UWB system description: EVALITA 2009

Lukáš Machlica and Jan Vaněk

University of West Bohemia, Pilsen, Univerzitní 8, Czech Republic,
{machlica, vanekyj}@kky.zcu.cz,
WWW home page: http://www.kky.zcu.cz/en

Abstract. The report describes two UWB systems submitted to the EVALITA 2009 evaluation campaign. Both systems are based on the UBM-GMM approach. Our main motivation laid in the investigation of complementarity of simple UBM-GMM systems in order to achieve a robust performance in different operating conditions, as proposed in the EVALITA 2009 contest. Results are presented on the development set as well as on the test set.

Key words: UBM-GMM, fusion, complementarity, speaker verification.

1 Introduction

We have submitted two systems to EVALITA 2009 evaluations. The primary system is a fusion of 8 UBM-GMM [1] based subsystems differing in parametrization, modelling and verification, as described in Section 2. The second system represents the subsystem (included in the primary system) with the best performance on the development set (see Section 4). Individual subsystems and their fusion were tuned on the development set, which layout description can be found in Section 3. The outputs of individual subsystems were combined as described in Section 4.1. At the end of Section 4, results obtained on the test set are given.

2 System Description

Altogether we proposed and tested 21 systems. However, finally we chose only 8 systems (according to the procedure described in Section 4.1), which tend to be most supplementary. In a sequel, only these 8 systems are going to be described because of lucidity. The rest of the systems were various combinations of parametrization, UBM, modelling and verification techniques presented in a sequel.

2.1 Parametrization

We have utilized several parametrizations based on MFCC or LFCC using 25 triangular band filters. First, a 25 ms hamming window was applied with a 10 ms window shift. Then, 20 cepstral coefficients excluding the zeroth were extracted. Several other improvements were further applied leading to 5 different parametrizations:

- $P_{LFCC}(\Delta, M, V)$, $P_{LFCC}(\Delta, \Delta\Delta, M)$, $P_{LFCC}(DCT2_{(3,9)}, M)$, $P_{MFCC}(\Delta, \Delta\Delta, M)$, $P_{MFCC}(DCT2_{(4,13)}, M, V)$,

where $\Delta, \Delta\Delta$ indicates that Δ and $\Delta\Delta$ coefficients were added; M, V represents the mean and variance normalization of the features performed as the last step in the feature extraction process. The $DCT2_{(P,wlength)}$ stands for a discrete cosine transformation in the time domain and it is used instead of the Δ coefficients. It consists in weighting of features 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

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

where $P = 1, \ldots, N$ is fixed for one window. Thus, when $P > 1$ features are weighted at first with a window with $P = 1$, then the features are weighted again with a window with $P = 2$ and so on. At the end, all such new features are added to the feature vector increasing its dimensionality to $dim(\mathbf{o}) + P \cdot dim(\mathbf{o})$, where $dim(\mathbf{o})$ is the dimensionality of unextended feature vectors - in our case $dim(\mathbf{o}) = 20$ (hence, the $DCT2_{(3,9)}$ leads to a feature vector with dimension 80). At the end, the features were downsampled with a factor of 2. Also a voice activity detector (VAD) was used to discard the non-speech frames.

2.2 Modelling

Ordinary maximum likelihood (ML) training was used to estimate the UBM parameters. Gender dependent UBMs with 256, 512 and 1024 mixtures were constructed. All the other models were adapted utilizing maximum likelihood linear regression (MLLR) [2] and maximum a-posteriori probability (MAP) adaptation. Hence, several techniques were used to estimate speaker model parameters. These techniques are denoted as

- $M(MAP(\mu))$, $M(MLLR(\mu), MAP(\mu))$, $M(MLLR(\mu), MAP(\mu, \sigma))$,

where μ, σ in the brackets refers to mean, eventually variance adaptation. We have utilized only one (global) MLLR matrix common for all the model means. $M(MLLR, MAP)$ stands for an adaptation, where MLLR is performed prior to MAP adaptation in order to refine the adaptation statistics. We found it quite useful mainly in situations, where only a few speaker data are provided for training [3]. In the case of MAP adaptation a relevance factor $\tau = 15$ was used.

