TUL NIST 2010 SRE System Description

Jan SILOVSKY

SpeechLab, Faculty of Mechatronics, Informatics and Interdisciplinary Studies, Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic

jan.silovsky@tul.cz

1. Introduction

Technical University of Liberec (TUL) submitted three systems to the NIST SRE 2010. All systems were applied only to the core test condition. The primary system is a JFA system. The first contrastive system is a UBM-GMM system with eigenchannel adaptation and the second contrastive system is an i-vectors based system. All systems are gender-dependent.

2. Common processing

2.1 Feature extraction and segmentation

All systems used the same type of short-term acoustic features. 19 Mel-frequency cepstral coefficients + c0 were extracted using 25 ms windows with shift of 10 ms, short-time gaussianized using window of 300 frames (3 s) and augmented with their first derivatives forming a 40-dimensional feature vector. The resulting vectors were mean normalized over the whole utterance.

For segmentation, we used the time information from ASR transcripts provided by NIST with some merging and padding of speech segments. Energy based detector was subsequently used for the telephone data to label silence regions determined by a high energy drop.

2.2 UBMs

All systems need a UBM for initial processing. Gender dependent UBMs were trained on both telephone and microphone data from previous NIST evaluations, ranging from SRE04 to SRE08. In total, 22,872 recordings (1,931 hours) from female speakers and 16,899 recordings (1,432 hours) from male speakers were used in training. UBMs with 1024 Gaussian components were trained using the EM algorithm with binary splitting and using 20 iterations for all models' sizes.

2.3 Sufficient statistics

Zero-, first- and second[1]-order sufficient statistics were computed and stored for all development data, training and test segments using the gender-specific UBMs. In all further processing, no information about speech signal other than these statistics is required. Estimation of system hyper-parameters, training of models and scoring of test segments was done using only these statistics.

3. JFA system (primary)

The JFA system is based on the joint factor analysis model introduced by Patrick Kenny [1, 2]. This model is based on the assumption that a recording can be represented by a speaker- and channel-dependent supervector $\boldsymbol{M}$ which can be decomposed as follows:

$$\boldsymbol{M} = \boldsymbol{m} + \boldsymbol{v}\boldsymbol{y} + \boldsymbol{d}\boldsymbol{z} + \boldsymbol{u}\boldsymbol{x}.$$

This system used UBMs with 1024 Gaussian components and hence supervectors $\boldsymbol{M}$ are 40960-dimensional.

3.1 Hyper-parameters estimation

The UBM's mean supervector was used as an estimation of the global mean supervector $\boldsymbol{m}$ and it was not re-estimated during the hyper-parameters estimation process. In estimation of all remaining hyper-parameters, we used only the data from SRE04, SRE05 and SRE06 which we will further refer to as to the background set. The data from SRE05 and SRE06 were considered as coming from one database because of the overlap of speakers and segments. We used decoupled estimation of the system hyper-parameters.

First, the eigenvoices space matrices $\boldsymbol{v}$ were estimated using both telephone and microphone data. For both channel-types, we used the data from those speakers for which at least 8 recordings are available per a given channel. The maximum number of recording used per speaker and channel-type was set to 32. In total, there were 9010 recordings from 593 male speakers and 11989 recordings from 819 female speakers used in estimation of 200 eigenvoices. We used seven iterations of maximum likelihood (ML) estimation and two iterations of minimum divergence (MD) estimation.

Next, eigenchannels space matrices $\boldsymbol{u}$ were estimated. More specifically, two channel-specific eigenchannels space matrices were estimated separately for telephone and microphone channel data and concatenated to form matrix $\boldsymbol{u}$. For both channel-types, a set of 100 eigenchannels was estimated using the data from speakers for which at least 8 recordings are available per a given channel. The maximum number of recordings used per speaker was set to 16 and 32 recordings for telephone and microphone data re-

[1] Please note that second-order statistics were extracted but not actually used, because we used the dot-product scoring in JFA system

spectively. The sets of recordings used to estimate the matrix $\boldsymbol{u}$ and matrix $\boldsymbol{v}$ shares most of recordings. In total, there were 6218 recordings from 490 male speakers and 8691 recordings from 704 female speakers used in telephone eigenchannels training and 2224 recordings from 82 male speakers and 2688 recordings from 98 female speakers used in microphone eigenchannels training. For the recordings, MAP point estimates of speaker factors were calculated using all recordings of the speaker and used within the eigenchannels training process to center the statistics. Again, we used seven iterations of ML estimation and two iterations of MD estimation.

