training data sparsity, techniques to improve recognizer performance even with small amount training data with longer duration units like the syllable exist[3], [4]. Boundary detection and the use of the same in the recognition framework have proven to be useful in improving recognition accuracy of phone-based continuous speech recognition systems[5], [6]. [7] describes a hybrid ANN/HMM syllable recognizer that tracks vowel phonemes based on discrete hidden markov models, multilayer perceptrons, heuristic rules and models segments between consecutive vowels. They reported a syllable recognition rate of 75.09% on the TIMIT database. Most of these techniques are based on the use of segmentation as obtained from statistical techniques that require significant amounts of train data. On the other hand, we look to propose a purely knowledge based technique that does not require any training apriori.

Previously, [8] developed two-level group delay segmentation to segment the speech signal into syllable-like units. But this technique requires significant tuning for every new database. In particular, the parameters had to be re-tuned when the syllable-rate varied significantly. In the first-level, gross segmentation gave polysyllable boundaries. In the second-level, polysyllables were re-segmented using a duration constraint. But when syllable rates vary significantly, duration constraints cannot be used effectively.

In this paper, different methods to make segmentation robust against variations in syllable rate are explored. First, a syllable rate based parameter lookup is created by mapping an approximate estimate of syllable rate to the correct resolution for segmentation. In another approach, Vowel Onset Points (VOP)'s detected using [9] are used to (i) determine the approximate syllable rate, (ii) disambiguate the syllable boundaries obtained using group delay (GD) segmentation.

Additionally, analysis on incorporating the syllable boundaries into the linguistic framework during recognition is reported.

The remainder of the paper is organized as follows. In Section II, the relationship between group delay segmentation and syllable rate is studied empirically. In Section III the proposed algorithms are discussed. Section IV explores their performance in a segmented syllable based continuous speech recognizer. Section V reports experimental details and conclusions are made in Section VI.

II. Group delay based segmentation of speech

The baseline group delay based segmentation algorithm uses a minimum phase signal derived from the short-term energy(STE) as if it were a magnitude spectrum. The high energy regions in the STE reflect the syllable nuclei and the valleys at either ends of the nuclei reflect the syllable boundaries. The algorithm follows [10].

1) Let $x[n]$ be the samples of a continuous speech utterance.
2) Compute the STE function $E[m]$ using overlapped windows. This can be viewed as the magnitude spectrum (from 0 to π) of some real-valued signal.
3) Using symmetry, extend the spectrum over the region $(-\pi, 0)$, and denote the entire spectrum by $\tilde{E}[k]$.
4) Compute the IDFT of $1/\tilde{E}[k]$ to give $\tilde{e}^i[n]$. This signal is the root cepstrum, the causal portion of which has the properties of a minimum phase signal.
5) Compute the minimum phase group delay function of the windowed causal sequence of $\tilde{e}^i[n]$ and call it as $\tilde{E}_{gd}[k]$. The size of the window (i.e., cepstral lifter) is denoted by N_c.

6) The location of the positive peaks in the minimum phase group delay function $\tilde{E}_{gd}[k]$ approximately correspond to syllable boundaries.

The parameter N_c (length of the cepstral lifter) determines the resolution of the boundaries in the speech signal.

$$N_c = \frac{\text{Length of the energy function}}{\text{Window scale factor}} \quad (1)$$

where window scale factor (WSF) is an integer > 1. Note that N_c is inversely proportional to WSF. If N_c is high, the resolution will be high and two very closely spaced boundaries can be resolved. If its too high, a boundary will appear between CV/CVC at the CV transition. Thus, the choice of WSF and in turn N_c depends on the syllable rate.

Figure 1 shows the segmentation of 3 different utterances of the text "*She had your dark suit in greasy wash water all year*" at three syllable rates 2.5, 4 and 5.5 syl/sec. WSF is fixed at 8. Observe that segmentation is correct only for the 4 syl/sec utterance, while over-segmentation occurs for 2.5 syl/sec utterance and under-segmentation occurs for 5.5 syl/sec utterance.

