UNSUPERVISED CLUSTERING OF SYLLABLES FOR LANGUAGE IDENTIFICATION

Subhadeep Dey and Hema Murthy

Department of Computer Science and Engineering, IIT Madras

sdey@cse.iitm.ac.in, hema@cse.iitm.ac.in

ABSTRACT

Automatic Language Recognition makes extensive use of phonotactics for identifying a language. The accuracy of phonotactic information depends upon the amount of data available for training. The state of the art approaches capture the phonotactics in terms of cross-lingual GMM tokens. The accuracy of such tokenisers crucially depends upon the availability of specific corpora. In this paper, we suggest an alternative to GMM tokens, namely, syllable based tokens. Syllables implicitly capture the phonotactics across phonemes in a language. Unsupervised Syllable tokenisation for language identification requires a) segmentation of speech into syllable-like units syllable level, and b) unsupervised modeling of the syllable tokens by Hidden Markov Models. The first issue is addressed by segmenting the wavform into syllable-like units using a well-established group delay based segmentation algorithm. To address the second issue, two different solutions are proposed, namely, (i) a top down clustering approach, which does not require significant parameter tuning, and is also robust, and (ii) a universal syllable approach. In this syllable models for every language are obtained from adapted universal syllable models. Experimental results on the OGI 1992 multilingual corpus and NIST 2003 LRE corpus show that the proposed approaches donot require significant tuning of parameters and the performance is comparable to that of a well-tuned baseline syllable tokenisation system.

Index Terms— top down syllable clustering, universal syllable models, unsupervised clustering, syllable segmentation

1. INTRODUCTION

Automatic spoken Language Identification (LID) is the task of classifying an utterance as belonging to one of the known languages. The state of the art LID system uses parallel phone recognition followed by language modeling (PPRLM) to capture the characteristics of a language. This system is referred to as the explicit LID system which needs labeled speech corpora. An alternative scalable approach is the implicit LID system which does not need annotated speech database. A popular implicit LID system is the Gaussian Mixture Model (GMM) tokenizer as described in [1]. It uses GMMs as the front end to tokenize an incoming speech utterance into cluster indices. These cluster indices are then used to create interpolated Language Models that discriminate among languages. The performance of this system is comparable to Parallel PPRLM system on the OGI-MLTS database.

The sub-word unit based LID system has consistently given good performance as evidenced from the NIST evaluations. Researchers have preferred to use phonemes over other sub-word units like syllables. Phonemes by itself cannot identify a language as languages share the same phoneme inventory. However sequence of phonemes contain the distinguishing characteristics and it has been found that trigram or higher n-gram phonemes statistics gives higher accuracy [2]. A syllable on an average contains three phoneme units. We hypothesize that syllables inherently contain distinguishing characteristics. To cite an example on the importance of syllables, if the domain of discourse is Indian languages, the presence of the syllable */zha/*, reduces the search space to two languages, namely Tamil and Malayalam. The other advantage of using syllables as a sub-word for building an LID system is that it can be automatically extracted from the speech signal.

Most LID systems [2] build statistical models with out explicitly capturing the acoustic characteristics of the units that make up a language. These systems build statistical models that capture the unique characteristics of the language using huge amount of data. The state of the art system [3] use the Call Friend Corpus to build the initial models for language identification. The Call Friend corpus is about 800 hours of data. Humans do not require such amount of data for identifying languages. We conjecture that it relies on signal processing cues to identify events [1] and optionally followed by statistical modeling. Since syllable is the basic unit of production, we arrive at a hybrid model, in that syllable boundaries are obtained using knowledge based signal processing. The recognition of syllables is performed using isolated-style HMMs.

