AGE AND GENDER RECOGNITION FOR TELEPHONE APPLICATIONS BASED ON GMM SUPERVECTORS AND SUPPORT VECTOR MACHINES

Tobias Bocklet[1], *Andreas Maier*[1], *Josef G. Bauer*[2], *Felix Burkhardt*[3], *Elmar Nöth*[1]

[1] Institute of Pattern Recognition, *University of Erlangen-Nuremberg*, Germany
[2] Siemens AG, CT IC5, München, Germany
[3] T-Systems Enterprise Service GmbH, SSC ENPS, Berlin, Germany

tobias.bocklet@informatik.stud.uni-erlangen.de,
noeth@informatik.uni-erlangen.de

ABSTRACT

This paper compares two approaches of automatic age and gender classification with 7 classes. The first approach are *Gaussian Mixture Models* (GMMs) with *Universal Background Models* (UBMs), which is well known for the task of speaker identification/verification. The training is performed by the EM algorithm or MAP adaptation respectively. For the second approach for each speaker of the test and training set a GMM model is trained. The means of each model are extracted and concatenated, which results in a GMM supervector for each speaker. These supervectors are then used in a support vector machine (SVM). Three different kernels were employed for the SVM approach: a polynomial kernel (with different polynomials), an RBF kernel and a linear GMM distance kernel, based on the KL divergence. With the SVM approach we improved the recognition rate to 74% ($p < 0.001$) and are in the same range as humans.

***Index Terms*—** Acoustic signal analysis, speaker classification, age, gender, Gaussian mixture models (GMM), support vector machine (SVM)

1. INTRODUCTION

The human voice not only provides the semantics of spoken words. It also contains speaker dependent characteristics. Examples for such non-verbal information are the identity, the gender, the emotional state or the age of a speaker. In telephone calls of everyday life we extract these speaker specific characteristics and adapt our speaking style to the person we are talking to. Apart from gender, in automatic speech recognition (ASR) information about speaker characteristics are rarely used. There are some approaches to identify dialogues with angry or unsatisfied users/callers [1]. But there are only a few approaches, that use the age of speakers in ASR systems [2, 3], although there are a lot of useful applications associated with this task. The age (combined with the gender) information can be used to adapt the ASR system to a certain customer. Other examples are the adaptation of the waiting queue music, the offer of age dependent advertisements to callers in the waiting queue or to change the speaking habits of the *text-to-speech module* of the ASR system. Statistical information on the age distribution of a caller group might also be an application.

In 2007 *T-Systems*, *Siemens AG*, *Deutsches Forschungszentrum für Künstliche Intelligenz* and *Sympalog Voice Solutions* compared four different age recognition systems on two corpora [4]. The most successful systems used *Mel Frequency Cepstral Coefficients* (MFFCs) and either performed multiple phoneme recognition or modeled the different age classes with *Gaussian Mixture Models* (GMMs).

In this paper we also use MFCCs as features and compare two different approaches. On the one hand a GMM - UBM (*Universal Background Model*) system, which has been shown to be very effective for the task of speaker identification [5, 6]. On the other hand we use Support Vector Machines (SVMs) with a GMM Supervector to identify the speaker's age. This approach was also published in terms of speaker identification/verification [7].

The outline of this article is organized as follows: Section 2 depicts the evaluation corpora on which the two systems are trained and tested. The baseline GMM-UBM system is described in Section 3. The basic framework for SVMs and the used kernel functions are summarized in Section 4. Section 4.2 shows the idea of GMM Supervectors and describes the SVM-based classification system. In Section 5 we show the results of the SVM-based approach and compare them to the baseline GMM system developed in our group and to the 128-dimensional GMM and parallel phone recognition (PPR) system of [4]. We also compared our results to the human baseline experiment mentioned in [4]. The paper finishes with a conclusion and a short outlook in Section 6.

2. CORPORA

The data was taken from the German *SpeechDat II* corpus which is annotated with gender and age labels as given by callers at the time of recording. The scenario of the corpus is telephone speech, where the speakers called an automatic recording system and read a set of words, sentences and digits. The used data was an age-balanced subset of the 4000 native German speakers. The training and test set is identical to [4]. For each class about 80 speakers were used for training. The training data consisted of 44 utterances per speaker.

