Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 53 (2011) 1172–1185

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Application of speaker- and language identification state-of-the-art techniques for emotion recognition[☆]

Marcel Kockmann[*], Lukáš Burget, Jan “Honza” Černocký

Brno University of Technology, Speech@FIT, Czech Republic

Available online 1 February 2011

Abstract

This paper describes our efforts of transferring feature extraction and statistical modeling techniques from the fields of speaker and language identification to the related field of emotion recognition. We give detailed insight to our acoustic and prosodic feature extraction and show how to apply Gaussian Mixture Modeling techniques on top of it. We focus on different flavors of Gaussian Mixture Models (GMMs), including more sophisticated approaches like discriminative training using Maximum-Mutual-Information (MMI) criterion and InterSession Variability (ISV) compensation. Both techniques show superior performance in language and speaker identification. Furthermore, we combine multiple system outputs by score-level fusion to exploit the complementary information in diverse systems. Our proposal is evaluated with several experiments on the FAU Aibo Emotion Corpus containing non-acted spontaneous emotional speech. Within the Interspeech 2009 Emotion Challenge we could achieve the best results for the 5-class task of the Open Performance Sub-Challenge with an unweighted average recall of 41.7%. Further additional experiments on the acted Berlin Database of Emotional Speech show the capability of intersession variability compensation for emotion recognition.

Keywords: Emotion recognition; Gaussian mixture models; Maximum-mutual-information; Intersession variability compensation; Score-level fusion

1. Introduction

Spoken emotion recognition is the problem of automatically recognizing the emotional state of a person from their speech. Different moods may change the attributes of the human voice, such as pitch, speaking-rate, and intonation.

In automatic speech processing these properties are usually represented using the appropriate parametrization of speech, so called features. Pattern recognition and machine learning algorithms can then be used to model certain characteristics of emotionally colored speech and recognize emotions in speech utterances. Typically, classifiers like Hidden-Markov-Models (HMMs), Gaussian Mixture Models (GMMs), Support Vector Machines (SVMs) or Neural Networks (NNs) (Bishop, 2006) are used.

While sensing the emotions of an individual from their speech is a relatively new research field in speech processing, a research community has formed in recent years and several methods have been applied successfully (Steidl, 2009; Vlasenko et al., 2007; Seppi et al., 2008; Batliner et al., 2006) and evaluated on special databases containing emotional speech (Ververidis and Kotropoulos, 2003).

Recently the usage of SVMs to directly model large-scale feature vectors has become the standard for emotion recognition (Schuller et al., 2007, 2009). These feature vectors contain diverse kinds of speech parametrization extracted on a per-utterance basis including acoustic, prosodic and voice quality features. Frame based features are usually modeled by HMMs to capture the temporal dynamics of the signal (Schuller et al., 2009).

Using these state-of-the-art techniques, accuracies of over 80% have been reported for emotion classification

[☆] This work was partly supported by European project MOBIO (FP7-214324), by Grant Agency of Czech Republic project No. 102/08/0707, and by the Czech Ministry of Education project No. MSM0021630528. Marcel Kockmann is supported by SVOX Deutschland GmbH, Munich, Germany.

[*] Corresponding author.
E-mail address: kockmann@fit.vutbr.cz (M. Kockmann).

0167-6393/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2011.01.007

tasks on acted non-spontaneous data (Schuller et al., 2006). However, on real life non-acted spontaneous emotionally colored data these accuracies drop drastically (below 40%) (Schuller et al., 2009).

Besides emotion recognition there are many diverse research fields with the goal of extracting certain attributes from speech. These include:

- What is spoken: Automatic Speech Recognition (ASR).
- Who is speaking: Speaker Identification (SID).
- Which language is used: Language Identification (LID).
- Which gender is the speaker: Gender identification (GID).
- What is the age of the speaker: Age Identification (AID).

In many of these fields (like SID, LID and GID) the use of Gaussian Mixture Models has established itself as the standard (Reynolds et al., 2000). HMMs, as used in ASR, are usually outperformed by GMMs (which are actually a HMM containing a single state) on text-independent tasks. Also, best results in all these fields are often obtained using more or less standard acoustic features extracted on a frame-based level, as used in ASR. This is somewhat illogical as features for ASR are optimized to blind out properties like speaker characteristics. Still, these tools seem to provide a good framework for diverse kinds of speech characterization.

As mentioned above, the state-of-the-art for emotion recognition has moved in a different direction. Gaussian mixture modeling of short-time acoustic features has been mostly replaced by Support Vector Machine classification. A similar trend was observed in the field of Speaker Verification as well. However, recent advances in Gaussian Mixture Modeling, like discriminative training or intersession variability compensation, has significantly raised the performance of GMM based systems and currently defines the state-of-the-art (Kinnunen and Li, 2010). This is the main motivation for our work. Our aim is to take basic and newly evolved features and modeling techniques, as used in current LID and SID systems and to apply them to the task of emotion recognition. By doing so we want to provide another view to the problem of emotion recognition. Further enhancement can then be expected by combining both approaches.

Through this paper, we will investigate standard spectral features based on Mel-Frequency-Cepstral-Coefficients (MFCC) (Davis and Mermelstein, 1980) as they are usually used in ASR. There have been many modifications of standard MFCC features to better fit the needs of SID and LID, like longer temporal context and speaker normalization. We will evaluate below some of these techniques for emotion recognition.

Furthermore, prosodic features (incorporating duration, pitch and energy) are often used to enhance the performance of MFCC based systems. Different from spectral features, prosodic features are usually extracted over a longer time span, like on a syllable basis. We examined a prosodic feature extraction method successfully used for GMM based speaker recognition (Kockmann and Burget, 2008).

All these features will be modeled using different flavors of Gaussian Mixture Models. It should be noted, that in all cases we model frame or syllable based features using models without any temporal dependencies. This statistical method of creating a “footprint” has been very successful. We will investigate in detail basic GMM approaches used in speaker and language identification. Furthermore, more sophisticated techniques evolved in the last few years are examined for their applicability in emotion recognition. These include discriminative training of GMMs and intersession variability compensation. Intersession variability for emotion recognition may refer to different acoustic conditions, different speakers or simply the spoken content of the utterance. All these attributes are a nuisance for the task of emotion recognition and we want to “ignore” them during modeling.

