Language-dependent Fusion for Language Identification

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

[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.com.au, ambi@ee.unsw.edu.au, fang.chen@nicta.com.au

Abstract

A novel fusion approach for Language Identification called Language-dependent Fusion (LDF) is presented in this paper. A fusion system is a hybrid system which fuses the results from several individual sub-systems which utilize varied features, models, and/or classifiers. In LDF, instead of applying single fixed weighting coefficients to each sub-system, which happens in conventional approach such as Linear Score Weighting (LSW), varied weighting coefficients are applied to not only each sub-system but also to each language. Furthermore, instead of the experimental and statistical approach, weighting coefficients are calculated from the performance of each language-pair, which reflects the difference among languages. Experiments conducted on the OGI-92 multi-language database demonstrate a remarkable improvement when compared to individual sub-systems (45.46% error rate reduction) and commonly used fusion techniques such as LSW (33.33% error rate reduction) in a 10-language setting. Other advantages of LDF are also discussed.

1. Introduction

Language Identification (LID), which is essential for multi-language speech recognition, has been well researched in recent years. When different models, such as Gaussian Mixture Model (GMM) (Zissman, 1996), Hidden Markov Model (HMM) (Zissman, 1996) and Support Vector Machine (SVM) (Campbell, Singer, & Torres-Carrasquillo, 2004), are applied and features, such as cepstrum (Zissman, 1996), prosody (Yin, Ambikairajah, & Chen, 2006) and phase (Allen, Ambikairajah, & Epps, 2006), are utilized individually, reasonable performance has been achieved in LID. As new cues which contribute to human audible intelligence are discovered continuously (Greenberg & Arai, 2004), more and more modern LID systems (Rong, Bin, Donglai, Haizhou, & Eng Siong, 2006; Singer, Torres-Carrasquillo, Gleason, Campbell, & Reynolds, 2003) utilize a 'hybrid' approach. In a typical hybrid system, the results from individual sub-systems which utilize different models or features (referred to as sub-system in following sections) are fused together to produce the final evaluation result. Using this method, the hybrid system could benefit from all of available cues, and may produce an improved performance. The key issue of the hybrid approach is to discover an appropriate and effective fusion scheme.

In recent research, several different fusing approaches have produced remarkable improvements when compared to a single system, including Linear Score Weighting (LSW) (Wang, 2004; Wong & Sridharan, 2001), Gaussian Mixture Model (GMM) (Torres-Carrasquillo et al., 2002) and Discriminant Factor Analysis (DFA) (Gutierrez, Rouas, & Andre-Obrecht, 2004). The basic idea behind these approaches involves applying weighting coefficients to the likelihood scores produced by individual sub-systems. However, most existing research was focused on language-independent fusion, not language-dependent fusion. In the former case, the weighting coefficients are only different for varied sub-systems, but always the same for different languages in a particular sub-system. In other words, particular language-related or language-dependent discriminating knowledge is not used at all in language-independent fusion. There is a similar

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

situation in the more recent Feature Combination approach (Yin et al., 2006) as well.

In this paper, a novel Language-dependent Fusion (LDF) technique is presented. In this approach, the weighting coefficients are varied for different languages. The coefficients are directly calculated from the sub-system performances, which reflect how one feature or model (represented by a sub-system) contributes to each particular language.

This technique is particularly useful and practical for LID, because all potential languages faced by an LID system are always known in advance, and not changed during the whole application period. The situation is different in other speech recognition tasks, e.g. speaker recognition.

For this paper, experiments have been conducted on the OGI-92 multi-language telephony speech database. The proposed Language-dependent Fusion technique introduces a 33.33% error-rate reduction in both 10-language and 4-language settings when compared to LSW. It also introduces a 45.46% and 39.96% error-rate reduction in 10-language and 4-language settings, when compared to the best performance of the individual sub-system. Further more, the LDF technique achieves a comparable or even better performance when compared to a Feature Combination system as well.

2. Language-dependent fusion

A language identification system typically consists of two major parts: the front-end, which extracts features from raw speech data, and the back-end, which models and classifies the distribution of features for each language. As more and more features and models have proven to be effective for LID, a hybrid system is required for combining different features and/or models together, thereby producing a higher overall performance.

