A HIERARCHICAL FRAMEWORK FOR LANGUAGE IDENTIFICATION

Saad Irtza[1,2], Vidhyasaharan Sethu[1], Haris Bavattichalil[1], Eliathamby Ambikairajah[1,2], Haizhou Li[3]
[1]School of Electrical Engineering and Telecommunications, UNSW Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Australia
[3]InstituteforInfocommResearch, A*STAR, Singapore
s.irtza@student.unsw.edu.au

ABSTRACT

Most current language recognition systems model different levels of information such as acoustic, prosodic, phonotactic, etc. independently and combine the model likelihoods in order to make a decision. However, these are single level systems that treat all languages identically and hence incapable of exploiting any similarities that may exist within groups of languages. In this paper, a hierarchical language identification (HLID) framework is proposed that involves a series of classification decisions at multiple levels involving language clusters of decreasing sizes with individual languages identified only at the final level. The performance of proposed hierarchical framework is compared with a state-of-the-art LID system on the NIST 2007 database and the results indicate that the proposed approach outperforms state-of-the-art systems.

Index Terms— Language identification, hierarchical framework, i-vector, PLLR

1. INTRODUCTION

The most widely adopted approaches to language identification use acoustic and phonotactic information [1-3]. Specifically, most current systems employ the i-vector framework trained on both acoustic and phonotactic front-ends. MFCCs continue to be one of the most commonly utilized acoustic front-end and recently Phone Log Likelihood Ratios (PLLRs) have shown to be a promising phonotactic front-end [4-8]. State-of-the-art LID systems also make use of score level fusion to combine individual systems based on different speech cues[4]. These systems are single level approaches where all language hypotheses are treated identically.

The proposed hierarchical structure is based on the observation that similarities between languages do exist (e.g., at a very broad level tonal and non-tonal languages can be separated in to two groups). Furthermore, it is easier to distinguish between languages that are significantly dissimilar than those that have a lot more similarities (e.g., It is much easier to distinguish between Dutch and Vietnamese than it is to distinguish between Dutch and Afrikaans). It has also been observed that the cues that are utilised to distinguish between languages depend on how similar they are (e.g., Prosodic cues are significantly better at distinguishing between a tonal and a non-tonal language than they are at distinguishing between two non-tonal languages) [2]. Preliminary work on the use of hierarchical structures have been proposed as an alternative to the traditional single level structure and has shown some promising performance [9, 10]. There are several approaches to creating hierarchical structures. For example, automatic language clustering algorithms can be used to form a hierarchical structure [9]. Alternatively, linguistic language families [11] can also form the basis for a hierarchical structure. Previously, language clustering based on a performance based distance measure using GMMs was used to form a binary hierarchical tree [9]. This was superseded by a clustering method based on model likelihood based distance [10]. Information based on language clusters have also been used effectively in NIST language pair evaluation tasks [12].

The proposed framework also determines the hierarchical structure based on automatic clustering of the languages. However, the structure is determined in such a manner that allows for different acoustic or phonotactic front-ends to be utilised for distinguishing between language families at different levels in the framework (or individual languages in the case of the last level).

2. PROPOSED HIERARCHICAL FRAMEWORK

A high level block diagram of the proposed hierarchical language identification (HLID) framework is shown in Figure 1. The root of the tree will consist of a single language group that contains all language hypothesis with nodes at each subsequent level representing a smaller cluster of languages such that the union of all the groups corresponding to the nodes with a single parent node will be the language group associated with the parent group. Each level of the tree will act as a language recognition system, with every node representing a possible hypothesis. This framework allows the most suitable front-end to be chosen at each level.

The specific structure of the hierarchical framework can be based on language families established in linguistics. However, in the work reported in this paper a computational approach was adopted and the structure was determined on the basis of automatic clustering of the languages. The primary aim of the proposed structure is to divide the normal LID task into several subtasks structured in a hierarchy such that the easier classification sub-tasks are

carried out at the initial levels and harder classification sub-tasks involving distinguishing between similar languages are carried out at lower levels.

