A NOVEL WEIGHTING TECHNIQUE FOR FUSING LANGUAGE IDENTIFICATION SYSTEMS BASED ON PAIR-WISE PERFORMANCES

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

One of the key research issues in modern Language Identification (LID) research is how best to combine multiple approaches with different features. Existing statistical fusion techniques are popular but have serious limitations when development data is insufficient, since the data is used for training the statistical fuser. In this paper we compare existing fusion techniques for LID systems and propose an alternative to reduce this problem. By deriving the language-specific weighting directly from pair-wise LID performance, a novel weighting approach is introduced and implemented. Experiments on the NIST LRE 2003 task (CallFriend database) and OGI-TS databases demonstrate that the proposed weighting technique outperforms other recent fusion techniques when the available development data is limited.

***Index Terms*—** Language identification, language recognition, fusion, weighting

1. INTRODUCTION

The purpose of Language Identification (LID) is to determine from segments of speech the language being spoken. LID is an important component of multilingual speech-based user interfaces. Recently, many cues that contribute to the intelligibility of spoken languages have been discovered[1], enabling discrimination between different languages, e.g. spectrum, prosody, phoneme, and group delay. Various speech classification systems have utilized these or related cues, and have proven effective for LID[2, 3]. To combine all this useful information together accurately and reliably, most state-of-the-art systems use some form of 'hybrid' approach. In this type of approach, either different features are mixed, known as 'Feature combination', or the likelihood scores produced by different 'primary' LID systems are fused to produce a new set of likelihood scores, known as 'Fusion'. The latter fusion approach is more widely utilized because it is more flexible when combining different types of primary systems.

The major challenge of fusion, then, is to decide how to produce the final likelihood scores, based on the likelihood scores produced by primary LID systems. Since the exact relationship between these two sets of likelihood scores is unknown, most fusion techniques try to model this relationship, based on the test cases on development dataset, and then apply the modeled relationship to the fusion process. Specifically, either a statistical classifier or an empirical weighting process is utilized.

In this paper, the most popular existing fusion techniques [3-5] in LID systems are analyzed and compared. An alternative fusion approach is introduced, where likelihood scores from the different primary LID systems are weighted and combined. These weights are directly derived from pair-wise LID performances on the available development dataset. The weights in this approach are not only different for each primary LID system, but also vary between different language hypotheses.

2. EXISTING FUSION-BASED LID

Normally a Language Identification system utilizes a single feature set and a single classifier. However, current LID systems incorporate several individual LID systems (known as 'Primary LID systems') by combining their likelihood scores. Therefore, each primary LID system is actually capable of identifying the language by itself. However, combined likelihood scores generally result in a higher identification rate. The likelihood scores produced by the primary LID systems are obtained by estimating the probability of a test segment belonging to a target language. A 'Language hypothesis' refers to the process of selecting a particular language as the target language when testing.

There are two major different types of fusion techniques[6]: empirical fusion, such as sum-based or product-based weighting; and statistical fusion, such as Gaussian Mixture Model (GMM) or Artificial Neural Network (ANN) fusion[3-5].

In empirical fusion, the final likelihood score when testing a given language hypothesis is either a weighted sum or product of the likelihood scores produced by the primary LID systems. An empirical process is used to find the optimum weighting coefficients which contribute to the highest performance, e.g. Linear Score Weighting (LSW) [4]. Sometimes performance related weightings are used, e.g. 'Matcher weighting' [7]. The sum and product based weighting techniques always give the likelihood scores for different language hypotheses the same weights, as long as they were produced by the same primary LID system.

GMM-based fusion perhaps is the most popular fusion technique in recent research[3]. In such systems, the likelihood scores produced by primary LID systems are used to train a GMM classifier. The likelihood scores obtained from this classifier are then used for final decision.

Although GMM-based fusion implicitly considers the difference between contributions towards different languages from the same primary LID system, the performance will depend on the size of the available development data. The overall performance of the LID system will deteriorate if the amount of development data is insufficient for adequate training of the GMM backend.

ANN classifiers can also be used to fuse likelihood scores produced by primary LID systems[8]. With regard to the network structure, one hidden layer and one output layer has been shown to achieve reasonable performance[8]. The number of perceptrons in hidden layers and the activation functions in each layer need to be optimized on the development dataset.

