Computer Speech and Language 27 (2013) 135–150

Speaker state recognition using an HMM-based feature extraction method☆

R. Gajšek*, F. Mihelič, S. Dobrišek

University of Ljubljana, Faculty of Electrical Engineering, Tržaška Cesta 25, 1000 Ljubljana, Slovenia

Received 1 May 2011; received in revised form 29 December 2011; accepted 18 January 2012
Available online 2 February 2012

Abstract

In this article we present an efficient approach to modeling the acoustic features for the tasks of recognizing various paralinguistic phenomena. Instead of the standard scheme of adapting the Universal Background Model (UBM), represented by the Gaussian Mixture Model (GMM), normally used to model the frame-level acoustic features, we propose to represent the UBM by building a monophone-based Hidden Markov Model (HMM). We present two approaches: transforming the monophone-based segmented HMM–UBM to a GMM–UBM and proceeding with the standard adaptation scheme, or to perform the adaptation directly on the HMM–UBM. Both approaches give superior results than the standard adaptation scheme (GMM–UBM) in both the emotion recognition task and the alcohol detection task. Furthermore, with the proposed method we were able to achieve better results than the current state-of-the-art systems in both tasks.

Keywords: Emotion recognition; Intoxication recognition; Hidden Markov Models; Universal Background Model; Model adaptation

1. Introduction

Augmenting a human–computer interaction (HCI) system with various paralinguistic recognition capabilities has recently gained a lot of attention from the speech processing community. As stated by Cowie et al. (2001), the speech communication between humans can be split into two channels, one transmitting the explicit information, and the other transmitting the implicit information. The explicit information channel represents "what" is being said and has been studied for years with the development of speech recognition systems. The implicit channel represents "how" it is being said and consists of different phenomena such as emotions, gender, age, stress, identity, etc. While some (identity, gender) have been studied in the past, others such as emotions or stress, have been somewhat neglected. But in order to insure that the communication with the artificial systems is perceived by humans as natural, the implicit channel of communication needs to be incorporated as well. Furthermore, the system's dialog manager could benefit greatly from the added information about the user's state such as age, gender, emotional state, etc., when estimating the type of reaction to the user's command. In certain circumstances, even other knowledge about the speaker state

is desired, for instance in cars, where intoxication detection could be used to prevent a person under the influence of alcohol from driving.

In this article we focus on the machine analysis of the speaker states. Specifically, we develop an efficient method for modeling the acoustic features in order to reliably recognize different speaker states. The applicability of our method for speaker state recognition tasks is shown by comparing our system's performance with the state-of-the-art systems and the reported results in two different recognition tasks. Firstly, the FAU Aibo Emotion Corpus (FAU-Aibo) (Steidl, 2009; Batliner et al., 2008) is used to assess the performance of our proposed method for recognizing emotional states. The database was used for the Interspeech 2009 Emotion Challenge and therefore reported results, using different approaches, exist (Schuller et al., in press). Secondly, the VINDAT database (Mihelič et al., 2003) consisting of speech under different levels of alcohol intoxication is used to further evaluate the proposed method. The speech diversity found in the selected databases (children vs. adults, spontaneous vs. scripted) further assures that a robust and reliable assessment can be made.

In various speaker state recognition tasks, the difference between the systems can be seen primarily in terms of the modeling approach, which is used to represent the given audio features. Schuller et al. (2009) classify the existing techniques into two classes: (i) frame-level modeling techniques, which build statistical models of feature vectors extracted from overlapping frames of a given utterance, and (ii) supra-segmental modeling techniques, where a number of statistical functionals are applied to the frame-level features of a particular utterance or a chunk of audio, yielding a single, high-dimensional feature vector per utterance. The low level acoustic features for both types of modeling techniques typically consist of spectral, prosodic and voice quality features (Busso et al., 2009; Schuller et al., 2007; Batliner et al., 2011). In the article we focus mainly on modeling the spectral features, specifically the Mel Frequency Cepstral Coefficients (MFCC), known for being one of the most prominent feature types in emotion recognition, as shown by Batliner et al. (2011) and Schuller et al. (2007). Furthermore, the results from the Interspeech 2009 Emotion Challenge (Schuller et al., 2009) and the Interspeech 2010 Paralinguistic Challenge (Schuller et al., 2010) show that systems using the MFCCs were producing the best results (Schuller et al., in press).

In the paper we will deal with the frame-level modeling approach, but experiments will be conducted using the supra-segmental modeling as well, in order to compare them. The Hidden Markov Model (HMM) and the Gaussian Mixture Model (GMM) are normally utilized for modeling the frame-level features. Besides the basic speech recognition task, where the HMM is predominantly used, the modeling technique based on the GMM is widely utilized in numerous areas of speech analysis, such as speaker recognition/verification (Reynolds et al., 2000), language identification (Burget et al., 2006), speech/non-speech detection (Žibert et al., 2006), and others. Recently, this type of modeling has been successfully applied to various paralinguistic tasks as well. State-of-the-art results were reported for the tasks of age and gender recognition (Kockmann et al., 2010), emotion recognition (Kockmann and Burget, Černocký, 2009; Dumouchel et al., 2009; Gajšek, Štruc and Mihelič, 2010), and affect recognition (Gajšek et al., 2010). The GMM is incorporated into a modeling scheme, either as a representation of a selected feature distribution for a particular class or as the Universal Background Model (UBM), characterizing the general feature space. In the first method, the GMM is trained for every class in the training data, either using only the data from the corresponding class, or in the case of discriminative training, the data from the corresponding class as positive samples and the rest of the training data as negative samples. In the second method, the UBM is trained with the objective to represent a general distribution of the features, regardless of the class or the particular recognition task at hand. Consequently, it can be used as a reference for how particular classes or individual samples are distributed in the feature space. Furthermore, it represents a starting point for calculating the distribution of a particular speech utterance, which is usually achieved by the Maximum A Posteriori (MAP) adaptation of the UBM. Reynolds et al. (2000) first introduced the approach of the UBM adaptation to tackle the speaker verification problem, where a likelihood function was realized using the GMM. Additionally, the authors state that no advantage was shown in the text-independent tasks if a more complex likelihood function (e.g. HMM) was selected. In this paper we present a modeling scheme where an improvement over the standard GMM–UBM approach can be achieved, even for the text-independent recognition tasks. The first novelty of our proposed method is to construct the GMM–UBM from the HMM. Instead of determining the clusters of the model purely statistically, such as with *K*-means clustering or the Linde–Buso–Gray (LBG) method (Linde et al., 1980), the training data, or only a part of the training data, can be used to train a monophone-based HMM. In Section 2.3 we present a transformation of the HMM to a GMM, which can be later used in a GMM–UBM adaptation scheme, following the same procedure as for the standard GMM adaptation (Reynolds et al., 2000). The segmented way of constructing the UBM based on the monophone HMM has the advantage of having the mixtures of the UBM model represent different monophones. Thus,

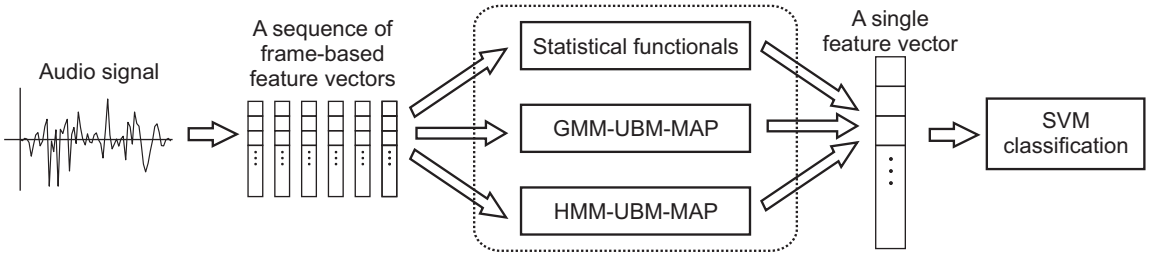

Fig. 1. Comparison of different approaches to modeling the acoustic features.

in the adaptation step the difference between the adaptation data (usually a single utterance) and the UBM is captured on a monophone level, with the MAP criteria ensuring that only mixtures, representing the monophones observed in the adaptation data, are updated. It is essential for the training of the monophone HMM that transcriptions are available for the training data, or that they can be produced by a speech recognition system. If transcriptions are available or can be acquired using a speech recognition system, we propose a method of adapting the HMM using the Maximum A Posteriori (MAP) criteria, normally used for speaker adaptation of the HMM (Lee and Gauvain, 1993). As we show in Section 4, when employing our proposed modifications, we were able to consistently outperform the standard GMM adaptation systems, as well as exceed the best reported results in the literature.