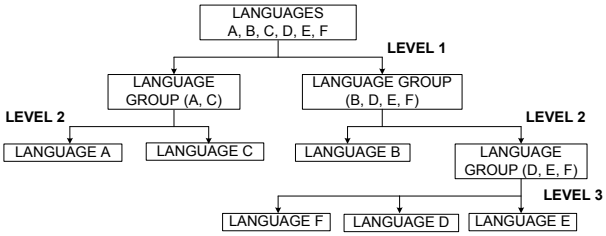

Figure 1: Hierarchical Framework for Language Identification

2.1 Language Clustering

Given a set of languages, clustering can either be top down, i.e., divisive clustering or bottom up, i.e., agglomerative clustering [9, 11]. As in [9], a bottom up agglomerative approach is undertaken in the proposed framework, with a different feature set chosen at each level. This clustering is based on a similarity measure that is defined in a suitable i-vector space. Specifically, the cosine similarity score (CSS) [13] is adopted and the similarity score between two languages (A and B) is computed as follows:

$$S^{\phi}(A,B) = \frac{L_A^{\phi} \cdot L_B^{\phi}}{\left\| L_A^{\phi} \right\| \left\| L_B^{\phi} \right\|} \quad (1)$$

Where, L_A^{ϕ} and L_B^{ϕ} are i-vectors based on front-end ϕ estimated using data from languages A and B and ϕ may denote MFCCs, PLPs, etc.

The Unweighted Pair-Group method of Average (UPGMA) [11] is used to extend this similarity score to a measure of similarity between two clusters of languages (C_1 and C_2) as follows:

$$S^{\phi}(C_1,C_2) = \frac{\sum_{m \in C_1, n \in C_2} S^{\phi}(m,n)}{n_{C_1} n_{C_2}} \quad (2)$$

Where, m denotes the languages belonging to cluster C_1, n denotes the languages belonging to cluster C_2, n_{C_1} is the total number of languages in C_1 and n_{C_2} is the total number of languages in C_2. It can be seen that $S^{\phi}(C_1,C_2)$ is computed as the average pairwise cosine similarity between all possible pairs of languages from C_1 and C_2.

The agglomerative clustering used to determine the structure of proposed hierarchical framework initially clusters the set of languages into a smaller set of language groups based on the above mentioned similarity measures estimated on one of a given set of front-ends. This set of language groups are then clustered again based on another front-end to give an even small set of language groups. This process is continued iteratively until a single cluster (group) of languages is obtained and each level of clustering corresponds to a level in the hierarchical structure.

In each level, given a front-end ϕ and a set of languages/language groups to be clustered, the clustering is carried out as follows:

Step 1: The pairwise similarity between all languages/language groups is computed and the pair (a, b) with the highest similarity score is assigned to a cluster C_m $(m = 1, 2, ...,)$ if and only if eqn (3) is satisfied.

$$S^{\phi}(a,b) > \alpha \quad (3)$$

Where, α is a fixed threshold that is selected empirically and in our work, $\alpha = 0.5$.

Step 2: C_m is expanded by considering the next language/language group, (c) that is most similar to either A or B and including it in C_m if and only if eqn (4) satisfied.

$$\forall_{i,j \in C_m} S^{\phi}(i,j) - \forall_{k \in C_m} S^{\phi}(k,c) < \beta \quad (4)$$

Where β is a fixed threshold that is selected empirically and in our work, $\beta = 0.05$.

Step 3: Here Step 2 is repeated until the cluster C_m is finalized;

Step 4: Steps 1 to 3 are repeated until all clusters are determined.

2.2 Experimental Setup

The LID experiments reported in this work are performed on the NIST 2007 LRE dataset. The dataset consists of conversational telephonic speech in 14 languages [14]. For training the language models and for the development purpose speech utterances derived from Call-Friend, NIST 2005 LRE and NIST 2007 LRE datasets are used. The distribution of the training, development and evaluation data used for each target language is the same as that described in [15]. Total duration of the training data used is approximately 968 hours. The development test set consists of 10 conversations selected randomly from each target language. The selected utterances are segmented to mimic 30sec test utterances. Final results are reported on 30 sec test set for the primary task in the NIST 2007 LRE dataset which consists of 2158 test trials.