2.3 Verification

Majority voting rule (MVR) was used for verification [4]. Denote $\mathbf{O} = \{\mathbf{o}_1, \ldots, \mathbf{o}_T\}$ a set of feature vectors and $L(\mathbf{o}_i|s)$, $L(\mathbf{o}_i|UBM)$ the log likelihood of $\mathbf{o}_i$ in the

$s-th$ speaker model and UBM model, respectively. The final decision D_s for the $s-th$ speaker has the form

$$D_s = 1/T \sum_{i=1}^{T} D_i, \;\; D_i = \begin{cases} 1 \text{ if } L(\mathbf{o}_i|s) > L(\mathbf{o}_i|UBM) \\ 0 \text{ otherwise} \end{cases} . \tag{2}$$

We have also utilized a slightly different version of MVR denoted as SMVR, where a sigmoid function is involved in the computation of D_i in (2), thus

$$D_i = \frac{1}{1 + exp\{\frac{-1.8}{\lambda}(L(\mathbf{o}_i|s) - L(\mathbf{o}_i|UBM))\}}, \tag{3}$$

where best results were obtained for $\lambda = 3$.

2.4 Individual systems

Now we are ready to describe each of the 8 systems used in the evaluation, they can be found in Table 1. The primary system submitted to EVALITA 2009 consisted of a fusion of these 8 systems. The secondary system is the system **SYS3** achieving the best performance on the development set (see Table 2).

Table 1. Description of systems used in the EVALITA 2009 evaluation campaign.

	parametrization	**UBM**	**modelling**	**verification**
SYS1	$P_{LFCC}(\Delta, M, V)$	512	$M(MAP(\mu))$	MVR
SYS2	$P_{LFCC}(\Delta, \Delta\Delta, M)$	512	$M(MLLR(\mu), MAP(\mu, \sigma))$	SMVR
SYS3	$P_{LFCC}(DCT2_{(3,9)}, M)$	512	$M(MLLR(\mu), MAP(\mu, \sigma))$	MVR
SYS4	$P_{LFCC}(DCT2_{(3,9)}, M)$	256	$M(MLLR(\mu), MAP(\mu, \sigma))$	SMVR
SYS5	$P_{LFCC}(DCT2_{(3,9)}, M)$	512	$M(MLLR(\mu), MAP(\mu, \sigma))$	SMVR
SYS6	$P_{LFCC}(DCT2_{(3,9)}, M)$	1024	$M(MLLR(\mu), MAP(\mu, \sigma))$	SMVR
SYS7	$P_{MFCC}(DCT2_{(4,13)}, M, V)$	512	$M(MLLR(\mu), MAP(\mu))$	MVR
SYS8	$P_{MFCC}(\Delta, \Delta\Delta, M)$	512	$M(MLLR(\mu), MAP(\mu, \sigma))$	MVR

3 Dataset

The data were recorded from landline (PSTN) or mobile (GSM) telephone channels and all spoken in the Italian language. Four data sets were provided by the organizers, namely

- Universal Background Model (UBM) data: 60 speakers (30 female + 30 male) with total duration of speech 1200 minutes.
- Training data: 100 speakers (50 female + 50 male) representing the genuine clients of the system, whereas 6 training conditions (TC1 - TC6) were given depending on the duration of the speech recording and on the telephone channel (PSTN or GSM).

– Development data: 642 ”non blind” sound files representing an additional access trials (321 female + 321 male).
– Test data: two test conditions were considered according to the length of the recording - short (TS1 - cca 10 seconds) and long (TS2 - cca 30 seconds). Both conditions involved recordings from PSTN as well as from GSM.

For our development purposes we divided the development data into four disjoint sets according to the gender (male/female) and channel (PSTN/GSM). Each recording from each set was then tested against each of the models from each of the training conditions (TC1-TC6). Loosely speaking, we constructed 4 trials set (male_PSTN, female_PSTN, male_GSM, female_GSM), where each sound recording from each set was tested against each model of the respective gender in each of the conditions TC1-TC6. Hence, we obtained $4 \times 6 = 24$ files with results (each containing cca 8000 trials) reflecting dependencies on cross-channel conditions and duration of train and test recordings.

4 Experiments and Results

We have exploited only the provided data (UBM and Development data - see Section 3). Genders were handled separately (gender dependent UBMs were trained). All development tests were performed on the development set, which was divided into 24 distinct subsets as described in Section 3. Hence, for each system 24 equal error rates (EERs) were obtained and the main focus was laid on the overall EER, which was computed as their mean. The results (EERs) of particular systems can be found in Table 2.