In Section 3 we present the FAU-Aibo and the VINDAT databases, which are used to assess the performance of our proposed method against the current state-of-the-art systems. Section 2 presents our proposed method of the HMM adaptation by first giving an overview of the well-known and widely-used MAP adaptation of the GMM, and later describing our proposed method of the HMM adaptation for speaker state recognition tasks. In Section 4, the conducted experiments are presented and the obtained results are given and discussed. Finally, our concluding remarks are given in Section 5.

2. Speaker state feature modeling

In this section our method of using the adapted HMM for speaker state recognition is presented. The goal of our proposed approach is similar to the supra-segmental method of calculating statistical functionals and to the method of adapting the GMM–UBM, which is to produce a new feature vector (referred to throughout the paper as the super-vector) containing as much discriminative information between the classes as possible. The steps shared between the described methods are the selection and extraction of the frame-level acoustic features and the classification step, as depicted in Fig. 1. In the following sub-sections, firstly, the acoustic front-end is presented, secondly, we give an overview of a well-known adaptation of the GMM, and finally, we describe the improvements we propose in the HMM-based adaptation and give the mathematical formulae used in the proposed HMM adaptation scheme for speaker state recognition tasks.

2.1. Acoustic front-end

The most frequently used acoustic features for building the GMMs and the HMMs are the Mel Frequency Cepstral Coefficients (MFCCs) (Pols, 1966; Davis and Mermelstein, 1980). Even though they were primarily developed for speech recognition and should, as such, be independent of the speaker, the speaker's state, gender, age and other paralinguistic information, they were successfully applied to a number of tasks of this kind (Bocklet et al., 2010; Feld et al., 2010; Batliner et al., 2011; Schuller et al., 2007, 2009d, 2009).

In our experiments, 0–12 MFCCs were used, calculated on 20 ms frames, with a delay of 10 ms between the frames. The filter bank consisted of 22 Mel-spaced filters. After the logarithm of the energy in every filter is calculated, the coefficients are decorrelated by applying the Discrete Cosines Transform. The 0th cepstral coefficient corresponds to the energy of the frame. In the majority of the tests, we used the MFCCs produced by the openSMILE feature extractor (Eyben et al., 2010), which is a part of the open-source Emotion and Affect Recognition (openEAR) toolkit (Eyben et al., 2009). When building the HMM acoustic model we employed the HTK toolkit's implementation of the MFCC extraction (Young et al., 2009). Decision was made purely for practical reasons since we had the framework for

Table 1
List of individual acoustic features (low level descriptors – LLD) used for the baseline and, in part, for our proposed system. The feature sets were defined by Schuller et al. (2011).

Spectral LLD
RASTA-style filtered auditory spectrum, bands 1–26 (0–8 kHz)
MFCC 1–12
Spectral energy 25–650 Hz, 1–4 kHz
Spectral roll off point 0.25, 0.50, 0.75, 0.90
Spectral flux, entropy, variance, skewness, kurtosis, slope

Energy related LLD	Voice related LLD
Sum of auditory spectrum (loudness)	F0
Sum of RASTA-style filtered auditory spectrum	Probability of voicing
RMS energy	Jitter (local, delta)
Zero-crossing rate	Shimmer (local)

running the HMM-based speech recognition engine available from our previous work. It should also be noted that the parameters of the extraction were set identically and therefore the difference in the MFCC features should be minor. In all experiments, the first order delta coefficients (ΔMFCCs) were added to the base MFCC features, yielding a final feature vector of size 26.

The second set of features used in our HMM-based experiments consists of pitch (F0), probability of voicing, Jitter (local, delta) and Shimmer (local). This set is identical to the voice related low level descriptors (LLDs) selected for the Interspeech 2011 Speaker State Challenge, presented in Schuller et al. (2011). Again, the openSMILE tool was used to calculate the frame-level features and the same extraction setup was used as for the Challenge, in order to enable a comparison between the different modeling techniques.

For the current state-of-the-art recognition system in the alcoholization recognition task we used the Interspeech 2011 Speaker State Challenge baseline system. As defined in Schuller et al. (2011) and shown in Table 1, it is based on three sets of acoustic features: 50 spectral LLDs (including 1-12 MFCCs), 4 energy related LLDs, and 5 voice related LLDs.

2.2. GMM-based speaker state feature modeling

The procedure of using the adapted GMMs in classification tasks consists of (i) the estimation of the GMM model from the training data, called the Universal Background Model (UBM), (ii) the adaptation of the UBM for every training and test sample, using only the corresponding sample's data, (iii) combining the parameters of each adapted model in a vector, and (iv) the use of these vectors as an input for the selected classification method.

Formally, the GMM is defined as a linear combination of several multivariate Gaussian probability density functions (PDFs), i.e,

$$p(\mathbf{x}|\theta) = \sum_{i=1}^{M} w_i\, p_i(\mathbf{x}) \tag{1}$$

where w_i denotes the weight associated with the ith Gaussian PDF $p_i(\mathbf{x})$

$$p_i(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right\}. \tag{2}$$

In the above equations, $\boldsymbol{\mu}_i$ denotes the mean vector of the ith Gaussian PDF, $\boldsymbol{\Sigma}_i$ denotes the covariance matrix of the ith Gaussian PDF, d stands for the dimensionality of the PDF, and $\theta = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, w_i\}$, for $i = 1, 2, \ldots, M$, represents the set of GMM parameters. Note that the M-component GMM is fully characterized by the values of its parameters

θ. There is also a constraint that all GMM weights must be between 0 and 1, and should sum to 1, as defined in Eq. (3).

$$\forall i : 0 \leq w_i \leq 1, \quad \sum_{i=1}^{M} w_i = 1 \tag{3}$$

Even though full covariance matrices are supported by the general GMM, diagonal covariance matrices are normally used since they are computationally more efficient. Also, the higher order GMM with diagonal covariance can equally well model arbitrary densities as a lower order GMM with full covariance matrices.

The parameters of the GMM are estimated via the Expectation–Maximization (EM) algorithm (Dempster et al., 1977), where the parameters are iteratively updated, satisfying the criteria of maximizing the likelihood of the model, given the training data. Once the parameter convergence is achieved the iterative process is finished. However, during the estimation process, the magnitude of the elements in the covariance matrices can become too small and cause numerical problems. Therefore, a floor or a minimum value, to which an element in the covariance matrix can be updated to, is usually set. Secondly, the EM algorithm is highly sensitive to the initial values of the parameters in the GMM. If the initial values are far away from the actual optimum, the local maximum, that the EM algorithm will iterate to, can still poorly represent the distribution of the training data. The *K*-means clustering or the LBG method are often used for the initialization of the GMM.

