Brno University of Technology System for Interspeech 2010 Paralinguistic Challenge

Marcel Kockmann, Lukáš Burget and Jan "Honza" Černocký

Brno University of Technology, Speech@FIT, Brno, Czech Republic

{kockmann|burget|cernocky}@fit.vutbr.cz

Abstract

This paper describes Brno University of Technology (BUT) system for the Interspeech 2010 Paralinguistic Challenge. Our submitted systems for the Age- and Gender-Sub-Challenges employ fusions of several sub-systems. We make use of our own acoustic frame-based feature sets, as well as the provided utterance-based acoustic, prosodic and voice quality features. Modeling is based on Gaussian Mixture Models (GMM) and Support Vector Machines (SVM), followed by linear Gaussian backends and logistic regression-based fusion. For a single sub-system, we obtain improvement of about 2% absolute, for both tasks, on the development-set. Our final fusion results in nearly 9% absolute improvement for the *Age* task and about 4.5% for the *Gender* task on the development set. On the final test set we obtain 3.5% and 2% absolute improvement, respectively.

Index Terms: Age- and gender-recognition, GMM, MMI, eigenvoice, fusion

1. Introduction

Age- and gender recognition is the problem of automatically recognizing the age and/or gender of a person from speech. The attributes of the human voice significantly differ between genders, and they also change during the life of a single person. This makes it possible, to apply pattern recognition and machine learning algorithms to appropriate parameterization of speech – so called features –, to automatically recognize age and gender from speech samples.

Our submissions to the Age- and Gender-Sub-Challenges of the Interspeech 2010 Paralinguistic Challenge [1] employ several sub-systems, based on standard acoustic features. Furthermore, we make use of the provided acoustic, prosodic and voice quality features. Models are based on simple GMMs, but we also investigate into more sophisticated approaches like Maximum-Mutual-Information (MMI)-training [2] or Joint-Factor-Analysis (JFA) [3]-based approaches. Furthermore, we use SVMs for classification, using diverse feature sets. Some use direct parameterizations of speech, while others process parameters of GMM models.

Overall, we will present 6 different sub-systems in Section 2, some of them outperforming the provided baseline systems by about 2% absolute. The best provided baseline results [1] on the development set are 44.24% unweighted accuracy (UA) for a combined 7-class *Age+Gender* task, 47.11% UA for the 4-class *Age* task and 77.28% UA for the 3-class *Gender* task. If not mentioned otherwise, all results are presented for the provided development-set [1].

In Section 3, we will present fusion experiments of all sub-systems, resulting in huge improvements over the baseline systems. The fused systems are also used as our submission on the independent test-set of the *Age* and *Gender* Sub-Challenges.

In Section 4, we will review our experiments and draw a conclusion to our approach.

The work was partly supported by European project MOBIO (FP7-214324), Grant Agency of Czech Republic project No. 102/08/0707, Czech Ministry of Education project No. MSM0021630528 and by BUT FIT grant No. FIT-10-S-2. Marcel Kockmann was supported by SVOX Deutschland GmbH.

2. Systems

2.1. MFCC-MAP-GMM

The most widely used features in speech processing are MFCCs [4]. We will use them as our basic features. MFCC vectors are generated every 10 ms on a 20 ms frame of speech weighted by a Hamming window. The output of a Mel filter bank with 25 bands is processed by Discrete Cosine Transform (DCT) and generates 13 cepstral coefficients including C0. Then, cepstral mean subtraction (CMS) is applied on each coefficient per utterance.

The temporal trajectories of individual cepstral coefficients are filtered using standard RelAtive SpecTrAl (RASTA) filter [5] to remove slow and very fast spectral changes which do not appear to be characteristic for natural speech.

Simple MFCCs do not model any temporal characteristics, which might be discriminative for age and gender recognition. For this purpose, we generate delta and double-delta regression coefficients of the static features. This results in 39 dimensional feature vectors containing information spanning a context of 9 frames.

For all our features, non-speech frames are discarded and only speech frames are considered in the following stages of training models and verification. Speech/non-speech segmentation is performed by our Hungarian phone recognizer [6].

