LANGUAGE IDENTIFICATION USING PARALLEL SYLLABLE-LIKE UNIT RECOGNITION

T.Nagarajan and Hema A.Murthy

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

ABSTRACT

Automatic spoken language identification (LID) is the task of identifying the language from a short utterance of the speech signal. The most successful approach to LID uses phone recognizers of several languages in parallel. The basic requirement to build Parallel Phone recognition (PPR) system is annotated corpora. In this paper, a novel approach is proposed for the LID task which uses parallel syllable-like unit recognizers, in a frame work similar to PPR approach in the literature. The difference is that unsupervised syllable models are built from the training data. The data is first segmented into syllable-like units. The syllable segments are then clustered using an incremental approach. This results in a set of syllable models for each language. Our initial results on OGI_MLTS corpora show that the performance is 69.5%. We further show that if only a subset of syllable models that are unique (in some sense), are considered, the performance improves to 75.9%.

1. INTRODUCTION

Language identification is the task of identifying the language from a short utterance of speech signal. Existing LID systems can be classified into two major categories, namely **Explicit and Implicit LID systems**, based on whether the system requires annotated corpora or not. Building annotated corpora for all the languages to be recognized, is both time consuming and expensive, requiring trained human annotators and substantial amount of supervision [1]. Therefore, eventhough the performance of the implicit LID systems are slightly inferior to that of explicit LID systems, unavailability of annotated corpora makes the implicit LID systems attractive.

One of the successful approaches for LID task is to use phone recognizers of several languages in parallel [2]. This approach requires annotated corpora of more than one language, although the annotated corpus need not be available for all the languages to be identified. In [3], a parallel subword recognition system for the LID task is proposed, in a frame work similar to the parallel phone recognition (PPR) approach in the literature [2], but without requiring annotated corpora.

Using phonemes as the basic sound unit for LID task may not be optimal in the sense that most of the phonemes are common between languages. In this case, the source of information that may be used for LID is the variation in the frequency of occurrence of the same phoneme in different languages. Only very few phonemes are unique for a particular language. If a longer sound unit, say syllable is used, then the number of unique syllables in any language is very high, which may be a potential information for discriminating languages. Kung-Pu Li [4] has shown that spectral features derived from syllabic units are reliable for distinguishing languages.

In this paper, by considering the syllable as a basic unit, we propose a parallel syllable-like unit recognition system for the LID task, in a frame work similar to the PPR approach in the literature [2], but without using annotated corpora.

The basic requirement for building syllable-like unit recognizers for all the languages to be identified, is an efficient segmentation algorithm. Earlier, we have proposed an algorithm [5], which segments the speech signal into syllable-like units. Recently, we have made several refinements [6] to improve the segmentation performance of the baseline algorithm [5]. Using this algorithm ([5] [6]) each language training utterances are first segmented into syllable-like units. Similar syllable segments are then grouped together and syllable models are trained incrementally. These language-dependent syllable models are then used for identifying the language of the unknown test utterances.

The remainder of this paper is organized as follows. In Section.2, we briefly review the segmentation procedure used to segment the speech signal into syllable-like units followed by a detailed description of the proposed approach for language identification. In Section.3, we analyze the performance of the proposed approach.

2. PARALLEL SYLLABLE-LIKE UNIT RECOGNITION

2.1. Syllable-like segmentation

The syllable is structurally divisible into three parts, the onset, nucleus, and coda [1]. 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 [6], we have proposed a method for segmenting the acoustic signal into syllable-like units with various refinements. Using this approach, all the training speech data of each language are segmented into syllable-like units, which gives $\mathcal{M}_l$ syllable segments, $\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_{\mathcal{M}_l}$ ($\mathcal{M}_l \approx 5000$) for the language $\mathcal{L}_l$. These syllable segments are used during the training process. 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 method described in the following Section.

2.2. Unsupervised Incremental HMMs

2.2.1. 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.