In the classification procedure based on the adapted GMMs, the UBM is estimated first. The UBM, represented as the GMM, is estimated using the data in the training set, disregarding the information about the classes, and as such represents a distribution of the features, which is independent of a specific recognition task. There are many approaches to constructing the final UBM, e.g., combining gender-specific GMMs or splitting the training features according to the acoustic environment, training the GMM for every subset, and then combining different GMMs to produce the final UBM (Reynolds et al., 2000). However, in our own experiments we did not observe any improvements over the simplest approach of estimating the GMM using all the training data. Once the UBM is estimated via the EM algorithm, it symbolizes the general acoustic space and is independent of any specific speaker state, which is the focus of our recognition task.

The next step is to adapt the UBM for every utterance from the training and test sets. The Maximum A Posteriori (MAP) estimation criterion is used to adapt the UBM to the utterance-specific GMM.

The mean of the utterance-specific GMM, concatenated into a vector, represents a new feature vector. While the GMM for a given test utterance could also be calculated directly from the set of feature vectors extracted from the utterance, adapting the UBM has three important advantages: (i) it ensures that the ordering of the GMM parameters in θ is the same as in the UBM for each computed GMM; (ii) it compensates for the insufficient amount of data in the given utterance; and (iii) it incorporates domain-specific knowledge into the computed GMM.

When computing an adapted GMM from the UBM, the first step is to determine the probabilistic alignment of a particular sample $Pr_i(\mathbf{x}_t)$ against all M UBM components, as follows:

$$Pr_i(\mathbf{x}_t) = \frac{w_i p_i(\mathbf{x}_t)}{\sum_{j=1}^{M} w_j p_j(\mathbf{x}_t)}, \tag{4}$$

where $p_i(\mathbf{x}_t)$ again denotes the Gaussian probability density function of feature vector $\mathbf{x}_t$ for the ith component of the GMM, t denotes the feature vector index with $t = 1, 2, \ldots, T$, T stands for the total number of the feature vectors extracted from the given utterance, and w_i represents the weight associated with the ith GMM component.

In the second step, the sufficient statistics for updating the mean feature vectors are computed. In general, the MAP estimation procedure updates the means, variances and weights of the GMM, but commonly the focus is only on updating the GMM's means. The statistics required for the MAP adaptation are

$$n_i = \sum_{t=1}^{T} Pr_i(\mathbf{x}_t), \tag{5}$$

and

$$\mathbf{E}_i(X) = \frac{1}{n_i} \sum_{t=1}^{T} Pr_i(\mathbf{x}_t)\mathbf{x}_t, \tag{6}$$

where n_i and $\mathbf{E}_i$ stand for the null and first order sufficient statistics, and X represents a set of frame-level feature vectors $X = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$.

So far the presented adaptation procedure is identical to the Expectation step when using the ML criterion in the EM algorithm. The difference compared to the ML-based procedure is shown in the Maximization step, where the updating rule becomes:

$$\hat{\boldsymbol{\mu}}_i = \alpha_i \mathbf{E}_i(X) + (1 - \alpha_i)\boldsymbol{\mu}_i, \tag{7}$$

as postulated by the MAP criterion. The adaptation parameter α_i, which controls the balance between the old values of the means and the new estimate, is computed as

$$\alpha_i = \frac{n_i}{n_i + \tau}, \tag{8}$$

where τ is the relevance factor, which is the same for all the components of the GMM. The value of the relevance factor is chosen experimentally and usually falls in the interval between 8 and 16 (Reynolds et al., 2000).

After sufficient iterations of the described procedure, the algorithm stops if the change in the component means is sufficiently small or a predefined number of iterations is reached. The size of this final GMM vector of means equals the dimension of the original feature vector, multiplied by the number of components of the GMM and therefore increases with the increase in the number of GMM components, i.e., M.

The vectors of means, produced by the MAP adaptation of the UBM per every utterance in the training and test sets, characterize the difference between the specific utterance and the "general" speech represented by the UBM. In our speaker state recognition tasks we hope that the specific speaker state traits we are trying to recognize will be differently represented in the adapted means. In order to evaluate if the adapted vectors of means are representative of the speaker's state, a classification method is employed. Support Vector Machines (SVMs) (Vapnik, 1998) are usually selected because of their recognition performance and their ability to handle feature vectors of large dimensions.

2.3. Transforming the HMM to the GMM–UBM

The procedure described in the previous section trains the UBM by pooling together all the training features, and thus discarding all the temporal information. Furthermore, the success of the EM algorithm in estimating the UBM depends heavily on the initialization, where some method of statistical clustering is employed. Even though the clusters or mixtures of Gaussians, which constitute the UBM, are derived via statistical clustering, they are often considered as representatives of some base "units" of speech. If the number of mixtures in the UBM is around 40–60, they represent monophones. In modern systems the number of mixtures exceeds 1000 and the representative base units become biphones or triphones.

The first improvement we propose in this article is to construct the GMM–UBM based on the modeling of the actual base units of speech, the monophones. If transcriptions are provided for a set of speech utterances, the procedure we propose consists of the following 2 steps: (i) train a simple HMM for every monophone, and (ii) combine Gaussian distributions from the HMM into the final GMM. The final GMM can then be used as the UBM for adaptation.

The HMM likelihood function for a speech sample represented by a set of short-time vector measurements $X = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ is defined as

$$p(X|\lambda) = \sum_{\mathbf{S}} \pi_s \prod_{t=1}^{T} a_{s_{t-1}s_t}\, p(\mathrm{x_t}|\boldsymbol{\theta}_{s_t}), \tag{9}$$

where π_s is the initial probability of the state s, $a_{s_{t-1}s_t}$ is the transition probability from state s_{t-1} to state s_t, $p(\mathrm{x_t}|\boldsymbol{\theta}_{s_t})$ is the observation density for state s_t (defined in Eq. 1) and the vector $\mathbf{S} = (s_1, \ldots, s_T)$ represents a state sequence. The summation in Eq. (9) is over all possible state sequences. The HMM parameters are represented as $\lambda = (\boldsymbol{\pi}, \boldsymbol{A}, \boldsymbol{\theta})$, where $\boldsymbol{\pi}$ is a vector of the initial state probabilities, $\boldsymbol{A} = [a_{ij}]$ is a state transition probability matrix, and $\boldsymbol{\theta}$ denotes the state observation density parameters. Once the HMM parameters are estimated using the iterative Baum–Welch algorithm (Baum et al., 1970) we propose a transformation of the HMM to the GMM. As described in Section 2.2 the GMM is defined by its vectors of means, covariance matrices and weights. The new GMM means and covariances can be easily obtained from the HMM parameters λ by taking the estimated densities from the HMM states. The weights for the new GMM are calculated based on the HMM state occupation statistics. These are calculated as a sum of the posterior

likelihoods of all the feature vectors for every state. This gives us a measure of how much training data is available for estimating the parameters of a particular state and thus seems like a reasonable measure for weighting the distributions from the HMM states. If Gaussians are shared between the states, the average occupation count for the states that use the particular Gaussian is calculated and used as a weight. In cases where we have more than one Gaussian per state the particular weight of a single Gaussian is multiplied with the corresponding state's occupation statistics in order to produce the final weight used for this Gaussian in the new GMM. The new weights are scaled to sum to 1, in order to satisfy the constraint of the GMM weights, defined in Eq. (3). The initial state probabilities $\boldsymbol{\pi}$ and the state transition probabilities A of the HMM are discarded and not considered in the transformation of the HMM to the GMM.

