BUT system for NIST 2008 speaker recognition evaluation

Lukáš Burget, Michal Fapšo, Valiantsina Hubeika, Ondřej Glembek, Martin Karafiát,
Marcel Kockmann, Pavel Matějka, Petr Schwarz and Jan "Honza" Černocký

Speech@FIT, Brno University of Technology, Czech Republic
{burget|ifapso|ihubeika|glembek|karafiat|kockmann|matejkap|schwarzp|cernocky}@fit.vutbr.cz

Abstract

This paper presents BUT system submitted to NIST 2008 SRE. It includes two subsystems based on Joint Factor Analysis (JFA) GMM/UBM and one based on SVM-GMM. The systems were developed on NIST SRE 2006 data, and the results are presented on NIST SRE 2008 evaluation data. We concentrate on the influence of side information in the calibration.

Index Terms: speaker recognition, joint factor analysis, NIST SRE 2008.

1. Introduction

The goal of this paper is to present a consolidated version of BUT system description with results obtained on SRE 2006 and 2008 data, and to discuss performances of individual systems as well as their fusion.

BUT submitted three systems to NIST SRE 2008 evaluations, only to the short2-short3 condition. Our primary submission was a fusion of three subsystems: (1) Gender-dependent Factor Analysis system with MFCC20⇒60 features and gender-dependent zt-norm. (2) Gender-independent JFA system with MFCC13⇒39 features and gender-dependent zt-norm. (3) SVM-CMLLR-MLLR system with gender-independent zt-norm. The first contrastive systems differed only in calibration and the second contrastive system was a simplified version of the primary one (no ASR use). In this paper, we will deal only with the primary system.

2. Data

The **training data** for UBMs, JFA and calibration is described below in respective sections.

The **development data** was based on NIST SRE 2006 evaluation data, especially the 1conv4w-1conv4w condition (phn-phn) which was the core condition in 2006 evaluation. The sets for other conditions (phn-mic, mic-phn, mic-mic) were defined by MIT-LL. The numbers of target trials and non-target trials are:

phn phn	. . . T=3618	N=52041
phn mic	. . . T=2518	N=21204
mic phn	. . . T=2534	N=20937
mic mic	. . . T=5064	N=146111,

where phn is the label for telephone segment and mic is the label for telephone conversation recorded through microphone. This same data was used to train fusion and calibration. However, our cross-validation experiments on two halves of the development set with non-overlapping speakers showed that the development results can be considered realistic.

The **evaluation data** was the official SRE 2008 evaluation data[1], with the following conditions:

1. only interview speech in train/test T=11508 N=22609
2. interview speech from the same microphone type in training and test T=583 N=1144
3. interview speech from different microphones types in training and test T=10925 N=21465
4. interview training speech and telephone test speech T=1101 N=10620
5. telephone training speech and non-interview microphone test speech T=1250 N=6132
6. only telephone speech in train/test T=2668 N=33152
7. only English telephone speech in training and test T=1226 N=16509
8. only English telephone speech spoken by a native U.S. English speaker in training and test T=607 N=7877

3. Feature extraction, segmentation

Two types of features were used, both derived with classical analysis window of 20 ms with shift of 10 ms:

Short time gaussianized MFCC 19 + energy augmented with their delta and double delta coefficients, making 60 dimensional feature vector. The system making use of these features is denoted **MFCC20⇒60**.

Short time gaussianized MFCC 12 + C0 augmented with their delta, double delta and triple delta coefficients. The dimensionality of the resulting features is reduced from 52 to 39 using HLDA. HLDA classes correspond to UBM Gaussians. System with these features is denoted **MFCC13⇒39**. These features were already used in our 2006 system [3]. Short-time gaussianization in both cases uses window of 300 frames (3 sec).

Speech/silence segmentation is performed by our Hungarian phone recognizer [1, 3], where all phoneme classes are linked to 'speech' class. Several heuristics based on short-term energy are used for two-channel telephone data to eliminate cross-talks [3]. For interview speech, the phone recognizer failed on the original data. Therefore, a Wiener filter[2] was applied and new

This work was partly supported by European projects MOBIO (FP7-214324) and AMIDA (FP6-033812), by Grant Agency of Czech Republic project No. 102/08/0707, and by Czech Ministry of Education project No. MSM0021630528. We would like to thank MIT-LL team for creating the sets with the microphone conditions, Niko Brummer and Albert Strasheim for valuable advices, Patrick Kenny for advices and his excellent paper [5] and Tomáš Kašpárek for managing the BUT "computing beast" with more then 500 CPUs and 7 disk servers.