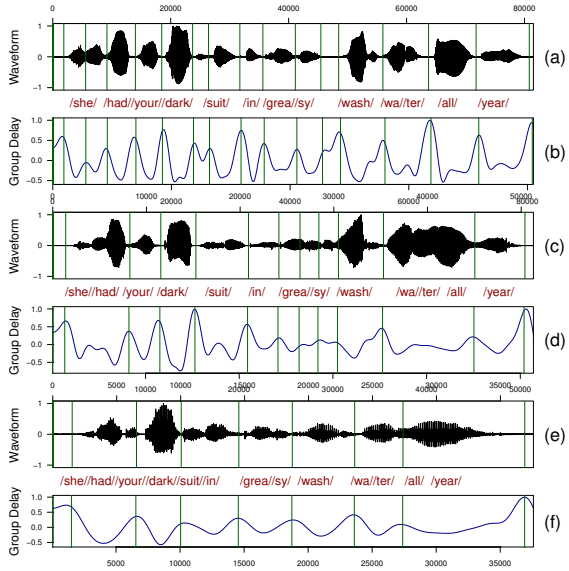

Fig. 1. Group delay plots for three utterances of the text "*She had your dark suit in greasy wash water all year*". (a) Utterance at 2.5 syl/sec (b) GD Plot for above utterance at WSF = 8 (c) Utterance at 4 syl/sec (d) GD Plot for above utterance at WSF = 8 (e) Utterance at 5.5 syl/sec (f) GD Plot for above utterance at WSF = 8

III. Robust Syllable Segmentation

Although the term 'syllable rate' applies to an entire speech utterance, considerable local variation in the utterance rate is present. Thus, a constant WSF for an entire utterance doesn't resolve all the boundaries.

As a first step, we seek to obtain a generalized variation of WSF with syllable rate. For this, the syllable rates in all corpora under consideration are first analyzed. The transcriptions available along with the utterances are used to estimate the syllable-rate. Next, the WSF that best segmented each utterance across the entire dataset is obtained. Eventually, bins of a constant WSF for a range of syllable rates is observed. This results in a syllable rate-WSF lookup table.

We now discuss two methods to estimate a syllable rate that can be used with the lookup to pick a WSF.

In the first method, the property that the energy is low at syllable-onset and coda is used to estimate the syllable rate (see Section III-A1). In the second method, the property that the syllable nucleus consists of a *single* vowel is used(see Section III-A2).

A. Syllable rate estimation

1) Lower Energy Threshold (LET) and Upper Energy Threshold (UET): Extensive analysis of the training data revealed that the average STE of speech utterances was correlated to *empirically defined* lower and upper STE thresholds. The syllable count is obtained by measuring the number of times the STE passes from the LET to the UET. Using STE directly for segmentation suffers as indicated in [11]. It is however sufficient to obtain an approximate estimate of syllable rate. To obtain the syllable rate, the following algorithm is employed.

LET and UET were estimated using the original utterance, the low-pass filtered utterance, the bandpass filtered utterance and the transcription. The text transcription from the training data was first syllabified. Using this information, the average STE and the corresponding LET and UET that estimated the correct syllable rate were obtained.

For each utterance in the test data,

1) Compute the average STE of the original, low-pass filtered and bandpass filtered utterance.
2) Use it as an index and obtain LET and UET for each case.
3) Measure how many times the STE of the test utterance goes from LET to UET for each case and get syllable count estimate.

This is divided by the duration to get the syllable rate.

2) VOP detection: A vowel detection process is run on an entire test utterance and the number of VOP's in the utterance gives us an estimate of the syllable count. The details of how vowels are detected is described in the next section.

B. Proposed Approach: Vowel Onset Point(VOP) detection

In spoken English, over 80% of the syllables are of the canonical CV, CVC, VC and V forms. The syllable nucleus is typically a sonorant, usually a vowel sound. Thus, if a segment represents a unique syllable, it is expected to have a single instant at which the onset of the vowel takes place. This quality can prove very useful in refining the segmentation process. That is, if a waveform is segmented with a much higher resolution than what is needed in the preliminary step, local variations that need the highest resolution are resolved but additional boundaries tend to show up at the CV transition between CV/CVC. A VOP detection can then be performed on the segments and those that reveal single VOP's are correctly segmented. Those that do not reveal any VOP's can be merged

with their neighbor, implying that they were the result of a split through a syllable and merging them with a neighbor results in a complete syllable.