There are two major methods for combining these different features and/or models in a hybrid system. One is feature-combination, another is fusion. In the former case, different features are simply concatenated before the model is trained. Apparently, only one model is trained based on all feature data in this case, which is an advantage because the joint feature distribution contains more useful information about language, but it is also a limitation because the combined features have to be similar and relative in feature space. In the latter case, the output scores produced by individual systems utilizing a single feature/model are fused together to produce a new set of scores for final evaluation. In this case, any type of system can be fused together. The sensitivity to the change of accent is similar between fusion techniques but may be different between fusion and feature combination, depends on the sensitivity of particular features to the change of accent. In this paper, the fusion based hybrid system is mainly focused, although a feature combination based system is deployed as a reference system.

In an LID system, a model for each language is trained from feature data in the training stage. In the evaluation stage, the same feature is extracted from unknown utterance, and compared to each existing language model. Normally, the distance between the testing utterance and language model is measured as a likelihood score. The language leading to maximum likelihood score is decided as the identification output. Therefore, the goal of fusion technique is basically to produce a reasonable likelihood score from several different likelihood scores produced by individual sub-systems.

Linear Score Weighting (LSW) is one of the most widely used fusion techniques in LID (Wong & Sridharan, 2001). In this approach, the final output score is calculated as follows:

$$Z_{LSW} = \sum_{i=1}^{N} w_i \cdot Y_i \qquad (1)$$

where Z_{LSW} is the final output score, Y_i is the output score generated by sub-system i, N is the number of sub-systems, w_i is the weighting coefficients for output score from sub-system i, and

$$\sum_{i=1}^{N} w_i = 1 \qquad (2)$$

Here, *a posteriori* probabilities are calculated as Y_i instead of likelihood score, for normalizing the scores among different sub-systems.

To find out the best w_i, different values are evaluated on the development dataset. The value with the highest performance is selected.

Apparently, in this case, w_i is only related to a sub-system. For a particular sub-system, the weighting coefficient is identical for all target languages. The assumption behind this is that there is no difference on the performance among varied languages in a particular sub-system.

However, this is not true in most cases. One particular sub-system (particular feature or model) usually demonstrates higher performance on some languages than others. To correctly decide the contribution ratio of each sub-system to particular language, a fusion scheme that considered this difference among languages is required. In this paper, a novel Language-dependent Fusion (LDF) approach is proposed and evaluated on a GMM based LID system.

For a typical GMM based LID system, in the

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

training stage, feature data (e.g. cepstral coefficients) is extracted from the raw speech data, and a separate GMM is trained for each language after that. In the testing stage, the likelihood scores between the target utterance features and language GMMs are calculated, the highest likelihood score indicates the identification result.

In an LDF system, the fusion process is applied at testing stage, more specifically, to likelihood score. The final likelihood score for each language is calculated as following:

$$LL_i = \sum_{f=1}^{M} W_{f,i} \cdot LL_{f,i} \quad (3)$$

where LL_i is the final weighted likelihood score of language i, f is the index of sub-system, $W_{f,i}$ is the weighting coefficient applied to sub-system f and language i, $LL_{f,i}$ is the likelihood score produced by sub-system f for language i, M is the total number of sub-systems. The final LID decision is made in usual way based on the final weighted likelihood scores. The LDF process described above is illustrated in Figure 1.

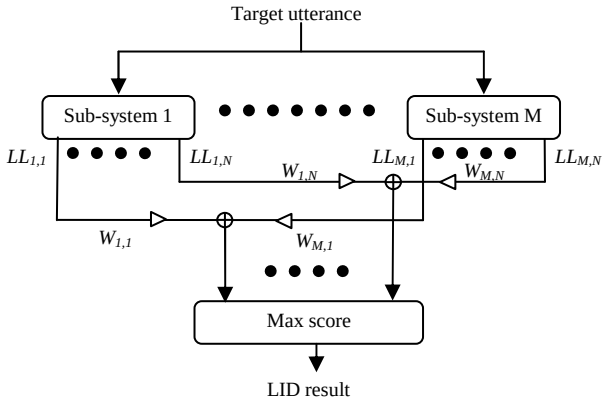

Figure 1: The process of Language-dependent Fusion (LDF)

