IMPROVEMENTS ON HIERARCHICAL LANGUAGE IDENTIFICATION BASED ON AUTOMATIC LANGUAGE CLUSTERING

Bo Yin[1,2], Eliathamby Ambikairajah[1,2], Fang Chen[2,1]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]National ICT Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia
bo.yin@student.unsw.edu.au, ambi@ee.unsw.edu.au, fang.chen@nicta.com.au

ABSTRACT

Hierarchical Language Identification (HLID) is a novel framework for combining multiple features or primary systems in language identification. In this paper, several key components of HLID are investigated and developed. Crossing Likelihood Ratio and Kullback-Leibler distance measures are introduced for faster and more accurate clustering. A novel feature selection scheme based on fusion is proposed to incorporate multiple features at each classification level. Further, a Phone Recognizer followed by Language Model (PRLM) system is introduced in addition to the other three acoustic systems to provide phonetic information. These proposed techniques improve the performance of HLID system to an EER of 6.3% on the NIST LRE 2003 30s task.

Index Terms – language identification, fusion, language clustering

1. INTRODUCTION

Language Identification (LID) has drawn much attention recently, due to the challenge of multi-lingual speech recognition. To identify which language is spoken from a speech utterance, traditionally, an individual language model is created for each possible language, and the utterance is classified by measuring the distances between this utterance and each one of language models. This basic idea works well with a single feature but needs to be expanded to benefit from multiple features. Much effort has been spent, therefore, utilizing fusion techniques to integrate varied LID systems which capture discriminative information from different features [1-3].

Fusion-based approaches are remarkably popular in modern LID systems for their ability to integrate acoustic and phonetic LID systems, which previously were competitive approaches. Among existing fusion techniques, the Gaussian Mixture Model (GMM)–based fusion technique is one of the best performing and most widely adopted approaches. In this approach, multiple LID systems using different classifiers or features (referred as 'Primary LID systems') are integrated [4]. The likelihood scores produced by the primary LID systems are concatenated to form the input vector of a low-mixture GMM classifier. The likelihood scores produced by this GMM classifier then are used as the output of the integrated fusion system.

However, the GMM fusion technique experiences difficulty in improving performance when the number of languages and features increases [5], because all language hypotheses are examined at a single level. This means the variation of the distances between languages in different feature spaces is not sufficiently considered.

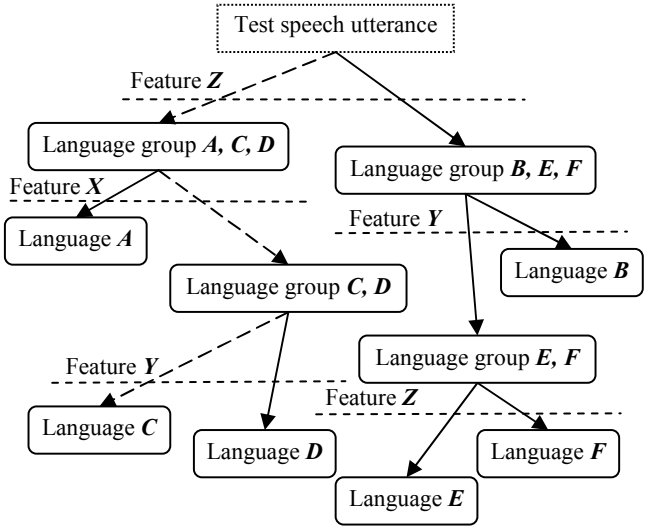

Figure 1: An example classification diagram for Hierarchical Language Identification

To improve matters, the Hierarchical LID (HLID) framework has been proposed [5]. This multi-level classification approach aims to ensure that the most discriminative hyper-plane in feature space is used at each classification level and therefore achieve the best performance. In the HLID framework language hypotheses are clustered hierarchically to form a tree structure (Fig. 1). In this structure, each leaf is an individual language hypothesis, and other nodes are language groups containing

language hypotheses from their child nodes. The test speech utterance is classified level-by-level, according to the most discriminative feature at each level (the dashed arrow lines in Fig. 1 shows an example of classification path).

To expose the most discriminative hyper-plane in feature space at each classification level, an unsupervised agglomerative clustering process has been proposed to create the hierarchical classification structure and to select the most discriminative feature at each level [5]. Based on a performance-based distance measure, the most similar languages or language groups are merged to form bigger groups level-by-level from bottom to top. This clustering process ensures that at each level the distances between language groups are always larger than the distances between the languages within a language group. Also, language group models are more precise and robust than individual language models because the languages within a group share similar characteristics (close to each other in feature space), and therefore the larger amount of training data (from all languages within the group) helps model training.