4.1 Complementarity Examination

We were searching for optimal weights w_i (in the sense of minimal overall EER), which would be used to weigh each systems output as defined in equation (4).

$$result_C = \sum_{i=1}^{S_{NUM}} w_i \cdot result_{SYS_i}, \tag{4}$$

where S_{NUM} stands for the number of employed systems (21 in our case), $result_{SYS_i}$ is the verification score of system SYS_i of one trial, and $result_C$ represents the combined output of involved systems. Weights $w_i, i = 1, \ldots, S_{NUM}$ were computed utilizing a gradient method, where the value of the criterion function (the overall EER) to be optimized was computed in three steps. First, the $result_C$ for each of the 24 trial sets (described in Section 3) was determined and secondly, the individual EERs for each set were computed. At last, the overall EER was estimated as a mean value of such individual EERs.

To reduce the number of exploited systems, the procedure was applied several times. After each estimation of weights, the system with smallest weight (missing any complementary information) was left out and the process repeated. The

Table 2. Equal Error Rates for **SYS1** – **SYS8**.

EERs	female		male		**EERs**	female		male	
SYS1	GSM	PSTN	GSM	PSTN	**SYS2**	GSM	PSTN	GSM	PSTN
TC1	18.23%	5.59%	19.38%	6.58%	TC1	19.54%	5.00%	22.24%	5.63%
TC2	9.88%	17.29%	7.57%	12.50%	TC2	8.26%	19.17%	7.35%	15.63%
TC3	14.24%	3.45%	16.72%	2.74%	TC3	14.27%	2.48%	17.14%	2.36%
TC4	5.59%	16.75%	5.72%	12.59%	TC4	5.50%	15.40%	4.38%	12.56%
TC5	10.71%	6.83%	10.40%	5.63%	TC5	8.19%	6.21%	7.81%	5.20%
TC6	6.76%	6.21%	5.70%	4.03%	TC6	5.41%	3.73%	5.63%	3.16%
	overall EER: 9.63%					**overall EER: 9.26%**			
SYS3					**SYS4**				
TC1	17.20%	3.73%	19.57%	3.13%	TC1	18.92%	5.22%	20.63%	4.38%
TC2	8.90%	19.76%	6.49%	13.41%	TC2	7.40%	17.39%	7.50%	13.13%
TC3	14.00%	2.64%	16.86%	2.35%	TC3	15.69%	2.13%	16.25%	2.84%
TC4	5.07%	13.30%	3.13%	12.12%	TC4	4.97%	14.82%	3.75%	11.88%
TC5	7.83%	5.59%	6.65%	4.08%	TC5	7.89%	6.21%	8.75%	3.74%
TC6	7.40%	3.73%	4.69%	2.50%	TC6	6.21%	4.35%	5.96%	2.50%
	overall EER: 8.50%					**overall EER: 8.85%**			
SYS5					**SYS6**				
TC1	17.69%	3.76%	24.06%	3.82%	TC1	18.83%	3.73%	21.86%	5.00%
TC2	8.33%	16.42%	6.62%	13.13%	TC2	8.70%	18.96%	6.87%	14.38%
TC3	13.66%	2.62%	16.03%	2.50%	TC3	13.55%	2.06%	17.94%	2.26%
TC4	4.93%	15.16%	3.86%	11.38%	TC4	4.58%	13.66%	3.56%	11.88%
TC5	7.64%	4.20%	7.50%	3.02%	TC5	8.92%	5.03%	7.05%	4.84%
TC6	6.37%	4.32%	5.58%	1.88%	TC6	6.75%	2.91%	5.14%	1.75%
	overall EER: 8.52%					**overall EER: 8.76%**			
SYS7					**SYS8**				
TC1	24.22%	9.32%	16.59%	5.19%	TC1	19.96%	7.52%	18.13%	6.23%
TC2	13.31%	25.27%	6.44%	10.07%	TC2	11.59%	23.46%	7.49%	10.22%
TC3	19.25%	4.74%	13.15%	3.75%	TC3	16.15%	4.17%	15.06%	3.40%
TC4	6.21%	17.78%	4.40%	9.38%	TC4	5.13%	17.78%	3.78%	9.44%
TC5	10.34%	10.25%	6.32%	5.00%	TC5	9.94%	7.77%	8.01%	4.95%
TC6	6.48%	6.69%	6.25%	3.75%	TC6	6.76%	4.99%	5.00%	3.01%
	overall EER: 10.17%					**overall EER: 9.58%**			

process ended when the overall EER begun to increase. We ended with 8 systems described in Table 1 with weights {0.3063, 0.0640, 0.1092, 0.1139, 0.0675, 0.0614, 0.1272, 0.1506} corresponding to {**SYS1**,...,**SYS8**}, results are shown in Table 3.

