IMPROVMENT OF LANGUAGE IDENTIFICATION PERFORMANCE BY AGGREGATED PHONE RECOGNIZER

Hosseini Amereii S.A., Homayounpour M.M.

Laboratory for Intelligent Sound and Speech Processing,
Amirkabir University of Technology
Hafez St., P.O. Box 15875-4413, Tehran, Iran
phone: +98 21 64542722, fax: +98 21 66495521, email: sabbas@aut.ac.ir, homayoun@aut.ac.ir
web: http://www.aut.ac.ir/official/homayoun

ABSTRACT

Two popular and better performing approaches to language Identification (LID) are Phone Recognition followed by Language Modeling (PRLM) and Parallel PRLM. In this paper, a new LID approach named Aggregated PRLM or APRLM is proposed. In PRLM based LID systems, only one phone recognizer is used, independently of the language targets. At the opposite, in PPRLM based LID systems, multiple phone recognizers are used, but always independently of the language targets. So it may happen that all phones of a language target don't occur in at least one of the tokenizers provided by the phone recognizers. In this paper, it is proposed that after the phone recognition step, to aggregate the phone sequences obtained by multiple phone recognizers and to provide a new phone sequence. Several language identification experiments were conducted and the proposed improvements were evaluated using OGI-MLTS corpus. Our results show that APRLM overcomes PPRLM about 1.3% in two language classification tasks.

1. INTRODUCTION

Automatic language identification (LID) is a process of determining the language identity corresponding to a given set of spoken queries. It is an important technology in many applications, such as spoken language translation, multilingual speech recognition, and spoken document retrieval [3]. In the past few decades, many statistical approaches to LID have been developed by exploiting recent advances in acoustic modeling of phone units and language modeling of n-grams of these phones. Acoustic phone models are used in language-dependent continuous phone recognition to convert speech utterances into sequences of phone symbols using phone language models. Then these acoustic and language scores are combined into language specific score for making an LID final decision [3, 1].

Syllable-like units have also been experimented [5]. To further improve the LID performance, other information, such as articulatory and acoustic [3], lexical knowledge and prosody [9], have also been integrated into LID systems. Zissman [1] experimentally showed that phonetic language models can sometimes be more powerful than the MFCC-based Gaussian Mixture Models (GMMs) [3]. In [10, 11] multi-layer kohonen self-organizing feature maps are applied for spoken language identification. Hierarchical LID (HLID) framework has been proposed and improved in [12, 13].

PRLM is an effective method for identifying the language of spoken messages. In this method the front-end phone recognizer can be trained for phones of a certain language and be used for recognition of other languages. Test language may have phones not included in the phone set of the front-end phone recognizer. Thus, it seems natural to look for a way to incorporate phones from more than one language into a PRLM-like system. For example, PPRLM has been proposed that run multiple PRLM systems in parallel with the single language front-end recognizers each trained for a different language. This approach requires that labelled training speech be available for more than one language. Although the labelled training speech does not need to be available for all or even any of the languages to be recognized [1].

PPRLM results are imperfect, if all phones of a certain language do not occur in at least one of the tokenizers of the PPRLM LID system. If PRLM do not cover all phones of a language, the PRLMs cannot model that language properly. Therefore, to obtain better results, for each language, PPRLM must contain a PRLM that covers all phones of that language.

Our proposed method aggregates multiple phone sequences that are tokenized by multiple phone recognizers. Then sequence score is computed by language models similar to PRLM. PRLM has single language phone tokenizer followed by language modeling, but our method, has Aggregated Phone Recognizer followed by Language Modeling (APRLM). Aggregated phone recognizer is a high performance phone recognizer that is constructed from multiple single language phone recognizers. Since tokenizer is the most important part of LID system, a high performance tokenizer can reduce the LID system error rate.

After this introduction, section 2 describes the APRLM proposed method. Section 3 presents dataset and the conducted experiments and finally section 4 concludes this paper.

