PHONEME RECOGNITION TUNING FOR LANGUAGE IDENTIFICATION SYSTEM

Pavel MATĚJKA, Doctoral Degree Programme (4)
Dept. of Radio Electronics, FEEC, BUT
E-mail: matejkap@feec.vutbr.cz

Supervised by: Dr. Milan Sigmund, Dr. Jan Černocký

ABSTRACT

This paper provides brief description of Language Identification (LID) system based on phoneme recognizer followed by language models (PRLM). Tuning phoneme recognizers for this task can increase performance of the whole system. Reported results are on data from NIST 2003 LID evaluation. Our system has Equal Error Rate (EER) 5.4% 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 meaningful units, because words in different languages differ and have different pronunciations. We can use a phoneme recognizer to tokenize speech into phonemes even if this recognizer is not trained on the target language. In this case such tokanization is closed to transcription 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. Summary of results, comparison with published results and conclusions are given in section 4.

2 DESCRIPTION OF THE SYSTEM

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 de-

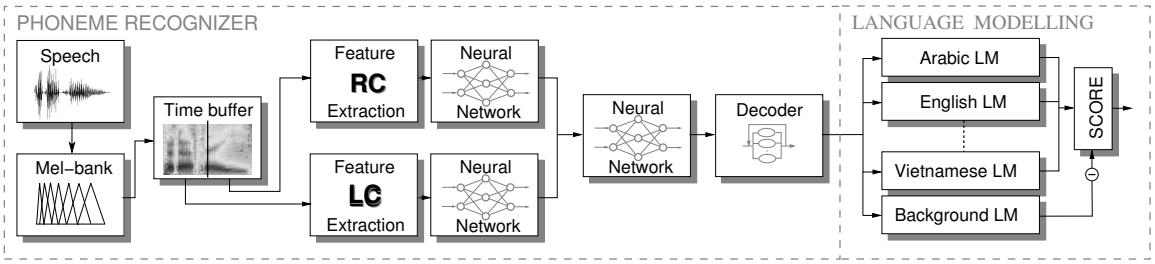

Figure 1: PRLM system based on phoneme recognizer with split temporal context

coder without any language model. An unconventional feature extraction technique based on long temporal patterns (TRAPs) [1] is used (see Figure 1).

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). 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 [2] for both context parts. Third NN functions as a merger and produces final set of posterior probabilities.

2.2 LANGUAGE MODEL - TRIGRAMS

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 trigrams 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.

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

Table 1: The twelve target languages

2.3 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 recognizers trained on different languages we have a system called Parallel PRLM (PPRLM).

3 EXPERIMENTS

3.1 DATABASES AND EVALUATION

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

Phoneme recognizer was trained on Czech part of SpeechDat-E corpus [1]. There is 9.72 hours of training data and 0.91 hours of testing data.

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

Test Data comes from NIST 2003 LID evaluation [4]. This data set consists of 80 segments 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[2] 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 referred to as the equal error rate (EER). The lower the EER value, the higher the accuracy of LID system.

3.2 PHONEME RECOGNIZER

The system was trained on the database described above. The length of 31 frames of the time trajectory for each filter bank was used for feature extraction. This length comes from the best results on phoneme accuracy [5]. Sentence mean normalization (Smn) was done on each critical band separately to perform channel normalization. **Phoneme accuracy** of the system **is 72.56%**. The test set was part of the database unseen in the training.

3.3 LANGUAGE IDENTIFICATION

Phoneme insertion penalty (see section 2.3) was tuned on NIST 1996 development and evaluation sets with minimal LID EER as criterion.

Phoneme recognizer described above was used to tokenize speech utterances. At first original stream of phonemes from phoneme recognizer was used for training and testing performance of system. Then we tested approaches with non-speech tokens mapped to silence. The non-speech tokens help us to improve acoustic model, but they are not relevant to the language model. There are two types of non-speech tokens in our system:

– **Int** . . . Intermittent noise. Environmental noise not generated by the speaker.

– **Spk** . . . Speaker noise (cough, lip-smack, etc.).

In the first experiment, the “Int” noise was mapped to silence, because this noise had not been produced by speaker and should not carry any information about language. Presented results show correctness of this assumption. In the second experiment even “Spk” noise was mapped to silence. This also improved the EER.