The three sets of frame based PLLR features of dimensionality 59, 50 and 43 were estimated using HU, RU and CZ TRAPs/NN phone recognisers [16] respectively. Following this, voice activity detection (VAD) is carried out by removing the frames whose highest PLLR value corresponds to the non-speech unit. PLLR features are augmented with dynamic coefficients. The openSMILE toolbox [17] was used to extract the 13 dimensional MFCCs and 13 dimensional PLPs, both augmented with 13-7-1-3 SDCs. The openSMILE toolbox was also used to carry out voice activity detection. All Universal Background Models (UBM) were Gaussian mixture models with 1024 components, estimated using Maximum Likelihood criteria (ML) and employing binary mixture component splitting. Total variability matrix (T-Matrix) is estimated as in [6]. I-vectors of 400 dimensions are used since they have shown promising results for language recognition [6].

2.3 Level wise Front-end Selection for Clustering

As previously mentioned, the primary aim of the proposed framework is to utilise a hierarchical structure where the

most suitable front-end is employed for classification at each level. Recent LID systems have used i-vectors derived from acoustic and phonotactic front ends [1, 4, 6, 7, 9, 10, 18] and these front ends have also been used in this work. Specifically 5 front-ends, namely MFCCs + SDC (13-7-1-3), PLPs, and PLLRs based on Hungarian (HU), Czech (CZ) and Russian (RU) phonemes, are considered at each level.

The question arises which features are to be used for clustering at each level? One possible solution is to try all possible combination of features at each hierarchy level. This is computationally expensive and exhaustive search of all front-ends will not be feasible as the number of front-ends increase in future work. As an alternative, a greedy method is implemented which starts by considering all the front-ends individually and proceeds with the one that minimises the number of clusters at each level. The number of clusters at each level is minimised in order to obtain the deepest hierarchical structure which in turn gives greater flexibility in choosing appropriate features.

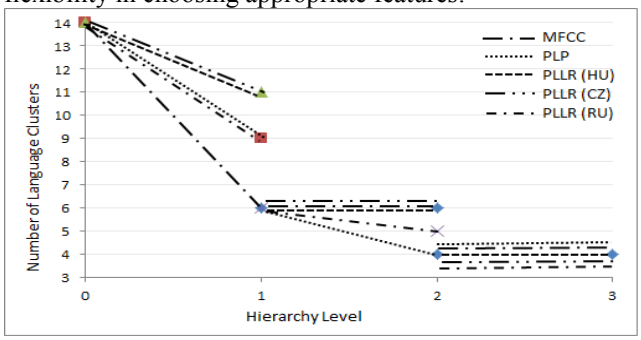

Figure 2: Language clustering using different features

Figure 2 compares the clusters obtained using all 5 front-ends on the 14 languages in the NIST 2007 database. It can be seen that the 14 languages are grouped into 11 clusters when based on the PLLR (CZ) and PLLR (HU) front-ends, 9 clusters when based on the PLLR (RU) and PLP front-ends and 6 clusters when based on the MFCC front-end. Consequently, the MFCC based clustering is used to define the bottom level of the hierarchical structure. Similarly the PLP front-end based clustering is used to define the next highest level since it reduces these 6 groups into 4 clusters. At the next stage of clustering, since none of the front-ends lead to any further reduction in cluster numbers, classification between 4 clusters at this stage becomes the first level of the hierarchical structure (shown in Figure 3).

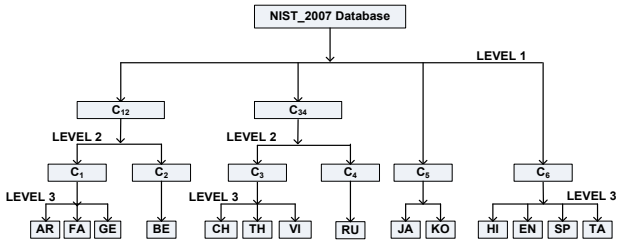

Figure 3: Language Hierarchical Tree Structure estimated from NIST 2007. It should be noted that the clusters $\mathbf{C_5}$ and $\mathbf{C_6}$ are identical at levels 1 and 2.

