AUTOMATIC SEGMENTATION AND LABELING OF CONTINUOUS SPEECH WITHOUT BOOTSTRAPPING

Hema A.Murthy, T.Nagarajan and N.Hemalatha

Dept. of Computer Science and Engineering
Indian Institute of Technology, Madras
email : hema@lantana.iitm.ernet.in

ABSTRACT

In this paper, a novel approach is proposed for automatically segmenting and transcribing continuous speech signal without the use of manually annotated speech corpora. The continuous speech signal is first segmented into syllable-like units by considering short-term energy as a magnitude spectrum of some arbitrary signal. Similar syllable segments are then grouped together using an unsupervised incremental clustering technique. Separate models are generated for each cluster of syllable segments. At this stage, labels are assigned for each group of syllable segments manually. The syllable models of these clusters are then used to transcribe/recognize the continuous speech signal of closed-set speakers as well open-set speakers. As a syllable recognizer, our initial results on Indian television news bulletins of the the languages Tamil and Telugu shows that the performance is 43.3% and 32.9% respectively.

1. INTRODUCTION

The last decade has witnessed substantial progress in speech recognition technology, with todays state-of-art systems being able to transcribe unrestricted broadcast news speech data with good accuracy. However, acoustic model development for these recognizers relies on the availability of large amounts of manually transcribed training data. Obtaining such data is both time-consuming and expensive, requiring trained human annotators and substantial amounts of supervision.

Researchers have tried several ways to reduce the manual effort and the expenditure of transcribing speech data for developing continuous speech recognizers with comparable accuracy. The most commonly used approach is bootstrapping. To increase the amount of training data, an existing speech recognizer is used to transcribe new data. This newly transcribed data is then used to retrain the model parameters. For a new language, if no such speech recognizer is available, few hours of data should be manually transcribed and this is used to build a recognizer. Which, in turn, is again used to increase the amount of training data by transcribing larger quantities of untranscribed data.

In [1], a low-cost recognizer trained with one hour of manually transcribed speech is used to recognize 72 hours of untranscribed acoustic data. These transcriptions are then used to train an improved recognizer. [2] uses an automatic approach to segmentation of labeled speech and labeling and segmentation of speech when only the orthographic transcription of speech is available. Lamel [3] has shown that the acoustic models can be initialized with as little as 10minutes of manually annotated data. The basic idea behind all the above mentioned works, is to use an existing speech recognizer to transcribe huge amount of untranscribed data, which can further be used to refine the trained models.

There are two basic problems which can be expected in the above mentioned methods. Firstly, if there is a mismatch in environment or language during transcription, the performance is expected to be very poor. If this newly transcribed data is going to be used for further refining the model parameters, convergence of models will be very slow and in some cases, convergence may be impossible. The immediate alternative to this problem is, manually transcribing part of the new data, building models using this and then transcribe the rest of the data. For transcribing even a small amount of data, we need trained human annotators. So, there is a need for a system to transcribe continuous speech signal without the benefit of a manually annotated speech corpus.

An automatic procedure for phonetic transcription of spontaneous speech has been developed in [4] which does not require transcription. In [4], articulatory-acoustic phonetic features are extracted from each frame of the speech signal and classification of phones is done by special purpose neural-networks. The output of these networks is processed by Viterbi-like decoder to produce a sequence of phonetic-segment labels along with boundary demarcations associated with each segment.

In this paper, we propose a novel approach for automatic segmentation and transcription of speech data without using any manually annotated corpus. The speech signal to be transcribed is first segmented into syllable-like units using an algorithm developed in our laboratory [5] which primarily uses the short-term energy function but after group delay based processing. Similar syllable segments are then grouped together using an unsupervised incremental clustering technique. Separate models are generated for each cluster of syllable segments. At this stage, a small amount of manual work is required to give an identity for each group of syllable segments. The syllable models of these clusters are then used to transcribe/recognize the continuous speech signal.

The remainder of this paper is as follows. In Section. 2, we briefly discuss the group delay based segmentation procedure which is adopted to segment the speech signal into syllable-like units followed by the proposed approach for automatic transcription of speech data in detail. In Section. 3, the performance of this approach is analyzed and criticized.

2. AUTOMATIC SEGMENTATION FOLLOWED BY LABELING

2.1. Speech corpus

For both training and testing, Indian television news bulletins of the languages Tamil and Telugu have been used [6]. During training, from both the languages, 4 speakers data, each of 15mins duration are used. During testing, 2 news bulletins of the languages Tamil and Telugu are used. The training and testing set speakers are different. Here, the total duration of the speech signal of each news bulletin is split into small segments of approximately 2.5s each.

