Age Determination of Children in Preschool and Primary School Age with GMM-based Supervectors and Support Vector Machines/Regression

Tobias Bocklet, Andreas Maier, Elmar Nöth

Institute of Pattern Recognition, *University of Erlangen-Nuremberg*, Germany

Abstract. This paper focuses on the automatic determination of the age of children in preschool and primary school age. For each child a *Gaussian Mixture Model* (GMM) is trained. As training method the *Maximum A Posteriori* adaptation (MAP) is used. MAP derives the speaker models from a *Universal Background Model* (UBM) and does not perform an independent parameter estimation. The means of each GMM are extracted and concatenated, which results in a so-called GMM supervector. These supervectors are then used as meta features for classification with *Support Vector Machines* (SVM) or for *Support Vector Regression* (SVR). With the classification system a precision of 83 % was achieved and a recall of 66 %. When the regression system was used to determine the age in years, a mean error of 0.8 years and a maximal error of 3 years was obtained. A regression with a monthly accuracy brought similar results.

1 Introduction

Automatic speech recognition of children's speech is much more sophisticated than speech recognition on adult's speech. This effect is often amplified by the lack of training data. Nevertheless, some approaches exist which try to compensate this drawback [1]. One remaining problem is the strong anatomic alteration of the vocal tract of children within a short period of time. An idea to face this problem is to use different acoustic models for different age classes of children [2]. The most appropriate acoustic model has to be selected before the automatic speech recognition can be performed. If the age of a child is not known a priori, the age has to be estimated out of the child's voice. This work focuses on this task.

In [3] it has been shown, that age recognition with SVM and 7 gender dependent classes outperforms different other classification ideas. The classification results of the SVM idea were in the same range as humans, and the precision even better. In this work we pick up the SVM idea and apply it to children in preschool and primary school age. As the next step we substitute the SVM in our recognition system by the SVR and are able to predict the age directly, i.e. no distinction into classes has to be done.

The outline of this paper is as follows: Section 2 describes the used corpora on which the systems are trained and tested. Section 3 gives a brief introduction into the GMM supervector idea, describes the employed cepstral features and depicts how the speaker models (GMMs) are trained. The basic ideas of SVM and SVR are summarized in Section 4 and Section 5. Section 6 shows the training and working sequence of the created systems. The results are summarized in Section 7.

2 Corpora

All children read the so-called PLAKSS test [4]. The test consists of 99 words. These words contain all phonemes/sounds of the German language and the most important conjunctions of them in different positions, i.e at the beginning, the end or within a word. The words are shown on a screen and are represented by pictograms. If the children are able to read they can used the shown word, otherwise they have to name the shown pictograms. Thus the PLAKSS text is very qualified for children in preschool and primary school age, which mostly are not able to read. All data was recorded with a system developed in our lab: the the PEAKS system [5].
For training and testing three different datasets have been available. One contains recordings of 38 children in preschool age. The data was recorded in a local preschool. The mean age of the children of this dataset is 5.7 ± 0.7 years.
The second dataset contains data of children in primary school age with a mean age 8.5 ± 1.4 years. The recordings took place in a local elementary school in August 2006 and April 2007. All in all 177 children were recorded in this school. The third set of recordings was collected in an elementary school in Hannover (April 2007). 128 children have been recorded with a mean age of 8.6 ± 1.1 years. A subset of each of these recordings was used for the experiments described in this paper. A total amount of 212 children has been chosen from all three different corpora to train the two systems. The children have been selected in order to create an almost age-balanced training set with respect to the following five classes:

- $<$ 7 years
- 7 years
- 8 years
- 9+10 years
- $>$ 10 years

The mean age of the training subset was 8.3 ± 2.4 years. The system was tested on a 100 speaker subset. Again the children have been selected to create an age-balanced test set. The average age in the test set was 8.4 ± 2.3.

3 GMM Supervector Idea - Metafeatures for the SVM

The creation of the GMM supervectors is shortly described in this section. The idea was first published in [6]. First feature vectors are extracted out of the speech signal. These features are then employed to train GMMs with M Gaussian densities [7]. Each GMM represents a different child. The mean values of each GMM are concatenated and used as meta features in SVMs. Each supervector is then labeled with the correct age. So each child it represented by one supervector. The principle of GMM supervector creation is displayed in Fig. 1.

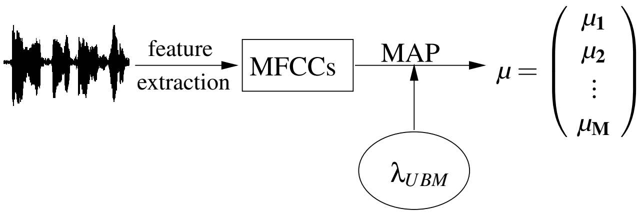
feature
extraction
MFCCs
MAP
$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_M \end{pmatrix}$
λ_{UBM}

Fig. 1. Training principle of GMM supervectors. After feature extraction the speaker model (GMM) is created. The concatenation of its mean vectors creates the GMM supervector. For details see [6].