Nagarajan [4] was the first to suggest the use of syllables for building implicit LID systems. The drawback of the approach used by Nagarajan is that the parameters of the incre-

[1] Events can be syllable, word boundaries.

mental HMM based clustering approach, significant tuning of parameters during training [5]. This is primarily because syllable models built from insufficient data suffer from modelling errors. Iterative tuning of parameters is required to ensure that robust syllable models are built [5]. To address this problem, we have explored two different approaches: i) top down approach approach to cluster syllables, and ii) a universal syllable model framework. While the performance [2] of the two proposed syllable approaches are worse than that of the baseline for OGI-MLTS corpus, it performs better than the baseline system for NIST 2003 LRE database. Although the performance is not comparable to that of state-of-the-art systems (3% EER for the NIST 2003 LRE database), it is important to note that we have only used the OGI-MLTS and NIST 2003 training and development corpora for building the models [3].

The rest of the paper is organised as follows. In section 2 we describe the dataset used. We briefly review the baseline segmentation algorithm and the baseline clustering algorithm in section 3.1 and 3.2 respectively. In Section 4, 5, we describe the top down syllable clustering algorithm and Universal Syllable framework respectively for LID. The experimental results on the OGI-MLTS corpus and NIST 2003 are presented in Section 6 and finally conclude in section 7.

2. SPEECH CORPORA

The Oregon Graduate Institute Multi language Telephone Speech (OGI-MLTS) and NIST 2003 Language Recognition Evaluation (LRE) corpora are used for performing experiments in language identification. The description of the databases are given below:

2.1. OGI-MLTS

This corpus [6] consists of spontaneous utterances in 11 languages: English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil and Vietnamese. The utterances were produced by 90 males and 40 females, in each of the languages over telephone lines.

2.2. NIST 2003 LRE

The development data of NIST 2003 LRE consists of conversational telephone speech in each of the target languages: Arabic, English, Farsi, French, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, Vietnamese. Development data are drawn from 1996 LRE development and the evaluation set. The NIST 2003 LRE evaluation data consists of 80 utterances in each of the target languages and additional 320 utterances from other corpora or other languages.

[2] In this paper identification accuracy is used as measure of performance.
[3] State of the art systems use Callfriend Database for Training Models.

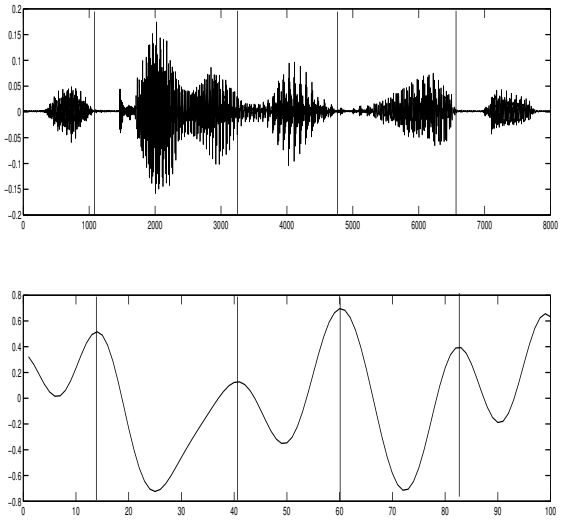

Fig. 1. *Segmentation of an utterance taken from the OGI database. The peaks in the second plot correspond to the locations of the syllable boundaries.*

3. BASELINE SYLLABLE BASED SYSTEM

It consists of (i) Segmenting utterances into syllable like units, (ii) Clustering similar sounding syllables and (iii) Testing against the syllable models.

3.1. Baseline segmentation algorithm

A syllable consists of a consonant, vowel and consonant (C^*VC^*). The vowel portion contains significant part of the energy of the syllable segment compared to that of the consonant parts. In [7], a syllable segmentation algorithm is proposed which uses short time energy as the criteria for the task. The variation in the short time energy function is smoothed by group delay functions [7]. Figure 1 shows the syllable boundaries obtained using the group delay function. The algorithm occasionally misses syllable boundaries. Sometimes therefore bisyllables are identified as a single unit. This is just as well, as it might correspond to a syllable sequence that are commonly found in a language.

Using the baseline segmentation algorithm, the training speech data of every language are segmented into syllable-like units resulting in M_l syllable like units for each language.

3.2. Baseline Syllable Clustering Algorithm

A syllable clustering algorithm has been proposed in [4] to cluster similar sounding syllables. From these similar sounding syllables, models for each of the languages are derived.