Finally, the diagonal matrices $\boldsymbol{d}$ describing the remaining variability were estimated. Here, the data from speakers for which less than 8 recordings but at least 5 recordings are available were selected. Thus, here we used disjunct set of speakers and recordings compared to the previous sets. In total, there were 392 recordings from 68 male speakers and 528 recordings from 89 female speakers. For the recordings, decoupled estimation of speaker and channel factors was performed. First, MAP point estimate of speaker factors was calculated using all recordings of the speaker and then MAP point estimate of channel factors was calculated for each recording. Again, we used seven iterations of ML estimation and two iterations of MD estimation.

3.2 Scoring

We used dot-product linear scoring as described in [3] instead of the integration over the whole distribution of channel factors [4]. The score for the trial is given as:

$$LLR_{lin}(\boldsymbol{O}|\boldsymbol{s},\boldsymbol{x}) = (\boldsymbol{vy}+\boldsymbol{dz})^{*}\boldsymbol{\Sigma}^{-1}(\boldsymbol{F}-\boldsymbol{Nm}-\boldsymbol{Nux})$$

where $\boldsymbol{O}$ represents the sequence of feature vectors extracted from the test segment, $\boldsymbol{s}$ is the speaker-dependent supervector ($\boldsymbol{s} = \boldsymbol{m} + \boldsymbol{vy} + \boldsymbol{dz}$) estimated in the speaker's enrollment, $\boldsymbol{x}$ is a vector of channel factors estimated for the test segment using the UBM, $\sum$ is diagonal supercovariance matrix (we used concatenation of UBM's covariance matrices as its estimate) and finally $\boldsymbol{N}$ is zero-order sufficient statistics of the test segment.

3.3 Score normalization

Gender dependent ZT-norm normalization was applied on the scores obtained by the linear scoring. For female trials, we used 205 T-norm models (64 trained on microphone data and 141 trained on telephone data) and 328 Z-norm recordings (79 drawn from microphone data and 249 from telephone data). For male trials, we used 154 T-norm models (50 trained on microphone data and 104 trained on telephone data) and 261 Z-norm recordings (69 drawn from microphone data and 192 from telephone data). We made no use of information about the channel-type of T-norm models and Z-norm segments during the score normalization process.

4. UBM-GMM system (1st alternate)

The UBM-GMM system is based on standard relevance MAP adaptation [5]. Speaker models are derived from a UBM by MAP adaptation of UBM's means with relevance factor 16. The implementation of this system closely follows the Niko Brummer's description of the linearized eigenchannel GMM system [6]. Eigenchannel adaptation was applied for channel compensation in both training of models and scoring of test segments. On contrary to the other systems, this system uses UBMs with only 512 Gaussian components.

Like for the JFA system, two eigenchannels space matrices were estimated for telephone and microphone speech data separately and concatenated. The same data sets as described in section 3.1 were used for training of eigenchannels but here only 50 eigenchannels were trained for each channel-type.

The gender-dependent ZT-norm was applied using the same sets of T-norm models and Z-norm recordings as described in section 3.3.

5. I-Vectors system (2nd alternate)

The I-Vectors system is based on representation of a recording in a low-dimensional total variability space using so called i-vectors [7]. For scoring, the raw cosine kernel distance between the i-vector of an enrollment utterance and the i-vector of a test segment is used. This system uses sufficient statistics derived using UBMs with 1024 Gaussian components.

The total variability matrices were estimated using both telephone and microphone data. For both channel-types, we used the data from those speakers for which at least 4 recordings are available per a given channel. The maximum number of recordings used per speaker was set to 4 and 8 recordings for telephone and microphone data respectively. In total, there were 2800 recordings from 618 male speakers and 3916 recordings from 881 female speakers used in estimation of total variability matrices. We used 300 total factors.

The total variability space is supposed to contain both speaker and channel variability. Several techniques to remove channel effects from i-vectors are described in [7]. We used the combination of Linear Discriminant Analysis (LDA) followed by the Within Class Covariance Normalization (WCCN). The LDA and WCCN projection matrices were estimated using the same data as the total variability space matrices. The LDA projection reduces the dimension of i-vectors from 300 to 200.

For this system, S-norm normalization as described in [8] is applied instead of ZT-norm normalization. The S-norm cohort is formed from utterances used in Z-norm normalization for the other systems.

6. Calibration

Output scores produced by all systems can be interpreted as log-likelihood-ratios. We used our perl port of Focal toolkit[2] for LLR calibration. The hard decisions were made using the threshold value 6.9.

The calibration is performed in two stages. First, gender-dependent channel-type conditioned calibration is performed based on the channel-type of the model's enrolment segment and the channel-type of a test segment. In next stage, gender conditioning is applied. Knowledge of whether or not a segment involves telephone channel transmission as well as knowledge of the gender of a target speaker is determined by categorization of data as provided by NIST.