2. DESCRIPTION OF THE SYSTEM

Phone recognizer is the most important part of an appropriate PRLM LID system. In this paper we propose to aggregate multiple sequences which are tokenized by multiple

phone recognizers using a voting procedure. HTK was used in our experiments for training and recognition [8].

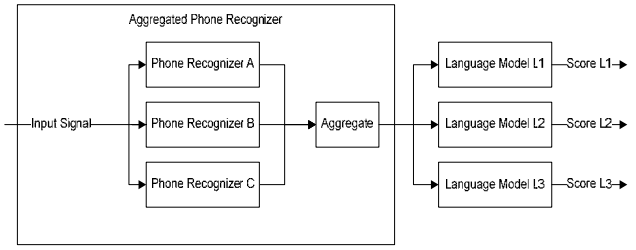

Figure 1 - Aggregated Phone Recognition followed by Language Modeling (APRLM) framework.

Sounds in languages to be identified do not always occur in the language used to train a phone recognizer. Due to this fact, PPRLM was proposed by Hazan [1]. Suppose that we want to identify a language L using a PPRLM including phone recognizers for languages A and B. Suppose language L include phone 'x' which is covered only by phone recognizer of language A and phone 'y' which covered only by phone recognizer of language B. Tokenization performance of speech signal from language L by A and B tokenizers are weak, since 'y' is not covered by A phone recognizer and 'x' is not also covered by B phone recognizer. Therefore results of both PRLMs would be weak. However, PPRLM can compensate this problem, but it is not adequate, because both two sequences is defective (sequence produced by A tokenizer does not include 'y' phones and B sequence does not include 'x' phones) and combination of them cannot satisfy LID task requirements. A sequence that contains all 'x' and 'y' phones together is needed to perform a better LID task. Aggregated phone recognizer is our solution; sequences produced by an aggregated tokenizer contain all phones existing in input utterances, if at least one tokenizer covers them.

The phone recognizer output includes the sequence of phones and their corresponding time interval and log-likelihoods. Substitution, insertion and deletion of phones are three errors that may occur in phone recognition. Also, start and end point of each recognized phone may be determined mistakenly. These errors are popular in PRLM and PPRLM and they are unavoidable. In this paper, we want to reduce these errors by voting among multiple sequences that have been tokenized by multiple phone recognizers.

In Aggregated PRLM (APRLM), aggregated phone recognizer is used to tokenize phone sequences of training and testing utterances. Language models are obtained by computing n-gram statistics of phone sequences. In recognition, score of any phone sequence is computed by each language model. Finally, scores obtained from language models determine test utterance language. APRLM is similar to PRLM in producing language models, computing test utterances scores and deciding about language.

For each phone, its start and end points and its score are obtained. Any sequence includes many phone observations. Each phone observation includes the phone name, its start and end points and its corresponding score. Here, we propose simple voting among phone observations in phone sequences; weighted voting can be used by considering score of each phone. The Aggregation algorithm is explained as follows:

2.1 Aggregation algorithm

Aggregation algorithm is performed in the following steps:

a- Create an empty sequence named aggregated sequence.

b- For each input utterance, tokenize it by at least three tokenizers. Each tokenizer produces one phone observation sequence.

c- At the first iteration of the algorithm choose the first phone observation in all phone observation sequences and choose the most frequent one and consider it as aggregated phone observation and add it to the aggregated phone sequence. Go to step e.

d- For the next iterations of the algorithm, choose the first unconsidered phone observation in all phone observation sequences and choose the most frequent one and consider it as aggregated phone observation and add it to the aggregated phone sequence.

e- Consider the aggregated phone observation in each phone observation sequence and choose the most frequent start point. Consider this start point as aggregated phone observation's start point.

f- Consider the aggregated phone observation in each phone observation sequence and choose the most frequent end point. Consider this end point as aggregated phone observation's end point.

g- In all sequences, each observation which its overlap with aggregated phone observation is more than half of the duration of aggregated phone observation will not be considered in determining the next aggregated phone observation.

h- The algorithm ends when all phone observations in all phone sequences are considered else go to step d.

i- End.

