Phonotactic Language Identification using High Quality Phoneme Recognition

Pavel Matějka, Petr Schwarz , Jan Černocký, Pavel Chytil

Brno University of Technology, Czech Republic
OGI School of Science & Engineering, OHSU, Portland, Oregon USA
{matejkap,schwarzp,cernocky}@fit.vutbr.cz,pchytil@bme.ogi.edu

Abstract

Phoneme Recognizers followed by Language Modeling (PRLM) have consistently yielded top performance in language identification (LID) task. Parallel ordering of PRLMs (PPRLM) improves performance even more. Since tokenizer is the most important part of LID system the high quality phoneme recognizer is employed. Two different multilingual databases for training phoneme recognizers are compared and the amount of sufficient training data is studied. Reported results are on data from NIST 2003 LID evaluation. Our four PRLM systems have Equal Error Rate (EER) of 2.4% on 12 languages task. This result compares favorably to the best known result from this task.

1. Introduction

Automatic language identification (LID) has increasing importance among speech processing applications. It can be used to route calls to human operators (commerce, emergency), preselect suitable speech recognition system (information systems) and has many uses in security applications.

The goal for Language Identification is to determine the language a particular speech segment was spoken. This work concentrates on phono-tactic approach to LID. The speech signal is first converted into a sequence of discrete sub-word (tokens) units that can characterize the language. In our case, these units are phonemes detected by a phoneme recognizer.

Conventional hidden Markov model (HMM) phoneme recognizer is used as a tokenizer for many phoneme recognition followed by language model (PRLM) LID systems [1, 2, 3]. In these systems, the phoneme recognizers are trained on OGI Stories corpus [4] designed as multilingual database. Unfortunately, this database does not contain enough data to train good phoneme recognizer to be used as tokenizer for LID. In [5], we have shown that the amount of training data is crucial for good performance of phoneme recognizer and we suggested to train the tokenizer on database where more transcribed data is available. We show that SpeechDat database [6] is suitable for this approach. Presented results show that **one** well trained PRLM on one language from SpeechDat outperforms six PRLM running in parallel trained on widely used OGI Stories database.

In section 2, description of the LID system is given. Section 3 presents data sets, evaluation method and experiments. Section 4 contains summary of results, comparison with published works and conclusions are presented.

2. Description of the System

Good phoneme recognizer is the most important part of an accurate PRLM LID system. We use a hybrid system based on Neural Networks (NN) and Viterbi decoder without any language model. The feature extraction makes use of long temporal context, known as TRAPs (temporal patterns) [7].

2.1. Phoneme recognizer - LCRC FeatureNet

The feature extraction uses Mel filter bank energies which are obtained in the conventional way. Temporal evolution of critical band spectral densities are taken. The temporal context is split into left and right context. This allows for more precise modeling of the whole trajectory while limiting the size of the model (number of weights in the NN) and redacting amount of necessary training data. Both parts are processed by discrete cosine transform (DCT) to de-correlate and reduce dimensionality. Two NNs are trained to produce the phoneme posterior probabilities for both context parts. Third NN functions as a merger and produces final set of posterior probabilities (see Figure 1). In [5], we have shown that this system outperfoms phoneme recognizers with GMM/HMM modelling.

2.2. Phono-tactic model - trigrams

Language model of 3rd order was used to capture phono-tactic statistics of each language. This was created by passing training speech of all target languages through phoneme recognizer and counting trigrams for each language separately. Phoneme insertion penalty (PIP) in the decoder is a constant which must be tuned for the specific task. This constant influences the output phoneme strings and can vary for different applications such as phoneme recognition or language identification. Here, it was tuned with the best LID performance as criterion. Problem of unseen trigrams is solved by replacing them by constant, which has to be experimentally tuned.

Arabic(Egyptian)	Japanese	Farsi
French(Canadian French)	German	Hindi
English(American)	Korean	Mandarin
Spanish(Latin American)	Tamil	Vietnamese

Table 1: The twelve target languages

2.3. Recognition

During recognition, the test sentence is passed through the phoneme recognizer. The resulting string of phonemes is processed by all phono-tactic models for each, the likelihoods of all trigrams are multiplied. Likelihoods are normalized over all languages. Finally we have scores for all target languages. Test sentence belongs to target language with the maximal score. If we merge output scores from several PRLM, each trained on different language, we have a system called Parallel PRLM denoted as PPRLM.

