UWB system description for NIST SRE 2010

Lukáš Machlica, Jan Vaněk

Department of Cybernetics, University of West Bohemia, Pilsen
machlica@kky.zcu.cz; vanekyj@kky.zcu.cz

1. Introduction

We were focusing on the core test. One primary and two contrastive systems were submitted. Primary system was a fusion of all GMM-UBM, SVM-GSV and SVM-GLDS systems described in this paper. The first contrastive system was a fusion of GMM-UBM based systems described in Section 3.1, the next was a fusion of SVM based systems described in Section 3.2. All the submitted scores can be interpreted as log-likelihood ratios. Submitted score files are *UWB_1_core_core_primary_llr* for primary system, *UWB_2_core_core_alternate_llr* for 1^{st} contrastive system, and *UWB_3_core_core_alternate_llr* for 2^{nd} contrastive system.

2. Signal Processing and Feature Extraction

At first, Voice Activity Detector (VAD) was applied on all of the data in order to discard non-speech frames. VAD was based on detection of energies in filter-banks located in the frequency domain. Local SNRs were estimated for each frame as a mean value of SNRs in each of the filter-banks, and global SNR was represented as a mean value of local SNRs computed across whole utterance. Frames with local SNR lower than the global SNR were marked as non-speech.

Extracted features were based on Linear Frequency Cepstral Coefficients (LFCCs). 20 LFCCs were extracted each 10 ms utilizing a 25 ms hamming window, these were than processed in 2 different ways

- $LFCC$ – Δ coefficients were added, hence 40 dimensional final feature vectors were obtained.
- $TDDCT$ – instead of the Δ coefficients discrete cosine transformation in the time domain was performed. Features were weighed with a window W of a constant length (specified by $wlength$ in samples) centered around a frame of interest. The shape of the window can be expressed as in (1), where $P = 0, \dots, N$. In our case, features were weighed with windows $P = 0, 1, 2$, and results of weighing were concatenated leading to a 60 dimensional feature vector.

$$W(i) = cos\left(\frac{i}{wlength} \cdot P \cdot \pi\right),\ i = 1, \dots, wlength \quad (1)$$

At the end, feature warping was carried out, and the final set of feature vectors was downsampled with a factor of 2.

3. Involved Systems

Two types of systems were involved, GMM based [1] utilizing $TDDCT$ feature extraction and SVM based utilizing $LFCC$ feature extraction. Genders were handled separately. In order to increase the robustness of submitted systems several systems of each kind (GMM, SVM) were trained and fused. Rather than using all the background data at once, data were divided into smaller sets, multiple Universal Background Models (UBMs) or SVM impostor sets were created, thus multiple models of one speaker were trained. An alternative insight may be interpreted as modeling the background population using a huge background model, which parts cover distinct regions of the feature space. We assume that such an approach makes the system less vulnerable to varying environment conditions, hence more robust. Mentioned approach facilitates the computation/parallelization since each UBM is handled separately aside from other UBMs, and not all the SVM impostors are required to be processed at once.

3.1. GMM-UBM Systems

In common 18 UBMs (for each gender) were trained differing in background data and number of mixtures. We assume that lower (higher) amount of mixtures may be preferable in some environmental conditions (unknown in advance) as higher (lower) uncertainty is present in the model. Hence, rather to utilize one specific number of mixtures we trained several UBMs with distinct amount of mixtures. Following background sets were used: Switchboard cellular part1, SRE04, SRE05, SRE06, SRE08 telephone condition and SRE08 interview condition. For each background set 3 UBMs were created differing in number of mixtures - 256, 512, 1024. To train one UBM Maximum Likelihood (ML) estimation preceded by distance based algorithm (in order to initialize the ML procedure) was used. Speaker models were adapted utilizing MAP adaptation with relevance factor $\tau = 14$ preceded by one iteration of MLLR adaptation. Only means were adapted.

3.2. SVM Systems

In order to strengthen the orientation of the separating hyperplane each training feature vectors were divided into disjoint sets containing 1000 feature vectors, and each set was mapped to a supervector (SV) separately. All of the systems utilized Nuisance Attribute Projection (NAP) [2]. NAP was trained on SRE04, SRE05 and SRE06 data, eigenvectors related to 256 highest eigenvalues were used to create the NAP matrix. Impostors were chosen also from SRE04, SRE05 and SRE06 data, however for each impostor set (SRE04, SRE05 and SRE06) one particular speaker model was trained (rather than pool the data into one impostor set). Hence, at the end 3 speaker models instead of 1 were obtained. SVMs were trained with SVMTorch [3] and only linear kernels were involved.

3.2.1. SVM-GSV Systems

GMMs were adapted from an UBM containing 512 mixtures with a relevance factor $\tau = 5$. UBM was trained on SRE04, SRE05 dataset. GMM means were concatenated, hence a 20480 dimensional GMM supervector (GSV) [4] was obtained.

3.2.2. SVM-GLDS Systems

Polynomial order 3 was assumed when constructing the Generalized Linear Discriminant Sequence (GLDS) kernel [5], thus the dimension of SVs was 12341.

4. Score Normalization

Only results obtained from GMM-UBM systems were TNormalized. The pre-cohort consisted of approximately 600 gender and channel (telephone/interview) dependent models chosen from SRE08, and for final cohort 40 models were selected according to their log-likelihood with regard to the test file.