The syllable clustering algorithm is divided into two phases as:

- Initial Cluster Selection: To initialize the parameters of the syllable models. Each of the syllable models are created from just one syllable instance with multiple frame rate.
- Incremental Training : Derive representative model for each of the language.

After the clustering process, we get H_l number of clusters for every language.

3.3. Testing

Various methods like acoustic likelihood, voting etc., can be used to evaluate the performance as described in [4]. We have used acoustic log-likelihood scores for evaluating system as it gives the best performance. An utterance is segmented into syllable like units and each of these syllables is scored against the syllable models of every language. Accumulate the log likelihood scores and pick the language that gives the maximum score.

3.4. Problem with the baseline system

The drawback of the baseline syllable based LID system is that in the initial cluster selection phase, the syllable models are built from a single syllable. Parameters of HMMs cannot be estimated accurately due to insufficient data. This is inspite of using multiple frame-rate and multiple frame-size approach suggested in [5]. Therefore, significant *manual* tuning of parameters is required during training for each of the languages.

To overcome this problem, in this paper, two approaches, namely, top down clustering approach and universal syllable model approach are proposed, which do not require significant tuning of parameters. These approaches are described in subsequent sections.

4. TOP DOWN CLUSTERING

The data insufficiency problem can be addressed by building models in a top down fashion and requiring that each of the models built have sufficient number of examples. The process is illustrated in Figure 2, the first node (root node) is created with sufficient amount of syllables instances. The parameters of the root node can be transformed by matrices A and $A^{'}$ to create two child nodes/models at first level and applying the process recursively appropriate number of syllable models can be obtained. The assumption is that syllable models at the leaf levels are pure in the sense that they contains more examples of the same syllable.

In the absence of labeled data computing these matrices is a big problem and hence we assume that all the matrices

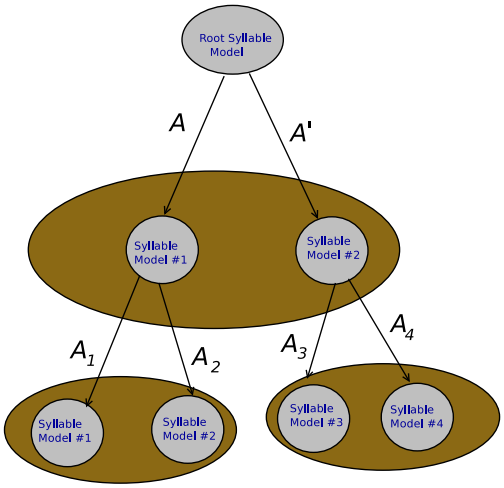

Fig. 2. Obtaining syllable models by top down clustering.

at all the levels of the tree are same i.e. $A = A^{'} = A_1 = A_2 = ... = A_L$. To further simplify the computation, only the means of the parent HMM nodes are transformed and updated; and means of the two child nodes are given by the equation $\mu_p \pm$ k σ_p where μ_p, σ_p are the mean and variance of the parent HMM model and k is a parameter that controls the spread of between the means. [4]. The algorithm is described in the following subsection.

4.1. Algorithm to build syllable models in top down fashion

Define $N_s = \frac{M_L}{M_\theta}$ be the effective number of syllables per model where M_θ is the number of models expected (input) ; M_i is the number of syllable models at i^{th} iteration; $\mu^{(x,j)}$ and $\sigma^{(x,j)}$ are the mean and variance parameters of the x^{th} syllable model's j^{th} state. [5]

- Initialise the algorithm with a root syllable model obtained from N_s syllables.

For ith iteration:

1. Create 2 syllable models from each syllable models with means $\mu^{(x,j)}$ + k $\sigma^{(x,j)}$ and $\mu^{(x,j)}$ - k $\sigma^{(x,j)}$ respectively from these M_i models.
2. Score i*N_s syllables against 2*M_i models and assign the index of the highest scoring syllable model.
3. Discard the models having less than $N_{s,min}$ syllables and the models after this step be $M_i^{'} \leq 2*M_i$
4. Re-estimate the parameters of these $M_i^{'}$ models and $M_i^{'}$ = M_i.