3. CLASSIFICATION

Recently Gaussian Probabilistic Linear Discriminant Analysis (GPLDA) based back-ends have been used in state-of-the-art LID systems [19] and is adopted as the classifier in all levels of the hierarchical structure as well as the back-end in the single-level systems to which the proposed framework is compared. The front-ends chosen are level specific. During classification, at each level of the hierarchical frame the log likelihood ratios of the possible classes are computed and propagated down to the next level and added to the log likelihood ratio in that level. For example, in the structure depicted in Figure 3, given a test utterance, the overall log likelihood ratio for German (GE) is the sum of the log likelihood ratios of C_{12} in level 1, of C_1 in level 2 and of GE in level 3. Finally the language corresponding to the test utterance is identified as the one that has the highest overall log likelihood ratio.

3.1 Level wise Front-end Selection for Classification

Given that the front-end used for clustering was the one that minimised the number of clusters. It is reasonable to assume that other front-ends will have more discriminative information at that level. For example, at the lowest level of clustering, MFCCs were used to combine 14 languages to form 6 clusters (Figure 2) and consequently level 3 of classification (Figure 3) is best carried out with the other front-ends. Specifically, the level-wise classification performance is evaluated on all combinations of the remaining front-ends and the best one chosen based on the development set.

For the experiments on the NIST 2007 database reported in this paper, the best performance at Level 3 was achieved using a combination of PLLR features from HU and CZ phone recognizers. In Level 2, the PLLR (HU) front-end was chosen and finally in Level 1, a combination of MFCCs, PLLR (HU) and PLLR (CZ) was chosen. All classification systems based on multiple front-ends utilise the i-vector concatenation framework [13].

3.2 Evaluation Metrics

Typically LID systems are evaluated using metrics such as accuracy and error rate [9], which treat all misclassifications identically. i.e., an English utterance misidentified as Spanish or Vietnamese are both penalised equally. However, in a hierarchical framework multiple levels of classification exist and it is reasonable to expect that misclassifications at higher levels are penalised more than those at lower levels. Therefore hierarchical Precision and hierarchical Recall (hP and hR) [20], which assign partial credit for correct classification at each level are used in addition to the traditional identification rate (IDR) in the work reported in this paper. These measures are calculated as follows:

$$IDR = \frac{T_C}{T_C + T_i} \tag{5}$$

Where T_C and T_i are the number of correctly and incorrectly identified test instances.

$$hP = \frac{\sum_i |\hat{C}_i \cap C_i|}{\sum_i |C_i|} \quad (6)$$

and

$$hR = \frac{\sum_i |\hat{C}_i \cap C_i|}{\sum_i |\hat{C}_i|} \quad (7)$$

Where $\hat{C}_i$ and C_i are sets of true and predicted class labels at all levels of the hierarchical structure for the i^{th} test utterance.

For a given test instance, IDR only depicts either it's correctly classified or not. Whereas from hierarchical precision, we can analyse at which hierarchy level the test instance is misclassified. For example, in this particular hierarchy (Figure 2), a single test instance can lead to one of four different values for hP: 0, 0.5, 0.67 or 1. If the test instance is misclassified at the1st, 2nd or 3rd level of hierarchy, then hP takes values of 0, 0.5 or 0.67 respectively. If the test instance is correctly classified, hP takes a value of 1.

3.3 Baseline System

The performance of the proposed HLID framework is compared to the baseline system that comprises of three LID sub-systems which are fused to give the final decision [15]. The three sub-systems are based on i-vectors estimated using front-ends that compute PLLRs using Hungarian, Russian and Czech phonemes. The baseline system and the proposed hierarchical system were both trained on the same training data.