Our first system is based on standard Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) paradigm [7]. Prior to any class-dependent model training, a class-independent model is trained on the pooled feature vectors of all training data of all classes. Following speaker recognition terminology, we call this Universal Background Model. Weights, means and variances of the UBM are trained in maximum-likelihood way with Expectation-Maximization (EM) algorithm.

The class models are determined by relevance MAP adaptation given all class features. During testing, instead of frame-based full log-likelihood evaluation, we perform an approximate fast linear scoring based on utterance statistics [8].

First experiments are carried out with separate models for

Table 1: *MFCC-MAP-GMM system. % UA for 4-class* Age *and 3-class* Gender *tasks. Separate and combined (7-class) systems with different number of Gaussian components.*

		64	128	256	512
Separate	Gender	73.25	73.80	74.74	**79.88**
	Age	44.41	44.95	46.19	47.48
Combined	Gender	71.72	74.9	75.58	78.45
	Age	45.04	46.88	47.85	**49.09**

Table 2: *MFCC-MAP-GMM system. Results for single combined 7-class system with task-specific backends.*

Task	Backend	% UA	% Impr.	% WA
Age-Gender	-	45.86	+1.62	45.96
Age	-	49.09	+1.98	48.27
	Gaussian	48.84	+1.73	47.20
Gender	-	78.45	+1.17	85.75
	Gaussian	79.47	+2.19	83.94

Age and *Gender* tasks. As shown in Table 1, the performance clearly enhances with growing number of Gaussian components. For both tasks, the best results are achieved with 512 components. Using a combined 7-class system with simple mapping to both, *Age* and *Gender* tasks, shows the same trend. We obtain the best result of 79.88% for *Gender* with a 3-class system, which is an absolute improvement of 2.6% over the best provided baseline. For *Age*, we get a significant improvement (over the 4-class system) by using a 7-class system with mapping to 4 classes, and obtain 49.09% UA. This outperforms the baseline by nearly 2% absolute. Following these experiments, we decided to use combined 7-class system, with task-specific backends for both, the *Age* and the *Gender* tasks. Instead of simple mapping, we use linear Gaussian backends [9], directly trained on the development set. Table 2 shows UAs and weighted accuracies (WA) for *Age+Gender*, *Age* and *Gender* tasks, using one single 512 component GMM with task-specific Gaussian backends. We also indicate the improvement on UA compared to the best provided baseline. While simple mapping seems better for the *Age* task, we get about 1% improvement on the *Gender* task, due to the backend.

Note, that we also performed experiments using channel compensation on top of the MAP-GMM system, similar to our submission to the 2009 Emotion Challenge [10], to compensate for channel mismatch. Although diverse recording conditions are clearly an issue in the *aGender* corpus, we could not observe significant and consistent gain due to the use of channel compensation. Again, the channel compensation seems to fail due to the short test utterances.

2.2. MFCC-MMI-GMM

For our second system, we use the same MFCC $\Delta\Delta$ features as described in Section 2.1. Aiming at complementary systems, we first used Shifted-Delta-Cepstra [11] for this modeling approach, but these caused a huge degradation in performance.

MMI is a discriminative training technique often applied to classification tasks like age or gender recognition [2], or similar [12]. For this approach, an initial set of models is trained per class under conventional Maximum Likelihood (ML) framework, as for the UBM, but using class specific feature vectors only. These serve as a starting point for further discriminative re-estimations of means and variances using Maximum Mutual Information criterion. For details on MMI and its implementation, see [13, 14].

Table 3: *MFCC-MMI-GMM system. % UA for 4-class* Age *and 3-class* Gender *tasks. Separate generative and discriminative systems with different number of Gaussian components.*

Age	4	8	16	32	64	128	256
ML	40.6	41.9	42.4	43.8	44.7	45.9	46.9
MMI	42.7	42.9	43.6	44.6	45.7	46.5	**47.2**
Gender	4	8	16	32	64	128	256
ML	64.7	71.7	74.0	76.0	77.3	78.1	79.3
MMI	67.9	72.0	74.5	76.1	77.2	78.1	**79.4**

Table 4: *MFCC-ML-GMM system. Results for single combined 7-class system with task-specific backends.*

Task	Backend	% UA	% Impr.	% WA
Age-Gender	-	45.24	+1.00	45.04
Age	-	48.46	+1.35	47.34
	Gaussian	49.16	+2.05	47.66
Gender	-	78.41	+1.13	85.21
	Gaussian	79.30	+2.02	84.29

For both models, verification is done frame-by-frame with full log-likelihood computation.

