BUT language recognition system for NIST 2007 evaluations

Pavel Matějka, Lukáš Burget, Ondřej Glembek, Petr Schwarz, Valiantsina Hubeika, Michal Fapšo, Tomáš Mikolov, Oldřich Plchot, and Jan "Honza" Černocký

Speech@FIT, Faculty of Information Technology, Brno University of Technology, Czech Republic
matejkap|burget|glembek|schwarzp|xhubei00|ifapso|imikolov|iplchot|cernocky@fit.vutbr.cz

Abstract

This paper describes Brno University of Technology (BUT) system for 2007 NIST Language recognition (LRE) evaluation. The system is a fusion of 4 acoustic and 9 phonotactic subsystems. We have investigated several new topics such as discriminatively trained language models in phonotactic systems, and eigen-channel adaptation in model and feature domain in acoustic systems. We also point out the importance of calibration and fusion. All results are presented on NIST 2007 LRE data.

Index terms: language recognition, phonotactic LRE, acoustic LRE, decision trees, intersession variability, system calibration, system fusion.

1. Introduction

NIST coordinated recent evaluations of automatic language recognition systems in 1996, 2003, 2005 and 2007. This paper describes Brno University of Technology (BUT) system for LRE 2007 [5]. The system is a fusion of 4 acoustic systems and 9 phonotactic ones. The work builds on our previous LRE 2005 system [11] but also brings several new sub-systems such as binary decision trees, discriminatively trained language models in phonotactic systems, and eigen-channel adaptation in model and feature domain in acoustic systems. This paper gives an overall presentation of all the systems and deals also with system calibration which is essential to obtain good system performance. For detailed information on our work on discriminative training and channel compensation for acoustic language recognition, refer to [8]. The advances in phonotactic language recognition are presented in [6].

Our submission was only for the 'closed set' condition of the 14-class 'General LR' test. All scores can be interpreted as log likelihood ratios.

2. Data and metric

Four kinds of data were used. Besides the "hot" evaluation data, the systems needed to be trained (training data), calibrated and tuned (Dev1 data) and tested (Dev2 data).

This work was partly supported by European projects AMIDA (FP6-033812), Caretaker (FP6-027231) and MOBIO (FP7-214324), by Grant Agency of Czech Republic under project No. 102/08/0707 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under project No. 201/2006. L. Burget was supported by Grant Agency of Czech Republic under project No. GP102/06/383.

We would like to thank to MIT Lincoln Labs for preparing the test and dev sets. Special thanks to Niko Brümmer for his excellent FoCal toolkit. Thanks MIT, Niko, and David van Leeuwen for valuable discussions and support to our group. Thanks also to AMIDA LVCSR team coordinated by Thomas Hain.

Evaluation data: The number of languages to be detected has significantly increased since the last evaluation in 2005. There are 14 languages that were used as detection targets in LRE07 [5]. The evaluation set contains test segments with three nominal durations of speech: 3, 10 and 30 seconds. Actual speech durations varied but were constrained to be within the ranges of 2-4 seconds, 7-13 seconds, and 25-35 seconds of actual speech contained in segments, respectively. The silence was not removed from speech so a segment could be much longer. Unlike previous evaluations, the nominal duration for each test segment was not identified. There were more than 7500 segments to identify.

Training data: The selection of training data was quite challenging and we used a variety of corpora distributed by LDC and ELRA to train our systems. The amounts of data ranged from 264 hours for English until mere 1.45 hours for Thai (Table 1).

Development data data for this evaluation were defined by MIT Lincoln Labs. They have nominal duration 3, 10 and 30 seconds. The sets were based on segments from previous evaluations plus additional segments extracted from longer files from training databases (which were not included in the training set). *Dev1 data* were based on NIST LRE 1996 and 2003 and additional segments from Fisher, CallHome and Mixer databases. The set contains 5165 trials and served for system tuning and calibration (especially the back-ends). *Dev2 data* were based on NIST LRE 2005 and additional segments from OGI stories and Mixer database. This set contains 5884 trials and served to test the system during the development.

