LANGUAGE IDENTIFICATION USING PARALLEL SUB-WORD RECOGNITION

A. K. V. Sai Jayram V. Ramasubramanian T. V. Sreenivas

Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore 560 012, India
email: tvsree@ece.iisc.ernet.in

ABSTRACT

Parallel sub-word recognition (PSWR) is a new model that has been proposed for language identification (LID) which does not need elaborate phonetic labeling of the speech data in a foreign language. The new approach performs a front-end tokenization in terms of sub-word units which are designed by automatic segmentation, segment clustering and segment HMM modeling. In this paper, we develop PSWR based LID in a framework similar to the parallel phone recognition (PPR) approach in the literature. This includes a front-end tokenizer and a back-end language model, for each of the language to be identified. Considering various combinations of the statistical evaluation scores, it is found that PSWR can perform as well as PPR, even with broad acoustic sub-word tokenization, thus making it an efficient alternative to the PPR system.

1. INTRODUCTION

Automatic language identification (LID) has become an important research problem over the last decade with several promising solutions [1], [2]. Among the various approaches to LID [2], the phone-recognition approach offers considerable promise, as it incorporates sufficient knowledge of the phonology of the languages to be identified. One of the main frameworks in the phone-recognition approaches is Parallel Phone Recognition (PPR) [3].

An N - language LID task is to classify an input speech utterance as belonging to one of N languages $\mathcal{L}_1, \ldots, \mathcal{L}_N$. The PPR system for this task has N paths, each with a front-end phone-recognizer (PR) followed by a back-end language model (LM) in that language. The front-end PR tokenizes the input utterance into a sequence of phone symbols. The back-end LM performs phonotactic analysis on the resulting phone sequence. Phonotactics refers to the language-dependent constraints on the sequences of phones and is modeled by an n-gram analysis, with typical systems using a bigram ($n = 2$) statistics. An input utterance is classified by a maximum likelihood decision on the N scores obtained by the front-end PR [4] or by the back-end LM or jointly by PR and LM of each language [3], [5].

In the PPR system, the front-end PR has to be trained on phonetically labeled data, usually obtained by manual labeling, to generate a 'phone HMM inventory' of that language. Thus, a PPR requires labeled training data for all the N languages in the task to train each of its N front-end PRs.

Recently, we have proposed a parallel sub-word recognition (PSWR) system for LID [4], which operates in a PPR framework but without requiring manually labeled phonetic data in any of the languages in the task. In this work, we studied the performance of PSWR on 6 languages in the OGI-TS database [6] and obtained an LID accuracy of 91.6% on training data and 67.5% on test data for 45-sec utterances using only the acoustic score from the front-end sub-word recognizers. These were comparable to the PPR system (87.9% on training data and 70% on test data), thus making the PSWR system an efficient alternative to the PPR system.

In this paper, we study the complete PSWR system with both the front-end sub-word recognizer (SWR) and the back-end language model (LM) of sub-word unit sequences tokenized by the front-end SWR. The resulting PSWR system can yield three types of scores, namely, the acoustic score, joint acoustic-language score and the language model score. We examine the effectiveness of these scores for a LID task of 6 languages in the OGI-TS database.

2. PARALLEL SUB-WORD RECOGNITION (PSWR)

The PSWR system uses a front-end sub-word recognizer (SWR) for each language in the task. For an N-language task, the PSWR system has N front-end SWRs. Each SWR has a language dependent sub-word unit (SWU) inventory. The important point to note is that the SWU inventory is obtained without the need for manually labeled training data. Each front-end SWR is followed by a back-end language model (LM) which performs phonotactic analysis. The LM is typically an n-gram analyzer of the SWU label sequence output by the SWR front-end. For a given input utterance, the PSWR system yields N scores which can be of the following types: i) Acoustic score, obtained by the front-end SWR, ii) Joint acoustic-language score, obtained by a joint decoding using both the front-end SWR and back-end LM, and iii) Language-model score, obtained from only the back-end LM. The input utterance is classified into one of N languages based on a maximum likelihood decision on the N scores of any of these three types.