Once the trained HMM model is converted into the GMM, the latter is used as the UBM. Next, the same MAP adaptation and the SVM classification steps as described in Section 2.2, are taken in order to generate the predictions for the test utterances. As we will show in Section 4, the HMM-derived GMM–UBM gives a superior recognition accuracy compared to the standard way of GMM–UBM estimation. Using our approach, only transcriptions of the training set (or a part of it) are needed in order to estimate the HMM. Usually, a big corpus is required to estimate a triphone HMM with a large number of total Gaussian densities. However, in our case the monophone HMM with a small number of Gaussians is adequate for subsequent adaptation, and can thus be trained using much smaller amount of data.

2.4. HMM-based speaker state feature modeling

In the previous section we presented our approach of producing the GMM–UBM, constructed from the monophone-based HMM. Once the HMM is estimated using the Baum–Welch algorithm, the HMM is transformed into the GMM–UBM and a new GMM is produced for every utterance by the MAP adaptation of the means. The means are then concatenated into a vector and classified by employing a SVM classification. However, if transcriptions are available or if a speech recognition engine can be employed to produce them, the MAP adaptation for every utterance can be performed on the HMM. The HMM trained on all the training data is then called the HMM–UBM, and using the MAP adaptation, the new HMM is calculated from the UBM-HMM. Like with the MAP adaption of the GMMs, only the means of the Gaussian densities in the states are adapted. The means are then concatenated into a vector, again following a similar procedure as with the means of the adapted GMMs, described in the previous sections.

Following the notation in Section 2.3, the adaptation formula for the mean vector of the mixture component i in the HMM state s is

$$\hat{\boldsymbol{\mu}}_{si} = \alpha_{si}\mathbf{E}_{si}(X) + (1-\alpha_{si})\boldsymbol{\mu}_{si}^{UBM}, \tag{10}$$

where $\boldsymbol{\mu}_{si}^{UBM}$ is the mean of the HMM–UBM s state and ith mixture, $\mathbf{E}_{si}(x)$ is the mean of the observed adaptation data and α_{si} is the adaptation parameter corresponding to the α_i parameter in the GMM adaptation (Eq. 8). Similarly, it is defined as

$$\alpha_{si} = \frac{n_{si}}{n_{si}+\tau}, \tag{11}$$

where τ is the relevance factor and n_{si} is the occupation likelihood of the adaptation data, defined as

$$n_{si} = \sum_{t=1}^{T} Pr_{si}(t). \tag{12}$$

The mean of the observed adaptation data $\mathbf{E}_{si}(x)$ from Eq. (10) is calculated as

$$\mathbf{E}_{si}(X) = \frac{\sum_{t=1}^{T} Pr_{si}(t)\mathbf{x}_t}{n_{si}}, \tag{13}$$

where Pr_{si} denotes the probability of occupying the mixture i of state s at time t, and $\mathbf{x}_t$ is the adaptation data feature vector at time t. One can see that the MAP adaptation of the means of the HMM (Eqs. ((10)–(13))) is an extension of the MAP adaptation of the GMM (Eqs. ((4)–(8)) from Section 2.2), hence, if there is only one state in the HMM, the equations become identical.

The adaptation formula for means (Eq. (10)) is made up of two terms, the UBM value of mean $\boldsymbol{\mu}_{si}^{UBM}$ and the calculated mean of the observed adaptation data $\overline{\boldsymbol{\mu}}_{si}$, both weighted by the combination of the occupation likelihood n_{si}

and the relevance factor τ. The distance between the new MAP estimated mean and the UBM initial mean is determined by the occupation likelihood n_{si} for the particular component i of state s. The higher the value of n_{si}, the more influence the adaptation data has on the new adapted mean. And vise-versa, if the value of the n_{si} for the i component is small, the value of the mean vector will remain similar to the initial UBM mean. Likewise, the relevance factor τ controls the weighting of the prior mean and the influence of the adaptation data as well. However, it is independent of the adaptation data, and thus controls mostly the speed of the convergence. It should be noted that Eqs. (10)–(12) are run iteratively until either the predefined number of iterations is reached, or the difference between the new and the old means is smaller than a predefined threshold. Thus, the notation $\boldsymbol{\mu}_{si}^{UBM}$ is appropriate only for the first iteration, when the value of the mean vector from the UBM is used, and should be, in later iterations, better denoted by $\boldsymbol{\mu}_{si}$, symbolizing the current value of the mean being updated. However, the symbol $\boldsymbol{\mu}_{si}^{UBM}$ is left in the equations since it better represents the idea of adapting the same HMM–UBM for each particular utterance.

Although the MAP adaptation can be used to update all the parameters of the HMM–UBM, we leave the weights, covariances, transition probabilities and state probabilities intact. Hence, all utterance-specific information or differences from the "general" speech, represented by the HMM–UBM, is captured in the new vectors of means. In this way, the exploitation of the means as a classification feature is enabled.

Once the above-described adaptation of the HMM–UBM for all train and test samples is finished, a set of new HMMs is obtained. These HMMs share the same values for all the parameters, except for the means of the Gaussian densities in the states. Next, for every utterance-specific HMM, the mean vectors are pooled together and combined in a new super-vector. The size of the super-vector is the number of Gaussians in the HMM–UBM times the dimension of the front-end acoustic feature vector. The set of super-vectors is obtained for the training and the test sets. These sets represent a suitable input to the classification, which is in our case realized by the SVM classification, like in the case of the GMM adaptation. Hence, the improvement of our HMM–UBM–MAP against the GMM–UBM–MAP modeling can be easily demonstrated.

3. Databases

In order to reliably assess the performance of our proposed method, two databases focusing on different speaker states were selected. The FAU-Aibo corpus and the VINDAT corpus, selected for the evaluation, differ in terms of a number of characteristics. While the FAU-Aibo corpus covers children's speech, the VINDAT corpus contains speech from adults. Next, in the FAU-Aibo corpus, the children speak spontaneously, whereas in the VINDAT database the sentences are predefined. Finally, the databases concentrate on different speaker states with FAU-Aibo focusing on emotionally colored speech, and VINDAT on speech under the influence of alcohol.

3.1. FAU-Aibo

FAU-Aibo corpus (Steidl, 2009; Batliner et al., 2008) contains spontaneous children's speech labeled according to the emotional states being expressed in the recordings. The Wizard-of-Oz type of scenario was designed, where children utter commands to the little robot dog Aibo, while a person in the back is actually controlling the motion of the robot. This setup is effective for inducing an emotional response from the participant since the controller can intentionally disobey the participant's commands. In order to have the same behavior of the robot for all the participants, its movement and reactions to certain tasks were predefined. The session included five object localization tasks with children directing the robot to the defined spot, and the main task of leading the robot through a predefined path and at certain stops commanding the robot to perform a particular action. On one hand, the predefined and disobedient actions of the robot dog lead to frustration and other negative emotions, while the actions, compliant with the children's commands, induce positive emotions.

The FAU-Aibo database contains the recordings of 51 children between the ages of 10 and 13, of which 30 were female and 21 were male. In our experiments we follow a 2-class protocol of the Interspeech 2009 Emotion Challenge described in Schuller et al. (2009). The class of negative emotions (NEG) consists of emotional labels *angry*, *touchy*, *reprimanding* and *emphatic*, while the idle class (IDL) consists of all non-negative emotional states. The sessions took place at two schools and according to the location of the recording the corpus is split into the training (school #1) and test (school #2) sets. The number of utterances per class is shown in Table 2, where an imbalance towards the idle class

Table 2
Number of instances in (a) the FAU-Aibo and in (b) the VINDAT.