To evaluate the performance of the proposed techniques we provide experiments on two independent emotional databases, one containing non-acted spontaneous speech and the other acted non-spontaneous speech. Results on the first database include our submission to the Interspeech 2009 Emotion Challenge (Kockmann et al., 2009) where we could achieve very good results using the techniques described above.

The paper is organized as follows: Section 2 describes the acoustic features we used in our experiments while Section 3 explains the prosodic features used. Section 4 gives detailed information on the Gaussian Mixture Models we used and their training and evaluation procedures. In Sections 5, 6 we present results to evaluate the proposed approaches for emotion recognition. In Section 7 we draw conclusions to our approaches and consider future research.

2. Spectral features

This section will introduce the used MFCC features and the additional techniques applied to make them more suitable for the given task.

2.1. Basic acoustic features

The most widely used features in speech processing are MFCCs (Davis and Mermelstein, 1980). They have been applied successfully for speech recognition as well as for speaker recognition and language identification. We will use them as our basic features for the emotion recognition task. MFCC vectors are generated every 10 ms on a 20 ms frame of speech weighted by a Hamming window. Fast-Fourier-Transform (FFT) output of each speech window is processed by a Mel filter bank with 25 bands. The output is transformed by Discrete Cosine Transform (DCT) and

13 cepstral coefficients including C0 are generated. C0 represents an energy measure of the speech window.

2.2. Channel normalization

The temporal trajectories of individual cepstral coefficients are filtered using a standard RelAtive SpecTrAl (RASTA) filter (Hermansky and Morgan, 1994) to remove slow and very fast spectral changes which do not appear to be characteristic for natural speech. We use the standard IIR filter:

$$H(z) = 0.1\frac{2 + z^{-1} - z^{-3} - 2z^{-4}}{z^{-4}(1 - 0.982z^{-1})}. \tag{1}$$

Furthermore, cepstral mean subtraction (CMS) is applied on each coefficient per utterance for simple channel normalization.

2.3. Speaker normalization

We do not want to model the characteristics of the individual speaker by the position of the formants based on the length of the vocal tract. We use Vocal Tract Length Normalization (VTLN) (Cohen et al., 1995) for simple speaker normalization. The spectrum (during FFT) is either compressed (usually for females) or expanded (for male speakers) based on a warping factor estimate for each utterance.

Warping factors for training and test data are estimated using a rather small GMM trained on all unnormalized training data to represent average characteristics of the target population. Warped MFCCs are then created for all files with warping factors in a range from 0.88–1.12 with a step-size of 0.02. This results in 13 feature sets: 6 compressed, 1 neutral and 6 expanded. The optimal warping factor per utterance is obtained by evaluating the likelihood of all warped instances against the unnormalized GMM and selecting the maximum. This way we select the factor that best fits the average speaker. The warped utterances are then used for standard model training. Refer to Section 4.1 for implementation details for GMM training and likelihood scoring. For spectrum manipulation we use a linear piecewise warping function with a warping cutoff of $0.875 \times N_f$, where N_f is the Nyquist frequency.

2.4. Temporal context

Simple MFCCs do not model any temporal characteristics which are most likely informative for emotion recognition. As our classifier also does not model feature sequences, we generate delta, double and triple delta regression coefficients of the static features to model co-articulations in speech. We use a standard formula (Young et al., 2006):

$$d_t = \frac{\sum_{\forall\Theta}\Theta(c_{t+\Theta} - c_{t-\Theta})}{2\sum_{\forall\Theta}\Theta^2} \tag{2}$$

with d_t being the regression coefficient of static coefficient c_t and the shift vector $\Theta = [2]$ for delta, $\Theta = [2,4]$ for double delta and $\Theta = [2,4,6]$ for triple deltas. This results in 26, 39 and 52 dimensional feature vectors containing information spanning a context of 5, 9 and 13 frames, respectively.

2.5. Shifted delta cepstra

The importance of an even broader temporal information has been shown for LID (Torres-Carrasquillo et al., 2002). The so-called Shifted Delta Cepstra (SDC) is created by stacking delta coefficients computed across multiple speech frames, as depicted in Fig. 1. Multiple delta coefficients with a shift of ±1 are computed for a context of ±10 frames, without overlap and concatenated in one feature vector.

For static features c_t shifted deltas are defined:

$$\Delta c_t = c_{(t+iP+d)} - c_{(t+iP-d)} \tag{3}$$

for $i = [-3\ldots0\ldots3]$ with shift $P = 3$ and the window shift $d = 1$ over which deltas are computed.

The basic features in our system are 7 static MFCC coefficients (including coefficient C0) concatenated with delta cepstra which totals 56 SDC coefficients per frame, spanning a context of 21 frames. This configuration has been successfully used in our language identification systems (Matejka et al., 2008, 2006).

2.6. Post processing: voice activity detection

For all our frame based spectral features, non-speech frames are discarded and only speech frames are considered in the following stages of training models and verification. Speech/non-speech segmentation is performed by our Hungarian phone recognizer (Schwarz et al., 2006). This step is performed based on the final feature vectors ensuring that RASTA and regression coefficients are correctly estimated.