Evaluation Metric: According to [5], basic pair-wise language recognition performance is computed for all target and non-target language pairs, and represented by miss and false alarm probabilities. From these and application-motivated costs, the average cost performance C_{avg} is computed [5] which is our primary evaluation metric. All results are reported as $100 \times C_{avg}$.

3. System description

The system was a fusion of 13 sub-systems: 4 acoustic ones and 9 phonotactic. All sub-system descriptions are completed with a "code" of the system for easy identification.

For all systems, the pre-processing was done by voice activity detector (VAD) based on our Hungarian phone recognizer with all the phoneme classes linked to 'speech' class. The silence is not used in acoustic systems.

3.1. Acoustic systems

The feature extraction was identical for all acoustic systems and was the same as in our LRE 2005 system [11]: 7 MFCC coeffi-

	sum	CF	CH	F	SRE	LDC07	OGI	OGI22	Other
Arabic	212	19.5	10.4	175	5.93	1.45		0.33	
Bengali	4.27				2.86	1.42			
Chinese	93.2	41.7	1.64	17.2	44.9	4.2	0.87	0.85	
English	264	39.8	4.68	162	34.9		6.77	0.52	15.6 (FAE)
Hindustani	23.5	19.6			0.64	1.32	1.53	0.42	
Spanish	54.3	43.8	6.71		2.63		1.18	0.38	
Farsi	22.7	21.2			0.03		1.00	0.42	
German	28.2	21.6	5.10				1.12	0.38	
Japanese	23.9	19.1	3.47				0.87	0.35	
Korean	19.7	18.4			0.09		0.72	0.5	
Russian	15.1				3.38	1.33		0.43	10.0 (SpDat)
Tamil	19.6	18.4					0.96	0.26	
Thai	1.45				0.15	1.23			
Vietnamese	21.6	20.6					0.79	0.27	
Other	62.5	20.7					1.10	3.29	37.4 (SpDat)

Table 1: Training data in hours for each language. Sources: **CF:** CallFriend, **CH:** CallHome, **F:** Fisher English Parts 1. and 2., Fisher Levantine Arabic, HKUST Mandarin, **SRE:** Mixer (data from NIST SRE 2004, 2005, 2006), **LDC07:** development data for NIST LRE 2007, **OGI:** OGI-multilingual, **OGI22:** OGI 22 languages, **FAE:** Foreign Accented English, **SpDat:** SpeechDat-East (see http://www.fee.vutbr.cz/SPEECHDAT-E or the ELRA/ELDA catalog).

cients (including coefficient C0) concatenated with SDC 7-1-3-7, which totals in 56 coefficients per frame.

Vocal-tract length normalization (VTLN) was done with the same models and in the same way as in NIST LRE 2005 [11]. The warping factors are estimated using single GMM (512 Gaussians), ML-trained on the whole CallFriend database (using all the languages). The model was trained in standard speaker adaptive training (SAT) fashion in four iterations of alternately re-estimating the model parameters and the warping factors for the training data.

3.1.1. GMM system with 2048 Gaussians per language and eigen-channel adaptation `GMM2048-eigchan`

The inspiration comes from our GMM system for speaker recognition [3] which follows conventional Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) paradigm and employs number of techniques that have previously proved to improve the GMM modeling [11].

Each language model is obtained by traditional *relevance MAP* adaptation of UBM using enrollment conversation. In the verification phase, standard Top-N Expected Log Likelihood Ratio (ELLR) scoring is used to obtain verification score, with $N = 10$. For each trial, both the model of target language and UBM are adapted to channel of test conversation using simple eigen-channel adaptation [1] prior to computing the log likelihood ratio score. We adopted the term 'eigen-channel' as used in speaker recognition (SRE) by Kenny [9]. The technique consists of eigen-channel subspace estimation (training phase) and eigen-channel adaptation (testing) and is described in detail in [3, 8].

3.1.2. GMM-MMI: `GMM256-MMI`

This subsystem uses GMM models with 256 Gaussians per language, where mean and variance parameters are re-estimated using Maximum Mutual Information criterion - the same as for LRE 2005 [11].

3.1.3. GMM-MMI with channel compensated features: `GMM256-MMI-chcf`

Similar set of GMM models with 256 Gaussians per language are trained with Maximum Mutual Information criterion. However, the features are first compensated using eigen-channel adaptation in feature domain [8].