ANN-based fusion considers the language-dependent contribution of primary LID systems. However, it faces a similar problem to GMM-based fusion, where the performance will deteriorate if sufficient development data is not available for training.

3. PAIR-WISE PERFORMANCE BASED WEIGHTING TECHNIQUE

Although statistical fusion techniques outperform empirical techniques in most cases, the performance is heavily dependent on the sufficiency of training for the fuser. To create a more robust fusion technique, we propose an alternative weighting technique which derives a set of language-dependent weighting coefficients from pair-wise LID performances. Each of these performances shows that how a specific feature benefits the discrimination between a particular pair of languages.

3.1. Measuring feature-specific contribution to each language

Modeling process in statistical fusion techniques are always involved in training a statistical model. Although the statistical model may be a wise choice for an unknown relationship, it doesn't directly take into account any available prior knowledge about the relationship.

When discriminating a particular language pair, the performance of a specific primary LID system shows the language-dependent contribution from this system. Put all the pair-wise LIDs between one particular language and each of the other languages into a group, the average performance of this group can be used to measure the average contribution of a specific primary LID system to that particular language. Based on these contributions a weighting scheme can be developed to better combine the output scores from different primary systems.

Table 1 shows an example of pair-wise LID performances from two primary LID systems with different features (MFCC and prosodic) on the OGI-TS database. It clearly shows that the contribution of one primary LID system to different languages varies, e.g. MFCCs contribute more than prosodic information to Tamil (ta) because of the lower error rates, and vice-versa to Japanese (ja). The contribution of a specific LID system to English can therefore be estimated as the average of all pair-wise performances in the first column.

Table 1. Pair-wise LID performance (error rate %) on OGI-TS database from two primary LID systems using MFCC/Prosodic features

	fa	vi	ja	fr	en	ge	ko	ma	sp	ta
fa	0.0/0.0	8.5/8.5	2.1/2.1	6.5/11.3	1.8/17.9	1.8/18.2	2.0/6.0	2.0/4.1	3.4/12.1	2.0/4.0
vi	8.5/8.5	0.0/0.0	0.0/0.0	3.8/5.7	4.3/6.4	4.3/2.2	4.9/4.9	2.5/2.5	4.1/6.1	7.3/2.4
ja	2.1/2.1	0.0/0.0	0.0/0.0	5.6/1.9	0.0/2.1	0.0/2.1	11.9/0.0	4.9/0.0	14.0/0.0	0.0/4.8
fr	6.5/11.3	3.8/5.7	5.6/1.9	0.0/0.0	1.6/14.5	11.5/16.4	3.6/17.9	9.1/1.8	3.1/20.3	5.4/19.6
en	1.8/17.9	4.3/6.4	0.0/2.1	1.6/14.5	0.0/0.0	1.8/16.4	4.0/10.0	2.0/6.1	5.2/8.6	2.0/14.0
ge	1.8/18.2	4.3/2.2	0.0/2.1	11.5/16.4	1.8/16.4	0.0/0.0	4.1/10.2	0.0/2.1	10.5/15.8	6.1/22.4
ko	2.0/6.0	4.9/4.9	11.9/0.0	3.6/17.9	4.0/10.0	4.1/10.2	0.0/0.0	2.3/7.0	5.8/26.9	2.3/13.6
ma	2.0/4.1	2.5/2.5	4.9/0.0	9.1/1.8	2.0/6.1	0.0/2.1	2.3/7.0	0.0/0.0	3.9/2.0	0.0/2.3
sp	3.4/12.1	4.1/6.1	14.0/0.0	3.1/20.3	5.2/8.6	10.5/15.8	5.8/26.9	3.9/2.0	0.0/0.0	3.8/15.4
ta	2.0/4.0	7.3/2.4	0.0/4.8	5.4/19.6	2.0/14.0	6.1/22.4	2.3/13.6	0.0/2.3	3.8/15.4	0.0/0.0
AVG.	**3.0/8.4**	**4.0/3.9**	**3.8/1.3**	**5.0/10.9**	**2.3/9.6**	**4.0/10.6**	**4.1/9.6**	**2.7/2.8**	**5.4/10.7**	**2.9/9.9**

3.2. Weighting scheme based on pair-wise performances

When applying the proposed weighting technique (see Figure 1), the final likelihood score for language hypothesis ***i*** is calculated as following:

$$L_i = \sum_{j=1}^{N} w_{ij} \cdot l_{ij} \qquad (1)$$

where $\boldsymbol{l_{ij}}$ is the likelihood score produced by the primary LID system ***j*** for language hypothesis ***i***, $\boldsymbol{w_{ij}}$ is the weighting coefficient applied to primary LID system ***j*** and language hypothesis ***i***, and ***N*** is the total number of primary LID systems.