[1] http://www.nist.gov/speech/tests/sre/2008/sre08_evalplan_release4.pdf

[2] http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=7673

phone strings were generated. All phoneme classes were linked to 'speech' class and no further post-processing was done. After that, we took ASR transcripts of the interviewer and, based on time-stamps provided by NIST, we removed his/her speech segments from our segmentation files, as we are interested in the interviewee speech. Note, that Wiener filtered signals were used only in the segmentation, the rest of feature extraction processed the original signals.

4. Joint factor analysis systems

4.1. FA-MFCC20⇒60 system

4.1.1. Universal background models

Two universal background models (UBMs) are trained on Switchboard II Phases 2 and 3, Switchboard Cellular Parts 1 and 2, and NIST SRE 2004 and 2005 telephone data. In total, there were 16307 recordings (574 hours) from 1307 female speakers and 13229 recordings (442 hours) from 1011 male speakers. Two gender-dependent UBMs with 2048 Gaussians were trained. We used 20 iterations of EM algorithm for up to 256 Gaussians and 25 iterations for 512 and more. No variance flooring was used.

4.1.2. Joint factor analysis

The Joint factor analysis (JFA) system closely follows the description of "Large Factor Analysis model" in Patrick Kenny's paper [5], with the speaker model represented by mean super-vector: $\mathbf{M} = \mathbf{m} + \mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{z} + \mathbf{U}\mathbf{x}$, where $\mathbf{m}$ is speaker-independent mean super-vector, $\mathbf{U}$ is a subspace with high intersession/channel variability (eigenchannels), $\mathbf{V}$ is a subspace with high speaker variability (eigenvoices) and $\mathbf{D}$ is a diagonal matrix describing remaining speaker variability not covered by $\mathbf{V}$.

The two gender-dependent UBMs are used to collect zero and first order statistic for training two gender-dependent JFA systems. The mean $\mathbf{m}$ of JFA equation was set to the UBM mean and, in contrary to [5], it was never re-trained. The super-vector of variances (diagonal of $\mathbf{\Sigma}$ from [5]) is also set to UBM values and not re-trained in the training of JFA.

First, for each JFA system, 300 eigenvoices are trained on the same data as UBM, although only speakers with more than 8 recordings were considered here. For the estimated eigenvoices, MAP estimates of speaker factors are obtained and fixed for the following training of eigenchannels. A set of 100 eigenchannels is trained on NIST SRE 2004 and 2005 telephone data (5029 and 4187 recordings of 376 females and 294 males speaker respectively). Another set of 100 eigenchannels is trained on SRE 2005 auxiliary microphone data (1619 and 1322 recordings of 52 females and 45 males speaker respectively). Both sets are concatenated. In contrary to Kenny's paper [5], the diagonal matrix describing the remaining speaker super-vector variability (matrix $\mathbf{D}$ in JFA equation) is estimated on top of eigenvoices and eigenchannels. A disjoint set of NIST SRE 2004 speakers with less than 8 recordings (277 and 82 recordings of 44 females and 13 males speaker respectively) is used for this purpose and MAP estimates of speaker and channel factors are fixed for estimating the diagonal matrix. To obtain speaker models, MAP estimates of all the factors are estimated on enrollment segments using Gauss-Seidel-like iterative method [6]. Unlike Kenny [5], we use only MAP estimates (not posterior distribution) of channel factors and standard 10-best Expected Log Likelihood Ratio for scoring.

4.1.3. Normalization

Scores are normalized using zt-norm. We used 221 females and 149 males z-norm segments, 200 females and 159 males t-norm models, together 729 segments derived each from one speaker of NIST SRE 2004 and 2005 data. Experiments have shown that in contrary to simple eigenchannel adaptation where we obtained only small improvement from zt-norm, for JFA system, gender-dependent zt-norm is crucial for good performance.