Vowels associate with distinct vocal tract shapes that manifest in the spectrum of the speech signal as peaks. Thus, by picking some of the largest peaks in the spectrum, the amplitudes of the formants may be estimated. The VOP's are obtained as follows [9]. The sum of ten largest peaks in the first half of a 256-point DFT of a speech utterance is plotted as a function of time. This is used as a representation of energy of spectral peaks. This is then enhanced and the VOP's are detected as peaks in the VOP Evidence Plot. Figure 2 demonstrates an example.

First, a value of WSF that gives high resolution is picked and segmentation is performed. This will result in spurious boundaries in addition to the correct boundaries [1] (See Fig 2 (a)). Next we detect the VOPs for the segments obtained using group delay segmentation (See Fig 2 (c)). Segments corresponding to a unique VOP are treated as a single syllable segment while segments with no VOP's are merged with their neighbor. Some segments showing excess of one VOP are re-segmented (see Fig 2(e)).

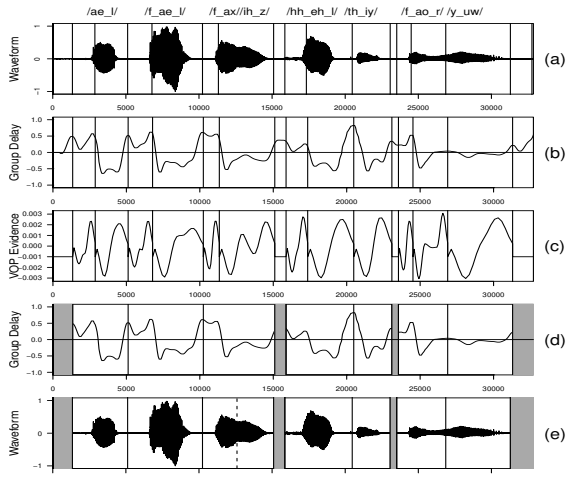

Fig. 2. segmentation using VOP detection (a) Segmentation with high resolution (b) Group delay function for above segmentation (c) VOP Evidence Plot (d) Group delay function after grouping (e) Final set of segments

IV. Segmented Syllable Based Speech Recognizer

A. Training

Speech is segmented into syllable like units using the algorithm described above. The corresponding text is segmented using rules. By mapping the segmented speech and text, syllable level annotations are obtained for the training data. Different examples of each syllable are used to build isolated-style syllable-HMMs. The block diagram for the training procedure is shown in Figure 3.

[1]It was observed in [11] that group delay based segmentation algorithm does not misplace boundaries

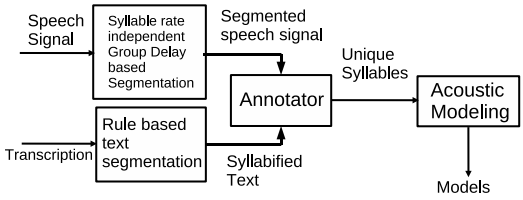

Fig. 3. Block diagram of training procedure

B. Testing

During testing, the utterances are segmented using the modified algorithm. The important difference between conventional recognizer and the proposed system is the use of segmentation in the linguistic framework. Conventional recognizers use the language information to derive the word output from the recognizer. The language models are specified as grammar or $\mathcal{N}$-gram language models. Here too, language models are used in a syllable-recognition framework *albeit with a difference*.

Fig. 4. A sample utterance

Consider the test utterance shown in Figure 4 with marked syllable boundaries. Figure 5 explains the search in the context of a conventional system versus that in the proposed system.