Table 3. Equal Error Rates for combination of systems **SYS1** – **SYS8** acquired on the development set. Weights were trained on the development set and following values were estimated: {0.3063, 0.0640, 0.1092, 0.1139, 0.0675, 0.0614, 0.1272, 0.1506} corresponding to **SYS1** – **SYS8**.

SYSC	female		male	
EERs	GSM	PSTN	GSM	PSTN
TC1	14.93%	2.48%	15.76%	3.18%
TC2	6.83%	14.64%	5.00%	8.75%
TC3	11.18%	1.45%	12.44%	1.51%
TC4	3.37%	13.04%	3.13%	8.21%
TC5	6.21%	4.18%	5.78%	2.58%
TC6	4.90%	3.64%	3.75%	1.62%
	overall EER: 6.61%			

4.2 Results Acquired on the Test set

Genders were handled separately (gender dependent UBMs were trained), however, channels (PSTN/GSM) were not detected. Results obtained in the EVALITA 2009 evaluation can be found in Table 4, where results for both genders were concatenated and the EER was computed for the joint set.

Table 4. EERs [%] and minDCFs obtained in the EVALITA 2009 evaluation campaign for primary and secondary system, where the overall EER, overall minDCF were computed as mean values of individual EERs, minDCFs, respectively. MinDCF was computed according to parameters specified in the EVALITA 2009 task guidelines.

PRIMARY SYSTEM

EERs	TC1	TC2	TC3	TC4	TC5	TC6
TS1	15.96	16.51	13.34	15.69	7.5	5.89
TS2	10.82	12.02	9.83	11.25	3.62	2.11
overall EER: 10.38						
mDCFs	TC1	TC2	TC3	TC4	TC5	TC6
TS1	0.40	0.40	0.33	0.38	0.18	0.14
TS2	0.27	0.28	0.22	0.28	0.09	0.06
overall minDCF: 0.25						

SECONDARY SYSTEM

EERs	TC1	TC2	TC3	TC4	TC5	TC6
TS1	20.48	21.48	17.98	19.71	10.32	7.54
TS2	15.37	17.43	13.12	15.4	4.81	3.32
overall EER: 13.91						
mDCFs	TC1	TC2	TC3	TC4	TC5	TC6
TS1	0.47	0.46	0.43	0.41	0.27	0.20
TS2	0.39	0.39	0.30	0.35	0.14	0.08
overall minDCF: 0.32						

5 Discussion

Experiments were devoted to the investigation of complementarity of simple UBM-GMM systems with varying parameters in parametrization, modelling and verification modules. After inspection of results in Tables 2, 3 and mainly in Table 4 it is quite obvious that the proposed combination of systems significantly improves the verification performance (robustness), the most in cases of channel mismatches (consider conditions TC1 – TC4). Such a behavior is understandable when the change in system parameters is significant. However, also slight changes in the modelling procedure (e.g. number of mixtures) or in the verification can still bring some additional information as proved the experiments.

Acknowledgments. This research was supported by the Ministry of Education of the Czech Republic, project No. 2C06020 and project No. MŠMT LC536.

References

1. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, pp. 19–41 (2000)
2. Gales, M.J.F.: Maximum Likelihood Linear Transformation for HMM-based Speech Recognition. Tech. Report, CUED/FINFENG/TR291, Cambridge Univ. (1997)
3. Zajíc, Z., Machlica, L., Müller, L.: Refinement approach for adaptation based on combination of MAP and fMLLR. Text, Speech and Dialogue (2009)
4. Padrta, A. and Radová, V.: Comparison of several speaker verification procedures based on GMM. Journal of the Acoustical Society of Korea, pp. 1777–1780 (2004)