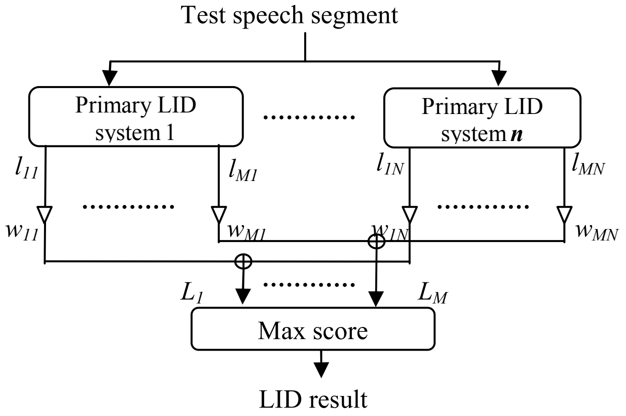

Figure 1. The proposed weighting scheme

The proposed weighting scheme is similar to the sum based weighting scheme described in section 2.1. However, in this weighting scheme different weights $\boldsymbol{w_{ij}}$ are applied not only to different primary LID systems ***j*** but also to different language hypotheses ***i*** produced by the same primary LID system.

Equation 1 can be viewed as a linear perceptron. However, instead of obtaining the weights by minimizing a generic cost function as is done in a linear perceptron, in our proposed scheme we determine the weights based on language-dependent contributions of the primary LID systems.

The goal of choosing weighting coefficients here is to estimate the language-dependent contribution on a primary LID system basis from the development data. Generally, for ***M*** languages and ***N*** primary LID systems, the weighting coefficient (which measures contribution) for language ***i*** and primary LID system ***j*** is calculated by:

$$w_{ij} = -\frac{1}{M}\sum_{k=1}^{M}\log(e_{ik,j}) \qquad (2)$$

where $\boldsymbol{e_{ik,j}}$ is the error-rate of language-pair ***i*** and ***k*** tested by primary LID system ***j***. A log function is used here for improving the significance of the variation of error-rates (which are mostly less than 0.2).

With the advantage of considering language-dependent contribution, the proposed weighting scheme is also shown to be more robust than existing major statistical fusion techniques (examined in section 4). Also, the training process requires less time than that of statistical fusion techniques.

4. EXPERIMENTS AND RESULTS

To compare different fusion techniques, four different fusion systems were implemented: (i) Linear Score Weighting (LSW); (ii) Gaussian Mixture Models (GMM); (iii) Artificial Neural Networks (ANN); and (iv) the proposed weighting technique. All of these fusion systems were implemented with the same two primary LID systems: a GMM LID system utilizing MFCC as features; and a GMM LID system utilizing prosodic features[2, 9]. The number of MFCC coefficients was chosen to be seven for optimal results[2]. The prosodic features consisted of pitch and log-energy. Although the primary LID systems were not directly comparable to state-of-the-art LID systems, they were acceptable in this work as the purpose was to compare fusion techniques, not the primary LID system performance *per se*.

For the LSW fusion system, optimal weighting coefficients were found empirically from testing on development dataset. The GMM fusion system utilizes a 16-mixture model for each language[3]. For the ANN fusion system a network was created with 36 inputs, one hidden layer, and 12 outputs. The optimum performance was obtained when using 16 tan-sigmoid perceptrons in the hidden layer, and softmax perceptrons in the output layer[8].

4.1. Experiments on OGI-TS database

The OGI-92 telephony speech database is a multi-language, multi-speaker corpus, composed of a minimum of 90 calls (approx. 2 minutes each, different speakers for different calls) in 10 languages. 50 of these calls were used as the training set, 10 or 20 as the development set (to test performance on different data-set sizes), and the remaining 20 calls were used as an evaluation set.

The results of different fusion systems on all 10 languages when using development datasets of different sizes are shown in Table 2. The optimized weighting coefficients of LSW fusion were 0.80 for the primary system with MFCC features and 0.20 for the system with prosodic features in the 20-call development data case, 0.85 and 0.15 in the 10-call development data case respectively. Two sets of target utterances, of 20 seconds and 10 seconds duration, were tested to investigate performance differences between utterances of differing duration.