When compared to a conventional LSW technique, there are two major differences in LDF: language-dependent weighting and non-normalization on likelihood score. As shown in Equation 1, in LDF, varied weighting coefficients are applied in case of different language instead of fixed value. The likelihood scores produced by sub-systems are directly used as input score without normalization instead of *a posteriori* probabilities, because the difference among those likelihood scores reflects the difference of contribution from the features/models behind sub-systems, which is needed by LDF.

Besides the commonly used LSW technique, another feature-combination-based system was also implemented and evaluated for comparison. In this feature combination approach (Yin et al., 2006), The input vector is concatenated from two individual features, and used for training GMM as a single feature vector. The remaining part of the system is exactly the same with individual sub-systems.

3. Weighting coefficients

An obvious question to LDF is how to calculate the weighting coefficients $W_{f,i}$ efficiently, to properly reflect the difference among languages. The possible solution may be to search global optimization values or statistical approaches like GMM fusion. However, these methods are either time consuming or not accurate enough with limited amount of data. In this paper, an intuitive way of calculating these language-dependent weighting coefficients is proposed.

The key idea behind the calculation of weighting coefficients is that the LID performance achieved by a particular sub-system on a particular language reflects the contribution of this sub-system (the feature/model behind this sub-system) to that particular language. Therefore, if the LID performance is evaluated on all possible language pairs, the average performance of all pairs related to one particular language could be a reasonable indicator for the effectiveness or contribution of the current sub-system to this particular language.

Based on the above idea, the weighting coefficient $W_{f,i}$ is calculated as following:

$$W_{f,i} = \frac{1}{N} \sum_{j=1}^{N} \log(P_{f,ij}) \quad (4)$$

where N is the total number of languages, $P_{f,ij}$ is the performance of sub-system f for language pair i and j. When i equals to j, $P_{f,ij}$ is set to 1.

In the case of this paper, the accuracy (correction-rate) of LID on language pair i and j produced by sub-system f is used as $P_{f,ij}$. Each combination of sub-system f and language pair i and j is evaluated on development dataset to produce $P_{f,ij}$. More specifically, two GMM based sub-systems are deployed in this paper. The only difference between these two sub-systems is the feature used.

When compared to other weighting coefficients calculation methods such as LSW, $W_{f,i}$ calculated here is not normalized among languages or sub-systems. The reason is similar to the non-normalized likelihood

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

explained in section 2. Additionally, the reason is also that the rejection of unknown language is not considered in the case of this paper. A simple normalization could be deployed if the rejection threshold is set. When compared to statistical approach, the calculation of weighting coefficients in LDF is very efficient because the training process in the statistical approach will take a long time when the amount of data is huge. Therefore, the calculation process of weighting coefficients in LDF is easy and intuitive.

4. Experiments and results

All experiments conducted in this paper are based on a GMM-UBM LID system (Allen, Ambikairajah, & Epps, 2005; Yin et al., 2006). The diagram of this system is illustrated in Figure 2. Although this original design was used directly as an individual sub-system, the testing stage is alternated for fusing likelihood scores from sub-systems together in LDF configuration (see Figure 1).

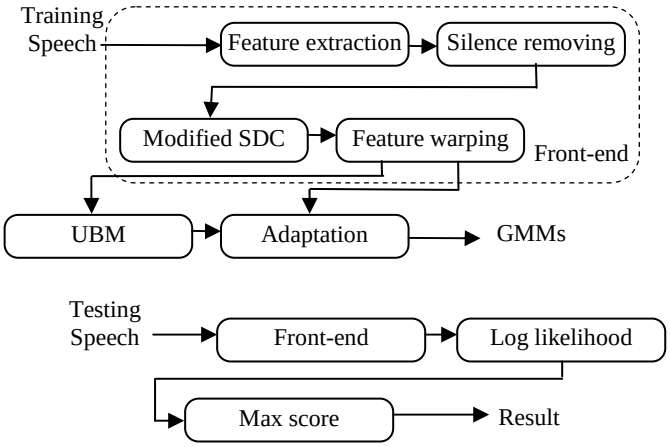

Figure 2: Diagram of UBM-GMM system for training (above) and testing (bottom)