The PSWR system thus consists of three components: i) Sub-word recognizer (SWR), ii) Language model (LM), and iii) Maximum - likelihood (ML) classifier. These are described in detail in the following sections with emphasis on the three types of scores stated above.

2.1. Sub-word recognizer (SWR)

2.1.1. Sub-word unit (SWU) inventory

Each of the N SWRs in a PSWR system has an inventory of sub-word units (SWUs) $\mathcal{P}_l = \{P_1, P_2, \ldots, P_L\}$ and corresponding sub-word HMM models [7] $\mathcal{H}_l = \{\lambda_1, \lambda_2, \ldots, \lambda_L\}$ for each language $\mathcal{L}_l$, $l = 1, \ldots, N$. In the PSWR system, the training phase involves the design of the SWU inventory $\mathcal{H}_l$ for a set of SWUs $\mathcal{P}_l$ of each language $\mathcal{L}_l$ in the LID task. The procedure for generating

the SWU inventory is essentially that used for acoustic SWU based speech recognition [8] and constitutes the training of the SWR in the PSWR system; this is discussed in detail in [4]. Briefly, it consists of the following steps for each language:

i) **Automatic segmentation:** The training utterances (in the form of MFCC vector sequence) is segmented into acoustic segments using the maximum-likelihood (ML) segmentation technique [9] for a required number of segments/second. This generates a large corpus of segments $\mathcal{S} = \{S_1, S_2, \ldots, S_M\}$.

ii) **Segment clustering:** The acoustic segments in the segment corpus $\mathcal{S}$ are represented by their respective centroids; these centroids are clustered into L clusters $\mathcal{C} = \{C_1, C_2, \ldots, C_L\}$ using the K-means algorithm. Each acoustic segment $S_m \in \mathcal{S}$ then belongs to one of L sub-word clusters.

iii) **Segment modeling:** Each cluster $C_i \in \mathcal{C}$ defines a class of acoustically similar segments and is treated as representing a notional sub-word unit P_i. The segments belonging to each sub-word class $P_i, i = 1, \ldots, L$ are modeled by an HMM [7]. This results in an inventory of L sub-word HMMs, $\mathcal{H}_l = \{\lambda_1, \lambda_2, \ldots, \lambda_L\}$ with the corresponding inventory of SWUs $\mathcal{P}_l = \{P_1, P_2, \ldots, P_L\}$ for the language $\mathcal{L}_l$. Each language has a different $\mathcal{P}_l$.

2.1.2. Sub-word tokenization

The front-end SWR in path l uses the sub-word inventory $\mathcal{H}_l$ to tokenize an input utterance into a sequence of sub-word unit labels by optimal decoding as in connected word recognition [7]. This decoding is an optimum 'connected sub-word recognition' problem, which generates a decoded sub-word sequence and the associated acoustic likelihood score corresponding to the decoded best path by the Viterbi search. This SWR decoding is performed for each of the N languages in PSWR.

Let the input utterance be a sequence of feature vectors $\mathbf{O} = (O_1, O_2, \ldots, O_T)$. $\mathbf{O}$ is tokenized by the SWR into an optimal sequence of K SWUs $(P_{\bar{1}}, \ldots, P_{\bar{k}}, \ldots, P_{\bar{K}})$, where $P_{\bar{k}} \in \mathcal{P}_l$. The likelihood associated with this optimal string of sub-word units is calculated as $\mathbf{P}_A(l) = P(\mathbf{O}|\mathcal{H}_l)$ (acoustic score) given by,

$$\mathbf{P}_A(l) = \max_{B,K} \sum_{k=1}^{K} \log(p(s_k|\lambda_{\bar{k}})) \tag{1}$$