#	IDL	NEG	$\sum$
(a) FAU-Aibo 2-class: negative vs. idle			
Train	6601	3358	9959
Test	5792	2465	8257
$\sum$	12393	5823	18,216

#	NON-ALCO	ALCO	$\sum$
(b) VINDAT 2-class: non-alcoholized vs. alcoholized			
All	450	421	871

(IDL) can be observed. The number of samples for the negative class (NEG) is approximately half of the number of samples for the IDL class, which is usually the case in databases of spontaneous emotions.

3.2. VINDAT

The VINDAT database (Mihelič et al., 2003) contains the recordings of people speaking at different levels of alcohol intoxication. Ten Slovene speakers, five men and five women, took part in the recording sessions. The average age of the adults was 35 years. The recording session for each speaker consisted of two parts. For the first part, fourteen Slovenian words were selected based on their demanding pronunciation (for example, the Slovene translation of otorhynolaryngologist – "otorinolaringologinja"). The selected words formed the center of a sentence, with meaningful words added left and right, in order to avoid the changes in speed and intonation that are usually found at the beginning and the end of spoken utterances. In the second part, the speakers repeated sentences previously read to them by the operator in an attempt to record speech closer to the natural speaking style.

The participants were recorded in three sessions based on the amount of the consumed alcohol. The alcohol levels were measured by a hand-held indicator, usually employed by the police for the inspection of drivers on the roads. The device measures the level of intoxication as the amount of milligrams per liter of exhaled air (‰). In the first session, the participants were sober with 0‰). Before the second and the third session, each speaker consumed a selected amount of alcoholic beverage and measured the level of alcoholization prior to recording session. Understandably, at least 15 min passed since the last drink was consumed, before each measurement was made in accordance with the instructions of the alcohol-level indicator. Half of the participants exhibited 0.5‰) (which is a legal limit for driving in Slovenia) prior to the second recording, and by the last session, all participants had an alcohol level above 0.5‰).

For our task of alcoholization recognition, all the utterances labeled as less than 0.5‰) were assigned to the non-alcoholized class (NON-ALCO) and the rest (equal or more than 0.5‰) were assigned to the alcoholized class (ALCO). The threshold was set in accordance with the local legal limit for driving, as well as in accordance with the Intoxication sub-challenge in Schuller et al. (2011) and the comparative database of Schiel et al. (2011). In the lower part of Table 2, the distribution among the NON-ALCO and the ALCO classes is presented and it can be seen that the classes are almost balanced. While the text uttered in the database's recordings was predefined and should as such match the content of the recordings, after listening to all the utterances, approximately 10% were not in accordance with the proposed text. These exceptions were carefully corrected enabling the use of transcriptions in our HMM adaptation.

4. Experimental results

In this section we present the experimental comparison between our proposed method of HMM–UBM–MAP adaptation and other state-of-the-art systems. As described in Section 3, two databases were selected for assessment, based on the fact that they have different types of speech (scripted vs. spontaneous), different "type"/age of speakers (children vs. adults) and focusing on capturing different speaker states. By evaluating our proposed method in the

Table 3
Comparison of classification results on the FAU-Aibo database in terms of unweighted and weighted average recall (%).

System	Recall	
	UW	WA
Interspeech 2009 winner: Dumouchel et al. (2009)	70.3	68.7
Interspeech 2009 majority voting of 5 best systems	71.2	70.4

System	Acoustic features	# Gaussians	UW	WA
HMM to GMM–UBM	0–12 MFCCs + Δ	120	69.3	70.6
HMM–UBM–MAP	0–12 MFCCs + Δ	120	70.3	70.6
rec-HMM–UBM–MAP	0-12 MFCCs + Δ	120	69.7	70.1
HMM–UBM–MAP	0–12 MFCCs + Δ	30	70.3	70.2
HMM–UBM–MAP	5 Voice related + Δ	150	65.7	63.0
HMM–UBM–MAP	MFCCs + voice related		71.5	71.6

alcoholization detection task and in the task of recognizing emotional states, we hope to reliably show the applicability of the HMM-adaptation method in various tasks of speaker state analysis.

In order to reliably compare different modeling techniques, the same SVM classification setup was used in all the experiments. Once a set of super-vectors was produced, a support vector classifier with a linear kernel was trained using a sequential minimal optimization algorithm (SMO) (Platt, 1999; Keerthi et al., 2001), as implemented in the well-known and freely available Weka toolkit (Hall et al., 2009). The SVM setup was identical to the 2011 Interspeech Speaker State Challenge baseline system (Schuller et al., 2011) and no tuning of the SVM parameters was performed, neither on the training nor on test data sets. The identical SVM setup enabled us to honestly compare our HMM-based modeling of the frame-level features with the approach of computing statistical functionals. In the FAU-Aibo corpus, the emotional classes are imbalanced in favor of the IDL class. Therefore, prior to the classification with the SVM, the balancing of the training data was done using the SMOTE algorithm (Nitesh and Chawla, 2002). In the VINDAT database, the sizes of both classes are almost identical and no manipulation with the number of instances took place before the SVM classification.

The criteria by which we evaluate the performance of each individual system is the unweighted average recall (UW recall). In speaker state tasks, we usually have an imbalance between the classes being recognized. Therefore, the UW recall measure is appropriate since it is independent of the size of a particular class that is being recognized. Besides the UW recall, the weighted average recall (WA recall), representing the accuracy of a system, is reported for all the experiments as well.

4.1. FAU-Aibo experiments and results

In the FAU-Aibo corpus we follow the 2-class protocol of the Interspeech 2009 Emotion Challenge, defining the training set as recordings from one school (named in the corpus as “Ohm”) and the test set from another school (named in the corpus “Mont”).

The results from the Interspeech Challenge form the baseline against which we compare the effectiveness of our proposed method. The first and the third best individual contributions in the 2-class problem based their systems on the GMM–UBM–MAP adaptation scheme in combination with the MFCC features (Dumouchel et al., 2009; Kockmann and Burget, Černocký, 2009). Furthermore, they incorporate other sub-systems, either based on other acoustic features as Dumouchel et al. (2009), or based on other modeling techniques as Kockmann and Burget (Černocký, 2009) using a score level fusion. The system by Vlasenko and Wendemuth (2009), which reported the second best result used a similar approach of modeling MFCCs on the monophone level with the HMMs, hence it will be used as a comparison as well. The best single result of the Challenge in the 2-class problem was by Dumouchel et al. (2009), achieving a 70.3% UW recall. The organisers of the challenge improved the result with the majority voting fusion of the best 5 participating systems and reported a result of 71.2% UW recall. Both results are presented in the upper part of Table 3.

The aim of the first experiments was to establish the ability to use the HMM training as a means of acquiring the GMM–UBM. The availability of transcriptions enabled us to build a 3-state left to right monophone HMM, using the training data. The acoustic features used were 0–12 MFCCs and their deltas. The distribution in each state of the HMM was represented by a single Gaussian, and with 40 allophones used in the FAU-Aibo dictionary and 3 states per model, the initial HMM comprised of 120 Gaussians. According to the procedure described in Section 2.3, the HMM was then transformed into the GMM–UBM. Next, the standard GMM–UBM–MAP procedure, presented in Section 2.2, was used to generate a set of super-vectors. The dimension of the super-vector is equal to the number of Gaussians times the dimension of the acoustic feature vector, thus $120 \times 26 = 3120$. A relevance factor τ was set to 16 and 5 iterations of the MAP adaptation were run for each train and test utterance. Employing the SVM classification, a result of 69.3% UW recall was achieved, as presented in Table 3 – the system named "HMM to GMM–UBM". The achieved accuracy is lower than the best result from the Challenge, but still represents a competitive result. The emphasis here should be made on the set of acoustic features used, since this result was achieved based on only the MFCCs and their deltas, whereas other systems with similiar results used a broader range of acoustic features. It should also be noted here that this procedure requires the transcriptions to be available only for the training set, or even a subset of the whole training set, used to build the initial HMM. After the HMM is transformed to the GMM and during the MAP adaptation of the training and test utterances, no information about what is being said is used.