4.2. FA-MFCC13⇒39 system

The second JFA system is similar to the previous one, with the following differences: (1) MFCC13⇒39 features are used, (2) UBM is gender-independent with 2048 Gaussians. We used 10 iterations of EM algorithm for each splitting. (3) In JFA, a single gender-independent model is used.

5. SVM CMLLR-MLLR system

In this system, the coefficients from constrained maximum likelihood linear regression (CMLLR) and maximum likelihood linear regression (MLLR) transforms estimated in an automatic speech recognition (ASR) system are classified by SVMs.

5.1. Segmentation and recognition

In this system, we used the time information from ASR transcripts provided by NIST. Because of time shift of phncall-mic data, forced alignment was done to find out correct timing of the words.

The ASR features are PLP with C0, delta coefficients up to third order, cepstral mean and variance normalization, and HLDA (dimensionality reduction from 52 to 39).

The core of AMI system submitted to NIST RT 2005 [7] was used in MLLR/CMLLR work. However, the models were re-trained on Fisher database using Minimum Phone Error rate criterion. Because of lack of time, we did not generate our own ASR transcriptions, but used the ASR output provided by NIST. Since NIST did not provide pronunciation dictionary, we used the AMI dictionary and generated the missing pronunciations using a grapheme-to-phoneme system with automatically trained rules. With this, we were able to generate the triphone alignment and to apply VTLN.

CMLLR and MLLR transforms are trained for each speaker. At first, CMLLR is trained with two classes (speech + silence). On the top of it, MLLR with three classes (2 speech classes obtained by automatic clustering on the ASR training data + silence) is estimated.

5.2. SVM and normalization

The transform matrices from CMLLR speech classes ($39 \times 39 \times 1 + 39$) and MLLR ($39 \times 39 \times 2 + 2 \times 39$) are concatenated to one super-vector with 4680 features. Rank normalization is applied.

The SVM used to classify super-vectors uses linear kernel. It is trained on one positive example from the target speaker. The negative examples are taken from NIST 2004 data and microphone data from NIST 2005. In the testing, the trial is scored by the respective SVM. The SVM training and scoring was built with LibSVM library[3].

zt-norm normalization was applied on the scores. The same

[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm

selection of speakers as for our JFA-systems was used (section 4.1.3) but the normalization was gender-independent.

6. Calibration and fusion

Each system was calibrated at first with the channel side-information, then with language side-information. Such calibrated sub-systems were fused by linear logistic regression (LLR).

Side **channel** information for each trial is given by its hard assignment provided by NIST: phn-phn, phn-mic, mic-phn, mic-mic. The **language** information is the English/non-English decision given by our phonotactic LID system[4]. Hard decisions (not language posteriors) were used as side-information. The side information is used as follows:

1. For each system:
   (a) Split trials by channel condition and calibrate scores using linear logistic regression (LLR) in each split separately
   (b) Split trials according to English/non-English decision and calibrate scores using LLR in each split separately
2. Fuse the calibrated scores of all subsystems using LLR without making use of any side information

For convenience, FoCal Bilinear toolkit by Niko Brümmer[5] was used, although we did not make use of its extensions over standard LLR.

We have seen that the use of side-information is helpful – it actually allowed to use the same unchanged subsystems for all the channels. Additional improvement could be further obtained by relying on language information provided by NIST instead of more realistic LID system.

7. Results

The importance of side-info based calibration and fusion is shown in Figure 1 (tel-tel condition). On 2006 data, the performances using NIST-provided language labels and LID are almost the same, on 2008 we see slight deterioration using a real LID system. It is however without any doubt that both channel- and language-conditioning substantially improve the system. Figure 2 (mic-mic) shows that the use of side information is beneficial also for other conditions.

Figure 3 compares different systems on the tel-tel condition on both 2006 and 2008 data. It shows clearly that the single FA-MFCC20⇒60 system performs almost as well as the fusion. On mic-mic condition (Figure 4), we see that FA-MFCC20⇒60 is outperformed by FA-MFCC13⇒39 and that the fusion is beneficial. We explain this by the fact that FA-MFCC20⇒60 has 3×more parameters than FA-MFCC13⇒39 and is possibly over-trained to telephone data primarily used for FA model training.

