BUT system description: NIST SRE 2008

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

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

1. Introduction

BUT submitted three systems to NIST SRE 2008 evaluations, only to the short2-short3 condition. The primary system is a fusion of three sub-systems: 2 based on MFCC and factor analysis and one making use of SVM scoring of CMLLR and MLLR matrices of an ASR system. The first contrastive systems differs only in callibration and the second contrastive system is a simplified version of the primary one (no ASR use).

2. Feature extraction, segmentation

We used two types of features:

- **MFCC19:** Short time gaussianized MFCC 19 + energy augmented with their delta an double delta coefficients, making 60 dimensional feature vector. The analysis window has 20 ms with shift of 10 ms.

- **MFCC12:** 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 corresponded to UBM Gaussians.

Short-time gaussianization in both cases uses window of 300 frames (3 sec). For the first frame, only 150 frames on the right are used and the window is growing till 300 while we move in time. When we approach the last frame, we use only 150 frames on the left side.

Speech/silence segmentation is performed by our Hungarian phoneme recognizer [1, 2], where all phoneme classes are linked to ’speech’ class. Segments labeled ’speech’ or ’silence’ are generated, but not merged yet to preserve smaller segments — a post-processing with two rules based on short time energy is applied first:

1. If the average energy in ’speech’ segment is 30 dB less than the maximum energy of the utterance, the segment is labeled as silence.

2. If the energy in the other channel is greater than maximum energy minus 3 dB in the processed channel, the segment is also labeled as silence.

After this post-processing, the resulting segments are merged together. Segments shorter than 20 frames are marked as silence. Only speech segments are used. In case of 1-channel files, rule #2 is not applied.

For the **interview data**, the processing described above resulted in very small amount of speech, mainly to complete failure of our phoneme recognizer. Therefore, they were segmented in a different way: first, a Wiener filter [1] was applied and new phoneme 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 removed his/her speech segments from our segmentation files based on time-stamps provided by NIST. Note, that Wiener filtered signals were used only in the segmentation, in the rest of feature extraction, original signals were used.

3. Factor analysis based sub-subsystem No. 1

3.1. Universal background models

Thwo 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 with 2048 Gaussians trained on MFCC19. We used 20 iterations of EM algorithm for each we do splitting up to 256 Gaussians and 25 iterations for 512 and up. No variance flooring was used.

3.2. Factor analysis – details

The Factor analysis (FA) system closely follows the description of “Large Factor Analysis model” in Patrick Kenny’s paper [3] with MFCC19 features. The two gender dependent UBMs are used to collect zero and first order statistic for training two gender dependent FA systems.

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

First, for each FA system, 300 eigenvoices[2] 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. On contrary to Kenny's paper [3], the diagonal matrix describing the remaining speaker super-vector variability (matrix $\mathbf{d}$ in [3]) is estimated on top of eigenvoices and eigenchannels. A disjunct 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 [4]. Unlike Kenny [3], we use only MAP estimates (not posterior distribution) of channel factors and standard 10-best Expected Log Likelihood Ratio for scoring.

3.3. Normalization

Finally, 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.

4. Factor analysis based sub-subsystem No. 2

The second FA system is similar to the previous one, with the following differences:

1. MFCC12 features are used.
2. Gender-independent UBM with 2048 Gaussians trained on MFCC12 is trained. We used 10 iterations of EM algorithm for each splitting.
3. the Factor Analysis is also gender independent
4. the ztnorm is gender dependent!

5. SVM CMLLR-MLLR

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

In this system, we used the time informataion 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.

5.2. Recognition system

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

The core of AMI system submitted to NIST RT 2005 [6] was used in MLLR/CMMLR 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 G2P 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.

5.3. SVM

The transform matrices from CMLLR speech classes ($39\times39\times1+39$) and MLLR ($39\times39\times2+2\times39$) are concatenated to one super-vector with 4680 features. The 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 [5] library.

5.4. Normalization

zt-norm normalization was applied on the scores. The same selection of speakers as for our FA-systems was used (section 3.3) but the normalization was gender-independent.

[2] We refer to "eigenvoices" and "eigenchannels" following the terminology defined in [3] although these sub-spaces are estimated using EM-algorithm, not PCA.