Next, we evaluated our proposed HMM–UBM–MAP adaptation scheme, presented in Section 2.4. Similar to the previous experiment, a 3-state monophone HMM was built on the training data, using the 26-dimension feature vector, consisting of MFCCs and their Δ. Instead of converting the HMM to the GMM and proceeding with the GMM–UBM–MAP method, the HMM is directly treated as the UBM and the adaptation is applied for every training and test utterance employing the same MAP criteria. In order to ensure comparability between the results the relevance factor was the same as in the GMM–UBM–MAP ($\tau = 16$), and after 5 iterations of MAP adaptation the final set of super-vectors was produced. Again, only one Gaussian per state was used, yielding a total of 120 Gaussians. Combined with the size of the acoustic feature vector (26), the dimension of the super-vector was 3120. The "HMM–UBM–MAP" system with the above-described parameters achieved an UW recall of 70.3%, equaling the best single result of the Interspeech 2009 Emotion Challenge. Again, it should be stressed that we were able to attain the same recognition accuracy, relying only on the MFCCs acoustic features and their Δ. However, the transcriptions for the train and the test utterances are necessary in order to use our proposed method. In the case of the FAU-Aibo, transcriptions were available for the whole corpus, but in real-life applications this is usually not the case. Therefore, we wanted to evaluate our proposed method of using adapted HMMs in sitution where accurate transcriptions for the test set would not be present. Discarding the corresponding transcriptions of the test set that were included in the corpus, we used a speech recognition system to obtain its recognized set of monophones for every utterance in the test set. Using the HTK toolkit's implementation of the Baum–Welch algorithem we built a monophone HMM with 3-states and a total number of 3000 Gaussians, employing the training set of the FAU-Aibo corpus. The 0–12 MFCCs, with their Δ and $\Delta\Delta$, as implemented in the HTK toolkit, comprised the feature vector for training the acoustic model. The monophone recognition accuracy achieved on the test using the trained HMM, was 75.8%. The new transcriptions produced by the recognition process were then used for the HMM adaptation of the test samples. A decline in the emotion recogition accuracy was expected due to the errors in transcription caused by the recognition system, but the difference was only slight with an UW recall of 69.7%, presented in Table 3 under the name "rec-HMM–UBM–MAP". This result shows that the proposed modeling scheme of adapted HMMs can also be used in situations where transcriptions are not provided, but can be produced by a speech recognition engine. In our work we did not focus primarily on the speech recognition task, thus a simple monophone acoustic model was used. With a better speech recognition engine the number of errors in the produced transcriptions could be reduced, enabling an even better emotion recognition performance.

In our experiments so far we used one Gaussian per state in the HMM, yielding a total of 120 Gaussians for the whole HMM. But the number of Gaussians representing the density of a particular state can be increased or reduced. Reduction is achieved by tying the distribution of particular states, which means some states share the same distribution. Using this procedure the number of states remains the same, but the number of Gaussian distributions is smaller consequently yielding a simplified model structure. The higher number of Gaussians is achieved by simply using more Gaussians to represent the distribution of a particular state. However, one must be aware that the final super-vector's size depends on the number of Gaussians used. Therefore, a double increase of the Gaussians will result in twice the size of the super-vector, which is already with 120 Gaussians at 3120. In order to determine the optimum number of Gaussians,

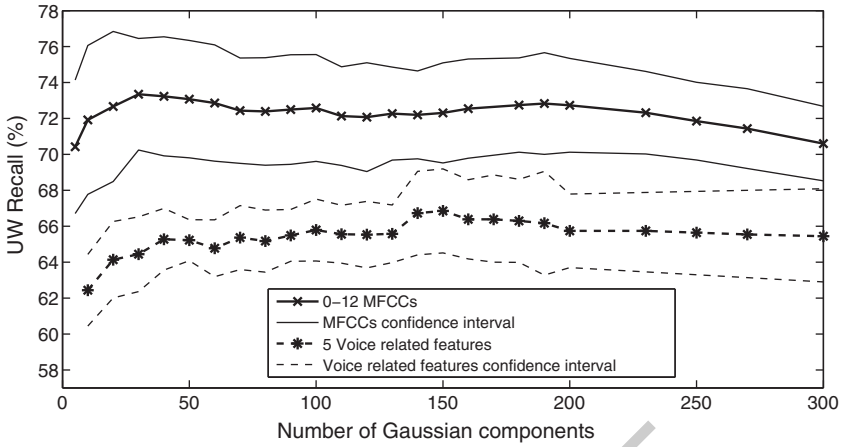

Fig. 2. The effect of different number of Gaussian components on performance of the HMM–UBM–MAP system, determined by a 10-fold cross validation over the training data (FAU-Aibo corpus). Confidence intervals shown represent confidence level of 95%.

we evaluated our HMM–UBM–MAP system in a 10-fold cross-validation on the training data. The reason for selecting such a protocol was to omit the use of test recordings in parameter tuning in order to simulate the conditions that the participants of the Interspeech 2009 Challenge faced. The results of the cross-fold evaluation are shown in Fig. 2 for the MFCC-based system and the system using voice related features, which is presented later. It can be seen that the performance of both systems changes only slightly throughout the range. However, for the final evaluation on the test set, the number of Gaussians was set to the value that gave the best result in the cross-fold evaluation. In the case of the MFCC-based system the optimum number of Gaussians for the cross-fold test on the training data was 30. On the test set an UW recall of 70.3% was achieved, when using 30 Gaussian. The recognition accuracy is identical to a system with 120 Gaussians, confirming our conclusions, drawn from Fig. 2, that the number of Gaussians, if within a range of approximately 10–200, does not have a major effect on the recognition performance of the system.

Until now all the experiments were based on the MFCCs. However, we wanted to evaluate our proposed modeling scheme on other sets of acoustic features, usually employed for the tasks of speaker state recognition. For the Interspeech 2011 Speaker State Challenge baseline system, a set of 5 voice related acoustic features and their first order Δ are used, which seemed appropriate for our experiments as well.

The F0-related features are not directly suitable for modeling on the frame level since parts of the recording where an unvoiced phone is uttered or during silence the value is always 0. Thus, when modeling the F0-related features with the GMM or HMM one or more mixtures will be estimated to the mean of 0, and while there is no variability around the mean value, the covariance matrix values will be getting smaller, causing numerical problems. To overcome this problem we added a Gaussian noise to the F0-related features, with the magnitude of 0.1% of the variance. Next, in the HMM–UBM–MAP process we added the MFCCs to the voice related feature set in order to compensate for training the HMM states that correspond to unvoiced monophones. Finally, we extracted the voice related components of the super-vector and evaluated their performance. Similary to the MFCC evaluation, we ran a 10-fold cross-validation on the training set, as shown in Fig. 2. Again, the performance does not change significantly throughout the range between 10 and 300 Gaussians. The peak performance on the training data cross-fold evaluation was observed when using 150 Gaussians. On the test data the system with the 150 Gaussians produced the UW recall of 65.7%.

In the final experiment on the FAU-Aibo database, the MFCC-based and the voice-related-based super-vectors, from the best individual systems, were concatenated in a form of an early fusion. Using the same SVM classification as in all the previous experiments, we achieved an UW recall of 71.5%, which is a competitive result to the Interspeech 2009 majority voting fusion of 5 best systems (71.2% UW recall), though not significantly better. However, at the significance level $\alpha = 0.05$, the difference between our result and the best individual system from the challenge is significant. A figure presenting the absolute improvements needed to state that the difference between the two results is significant can be found in Schuller et al. (in press).

