Phonotactic Language Identification*

Pavel Matějka[1,2], Petr Schwarz[2], Jan Černocký[2], Pavel Chytil[1,3]

[1] Faculty of Electrical Eng. and Comm., Brno University of Technology
[2] Faculty of Information Technology, Brno University of Technology
[3] Department of Biomedical Engineering, OGI school of Science & Technology, OHSU
E-mail: matejkap@feec.vutbr.cz

This paper provides brief description of Language Identification (LID) system based on phoneme recognizer followed by language models (PRLM). Reported results are on data from NIST 2003 LID evaluation. Our system has Equal Error Rate (EER) 4.8% on task with 12 languages. This result compares favorably to the best known Parallel PRLM results from this evaluation.

1 Introduction

The goal for LID is to determine the language of particular speech segment. This work concentrates on phono-tactic approach to language identification. Speech signal is first converted into a sequence of meaningful discrete sub-word units (tokens) that can characterize language. In our case, these units are phonemes detected by a phoneme recognizer. The phoneme strings are modeled by statistical language model. We can consider phonemes as a meaningful units, because words in different languages differ and have different pronunciation. We can use phoneme recognizer to tokenize speech into phonemes even if phoneme recognizer is not trained on the target language. In this case such transcription is like pronunciation of the unknown language by phonemes from language the tokenizer was trained on.

This article is description of our baseline system. In section 2 description of whole LID system is given. In section 3 we describe data, evaluation method and experiments. Results summarization, comparison and conclusion is given in section 4.

*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, by industrial grant from CAMEA Ltd. Jan Černocký was supported by post-doctoral grant of Grant Agency of Czech Republic No. GA102/02/D108.

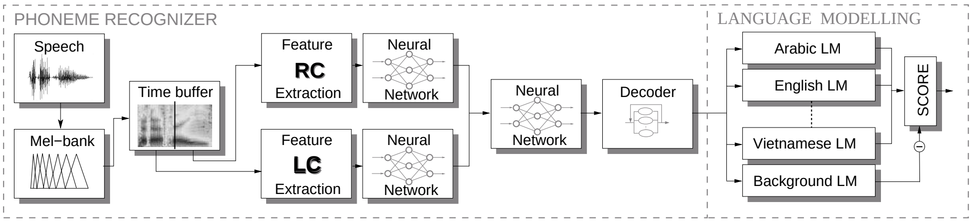

Figure 1: PRLM system based on phoneme recognizer with split temporal context

2 System Description

Good tokenizer is the most important part of an accurate LID system. We use a phoneme recognizer – a hybrid system based on Neural Networks (NN) and Viterbi decoder without any language model. An unconventional feature extraction technique based on long temporal context [1] is used.

Phoneme recognizer - FeatureNet

The feature extraction uses Mel filter bank energies which are obtained in the conventional way. In the first system, temporal evolution of critical band spectral densities are taken, windowed and projected (with dimensionality reduction) by discrete cosine transform (DCT). The feature vector which is feed to NN is created by concatenation of vectors over all filter bank energies. NN is trained to produce estimates of phoneme posterior probabilities.

Phoneme recognizer - LCRC FeatureNet

More complicated structure was chosen for the second system. 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). Both parts are processed by 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.

Language model - 3-grams

Language model of 3rd order was used to capture phonotactic statistics of each language. It was created by passing training speech of all target languages by phoneme recognizer and counting 3-grams for each language separately. Phoneme insertion penalty (PIP) in the decoder is a constant which has to 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. It was tuned for this task.

Recognition During recognition, the test sentence is passed through the phoneme recognizer. Sums of the likelihoods of the phoneme 3-grams are calculated from all language models of target languages separately. We have scores for all target languages at the end. Test sentence belongs to target language with maximal score. This system is generally known as Phoneme Recognizer followed by Language Models (PRLM). If we merge output scores of several PRLM with phoneme recognizer trained on different language we have system called Parallel PRLM (PPRLM).

data	Czech	Hungarian	Polish	Russian
train [hours]	9.72	7.86	9.49	14.02
test [hours]	0.91	0.77	0.88	1.57
phoneme set [-]	46	62	41	53

Table 1: Amounts of data [hours] used to train phoneme recognizers

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

Table 2: The twelve target languages

3 Experiments

Two different types of phoneme recognizers are evaluated in terms of phoneme accuracy and language identification error rate.