To further improve initial HLID alternative distance measures are explored in this paper, an enhanced feature selection scheme is proposed, and a phonetic LID system is added as another primary LID system.

2. ALTERNATIVE DISTANCE MEASURES

Distance measures play a key role in HLID, because language clustering is based solely on the distance between languages and/or language groups. Specifically, an objective distance measure is particularly required for unsupervised language clustering. Three different types of distance measures are discussed and compared below.

2.1. Performance-based distance

The most straight-forward performance-based distance between two languages or language groups is probably the accuracy of identifying them from each other. Higher accuracy indicates that they are easier to discriminate (larger distance). This distance can be defined as [5]:

$$d_{PERF}(\lambda_x, \lambda_y) = A_{LID}(\lambda_x, \lambda_y) \mid \lambda_{UBM} \qquad (1)$$

where $d_{PERF}(\lambda_x, \lambda_y)$ is the distance between cluster λ_x and λ_y, $A_{LID}(\lambda_x, \lambda_y)$ is the LID accuracy of these two clusters, and λ_{UBM} is the Universal Background Model (UBM) trained from all clusters. Here a cluster is either a single language or a language group.

To calculate this distance, a series of pair-wise LID experiments are performed at each classification level. To reduce the computation cost, all cluster models are adapted from the same Universal Background Model (UBM) which is trained on all available training data.

The performance-based distance relies on the evaluation on the development dataset. It may be unstable if the development data is insufficient. Furthermore, the computation cost is relatively high because the complete LID training/evaluation process is required at each classification level.

2.2. Likelihood-based distance

The likelihood ratio is a commonly used distance measure for statistical models such as GMMs. The basic idea of this distance is to measure the dissimilarity between two models by computing and comparing the likelihood scores from each model to the same test data. Two common likelihood ratio based distances are Generalized Likelihood Ratio (GLR) and Cross Likelihood Ratio (CLR) [6] distances. They are defined as:

$$d_{GLR}(\lambda_x, \lambda_y) = \log\left(\frac{L(X \mid \lambda_x) \cdot L(Y \mid \lambda_y)}{L(X \cup Y \mid \lambda_{X \cup Y})}\right) \qquad (2)$$

And

$$d_{CLR}(\lambda_x, \lambda_y) = \log\left(\frac{L(X \mid \lambda_x)}{L(X \mid \lambda_y)}\right) + \log\left(\frac{L(Y \mid \lambda_y)}{L(Y \mid \lambda_x)}\right) \qquad (3)$$

where $L(.)$ is the likelihood function, X, Y and $X \cup Y$ are feature vectors from two different clusters and their combination, and λ_x, λ_y and $\lambda_{X \cup Y}$ are the GMMs trained on each cluster and their combination.

While it is true that CLR is less accurate than GLR [6], the advantage of CLR compared to GLR is the lower computation cost since there is no need to train a new model such as $\lambda_{X \cup Y}$.

2.3. Model-based distance

Since both the performance-based and likelihood-based distances are calculated from data samples, runtime error may be introduced due to data bias/insufficiency, and the computation cost is high since calculations are done sample by sample. It would be more effective and more efficient if the distance can be directly measured between cluster models without any data. This type of distance is referred to as model-based distance.

In case of GMMs, methods for measuring distance between single Gaussians have been well developed, ranging from the simple Euclidean distance $\|\mu_1 - \mu_2\|$ to the more sophisticated Kullback-Leibler divergence (KL divergence) [7]:

$$d_{KL}(\lambda_1, \lambda_2) = \frac{\sigma_1}{\sigma_2} + \frac{\sigma_2}{\sigma_1} + \frac{(\mu_1 - \mu_2)^2}{\sigma_1} + \frac{(\mu_1 - \mu_2)^2}{\sigma_2} \qquad (4)$$

where μ_1, μ_2 and σ_1, σ_2 are means and variances of the Gaussians λ_1 and λ_2 respectively. Strictly, KL divergence is not a distance as it fails the triangle inequality, yet in this

paper we refer it as a distance measure (KL distance) as it does measure the similarity between Gaussians.