In order to simulate a mismatched condition of training and test data we also evaluated the system on a 23 speaker subset of the *VoiceClass* corpus. This is a dataset collected by *Deutsche Telekom* and it consists of 660 native German speakers. These speakers also called an automatic recording system and talked about their favorite dish. For each speaker between 5 and 30 seconds of speech data was available. The age structure is not balanced, i.e children and youth speakers are represented significantly higher than senior speakers. For this corpus also labels of gender and age were available for each speaker.

The labels of the training and test sets were used, to build up the

1-4244-1484-9/08/$25.00 ©2008 IEEE

following 7 gender-dependent age classes:

- Children (C): ≤ 13 years
- Young male (YM) and female (YF) speakers: 14-19 years
- Adult male (AM) and female (AF) speakers: 20-64 years
- Senior male (SM) and female (SF) speakers: ≥ 65 years

3. BASELINE GMM SYSTEM DESCRIPTION

For the baseline system we use a GMM-UBM system. Each of the 7 classes is modeled by a *Gaussian Mixture Model* (GMM), composed of M unimodal Gaussian densities:

$$p(\boldsymbol{c}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{M} \omega_i p_i(\boldsymbol{x}|\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}), \qquad (1)$$

where ω_i denotes the weight, $\boldsymbol{\Sigma_i}$ the covariance matrix and $\boldsymbol{\mu_i}$ the mean vector of the i-th Gaussian density. We varied the number of mixtures M from 16 to 512 in 2^x steps. For classification a standard *Gaussian Mixture Classifier* is used. The classifier calculates for each feature vector of a specific test speaker the allocation probability for each GMM-age model. This is done for each frame of one utterance. The probabilities of each age model are then accumulated. The model which achieved the highest value is expected to be the correct one.

3.1. Feature Extraction

As features the commonly used *Mel Frequency Cepstrum Coefficients* (MFCCs) are used. They examine each of the 18 Mel-bands but only consider a time window of 16 ms with a time shift of 10 ms. This gives us a feature vector with 24 components (log energy, MFCC(1)-(11)). Furthermore the first order derivatives are computed by a regression line over 5 consecutive frames.

3.2. Training

The training process is shown in Figure 1. After extraction of the MFCCs a *Universal Background Model* (UBM) is created by employing all the available training data, using the *Expectation-Maximization* (EM) algorithm [8]. The UBM is then employed as an initial model for a standard EM training with the age dependent training data or for the *Maximum A Posteriori* (MAP) adaptation [9]. Both algorithms take the UBM as an initial model and create one GMM for each age class. MAP adaptation calculates the age-dependent Gaussian mixture components by a single iteration step and combines them with the UBM parameters. The number of iterations in the EM training was set to 10.

4. SUPPORT VECTOR MACHINES

4.1. SVM Classification

The Support Vector Machine (SVM) [10] performs a binary classification $y \in (-1, 1)$ based on hyperplane separation. The separator is chosen in order to maximize the distances between the hyperplane and the closest training vectors, which are called *support vectors*. By the use of kernel functions $K(\boldsymbol{x_i}, \boldsymbol{x_j})$, which satisfy the Mercer condition, the SVM can be extended to non-linear boundaries:

$$f(\boldsymbol{x}) = \sum_{i=1}^{L} \lambda_i y_i K(\boldsymbol{x}, \boldsymbol{x_i}) + d \qquad (2)$$

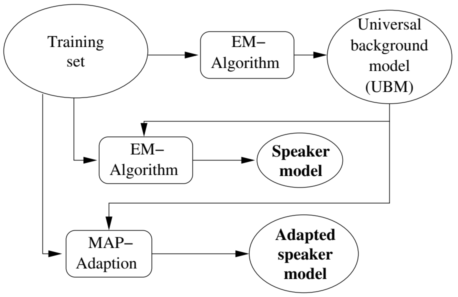

Fig. 1. Training of the GMM Baseline System

where y_i are the target values and $\boldsymbol{x_i}$ are the support vectors. λ_i have to be determined in the training process. L denotes the number of support vectors and d is a (learned) constant. The task of this paper is a 7-class age identification. So the binary SVM has to be extended. The simplest way is to separate each age class from all others. Therefore $N \times (N-1)/2$ classifier are created, each of them separating two classes.