7. Processing time and hardware

The processing time is measured on a machine with Intel Core i7 920 CPU (@2.66GHz) and 3 GB RAM (DDR3@1.6GHz). Gathering of sufficient statistics was performed completely in a single-threaded way on one core, other operations were performed in Matlab and thus some of them run on two cores. Table 1 reports real-time factors of the processing times in some sub-tasks performed by the systems as well as the total times required to process the evaluation data.

Real-time factors	Training Models	Scoring Trials
Gathering sufficient statistics		
512 Gaussians	0.007	0.007
1024 Gaussians	0.012	0.012
JFA system		
raw training / scoring	0.004	0.003
+ ZT-norm	0.012	0.008
+ gathering statistics (total)	0.024	0.019
UBM-GMM system		
raw training / scoring	0.0008	0.0012
+ ZT-norm	0.001	0.002
+ gathering statistics (total)	0.008	0.009
I-Vectors system		
raw training / scoring + S-norm	0.003	0.003
+ gathering statistics (total)	0.015	0.015

Tab. 1. Real-time factors for systems.

8. Development results

We used the core condition of the NIST SRE08 for development experiments. Tables 2, 3 and 4 summarize results achieved in the subsets of the core test trials as defined by NIST SRE08 evaluation plan[3]. The JFA system performed best in all subsets of trials except for the det5 subset, which refers to the trials using telephone training speech and non-interview microphone test speech. We believe that this singularity is caused by the composition of Z- and T-norm sets, because for the other subsets of trials we observed improvement of results after application of ZT-norm normalization, but no such effect was observed for det5 subset. The other systems performed worse but the results were more balanced across all the subsets of trials.

	det1	det4	det5	det6	det7
JFA system					
EER [%]	5.48	7.65	10.94	6.55	3.55
UBM-GMM system					
EER [%]	7.66	8.40	8.42	8.16	4.43
I-Vectors system					
EER [%]	7.51	8.99	9.26	8.70	5.58

Tab. 2. Results on female part of the development evaluation set (NIST SRE08 data, core condition)

	det1	det4	det5	det6	det7
JFA system					
EER [%]	3.64	6.40	8.75	5.49	2.96
UBM-GMM system					
EER [%]	5.16	5.69	6.84	4.80	2.73
I-Vectors system					
EER [%]	6.43	5.29	7.19	7.55	6.38

Tab. 3. Results on male part of the development evaluation set (NIST SRE08 data, core condition)

	det1	det4	det5	det6	det7
JFA system					
EER [%]	4.74	7.14	9.85	6.19	3.18
C_{llr}[4]	0.183	0.267	0.371	0.249	0.144
UBM-GMM system					
EER [%]	6.62	7.24	7.68	6.98	3.90
C_{llr}	0.239	0.268	0.288	0.278	0.175
I-Vectors system					
EER [%]	7.06	7.58	8.23	8.37	5.78
C_{llr}	0.255	0.279	0.358	0.315	0.233

Tab. 4. Results on complete development evaluation set (NIST SRE08 data, core condition)

[2] http://www.dsp.sun.ac.za/~nbrummer/focal/

[3] http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf

[4] Please note that systems were both calibrated and evaluated on the development set

Acknowledgements

This work was supported by project of the Czech Ministry of the Interior (project no. VD20072010B160) and by the Czech Grant Agency (grant no. 102/08/0707).

References

[1] Kenny, P.: Joint factor analysis of speaker and session variability: Theory and algorithms - Technical report CRIM-06/08-13, Montreal, CRIM, 2005.

[2] 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.

[3] Glembek, O., Burget, L., Dehak, N., Brummer, N., and Kenny, P.: Comparison of Scoring Methods used in Speaker Recognition with Joint Factor Analysis. In Proc ICASSP 2009, Taipei, Taiwan, April 2009.

[4] Kenny, P., Boulianne, G., Ouellet, P. and Dumouchel, P.A..: Joint factor analysis versus eigenchannels in speaker recognition. IEEE Transactions on Audio, Speech and Language Processing 15 (4), pp. 1435-1447, May 2007.

[5] Reynolds, D., Quatieri, T. and Dunn, R.: Speaker verification using adapted Gaussian mixture models. Digital Signal Processing, vol. 10, pp. 19–41, 2000.

[6] Strasheim, A. and Brummer, N.: SUNSDV System Description: NIST SRE 2008. In Proc. NIST Speaker Recognition Evaluation 2008, Montreal, Canada, Jun. 2008.

[7] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P and Ouellet, P.: Front-End Factor Analysis for Speaker Verification. Submitted to IEEE Transactions on Audio, Speech and Language Processing, November 2009.

[8] Brummer, N. and Strasheim, A.: AGNITIO's Speaker Recognition System for EVALITA 2009. http://evalita.fbk.eu/reports/Speaker%20Identity%20Verification/Application/SIV_APPLICATION_AGNITIO.pdf