CONTINUOUS SPEECH RECOGNITION USING AUTOMATICALLY SEGMENTED DATA AT SYLLABIC UNITS

V. Kamakshi Prasad *T. Nagarajan* *Hema A. Murthy*
Indian Institute of Technology, Madras, Chennai - 600 036
e-mail : hema@lantana.iitm.ernet.in

Abstract: We propose an alternative approach for continuous speech recognition where the segmentation and recognition tasks are separated. Syllable is considered as a unit for both segmentation and recognition. Using minimum phase group delay function based approach, the speech signal is segmented at boundaries of syllabic units and a syllable based isolated style HMM recognition system has been implemented for two Indian languages. To address the errors in recognition due to shift in segment boundaries and merger of syllabic units, Viterbi algorithm based approaches are proposed.

1 Introduction

Speech does not contain any reliable demarcation at the boundaries of sub-word units. Further the articulatory configuration of neighbouring phonetic units affect the articulation of a phonetic unit which makes it difficult to segment the speech signal into discrete phonetic units.

Present day continuous speech recognition systems consider phoneme as a basic unit and use syntactic and semantic rules of the language. Such recognition systems are primarily language or task dependent. The phonemes are extremely context sensitive. Hence is not a good choice as the basis for speech models for recognition.

Over the last decade, researchers have started using larger units than phonemes in speech recognition. For example in [1], a character spotting approach to speech recognition is employed. Syllable is an intuitive unit for representation as the variation observed is more systematic at the level of the syllable than at the level of phoneme [2]. Syllable is acoustically and perceptually more stable unit and variations at syllable level are more systematic. Hence in our work, syllable is considered as a basic unit for segmentation and recognition. Using minimum phase group delay function, the speech signal is segmented at boundaries of syllabic units. The segments thus obtained, are evaluated against the hidden Markov models (HMMs) trained apriori, for each of the syllables. Advantage of such a recognition system is, it is task and language independent and errors found in recognition if any, their effect is localized.

In Section 2, motivation for segmenting the speech signal at syllabic boundaries is presented and in Section 3, a segmentation approach is presented to segment the continuous speech signal. In Section 4, the proposed continuous speech recognition system is discussed and the conclusions are presented in Section 5.

2 Motivation for segmentation of continuous speech

To determine whether an isolated style recognition system trained using the database excised from continuous speech would suffice, the following experiment was performed on the speech database collected for two Indian languages [3]. The Indian languages are syllable centric in that acoustic variations are more systematic at syllable boundaries.

Hidden Markov models (HMMs) for each syllable were trained using the manually segmented syllable database excised from the continuous speech signal and tested against the syllabic segments which were again manually segmented. Database used for training and testing are different. The recognition performance of the manually segmented database using HMM based system is shown in Table 1. The average recognition performance using 10-best criteria is as high as 94%. Hence, it can be concluded that for the development of the continuous speech recognition system, if the speech signal can be segmented at syllable boundaries by an auto-

0-7803-7488-6/02/$17.00 © 2002 IEEE.

matic segmentation algorithm, a simple isolated style syllable recognition system can be implemented to recognize each syllable. An advantage of such an implementation is, there are no intensive dynamic programming based computations and error if any, its effect is localized.

3 Segmentation of speech signal using minimum phase group delay function

The group delay function derived from the minimum phase signal is called "minimum phase group delay function". The minimum phase group delay function has the property that it resolves the poles and zeroes where poles correspond to peaks and zeroes correspond to valleys. Non-minimum phase signals do not posses this property. This is the primary motivation for converting any non-minimum phase signal to a minimum phase signal.

Table 1. Recognition performance on manually segmented data of one complete news bulletin of one Indian language (Telugu) of duration 15 minutes. (The recognition system consists of 244 syllable models).

n-best Result	1	2	3	4	5	6	10
All Syllables recognition (%)	65	78	84	87	89	91	94

3.1 Segmentation of an arbitrary positive function

In [4], it was shown that a minimum phase signal can be derived from any magnitude spectrum. Further it was shown that any arbitrary positive function symmetrised about y-axis, can be considered as a magnitude spectrum and a minimum phase signal can be derived from the same. Combining the properties of group delay function and minimum phase signal, the valleys in the positive function can be obtained. These valleys may also be considered as segmentation point. The procedure is summarised as follows:

- Let $e(n)$ be an arbitrary positive function.
- Construct the symmetric part of the sequence by producing a lateral inversion of this sequence about the Y-axis. Let this sequence be $E(k)$. This may be viewed as an arbitrary magnitude spectrum.
- Compute $1/E(k)$ [1]. Let it be denoted as $\tilde{E}(k)$.
- Compute the inverse DFT of the sequence $\tilde{E}(k)$. This resultant sequence $\tilde{e}(m)$, is the root cepstrum and the causal portion of it resembles the properties of the minimum phase signal [5].
- Compute the minimum phase group delay function of the windowed causal sequence of $\tilde{e}(m)$ [5].
- The location of peaks in the minimum phase group delay function correspond to valleys of the given arbitrary positive function $e(n)$.