3.1.4. SVM on GMM super-vectors: `GMM512-SVM`

In this system, GMM super-vectors (concatenated GMM mean vectors obtained by MAP adapting UBM to given speech segment) are extracted not only from target-model training speech segments, but also for all other background and test speech segments. In other words, each speech segment is represented by a single GMM super-vector. The target and background super-vectors are then used to train support vector machine (SVM) models of target languages against which the test super-vectors are scored. The SVM uses a linear kernel in super-vector space. Each SVM is trained using all available positive examples from the target language, and many negative examples from other languages.

3.2. Phonotactic systems

The phonotactic systems were based on 3 phone recognizers: two ANN/HMM hybrids and one based on GMM/HMM context-dependent models.

The two *hybrid phone recognizers* [11] are based on ANN/HMM approach, where artificial neural networks (ANN) are used to estimate posterior probabilities of phones from Mel filter bank log energies using split left and right contexts (LCRC) of 310ms around the current frame. They were trained on Hungarian and Russian SpeechDat-E databases.

One *GMM/HMM phone recognizer* was based on context-dependent state-clustered triphone models, which are trained in similar way as the models used in AMI/AMIDA LVCSR [7]. The models were trained using 2000 hours of English telephone conversational speech data from Fisher, Switchboard and CallHome. The features are 13 PLP coefficients augmented with their first, second and third derivatives projected into 39 dimensional space using HLDA transformation. The models are trained discriminatively using MPE criterion. VTLN and CM-

LLR adaptations are used for both training and recognition in SAT fashion. The triphones were used for phone recognition with a bi-gram phonotactic model trained on English-only data.

All the recognizers were able to produce phone strings as well as phone lattices. In case of lattices, posterior-weighted counts ("soft-counts") were used. Detailed description of our phonotactic systems can be found in [6].

3.2.1. 4-gram language model based on strings: `HU_strLM, EN_strLM`

These systems use 4-gram model estimated on phone strings from the Hungarian LCRC and English GMM/HMM phone recognizers. In the case of Hungarian phone recognizer, the LM for each language was derived by interpolating several LMs. In the case of English phone recognizer, the final target language LMs were interpolated with single LM trained on all languages together. This was helpful because of the limited amount of data to train LMs. The interpolation weights were tuned to give minimal perplexity on Dev1 set. Witten-Bell smoothing was used and pruning using minimal count was applied.

3.2.2. 3-gram language model on lattice counts: `HU_LM, RU_LM`

The phonotactic models were based on soft-counts but they were *adapted* from "UBM" trained on all data in the same way as in decision tree based phonotactic models [10].

3.2.3. Binary decision trees on lattice counts: `HU_TREE_A3E7M5S3G2_FA, RU_Tree and EN_Tree`

In all our systems, binary decision tree language modeling was based on creating a single language independent tree (referred to as "UBM") and adapting its distributions to individual language training data, as described by Navratil [10]. While the sub-systems built on Russian LCRC and English GMM/HMM phone recognizers use this basic approach only, the Hungarian output was processed in a more complex way using Multi-models and applying factor analysis for intersession compensation of phonotactic statistics.

Multi-models: Instead of merging all resources (databases) of one language together for a UBM adaptation, those resources with large amount of data were "hand-clustered", and a single LM was created for each of these clusters (e.g. 7 LMs for English, see the abbreviation `A3E7M5S3G2`). Such hand-clustering reflected some specifics such as foreign-accented English, different dialects, etc. A linear back-end was used to post-process these individual outputs to come up with one score per language.

Factor analysis is a method we have proposed to compensate for inter-session variation in decision tree modeling. It operates on the leaf distributions by taking into account undesired variability within languages, similarly as in the eigen-channel compensation of acoustic systems [6].

3.2.4. 3-gram lattice counts as super-vectors to SVM: `HU_SVM-3gram counts`

In this subsystem, the trigram-lattice-counts from Hungarian phone recognizer were used as a super-vectors for subsequent classification by SVMs, similarly as in MIT's work [4].