4.2. GMM Supervector Classification

A GMM supervector is created by concatenating the M 24-dimensional mean vectors of a speaker model (Eq. 1). The supervectors are built for every speaker and a label for one of the seven classes is assigned to each vector. In the baseline system we derive a GMM from the UBM for each age class. For the supervector classification approach we use the same UBM and adapt for every speaker of the training and test set a GMM by EM training or MAP adaptation. We treated several aspects of adaptation: We used full covariance matrices, diagonal covariance matrices and we also considered only adapting the mean values. The GMM supervectors can be regarded as a mapping from the utterance of a speaker (in our case the MFCCs) to a high-dimensional feature vector. The supervectors are then used as support vectors and are taken as input for SVM training.

4.3. Employed Kernels

In this paper we applied three different kernel types: the polynomial kernel Eq. (3), the radial basis function (RBF) kernel (Eq. 4) and a GMM-based distance kernel (Eq. 6), which is derived from the KL divergence. This kernel is also very similar to the *Mahalanobis distance*.

$$\begin{aligned} K(\boldsymbol{x_i}, \boldsymbol{x_j}) &= (\boldsymbol{x_i}^T \boldsymbol{x_j} + 1)^n & (3) \\ K(\boldsymbol{x_i}, \boldsymbol{x_j}) &= exp\left[\frac{1}{2}\left(\frac{\|(\boldsymbol{x_i} - \boldsymbol{x_j}\|}{\psi}\right)^2\right] & (4) \end{aligned}$$

n in Eq. 3 defines the polynomial order and ψ in Eq. 4 denotes the width of the radial basis function. These kernels are commonly used in the case of SVM-based classification.

For Gaussian densities (created with mean-adapted MAP) an adequate kernel exists [7]. It is an approximation of the KL divergence

Densities	EM-f	EM-d
32	35% 35%	19% 25%
64	46% 43%	18% 24%
128	41% 42%	43% 32%
256	37% 42%	43% 34%
512	44% 45%	48% 40%

Densities	MAP-f	MAP-d	MAP-dM
32	29% 26%	29% 26%	44% 38%
64	43% 41%	30% 28%	33% 30%
128	45% 40%	40% 36%	**49% 41%**
256	45% 40%	39% 37%	44% 39%
512	44% 41%	43% 42%	46% 43%

Table 1. Precision and recall on the *SpeechDat II* corpus with different training algorithms (EM-f ↔ EM with full covariance matrices; EM-d ↔ EM with diagonal covariance matrices; MAP-f ↔ MAP with full covariance matrices; MAP-d ↔ MAP with diagonal covariance matrix; MAP-dM ↔ MAP with diagonal covariance matrices [only means are adapted])

[11] which can be rewritten in closed form as

$$K(\boldsymbol{\mu}^{\boldsymbol{a}}, \boldsymbol{\mu}^{\boldsymbol{b}}) = \sum_{i=1}^{N} \omega(\boldsymbol{\mu}_{\boldsymbol{i}}^{a})^T \boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1} (\boldsymbol{\mu}_{\boldsymbol{i}}^{b}) \quad (5)$$

$$= \sum_{i=1}^{N} \left(\sqrt{\omega}\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1/2} \boldsymbol{\mu}_{\boldsymbol{i}}^{a}\right)^T \left(\sqrt{\omega}\boldsymbol{\Sigma}_{\boldsymbol{i}}^{-1/2} \boldsymbol{\mu}_{\boldsymbol{i}}^{b}\right). \quad (6)$$

5. EXPERIMENTAL RESULTS

In this work we performed age recognition experiments on two different corpora: the *SpeechDat II* corpus and the *VoiceClass* corpus provided by *Deutsche Telekom*.
First we performed preliminary experiments (Section 5.1) in order to determine the best parameters for the GMM-UBM system. A second set of preliminary experiments selected the SVM kernel with the best performance. Section 5.2 compares the recognition results of the GMM-UBM system and the supervector-based SVM approach of our lab with the best results achieved in [4].