The basic KL distance is expanded to allow the measuring of distance between two mixture collections (entire GMMs) in [7]. The distance between two GMMs λ_x and λ_y (referred as 'Model-KL' distance) is defined as:

$$d_{GMM}(\lambda_x, \lambda_y) = \frac{\sum_{i=1}^{M} W_i^x + \sum_{j=1}^{N} W_j^y}{\sum_{i=1}^{M} w_i^x + \sum_{j=1}^{N} w_j^y} \quad (5)$$

where

$$W_i^x = w_i^x \cdot \min_{j=1..N}(d_{KL}(\lambda_i^x, \lambda_j^y)) \quad (6)$$

$$W_j^y = w_j^y \cdot \min_{j=1..N}(d_{KL}(\lambda_j^y, \lambda_i^x)) \quad (7)$$

The KL distance between two single mixtures is [6]:

$$d_{KL}(\lambda_i^x, \lambda_j^y) = 0.5 \cdot \left(\vec{\mu}_j^y - \vec{\mu}_i^x\right)^T \left(\frac{1}{\Sigma^y} + \frac{1}{\Sigma^x}\right)\left(\vec{\mu}_j^y - \vec{\mu}_i^x\right) + 0.5 \cdot tr\left(\frac{\Sigma^x}{\Sigma^y} + \frac{\Sigma^y}{\Sigma^x} - 2 \cdot I\right) \quad (8)$$

where w_i^x is the weight of λ_i^x, the ***i***th mixture of model ***x***. Σ is the covariance matrix. ***I*** is the identity matrix and ***tr(.)*** is the trace function.

Although model-KL distance measure is faster and more accurate than other distance measures, it is relatively difficult to apply to models other than a mixture model.

3. FUSION-BASED FEATURE SELECTION

In the single feature selection scheme [5], the most discriminative feature (or primary system) is selected as the only effective feature in each classification level. Although this feature selection process introduces one of the most important advantages of HLID – the most discriminative information of that particular classification level is used and emphasized – it is also true that the less valuable (but still useful) information provided by other features is discarded.

To improve on this, a fusion-based feature selection scheme is proposed. In this scheme, the classifier which utilizes only the most discriminative feature (shown as Feature X, Y, etc. in Figure 1) is replaced by a GMM-fusion-based classifier to integrate different primary language identification systems at each classification level. The likelihood scores produced by the primary LID systems are concatenated to form the input vector of the fuser which utilizes 16-mixture GMMs. The output of each fuser is used as the classification result of the corresponding level. In this fusion-based feature selection scheme, each GMM fuser is separately trained on the language/language groups in the associated classification level.

4. PHONETIC PRIMARY SYSTEM

Although Parallel Phone Recognizer followed by Language Model (PPRLM) is more popular these days, the Phone Recognizer followed by Language Model (PRLM) system is still one of the most classical and effective LID systems utilizing phonetic information. Considering that the purpose of introducing another primary system is to investigate whether a phonetic system will benefit the HLID performance, a simpler PRLM system is acceptable in this research. In this paper, a PRLM system is developed which utilizes a uni-phone recognizer and a series of bi-gram language models, as shown in Fig. 2.

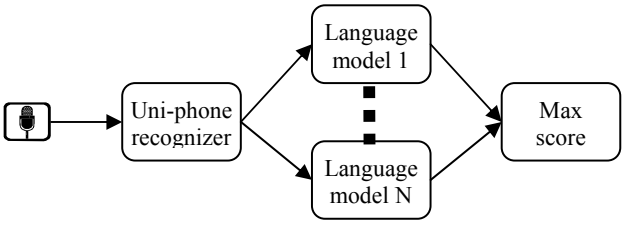

Figure 2: Diagram of proposed PRLM system

The uni-phone recognizer is created based on neural networks [8], which are trained on labeled data from the OGI-TS speech database. The training data contains around 1 hour of speech from six languages (English, German, Hindi, Japanese, Mandarin, and Spanish). The phoneme sets of these six languages are merged and the training data from all six languages is used for training the single phone recognizer. For each of the 12 languages in the CallFriend database, individual bi-gram language models are then trained on the phoneme sequences produced by the uni-phone recognizer from the CallFriend training set. During evaluation, the final decision is made by selecting the highest scored model among all language models.

5. EXPERIMENTS

Four primary LID systems were developed for the experiments in this paper. Three of them were acoustic LID systems accepting different speech features. They were all based on the same GMM-based classifier with 256 mixtures, Univeral Background Model (UBM) adaptation and fast scoring [9]. The features used by these systems varied from MFCC with 7 coefficients (primary LID system 1), pitch and intensity (system 2), to the concatenation of these features (system 3) [5]. In all three systems, the features along with the corresponding Shifted Delta Coefficients (SDC) were normalized by segmental histogram equalization. The PRLM system described in the previous sections was used as System 4.