Table 4
Comparison of classification results on the VINDAT database in terms of unweighted and weighted average recall (%).

System	Acoustic features	# Gaussians	Recall	
			UW	WA
Interspeech 2011 Speaker State Challenge baseline			67.1	62.5
GMM–UBM–MAP	0–12 MFCCs +Δ	64	66.8	62.2
GMM–UBM–MAP	0–12 MFCCs +Δ	128	65.0	60.4
HMM to GMM–UBM	0–12 MFCCs +Δ	123	67.3	62.6
rec-HMM–UBM–MAP	0–12 MFCCs +Δ	123	68.7	64.4
HMM–UBM–MAP	0–12 MFCCs +Δ	123	70.9	66.5

4.2. VINDAT experiments and results

The VINDAT corpus features a smaller amount of data with a total number of 871 samples. Therefore, a 5-fold stratified cross-validation was used in all the experiments. Out of 10 people (5 male and 5 female), 1 male and 1 female were selected for the test set in each fold, with the rest (4 male and 4 female) forming the training set in the corresponding fold. With this we have a speaker independent protocol, which would guarantee a robust and reliable assessment of the competing systems. The reported UW and WA recalls are averages over the results from all 5 folds.

As a reference system, the Interspeech 2011 Speaker State Challenge baseline was used (Schuller et al., 2011). One of the sub-challenges deals with the speaker's intoxication based on the amount of consumed alcohol, thus the system from the challenge seems an appropriate baseline in order to assess the performance of our proposed method. The acoustic feature set consists of 4 energy-related, 5 voice related and 50 spectral acoustic features, which are then modeled by a number of functionals on the utterance level. The produced functional set per each utterance represents an input to the SVM classification. The UW recall achieved by the baseline system is 67.1%, as presented in Table 4. Interestingly, the recognition performance is very close to the one reported in the Interspeech 2011 Challenge, where a different and much larger database of recordings under alcohol intoxication (Schiel et al., 2011) is used. The similarity in the results shows that the VINDAT corpus, although smaller in size, can be used as a reasonable representation of the variability between normal speech and speech under the influence of alcohol.

The second experiment conducted on the VINDAT was aimed at assessing the standard GMM–UBM–MAP modeling scheme, described in Section 2.2. The acoustic features consisted of 0–12 MFCCs and their Δ, following the same setup as the FAU-Aibo experiments. A different number of GMM mixtures was evaluated, with the 64-component GMM producing the highest UW recall of 66.8%. A larger number of GMM mixtures was evaluated as well, but the recognition performance was lower, as shown in Table 4 for the case of 128 mixtures. The result is lower than the baseline, which is in accordance with our experiments on the ALC corpus, conducted as a part of our participation in the Interspeech 2011 Speaker State Challenge (Gajšek et al., 2011).

Next, our proposed method of constructing the GMM–UBM based on the HMM was evaluated. The number of allophones used in the dictionary of the VINDAT corpus is 41, combined with one Gaussian per state in the 3-state HMM, the total number of Gaussian components extracted from the HMM is 123. With the "HMM to GMM–UBM" system we achieved a slightly better result than the baseline, with the UW recall of 67.3%, as presented in Table 4.

The ability to use the proposed HMM–UBM recognition system in scenarios where transcriptions are not available was evaluated using a speech recognition system for the Slovenian language. The monophone acoustic model used in the speech recognition system was trained using Gopolis, VNTV and K211d corpora (described in Mihelič et al. (2003)) omitting the VINDAT recordings. This is a different approach to the FAU-Aibo corpus where automatic transcriptions were generated using an acoustic model trained on the training data of the FAU-Aibo corpus. No language model was used in the recognition process. After the recognition process the generated transcriptions were used with the HMM–UBM–MAP system. Using the same number of Gaussians (123) we achieved a competitive result with the UW recall of 68.7%.

Finally, we tested the HMM–UBM–MAP method using the manual annotations, which gave the best result in the emotion recognition task in Section 4.1. A noticeable improvement was observed with the UW recall of 70.9%, surpassing the baseline by almost 4% absolute. This is a significant improvement over the baseline system, determined by a one-tailed significance test at the significance level $\alpha = 0.01$. The baseline statistical feature's dimension is 4368,

whereas our HMM–UBM–MAP derived super-vector has a lower dimension with the number of Gaussians (123) times the dimension of the acoustic feature vector (26) equaling 3198. Therefore, it should be noted that the relatively small number of training samples (approximatly 700) in combination with the rather large SVM input vectors favors our approach with the smaller vector size. The difference in performance between the reference system and the HMM–UBM–MAP method is expected to be lower with more training data available, but overall both should improve. This can be observed, for instance in Schuller et al. (2011), where the SVM classification results improve consistently with more training data available. The reported results are consistent with the authors' experiments on the ALC database, as a part of the Interspeech 2011 Speaker State Challenge participation, showing that the proposed type of modeling of the acoustic features is suitable, and in combination with the SVM classifier leads to a state-of-the-art results.

5. Conclusions

In this article, an efficient method for modeling the acoustic features needed to recognize various paralinguistic states is presented. Building on the established scheme of modeling the features using the adaptation of the Universal Background Model (UBM), usually represented by a Gaussian Mixture Model (GMM), we take a different approach to constructing the UBM. Instead of training the GMM as the UBM, a monophone-based HMM is used to represent the UBM. In situations where transcriptions cannot be made available for the test set, we propose to transform the HMM–UBM to the GMM–UBM, retaining the segmental nature of the trained HMM, which is being constructed on the monophone level. If transcriptions are available, or can be produced by a speech recognition system, we propose that the adaptation of the monophone HMM is employed. In all cases, only the mean vectors of the Gaussians are adapted, thus producing the super-vectors, which are later used as an input to the Support Vector Machines (SVMs) classification.

The proposed modifications are evaluated by using two databases, focusing on the speech of different speaker states. The advantage of our method in recognizing emotional states is shown by using the FAU-Aibo corpus of spontaneous children's speech. The corpus was used in the Interspeech 2009 Emotion Challenge, thus the results from the competition present a reliable reference. With our system, we are able to achieve a better performance than all the individual contributions in the 2-class protocol. Furthermore, our unweighted average recall of 71.5% is higher than the majority voting fusion of 5-best systems competing in the challenge, which to our knowledge is the highest reported result in the literature.

In the second task we employed our system in the alcohol detection task, using the VINDAT corpus. Here, the reference system was taken from the Interspeech 2011 Speaker State Challenge, where one of the sub-challenges is the recognition of alcoholization. First, we evaluated the standard adaption of the GMMs, which gave a lower recognition rate than the reference system. With our proposed method we were able to surpass the reference system by almost 4% absolute, achieving an unweighted average recall of 70.9%. Interestingly, the results from our experiments are very close to the baseline results for the Interspeech challenge, leading us to believe that the VINDAT corpus, regardless of its smaller size, contains an adequate representation of speech under different levels of alcohol intoxication.

The observed improvement of the HMM-based construction of the UBM over the GMM-based, is in our opinion due to the way the clusters are formed in the UBM. In the case of the GMM, the model is estimated using the Expectation–Maximization algorithm, where clusters are formed in an unsupervised manner and do not necessarily represent the actual base units of speech. In the case of the monophone HMM, the clusters are formed based on the phonetic transcriptions and the final UBM model represents a general distribution of the features for each base unit of speech. Hence, during the UBM adaptation only the parameters of the base units which are represented in the current recording are updated. In this way we get a new, adapted HMM, which differs from the UBM only in the clusters representing the base units found in the corresponding recording. Therefore, the super-vector, extracted from the means of the newly adapted HMM, represents the distribution of the base units contained in the recording, normalized by the UBM. Effectively, this means that we are comparing how the distributions of the base units differ in different recordings. We believe that this additional constraint in the HMM estimation, that clusters are formed based on the base units of speech, assists in an improved representation of speech by the UBM. Furthermore, the segmental nature of the HMM–UBM construction enables a future evaluation of using only parts of the whole UBM for the adaptation. For instance, if we know that the acoustic features being modeled should not differ in the parts of speech represented by certain monophones, the corresponding monophones' mixtures could be discarded, reducing the size of the super-vector and consequently improving the performance.