Tables 1 and 2 contain complete results for both the development and evaluation data.

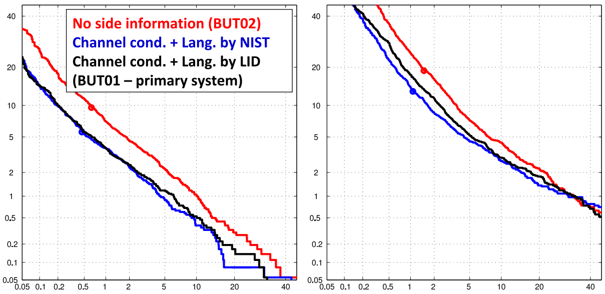

Figure 1: Side-info based calibration and fusion, tel-tel trials. Left panel: SRE 2006 (all trials, det1), right panel: SRE 2008 (all trials, det6).

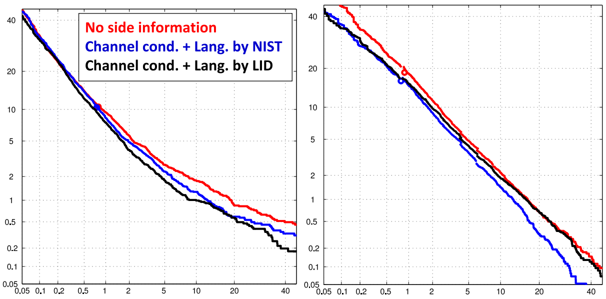

Figure 2: Side-info based calibration and fusion, mic-mic trials. Left panel: SRE 2006 (trial list defined by MIT), right panel: SRE 2008 (det1).

8. Conclusions

JFA systems built according to recipe from [5] perform excellently. From Fig. 3, it is obvious that it was hard to find another complementary system that would contribute to fusion of our two JFA systems. Especially for the phn-phn condition, a single JFA system is as good as system combination. We have however seen more improvement from the SVM CMLLR-MLLR system for other conditions. The dominance of JFA was also the reason why other techniques investigated for the NIST 2008 SRE at BUT (FA modeling prosodic and cepstral contours, SVM on phonotactics, etc.) did not make it to our final submission.

Although our system was primarily trained on and tuned for telephone data, JFA subsystems can be simply augmented with eigenchannels trained on microphone data (as also proposed in [5]), which makes the system performing well also on microphone conditions.

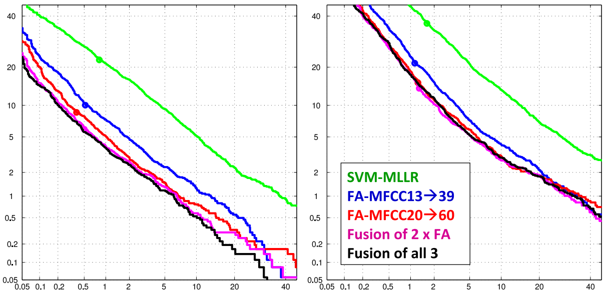

Figure 3: Subsystems and fusion - tel-tel trials. Left panel: SRE 2006 (all trials, det1), right panel: SRE 2008 (all trials, det6).

Another significant improvement was obtained by training

[4] based only on strings, see [2] and our web-demo `http://speech.fit.vutbr.cz/lid-demo/`

[5] `http://niko.brummer.googlepages.com/focalbilinear`

system	det1 - all trials				det3 - English only			
	phn-phn	phn-mic	mic-phn	mic-mic	phn-phn	phn-mic	mic-phn	mic-mic
FA-MFCC20⇒60	1.34 2.60	1.27 2.86	1.71 4.03	2.89 5.20	0.78 1.50	1.23 2.72	1.61 3.92	2.85 5.21
FA-MFCC13⇒39	1.79 3.59	1.31 3.18	1.69 4.85	2.05 4.17	0.81 1.74	1.25 3.05	1.56 4.04	2.00 4.18
SVM CMLLR2-MLLR3	3.38 7.66	2.08 5.36	2.62 7.17	3.88 8.04	1.60 3.34	1.94 4.87	2.36 6.16	3.75 7.74
Fusion of 2	1.09 2.30	0.95 2.26	1.33 3.35	1.74 3.06	0.64 1.26	0.92 2.10	1.22 2.94	1.70 3.02
Fusion of 3	1.05 2.24	0.75 1.75	1.08 3.03	1.68 2.98	0.55 1.08	0.71 1.57	0.98 2.64	1.65 2.89