3. Prosodic features

Prosodic information based on the lexical context might be useful for this task and is complementary to the acoustic short time features. For this purpose, we use our detector of syllable-based feature contours as presented in (Kockmann and Burget, 2008). It processes classical prosodic features like duration, pitch and energy in a syllable-like temporal context. The trajectories of each feature are continuously modeled over the time span of a syllable and are represented by discrete cosine transformation (DCT) coefficients, as depicted in Fig. 2. The pseudo-syllable segmentation is based on a phone recognizer where vowels are considered as nuclei for the syllables. The segments are non-overlapping and undefined frames are discarded prior to DCT approximation. Additionally, we also capture the temporal contours of MFCCs and form a single feature vector out of duration, pitch, energy and

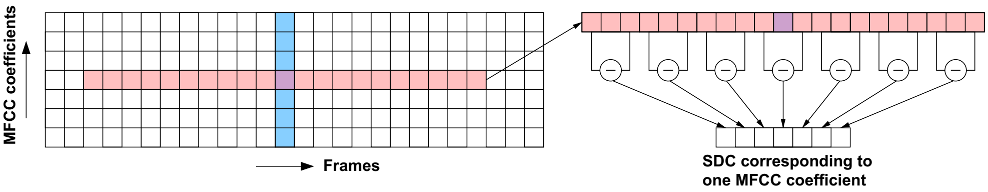

Fig. 1. Computation of SDC features for a single static feature stream, incorporating 21 consecutive static MFCCs, results in 7-dimensional SDC vector for each frame.

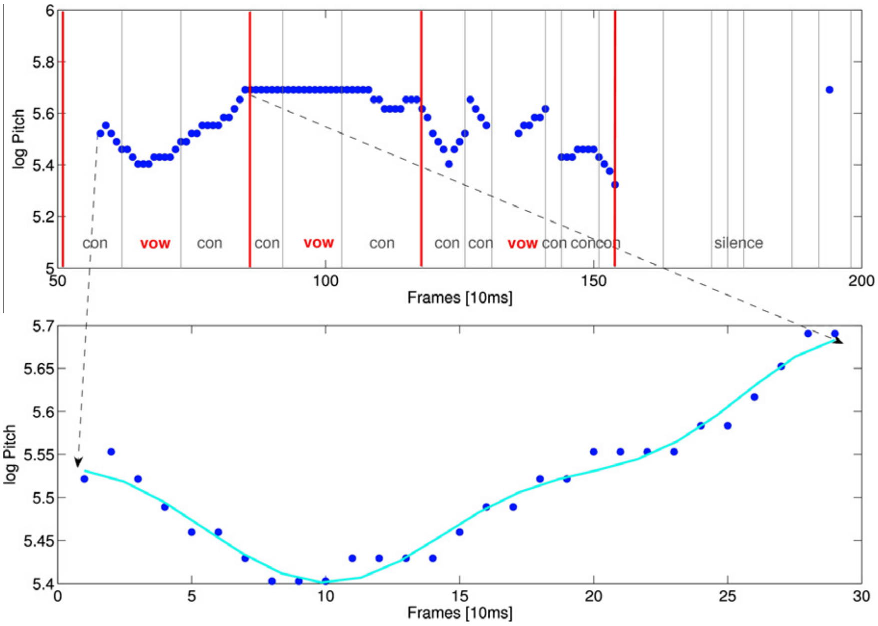

Fig. 2. Example of a pitch contour over a syllable consisting of three phones. Top: Original pitch values with phone and pseudo-syllable boundaries (horizontal lines). Bottom: Original (points) and DCT approximated curve (solid line).

the MFCC contours. Frame-based pitch and energy are generated first and are mean subtracted over the voiced part of the utterance before approximating the temporal trajectory. We use the syllable duration (number of frames) and 6 DCT coefficients per feature contour which results in 13-dimensional vectors for the prosodic and 85-dimensional vectors for the combined prosodic and MFCC contours.

4. Classifier

In this section, we introduce four statistical models that are used in our experimental part. The first two are flavors of Universal Background Model (UBM)-GMM models as used in speaker verification, with and without session variability compensation. The third and fourth are classical GMMs trained in generative and discriminative manner as often used in language identification. We provide most of the needed formulas to easily allow the reader to reproduce our results.

4.1. UBM based models

Our first two GMM systems are based on a standard Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) paradigm (Reynolds et al., 2000). All GMMs used are multivariate with dimension D and using diagonal co-variances.

Prior to any class-dependent model training a class-independent model is trained on the pooled feature vectors $\boldsymbol{o}$ of all development data of all classes. Following speaker recognition terminology we call this a Universal Background Model. Weights π, means $\boldsymbol{\mu}$ and variances $\boldsymbol{\Sigma}$ of the UBM are trained in a maximum-likelihood way with an Expectation-Maximization (EM) algorithm (Bishop, 2006).

EM is an iterative algorithm that alternates between estimating the responsibilities $\gamma_k(n)$ (E-Step, alignment of frame $n = 1 \ldots N$ to Gaussian components $k = 1 \ldots K$) and re-estimation of the parameters using the current responsibilities (M-Step):

E-Step:

$$\gamma_k(n) = \frac{\pi_k \mathcal{N}(\boldsymbol{o}_n|\boldsymbol{\mu}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_k \mathcal{N}(\boldsymbol{o}_n|\boldsymbol{\mu}_k, \Sigma_k)}. \tag{4}$$

M-Step:

$$\boldsymbol{\mu}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(n) \boldsymbol{o}_n, \tag{5}$$

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma_k(n) \left(\boldsymbol{o}_n - \boldsymbol{\mu}_k^{new}\right) \left(\boldsymbol{o}_n - \boldsymbol{\mu}_k^{new}\right)^T, \tag{6}$$

$$\boldsymbol{\pi}_k^{new} = \frac{N_k}{N} \tag{7}$$

with

$$N_k = \sum_{n=1}^{N} \gamma_k(n) \tag{8}$$

and likelihood function $\mathcal{N}(\boldsymbol{o}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is

$$\frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}_k|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{o}_n - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\boldsymbol{o}_n - \boldsymbol{\mu}_k)\right\} \tag{9}$$

for feature vector $\boldsymbol{o}_n$ with feature dimension D.

Data log-likelihood for the whole GMM and all data $\boldsymbol{o}$

$$\ln p(\boldsymbol{o}|\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^{N} \ln \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{o}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{10}$$

is checked for convergence after each iteration.

For UBM training we initialize a single Gaussian component with a global mean and variance of all background data and keep splitting the components in two (after several iterations when convergence of data log-likelihood is achieved) until the final size is reached. For this purpose, the copied weights π are halved, variances $\boldsymbol{\Sigma}$ are kept and copied means $\boldsymbol{\mu}$ are shifted by $\pm 0.2\sqrt{\boldsymbol{\Sigma}}$.