Two different feature sets, prosodic feature and Mel-Frequency Cepstral Coefficients (MFCC), are utilized to construct two sub-systems (the formal sub-system will be referred as sub-system MFCC, and the latter sub-system will be referred as sub-system PRO in the following sections). More detailed, the prosodic feature consists of pitch and log-energy values calculated on a frame basis, and the number of MFCC coefficients is seven for optimal results (Yin et al., 2006).

For comparison purpose, besides LDF, LSW and feature combination systems were also implemented in same system structure.

The OGI-92 telephony speech database is used in all experiments. This database is a multi-language, multi-speaker database, composed of an average 122 calls (approx. 2 minutes each, different speakers for different calls) in each of 11 languages.

The available speech data is allocated as follows: for each language, only 90 calls are used in total to ensure the same amount of data is available for each language. All this data is partitioned into four datasets: dataset A (30 calls), dataset B (20 calls), dataset C (20 calls), and dataset D (20 calls). When calculating the weighting coefficients, dataset A was used for UBM training, datasets A+B (training data) were used for GMM adaptation, dataset C (development data) was used for evaluation. When preparing the sub-system results for fusion, datasets A+B were used for UBM training, datasets A+B+C were used for GMM adaptation, dataset D (evaluation data) was used for evaluation. Separate experiments were conducted on sub-system MFCC and sub-system PRO with the same data.

Since the purpose of these experiments is to compare the performances of different sub-systems achieved on different language pairs, and to compare the performances between LDF and other fusion techniques, the system and database are acceptable.

4.1. Experiments on ten languages

To research the LDF performance on most languages faced by LID, experiments on ten languages (English, Farsi, French, German, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese) were conducted.

The first task is to acquire the weighting coefficients. Two sub-systems were tested on development data at this stage. Each language pair was evaluated and the correction rate of each evaluation was used as performance value *P* (see section 3).

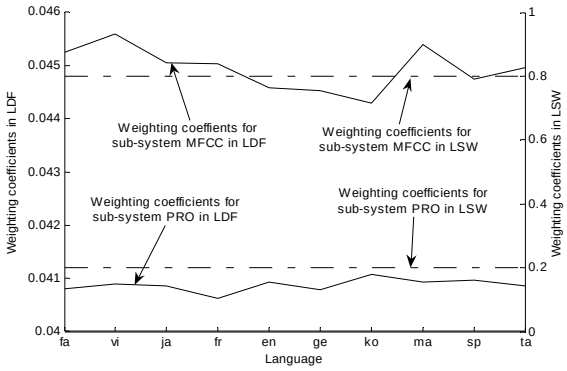

Figure 3: Weighting coefficients produced by LDF and LSW

In the second task, all weighting coefficients were calculated according to Equation 4. The weighting coefficients produced by LDF and LSW are demonstrated in Figure 3. Apparently, LDF presented more language related information than LSW.

Since the weighting coefficients had been acquired, the weighting process could be applied. At this stage, all individual sub-systems were trained and evaluated on

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

the whole data set. Fused likelihood scores are produced by calculating Equation 3. The final result is decided by picking the maximum likelihood score. An example of the likelihood score produced by individual sub-systems and LDF system is shown in Figure 4.

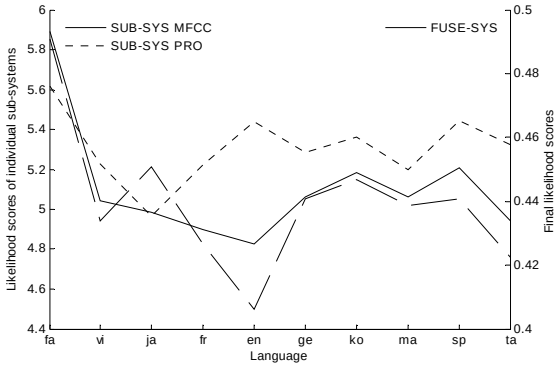

Figure 4: An example of likelihood scores produced by individual sub-systems and LDF system (FUSE-SYS)

It's clear from the example in Figure 4 that different sub-systems contributed differently for different languages. The final decision is not always decided solely by one particular sub-system.