$B = (b_0, b_1, \ldots, b_K)$, with $b_0 = 0$ and $b_K = T$, are the segment boundaries for any segmentation of the T frames of $\mathbf{O}$. The k^{th} segment $s_k = (O_{b_{k-1}+1}, \ldots, O_{b_k})$ is associated with the SWU model $\lambda_{\bar{k}}$, where, $\lambda_{\bar{k}}$ is the HMM model which has the maximum likelihood of generating segment s_k, from among $\mathcal{H}_l = \{\lambda_1, \lambda_2, \ldots, \lambda_L\}$; $p(s_k|\lambda_{\bar{k}})$ is the corresponding HMM likelihood and $P_{\bar{k}}$ is the corresponding SWU.

2.2. Language-model (LM)

The back-end LM in each of the PSWR system paths performs phonotactic analysis on the SWU label sequence generated by the front-end SWR; typically, it evaluates the likelihood of the SWU sequence using a bigram distribution.

Let the front-end SWR of language $\mathcal{L}_l$ tokenize the input utterance into a sequence of K SWU labels $(P_{\bar{1}}, \ldots, P_{\bar{k}}, \ldots, P_{\bar{K}})$, as given by (1). The LM likelihood $\mathbf{P}_L(l) = P(\mathbf{O}|\mathcal{B}_l)$, using a bigram model $\mathcal{B}_l$ of language $\mathcal{L}_l$, is given by,

$$\mathbf{P}_L(l) = \sum_{k=2}^{K} \log p(P_{\bar{k}}|P_{\overline{k-1}}, \mathcal{L}_l) \tag{2}$$

where $P_{\bar{k}}$ and $P_{\overline{k-1}}$ are consecutive symbols observed in the tokenized SWU stream. The bigram model for language $\mathcal{L}_l$ is given by $\mathcal{B}_l = \{p_l(i,j)\} = \{p(P_j|P_i)\}, i, j = 1, \ldots, L$, where $p_l(i,j)$ is used to evaluate $p(P_{\bar{k}}|P_{\overline{k-1}}, \mathcal{L}_l)$ in (2) when $P_{\bar{k}} = P_j$ and $P_{\overline{k-1}} = P_i$. $\mathcal{B}_l$ is learnt from the SWU labels obtained from tokenization of training utterances of language $\mathcal{L}_l$ using the SWR front-end of language $\mathcal{L}_l$.

2.3. Joint acoustic-phonotactic decoding

While sub-word recognition (SWR) and language modeling (LM) are described above as done independently, they can be combined into one step in the PSWR configuration; i.e., it is possible to integrate the acoustic/phonotactic models so that language-specific phonotactic constraints can be used during the Viterbi decoding process of SWR rather than applying the LM (phonotactic) constraints after the sub-word recognition is complete. This is referred to as *joint acoustic-phonotactic decoding* or simply as *joint decoding.* This was first used in [3] and is being extended here to joint decoding with sub-word units. The most likely sub-word sequence obtained by joint decoding and the corresponding acoustic - phonotactic likelihood measure optimally combines both the acoustic likelihood and the language model likelihood (obtained by bigram models).

In the case of joint acoustic-phonotactic decoding, the input utterance $\mathbf{O} = (O_1, O_2, \ldots, O_T)$ is tokenized by the SWR and LM jointly for each language $\mathcal{L}_l, l = 1, \ldots, N$. The optimal sequence of K SWUs $(P_{\hat{1}}, \ldots, P_{\hat{k}}, \ldots, P_{\hat{K}})$, obtained by joint decoding by the SWR and LM of language $\mathcal{L}_l$, maximizes the joint acoustic-phonotactic likelihood (or acoustic-language score) $\mathbf{P}_{AL}(l) = P(\mathbf{O}|\mathcal{H}_l, \mathcal{B}_l)$, as given by,

$$\mathbf{P}_{AL}(l) = \max_{B,\hat{\lambda},K} \left\{ \mathbf{P}_1 + \sum_{k=2}^{K} \log[p(s_k|\lambda_{\hat{k}}) \cdot p(P_{\hat{k}}|\widehat{P_{k-1}})] \right\} \tag{3}$$