Following UBM training, the individual emotion-class models are obtained by relevance Maximum-A-Posteriori (MAP) adaptation (Reynolds et al., 2000) of the mean parameters using class specific feature vectors only. Weights and variances are kept fix. The UBM mean serves as a prior for posterior distribution of class model means and the relevance factor further restricts their movement. The point estimate of the posterior mean distribution can be seen as a compromise between the prior (UBM) mean and the maximum likelihood solution (using feature vectors for emotion class e only):

$$\boldsymbol{\mu}_{ek}^{MAP} = \boldsymbol{\alpha}_k \boldsymbol{\mu}_{ek}^{ML} + (1 - \boldsymbol{\alpha}_k) \boldsymbol{\mu}_k^{UBM} \tag{11}$$

with adaptation coefficients

$$\alpha_k = \frac{\sum_{n=1}^{N} \gamma_k(n)}{\sum_{n=1}^{N} \gamma_k(n) + \tau} \tag{12}$$

and relevance factor $\tau = 16$. If some components are not occupied at all by the training data, the parameters keep their prior values; while for unlimited amount of data the MAP estimate would equal the ML estimate.

During testing the models are evaluated using the log-likelihood ratio (LLR) between the class model- and the UBM log-likelihood for the test data, evaluating Eq. (10) for both the class model and UBM. For computational efficiency, only top scoring Gaussians (determined based on the UBM) are evaluated for the class models per frame. We will call this model simply *GMM-UBM* model.

The described GMM-UBM framework can be expanded to cope with intersession variability (e.g. different channel, language, gender, etc. between training and test utterances). This technique allows us to adapt the supervector of means (concatenated mean parameters of all Gaussian components) in directions of large intersession variability during verification to better match the test utterance.

In Fig. 3 we try to visualize the meaning of this technique for emotion recognition on a simple toy example. We assume GMMs containing a single mixture component each in a two dimensional feature space. The figure shows only the mean parameters of the GMMs. We should assume two utterances for each of the three emotion classes *Anger (black star), Neutral (cyan diamond)* and *Joy (magenta x-mark)*.

After the training of the UBM (blue cross) on all utterances we do one additional ML iteration using data from each utterance only. The new mean parameter ML estimates for each utterance are depicted in the figure, same colors belong to same emotion classes. It can be observed that most of the variability between different utterances belonging to the same emotion classes can be projected on a one-dimensional latent space (Intersession variability direction, dash-dotted line). This subspace can be robustly estimated on many diverse utterances belonging to different emotion classes (after UBM training, prior to class model training). Emotion class models are then derived by

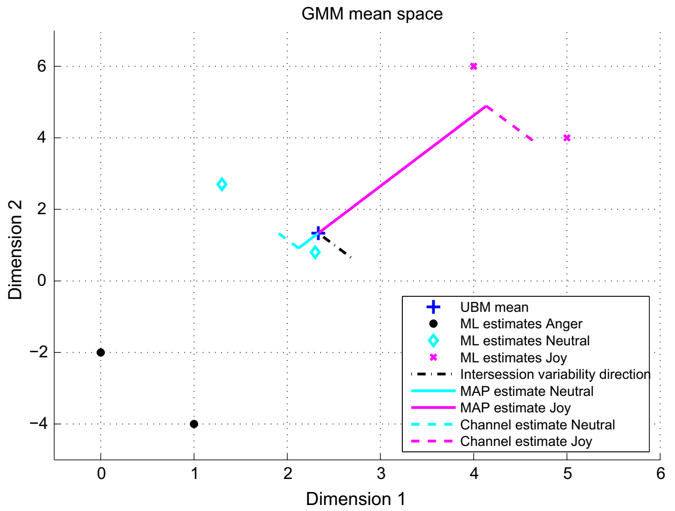

Fig. 3. Toy example of intersession variability compensation in a 2D mean parameter space. 1D subspace is estimated based on differences between utterances belonging to the same class. Model parameters can be moved along this space during verification to adapt to the test environment.

standard MAP adaptation as for the *GMM-UBM* model (shown for Neutral and Joy in plot).

During verification, the MAP adapted means of the model to be tested can be moved along the intersession variability subspace to adapt to the condition in the test utterance (acoustic condition, gender, linguistic content, etc.). This is illustrated for two utterances tested against class models for *Joy* and *Neutral* by the dashed lines drawn from the top of the solid lines (MAP estimate).

In a real application the subspace usually maps out from a very high dimensional supervector space (up to 100,000 dimensions) down to a low dimensional latent space (e.g. 50 dimensions) allowing it to robustly adapt model parameters on small amounts of data.

The adapted mean supervector can be represented as

$$\boldsymbol{m}_e + \boldsymbol{U}\boldsymbol{x}_n \tag{13}$$

and is distributed with a mean of $\boldsymbol{m}_e$ and a co-variance of $\boldsymbol{U}\boldsymbol{U}^T$. $\boldsymbol{m}_e$ is the class (emotion) dependent supervector of MAP adapted means (from standard *GMM-UBM* model). $\boldsymbol{U}$ defines the low-dimensional subspace matrix (size $DK \times S$ with subspace size $S \ll DK$) of the full GMM space with high intersession variability. The utterance dependent factors $\boldsymbol{x}_n$ define the shift of the model parameters within the subspace. These factors are assumed to be normally distributed random variables making the whole thing a probabilistic model.

The subspace is usually estimated for on a large amount of data (similar to UBM), either using Principle Component Analysis (PCA) (Burget et al., 2007) or by an EM algorithm (Kenny et al., 2008). Please refer to these citations for detailed descriptions.