1. From the $\mathcal{M}_l$ syllable segments of language $\mathcal{L}_l$, a subset ($\mathcal{N}1$) syllable segments, $\mathcal{S}_1, \mathcal{S}_2, ...\mathcal{S}_{\mathcal{N}1}$, where $\mathcal{N}1 < \mathcal{M}_l$, are taken for initialization.
2. Features (13 dimensional MFCC + 13 delta + 13 acceleration coefficients, after Cepstral Mean Subtraction) are extracted from these $\mathcal{N}1$ 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. $\mathcal{N}1$ *Hidden Markov Models* ($\lambda_1, \lambda_2, ..., \lambda_{\mathcal{N}1}$) 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 syllable segment but with different resolutions), then the HMM parameters are estimated.
4. Using Viterbi decoding process, the same $\mathcal{N}1$ syllable segments are decoded using 2-best criteria, resulting in $\mathcal{N}1$ pairs of syllable segments ($p_1, p_2, ..., p_{\mathcal{N}1}$).

$$p_i = [arg \max^{1}_{1 \le i \le \mathcal{N}1} \mathcal{P}(\mathcal{O}/\lambda_i), \quad arg \max^{2}_{1 \le i \le \mathcal{N}1} \mathcal{P}(\mathcal{O}/\lambda_i)] \tag{1}$$

where,
- p_i is the i^{th} pair of syllable segments (where $1 \le i \le \mathcal{N}_1$)
- $\mathcal{P}(\mathcal{O}/\lambda_i)$ is the probability of the observation sequence $\mathcal{O}$ ($o_1 o_2 ... o_n$) 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 $\mathcal{N}1$ cases, both 1-best and 2-best syllable segments are identical/similar. This step gives $\mathcal{N}1$ pairs of syllable segments.

5. Among $\mathcal{N}1$ pairs ($p_1, p_2, ..., p_{\mathcal{N}1}$), 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 $\mathcal{N}2$ clusters ($\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{\mathcal{N}2}$). In the next step, the model parameters are re-estimated incrementally.

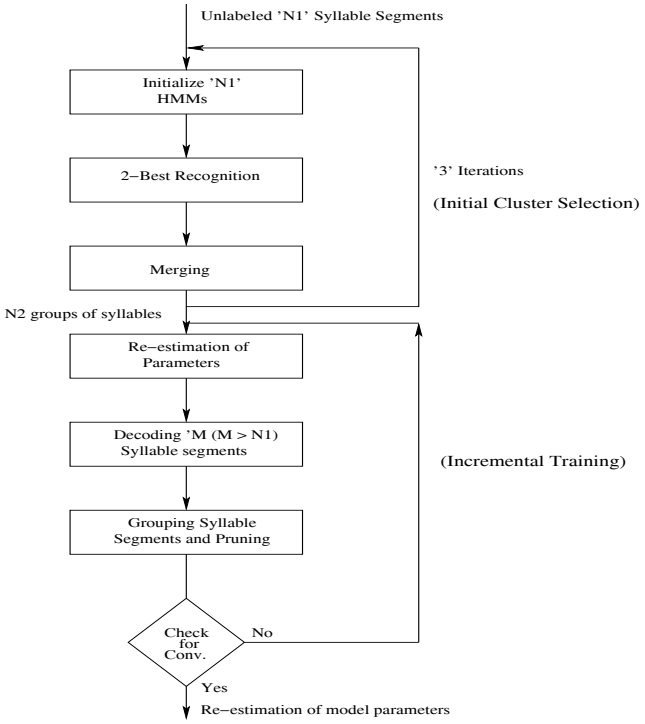

Fig. 1. *Flow chart : Unsupervised and Incremental HMM*

2.2.2. Incremental training

After selecting the initial clusters $(\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{\mathcal{N}2})$, 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. This training strategy must be contrasted to 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 derived $(\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{\mathcal{N}2})$ 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 $(\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_{\mathcal{M}_l})$ 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.