[4]The Gaussian of HMM's of each state are assumed to have diagonal covariance matrix

[5]We assume that each state of the HMM has only 1 mixture

Repeat Step $1-4$ until $M_i \leq M_\theta$. The algorithm converges when N_s - $N_{s,min} > 10$ with $N_{s,min} \geq 10$ so that models have sufficient examples.

5. UNIVERSAL SYLLABLE MODEL (USM)

The success of Maximum a Posteriori (MAP) adaptation in Speaker Verification [8] motivates us to explore this approach in syllable based LID system. The steps to obtain each of the class model in MAP adaptation are

- Pool data from all the classes and build a single Gaussian Mixture Model (GMM) which is called Universal Background Model (UBM).
- Derive a class model 'l' by adapting the parameters (only means μ_i^l are adapted) of the UBM using that class training data as:

$$\mu_i^l = \alpha E_i(x) + (1-\alpha)\mu_i^{ubm} \quad (1)$$

where μ_i^{ubm} is the mean of the UBM, $E_i(x)$ is the new estimate of mean and α is the adaptation coefficient controlling the balance between the old and the new mean. In this paper, we explore HMM adaptation framework similar to GMM-UBM framework in remainder of the section.

5.1. Universal Syllable Model

The major steps in this approach are (i) creating the universal syllable set and (ii) deriving models for each of the languages. We describe the method to create the same below:

- Randomly select N syllables from each of the languages to form an universal syllable inventory (s_1, s_2, s_3, $\cdots$, s_N). These syllables are then clustered using the baseline syllable clustering algorithm. The clustering process will result in 'L' clusters which we refer to as universal syllable model set.

We then derive models for each of the languages as follows

- The training utterances of every language are segmented into syllables. Each syllable is scored against the universal syllable model and then it is assigned to the highest scoring syllable model.
- adapt the means of GMMs in every state using the result of previous step.

The main advantage of this method is that it reduces the training time by $\approx \frac{1}{7}^{th}$ to that of baseline syllable based system.

6. EXPERIMENTS

In this section we describe the results of the baseline syllable and universal syllable based LID systems. For OGI database, 40 utterances of 45 seconds duration each are used for training, 20 utterances of 45 seconds each are used for testing and development. For NIST 2003 LRE, we have considered only the 80 utterances each of 30 seconds duration from every language corresponding to primary condition as evaluation data. Since CallFriend database is not avaliable to us, therefore we have used OGI-MLTS and 1996 development and evaluation data as training data for NIST 2003 LRE.

We conducted closed set language identification on OGI-MLTS [6] and NIST 2003 LRE [9] databases respectively. The language identification results are shown in Table 1.

Table 1. *Identification Accuracy of LID systems on OGI-MLTS and NIST 2003 LRE.*

LID System	OGI-MLTS	NIST 2003 LRE
GMM-UBM	45 %	35%
Baseline System	62.72 %	39.40 %
USM	55 %	43.54 %
USM Training and Dev	63.63 %	-
Top Down Clustering	58.63 %	**45.62** %
Top Down Clustering Training And Dev	**68.18** %	-

6.1. Baseline Syllable based LID system

For OGI-MLTS database, models are built from 5000 syllables from each of the languages. For each of the languages we get 370-390 representative models after the clustering process. We observed that the best performance is **62.72 %** when each of the syllable models is created with 5 states, 1 mixture HMM. We have observed that increasing the syllable inventory size does not improve performance and 5000 syllables is optimal.

Owing to limited amount of training data the performance on 2003 NIST LRE is only **39.4 %**. Since development data is not available, parameters of best performing system on OGI-MLTS database are used.

We compare syllable based LID system with state of the art GMM-UBM system using shifted delta cepstral (SDC) features as described in [3]. SDC features are extracted from the speech data with 7 dimensional MFCC features and 49 dimension delta MFCC appended to obtain a 56 dimension feature vector. An accuracy of **45.9%** and **35.65 %** are obtained on OGI-MLTS and 2003 NIST-LRE database respectively. We therefore conclude that syllable based LID system works better than SDC based GMM-UBM system on this limited amount of data.