Once the subspace is estimated, emotion models (or UBM) can be adapted by shifting its mean supervector in the directions given by an intersession variability subspace to better fit the test utterance data. Mathematically, this can be expressed as finding the factors $\boldsymbol{x}_r$, that maximize the following MAP criterion:

$$p(\boldsymbol{o}_r|\boldsymbol{m}_e + \boldsymbol{U}\boldsymbol{x}_r)\mathcal{N}(\boldsymbol{x}_r; \boldsymbol{0}, \mathbf{I}), \tag{14}$$

where $p(\boldsymbol{o}_r|\boldsymbol{m}_e + \boldsymbol{U}\boldsymbol{x}_r)$ is the likelihood of the test conversation r given the adapted supervector (model) and $\mathcal{N}(\cdot; \boldsymbol{0}, \mathbf{I})$ denotes a normally distributed vector. Assuming a fixed occupation of Gaussian mixture components (responsibilities) by test conversation frames, $\boldsymbol{o}_n$, $n = 1, \ldots, N$, it can be shown (Brümmer, 2004) that $\boldsymbol{x}_r$ maximizing criterion (14) is given by:

$$\boldsymbol{x}_r = \boldsymbol{A}^{-1} \sum_{k=1}^{K} \boldsymbol{U}_k^T \sum_{n=1}^{N_r} \gamma_k(n) \frac{\boldsymbol{o}_n - \boldsymbol{\mu}_k}{\sigma_k}, \tag{15}$$

where $\boldsymbol{U}_k$ is the $D \times S$ part of matrix $\boldsymbol{U}$ corresponding to kth mixture component; $\gamma_k(n)$ is the probability of occupation mixture component k at time n, $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$ are the mixture component's mean and standard deviation vectors and

$$\boldsymbol{A} = \boldsymbol{I} + \sum_{k=1}^{K} \boldsymbol{U}_k^T \boldsymbol{U}_k \sum_{n=1}^{N_r} \gamma_k(n). \tag{16}$$

In our implementation, occupation probabilities, $\gamma_k(n)$, are computed using UBM and assumed to be fixed for a given test conversation. This allows one to pre-compute matrix $\boldsymbol{A}^{-1}$ only once for each test conversation.

Note, that both model and UBM means are adapted to the test utterance and afterwards scoring is done exactly as for the *UBM-GMM* model (LLR).

We will call this model incorporating intersession variability compensation *ISV* model.

4.2. Generative and discriminative GMMs

Emotion recognition is a closed-set identification task (similar to Language identification) and usually large amounts of data are available to train the separate class models. In this section we propose to train each class model using an EM algorithm as described in the previous section for the UBM. Our assumption is that we have enough data to robustly estimate weight, mean and variance parameters for each emotion class individually.

Furthermore, we propose to re-estimate the model parameters using a discriminative training technique successfully applied to language identification (Matejka et al., 2006).

As depicted in Fig. 4 discriminative techniques aim to precisely model the boundary between the competing models in such a way that the correct estimation of class affiliation is improved rather than maximizing the likelihood of the training data. This way model parameters are mostly used to estimate precisely the boundaries between separable regions in the features space. Highly overlapping areas are neglected.

Our first set of models is trained per class under the conventional Maximum Likelihood (ML) framework, as used for the UBM (see Section 4.1, Eqs. (4)–(10)), but only using class specific data. Note, that we re-estimate not only means, but also weights and variances per emotion class. We will call these models simply *ML* models.

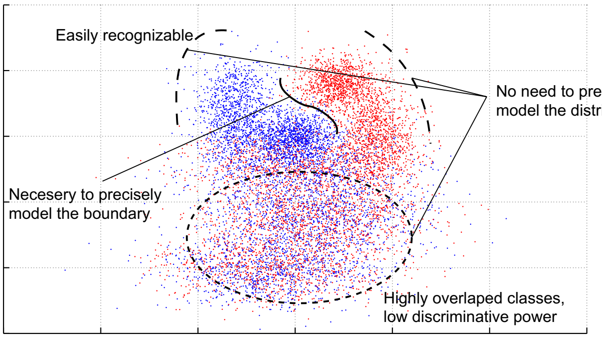

Fig. 4. Effect of discriminative training for two classes in 2D feature space. The model parameters are used to precisely model the boundary between separable data while highly overlapping areas are neglected.

These serve as a starting point for further discriminative re-estimations of means and variances using the Maximum Mutual Information (MMI) criterion.

Unlike in the case of ML training, which aims to maximize the overall likelihood of training data given the transcriptions, the MMI objective is to maximize the posterior probability of correctly recognizing all training segments (utterances):

$$\mathcal{F}_{MMI} = \sum_{r=1}^{R} \ln \frac{p(\boldsymbol{o}_r|\boldsymbol{\mu}_{e^+}, \boldsymbol{\Sigma}_{e^+}, \boldsymbol{\pi}_{e^+})}{\sum_{e=1}^{E} p(\boldsymbol{o}_r|\boldsymbol{\mu}_e, \boldsymbol{\Sigma}_e, \boldsymbol{\pi}_e)}. \tag{17}$$

where the numerator is the likelihood of r-th training segment $\boldsymbol{o}_r$; given the correct emotion class model of the segment, e^+; R is the number of training segments and the denominator represents the overall probability density, $p(\boldsymbol{o}_r)$ (likelihood given any emotion class). So, the MMI parameter re-estimates aim to maximize the ration between true class likelihood and overall likelihood of each segment.