With the exception of the labeled corpus used for training the phone recognizer in the PRLM system, all experiments were conducted on the CallFriend database, which contains 60 half-hour telephone conversations for each of its 12 languages. This database was separated into three equally sized sets to act as training, developing and testing data sets. To save time, a smaller dataset was selected from the CallFriend training sets for the clustering

process. This smaller set consisted of around 2 hours of speech for each language, and was well balanced with different speakers.

Three Hierarchical LID models (incl. clustering structures, fusers at each level, and language/language group models) were created on all four primary systems: one used the single feature selection with pair-wise LID accuracy based distance; the other two used fusion-based feature selection with the accuracy based and the Cross Likelihood Ratio (CLR) distance respectively. Two additional HLID models were created on the three acoustic primary systems for comparing the model Kullback-Leibler (model-KL) and CLR distance measures, because the model-KL distance measure is relatively difficult to apply to bi-gram based language models. The cluster structures of these models are slightly varied.

Individual primary systems, baseline GMM fusion system and the Hierarchical LID systems under different configurations were evaluated on the NIST LRE 2003 task (30s only, primary condition, no Russian). The performances are reported in Table 1.

Table 1: Equal Error Rate (EER) of varied systems in NIST LRE 2003 30s tasks

SYSTEM	EER%
Primary system 1 (MFCC)	11.9
Primary system 2 (Pitch+Intensity)	25.3
Primary system 3 (MFCC+Pitch+Intensity)	9.2
Primary system 4 (PRLM)	14.6
GMM fusion system (incl. all primary systems)	7.5
HLID system (incl. all primary systems) - single feature selection - pair-wise LID accuracy based distance measure	7.1
HLID system (incl. all primary systems) - fusion based feature selection - pair-wise LID accuracy based distance measure	6.4
HLID system (incl. all primary systems) - fusion based feature selection - CLR distance measure	6.3
HLID system (incl. primary system 1, 2, 3) - fusion based feature selection - CLR distance measure	8.5
HLID system (incl. primary system 1, 2, 3) - fusion based feature selection - model-KL distance measure	8.3

It can be observed that the HLID systems perform better than the conventional GMM fusion system. The novel fusion-based feature selection provides a relative improvement of 9.8% to the EER when compared to the existing single feature selection technique. The CLR and model-KL distance measures perform similarly while the CLR measure outperforms the pair-wise LID accuracy based distance measure. The introduction of the PRLM system improves the overall performance further to 6.3% EER, which is comparable to other modern LID systems.

6. CONCLUSION

Hierarchical LID is a novel framework for combining multiple features or primary systems in language identification. In this paper, several improvements on HLID have been achieved. While both the CLR and model-KL distance measures outperform the existing performance-based distance measure, the model-KL distance measure requires much less computation and therefore reduces training time. The proposed novel fusion-based feature selection technique also shows a remarkable improvement compared to the existing single feature selection. Introducing a PRLM system further improves the performance. The best performing HLID system achieves a 16.0% relative improvement compared to the baseline system utilizing the popular GMM-based fusion technique in the NIST LRE 2003 30s task.

7. REFERENCES

[1] J. Gutierrez, J. L. Rouas, and R. Andre-Obrecht, "Fusing language identification systems using performance confidence indexes," IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal Canada, 2004.

[2] E. Wong and S. Sridharan, "Fusion of Output Scores on Language Identification System," Workshop on Multilingual Speech and Language Processing, Aalborg Denmark, 2001.

[3] B. Yin, E. Ambikairajah, and F. Chen, "A Novel Weighting Technique for Fusing Language Identification Systems based on Pair-wise Performances," IEEE workshop on Automatic Speech Recognition and Understanding (ASRU), Tokyo, Japan, 2007.

[4] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, Phonetic, and Discriminative approaches to Automatic Language Identification," EuroSpeech, Geneva, Switzerland, 2003.

[5] B. Yin, E. Ambikairajah, and F. Chen, "Hierarchical Language Identification based on Automatic Language Clustering," InterSpeech - EuroSpeech, Antwerp, Belgium, 2007.

[6] T. Stadelmann and B. Freisleben, "Fast and Robust Speaker Clustering Using the Earth Mover'S Distance and Mixmax Models," Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, 2006.

[7] H. S. M. Beigi, S. H. Maes, and J. S. Sorensen, "A distance measure between collections of distributions and its application to speaker recognition," Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on, 1998.

[8] P. Schwarz, P. Matějka, and J. Černocký, "Towards Lower Error Rates In Phoneme Recognition," TSD 2004, Brno, Czech Republic, 2004.

[9] B. Yin, E. Ambikairajah, and F. Chen, "Combining Prosodic and Cepstral Features in Language Identification," IEEE International Conference on Pattern Recognition, Hong Kong, China, 2006.