3.3. Normalization and Calibration

All systems were first processed by linear back-end and then fused (or calibrated) using multi-class linear logistic regression [2]. Both linear back-end and fusion parameters were trained on Dev1 data. The FoCal Multi-class toolkit by Niko Brümmer was used for the pre-processing and fusion.

	GMM2048-MMI-chcf			EN_Tree_all		
	30	10	3	30	10	3
No back-end	5.75	9.45	18.44	9.02	14.21	24.37
LLR	3.49	7.90	17.65	3.96	10.83	22.97
LDA	2.88	7.42	16.94	3.85	10.55	22.58
LDA+LLR	2.41	7.02	16.90	3.54	10.69	22.66

Table 2: Effects of calibration.

Acoustic	30	10	3
GMM256-MMI	4.15	8.61	18.43
GMM256-MMI-chcf	3.73	9.81	20.98
GMM2048-eigchan	2.76	7.38	17.14
GMM512-SVM	3.80	8.77	20.14
Phonotactic	30	10	3
HU_LM	5.54	11.75	23.54
HU_TREE_A3E7M5S3G2_FA	4.52	10.35	23.66
HU_strLM (4-gram)	6.35	13.86	27.12
HU_LM-MMI (2-gram)	6.85	14.27	26.37
HU_SVM-3gram-counts	5.41	13.26	26.92
RU_LM	6.06	13.04	24.47
RU_Tree	6.31	12.99	24.51
EN_Tree	4.56	12.32	24.54
EN_strLM (4-gram)	5.83	14.62	27.24

Table 3: Performance of individual subsystems submitted to NIST LRE 2007.

4. Experiments

The attention is first given to the calibration and fusion, since it had most of impact in our post-evaluation analysis.

The effect of calibration is demonstrated on one acoustic and one phonotactic system, the best one from each category: GMM with 2048 Gaussian mixture model trained using MMI criterion on channel compensated features (`GMM2048-MMI-chcf`) and Binary decision trees on posterior weighted counts from English HMM phone recognizer trained on full data (`EN_Tree_all` – see below for the difference from `EN_Tree`). Three calibration schemes were compared: (1) Linear back-end (LDA), (2) Linear Logistic Regression (LLR) and (3) LDA followed by LLR. With proper calibration, it is possible to reduce the error by 60% for 30 second condition (Table 2).

Detailed report of the results of all subsystems that were part of our submission is given in Table 3. All systems were calibrated using LDA+LLR. For the submission, the fusion of systems calibrated by LDA+LLR was done by LLR, the results are in the first line of Table 5.

In the post-evaluation analysis, we have concentrated on the following three topics:

Improving the acoustic system: 2048 Gaussians, eigen-channel compensated features and MMI training produced the best performing acoustic subsystem `GMM2048-MMI-chcf` and also the best performing stand-alone subsystem in our post evaluation work with $100 \times C_{avg} = 2.41$ on 30s segments [8].

Full training data for decision tree: with the English phone recognizer based on LVCSR, we were able to process only 3 hours per language for the submission, resulting in $100 \times C_{avg} = 4.56$ (Table 3). By processing all 450 hours

	LLR fusion			LDA fusion		
calibrated on	30	10	3	30	10	3
Dev1-30+Dev1-10	2.01	4.74	14.20	1.71	5.21	16.39
Dev1 - duration dependent	1.94	4.87	13.84	2.02	5.58	16.06
Dev2 - duration dependent	1.61	4.61	14.24	2.38	5.32	18.73
Dev1+Dev2 - duration dependent	1.41	4.43	12.98	1.31	4.51	14.69

Table 4: Different fusions.

	100xCavg			Cllr avg			Cllr multiclass		
	30	10	3	30	10	3	30	10	3
Submitted (LLR fusion)	2.01	4.74	14.20	.075	.184	.761	.284	.663	2.357
New Calibration (LLR fusion)	1.41	4.43	12.98	.056	.166	.447	.212	.614	1.671
Post-evaluation system (LLR fusion)	1.30	4.12	12.53	.051	.156	.433	.191	.577	1.615
Best 3 systems (LDA fusion)	1.28	4.63	13.53	.053	.161	.459	.187	.605	1.718

Table 5: Recapitulation of results using LLR and LDA fusions.

of training data and re-training decision trees, we obtained 3.54 (the system is denoted `EN_Tree_all`) [6].