2.2.3. 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 be reduced 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 produces $\mathcal{N}3$ syllable clusters $(\mathcal{C}_1, \mathcal{C}_2, ..., \mathcal{C}_{\mathcal{N}3})$, and in turn $\mathcal{N}3$ syllable models $(\lambda_1, \lambda_2, ..., \lambda_{\mathcal{N}3})$.

The above mentioned process is repeated and the syllable models are trained for each language $\mathcal{L}_l$ separately. Here, the total training process is unsupervised and so these clusters/syllable models do not have an identity that has a bearing on their acoustic manifestation.

2.3. Language Identification (LID)

2.3.1. LID system using acoustic likelihood

Syllable models are created for each language $\mathcal{L}_l$ using unsupervised and incremental HMM, as explained in Section. 2.2. The clustering process automatically derives $\mathcal{N}3$ syllable models $(\lambda_1, \lambda_2, ..., \lambda_{\mathcal{N}3})$ for each language $\mathcal{L}_l$. During testing, each of the 45sec test utterances is segmented into syllable-like units. This results in $\mathcal{K}$ syllable segments $(\mathcal{S}_1, \mathcal{S}_2, ..., \mathcal{S}_{\mathcal{K}})$. Using **Viterbi decoding** procedure, the syllables are decoded:

$$p(\mathcal{S}_k/\mathcal{L}_l) = \max_{i=1,2,...,\mathcal{N}_3} p(\mathcal{S}_k/\lambda_i) \quad (2)$$

where $k = 1, 2, 3, ..., \mathcal{K}$.
For each language $\mathcal{L}_l$, the acoustic log-likelihood score $\mathcal{P}_l$ is calculated.

$$\mathcal{P}_l = \sum_{k=1}^{\mathcal{K}} \log\ [p(\mathcal{S}_k/\mathcal{L}_l)] \quad (3)$$

where $l = 1, 2, 3, ..., \mathcal{N}$.
The language of the test utterance is identified as,

$$l^* = arg \max_{l=1,2,...,\mathcal{N}} (\mathcal{P}_l) \quad (4)$$

2.3.2. LID system using unique syllable segments

In this work, since syllable is considered as the basic sound unit for recognition, for each language, say $\mathcal{L}_i$, the number of unique syllable models when compared to the other language, say $\mathcal{L}_j$, is expected to be high. To find out these unique syllable models of $\mathcal{L}_i$, say $\mathcal{S}_{\mathcal{U}}^{\mathcal{L}i}$ when compared to $\mathcal{L}_j$, the following experiment is carried out. The training syllable segments of the language $\mathcal{L}_j$ is decoded by the syllable models $(\mathcal{S}_{\mathcal{W}}^{\mathcal{L}i})$ of the language $\mathcal{L}_i$. It is found that, all the syllable segments of $\mathcal{L}_j$ are clustered to only half of the syllable models $(\mathcal{S}_{\mathcal{C}}^{\mathcal{L}i,j})$ of $\mathcal{L}_i$ (Fig.(2)). Here, the unique syllable models of $\mathcal{L}_i$ is given by,

$$\mathcal{S}_{\mathcal{U}}^{\mathcal{L}i} = \mathcal{S}_{\mathcal{W}}^{\mathcal{L}i} \cap \mathcal{S}_{\mathcal{C}}^{\mathcal{L}i,j} \quad (5)$$

where,
$\mathcal{S}_{\mathcal{U}}^{\mathcal{L}i}$ - Unique syllable models belonging to language $\mathcal{L}_i$.

$\mathcal{S}_{\mathcal{W}}^{\mathcal{L}i}$ - The universe of syllable models belonging to language $\mathcal{L}_i$.

$\mathcal{S}_{\mathcal{C}}^{\mathcal{L}i,j}$ - Common syllable models belonging to languages $\mathcal{L}_i$ and $\mathcal{L}_j$.