2.2. Syllable-like segmentation

The syllable is structurally divisible into three parts, the onset, nucleus, and coda [4]. Although many syllables contain all the three elements, say **CVC**, a significant portion contain one element typically, **V** or two elements **CV or VC**. In [7], we have proposed a method for segmenting the acoustic signal into syllable-like units with various refinements [5]. Using this approach, four speakers speech data are segmented into syllable-like units, which gives *M* syllable segments, $s_1, s_2, ..., s_M$ (*M* = 8000). These syllable segments are used during the training process (Fig.1). The training process is similar to conventional clustering technique but instead of clustering the feature vectors at frame level, it is done at syllable segment level using the the method described in the following section.

2.3. Initial cluster selection

For any iterative training process, the assumed initial condition is crucial for the speed of convergence. After having all the syllable segments, the first task is to select some unique syllable segments or groups of unique syllable segments for training. The initial groups of syllable segments should be carefully selected to ensure fast convergence. At the initial stage itself, if the selected group of syllable segments are unique, the convergence may be accelerated during iterative training. For selecting such initial clusters, the following procedure is adopted.

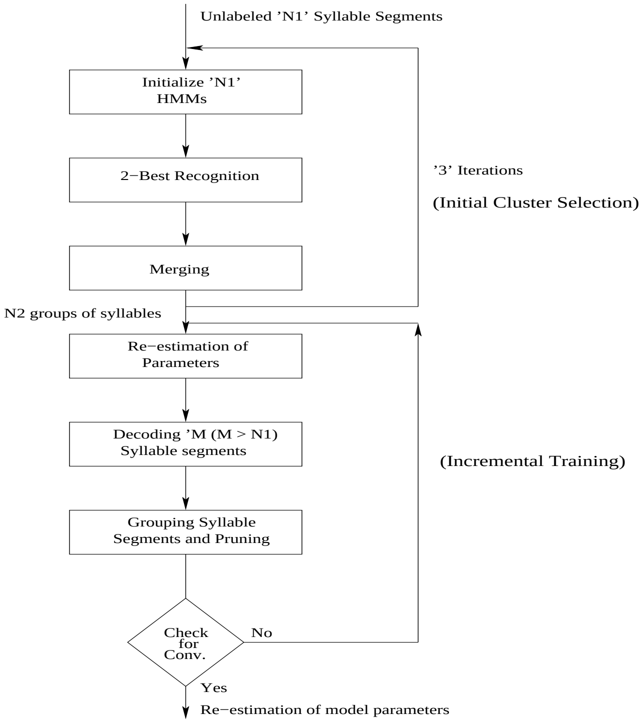
Unlabeled 'N1' Syllable Segments
Initialize 'N1' HMMs
2−Best Recognition
Merging
'3' Iterations
(Initial Cluster Selection)
N2 groups of syllables
Re−estimation of Parameters
Decoding 'M (M > N1) Syllable segments
(Incremental Training)
Grouping Syllable Segments and Pruning
Check for Conv.
No
Yes
Re−estimation of model parameters

Fig. 1. *Flow chart : Unsupervised Incremental HMM*

1. From the M syllable segments, first N1 syllable segments $s_1, s_2, ...s_{N1}$, where $N1 < M$, are taken for initialization.
2. Features (13 dimensioanl MFCC + 13 delta + 13 acceleration) are extracted from these N1 syllable segments with multiple resolutions (ie., with different window sizes and frame shifts). Multi-resolution feature extraction ensures a reasonable variance for each Gaussian mixture in the models.
3. N1 *Hidden Markov Models* $(\lambda_{s1}, \lambda_{s2}, ..., \lambda_{sN1})$ are initialized. To initialize model parameters, the Viterbi algorithm is used to find the most likely state sequence corresponding to each of the training examples (here, each of the feature vectors derived from the same syl-

lable segment but with different resolutions), then the HMM parameters are estimated.