For comparison purpose, different fusion techniques were tested on the same data. In LSW fusion, an optimal weighting coefficient was found to be 0.8 for sub-system MFCC and 0.2 for sub-system PRO.

The results of individual sub-systems, LDF system, LSW system and Feature-Combination system (FC-SYS) are shown in Table 1. Two sets of target utterance, 20 seconds and 10 seconds, were tested to demonstrate performance difference between utterances in varied length.

Table 1: Error rate of individual sub-systems and different fusion systems in 20s and 10s utterance (10 languages)

SYSTEM TYPE		ERR%-20s	ERR%-10s
SUB-SYS	MFCC	17.53%	26.85%
	PRO	33.86%	44.65%
FUSE-SYS	LDF	9.56%	19.11%
	LSW	14.34%	24.22%
FC-SYS		8.76%	19.28%

It is clear that in the 20s case LDF achieved a 45.46% error-rate reduction when compared to the individual sub-system with highest performance (SUB-SYS MFCC in this case), and outperformed LSW by a 33.33% error-rate reduction. Another interesting observation is that the performance of LDF is slightly lower than the feature combination approach(Yin et al., 2006). This may be explained by the fact that feature combination already produced a good joint distribution on the discrimination information presented by two individual features. However, the feature combination could only be used in the same frame-rate based features, not two sub-systems utilizing totally different models.

In the 10s case, although the performances dropped, the LDF system still outperformed the LSW system and was comparable with the feature combination system.

The computational costs for different systems are varied, but not much different. In 20s case, the ratio of the training time of SUB-SYS PRO, SUB-SYS MFCC, and FC-SYS is 0.3:1.0:1.6. For LSW system, finding the optimal weighting coefficients may take some evaluation time. For LDF system, the calculation of weighting coefficients is very efficient. On the other hand, the evaluation time of each method is quite similar. Therefore, the LDF system is overall most efficient among all these approaches.

4.2. Experiments on four languages

As revealed in previous experiments, the same feature may contribute differently to different languages. To research the LDF performance among different language groups, another set of experiments was conducted. In these experiments, two groups of four languages were chosen. The first group consisted of the languages with the four highest weighting coefficients (i.e. best average performance against all other languages). Oppositely, the second group consisted of languages with the four lowest weighting coefficients. Obviously, a reference sub-system had to be selected as the source of weighting coefficients. Because the performance of sub-system PRO shows more distinct variation or discrimination among languages, the sub-system PRO is selected as the reference in this case. According to this criterion, the first group consisted of Japanese, Mandarin, Vietnamese, and Farsi, and the second group consisted of Tamil, German, Spanish, and French. The results and comparison to other fusion techniques are shown in Table 2 and 3. Similar to the 10 languages experiment, two sets of target utterances, 20 seconds and 10 seconds, were tested.

Table 2: Error rate of individual sub-systems and different fusion systems in 20s and 10s utterance (Japanese, Mandarin, Vietnam, Farsi)

SYSTEM TYPE		ERR%-20s	ERR%-10s
SUB-SYS	MFCC	5.68%	14.48%
	PRO	6.82%	16.29%
FUSE-SYS	LDF	3.41%	7.24%
	LSW	4.55%	11.31%
FC-SYS		4.55%	9.50%

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Table 3: Error rate of individual sub-systems and different fusion systems in 20s and 10s utterance (Tamil, German, Spain, French)

SYSTEM TYPE		ERR%-20s	ERR%-10s
SUB-SYS	MFCC	11.50%	17.98%
	PRO	32.74%	38.95%
FUSE-SYS	LDF	7.08%	16.85%
	LSW	10.62%	17.60%
FC-SYS		7.08%	14.61%

The results show the LDF approach is still effective on a smaller language group either for high discrimination groups or low discrimination groups. In particular, LDF achieves an equal or even better performance than feature combination approach, without the requirement of re-training system.