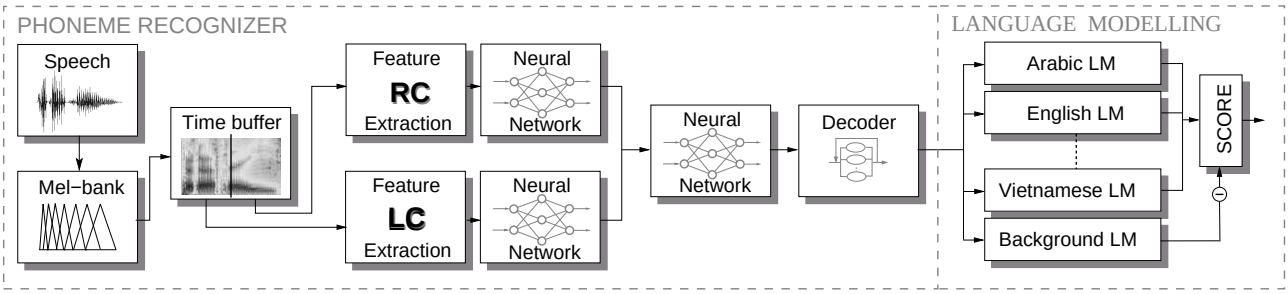

Figure 1: PRLM system based on phoneme recognizer with split temporal context

database	**SpeechDat-E**				**OGI Stories**					
language	CZE	HUN	POL	RUS	ENG	GER	HIN	JAP	MAN	SPA
train [hours]	9.72	7.86	9.49	14.02	1.71	0.97	0.71	0.65	0.43	1.10
test [hours]	2.26	1.97	2.34	3.89	0.42	0.24	0.17	0.15	0.11	0.26
cv [hours]	0.91	0.77	0.88	1.57	0.16	0.10	0.07	0.06	0.03	0.10
phoneme set [-]	46	62	41	53	40	44	47	30	45	39

Table 2: Amounts of speech data used to train phoneme recognizers

3. Experiments

3.1. Databases and Evaluation

All data used for experiments were conversational data recorded over telephone line.

Ten **phoneme recognizers** were trained on 4 languages from SpeechDat-E corpus [6] – Czech (CZE), Hungarian (HUN), Polish (POL) and Russian (RUS) and on 6 languages from OGI Stories Multilingual corpus [4] – English (ENG), German (GER), Hindi (HIN), Japanese (JAP), Mandarin (MAN) and Spanish (SPA). The amounts of data in training, cross-validation and test parts are in Table 2.

Language models were trained on the CallFriend Corpus [8]. Each of 12 target languages (Table 1) contains 20 complete half-hour conversations.

Test Data comes from NIST 2003 LID evaluation [9]. This data set consists of 80 segments with duration of 3, 10 and 30 second duration in each of 12 target languages (Table 1). This data comes from conversations collected for the CallFriend Corpus but not included in its publicly released version. In addition, there are four additional sets of 80 segments of each duration selected from other LDC[1] supplied conversational speech sources, namely Russian, Japanese, English, and cellular English.

Evaluation – Probabilities of false alarms and miss rejections are evaluated as a function of detection threshold. The point where these values are equal is referred to as the equal error rate (EER). The lower the EER value, the higher the accuracy of LID system.

3.2. Phoneme Recognizer generalities

All systems were trained on the databases described above. The length of 31 frames of the time trajectory in feature extraction was used in each critical band. This length was chosen based on experiments aiming at maximizing the phoneme accuracy [10]. Sentence mean normalization (Smn) was done on each critical band separately to perform channel normalization.

[1]Linguistic Data Consortium, *http://www.ldc.upenn.edu*

3.3. OGI Phoneme Recognizers

OGI Stories [4] is widely used for training phoneme recognizers for LID applications [1, 2, 3]. For each language from this database we divided data into three parts. Recognizers were trained on the training part. The increase in classification error on cross-validation part was used in NN as a stopping criterion to avoid over-training. Testing of performance was done on the test part. Results of these phoneme recognizers in terms of phoneme error rate are in Table 3. Unfortunately, the amount of transcribed data per language is only about one hour (see Table 2) which is not enough to train phoneme recognizer properly [5]. We are not looking on phoneme error rate in language identification therefore we can use all data from database for training phoneme recognizers and evaluate LID error rate. We merged train and test sets together. Our test sets were about 15 minutes in average but this represents about 20% of transcribed data! After this move we can no more evaluate phoneme error rate on the test data, as it was seen in the training. However, we can compare results on cross-validation part and see at least tendencies of phoneme recognizers (see Tale 3). The influence on LID EER is shown Table 4. The results prove correctness of our assumption: more data leads to better phoneme recognizer and lower LID EER.

Fusing of our PRLM systems was made by linear combination. The weights were tuned to optional LID EER on NIST 1996 development and evaluation data [11].