4. Using Viterbi decoding process, the same N1 syllable segments are decoded using 2-best criteria, resulting in $N1$ pairs of syllable segments $(p_1, p_2, ..., p_{N1})$.

$$p_i = [arg \max_{1 \leq i \leq N1}^{1} P(O/\lambda_i) \quad arg \max_{1 \leq i \leq N1}^{2} P(O/\lambda_i)] \tag{1}$$

where,
- p_i is the i^{th} pair of syllable segments (where $1 \leq i \leq N_1$)
- $P(O/\lambda_i)$ is the probability of the observation sequence O $(o_1 o_2 ... o_n)$ for the given model λ_i
- max^1 and max^2 denotes the 1-best and 2-best results respectively.
Interestingly, in almost all the N1 cases, both 1-best and 2-best syllable segments are identical/similar. This step gives N1 pairs of syllable segments.

5. Among N1 pairs $(p_1, p_2, ..., p_{N1})$, if a syllable segment is found to be repeated in more than one pair, the other pairs are removed, and the number of models is thus pruned.
6. New models are created with these reduced number of pairs.
7. Steps 4-6 are repeated for **m** times (here, m = 3). After **m** iterations, each cluster will have 2^m syllable segments grouped together.

This initial cluster selection procedure will lead to $N2$ clusters $(c_1, c_2, ..., c_{N2})$. In the next step, the model parameters are re-estimated incrementally.

2.4. Incremental training

After selecting the initial clusters $(c_1, c_2, ..., c_{N2})$, where the models are only initialized, the parameters of the models of each of the clusters is re-estimated using Baum-Welch re-estimation procedure. This training procedure is referred to as **incremental training**. It is considered incremental because the HMM parameters are adjusted before all the training data has been considered. The training strategy must be contrasted with conventional batch training where the models are updated only after all the data in the training set are processed. The steps followed for this incremental training are given below.

1. The model parameters of the initial clusters $(c_1, c_2, ..., c_{N2})$ derived from the previous step are re-estimated using Baum-Welch re-estimation. Each model is a 5 state 3 Gaussian mixtures/state HMMs.
2. The new models are used to decode all the syllable segments $(s_1, s_2, ..., s_M)$ using Viterbi decoding.
3. Clustering is done based on the decoded sequence.
4. If a particular cluster is found to have less than ϵ (Here,$\epsilon = 3$) syllable segments, that cluster is removed and number of models is reduced by one.
5. Steps 1-3 are repeated until convergence is met.

The convergence criteria followed in this approach is explained below.

2.5. Convergence criteria

In each iteration as the model parameters are re-estimated and the syllable segments are re-clustered, the number of syllable segments which migrate from one cluster to another is expected to reduce at each iteration. The convergence criteria followed for the incremental training is based on 'number of migrations between clusters'. The convergence is said to be met if the number of migrations between clusters reaches zero. When this condition is met, the incremental training procedure terminates. This incremental training process will produce $N3$ syllable clusters $(c_1, c_2, ..., c_{N3})$, and in turn $N3$ syllable models $(\lambda_{s1}, \lambda_{s2}, ..., \lambda_{sN3})$. Interestingly, the syllable segments in each cluster now are found to be identical/similar in almost all the clusters, with few exceptions. And the models for these new clusters can be used for labeling. But since the total training process is unsupervised, these clusters do not have any identity.

2.6. Labeling clusters and Transcription

For using the above derived syllable models for transcription/recognition tasks, it is required to assign a label for each of clusters. By manually listening to the syllable segments in each of the clusters, a proper label is assigned by listening to ONLY the syllable segments in each cluster ONCE. This may be contrasted with listening to continuous speech, identifying syllable boundaries and assigning labels manually.

Now, the models are ready with labels assigned to it and they can be used to transcribe/recognize speech data in a conventional way. For performance analysis, the following categories are identified (Fig.2). In Fig.2, the group delay function derived from the band-pass filtered speech signal alone is shown for reference (Fig.2b).

- Whole syllable (the syllable can be either **CVC or CV**) is correct (marked as *syl*).
- Vowel followed by consonant (**VC**) is correct, i.e., if any of the consonant is wrong (marked as *vc*).
- In the whole syllable, Vowel (**V**) alone is correct (marked as *v*).
- In the whole syllable, one of the consonants (**C**) alone is correct (marked as *c*).
- All the sound units in the syllable are wrongly recognized (marked as *W*).