4. RESULTS

Figure 4 shows the baseline and HLID system performance in terms of identification rates (IDR) of all 14 target languages in NIST 2007. The average identification rate achieved by the proposed hierarchical framework across all 14 languages is 96.02% which is significantly higher than the 90.36% identification rate achieved by the baseline system.

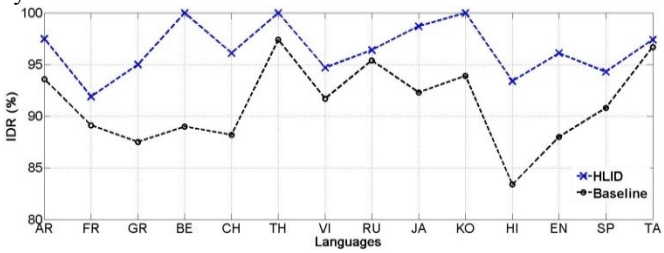

Figure 4: System performance in terms of Identification rate (IDR)

Table 1 shows the performance of HLID system in terms of hP and hR computed over all test utterances. It can be seen that both hierarchical precision and hierarchical recall are consistently high across all 14 languages. Finally, in order to determine if the level-wise feature selection provided any advantage, the number of misclassifications in the individual classification problems in each stage of the hierarchical structure is compared to that obtained by the baseline system, and is shown in Table 2. For instance, when comparing at level 1 (refer Figure 2), the number of misclassification between the clusters C_{12}, C_{34}, C_5 and C_6 in the hierarchical structure is compared to the number of instances where a test utterance corresponding to one of the languages belonging to each of these clusters is misclassified as corresponding to a language that is in another cluster. It should be noted that in this case, misclassifications between languages within the same cluster are ignored. i.e., misclassifications between AR, FA, GE and BE are ignored since all 4 languages correspond to the same cluster, C_{12}.

Table 1: HLID performance in terms of hierarchical precision

Languages	hP (%)	hR (%)	Languages	hP (%)	hR (%)
Arabic (AR)	95.1	90.0	Russian (RU)	95.9	92.5
Farsi (FR)	96.2	91.0	Japanese (JA)	96.1	92.9
German (GR)	96.0	90.4	Korean (KO)	100	100
Bengali (BE)	100	100	Hindi (HI)	92.3	90.1
Chinese (CH)	97.9	96.5	English (EN)	93.5	91.9
Thai (TH)	100	100	Spanish (SP)	91.7	90.4
Vietnamese (VI)	91.5	90.6	Tamil (TA)	97.7	96.4

Table 2 lists these misclassification error rates for all the individual classification problems in all 3 levels of the hierarchical structure. The use of hierarchical structure reduces the average misclassification by 68.2%, 80.0% and 84.25% in 1st, 2nd and 3rdlevel respectively compared to baseline.

Table 2: Misclassifications between languages/clusters

Level	Confusion between Clusters	Misclassifications errorrate		Error reduction
		Baseline	HLID	
1	C_{12},C_{34},C_5,C_6	5.9 %	1.9 %	4 %
2	C_1,C_2	1.5 %	0.62 %	0.9 %
2	C_3,C_4	1.0 %	0 %	1 %
3	AR, GR, FR	4.2 %	0.83 %	3.4 %
3	CH, TH, VI	6.2 %	1.5 %	4.7 %
3	JA, KO	0 %	0 %	0 %
3	HI, EN, SP, TA	8.2 %	1.4 %	6.8 %

5. CONCLUSION

This paper proposes a novel hierarchical framework for language identification. Specifically, it combines automatic clustering of languages to form the hierarchical structure with the selection of a suitable level-wise front ends and thus does not make the common assumption that any one front-end is the best choice for the identification of all languages. Finally the log-likelihood ratios at all levels of the hierarchical structure are propagated down to the lowest level representing the target languages and final identification is based on these combined log-likelihood ratios. All experimental results suggest that the proposed hierarchical framework outperforms the non-hierarchical baseline which is one of the best performing language identification system currently reported in the literature.

7. REFERENCES

[1] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: A tutorial," Circuits and Systems Magazine, IEEE, vol. 11, pp. 82-108, 2011.