6. Calibration and fusion

We used FOCAL toolkit[3] for LLR side-information-conditional calibration and fusion. The output scores can be interpreted as detection log-likelihood-ratios. The hard decision was made by using the Bayes threshold 2.29.

We used two kinds of side information in calibration in the following order:

6.1. Channel type conditioning

First, we calibrated the subsystems with side information about channel provided by NIST which categorized each trial into one of four classes: phonecall/phonecall, mic/phonecall, phonecall/mic, mic/mic.

6.2. Language type conditioning

On the top of channel-type conditioned calibration, another calibration is done with the side information on the language of training and test segments in a trial: eng/eng, eng/noneng, noneng/eng and noneng/noneng. The language was automatically detected by our phonotactic LID system (based only on strings, see [2] and a our web-demo[4]). Hard decisions (not language posteriors) were used as side-information.

6.3. Fusion

Log Likelihood Regression fusion is used on the top of calibrated systems.

7. Submitted systems

Our primary system is a fusion of three subsystems:

- Gender dependent Factor Analysis system with MFCC19 features and gender dependent zt-norm.
- Gender independent Factor Analysis system with MFCC12 features and gender dependent zt-norm.
- SVM-CMLLR-MLLR system with gender independent zt-norm.

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

7.1. First contrastive system

The same system as the primary one but without the channel and language conditioning in the calibration stage.

7.2. Second contrastive system

The same system as the primary one but without the SVM-CMLLR-MLLR subsystem.

8. Results on development set

This section reports the results on our development set extracted from NIST SRE 2006 data. The 1conv4w-1conv4w condition was the core condition in 2006 evaluation defined by NIST. Other conditions were defined by MIT. The numbers of trials and non-trials are:

- phn phn . . . Targets=3618 Nontargets=52041
- phn mic . . . Targets=2518 Nontargets=21204
- mic phn . . . Targets=2534 Nontargets=20937
- mic mic . . . Targets=5064 Nontargets=146111

Where the phn is the label for telephone segment and mic is the label for telephone conversation recorded through microphone. The results of our three submitted systems are reported in Table 1. These results use trained fusion and calibration on the same data. However, our cross-validation experiments on two halves of the development set with non-overlapping speakers showed the realistic results.

9. Speed and resources

Real time factors are estimated on a standard PC Intel Xeon 3,2GHz or similar.

For each FA-based system, the real time factor is approximately 0.5×RT.

For SVM CMLLR-MLLR system, the real time factor is approximately 2.5×RT. Memory requirement for testing is 2 GB.

10. Acknowledgments

We would like to thank to:

- The MIT for creating the sets with the microphone conditions.
- Niko Brummer and Albert Strasheim for valuable advices.
- Patrick Kenny for advices and his excellent paper “A Study of Inter-Speaker Variability in Speaker Verification” [3].
- Tomáš Kašpárek for managing the BUT “computing beast” with more then 500 CPUs and 7 disk servers.

This work was partly supported by European projects AMIDA (FP6-033812) and MOBIO (FP7-214324), by

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

[4] `http://speech.fit.vutbr.cz/lid-demo/`

Primary system					
	phn-phn	phn-mic	mic-phn	mic-mic	all
DCF	0.0105	0.0075	0.0108	0.0168	0.0128
EER [%]	2.24	1.75	3.03	2.98	2.59
1st contrastive					
DCF	0.0157	0.0103	0.0138	0.0183	0.0232
EER [%]	3.27	2.70	4.14	3.64	4.73
2nd contrastive					
DCF	0.0109	0.0095	0.0133	0.0174	0.0137
EER [%]	2.30	2.26	3.35	3.06	2.82

Table 1: Results on our development set

Grant Agency of Czech Republic under project No. 102/08/0707, by Czech Ministry of Education under project No. MSM0021630528 and by Czech Ministry of Defense. The hardware used in this work was partially provided by CESNET under project No. 201/2006.

11. References

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

[2] Matějka P., Burget L., Schwarz P. and Černocký J., 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] Kenny, P., Ouellet, P., Dehak, N., Gupta, V., and Dumouchel, P.: “A Study of Inter-Speaker Variability in Speaker Verification” , IEEE Transactions on Audio, Speech and Language Processing, July 2008.

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

[5] Chih-Chung Chang and Chih-Jen Lin, LIBSVM: a library for support vector machines, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[6] 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.