Once more, we train separate systems for both tasks. Table 3 shows the influence of number of components on this model. Again, we obtain the best results with the largest model we evaluate, for both tasks. For the MMI systems, we only see significant improvement for small number of Gaussians. While retraining the big GMMs using MMI keeps increasing both, the objective function and the number of correct training segments, this seems to overtrain the model, as we even see degradation in performance on the development set after many iterations.

Following this, we will rather use a big, purely ML trained model as our sub-system. Table 4 shows results for a combined 256 component model, followed by task-specific backends. Again, we obtain consistent improvements of up to 2% over the baselines.

2.3. SMILE-SVM

For our third sub-system, we use the utterance-based features provided by the organizers [1]. These are 450-dimensional features including acoustic, prosodic and voice quality features. We discard dimensions 190 and 198, as they do not seem to contain meaningful values. Further, all coefficients are rank-normalized, using 1000 bins estimated on the distribution of the background data.

We use an SVM for this sub-system as a classifier. We directly train a multi-class SVM using *libsvm* [15], that internally trains *1-against-1* SVMs for all necessary combinations.

Evaluation for the 7-class task can directly be done by the classification output of the multi-class SVM. In order to use one combined SVM also for Age- and Gender-tasks (and to provide probability outputs per class), we can apply additional backends on the *1-against-1* decision values. We use task-specific linear Gaussian backends [9], that map the 21-dimensional out-

Table 5: *SMILE-SVM system. Results for single combined 7-class system with task-specific backends.*

Task	Backend	% UA	% Impr.	% WA
Age-Gender	-	42.81	-1.43	42.71
	Gaussian	47.55	+3.31	47.44
Age	Gaussian	48.46	+1.35	46.19
Gender	Gaussian	78.20	+0.92	81.04

Table 6: *PLP-GMM-SVM system. Results for single combined 7-class system with task-specific backends.*

Task	Backend	% UA	% Impr.	% WA
Age-Gender	-	41.58	-2.66	41.47
	Gaussian	43.40	-0.84	43.54
Age	Gaussian	45.59	-1.52	43.34
Gender	Gaussian	75.50	-1.78	77.92

puts of the SVM to 7-, 4- and 3-dimensional values, respectively. The parameters of the backends are trained directly on the development-set.

Results for this system are shown in Table 5. While we get worse performance than the baseline system, using the classification output of the SVM, we obtain significant improvements using the linear backends. Especially for the combined 7-class task, we achieve 47.55% UA, which is an improvement of 3.3% over the baseline.

2.4. PLP-GMM-SVM

For our fourth sub-system, we use Perceptual-Linear-Predictive (PLP) [16] features. 12th order linear-predictive coefficients are estimated on the same mel-filterbank outputs as used in Section 2.1. Further, CMS is applied per utterance and the features are augmented with their first order derivatives, resulting in 26-dimensional features. Non-speech frames are discarded by our VAD.

A 256 component UBM is trained on the pooled feature vectors for all classes. Afterwards, we gather maximum-likelihood zero order sufficient statistics for all utterances, based on the UBM. The 256-dimensional vectors are normalized by number of frames and further rank-normalized, as described in Section 2.3

We use the same SVM-approach for classification, as used in the SMILE-SVM system (Section 2.3).

Table 6 shows generally worse performance than the preceding systems. For all tasks, we obtain about 1-2% worse UA than the baselines. Still, the system might give complementary information to the fusion, due to its diverse features and modeling approach.

2.5. MFCC-JFA-Eigenvoice-SVM

Our fifth system is also based on the MFCC-$\Delta\Delta$ features as described in Section 2.1. However, the actual features used for classification, are estimated hyper-parameters of a Joint Factor Analysis (JFA)-based speaker recognition system [3].