[2] H. Li, B. Ma, and K. A. Lee, "Spoken language recognition: from fundamentals to practice," Proceedings of the IEEE, vol. 101, pp. 1136-1159, 2013.

[3] D. A. Reynolds, W. M. Campbell, W. Shen, and E. Singer, "Automatic language recognition via spectral and token based approaches," in Springer Handbook of Speech Processing, ed: Springer, 2008, pp. 811-824.

[4] L. D'Haro, R. Cordoba, C. Salamea, and J. Echeverry, "Extended phone log-likelihood ratio features and acoustic-based i-vectors for language recognition," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014, pp. 5342-5346.

[5] M. Diez, A. Varona, M. Penagarikano, L. J. Rodriguez-Fuentes, and G. Bordel, "New Insight into the use of Phone Log-Likelihood Ratios as Features for Language Recognition," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[6] M. Díez, A. Varona, M. Peñagarikano, L. J. Rodríguez-Fuentes, and G. Bordel, "On the use of phone log-likelihood ratios as features in spoken language recognition," in SLT, 2012, pp. 274-279.

[7] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in INTERSPEECH, 2011, pp. 857-860.

[8] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen, "iVector Approach to Phonotactic Language Recognition," in INTERSPEECH, 2011, pp. 2913-2916.

[9] B. Yin, E. Ambikairajah, and F. Chen, "Hierarchical language identification based on automatic language clustering," in INTERSPEECH, 2007, pp. 178-181.

[10] B. Yin, E. Ambikairajah, and F. Chen, "Improvements on hierarchical language identification based on automatic language clustering," in Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, 2008, pp. 4241-4244.

[11] A. Bouguettaya, Q. Yu, X. Liu, X. Zhou, and A. Song, "Efficient agglomerative hierarchical clustering," Expert Systems with Applications, vol. 42, pp. 2785-2797, 2015.

[12] J. Bing, S. Yan, and D. Li-Rong, "Exploiting language cluster information for language pair identification," in Audio, Language and Image Processing (ICALIP), 2012 International Conference on, 2012, pp. 1005-1009.

[13] Z.-Y. Li, W.-Q. Zhang, L. He, and J. Liu, "Complementary combination in i-vector level for language recognition," in Odyssey 2012-The Speaker and Language Recognition Workshop, 2012.

[14] L. F. D'Haro, R. Cordoba, C. Salamea, and J. Ferreiros, "Language Recognition using Phonotactic-based Shifted Delta Coefficients and Multiple Phone Recognizers," in Fifteenth Annual Conference of the International Speech Communication Association, 2014.

[15] S. Irtza, V. Sethu, P. Le, E. Ambikairajah and H. Li " Phonemes Frequency based PLLR Dimensionality Reduction for Language Recognition," accepted in INTERSPEECH, 2015.

[16] S. Kiritchenko, S. Matwin, and F. Famili, "Functional annotation of genes using hierarchical text categorization," in bioLINK sig, ISMB, 2005.

[17] M. Diez, A. Varona, M. Penagarikano, L. J. Rodríguez-Fuentes, and G. Bordel, "Dimensionality reduction of phone log-likelihood ratio features for spoken language recognition," in INTERSPEECH, 2013, pp. 64-68.

[18] P. Matejka, L. Burget, O. Glembek, P. Schwarz, V. Hubeika, M. Fapso, et al., "BUT system description for NIST LRE 2007," in Proc. 2007 NIST Language Recognition Evaluation Workshop, 2007, pp. 1-5.

[19] P. Schwarz, "Phoneme recognition based on long temporal context," Ph.D. dissertation, Faculty of Information Technology, Brno University of Technology, http://www.fit.vutbr.cz/, Brno, Czech Republic, 2008.

[20] Florian Eyben, Martin Wöllmer, Björn Schuller: "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor", In Proc. ACM Multimedia (MM), ACM, Florence, Italy, ACM, ISBN 978-1-60558-933-6, pp. 1459-1462, October 2010. doi:10.1145/1873951.1874246