3.1 Feature Extraction

As features the commonly used *Mel Frequency Cepstrum Coefficients* (MFCCs) are used. They examine the 18 Mel-bands and consider a short time window of 16 ms with a time shift of 10 ms. The first 12 MFCCs are used and the first order derivatives are computed by a regression line over 5 consecutive frames. Th final feature vector has 24 components (log energy, MFCC(1)-(11) and the first derivatives).

3.2 Gaussian Mixture Models

The basic idea behind the GMM supervector approach is to model every speaker with a different GMM, which is a standard approach in speaker identification/verification [8]. A GMM is composed of M unimodal Gaussian densities:

$$\begin{aligned} p(\boldsymbol{c}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \sum_{i=1}^{M} \omega_i p_i(\boldsymbol{c}|\boldsymbol{\mu_i},\boldsymbol{\Sigma_i}) && (1) \\ &= \sum_{i=1}^{M} \omega_i \cdot \frac{1}{(2\pi)^{D/2}|\Sigma_i|(1/2)} e^{-(1/2)(\boldsymbol{c}-\boldsymbol{\mu_i})^T \Sigma_i^{-1}(\boldsymbol{c}-\boldsymbol{\mu_i})}, && (2) \end{aligned}$$

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.

3.3 Training

After extraction of the MFCCs a UBM is created with all the available training data, using the EM algorithm [9]. The UBM is then employed as an initial model either for an EM training or for the MAP adaptation [10]. The MAP adaptation adapts the UBM to the speaker dependent training data, so that for every child a GMM is created. MAP adaptation calculates the Gaussian mixture components by a single iteration step and combines them with the UBM parameters. The number of EM iterations for the UBM training was set to 10.

4 Classification with Support Vector Machines

The SVM [11] performs a binary classification $y \in (-1, 1)$ based on a hyperplane separation. The separator is chosen in order to maximize the distances (margin) between the hyperplane that separates the two classes 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:

$$f(\boldsymbol{x}) = \sum_{i=1}^{L} \lambda_i y_i K(\boldsymbol{x}, \boldsymbol{x_i}) + d \tag{3}$$

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. One task of this paper is a 5-class age classification. So the binary SVM has to be extended. The simplest way is to separate each age class from all others. Therefore $N \times (N-1)/2$ classifiers are created, each of them separating two classes. The scores of these classifiers are then combined.

5 Prediction with Support Vector Regression

The general idea of regression is to use the vectors of the training set to approximate a function, which tries to predict the target value of a given vector of the test set. SVMs can also be used for regression. The method is then called SVR. A detailed description is given in [12].

Due to the fact, that no binary classification has to be performed, the so called "ε-tube" is defined. ε describes the deviation, which is allowed between the training vectors and the regression line (in positive and negative direction). Similar to the classification task, not all training vectors are needed to select the most appropriate ε-tube, but only a subset of them, i.e the support vectors. These vectors lie outside the ε-tube. Equal to the SVM labels y_i are needed for each training vector $\boldsymbol{x_i}$. In this case y_i denotes the target value, i.e. the age of a given child. The goal of SVR training is to find a function $f(\boldsymbol{x_i}) = \boldsymbol{x_i} \cdot \boldsymbol{w} + b$, that has at most ε deviation and is as flat as possible. Flatness in this case means to gather a small $\boldsymbol{w}$. This can be formulated as an optimization problem, where the norm ($||\boldsymbol{w}||$) has to be minimized. By introducing Lagrangian multipliers the optimization problem can be formulated and rewritten in a dual optimization problem. The prediction of the target value $\hat{y}$ of a given test vector $\boldsymbol{x}$ is then determined by

$$\hat{y} = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)(\boldsymbol{x_i} \cdot \boldsymbol{x}) + b. \tag{4}$$

α_i and α_i^* are Lagrangian multipliers and $\boldsymbol{x_i}$ denotes the training vectors. To extend the SVR to the non-linear case $\boldsymbol{x_i} \cdot \boldsymbol{x}$ again can be extended by kernel functions $K(\boldsymbol{x_i}, \boldsymbol{x})$.