5. Fusion

In order to fuse the results the linear logistic regression from FoCal toolkit [6] was used. Fusion weights were computed for each type of channel transmission in training and test segment (phonecall-interview, interview-phonecall, interview-interview, phonecall-phonecall) on the SRE08 set.

6. Development Experiments

All the development experiments were performed on the core test of the SRE08 set. Following conditions were examined individually

- interview speech in training and test (int-int)
- telephone speech in training and test (phn-phn)
- telephone speech in training and interview speech in test (phn-int)
- interview speech in training and telephone speech in test (int-phn)

Development results are depicted in Figure 1 - Figure 3. Decision Cost Function (DCF) was computed according to values given in the SRE08 evaluation plan, hence $C_{Miss} = 10$, $C_{FalseAlarm} = 1$, $P_{Target} = 0.01$.

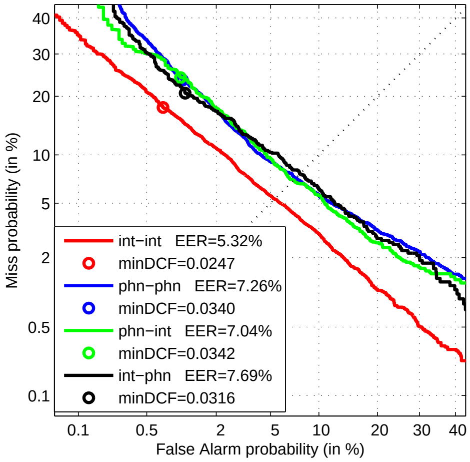

Figure 1: Primary system.

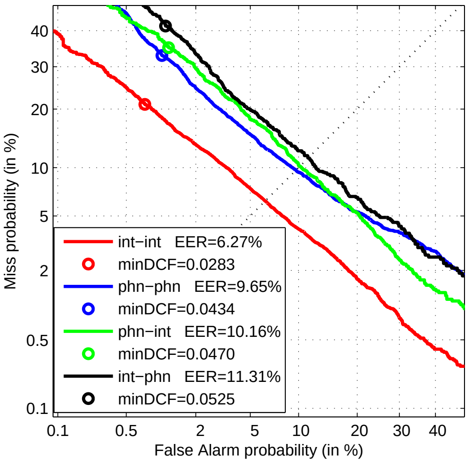

Figure 2: 1^{st} contrastive system.

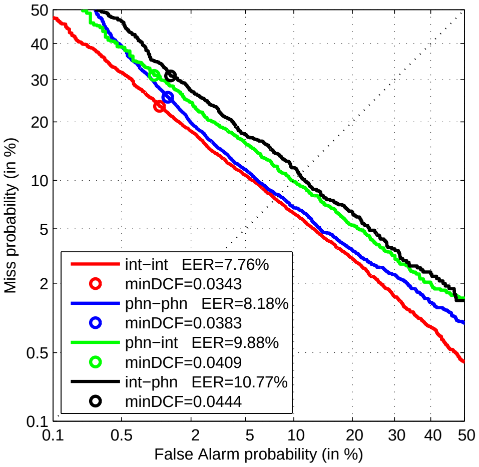

Figure 3: 2^{nd} contrastive system.

	enrollment [x RT]	**memory demands**	**verify** [x RT]
GMM-UBM	0.0072	2 MB	0.0019
SVM-GSV	0.1240	1.8 GB	6.6e-5
SVM-GLDS	0.0566	0.6 GB	3.7e-5
primary	0.5994	1.8 GB	0.0402
1^{st} contrastive	0.0792	2 MB	0.0399
2^{nd} contrastive	0.5238	1.8 GB	3.1e-4

Table 1: The CPU execution time that was required to process the evaluation data.

7. CPU Execution Time

CPU execution time and memory demands can be found in Table 1, they relate to a 2.39 GHz Intel 4 GB RAM PC. Only memory demands for speaker model training are mentioned, verification consumed at most 2 MB of memory.

8. Summary

In common 24 models for one speaker were trained (18*GMM-UBM + 3*SVM-GSV + 3*SVM-GLDS) as described in Section 3. GMM-UBM systems utilized $TDDCT$, and SVM used $LFCC$ feature extraction. Primary system involved all the speaker models, where outputs of individual systems were fused at the end. 1^{st} contrastive system utilized 21 GMM-UBM models and 2^{nd} contrastive system made use of all the NAP compensated SVM models, hence 3 SVM-GSV and 3 SVM-GLDS. Scores from GMM-UBM system were TNormalized.

9. Acknowledgements

This work was supported by the Grant Agency of the Czech Republic, project No. GAČR 102/08/0707, and by the Ministry of Education of the Czech Republic, project No. MŠMT LC536.

10. References

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1-3):19 – 41, 2000.

[2] A. Solomonoff, C. Quillen, and W.M. Campbell. Channel compensation for svm speaker recognition. *ODYSSEY*, pages 57–62, 2004.

[3] Ronan Collobert, Samy Bengio, and C. Williamson. Svmtorch: Support vector machines for large-scale regression problems. *Journal of Machine Learning Research*, 1:143–160, 2001.

[4] W.M. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff. Svm based speaker verification using a gmm supervector kernel and nap variability compensation. *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings*, 1:I–I, 2006.

[5] W.M. Campbell. Generalized linear discriminant sequence kernels for speaker recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing. ICASSP'02*, 1:I–161–I–164, 2002.

[6] N. Brummer. Focal: Tools for fusion and calibration of automatic speaker detection systems. 2006. Available at: http://sites.google.com/site/nikobrummer/focal.