3.4. SpeechDat-E Phoneme Recognizers

Ten hours of training data per language was used (initial and final silence is not considered and records were selected from phonetically balanced sentences). It is ten times more data for training phoneme recognizers than in OGI Stories and we suppose this is a sufficient amount of data for training. Note, that no language from SpeechDat-E matches with any of the target languages. Such tokanization is closed to transcription of an unknown language by phonemes from language the tokenizer was trained on.

Differences of phoneme error rates between phoneme rec-

Language		ENG	GER	HIN	JAP	MAN	SPA
baseline	test	45.26	46.10	45.74	41.19	49.93	39.55
	cv	53.50	53.66	47.50	40.66	48.24	40.70
retrained	cv	52.81	56.76	47.31	35.78	44.03	38.89

Table 3: Phoneme error rate [%] on OGI Stories

Language	ENG	GER	HIN	JAP	MAN	SPA	fusion
PPRLM BUT-OGI = baseline	11.83	11.67	9.75	11.42	15.08	14.08	6.92
PPRLM BUT-OGIretrained	10.58	10.33	8.92	9.08	12.83	11.33	5.58

Table 4: EER [%] of single PRLMs trained on OGI Stories and tested on 30second task from NIST 2003 LID evaluation

ognizers trained on different amounts of data are shown in Table 5. If we compare the systems trained on 1 hour and 10 hours, there is absolute difference of almost 7.5% on phoneme level and 3.75% on LID EER on 30 second task. When we see the DET curves in Figure 2, it is evident that the first system is not well trained which may allow us to suspect all the phoneme recognizers trained on OGI multilingual database to be also badly trained.

training data [hours]	PER [%]	EER 30sec [%]	EER 10sec [%]	EER 3sec [%]
1	34.89	9.17	18.08	28.92
3	30.82	6.50	15.75	27.00
5	29.57	5.67	15.17	26.42
7	28.27	5.42	14.25	26.83
10	27.44	5.42	14.17	26.00

Table 5: Influence of number of training data for phoneme recognizer on PER[%] and EER[%] (LID NIST 2003) with SpeechDat-E Czech phoneme recognizer

Final phoneme error rates of phoneme recognizers trained on four languages from SpeechDat-E corpus are in Table 6. These phoneme recognizers produce LID EER presented in Table 7.

Language	CZE	HUN	POL	RUS
PER	27.44	41.96	39.91	35.92

Table 6: Phoneme error rate [%] on SpeechDat-E corpus

Language	CZE	HUN	POL	RUS	fusion
EER [%]	5.42	4.42	6.75	4.75	2.42

Table 7: EER [%] of single PRLMs trained on SpeechDat-E and tested on 30second task from NIST 2003 LID evaluation

Table 8 gives us comparison with the results of best known systems from literature. All PRLM systems used in experiment were tuned on NIST 1996 LID evaluation data [12]. Testing was performed on NIST 2003 LID evaluation data. Results of

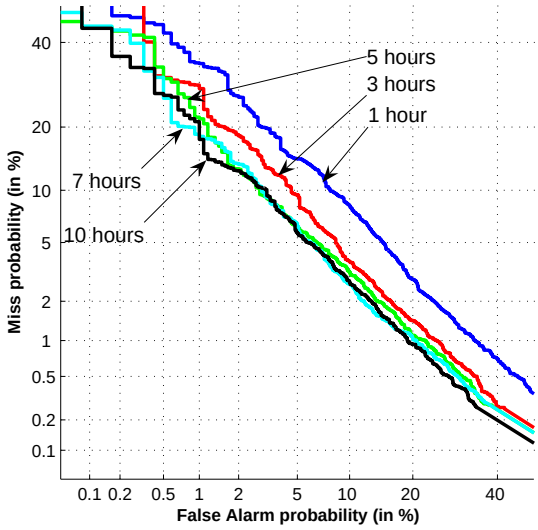

Figure 2: DET Plots of systems trained on different amount of data on Czech Language on 30sec task from NIST LID 2003

OGI[2] and MIT[3] PPRLM [3] trained on 6 languages from OGI stories are in the first rows of the table. System labeled "FUSE MIT" was based on merging of output of PPRLM mentioned above, Gaussian Mixture Model and Support Vector Machine trained on acoustic features [3]. Our PPRLM system trained on OGI Stories (PPRLM BUT-OGIretrained) outperforms OGI and MIT ones by about 1% on the 30 second task, which is significantly better on the level of 95% from Gaussian approximation. This is proving the superiority of our LCRC FeatureNet phoneme recognizer. When we train **one** PRLM system on SpeechDat-E databases (PPRLM BUT-SPDAT) it outperforms our system trained on OGI Stories and also the MIT PPRLM system. Our best result was achieved with PPRLM trained on four languages from SpeechDat-E database – this system favorably compares to system "FUSE MIT".