Using this approach, the number of unique syllable models for all $\mathcal{L}_i$,$\mathcal{L}_j$ (where $i \neq j$) is obtained. During testing, each test utterance is first segmented into syllable-like segments. These syllable segments are then decoded using the syllable models of each pair of languages, say $\mathcal{L}_i$ and $\mathcal{L}_j$. After decoding, for each of the languages in the pair, the number of unique syllable segments are found using $\mathcal{S}_{\mathcal{U}}^{\mathcal{L}i}$ and $\mathcal{S}_{\mathcal{U}}^{\mathcal{L}j}$. The language which gets maximum number of unique syllable segments is noted as the winner for the test utterance, in that pair of languages.

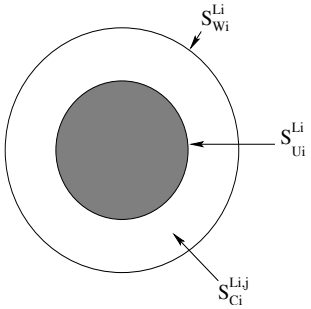

Fig. 2. *Unique syllable segments.*

Since the results of the above mentioned methods are complementary in some cases, it is decided to go for **one followed by another** approach. For the 2-best languages declared by the LID system using acoustic likelihood, the LID system using unique syllable segments approach is used.

3. PERFORMANCE ANALYSIS

The Oregon Graduate Institute Multi-language Telephone Speech Corpus, which is designed specifically for LID research, is used for both training and testing. This corpus currently consists of spontaneous utterances in 11 languages: English (En), Farsi (Fa), French (Fr), German (Ge), Hindi (Hi), Japanese (Ja), Korean (Ko), Mandarin (Ma), Spanish (Sp), Tamil (Ta) and Vietnamese(Vi). The utterances were produced by $\sim$90 male and $\sim$40 female, in each language over real telephone lines. To maintain the homogeneity in training and testing across languages, for each language 30 speakers are used for training and 20 speakers are used for testing. All the training and test set speakers are different. The 1-best performance of the proposed approach for 11 languages is given in Table.1 for 45sec and 10sec test utterances.

When the failure cases are manually analyzed, it is noticed that the error in identifying the languages correctly, is either because of the low-quality of the speech signal or accent variation. In particular, for the language Tamil, majority of the failure cases are found to be of the Srilankan accent.

4. CONCLUSION

In this paper, a novel approach is proposed for language identification, which does not require annotated corpora. It is shown that syllable models can be trained incrementally, without any supervision. The performance of the proposed approach for language identification shows that, syllable may also be considered as a potential sound-unit for language identification task. At the syllable level, since the number of unique sound-units is to be high, it is further shown that incorporating this information into the system improves the performance considerably.

Table 1. Language-wise performance (in %) comparison. (A) - Using acoustic likelihood, (B) - Using acoustic likelihood and unique syllables.

Language	45sec tests		10sec tests	
	A	B	A	B
En	85	80	80	80
Fa	75	80	65	70
Fr	95	90	85	95
Ge	65	85	65	65
Hi	80	85	80	75
Ja	65	90	55	60
Ko	60	60	55	55
Ma	40	40	40	40
Sp	80	85	70	80
Ta	55	65	55	55
Vi	65	80	60	65
Average	69.5	75.9	64.5	67.2

5. REFERENCES

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

[2] Marc A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech," *IEEE Trans. Speech, Audio Processing*, vol. 4, no. 1, pp. 31–44, Jan. 1996.

[3] A.K.V. SaiJayaram and V.Ramasubramanian and T.V.Sreenivas, "Language identification using parallel sub-word recognition," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 2003, pp. 32–35.

[4] Kung-Pu Li, "Automatic language identification using syllabic features," in *Proceedings of IEEE Int. Conf. Acoust., Speech, and Signal Processing*, 1994, pp. 297–300.

[5] V.Kamakshi Prasad, T.Nagarajan, and Hema A.Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions, to appear in speech communication," 2003.

[6] T.Nagarajan and Hema A.Murthy and Rajesh M. Hegde, "Segmentation of speech into syllable-like units," in *Eurospeech*, Geneva, 2003, pp. 2893–2896.