When using 20 second utterances, and when the size of development dataset was reduced from 20 calls to 10 calls, the proposed weighting system did not experience any drop in performance (9.3% error-rate) while the performance of the other fusion systems dropped significantly (Table 2). Similar trends were observed when using 10 second utterances. The proposed weighting system achieved the highest performance in all situations.

Table 2. **Error rates** for different fusion systems, with 20 and 10 call development datasets, for 20s and 10s duration utterances

System	Size of dev. data (no. of calls)	20secs	10secs
LSW fusion	20	11.0%	18.3%
	10	11.8%	18.5%
GMM fusion	20	12.7%	19.7%
	10	14.5%	21.0%
ANN fusion	20	13.4%	22.4%
	10	16.7%	27.1%
Proposed weighting system	**20**	**9.3%**	**15.1%**
	10	**9.3%**	**15.4%**

4.2. Experiments on CallFriend database

The experiments were repeated using the CallFriend database, based on the recommendations in NIST LRE 2003 tasks[10]. This experiment involved tasks of three durations (3s, 10s, 30s) in 12 languages (English, Arabic, Farsi, French, Mandarin, German, Hindi, Japanese, Spanish, Korean, Tamil and Vietnamese). The CallFriend database contains 60 calls of 30-minute conversation for each language. The calls were separated into three 20-call sets for training, development, and testing purposes. Similarly, the evaluations were repeated with different sized sets of development data (20 and 10 calls). The optimal weighting coefficients for LSW fusion obtained were found to be the same values as those obtained for the OGI database-based experiments. The results are presented (Table 3) as Equal Error Rate (EER), where a lower score indicates better performance.

Table 3. **EER** performance on NIST LRE03 tasks

System	Size of dev. data (no. of calls)	30secs	10secs	3secs
LSW fusion	20	17.6%	21.2%	28.7%
	10	18.0%	22.7%	28.7%
GMM fusion	20	18.8%	22.8%	27.9%
	10	19.3%	23.3%	29.8%
ANN fusion	20	19.4%	25.4%	30.2%
	10	22.1%	29.5%	32.1%
Proposed weighting system	**20**	**15.3%**	**19.8%**	**27.3%**
	10	**15.4%**	**19.9%**	**27.3%**

The above results clearly show a trend similar to the OGI database-based experiments. When the development data size is reduced, the performance degradation of the other fusion systems was more significant than those of the proposed system. Overall, the proposed weighting system achieved comparable or higher performance than other fusion systems to which it is compared in this paper.

5. CONCLUSIONS

In this paper, we compared different existing fusion techniques in LID and proposed an alternative technique. In this proposed weighting technique, the pair-wise LID performances are utilized to measure the language-dependent contribution from each of the primary LID systems. The measured weights then are applied to combine the output scores from different primary systems. The proposed system achieves a more robust performance than the other techniques when the size of development data is reduced. In both OGI and NIST LRE03 tasks, the proposed system shows a comparable (or slightly higher) performance when using the standard development data size, and a higher performance when using a reduced development data size.

ACKNOWLEDGEMENT

This research is supported by National ICT Australia Ltd. (NICTA).

REFERENCES

[1] S. Greenberg and T. Arai, "What are the Essential Cues for Understanding Spoken Languages?," *IEICE Transaction on Information & System*, vol. E87-D, pp. 1059, 2004.

[2] 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.

[3] 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.

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

[5] T. Rong, M. Bin, Z. Donglai, L. Haizhou, and C. Eng Siong, "Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification," IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France, 2006.

[6] 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.

[7] R. Snelick, U. Uludag, A. Mink, M. Indovina, and A. Jain, "Large Scale Evaluation of Multimodal Biometric Authentication Using State-of-the-Art Systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 450-455, 2005.

[8] J. L. Gauvain, A. Messaoudi, and H. Schwenk, "Language Recognition Using Phone Lattices," ICSLP, Jeju island, 2004.

[9] F. Allen, E. Ambikairajah, and J. Epps, "Language Identification Using Warping and the Shifted Delta Cepstrum," IEEE International Workshop on Multimedia Signal Processing, Shanghai, China, 2005.

[10] "NIST Language Recognition Evaluation 2003," http://www.itl.nist.gov/iad/894.01/tests/lang/2003/index.htm