For more details, the algorithm is explained by an example given in table 1. In table 1, 4 sequences are aggregated. Each observation is shown by {phone (start point, end point)}. Underlined expressions show incorrectness in observations. ASeq means aggregated sequence. In this table it can be observed that how insertion, deletion and substitution errors are reduced by aggregation algorithm. Phone time alignments are also provided by aggregation procedure.

Table 1 – Aggregating four phone sequences

Seq.1	a(0,9), c(10,16), d(17,19), z(20,29), s(30,38)
Seq.2	a(0,8), *f(9,12)*, c(13,16), d(17,21), x(22,29), g(30,35)
Seq.3	g(0,10), c(11,17), d(18,21), x(22,26), u(27,28), s(29,34)
Seq.4	a(0,10), g(11,16), r(17,21), x(22,35){deletion}
ASeq.	a(0,10), c(11,16), d(17,21), x(22,29), s(30,35)

First observation in Seq1 is "a (0, 9)", in Seq2 is "a (0, 8)", in Seq3 is "g (0, 10)" and in Seq4 is "a (0, 10)". By voting among the first observations in all sequences, "a (0, 10)" will be selected. Now, "c (10, 16)", "f (9, 12)", "c (11, 17)" and "g (11, 16)" are the first unconsidered observations of sequences Seq1, Seq2, Seq3 and Seq4 respectively. "c (11, 16)" is selected as the next aggregated observation. By continuing the

aggregation algorithm, "d (17, 21)", "x (22, 29)" and "s (30, 35)" will be selected in the next iterations of the aggregation algorithm.

3. EXPERIMENTS

3.1. Database

The Oregon Graduate Institute Multi-Language Telephone Speech (OGI-MLTS) corpus [6, 7] was used to evaluate the performance of proposed method versus conventional PRLM and PPRLM methods. This corpus was divided into four segments: initial training set, development set, evaluation set and extended training set. Initial training set is phonetically labelled [6]. In this paper ten phone recognizers were trained using phonetically labelled messages of initial training segment of the OGI-MLTS corpus.

3.2. Conducted experiments

In our experiments, we trained ten phone tokenizers; therefore we have ten phone observation sequences for each utterance in training or test set. One phone tokenization output is created for each utterance by aggregating the phone observation sequences of all ten tokenizers. Language models are extracted using phone sequences obtained by tokenizing a sufficient number of utterances by each language tokenizer. Utterances from all ten languages are tokenized by each of ten language tokenizers and so 100 language models are obtained. We have also 10 language models for the aggregated phone observation sequences (one language model is extracted from the phone sequence obtained by aggregation of phone observations when utterances from each language are tokenized).

3.2.1 Pair-wise experiments

In this section, language-pair identification results are presented. Two PRLMs were used: one by English phone recognizer and other by Farsi phone recognizer. PPRLM and APRLM use all ten language phone recognizers. In each work, we classify languages pair wisely. Therefore, 90 experiments (10* 9 pair languages) for each approach exist. Table 2 presents the average of these 90 experiments. More details are presented in tables 3,4 and 5. In table 2 the results of evaluating the PRLM using English or Farsi tokenizer, PPRLM with ten tokenizers in ten languages and aggregated PRLM with an aggregated tokenizer are presented.