3. PERFORMANCE ANALYSIS AND CRITICISMS

For performance analysis of our system for transcription, two closed-set and two open-set speakers data, from Tamil and Telugu, are considered. Using the group delay based segmentation approach, these speech data are segmented into syllable-like units. This gives $\approx$ 4000 syllable segments for each of the closed-set and open-set speakers data. As a syllable recognizer, for the closed-set and open-set speakers of Tamil (Telugu), the average performance is $\approx$ 49%(44%) and $\approx$ 43%(33%) respectively (Table.1). The preliminary results are quite promising. But, there are some noticeable problems with the clusters derived. Eventhough many of the cluster segments have identical syllable segments or at least similar syllable segments, very few clusters have syllable segments which are totally unrelated to each other. Another noticeable problem here is because of errors in speech segmentation. Problems with presence of semi-vowels in the middle of the word is, to the maximum extent, handled by the group delay function derived from the band-pass filtered speech signal ([5]). But in some cases, specifically for the semi-vowel **"r"** the segmentation is inaccurate. This results in two syllables in one segment. Post processing of segment boundaries and splitting the clusters which contain similar syllable segments can be taken up as future research to improve the recognition/transcription performance.

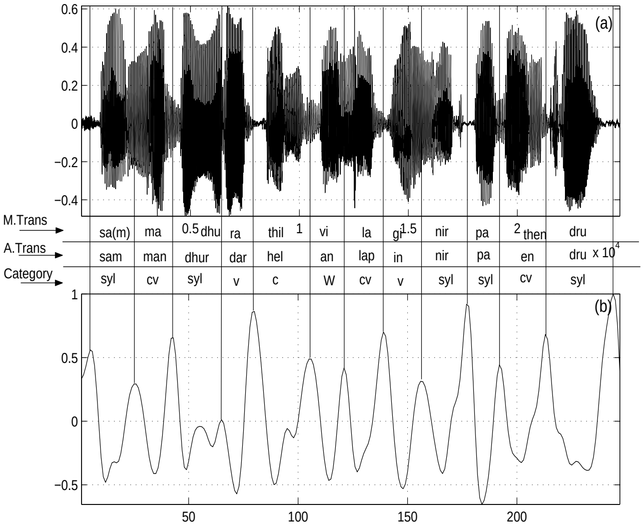

Fig. 2. *Performance analysis - an example. (a) The speech signal (b) group delay function derived from band-pass filtered speech signal. M.Trans – Manual transcription. A.Trans – Automatic transcription. Category – Type of unit recognized.*

Sound units	Tamil		Telugu	
	A	B	A	B
CVC + CV	49.24	43.33	43.6	32.9
VC	11.26	8.36	8.7	8.4
Vowel only	27.30	31.0	29.5	35.2
Consonant only	3.25	4.28	6.7	11.4
False	8.95	13.03	11.5	12.1

Table 1. Transcription/recognition performance (in %) . (A) Closed-set speakers data, (B) Open-set speakers data.

4. CONCLUSIONS

We have described a novel approach for segmenting and transcribing continuous speech signal without the benefit of a manually annotated speech corpus. We have shown that, as a open-set syllable recognizer, the performance of the proposed approach is 43.3% and 32.9% for the languages Tamil and Telugu respectively. The results shows that the proposed approach can be used as transcriber and as well as recognizer. The preliminary results are quite promising and are comparable in performance to the conventional batch training procedure, which uses manually annotated corpora.

5. REFERENCES

[1] Frank Wessel and Herman Ney, "Unsupervised training of acoustic models for large vocabulary continuous speech recognition," *IEEE workshop on ASRU*, pp. 307-310, Dec 2001.

[2] A.Ljolje and M.D.Riley, "Automatic segmentation and labeling of speech", *ICASSP-1991*, Vol.1, pp. 473-476, April 1991.

[3] Lori Lamel, Jean-Luc Gauvain and Gilles Adda, "Unsupervised acoustic model training", *ICASSP-2002*, Vol.1, pp. 877-880. May 2002.

[4] Shuangyu Chang, Lokendra Shastri and Steven Greenberg, "Automatic phonetic transcription of spontaneous speech (American English)", *ICSLP - 2000*.

[5] T.Nagarajan, Hema A.Murthy and Rajesh M.Hegde, "Segmentation speech into syllable-like units", *EUROSPEECH-2003*, pp. 2893-2896, Sep 2003.

[6] *Database for Indian Languages*, India, Speech and Vision Laboratory, IIT Madras, Chennai - 2001.

[7] V.Kamakshi Prasad, T.Nagarajan and Hema A.Murth, "Automatic segmentation of continuous speech using minimum phase group delay functions, *to appear in Speech Communications*