5.1. Preliminary Experiments

5.1.1. GMM-UBM system

We examined the influence of the number of Gaussian Densities, the training algorithm (EM, MAP) and the form of the covariance matrices (full and diagonal) on the recognition results. In the case of MAP adaptation we adapted all GMM-components ($\omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}$) or only the means respectively. The results are shown in Table 1. For our baseline systems the best results where achieved by a MAP-trained GMM with 128 Gaussian densities. Only the mean vectors of the model were adapted.

5.1.2. SVM system

Table 2 summarizes the overall precision and recall of the supervector-based SVM system with the different kernels described in Section 4.3. It can be seen, that the adjustment of the kernel parameters is very important (especially for the RBF Kernel). The best results were achieved with MAP adaptation. The results reached with full covariance matrices and 64 Gaussian densities are comparable to diagonal covariances and 512 Gaussian densities. But with 512 Gaussian densities, MAP adaptation, full covariance matrices and a linear kernel we achieved a recall of 74% and a precision of 77%.

Kernel	full	dia	diaMean
EM Training 64 Densities			
poly e=1	63% 61%	49% 47%	–
poly e=3	62% 60%	49% 48%	–
RBF 0.01	29% 50%	23% 38%	–
RBF 0.1	65% 41%	25% 38%	–
KL-based	41% 43%	47% 48%	–
512 Densities			
poly e=1	– –	64% 61%	–
poly e=3	– –	66% 64%	–
RBF 0.01	– –	09% 15%	–
RBF 0.1	– –	26% 43%	–
KL-based	– –	53% 52%	–
MAP Adaptation 64 Densities			
poly e=1	66% 65%	59% 56%	58% 55%
poly e=3	**66% 66%**	59% 55%	56% 53%
RBF 0.01	44% 49%	25% 42%	21% 36%
RBF 0.1	53% 51%	56% 46%	52% 45%
KL-based	47% 48%	58% 57%	57% 57%
512 Densities			
poly e=1	**77% 74%**	66% 63%	66% 64%
poly e=3	75% 74%	67% 63%	68% 66%
RBF 0.01	21% 24%	26% 19%	26% 19%
RBF 0.1	59% 57%	61% 56%	66% 60%
KL-based	57% 60%	55% 53%	56% 54%

Table 2. Precision and recall on the *SpeechDat II* corpus with different kernels and training (full ↔ full covariance matrices; dia ↔ diagonal covariance matrices; mean ↔ diagonal covariance matrix with only adapting the mean vector)

5.2. SVM vs GMM system

Table 3 shows the evaluation results on the two different corpora. For the *SpeechDat II* corpus, the accuracy can be improved – compared to our GMM-UBM system – by the supervector-based SVM system by 57% from 49% to 77%. The recall of this approach was 74%, and the recall of the best GMM-UBM system was 41%. This is a relative improvement of 80% (significant with $p < 0.001$). Compared to the PPR system of [4] the precision of our SVM system is 43% higher and the recall 35% respectively. This is significant with $p < 0.001$. The confusion matrices of the two systems on the *SpeechDat II* corpus are tabulated in Table 4 and Table 5. The confusions of the SVM-system (Table 5) are more balanced and way more intuitive than those of the GMM-UBM system Table 4.
If we compare the performance of the human listeners to the SVM approach, both the recall and the precision of the SVM approach are higher. The differences in precision between human and machine are significant with $p < 0.001$. The differences in recall are

	SpeechDat II		*VoiceClass*	
System	precision	recall	precision	recall
GMM ([4])	42%	46%	64%	65%
PPR	54%	55%	60%	58%
GMM-UBM	49%	41%	65%	63%
SVM	77%	74%	61%	60%
HUMAN	55%	69%	–	–