Table 2 – LID accuracy on 45s utterances

LID	L VS. L' (average)
PRLM (English Tokenizer)	82.9%
PRLM (Farsi Tokenizer)	81.5%
APRLM (Aggregated PRLM)	84.7%
PPRLM	83.4%

Right column in this table shows average accuracy of LID task between two languages, for example English vs. Farsi or French vs. German and so on. The results show that APRLM overcomes other approaches. Aggregated phone recognition improves LID performance in APRLM because its tokenization performance is better than other classic phone recognizers. APRLM covers more phones than single phone recognizer or multiple independent phone recognizers. Table 3 shows LID task on Japanese language versus other nine languages. Superiority of APRLM is evident in this table. Table 4 and table 5 show results of identifying French and German languages versus other nine languages respectively. In these tables, the superiority of APRLM can be observed.

Table 3 – LID accuracy Japanese versus Others

Language	PRLM (EN)	PRLM (FA)	PPRLM	APRLM
English	87%	76%	79%	84%
Farsi	87%	84%	90%	95%
French	84%	82%	92%	90%
German	89%	85%	87%	92%
Korean	92%	81%	89%	92%
Mandarin	84%	92%	83%	90%
Spanish	61%	66%	68%	72%
Tamil	92%	95%	92%	95%
Vietnamese	83%	85%	88%	95%
Average	**84%**	**83%**	**85%**	**89%**

Table 4 – LID accuracy French versus Others

Language	PRLM (EN)	PRLM (FA)	PPRLM	APRLM
English	74%	58%	79%	68%
Farsi	69%	71%	74%	67%
German	80%	72%	70%	80%
Japanese	84%	82%	92%	90%
Korean	83%	84%	86%	84%
Mandarin	78%	87%	81%	76%
Spanish	82%	83%	77%	89%
Tamil	100%	100%	97%	100%
Vietnamese	91%	91%	94%	94%
Average	**82%**	**81%**	**83%**	**83%**

Table 5 – LID accuracy German versus Others

Language	PRLM (EN)	PRLM (FA)	PPRLM	APRLM
English	67%	70%	75%	72%
Farsi	75%	63%	75%	70%
French	80%	72%	70%	80%
Japanese	90%	85%	87%	92%
Korean	76%	74%	85%	77%
Mandarin	76%	79%	79%	82%
Spanish	73%	68%	72%	90%
Tamil	100%	95%	94%	98%
Vietnamese	86%	82%	83%	92%
Average	**80%**	**76%**	**80%**	**83%**

3.2.2 Ten language Identification

Table 6 shows language identification results on ten languages. APRLM in ten languages identification overcomes PPRLM with different configuration, also. Increasing PPRLM tokenizers always don't improve LID performance, table 6 shows this claim precisely. PPRLM with two tokenizers in table 6, includes only English and Farsi phone recog-

nizers for sequencing input signal, and PPRLM with ten tokenizers includes phone recognizers of all languages.

Table 6 – LID accuracy on ten languages

LID system	Classification accuracy
PPRLM (two tokenizers)	62.6%
PPRLM (ten tokenizers)	58.9%
APRLM	64.5%

This paper claims that proposed method (APRLM) overcomes PPRLM, because its frond-end is more powerful than PPRLM tokenizers. Table 7, table 8 and table 9 are good evidences for this purpose. These tables show tokenizing results on Japanese, German and French utterances respectively. Correctness and accuracy of tokenizing task are computed as follows:

$$\text{Correctness} = \frac{N - D - S}{N} * 100 \qquad (1)$$

$$\text{Accuracy} = \frac{N - D - S - I}{N} * 100 \qquad (2)$$

Where N is the total number of phones; D, I, and S are the number of deletions, insertions and substitutions of phones respectively.