6 Classification and Regression System

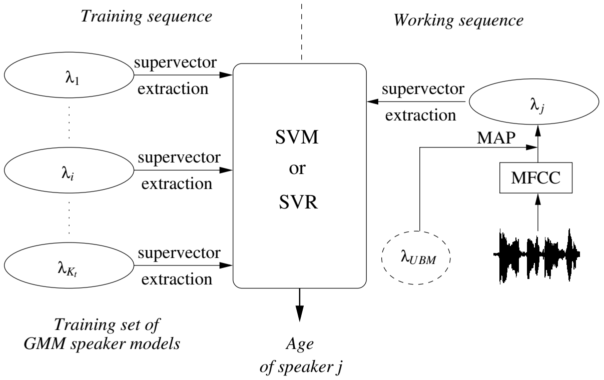

Fig. 2. Principle of the classification (SVM) and regression (SVR) system.

Fig. 2 shows the principle of the training and the working sequence of both systems. For the classification task a SVM is used. For the regression task the SVM is replaced by a SVR, which facilitates the possibility to predict the real age of a child in detail. For each speaker of the training set a GMM λ is created. The mean vectors of each GMM are concatenated and used as input vectors for the SVM/SVR training. These are denoted as GMM supervectors and can be regarded as a mapping from the utterance of a speaker (the MFCCs) to a higher-dimensional feature vector, which represents the speaker himself or characteristics of this speaker, e.g. the age. If the age (or the group membership) of a speaker j should be determined, a speaker model λ_j is derived out of the background model λ_{UBM}. Note that the background model is the same one, which is used for creating the GMMs of the training speakers. Again the GMM supervector is created for speaker j and the SVM or SVR is performed. The SVM determines the expected class membership of speaker j and the SVR directly predicts his or her age. The SVM and SVR implementations of the WEKA toolkit [13] were used.

7 Experiments and Results

This paper focuses on the automatic determination of the age of children. We performed preliminary experiments which focused on selection of the best parameters for the speaker model creation. Another set of preliminary experiments were performed to determine the best kernel for the SVM/SVR system and the combination of GMM supervectors. The results of the SVM system are described in Section 7.2 and the results of the SVR system are depicted in Section 7.3.

7.1 Preliminary Experiments

We examined the influence of the number of Gaussian densities, the training algorithm (MAP, EM) and the influence of the different forms of covariance matrices (full or diagonal) on the recognition results. In case of MAP adaptation we evaluated the use of different adaptation parameters (for details see [8]) and the kind of training, i.e. adapting all GMM parameters ($\omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}$) or only the means respectively. The best results were achieved with GMMs with MAP adaptation, where 256 Gaussian densities with full covariance matrices were adapted. The MAP adaptation parameter α was only of minor importance for the recognition results. We achieved comparable results for different values of α.

The second set of experiments focused on the determination of the most appropriate kernel for SVR/SVM. We regarded linear, exponential and RBF kernels. We varied the polynomial order of the exponential kernel and the width of the radial basis function for the RBF kernel. Besides these standard kernels we also examined a GMM-based distance kernel, which is based on the KL divergence. This kernel is often used in speaker verification applications [14]. The best results were achieved with the most simple one – the linear kernel.

The composition of the GMM supervector was also a task in the preliminary experiments. We focused on using other information than the mean vectors in it. So we additionally added the weights of the GMM components and the diagonal of the covariance matrices to the supervector. This results in a higher dimension, but not in a higher recognition result.

7.2 Classification Results

For the classification task the children were assigned to five different age classes (Section 2). The confusion matrix is shown in Table 1. It can be seen that all children, who belong to the class *9+10* and *>10* are classified correctly. However only 6 % children (in total 1 of 16 children) of the age of 8 are classified correctly. The recognition results of children with the age of 7 or smaller are classified correctly with about 60 %. The overall recall of the classification system was 66 % and the overall precision was 83 %. This has been calculated by an unweighted mean of the class-wise precision and recall.

	<7	7	8	9+10	>10
<7	**60**	33		7	
7	5	**55**		35	5
8		20	**6**	47	27
9+10			3	**97**	
>10					**100**

Table 1. Relative confusion matrix of the SVM system; The y-axis determines the actual class and the x-axis determines the recognized class.

	<7	7	8	9+10	>10
<7	**66**	20	13		
7	10	**55**	15	20	
8	6	13	**40**	40	
9+10			25	**74**	
>10				20	**80**

Table 2. Relative confusion matrix of the SVR system. After the age is determined the children are assigned to the specific class

7.3 Regression Results

The advantage of the regression system is, that the children do not have to be assigned to different age classes. The determination of the age is performed in two different ways: with an accuracy in years and an accuracy in months.
When meassuring the accuracy in years, the maximum deviation is three years and the mean deviation was determined to be 0.8 years. We calculated the Pearson and the Spearman correlation: 0.87 and 0.83 respectively. Additionally we assigned the recognized age to the different age classes which were used for the SVM system. So we were able to build up a confusion matrix and to compare the SVM and SVR results. The confusion matrix of the SVR system is shown in Table 2. The overall precision of this approach was 68 % and the recall was 63 % respectively. The recall of the SVM and the SVR system are identically, but the precision of the SVM is much higher. Nevertheless, the confusion matrix of the SVR system is more balanced and way more intuitive, i.e. 40 % of the 8 year old children are classified correctly and the results for children with nine years or older are in the same range as the results for the other classes.