In the 20s case, for the high discrimination group, LDF introduced a 39.96% error rate reduction to the best sub-system (sub-system MFCC), and outperformed LSW and feature combination by 25.05% in error rate. For the low discrimination group, LDF achieved a similar performance of feature combination, introduced a 38.43% error rate reduction to the best sub-system (sub-system MFCC), and outperformed LSW by 33.33%.

Similarly, in 10s case, although all performances dropped the LDF still outperformed individual sub-systems and LSW system, and achieved comparable performance when compared to the feature combination system.

5. Conclusions

From the analysis and experiments conducted in this paper, the following conclusion can be drawn.

In a hybrid LID system, not only do different sub-systems contribute differently to all languages, but the same sub-system (utilizing a particular feature and/or model) may also contribute differently to different languages.

The Language-dependent Fusion (LDF) technique is an effective method for combining different features or fusing different systems. Using LDF introduces a remarkable improvement when compared to individual sub-systems and other commonly-used fusion techniques, especially in a limited situation. Additionally, the weighting coefficients in LDF are easily and efficiently calculated. There is no extra manual labeling or time-consuming training required.

The LDF is especially useful and suitable for LID because the potential languages faced by LID system are always known in advance, and not changed during the whole application period. Therefore, it is possible and practical to pre-acquire all weighting coefficients.

Further research on the relation between particular feature and language discrimination is scheduled. The additional experiments will be conducted and presented according to NIST LRE guidelines for more precise comparison.

6. References

Allen, F., Ambikairajah, E., & Epps, J. (2005). *Language identification using warping and the shifted delta cepstrum.* Paper presented at the IEEE International Workshop on Multimedia Signal Processing, Shanghai, China.

Allen, F., Ambikairajah, E., & Epps, J. (2006). *Warped magnitude and phase-based features for language identification.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France.

Campbell, W., Singer, E., & Torres-Carrasquillo, P. (2004). *Language recognition with support vector machines.* Paper presented at the ODYSSEY - The Speaker and Language Recognition Workshop, Toledo, Spain.

Greenberg, S., & Arai, T. (2004). What are the essential cues for understanding spoken languages? *IEICE Transaction on Information & System, E87-D*(5), 1059.

Gutierrez, J., Rouas, J. L., & Andre-Obrecht, R. (2004). *Fusing language identification systems using performance confidence indexes.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Montreal Canada.

Rong, T., Bin, M., Donglai, Z., Haizhou, L., & Eng Siong, C. (2006). *Integrating acoustic, prosodic and phonotactic features for spoken language identification.* Paper presented at the IEEE International Conference on Acoustics, Speech, and Signal Processing, Toulouse, France.

Singer, E., Torres-Carrasquillo, P. A., Gleason, T. P., Campbell, W. M., & Reynolds, D. A. (2003). *Acoustic, phonetic, and discriminative approaches to automatic language identification.* Paper presented at the EuroSpeech, Geneva, Switzerland.

Torres-Carrasquillo, P. A., Singer, E., Kohler, M. A., Greene, R. J., Reynolds, D. A., & J.R. Deller, J. (2002). *Approaches to language identification using gaussian mixture models and shifted delta cepstral features.* Paper presented at the International Conference on Spoken Language Processing, Denver, USA.

Wang, E. (2004). *Automatic spoken language identification utilizing acoustic and phonetic speech information.* Unpublished Ph.D. Thesis, Queensland University of Technology, Australia.

Wong, E., & Sridharan, S. (2001). *Fusion of output scores on language identification system.* Paper presented at the Workshop on Multilingual Speech and Language Processing, Aalborg Denmark.

Yin, B., Ambikairajah, E., & Chen, F. (2006). *Combining prosodic and cepstral features in language identification.* Paper presented at the IEEE International Conference on Pattern Recognition, Hong Kong, China.

Zissman, M. A. (1996). Comparison of four approaches to automatic language identification of telephone speech. *IEEE Transactions on Speech and Audio Processing, 4*(1), 31-44.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review