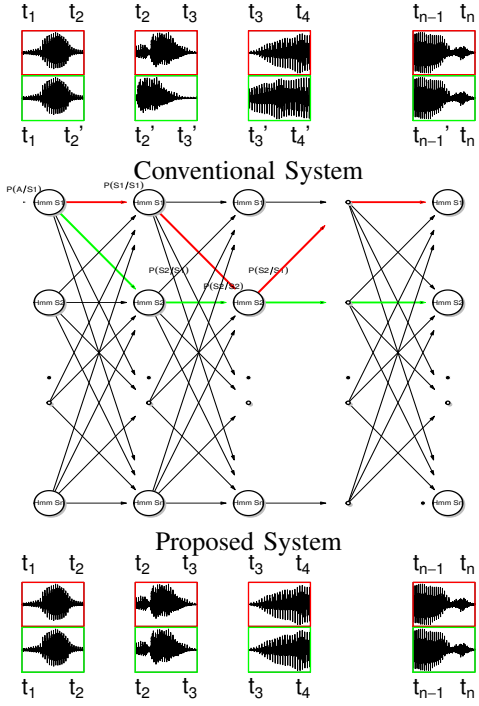

Fig. 5. Viterbi decoding in a conventional system vs. Viterbi decoding of the proposed system

Let Hyp I and Hyp II be two hypotheses that are generated

by the language model. Let the paths colored in red and green represent the paths taken by the Viterbi decoder for each of the hypotheses, respectively. The waveforms on the top and bottom show the segmentation of the waveform that may result during the decoding process. The segment boundaries obtained can be different (in the figure, this is exaggerated *only* to make a point). This is because different HMMs are active at the same time for the two different hypotheses. Language models are accessed whenever an HMM reaches the final state. This is required to prune the Viterbi Beam search. Clearly in this example, the language models are accessed at time t_2, t_3, t_4, t_{n-1} for Hyp I and at times $t_2^{'}$, $t_3^{'}$, $t_4^{'}$, $t_{n-1}^{'}$ for Hyp II. In principle, in Traditional Language Modeling(TLM), the language model can be accessed at the rate at which features are generated. The proposed approach is again illustrated in the same example. Figure 5 shows the incorporation of the acoustic information into the language modeling framework. Viterbi decoding is again employed here. The difference is that because the segmentation is supplied to the recognizer, it needs to access the language model for Viterbi decoding only at segment boundaries.

To understand its effect on complexity, we discuss a sample utterance "*Salvation reconsidered*". Figure 6 reveals the number of active HMM states corresponding to each frame with and without using boundary information. In this example, the average number of active states corresponding to a conventional system is $\approx$ 15000 states/frame. The average number of active HMM states corresponding to the proposed system is a mere $\approx$ 6000 states/frame. The basic difference in the proposed approach is that the LM model is accessed at *fixed time instants* across *all* hypotheses.

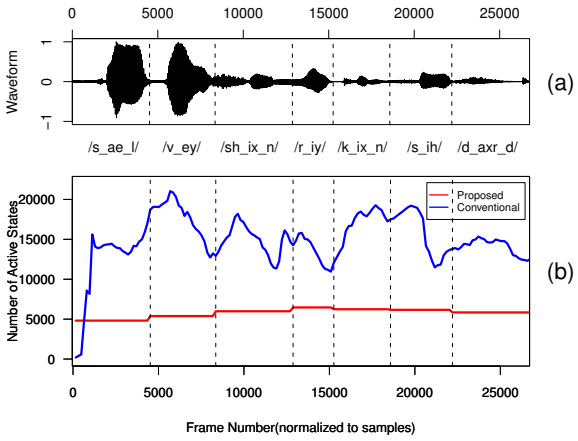

Fig. 6. Number of active HMM states per frame of proposed system vs. conventional system

V. Experiments

A. Conditions

For our experiments we chose the TIMIT[12] and NTIMIT[13] databases. They consist of phoneme level transcriptions. The dictionary is first syllabified using NIST syllabification software[14] available from NIST. The NIST syllabification software [14], is based on rules that define possible syllable-initial and syllable-final consonant clusters, as well as prohibited syllable-initial consonant clusters. The database contains 2 SA sentences per speaker which are same across all the 630 speakers. SA sentences are removed from both train and test databases as they introduce unfair bias. A total of 3570 unique syllables are present in the training data and 986 unique syllables in the core test set. Test syllables which are not in the training data are replaced with corresponding phonemes. Isolated style, continuous, left-to-right, 5 state, 3 mixture HMM's are built for each syllable.