So called *eigenvoices* are used in JFA-based speaker recognition systems. They define a low-dimensional sub-space of the full GMM-space, in which speaker-variability is high. So called *speaker factors* – normally distributed vectors – are estimated during speaker-enrollment and control the position of the speaker model within the *eigenvoice*-space. We use theses vectors, estimated per utterance, as input to an SVM. They should be highly discriminative for speaker characteristics, like age and gender.

Table 7: *MFFC-JFA-eigenvoice-SVM system. Results for single combined 7-class system with task-specific backends.*

Task	Backend	% UA	% Impr.	% WA
Age-Gender	-	41.97	-2.27	41.67
	Gaussian	45.26	+1.02	45.35
Age	Gaussian	47.99	+0.88	45.50
Gender	Gaussian	76.06	-1.22	78.11

Starting with a 256 component UBM as used in Section 2.1, we train the JFA hyper-parameters using the *2004 NIST Speaker Recognition Evaluation* corpus [17]. We train 50 eigenvoices, using utterances from 309 speakers of different age and gender. Afterwards, we estimate 50-dimensional *speaker-factor* vectors for all *aGender* training and test utterances. These are then used to train and test multi-class SVM, as described in Section 2.3.

Results on the development-set are shown in Table 7. Using the Gaussian backends, we obtain improvements of about 1% UA over the baselines for the combined *Age+Gender*, as well as the *Age* task. The performance for the *Gender* task stays below the baseline (76% UA).

2.6. MFCC-JFA-Anchor-SVM

Similar to the latter system, this one is based on the MFCC-$\Delta\Delta$ features as described in Section 2.1 , but the actual features used for classification, are scores of a Joint Factor Analysis (JFA)-based speaker recognition system [3].

The idea is, to use a representative part of the training speakers as *anchor models*, and the scores for each utterance (from second part of training-set, as well as development and test) against the *anchor models* as features.

We use the same JFA system as described in the latter section, but train additional *eigenchannels*. As the *aGender* corpus consists of telephone calls from different sessions, we hope to compensate for channel mismatch, using the additional eigenchannels. Afterwards, the *aGender* training set is split into two parts, each consisting of approximately the same number of speakers per class. All utterances of the 235 speakers in the *anchor-model*-set are concatenated per speaker, and a separate speaker model is enrolled using the JFA system.

Following this, we score the second part of the training data, as well as the development and test data, using the JFA system against the enrolled *anchor-models*. We use a JFA system to both, rapidly adapt the enrolled models and to compensate for channel-mismatch. The 235-dimensional vectors for all utterances in the second part of the training set are then used to train multi-class SVM (according to Section 2.3), without further normalization. Results for the this system are shown in Table 8. We obtain similar accuracies compared to the baselines for the *Age+Gender* and *Age* tasks. For the *Gender* task we obtain the worst performance so far, with 74.54% UA.

3. Submission

For the submission, we hand in for the independent test-set, we perform calibrated fusion of several sub-systems using multi-class logistic regression toolkit [9]. The fusion parameters are directly trained on the development set and are then applied to

Table 8: *MFFC-JFA-Anchor-SVM system. Results for single combined 7-class system with task-specific backends.*

Task	Backend	% UA	% Impr.	% WA
Age-Gender	-	39.59	-4.65	39.40
	Gaussian	44.02	-0.22	43.75
Age	Gaussian	46.88	-0.23	44.61
Gender	Gaussian	74.54	-2.74	76.79

Table 9: *Calibrated fusion of several sub-systems for the submission. Results on the development (DEV) and evaluation (EVAL) sets.*

	Task	Model	Sys.	% UA	% Impr.	% WA
DEV	Age-Gender	7-class	1-6	53.86	+9.62	54.20
	Age	7-class	1-6	56.03	+8.92	55.29
		4-class	1-5	52.88	+5.77	50.87
	Gender	7-class	1-6	81.57	+4.29	87.14
		3-class	1-5	81.82	+4.54	84.85
EVAL	Age	7-class	1-6	52.35	+3.44	51.17
	Gender	7-class	1-6	83.14	+1.93	85.66

the test data.

Results for our final systems on the development-set, are shown in Table 9. For the combined *Age+Gender* task, we obtain a huge improvement due to the fusion of all 6 sub-systems described in Section 2. We reach 53.86% UA, which is an improvement of 9.62% absolute over the baseline.