The second regression experiment regarded the age determination on a monthly accuracy level. Therefore the labels of the GMM supervectors have been changed. The results are shown in Table 3. The age of 65 children lies within a deviation of 12 month. The mean deviation for this experiment was 10.3 months. The maximal error was 35 months. The Pearson correlation was measured to be 0.89 and the Spearman correlation 0.83 %.

8 Outlook and Conclusion

This paper focused on the automatic determination of the age of children in preschool and primary school age. Therefore, we created GMM supervectors for every child of the training and test set and either performed a classification with SVM or a SV regression. The classification approach had problems in determining children with the age of eight years. This lack could be improved by application of the SVR and a class determination afterwards. But this system

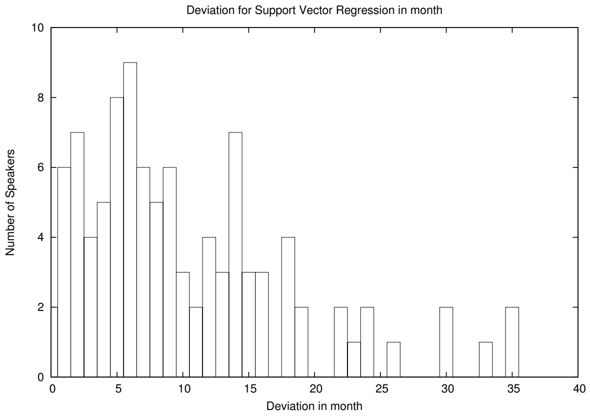

Fig. 3. Deviation between the age in months and the predicted age in months based on SVR.

was not as precise as the SVM system.
The regression could also be performed with a monthly accuracy, which means the training labels described the age of a child in month. With this idea the age of 65 % of the children was determined with a deviation of less than 12 months. One field of application for the age detection of children is to select a proper acoustic model for an adjacent speech recognition. Another idea would be to estimate or to assess whether the acoustic development of a child is appropriate for his or her age or not. But therefore a training set has to be created, which contains for each age appropriate reference speakers. Future work will focus on this task.

References

1. G. Stemmer, C. Hacker, S. Steidl, and E. Nöth, "Acoustic Normalization of Children's Speech," in *Interspeech Proceedings 2003*, Geneva, Switzerland, 2003, pp. 1313–1316.
2. T. Cincarek, I. Shindo, T. Toda, H. Saruwatari, and K. Shikano, "Development of Preschool Children Subsystem for ASR and QA in a Real-Environment Speech-oriented Guidance Task," in *Proceedings Interspeech 2007*, 2007, pp. 1469–1472.
3. T.Bocklet, A.Maier, J.G. Bauer, F. Burkhardt, and E. Nöth, "Age and Gender Recognition for Telephone Applications based on GMM Supervectors and Support Vector Machines," in *ICASSP 2008 Proceedings, IEEE International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, Nevada, USA, 2008, to appear.
4. A. V. Fox, *PLAKSS - Psycholinguistische Analyse kindlicher Sprechstörungen*, Swets & Zeitlinger, Frankfurt a.M., 2002.
5. A. Maier, E. Nöth, A. Batliner, E. Nkenke, and M. Schuster, "Fully Automatic Assessment of Speech of Children with Cleft Lip and Palate," *Informatica*, vol. 30, no. 4, pp. 477–482, 2006.

6. 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.
7. T. Bocklet, A. Maier, and E. Nöth, “Text-independent Speaker Identification using Temporal Patterns,” in *10th International Conf. on Text, Speech and Dialogue (TSD)*, V. Matoušek and P. Mautner, Eds., Berlin, Heidelberg, New York, 2007, vol. 4629 of *Lecture Notes in Artificial Intelligence*, pp. 318–325, Springer.
8. Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, “Speaker Verification using Adapted Gaussian Mixture Models,” *Digital Signal Processing*, pp. 19–41, 2000.
9. 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.
10. 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.
11. 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.
12. A.J. Smola and B. Schölkopf, “A tutorial on support vector regression,” *Statistics and Computing*, vol. 14, pp. 199–222, 2004.
13. I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, “Weka: Practical machine learning tools and techniques with java implementations,” 1999.
14. 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.