B. Results

The syllable rates observed across the TIMIT and NTIMIT databases is as shown in the histogram in Figure 7. They vary from 1.5 syl/sec to 7.5 syl/sec and average at around 4.5 syl/sec. The correspondence between syllable rate and WSF obtained empirically is displayed in Table I.

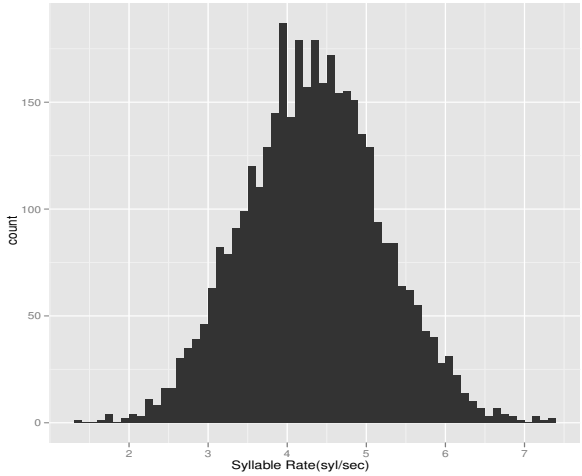

Fig. 7. Histogram of syllable rates observed across the TIMIT and NTIMIT databases

TABLE I
TABLE LOOKUP FOR SYLLABLE RATE-WSF CORRESPONDENCE

Rate(syl/s)	from	-	1.30	2.00	2.70
	to	1.30	2.00	2.70	3.30
	WSF	13	12	11	10
Rate(syl/s)	from	3.30	3.93	4.50	4.85
	to	3.93	4.50	4.85	5.95
	WSF	9	8	7	6
Rate(syl/s)	from	5.95	6.08	6.80	7.50
	to	6.08	6.80	7.50	-
	WSF	5	4	3	2

Figure 8 shows the scatter plots for the estimated syllable rate versus the transcribed syllable rate for the two methods discussed on the TIMIT and NTIMIT corpora. Observe the linear relationship on the scatter plots. The blue boxes represent the WSF bins and all points that fall within the boxes are segmented with the correct WSF. The performance of

segmentation using the fixed WSF method, the lookup based method and the VOP detection based method (Section III-B) is as shown in Table II.

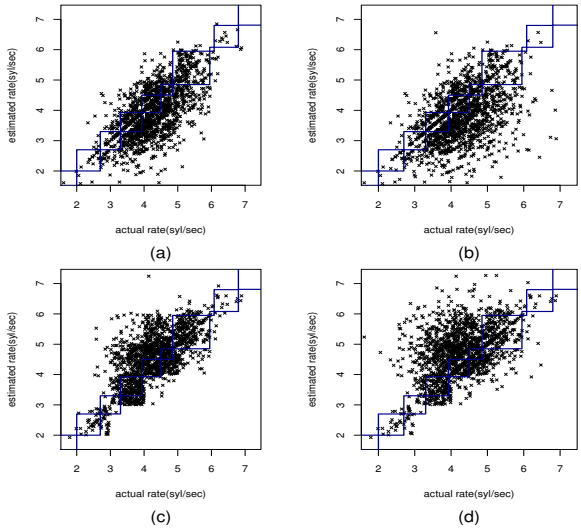

Fig. 8. Scatter plots of estimated syllable rate versus transcribed syllable rate for entire test set in (a) TIMIT using LET-UET thresholds (b) NTIMIT using LET-UET thresholds (c) TIMIT using VOP detection (d) NTIMIT using VOP detection. The blue boxes represent the WSF bins. Points falling within the bins are segmented with the correct WSF.