where B and s_k are as given in (1); $\mathbf{P}_1 = \log p(s_1|\lambda_{\hat{1}})$, $\hat{\lambda} = (\lambda_{\hat{1}}, \ldots, \lambda_{\hat{k}}, \ldots, \lambda_{\hat{K}})$ is any sequence of $\lambda_{\hat{k}} \in \mathcal{H}_l, k = 1, \ldots, K$ and $(P_{\hat{1}}, \ldots, P_{\hat{k}}, \ldots, P_{\hat{K}})$ is the corresponding sub - word sequence. $(P_{\bar{1}}, \ldots, P_{\bar{k}}, \ldots, P_{\bar{K}})$ is the SWU sequence corresponding to the optimal $\bar{\lambda} = \{\lambda_{\bar{k}}\}_{k=1}^{K}$ which maximizes $\mathbf{P}_{AL}(l)$ in (3).

2.4. Language-model (LM) scores

The bigram model $\mathcal{B}_l$ used in (2) is estimated from SWU labels obtained from decoding of training utterances of language $\mathcal{L}_l$ by the front-end SWR of language $\mathcal{L}_l$, as given by (1); this does not use any LM (bigram) constraint in the SWU decoding as done in (3). The LM score in (2) using this $\mathcal{B}_l$ on the SWU sequence obtained by (1), is referred here as the "Language-Model score – Decoupled" and is denoted by $\mathbf{P}_{LD}(l)$.

For joint decoding (3), the bigram model $\mathcal{B}_l$ (which provides $p(P_{\hat{k}}|\widehat{P_{k-1}})$ in (3)) is estimated from the 'reference SWU labels' of training utterances, obtained as follows during the SWR training for language $\mathcal{L}_l$ (Sec. 2.1.1): The training utterances are segmented using ML-segmentation into a sequence of segments $(S_1, S_2, \ldots, S_M)$; each of these segments belongs to some cluster from $\mathcal{C} = \{C_1, C_2, \ldots, C_L\}$ and is labeled by the corresponding SWU from $\mathcal{P}_l = \{P_1, P_2, \ldots, P_L\}$. The resulting SWU label sequence gives the 'reference SWU labels' of the training utterances. Use of the $\mathcal{B}_l$ (estimated from these reference SWU labels) in (3) constrains the joint decoding to decode an utterance into a SWU se-

quence which better matches the reference SWU label sequence of the utterance. The LM score computed by (2) when applied on the SWU sequence obtained by (3) is referred here as "Language-Model score – Joint" and is denoted by $\mathbf{P}_{LJ}(l)$.

In a PPR system, the bigram model $\mathcal{B}_l$ used for joint decoding (3) is estimated from the manually labeled phone sequence of the training data. In the PSWR system, the reference SWU labels, derived by automatic segmentation and labeling of training data, serves the role of the manual phone labels in the PPR system.

2.5. Maximum-likelihood classifier (MLC)

We have described four types of scores in PSWR for an input utterance: i) Acoustic score $\mathbf{P}_A(l)$ (Eqn. (1)), ii) Joint acoustic-phonotactic (or acoustic-language) score $\mathbf{P}_{AL}(l)$ (Eqn. (3)), iii) Language-model score – Decoupled $\mathbf{P}_{LD}(l)$ (Sec. 2.4), and iv) Language-model score – Joint $\mathbf{P}_{LJ}(l)$ (Sec. 2.4). Denoting any of these four scores by $\mathbf{P}(l)$, the maximum - likelihood (ML) classifier identifies the language of the input utterance as $\mathcal{L}_{l^*}$ which has the highest likelihood (score) $\mathbf{P}(l)$, i.e.,