3.2 Processing short term energy function for segmentation

In Section 3.1, it was shown that significant events, namely, location of valleys for any arbitrary positive function can be obtained using the group delay function. Since the short term energy function is a positive function, it can also be processed in a manner similar to that of processing an arbitrary magnitude spectrum.

Fig. 1(b) shows the short term energy function of the speech signal shown in Fig. 1(a) and the segment boundaries for Fig.1(a) can be determined by extracting the valleys in the short term energy function. If the duration of the utterance is denoted by T, then in Fig. 1(b), the range $(0 - T)$ is replaced by $(0 - \pi)$ and symmetrised about $Y - axis$ to resemble the magnitude spectrum. Using the procedure explained in Section 3.1, the boundaries at syllabic segments are computed. These boundaries are shown in Fig. 1(c) with thick vertical lines which pass through the peaks in minimum phase group delay function.

4 Continuous speech recognition using segmented data

HMM based recognition system has been implemented for continuous speech recognition. Models are trained using the manually segmented database. Separate models are trained for each of the syllables. Although there exist thousands of distinct syllables in any language, most frequently occurred

[1] Positive function is inverted to overcome the error in location of valleys due to truncation in the cepstral domain

syllables are only few hundreds. Specifically, in Telugu language, there are only 244 syllables with frequency of occurrence greater than or equal to 50 out of 2450 distinct syllables in the entire database [3] of four hours and comprising of 98% of the total syllables present in the entire database. Hence even if the models are trained for only those syllables which contain atleast 50 utterances, there will not be much degradation in the performance.

4.1 Syllable recognition using HMMs

A separate 5 state left to right HMM with 3 mixtures per state has been trained for each of the syllables using the feature vectors obtained from the manually segmented database excised from Telugu utterances of different lengths. Syllable segments excised from continuous speech signal contain heterogeneous spectral content at both ends due to co-articulation present at both ends of the segment. Hence the system trained using syllable segments captures the variability at both the ends and becomes robust against different phonetic contexts. Mel-frequency cepstral coefficients (MFCCs) are extracted from the speech signal using a window size of 25 ms and a frame shift of 10 ms.

4.2 Performance evaluation

The test utterance is segmented using the approach described in Section 3 and each segment is tested against the syllable HMMs to find the corresponding HMM with maximum likelihood value. Fig.1 demonstrates the segmentation of one speech utterance from Telugu news bulletin. Thick vertical lines in Fig.1(c) which pass through the peaks of the minimum phase group delay function, denote the segment boundaries obtained by the automatic segmentation algorithm and dotted vertical lines denote the locations corresponding to manual segment boundaries of the speech signal. The transcription below Fig.1(c) denotes the recognised syllable for the corresponding segment. Spurious peaks and peaks correspond to actual boundaries below the threshold, are denoted respectively with symbols "N" and "Y". These symbols are indicated on top of Fig.1(c).

The recognition performance of one complete news bulletin of duration 15 minutes is shown in (Table 2, column 2). The news bulletin was divided into 371 phrases for convenience and duration of each phrase is around 3 seconds. The length of the continuous speech signal used

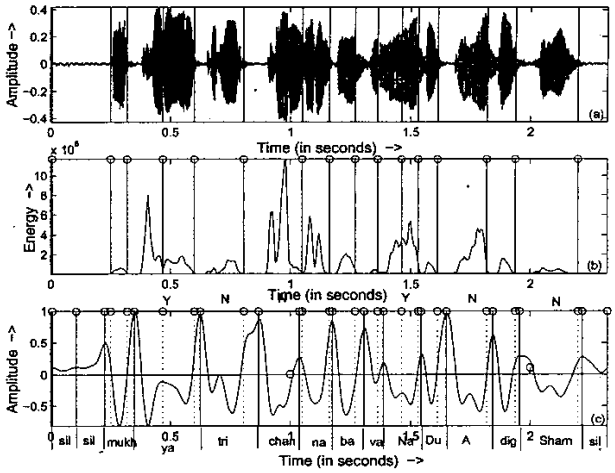

Fig. 1. a) Speech utterance "mukhya mantri chandrababu naayudu aadesham" b) Short term energy function c) minimum phase group delay function with segment boundaries.