TABLE II
SYLLABLE SEGMENTATION ACCURACY USING VARIOUS METHODS

	Corpora	% FR	% FA
w/o syllable-rate info	TIMIT	21.37%	16.95%
with syllable-rate info	TIMIT	8.68%	4.74%
with VOP detection	TIMIT	6.91%	3.12%
w/o syllable-rate info	NTIMIT	33.12%	37.76%
with syllable-rate info	NTIMIT	12.57%	14.21%
with VOP detection	NTIMIT	10.46%	9.65%

The WER's obtained for the TIMIT and NTIMIT corpora are shown in Table III. . *2-level GD* refers to the method discussed in [8]. *Lookup Based GD* refers to our first method where the approximate syllable-rate is estimated and the WSF value is looked up. *VOP detection + GD* corresponds to the system discussed in Section III-B, where the syllable boundaries are verified using VOP detection. The performance of the 2-level GD based system is poor as the WSF value is *fixed* for the entire data. Using a lookup table performs fairly well but requires that syllable-rate of the utterance be estimated first. This is prone to errors with a basic syllable estimation method. The VOP+GD based segmentation approach gives significantly better results and doesn't require any such estimation. The flat-start recognizer does not use boundary information and significant difference in WER is observed between both systems.

VI. SUMMARY

In this paper, we have proposed a robust syllable segmentation technique that uses VOP detection in tandem with group delay segmentation. The modified algorithm is used in a syllable-centric continuous speech recognizer based on the TIMIT and NTIMIT corpora and significant improvement in WER's are observed. Additionally, using the segmentation information in the linguistic framework improved the performance of the recognizer.

TABLE III
WER'S ON THE TIMIT AND NTIMIT CORPORA FOR EACH TYPE OF SYSTEM

System	TIMIT %WER	NTIMIT %WER
2-level GD	42.3%	59.7%
Lookup based GD	5.2%	24.3%
VOP detection + GD	4.4%	21.2%
flat-start recognizer	13%	36%

REFERENCES

[1] Steven Greenberg, "On the origins of speech intelligibility in the real world," in *ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, Pont-a-Mousson*, 1997, pp. 23–32.

[2] D.W. Massaro, "Perceptual units in speech recognition," in *Journal of Experimental Psychology*, 1974, vol. 102, pp. 349–353.

[3] Abhinav Sethy and Shrikanth Narayanan, "A syllable based approach for improved recognition of spoken names," in *Proceedings of the ISCA Pronunciation Modeling Workshop, Estes Park*, 2002, pp. 30–35.

[4] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G.R. Doddington, "Syllable-based large vocabulary continuous speech recognition," in *IEEE Transactions on Speech and Audio Processing*, 2001, vol. 9,no. 4, pp. 358–366.

[5] Guillaume Gravier and Daniel Moraru, "Towards phonetically-driven hidden markov models: Can we incorporate phonetic landmarks in hmm-based asr?," vol. 4885/2007, pp. 161–168, 2007.

[6] Wu. Su-Lin, M. L. Shire, S. Greenberg, and N. Morgan, "Integrating syllable boundary information into speech recognition," in *In ICASSP*, Munich, april 1997, vol. 2, pp. 987–990.

[7] John Sirigos, Nikos Fakotakis, and George Kokkinakis, "A hybrid syllable recognition system based on vowel spotting," in *Speech Communication*, 2002, vol. 38,Issues 3-4, pp. 427–440.

[8] A. Lakshmi and Hema A. Murthy, "A syllable based continuous speech recognizer for tamil," *In Interspeech*, pp. 1878–1881, 2006.

[9] S. R. Mahadeva Prasanna, B. V. Sandeep Reddy, and P. Krishnamoorthy, "Vowel onset point detection using source, spectral peaks and modulation spectrum energies," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, May 2009.

[10] V. Kamakshi Prasad, T. Nagarajan, and Hema A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communication*, vol. 42, pp. 429–446, 2004.

[11] V. K. Prasad, *Segmentation and Recognition of Continuous Speech*, PhD dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Madras, India, May 2002.

[12] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgrena, and Victor Zue, "TIMIT acoustic-phonetic continuous speech corpus. linguistic data consortium," 1993.

[13] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephonebandwidth speech database," *In ICASSP*, pp. 109–112, April 1990.

[14] M. Fisher, *Syllabification Software*, The Spoken Natural Language Processing Group, National Institute of Standards and Technology, Gaithersburg, Maryland, U.S.A.