$$l^* = \arg \max_{l=1,\ldots,N} \mathbf{P}(l) \tag{4}$$

It has been observed [3] that the log-likelihood scores $\mathbf{P}_{AL}(l)$ were biased in favor of one language over another or even all languages. This is corrected using a bias-removal method, which subtracts a bias $b(l)$ from $\mathbf{P}_{AL}(l)$ of the language $\mathcal{L}_l$ path, before performing ML classification by (4). The bias $b(l)$ is determined as the average of $\mathbf{P}_{AL}(l)$ obtained from utterances of all languages input to language $\mathcal{L}_l$ path. We have observed this bias problem in all four scores $\mathbf{P}_A(l)$, $\mathbf{P}_{AL}(l)$, $\mathbf{P}_{LD}(l)$ and $\mathbf{P}_{LJ}(l)$ to varying extents and have used the bias-removal method of [3] for all the four scores in both PPR and PSWR systems. Note that the PPR system also has the same four types of scores as discussed in this paper for PSWR and in [5], though not dealt with in [3] (which is also the only other earlier work on PPR reported so far). We have interpreted this bias-problem in [5] and also proposed an alternate method for bias-removal.

3. EXPERIMENTS AND RESULTS

We treat the PPR system [3] as a baseline system with which to compare the performance of the PSWR system. PPR is described briefly in Sec. 1. We present here results comparing the performance of the PPR and PSWR systems for the four different types of scores: Acoustic score ($\mathbf{P}_A$), Joint acoustic-language score ($\mathbf{P}_{AL}$), Language model score – Decoupled ($\mathbf{P}_{LD}$), and Language model score – Joint ($\mathbf{P}_{LJ}$).

3.1. Database

We use the Oregon Graduate Institute Multi-language Telephone Speech (OGI-TS) corpus [6] for evaluation. The OGI-TS has a total of 11 languages, out of which 6 languages – English (EN), German (GE), Hindi (HI), Japanese (JA), Mandarin (MA) and Spanish (SP) – have phonetic labels. We evaluate both the PPR and PSWR systems using these 6 languages of the OGI-TS database.

Both the PPR and PSWR systems are trained on 50 'story-bt' (story-before-the-tone) utterances per language spoken by 50 different speakers. Both the systems are tested using 20 'story-bt' utterances per language outside the training data; the training and test utterances are each 45 seconds long.

3.2. Parameters of PSWR system

The ML segmentation technique (Sec. 2.1.1), can segment an input utterance into a pre-specified number of segments [9],[4]. We specify this as $m = Rt$, where R is the number of segments per second and t is the duration of the input utterance in seconds. R is used as a parameter to control the segmentation rate of the ML segmentation. R takes values as 2, 5, 10 and 20 seg/sec. $R = 2$ and 5 correspond to a coarse segmentation and can generate broad-phonetic segments and phone strings. $R = 10$ gives phone-like segmentation as the phone rate in normal speech is about 10 phones/sec; $R = 20$ results in a fine segmentation and produces sub-phonemic segments.

The sub-word inventory size L (Sec. 2.1.1), controls the resolution of the acoustic space; L is varied as 10, 30, 50 and 100. Small L ($L = 10$) corresponds to a coarse clustering and generates HMM models of broad-phonetic categories. $L = 30$ and 50 generate phone-like units in the inventory as languages typically have phone set sizes in this range. $L = 100$ yields a finer clustering of the acoustic space.

3.3. Model building

Speech data is parameterized every 20ms with a frame shift of 10ms. Each frame of speech is first pre - emphasized by $(1 - 0.95z^{-1})$ and then windowed by a Hamming window. The pre-emphasized and windowed frame is then used for MFCC parameter estimation.

Training of a PPR system consists of generating an inventory of phone HMM models of size determined by the number of different phonetic units in the corpus labeling scheme for each of the 6 languages. Whereas in a PSWR system, the training consists of generating an inventory of sub-word HMM models of any desired size L (Sec. 2.1.1). Both these training procedures are implemented using HTK [10].