It can be shown (Povey, 2003) that the MMI objective function (17) is increased by re-estimating model parameters using extended Baum-Welch algorithm (similar to standard EM training) with the following formula for updating mean and variances:

$$\mu_{ek}^{new} = \frac{\theta_{ek}^{num}(\boldsymbol{o}) - \theta_{ek}^{den}(\boldsymbol{o}) + 2\gamma_{ek}^{den}\mu}{\gamma_{ek}^{num} + \gamma_{ek}^{den}}, \tag{18}$$

$$\Sigma_{ek}^{new} = \frac{\theta_{ek}^{num}(\boldsymbol{o}^2) - \theta_{ek}^{den}(\boldsymbol{o}^2) + 2\gamma_{ek}^{den}\left(\Sigma_{ek} + \mu_{ek}^2\right)}{\gamma_{ek}^{num} + \gamma_{ek}^{den}} - \mu_{ek}^{new^2}. \tag{19}$$

The terms:

$$\theta_{ek}^{num}(\boldsymbol{o}) = \sum_{r=1}^{R}\sum_{n=1}^{N_r} \gamma_{ekr}^{num}(n)\boldsymbol{o}_r(n), \tag{20}$$

$$\theta_{ek}^{num}(\boldsymbol{o}^2) = \sum_{r=1}^{R}\sum_{n=1}^{N_r} \gamma_{ekr}^{num}(n)\boldsymbol{o}_r(n)^2,$$

$$\gamma_{ek}^{num}(\boldsymbol{o}) = \sum_{r=1}^{R}\sum_{n=1}^{N_r} \gamma_{ekr}^{num}(n),$$

are mixture component specific first and second order statistics and occupation counts corresponding to the numerator of the objective function (17). Denominator statistics can be expressed by similar equations, where all superscripts *num* are merely replaced by *den*. Note that the numerator statistic are ordinary ML statistics. Therefore, the numerator posterior probability of occupying mixture component ek by n-th frame of training segment r,

$$\gamma_{ekr}^{num}(n) = \begin{cases} \gamma_{ekr}(n) & \text{for } e = e^+, \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

is non-zero only for mixture components corresponding to the correct emotion class. To estimate the posterior probabilities for the denominator:

$$\gamma_{ekr}^{den}(n) = \gamma_{ekr}(n)\frac{p(\boldsymbol{o}_r|\boldsymbol{\mu}_e, \boldsymbol{\Sigma}_e, \boldsymbol{\pi}_e)}{\sum_{q=1}^{E} p(\boldsymbol{o}_r|\boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q, \boldsymbol{\pi}_q)}. \tag{22}$$

Note, that the fraction on the right-hand side is the posterior probability of the current emotion class given the whole segment that n belongs to.

Finally,

$$\gamma_{ekr}(n) = \frac{\pi_{ek}\mathcal{N}(\boldsymbol{o}_r(n)|\mu_{ek}, \Sigma_{ek})}{\sum_{j=1}^{K}\pi_{ej}\mathcal{N}(\boldsymbol{o}_r(n)|\mu_{ej}, \Sigma_{ej})} \tag{23}$$

where π_{ek} is mixture component weight and K is the number of mixture components in model e.

Starting from the ML models of final size, the mean and variance parameters are re-estimated using MMI for several iterations.

For both models, verification is done frame-by-frame for the test utterance with full log-likelihood computation according to Eq. (10). Note, that we always evaluate all Gaussian components for these two model types.

5. Experiments on the FAU Aibo emotion corpus

In this section we present experimental results to evaluate the techniques presented in Sections 2–4. All used feature configurations and classifiers are summarized in Table 1. Experiments on feature types and modeling techniques are performed on the FAU Aibo corpus.

5.1. Database

The FAU AIBO database is a corpus with recordings of children of age 10 to 13 interacting with a pet robot called Aibo. The emotionally colored speech is non-rehearsed, as the children believed that the robot was following their commands, so their reactions evoke emotions due to behavior or misbehavior. Actually, the actions of the robot were in a fixed order, controlled by an operator and similar for all participants.

The whole corpus consists of 9.2 hours of high quality speech which was annotated by human labelers and assigned to emotional classes by majority voting. All sessions are split on a chunk level to achieve homogeneity of emotional state within a unit and results in about 18,000 chunks.

The database was recorded at two different schools, consisting in a total number of recordings of 51 children. The first portion consists of 13 male and 13 female speakers. Within the Emotion Challenge 2009, the first part was provided as a combined training and development set, while the second part was defined to be the test set. The emotion labels for the second part were not provided and results could only be evaluated within the Interspeech 2009 Emotion Challenge. As a consequence, we will provide two different results in this chapter. First, we will describe progress in system development on our own defined development set and afterwards, we will give the official results obtained in the challenge with our final systems.

All annotated emotion labels of each chunk were mapped to two broader sets of emotions: A 5-class set

Table 1
Summary of feature sets and model types used in experimental part.

Feature type	Description	Dimension
MFCC	C0+12 MFCCs with CMS, VAD	13
RASTA	C0+12 MFCCs with RASTA and CMS, VAD	13
RASTA-Δ	C0+12 MFCCs with RASTA and deltas, CMS, VAD	26
RASTA-ΔΔ	C0+12 MFCCs with RASTA, deltas and double deltas, CMS, VAD	39
RASTA-ΔΔΔ	C0+12 MFCCs with RASTA, deltas, double and triple deltas, CMS, VAD	52
SDC	C0+6 MFCCs+delta cepstra over 21 frames, CMS, VAD	56
DPE	Duration+syllable contours (6 DCT coefficients each) for pitch and energy	13
DPEC	Duration+syllable contours (6 DCT coefficients each) for pitch, energy and MFCCs	85
Model type	**Description**	**Components**
GMM-UBM	GMM with MAP adapted means from UBM	8–128
ISV	GMM with MAP adapted means from UBM and intersession variability compensation	8–128
ML	ML-trained GMM (weights, means, variances)	16–128
MMI	ML-trained GMM (weights, means, variances) with further MMI training (means, variances)	16–128

containing **A**nger, **E**mphatic, **N**eutral, **P**ositive and **R**est, and a 2-class set comprising **NEG**ative and **IDL**e. Detailed information on the database and its design is given in (Steidl, 2009). To keep our experimental part clear for the reader we present only results on the 5-class task (obviously the more difficult task).

The total number of chunks available for training/development of the 5-class models are in Table 2. Note that the numbers differ from Schuller et al. (2009), as our voice activity detection did not identify any speech frames for several chunks.