6.2. Top Down Clustering

For OGI-MLTS, the root syllable model is created with 5 states and 1 mixture HMM. The optimal value of k is found to be 0.01 and it is obtained by optimizing the performance on the development data. The best performance is obtained with M_θ = 128 clusters. On increasing the training data, the performance increased by $\approx 10\ \%$. The performance is attributed to the fact that larger number of syllable models better capture the variability of the language.

In 2003 NIST database, 1996 development and evaluation data were used for building the syllable models. The best performance is obtained with M_θ = 512 clusters and $N_s \approx 30$ syllables. The performance degrades as M_θ is increased to 1024 or M_θ reduced to 128 and 256 clusters.

6.3. Universal Syllable Model

We choose 500 syllables from every language to form the universal set of syllables and each of these syllables is represented by 5 states and 1 mixture HMM. These models are then clustered using the baseline clustering algorithm to obtain 445 syllable models. These syllable models are then adapted for every language. After adaptation we obtain between 395 to 405 syllable models for every language. For OGI-MLTS corpora the language recognition performance is **55 %**. To evaluate the performance of universal syllable models on amount of training data we use training and development data for the adaptation process and we obtain accuracy of **63.63 %** (see Table 1, row 4, column 1).

For NIST 2003 LRE, we use the 1996 development and evaluation data for adaptation process to obtain syllable models for every language. We have used english syllable model trained using the OGI-MLTS corpus with the baseline syllable LID system as the universal syllable models.These models are adapted for every language. The performance is found to be **43.2 %**. To evaluate the effect of choosing universal syllable models we used the following language model trained in OGI-MLTS obtained by the baseline system as the universal syllable models: Farsi syllable model and French syllable model. We found that these syllable models the average performance is almost the same.

6.4. Discussion

For the OGI database, when training and development data are used to build the models, the performance of the baseline system decreased. The reason for this behavior is that the parameters of the baseline system were optimized for this database. The number of parameters in the universal syllable models approach is the same as the baseline syllable system with the exception of relevance factor [8] which is set to default value of 10.

7. CONCLUSION

In this paper we have described the universal syllable model approach to built LID system. Top down clustering approach and universal syllable model approach to language identification are explored in this paper. It is observed that top down clustering approach performs better than universal syllable model approach in OGI-MLTS and NIST-2003 databases. Nevertheless, training time for universal syllable approach is significantly less than than of the baseline system and top down clustering.

8. REFERENCES

[1] Pedro A. Torres-carrasquillo, Elliot Singer, Mary A. Kohler, and J. R. Deller, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. ICSLP*, 2002, pp. 89–92.

[2] Ricardo de Córdoba, Luis Fernando D'Haro, Fernando F. Fernández-Martínez, Javier Macías Guarasa, and Javier Ferreiros, "Language identification based on n-gram frequency ranking," in *INTERSPEECH*, 2007, pp. 354–357.

[3] Xi Zhou, Jiri Navratil, Jason W. Pelecanos, Ganesh N. Ramaswamy, and Thomas S. Huang, "Intersession variability compensation for language detection," in *ICASSP*, 2008, pp. 4157–4160.

[4] T. Nagarajan and H.A. Murthy, "Language identification using parallel syllable-like unit recognition," in *ICASSP*, 2004, pp. 401–404.

[5] Nagarajan, T., *Implicit Systems for Spoken Language Identification*, PhD dissertation, Indian Institute of Technology Madras, Department of Computer Science and Engg., Madras, India, May 2004.

[6] Yeshwant K. Muthusamy, Ronald A. Cole, and Beatrice T. Oshika, "The ogi multi-language telephone speech corpus," 1992, pp. 895–898.

[7] V. Kamakshi Prasad, T. Nagarajan, and Hema A. Murthy, "Automatic segmentation of continuous speech using minimum phase group delay functions," *Speech Communication*, vol. 42, no. 3-4, pp. 429–446, 2004.

[8] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[9] "2003 nist lre plan," `http://www.nist.gov/speech/tests,2003`, 2003.