The HMMs for both PPR and PSWR use a 26 - dimensional parameter vector consisting of 12 MFCC, 12 delta-MFCC, energy and delta energy. The HMMs are 3 state, left to right models. For PPR system the best performance was obtained with 6 Gaussian mixtures per state. For PSWR system we use 9 Gaussian mixtures per state, for the various (R, L) studied here. Only diagonal covariances are used for all mixture components.

3.4. Results

Fig. 1 shows the average LID accuracy for the 6 languages (EN, GE, HI, JA, MA, SP), for both PPR and PSWR systems for the 4 different scores: (a) Acoustic score ($\mathbf{P}_A$), (b) Joint acoustic-language score ($\mathbf{P}_{AL}$), (c) Language-model score – Decoupled ($\mathbf{P}_{LD}$), and (d) Language-model score – Joint ($\mathbf{P}_{LJ}$). These results are for test data of 20 utterances / language each 45 sec long (story-bt utterances). For PSWR, there are two parameters for each of the four scores: i) Segmentation rate R, varying as $R = 2, 5, 10$ and 20 segments/sec, and ii) SWU inventory size L, varying as $L = 10, 30, 50$ and 100. Each plot ((a) to (d)) shows the LID accuracy (y-axis) vs $L = 10, 30, 50, 100$ (x-axis) for different $R = 2, 5, 10, 20$, as given in the legend in Fig. 1. The PPR system does not have any parameters and the LID accuracy is shown as a thick-dark line for each of the four scores.

The following can be observed from this figure: **i)** PPR performs equally well (with an LID accuracy of 70%) for all the four scores with slight advantage for $\mathbf{P}_{LD}$ or $\mathbf{P}_{LJ}$. **ii)** PSWR offers

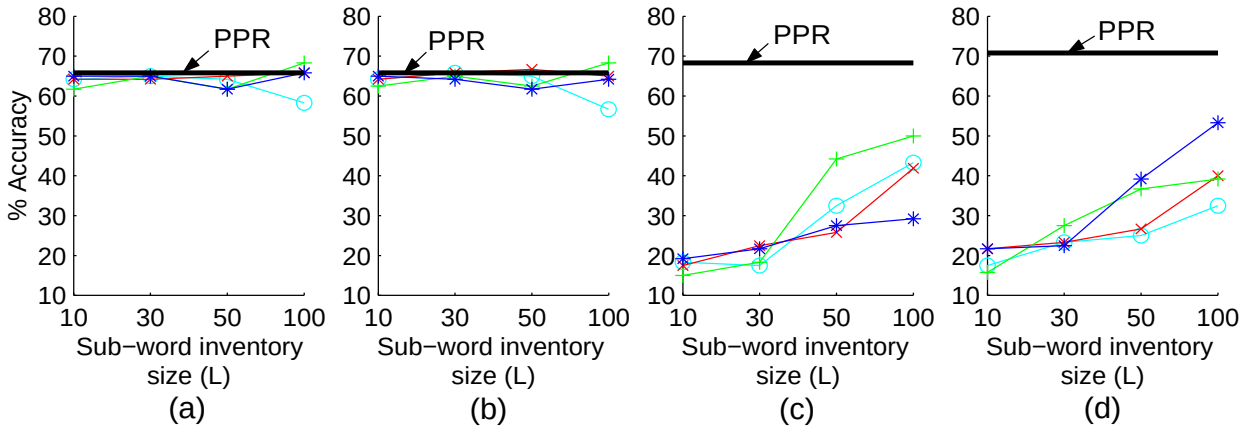

Fig. 1. Average LID accuracy of PPR and PSWR for different scores: (a) Acoustic score ($\mathbf{P}_A$), (b) Joint acoustic-language score ($\mathbf{P}_{AL}$), (c) Language-model score – Decoupled ($\mathbf{P}_{LD}$) and (d) Language-model score – Joint ($\mathbf{P}_{LJ}$). Test data: 20 utterances / language. Utterance length: 45 sec. Thick dark line: PPR. Legend for PSWR: R=2 (O), R=5 (X), R=10 (+), R=20 (*).