5.2. Development set

We use subsets of the training data for system development. We use a full jackknifing approach for the whole training set. Thirteen splits are created out of the training set, each excluding 1 male and 1 female (so speaker in training and test are always distinct), resulting in circa 700 chunks for the testing of each split. We train a separate system for each split on the remaining chunks. This is a very expensive procedure, but this way we can use all available data for training and testing, while the training and test portions are always distinct. Results are presented in terms of two accuracies: The *Weighted Accuracy (WA)* means the percentage of correctly recognized chunks, in the total for all chunks over all classes of the development data. The *Unweighted Accuracy (UA)* means the percentage of correctly recognized chunks per class, which are then averaged over all classes. As the class affiliation is highly unbalanced (see Table 2), we will use the unweighted accuracy as our primary measure for system development.

Table 2
Number of chunks in the AIBO corpus development set to train each classifier for 5 classes.

Anger	Emphatic	Neutral	Positive	Rest	∑
830	1890	5024	616	642	9002

Table 3
Results for static MFCCs features with longer temporal context using *GMM-UBM* with 64 components [%].

Static			Longer context		
Feature	UA	WA	Feature	UA	WA
MFCC	36.4	40.4	RASTA-Δ	41.8	41.3
RASTA	**37.4**	40.9	RASTA-ΔΔ	**43.5**	42.9
			RASTA-ΔΔΔ	42.6	40.7
			SDC	41.9	41.0

5.3. Spectral features

We start with investigations of spectral features using a fixed classifier to compare the performance of the different feature sets. We use a *GMM-UBM* system for this purpose. Preliminary experiments indicate that 64 Gaussians work well for the first *GMM-UBM* system.

As we are using an adaptation from the background to class model it is important to define a balanced set for the UBM training due to the unbalanced amount of class affiliation in the training data (see Table 2). Otherwise, the background model would be biased to the more dominant classes (Neutral and Emphatic) and adapted models for the under-represented classes might be poor. For this purpose, we select 500 chunks from each of the 5 classes to train a model that serves as the UBM. Emotion class models are then obtained by relevance MAP adaptation of the mean parameters.

Results are presented in the left column of Table 3. With 36.4%, the unweighted accuracy is very low for the simple MFCC features. Still, these results correspond with the results reported in a similar test set of the AIBO corpus for a frame based HMM system (Schuller et al., 2009). A significant[1] improvement is achieved through the use of a simple RASTA filter.

The use of Vocal Tract Length Normalization did not give conclusive results and no significant gains could be

[1] At a significance level of $\alpha = 0.1$.

Table 4
Results for syllable based feature contours modeled by 64 component *GMM-UBM* [%].

Feature	UA	WA
DPE	32.3	39.6
DPEC	**36.0**	38.3

achieved. The ineffectiveness of VTLN might be explained by an analysis of the observed distribution of the warping factors. Generally, for data of adults a separation of males and females in the form of a bimodal distribution can be observed. In our experiments there was no separation of warping factors for male and female speech at all which might be caused by the fact that we handle children's speech. Probably a more adjusted grid search than using warping factors 0.88–1.12 (which is somehow optimized for adult speech) can be more effective. As a consequence, we use only RASTA processed features for the following experiments.

As a next step we apply techniques to cover a broader temporal context. Up to now features only model a quasi-static period of approximately 30 ms. We augment the RASTA features with their delta (RASTA-Δ), double-delta (RASTA ΔΔ) and triple-delta (RASTA-ΔΔΔ) regression coefficients.

Results are presented in the column on the right of Table 3. Significant improvements are obtained through all examined configurations and the task clearly benefits from broadening the temporal context. The best results are achieved with RASTA-ΔΔ features which significantly outperform the single delta and SDC features.

One interpretation might be that enlarging the context keeps improving the accuracy but triple delta and SDC feature dimensions are already too high for this scenario. Keep in mind that a higher feature dimension also raises the free parameters in the model dramatically.

Following these experiments, we will use the RASTA-ΔΔ coefficients as our primary spectral feature set.

5.4. Prosodic features

The following experiments are performed to evaluate the prosodic features proposed in Section 3.

We use the same *GMM-UBM* model type as for our previous experiments with features containing the following feature subsets: duration and temporal contours of pitch and energy (DPE) and duration, pitch, energy and MFCC temporal contours (DPEC, see also Table 1).

Results are presented in Table 4. The best results of 36% UA are achieved with the DPEC features. These features show a similar performance as the simple MFCC features without any temporal context. However, the spectral frame-based features incorporating an equal temporal context still perform significantly better. This is a result we also observe in speaker or language identification. High-level features like these usually perform worse on their own but add complementary information. This is then exploited by score-level fusion of the diverse recognition systems.

Another reason for the huge degradation might be the fact that we use statistical classifiers with very little data. As these features are based on syllable regions spanning a context of up to several hundred milliseconds, often only a few or no feature vectors can be extracted per utterance. Clearly, the performance of this feature type suffers greatly from the fact that the test utterances are very short in the AIBO corpus.

5.5. GMM-UBM models

Now we start evaluating the modeling techniques proposed in Section 4. For this purpose, we will use the spectral RASTA-ΔΔ features that performed best in the previous section.

After selecting 64 Gaussians somehow ad-hoc for the initial feature experiments, additional experiments are carried out to find optimal sizes for *GMM-UBM* as well as for *ML* systems for this task.

The use of up to 2048 Gaussian components is typical in high-performing speaker and language identification systems, where much more data is available for each class or for the background model (Burget et al., 2007; Matejka et al., 2006). The used databases of emotional speech are rather small, so (1) we have little data to train the background model and the class model; and (2) the test utterances are also quite short (only up to several seconds). For this reason, we expect the optimum GMM size to be much smaller than for SID/LID systems.

As we use an EM training algorithm that splits Gaussian components after some iterations, we evaluate GMM sizes from 8 to 128, doubling the size after each step. It should be noted that we will also provide class-specific accuracies in this section to show the relation of the GMM size and the amount of available training data.

Results in Table 5 for a GMM-UBM system indicate that a size of 64–128 components is optimal for this task. Using a larger number of mixture components did not increase UA. WA usually kept rising as the major classes (like Neutral) benefit from larger amount of model parameters while the others get overtrained.

5.6. ML models

Furthermore, for the proposed model types in Section 4.2 an independent GMM is trained for each class on the

Table 5
Unweighted, weighted and class specific accuracies for different GMM sizes with RASTA-ΔΔ features for *GMM-UBM*.