for testing varies very widely, ranging from 5 syllables to 20 syllables. It is observed that if the desired syllable is not in the first position, in most of the instances, it is one among the first few alternatives. For syllables with short and long vowels separate models have been trained. As a result, the syllable with short vowel sound are some times recognised as syllable with long vowel and vice versa (eg. ka recognised as kA). A similar observation is found with syllables with nasal stop consonants of velar ($\tilde{N}$), alveolar (N) and dental (n) classes, although separate models are trained for each of these syllables. Such recognitions are categorised as 'similar syllables' in Table 2. In some cases only vowel part of the syllable is recognised correctly (eg. ka recognised as ta) and in some cases only consonant part of the syllable is recognised correctly (eg. ka recognised as ki). Such findings are also listed in Table 2. To detect the peaks, which correspond to actual boundaries, below the threshold, we propose a method as explained in the following subsection.

4.3 Processing missed segments

At every instance where a peak is below the threshold two alternative recognitions are formulated. The first alternative corresponds to the recognized syllable when the peak below the threshold is not considered. The second alternative

Table 2. Recognition performance of syllable segments. (I) Isolated style recognition, (II) Processing the peaks below the threshold also, and (III) Using local optimisation

Category	(I)	(II)	(III)
Whole syllable	39.6	40.8	46.1
Similar syllables	12.8	13.2	16.2
Only vowel part	5.1	5.6	5.8
Only consonant part	3.2	3.4	3.9
Total	60.7	63.0	72.0

corresponds to the case when the peak below the threshold is also considered as a segmentation point. Now the problem is transformed into a problem of finding the maximum likelihood value using Viterbi algorithm when two alternative syllable strings are given where the first syllable string consists of one syllable and the second consists of two syllables. If the maximum likelihood value is associated with the second string, the peak below the threshold is also considered as a segmentation point. Improvement in the performance after processing for the missed segments is shown in (Table 2, column 3). To address the error due to co-articulation, the approach as explained in the following subsection is used.

4.4 Local optimisation

It is found that the segment boundaries obtained by the proposed algorithm, in few cases, is shifted by few milliseconds. As a result the desired basic unit is likely to be one in the $n - best$ syllables. Hence if the decoder generates n most likely alternative basic unit outputs for each segment, the language model could re-score the sentences according to the grammar or semantics. But in a language independent recognition system as no syntactic or semantic knowledge is available, it is assumed that any basic unit can follow any other basic unit. Hence by enumerating the different permutations possible with n basic units at each position, a list of strings which are sequences of basic units, is found. As the desired syllable string is more likely to be one among this list, the objective now is to find the most appropriate string of basic units.

Now finding the string of phonetic units with maximum likelihood is transformed into a problem of finding optimal state sequence, where each state represents one syllable HMM. The output state sequence will correspond to only one of the strings that can be obtained by a concatenation of the syllabic unit from the given vocabulary. The results are shown in (Table 2, column 4). A similar performance was found for Tamil database.

5 Conclusions

A new approach for continuous speech recognition is proposed. The speech signal is segmented at syllable boundaries and HMM based isolated style recognition system is implemented. To address the errors in segmentation and recognition, Viterbi algorithm based approaches are proposed. The approach proposed in this paper is more suitable for the implementation of task independent speech recognition system.

6 References

[1] P. Eswar, C. Chandra Sekhar, S. K. Gupta, B. Yegnanarayana, and K. Nagamma Reddy, "An acoustic-phonetic expert for analysis and processing of continuous speech in Hindi," in *Proc. European Conference on Speech Technology*, (Edinburgh), pp. 369–372, 1987.

[2] Steven Greenberg, "Speaking in short hand - a syllable-centric perspective for understanding pronunciation variation," *Speech Communications*, vol. 29, pp. 159–176, 1999.

[3] *Database for Indian languages*. Speech and Vision Lab, IIT Madras, Chennai: India, 2001.

[4] T. Nagarajan, V. Kamakshi Prasad, and Hema A. Murthy, "Minimum phase signal derived from the magnitude spectrum and its application to speech segmentation," in *6th biennial conference proceedings on speech communications*, (IISc, Bangalore, India), pp. 95–101, July 2001.

[5] Hema A. Murthy and B. Yegnanarayana, "Formant extraction from minimum phase group delay function," in *Speech Comm.*, vol. 10, pp. 209–221, August 1991.