Figure 3 shows DET plots of our best PPRLMs for all three test – segment durations.

[2]OGI School of Science & Engineering, *http://cslu.cse.ogi.edu*
[3]Massachusetts Institute of Technology, *http://web.mit.edu*

SYSTEM EER(%)	1996			2003		
	30s	10s	3s	30s	10s	3s
PPRLM OGI	–	–	–	7.7	11.9	22.6
PPRLM MIT	5.6	11.9	24.6	6.6	14.2	25.5
FUSE MIT	2.7	6.9	17.4	2.8	7.8	20.3
PPRLM BUT-OGI	5.16	9.85	19.69	6.92	11.67	22.17
PPRLM BUT-OGIretrained	4.29	8.79	18.63	5.58	11.08	21.58
PPRLM BUT-SPDAT	**1.48**	**5.66**	**15.83**	**2.42**	**8.08**	**19.08**

Table 8: Comparison of EER [%] on NIST 1996 and 2003 evaluations

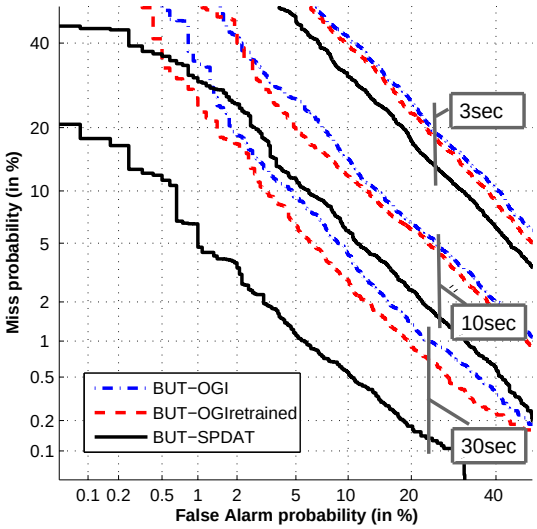

Figure 3: Comparison of DET Plots of our best systems on 2003 LID NIST evaluation

4. Conclusions

Presented results show dependencies between phoneme error rate and equal error rate in LID and necessity of huge amount of training data to train the tokenizer well. Three out of four well trained PRLM system outperform PPRLM system trained on 6 languages from OGI Stories. The conclusion of experiments is that it is better to have less well trained tokenizers then more poor ones. Our PPRLM trained on SpeechDat-E database outperformed all other PPRLM trained on OGI Stories by about 3%. With this PPRLM, we obtained also better results than MIT system integrating also acoustic scoring.

5. Acknowledgments

This work was partially supported by EC project Augmented Multi-party Interaction (AMI), No. 506811, Grant Agency of Czech Republic under project No. 102/05/0278 and by industrial grant from CAMEA Ltd. Jan Černocký was supported by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

6. References

[1] M.A. Zissman, "Comparison of four approaches to automatic language identification of telephone speech.," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 4, no. 1, pp. 31–44, 1996.

[2] Barnard E. Yan, Y., "An approach to automatic language identification based on language-dependent phone recognition," in *Proc. ICASSP 1995*, May 1995, pp. 3511–3514.

[3] Gleason T.P. Campbell W.M. Reynolds D.A. Singer E., Torres-Carrasquillo P.A., "Acoustic,phonetic,and discriminative approaches to automatic language identification," in *Proc. Eurospeech 2003*, Sept. 2003, pp. 1345–1348.

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

[5] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *Proc. TSD 2004*, Brno, Czech Republic, Sept. 2004, number ISBN 87-90834-09-7, pp. 465–472.

[6] "SpeechDat-East project," http://www.fee.vutbr.cz/SPEECHDAT-E.

[7] H. Hermansky and S. Sharma, "Temporal patterns (traps) in ASR of noisy speech," in *Proc. ICASSP 1999*, Phoenix, Arizona, Mar. 1999, pp. 2427–2431.

[8] "Callfriend corpus, telephone speech of 15 different languages or dialects," www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone.

[9] Przybocki-M.A. Martin, A.F., "NIST 2003 language recognition evaluation," in *Proc. Eurospeech 2003*, Sept. 2003, pp. 1341–1344.

[10] P. Schwarz, P. Matějka, and J. Černocký, "Recognition of phoneme strings using TRAP technique," in *Proc. Eurospeech 2003*, Geneva, Switzerland, Sept. 2003, pp. 825–828.

[11] "National institute of standard and technology," http://www.nist.gov.

[12] P. Matějka, "Tuning phonotactic language identificaion system," Tech. Rep., Brno University of Technology, Department of Computer Graphics and Multimedia, Feb. 2005.