Table 7 Tokenizing Japanese speech utterances in train set

Tokenizer	N	I	D	S	Correctness	accuracy
EN Tokenizer	2066	258	294	239	74.20%	61.71%
FA Tokenizer	2066	232	307	242	73.43%	62.20%
Aggregated Tokenizer	2066	54	389	125	75.12%	72.51%

Table 8 Tokenizing German speech utterances in train set

Tokenizer	N	I	D	S	Correctness	accuracy
EN Tokenizer	3406	347	503	418	72.96%	62.77%
FA Tokenizer	3406	377	524	425	72.14%	61.07%
Aggregated Tokenizer	3406	115	698	255	72.02%	68.64%

Table 9 Tokenizing French speech utterances in train set

Tokenizer	N	I	D	S	Correctness	accuracy
EN Tokenizer	2528	388	288	375	73.77%	58.43%
FA Tokenizer	2528	349	316	335	74.25%	60.44%
Aggregated Tokenizer	2528	101	424	219	74.56%	70.57%

4. CONCLUSION

Presented results show the importance of phone recognizer in LID systems. Our proposed method named Aggregated PRLM outperforms both PRLM and PPRLM approaches. The conclusion of experiments is that it is better to have an aggregated tokenizer instead of single or multiple language-dependent tokenizers. In the PPRLM systems, multiple phone recognizer work independently and they have no mutual effect, but in aggregated tokenizer, multiple phone recognizers work independently and tokenizers aid each other. PPRLM is powerful in LID tasks if at least one of its tokenizers covers phones of the test language properly; otherwise its performance is not satisfiable. But to have better results in APRLM, it is adequate to cover phones of the test language by all its tokenizers together. The experiments results support this claim.

ACKNOWLEDGMENTS

The authors would like to thank Iran Telecommunication Research Center (ITRC) for supporting this work under contract No. T/500/14939.

REFERENCES

[1] M. A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech, " *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 1, pp. 31-44, 1996.

[2] H. Li and B. Ma "A Phonotactic Language Model for Spoken Language Identification, " Proc. ACL, 2005.

[3] B. Ma and H. Li "An Acoustic Segment Modeling Approach to Automatic Language Identification, " in *Proc. INTERSPEECH 2005*, Lisbon, Portugal, September 4-8. 2005, pp. 2829–2832.

[4] P. Matejka, P. Schwarz, J. ˇCernock´y and P. Chytil, "Phonotactic Language Identification using High Quality Phoneme Recognition, " in *proc INTERSPEECH 2005*, Lisbon, Portugal, September 4-8. 2005, pp. 2237-2240.

[5] T. Nagarajan and Hema A.Murthy, "Language Identification Using Parallel Syllable-Like Unit Recognition, " in *proc ICASSP 2004*, pp. I-401 – I-404.

[6] Y. K. Muthusamy, R. A. Cole, and B. T. Oshika, "The OGI Multilanguage telephone speech corpus," in *Proc. ICSLP*, Oct. 1992, pp. 895–898.

[7] "OGI multi language telephone speech," http://www.cslu.ogi.edu/corpora/mlts/.

[8] S. Young et al., "*The HTK Book*," Cambridge University Engineering Department," 2005.

[9] J. L. Rouas, "Modeling Long and Short-Trem Prosody for Language Identification," in *proc INTERSPEECH 2005*, Lisbon, Portugal, September 4-8. 2005, pp. 2257- 2260.

[10] Wang, L., Ambikairajah, E., and Choi, Eric, H.C., "Multi-Layer Kohonen Self-Organizing Feature Map for Language Identification," in *InterSpeech 2007*, August 27-31, Antwerp, Belgium, pp. 174-177.

[11] Wang, L., Ambikairajah, E., and Choi, Eric, H.C., "A Comparisonal Study of the Multi-Layer Kohonen Self-Organizing Feature Map for Spoken Language Identification," in *proc. ASRU 2007*, pp. 402- 407

[12] B. Yin, E. Ambikairajah, and F. Chen, "Hierarchical Language Identification based on Automatic Language Clustering," in *proc. InterSpeech–EuroSpeech 2007*, Antwerp, Belgium, 2007, pp. 178-181.

[13] Yin. B, Ambikairajah E., "Improvements on Hierarchical Language Identification Based on Automatic Language Clustering," in *proc ICASSP 2008*, pp. 4241- 4244.