Table 1: Summary of results on 2006 data. For each system, the first line contains 100×DCF, the second line EER in [%].

system-metric	det1	det2	det3	det4	det5	**det6**	det7	det8
FA-MFCC20⇒60	4.01 8.11	1.00 2.73	3.97 8.00	3.00 7.50	3.09 7.13	2.95 5.71	1.40 2.85	1.38 2.79
FA-MFCC13⇒39	2.55 4.70	0.34 1.21	2.62 4.78	2.85 6.97	2.54 6.10	3.68 6.54	1.38 2.68	1.28 2.46
SVM CMLLR2-MLLR3	4.78 11.31	1.72 5.15	4.80 11.33	4.84 11.44	3.66 9.92	5.77 12.18	2.59 6.69	2.80 6.94
Fusion of 2	2.75 5.35	0.38 1.04	2.76 5.44	2.69 6.06	2.19 5.37	2.67 5.11	1.12 2.52	1.13 2.30
Fusion of 3	2.43 4.67	0.38 1.39	2.47 4.72	2.12 5.16	2.01 4.89	2.72 5.14	1.04 2.28	1.05 2.14

Table 2: Summary of results on 2008 data. For each system, the first line contains 100×DCF, the second line EER in [%].

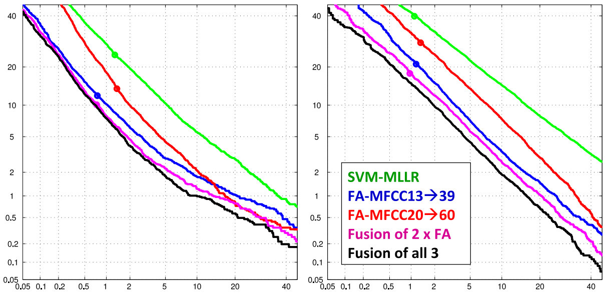

Figure 4: Subsystems and fusion - mic-mic trials. Left panel: SRE 2006 (trial list defined by MIT), right panel: SRE 2008 (det1).

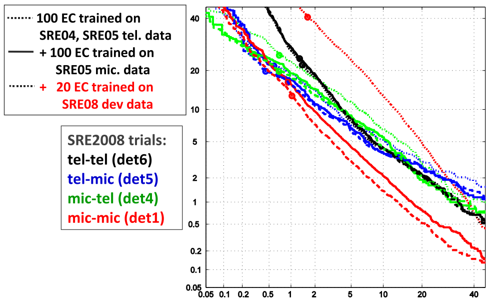

Figure 5: Training additional eigenchannels on SRE08 data. FA-MFCC13⇒39 system.

additional eigenchannels on data with matching channel condition, even thought there was very limited amount of such data provided by NIST (Figure 5).

9. References

[1] P. Schwarz, P. Matějka and J. Černocký: Hierarchical Structures of Neural Networks for Phoneme Recognition, In Proceedings of ICASSP 2006, May 2006, Toulouse, France

[2] P. Matějka, L. Burget, P. Schwarz and J. Černocký: Brno University of Technology System for NIST 2005 Language Recognition Evaluation. Odyssey: The Speaker and Language Recognition Workshop, San Juan, Puerto Rico, June 2006.

[3] L. Burget, P. Matějka, P. Schwarz, O. Glembek and J. Černocký: 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.

[4] 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.

[5] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel: A Study of Inter-Speaker Variability in Speaker Verification, IEEE Transactions on Audio, Speech and Language Processing, July 2008.

[6] R. Vogt, and S. Sridharan: Explicit Modelling of Session Variability for Speaker Verification. Computer Speech & Language 22(1), 2008, pp. 17-38.

[7] T. Hain et al.: The 2005 AMI system for the transcription of speech in meetings, in Proc. Rich Transcription 2005 Spring Meeting Recognition Evaluation Workshop, Edinburgh, July 2005.