In the future we would like to assess the ability of modeling other types of features, known to provide additional information of the speaker's state. Also, we would like to evaluate the proposed method in other tasks from the broad range of paralinguistic phenomena, where we believe similar improvements can be made.

References

Batliner, A., Steidl, S., Hacker, C., Nöth, E., 2008. Private emotions versus social interaction: a data-driven approach towards analysing emotion in speech. User Model. User-Adapted Interact. 18, 175–206.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N., 2011. Whodunnit – searching for the most important feature types signalling emotion-related user states in speech. Comput. Speech Lang. 25, 4–28.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. 41, 164–171.

Bocklet, T., Stemmer, G., Zeißler, V., Nöth, E., 2010. Age and gender recognition based on multiple systems – early vs. late fusion. In: INTERSPEECH, pp. 2830–2833.

Burget, L., Matějka, P., Černocký, J., 2006. Discriminative training techniques for acoustic language identification. In: Proceedings of ICASSP 2006, pp. 209–212.

Busso, C., Lee, S., Narayanan, S., 2009. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE Trans. Audio Speech Lang. Proc. 17, 582–596.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylorr, J., 2001. Emotion recognition in human–computer interaction. IEEE Signal Process Mag. 18 (1), 32–80.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoust. Speech Signal Proc. 28, 357–366.

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. Ser. B (Methodological) 39, 1–38.

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., Boufaden, N., 2009. Cepstral and long-term features for emotion recognition. In: Proc. INTERSPEECH 2009, Brighton, UK, ISCA, pp. 344–347.

Eyben, F., Wöllmer, M., Schuller, B., 2009. openEAR – introducing the munich open-source emotion and affect recognition toolkit. In: Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009 (ACII 2009), Amsterdam, The Netherlands, IEEE, pp. 576–581.

Eyben, F., Wöllmer, M., Schuller, B., 2010. openSMILE – the munich versatile and fast open-source audio feature extractor. In: Proc. ACM Multimedia (MM), Florence, Italy, pp. 1459–1462.

Feld, M., Burkhardt, F., Müller, C., 2010. Automatic speaker age and gender recognition in the car for tailoring dialog and mobile services. In: INTERSPEECH, ISCA, pp. 2834–2837.

Gajšek, R., Dobrišek, S., Mihelič, F., 2011. University of ljubljana system for interspeech 2011 speaker state challenge. In: INTERSPEECH 2011, ISCA, pp. 3297–3300.

Gajšek, R., Štruc, V., Mihelič, F., 2010. Multi-modal emotion recognition using canonical correlations and acoustic features. In: Int. Conf. on Pattern Recognition 2010, IEEE Computer Society, pp. 4133–4136.

Gajšek, R., Žibert, J., Justin, T., Štruc, V., Vesnicer, B., Mihelič, F., 2010. Gender and affect recognition based on GMM and GMM–UBM modeling with relevance MAP estimation. In: INTERSPEECH-2010, pp. 2810–2813.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H., 2009. The WEKA data mining software: an update. SIGKDD Expl. 11, 10–18.

Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K., 2001. Improvements to Platt's SMO algorithm for SVM classifier design. Neural Comput. 13 (3), 637–649.

Kockmann, M., Burget, L., Černocký, J., 2009. Brno University of Technology System for Interspeech 2009 Emotion Challenge. In: Proc. INTERSPEECH 2009, ISCA, pp. 348–351.

Kockmann, M., Burget, L., Černocký, J., 2010. Brno University of Technology System for Interspeech 2010 Paralinguistic Challenge. In: Proc. INTERSPEECH 2010, ISCA, pp. 2822–2825.

Lee, C.H., Gauvain, J.L., 1993. Speaker adaptation based on MAP estimation of HMM parameters. In: Acoustics, Speech, and Signal Processing, 1993 IEEE International Conference on ICASSP-93, vol. 2, pp. 558–561.

Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantizer design. IEEE Trans. Commun. 28, 84–95.

Mihelič, F., Gros, J., Dobrišek, S., Žibert, J., Pavešić, N., 2003. Spoken Language Resources at LUKS of the University of Ljubljana. Int. J Speech Technol. 6, 221–232.

Nitesh, V., Chawla, E.A., 2002. Synthetic minority over-sampling technique. J. Artif. Intel. Res. 16, 321–357.

Platt, J.C., 1999. Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf, B., Burges, C., Smola, A. (Eds.), Advances in Kernel Methods – Support Vector Learning. MIT, Cambridge, MA, USA, pp. 185–208.

Pols, L.C.W., 1966. Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words. Ph.D. thesis. Free University, Amsterdam, Netherlands.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker Verification Using Adapted Gaussian Mixture Models. Digit. Signal Proc. 10, 19–41.

Schiel, F., Heinrich, C., Barfüsser, S., 2011. Alcohol language corpus: the first public corpus of alcoholized german speech. Lang. Resourc. Eval., 1–19.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2007. The relevance of feature type for the automatic classification of emotional user states: low level descriptors and functionals. In: INTERSPEECH, ISCA, Antwerp, Belgium, pp. 2253–2256.

Schuller, B., Batliner, A., Steidl, S., Seppi, D., 2011. Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge. Speech Communication, in press.

Schuller, B., Hage, C., Schuller, D., Rigoll, G., 2009. 'Mister D.J., Cheer me up!': musical and textual features for automatic mood classification. J. New Music Res. (JNMR) 38 (4), Taylor & Francis.

Schuller, B., Steidl, S., Batliner, A., 2009. The INTERSPEECH 2009 emotion challenge. In: INTERSPEECH 2009, ISCA, Brighton, UK, pp. 312–315.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2010. The INTERSPEECH 2010 paralinguistic challenge. In: ISCA (Ed.), INTERSPEECH 2010, ISCA. , pp. 2794–2797.

Schuller, B., Steidl, S., Batliner, A., Burkhardt, F., Devillers, L., Müller, C., Narayanan, S., 2011. The INTERSPEECH 2011 speaker state challenge. In: INTERSPEECH 2011, ISCA, pp. 3201–3204.

Schuller, B., Vlasenko, B., Eyben, F., Rigoll, G., Wendemuth, A., 2009. Acoustic emotion recognition: A benchmark comparison of performances. In: Proceedings of Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 552–557.

Schuller, B., Wöllmer, M., Eyben, F., Rigoll, G., 2009d. The role of prosody in affective speech, linguistic insights, studies in language and communication. J. New Music Res. 97, 285–307.

Steidl, S., 2009. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech. Logos Verlag, Berlin.

Vapnik, V.N., 1998. Statistical Learning Theory. Wiley-Interscience.

Vlasenko, B., Wendemuth, A., 2009. Processing affected speech within human–machine interaction. In: Proc. INTERSPEECH 2009, Brighton, UK, ISCA, pp. 2039–2042.

Žibert, J., Pavešić, N., Mihelič, F., 2006. Speech/non-speech segmentation based on phoneme recognition features. EURASIP J. Appl. Signal Proc., 1–13.

Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.C., 2009. The HTK Book, Version 3.4.1. Cambridge University Engineering Department, Cambridge, UK.