Fusion: LLR fusion trained on join set of 10 and 30 seconds from Dev1 set was used at the time of submission. The LDA fusion[1] produced approximately the same results (Table 4) — it was better for the 30 seconds condition, but LLR outperformed it for 10 and 3 seconds. Therefore, we experimented with duration-dependent fusion: for each 30s, 10s, 3s condition, different calibration and fusion was trained on development segments of matching duration. This turned out to perform better. There are three data sets on which we can train the fusion: Dev1, Dev2 and Dev1+Dev2. Dev2 data are closer to the evaluation data than Dev1, therefore the results are better for this set for LLR fusion. But if we use Dev1 and Dev2 sets together (2 times more training examples), both LLR and LDA fusions are better and produce similar results. The post-evaluation fusions are calibrated on Dev1+Dev2 set, and are duration-dependent.

Table 5 compares results of submitted and post-evaluation systems. The last line stands for a "light" system which contains only 3 best sub-systems: `GMM2048-MMI-chcf`, `EN_Tree_all` and `HU_TREE_A3E7M5S3G2_LFA`.

5. Conclusions

There are several statements we can make based on the above results: The first one is about back-ends and fusion. It is necessary to have lots of calibration data as closed as possible to the evaluation. This is at least as important as having good systems that are calibrated. In acoustic system it is beneficial to use a combination of all successful techniques: a lot of Gaussian components, channel compensation and Maximum Mutual Information training. In phonotactic system, we have confirmed, that the accuracy of the phone recognizer is crucial for good performance of LRE. Also, we found adaptation from UBM advantageous both for classical LM and tree-based approaches, and we investigated into intersession compensation in phonotactic models using factor analysis. The biggest challenge in LRE nowadays is the availability and quality of training data. Resources such as Fisher do not exist for most languages and we have to recur to collection of data from other sources. Initial experiments in acquisition of telephone data from broadcasts [12] showed promising results and we continue work in this direction.

[1] Note, that for LDA fusion, LLR was applied at the end to calibrate the whole system.

6. References

[1] N. Brümmer, "Spescom DataVoice NIST 2004 system description," in *Proc. NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, June 2004.

[2] N. Brümmer, et al.: Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006, In: IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 7, 2007, pp. 2072-2084, ISSN 1558-7916.

[3] L. Burget, et al.: Analysis of feature extraction and channel compensation in GMM speaker recognition system, In: IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 7, 2007, pp. 1979-1986, ISSN 1558-7916.

[4] W.M. Campbell, F. Richardson, and D.A. Reynolds: Language Recognition with Word Lattices and Support Vector Machines, in Proc. ICASSP 2007.

[5] The 2007 NIST Language recognition evaluation plan (LRE07), available from http://www.nist.gov/speech/tests/lre/2007/LRE07EvalPlan-v8b.pdf.

[6] O. Glembek et al.: Advances in phonotactic language recognition, accepted to Interspeech 2008, Brisbane, Australia.

[7] T. Hain, et al.: The AMI System for the Transcription of Speech in Meetings, In: Proc. ICASSP 2007, Hononulu, 2007, pp. 357-360.

[8] V. Hubeika et al.: Discriminative training and channel compensation for acoustic language recognition, accepted to Interspeech 2008, Brisbane, Australia.

[9] P. Kenny and P. Dumouchel, "Disentangling speaker and channel effects in speaker verification," in *Proc. ICASSP 2004*, Montreal, Canada, May 2004, vol. 1, pp. 47–40.

[10] J. Navratil: "Recent advances in phonotactic language recognition using binary-decision trees," in *Proc. ICSLP 2006*, Pittsburgh, PA, October 2006

[11] P. Matejka, et al.: Brno University of Technology System for NIST 2005 Language Recognition Evaluation, in Proc. Odyssey 2006, San Juan, Puerto Rico, USA, June 2006.

[12] O. Plchot et al.: Acquisition of Telephone Data from Radio Broadcasts with Applications to Language Recognition, accepted to TSD 2008, Brno, Czech Republic.