Table 3. Overall precision and recall for the best two systems of [4] (GMM and parallel phone recognizer [PPR]) and of our two systems; tested on the two different corpora; the last row shows the performance of human listeners

ac\cl	C	YF	AF	SF	YM	AM	SM
C	**83**		8		8		
YF	55	**20**	15	5			5
AF	10		**30**	35		5	20
SF	25	4	8	**33**		8	21
YM	5		5	10	**30**	5	45
AM	16		5		5	**47**	26
SM	28			6	6	17	**44**

Table 4. Relative confusion matrix of the best GMM-UBM system (see text) on the *SpeechDat II* corpus; the columns contain the actual age (ac) and the rows contain the classified age (cl) (overall precision 49%)

not significant ($p > 0.1$). Note that the *F-measure* [12] of the SVM-system leads to higher values than the F-measure calculated on the results of the human listeners (with weighs of 0.5, 1 and 2).
To compare the robustness of the two approaches against data from different domains and channels, we used the already trained GMMs (or SVMs respectively) and tested on the *VoiceClass* database. The robustness of both of our systems seems to be good. The differences of the 4 approaches are negligible.

6. CONCLUSION

We applied the GMM supervector-based SVM approach to the field of automatic age recognition in combination with gender recognition. We compared this approach to the GMM-UBM approach, which is state-of-the art for the task of text-independent speaker identification, and to the PPR system of [4]. We only investigated spectral features. The SVM systems outperformed all of these approaches for the same domain corpus. Compared to the best system of [4] (PPR) we improved the accuracy by 43% and the recall by 35% (significance: $p < 0.001$).

ac\cl	C	YF	AF	SF	YM	AM	SM
C	**66**	33					
YF	5	**75**	20				
AF			**75**	25			
SF		4	20	**75**			
YM					**85**	15	
AM					15	**78**	5
SM				5	5	27	**61**

Table 5. Relative confusion matrix of the best GMM supervector-based SVM system (see text) on the *SpeechDat II* corpus; the columns contain the actual age (ac) and the rows contain the classified age (cl) (overall precision 77%)

7. REFERENCES

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, “Emotion Recognition in Human-Computer Interaction,” *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] N. Minematsu, K. Yamauchi, and K. Hirose, “Automatic estimation of perceptual age using speaker modeling techniques,” in *Proceedings Interspeech 2003*, Geneva, Switzerland, 2003, pp. 3005 – 3008.

[3] C. Müller, F. Wittig, and J. Baus, “Exploiting Speech for Recognizing Elderly Users to Respond to their Special Needs,” in *Proceedings Interspeech 2003*, Geneva, Switzerland, 2003, pp. 1305 – 1308.

[4] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Müller, R. Huber, B. Andrassy, J.G. Bauer, and B. Littel, “Comparison of Four Approaches to Age and Gender Recognition for Telephone Applications,” in *ICASSP 2007 Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, Hawai’i, USA, 2007, vol. 4, pp. 1089 – 1092.

[5] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, pp. 19–41, 2000.

[6] Douglas A. Reynolds, “An Overview of Automatic Speaker Recognition Technology,” in *ICASSP 2002 Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Orlanda, Florida, USA, 2002, vol. 4, pp. 4072–4075.

[7] W. M. Campbell, D. E. Sturim, and D. A. Reynolds, “Support Vector Machines Using GMM Supervectors for Speaker Verification,” *Signal Processing Letters, IEEE*, vol. 13, pp. 308–311, 2006.

[8] A. Dempster, N. Laird, and D. Rubin, “Maximum Likelihood from Incomplete Data via the EM Algorithm,” *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[9] J.L. Gauvain and C.H. Lee, “Maximum A-Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains,” *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 291–298, 1994.

[10] C. J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[11] R. Dehak, N. Dehak, P. Kenny, and P. Dumouchel, “Linear and Non Linear Kernel GMM Support Vector Machines for Speaker Verification,” in *Proceedings Interspeech 2007*, Antwerp, Belgium, 2007.

[12] C. J. van Rijsbergen, *INFORMATION RETRIEVAL*, Butterworths, London, 2nd edition, 1979.