Using the scores from systems with *Age*-specific backends, fusion of these 4-dimensional score-vectors results in less improvement. We even obtain a worse UA of 52.88% for the 4-class task, than for the 7-class task. Here, we exclude subsystem 6, as it degrades the performance on the development set. However, simple mapping of the 7-class-system-fusion gives much better results on the *Age* task. We obtain 56% UA, an improvement of nearly 9% absolute over the provided baseline.

For the *Gender* task we obtain similar results for both fused systems. Generally, we obtain less improvement on this task, due to the fusion. The fused system only performs about 2% UA absolute better, than the best single system. Finally, we reach an UA of 81.82%, which is an improvement of 4.5% over the best provided baseline.

Following this, we use a single 7-class system for the submission. We obtain 52.35% UA on the *Age* task and 83.14% on the *Gender* task, which is an improvement of 3.44% and 1.93%, respectively, over the baselines. Note, different from [1], we do not include development data into the training for the final system, mainly because we train the linear backends and the fusion on the development set and further would not be able to evaluate the system.

4. Conclusions

While we obtain huge improvements on the development set (especially for the *Age* task), we achieve smaller improvements on the test set. This trend is contrary to the results reported for the provided baseline systems [1]. This may have several reasons: First of all, our results on the development set may be over-optimistic, due to the "optimal" training of backends and fusion and may not generalize that well on the test set. Second, for the test set, the provided baselines use nearly twice as much training data by adding the development data for training. This makes these results hard to compare to ours, as it was not possible to exploit all the data, while being able to train backends and fusion properly (at least without making use of expensive cross-fold-validations). Still, our single GMM based systems outperform the baselines without any backend or calibration/fusion on the development set.

5. References

[1] B. Schuller *et al.*, "The INTERSPEECH 2010 Paralinguistic Challenge," *Submitted to Interspeech (2010), ISCA, Makuhari, Japan*, 2010.

[2] V. Hubeika *et al.*, "Maximum likelihood and maximum mutual information training in gender and age recognition system," 2007.

[3] P. Kenny *et al.*, "A study of inter-speaker variability in speaker verification," *IEEE Trans. Audio*, Jan 2008.

[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, no. 4, pp. 357–366, Jul 1980.

[5] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE transactions on speech and audio processing*, vol. 2, no. 4, pp. 578–589, 1994.

[6] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," *Proceedings of ICASSP 2006, Toulouse*, pp. 325–328, Mar 2006.

[7] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[8] O. Glembek *et al.*, "Comparison of scoring methods used in speaker recognition with Joint Factor Analysis," *Proc. of ICASSP, Taipei*, 2009.

[9] N. Brummer, "Focal multi-class – tools for evaluation, calibration and fusion of, and decision-making with, multi-class statistical pattern recognition scores," *Online on: http://sites.google.com/site/nikobrummer/focalmulticlass*, 2007.

[10] M. Kockmann, L. Burget, and J. Cernocky, "Brno University of Technology system for Interspeech 2009 Emotion Challenge," *Proc. Interspeech, Brighton*, pp. 348–351, 2009.

[11] P. Torres-Carrasquillo *et al.*, "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," *Seventh International Conference on Spoken Language Processing*, 2002.

[12] P. Matejka *et al.*, "Brno University of Technology System for NIST 2005 language recognition evaluation," *Proc. Odyssey*, 2006.

[13] D. Povey, "Discriminative training for large vocabulary speech recognition," *PhD thesis, Cambridge University Engineering Dept, 2003*, p. 172, Jun 2003.

[14] L. Burget, P. Matejka, and J. Cernocky, "Discriminative training techniques for acoustic language identification," *In: Proceedings of ICASSP 2006, Toulouse, FR*, pp. 209–212, Apr 2006.

[15] C. Chang and C. Lin, "Libsvm : a library for support vector machines, 2001," *Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[16] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America, vol. 87, no. 4, pp. 1738–1752*, 1990.

[17] A. Martin and M. Przybocki, "2004 NIST speaker recognition evaluation," *Linguistic Data Consortium, LDC2006S44, Philadelphia*, 2006.