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

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

Results of tuning phoneme insertion penalty for this task are in Table 2 – they were obtained on NIST 1996 LID evaluation data. We don't know the length of incoming utterance in the real applications therefore we have to choose one PIP for the system. Ideal PIP is -3.0 for baseline "Raw" and "Int$-$>Sil" system and -1.5 for system with merged all silence tokens "Int + Spk$-$>Sil" (see Table 2).

Another tuning needs to be done in the language modeling part of the system. Some trigrams don't occur in the training of target language. Dealing with this is also important. Experimental penalty for unseen trigrams was tuned on NIST 1996 LID evaluation. Results are presented in Table 2 and are denoted as "Tuned" systems.

PIP	duration[s]	-1.0	-1.5	-2.0	-2.5	3.0	-3.5
Raw (Baseline)	30	5.04	5.15	5.08	5.07	**4.79**	4.87
Int $-$ > Sil	30	4.62	4.70	4.70	4.73	**4.51**	4.60
Int + Spk $-$ > Sil	30	4.30	**4.25**	4.36	4.28	4.43	4.28
Tuned - Raw	30	4.28	4.12	4.19	4.11	**4.08**	4.11
Tuned - Int $-$ > Sil	30	3.93	4.10	4.14	3.94	**3.80**	3.87
Tuned - Int + Spk $-$ > Sil	30	**3.59**	3.79	3.95	3.83	3.60	3.65

Table 2: EER [%] on NIST 1996 LID evaluation for different phoneme insertion penalties and 3-grams language model

Table 3 gives us comparison with the results of best known systems from literature. Testing was performed on NIST 2003 LID evaluation data with tuned penalty (last three lines), which wasn't seen during the training and tuning. Results of OGI[3] and MIT[4] PPRLM [6] trained on 6 languages from OGI stories [7] are in the first lines of the table. Result of our baseline system with tuned PIP is between MIT and OGI ones in the 30 second task. Results on shorter utterances are worse. Our best system is about 1% better than the MIT one as one of the best system on the world.

Evaluation	**1996**			**2003**		
SYSTEM EER(%)	**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
Raw (Baseline)	4.96	14.05	31.07	7.25	17.75	31.75
Int $-$ > Sil	4.43	12.92	25.48	6.50	15.92	26.75
Int + Spk $-$ > Sil	3.96	12.32	25.68	6.42	14.83	26.42
Tuned Raw	4.23	12.92	29.87	6.25	16.25	31.42
Tuned Int $-$ > Sil	4.02	11.98	24.68	5.75	14.67	26.25
Tuned Int + Spk $-$ > Sil	3.69	10.85	23.75	5.42	14.17	26.00

Table 3: Comparison of EER [%] on NIST evaluations with different PPRLM systems

[3]OGI School of Science & Engineering, *http://cslu.cse.ogi.edu*

[4]Massachusetts Institute of Technology, *http://web.mit.edu*

4 CONCLUSION

The PRLM system based on one phoneme recognizer trained on Czech language favorably compares to MIT and OGI PPRLM systems. The difference between our best PRLM and MIT PPRLM is more than 1%. This result is significantly better on the level of 95% from Gaussian approximation. If we take into account that we use only one phoneme recognizer and MIT and OGI use six of them and merge them together then our results are encouraging for further investigation. Using MIT and OGI structure to implement PPRLM [8] is one way to improve system performance [9].

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. Thanks to Pavel Chytil from Department of Biomedical Engineering, OGI school of Science & Technology, OHSU, Oregon USA for help with this task.

REFERENCES

[1] Schwarz, P., Matějka, P., Černocký, J.: "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] Bourlard, H., Morgan, N.: *Connectionist speech recognition: A hybrid approach.*, Kluwer Academic Publishers, Boston, USA, 1994.

[3] "Callfriend corpus, telephone speech of 15 different languages or dialects," www.ldc.upenn.edu/Catalog/byType.jsp#speech.telephone, Jan. 2005.

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

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

[6] 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.

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

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

[9] Matějka, P., Schwarz, P., Černocký, J. Chytil, P.: "Language identification using high quality phoneme recognition," accepted to RADIOELEKTRONIKA 2005.