3.1 Databases and Evaluation

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

Four **phoneme recognizers** were trained on 4 out of 5 languages from SpeechDat-E corpus [2] – Czech, Hungarian, Polish and Russian. The amounts of data are in Table 1.

Language models (3grams) were trained on data from CallFriend Corpus [1] . Each of 12 target languages (Table 2) contains 20 complete half-hour conversations.

Test Data come from NIST 2003 LID evaluation [3]. This data set consists of 80 segments of 3, 10 and 30 second duration in each of 12 target languages (Table 2). This data come 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 supplied conversational speech sources, namely Russian, Japanese, English, and cellular English.

Evaluation False alarms and miss rejections are counted with different threshold. If this numbers are equal the common value is referred to as the equal error rate (EER). The value indicates that the proportion of false alarm is equal to the proportion of miss rejections. The lower the EER value, the higher the accuracy of LID system.

3.2 Phoneme Recognition

Systems were trained on the databases described above. The length of 31 frames of the time trajectory was used for feature extraction. This length comes from the best results on phoneme accuracy [4]. Sentence mean normalization (Smn) on each critical band separately was done to perform channel normalization. Results of the systems trained on different languages are in Table 3. Test set was part of the same language of course unseen in the training. Phoneme Insertion Penalty (PIP) was tuned to maximal phoneme accuracy.

SYSTEM [%]	**FN**	**FNSmn**	**LCRC**	**LCRCSmn**
SPDAT Czech	64.97	67.07	70.82	72.56
SPDAT Russian	50.70	52.19	56.91	58.04
SPDAT Polish	54.48	55.58	59.98	60.09
SPDAT Hungarian	58.55	59.26	63.64	64.08

Table 3: Accuracy of phoneme recognition on different languages

3.3 Language Identification

Phoneme recognizers described above were used to tokenize speech utterances. All non-speech tokens were merged to one.

Table 4 describes results of PRLM system with two types of phoneme recognizers trained on particular language. Phoneme insertion penalty (PIP) for these recognizers was tuned on NIST 1996 development and evaluation sets with minimal LID EER as criterion for each system separately.

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

SYSTEM	**CZECH**		**POLAND**		**HUNGARIAN**		**RUSSIAN**	
length[s]	FN	LCRC	FN	LCRC	FN	LCRC	FN	LCRC
PIP	-1.5	-1.5	-1	-1	-1	-1	-1	-3
30	7.42	6.42	7.83	7.83	6.83	**4.83**	7.58	6.25
10	15.00	14.83	15.83	17.08	15.42	**13.00**	14.83	14.67
3	25.75	26.42	25.92	27.83	26.00	25.83	**25.08**	25.58

Table 4: EER - Equal Error Rate [%] on NIST 2003 LID evaluation

SYSTEM EER(%)	**1996**			**2003**		
	30s	**10s**	**3s**	**30s**	**10s**	**3s**
MIT	5.6	11.9	24.6	6.6	14.2	25.5
OGI	–	–	–	7.71	11.88	22.60
1x LCRC (Hungarian)	2.82	10.19	23.49	4.83	13.00	25.83
4x LCRC	2.15	6.32	16.83	3.00	9.08	19.58

Table 5: Comparison of EER [%] on NIST evaluations of different PPRLM systems

Table 5 gives us comparison with the results of known systems from literature. Results of OGI and MIT PPRLM trained on 6 languages from OGI stories [5] are in the first lines of the table. Our best system based on one Hungarian phoneme recognizer is on the next line. Fusion of 4 systems based on phoneme tokenization (PPRLM) is on the last line of the table.

4 Conclusion

The PRLM system based on one phoneme recognizer trained on Hungarian outperforms MIT and OGI PPRLM system with six phoneme recognizers trained on OGI stories. The difference is more than 1.5% absolute and it is significantly better on the level of 95% from Gaussian approximation. The best results were obtained by fusing of four phoneme recognizers trained on Czech, Hungarian, Polish and Russian in PPRLM system. Performance of this system is 3.00% and it is significantly better then the result with one recognizer.

References

[1] 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.

[2] "SpeechDat-East project," http://www.fee.vutbr.cz/SPEECHDAT-E, Jan. 2004.

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

[4] 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.

[5] "OGI multi language telephone speech," http://www.cslu.ogi.edu/corpora/mlts/, Jan. 2004.