a performance comparable to PPR for both $\mathbf{P}_A$ and $\mathbf{P}_{AL}$ for all the SWU inventory sizes L and segmentation rates R. **iii)** The advantage of PSWR is the control on the granularity of the acoustic space in terms of variable size of SWUs obtained for different R and L. We had expected this to provide a means of determining whether any particular granularity is optimal for discrimination of languages. However, while $\mathbf{P}_A$ and $\mathbf{P}_{AL}$ showed no particular dependence on this granularity, $\mathbf{P}_{LD}$ (or $\mathbf{P}_{LJ}$) seems to require larger SWU inventory size (i.e., fine acoustic resolution) for higher PSWR performance. **iv)** Given that $\mathbf{P}_{AL}$ integrates both the acoustic and phonotactic information, and represents the complete PPR (and PSWR) system, PSWR performs as well as PPR, making it an efficient alternative to PPR, with the advantage of not requiring manually labeled training data for any of the languages in the task. **v)** The acoustic score $\mathbf{P}_A$ alone is also seen to be consistently high for both PPR and PSWR, indicating the possibility of better LID performance with improved acoustic modeling. **vi)** With regard to LM scores $\mathbf{P}_{LD}$ and $\mathbf{P}_{LJ}$, PSWR is distinctly poorer than PPR. It appears that the SWU inventory is not sufficiently unique across languages. This can also be ascribed to accumulation of errors from the sub-word tokenization stage.

4. CONCLUSIONS

We have proposed a parallel sub-word recognition system (PSWR) as an alternative to the parallel phone recognition (PPR) system for language identification (LID). The sub-word recognizer (SWR) used in the PSWR system can be obtained from training data without phonetic transcription in any of the languages in the task. It is based on automatic segmentation followed by segment clustering and segment HMM modeling. We have studied the complete PSWR system including the language modeling of sub-word unit sequences. The resulting PSWR system yields three types of scores, namely, acoustic score, joint acoustic-language score and language-model score. We have examined the effectiveness of these scores for a LID task of 6 languages in the OGI-TS database. We find that the PSWR performs comparably to PPR for both the acoustic and acoustic-language scores with an LID accuracy of 70% on test data (45 sec long), thus making it an efficient alternative to the PPR system.

5. REFERENCES

[1] Y. K. Muthusamy, E. Barnard, and R. A. Cole. Reviewing automatic language identification. *IEEE Signal Processing Magazine*, 11(4):33–41, Oct 1994.

[2] M. A. Zissman and K. M. Berkling. Automatic language identification. *Speech Communication*, 35(1-2):115–124, Aug 2001.

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

[4] A. K. V. Sai Jayram, V. Ramasubramanian, and T. V. Sreenivas. Automatic language identification using acoustic sub-word units. In *Proc. ICSLP*, pages 81–84, Denver, Colorado, Sep 2002.

[5] V. Ramasubramanian, A. K. V. Sai Jayram, , and T. V. Sreenivas. Language identification using parallel phone recognition. In *Workshop on Spoken Language Processing*, Jan 2003. Tata Institute of Fundamental Research. (submitted).

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika. The OGI multi-language telephone speech corpus. In *Proc. ICSLP*, pages 895–898, 1992.

[7] L. R. Rabiner and B.-H. Juang. *Fundamentals of speech recognition*. Prentice Hall, 1993.

[8] C. H. Lee, F. K. Soong, and B. H. Juang. A segment model based approach to speech recognition. In *Proc. ICASSP*, pages 501–504, New York, April 1988.

[9] A. K. V. Sai Jayram, V. Ramasubramanian, and T. V. Sreenivas. Robust parameters for automatic segmentation of speech. In *Proc. ICASSP*, pages I–513–I–516, Orlando, Florida, May 2002.

[10] S. J. Young et al. *The HTK Book*. Cambridge University Engineering Department, 2001.