GMM size	UA	WA	A	E	N	P	R
8	40.8	37.8	62.3	33.5	35.9	69.0	3.1
16	41.7	41.3	62.1	33.3	42.5	67.1	3.6
32	42.4	43.6	59.6	37.1	45.8	66.1	3.4
64	43.5	42.9	60.8	36.5	43.6	71.4	5.3
128	**43.6**	43.7	61.2	35.8	45.3	70.8	5.0

available target training data only, without any adaptation from a background model. According to Table 2 we only have several hundred chunks available for some classes. For this purpose and to further confirm the GMM size, we proceed with experiments using a different number of Gaussian components for simple *ML* trained models without any background model adaptation.

We evaluate sizes of 16 to 128 Gaussians. The results are presented in Table 6.

Interestingly, for these models we see a similar trend as for the MAP adapted models. We obtain the best results of about 44% UA with 32 and 64 Gaussians. Again, 64 Gaussians seem to be a good choice. Also the models for which only small amount of data is available, such as **A**, **P** or **R**, already seem to get overtrained with 128 Gaussians.

If we compare the results for the *GMM-UBM* system and the *ML* system in Tables 5 and 6 we observe a similar overall performance. Comparing same sized models, we see that the *ML* models are significantly better for the smaller models. This seems reasonable as the small *ML* models might have more discriminative power due to their individual weight and variance parameters. However, for the larger models the amount of training data might still be too small to estimate all these parameters robustly.

5.7. MMI models

After evaluating the two basic GMM models we move on with experiments using more sophisticated modeling approaches.

First, we use the MMI criterion to retrain all generative class GMMs (*ML* models) to discriminative models. This is done in addition to 10 iterations, always increasing the MMI objective function in (17). Comparing the numbers in Table 7 for *MMI* models with previous *ML* experiments (see Table 6) gives somewhat disappointing results.

Except for the small GMM with 16 Gaussians (not significant), all other recognition rates even decrease due to MMI training. This loss of performance is also not significant but seems to show a trend. Only when looking at very small number of Gaussians (e.g. 2) we could spot a significant gain due to MMI, but these models obviously perform much worse than the larger ones.

It should be mentioned here that the proposed technique of discriminative re-training of models leads to huge improvements on NIST evaluation sets for language identification (Matejka et al., 2006) with similar number of classes. More than 50% improvement can be achieved on 30 s long test utterances. Interestingly, on 3 s long utterances (which is more similar to our scenario here) the gain also reduces to less than 10% relative. Another difference is the amount of data to train the class models, which is much higher (hundreds of hours per class) in the case of the NIST LID task (NIST, 2005).

Table 6
Unweighted, weighted and class specific accuracies for different GMM sizes with RASTA-ΔΔ features for the *ML* model.

GMM size	UA	WA	A	E	N	P	R
16	42.7	46.2	54.0	40.2	47.5	46.4	16.2
32	**44.3**	48.2	55.9	45.5	51.5	46.3	22.1
64	44.0	49.2	51.5	45.0	54.1	46.3	23.1
128	42.8	51.3	48.6	43.6	59.6	42.9	19.2

Table 7
Unweighted, weighted and class specific accuracies for different GMM sizes with RASTA-ΔΔ features for the *MMI* model.

GMM size	UA	WA	A	E	N	P	R
16	42.9	46.9	53.6	49.1	48.8	44.5	18.5
32	**44.2**	48.5	53.4	45.7	52.3	44.5	25.0
64	43.7	49.5	49.5	45.1	55.0	44.0	24.8
128	42.2	51.4	47.1	43.4	60.5	39.9	20.2

5.8. ISV models

In the following experiments we want to evaluate the intersession variability compensation approach as proposed in Section 4.1. The system is mainly a *GMM-UBM* system as used in the initial feature experiments with additional intersession variability compensation during testing.

As a first step the low dimensional subspace defining the directions of intersession variability has to be estimated on the training data. The usage of the available training data is crucial during this step and defines what kinds of intersession variability can be compensated for.

The AIBO database comprises many chunks for the same class and the same speaker. So we can learn differences according to acoustic environment, speaker or linguistic content. Our main assumption is that we do not have many channel effects caused by different microphones or transmission channels. As all recordings are done using the same equipment in the same room, the within-class-covariance will mainly cover speaker and intrinsic variations (Shriberg et al., 2009). Still, acoustic channel compensation might be an issue for the test set as this is recorded in a different school under different acoustic conditions.

As the segments are rather short in this database we use a method to learn more reliable subspace directions. We concatenate all segments belonging to the same speaker and class and estimate $\boldsymbol{U}$ as to describe the difference between speakers. This way our intersession variability compensation serves more as a speaker compensation than an acoustic channel compensation.

Before starting the subspace training, we initialize $\boldsymbol{U}$ by PCA (Burget et al., 2007) to ensure a good starting point and faster convergence. Then we iteratively re-train $\boldsymbol{U}$ in 10 iterations.

Once the subspace is estimated, emotion class models are trained by relevance MAP adaptation exactly as for the *GMM-UBM* models. Also, the scoring part itself (LLR) is the same. The only difference is that we adapt the obtained MAP means towards the test utterance along the low-dimensional subspace $\boldsymbol{U}$. This is done by estimating the "channel" factors $\boldsymbol{x}$ for each test utterance using Eqs. (15), (16).

We perform several experiments to determine the optimal number S of intersession variability directions (size of the subspace). Fig. 5 shows unweighted accuracies for up to 5 subspace directions. We can observe that using more than 1 eigenchannel always decreases the performance. We get non-significant improvement over the relevance MAP model with 44.2% for 1 eigenchannel (dashed line), but it drops consistently when increasing the number of subspace directions, which significantly decreases the performance.

One explanation might be that the test utterances in this corpus are simply to short (often below one second of speech) to reliably estimate the $\boldsymbol{x}$ factors that control the adaptation of the model mean parameters. Similar degradation of intersession compensation techniques due to small amount of test data has been observed for speaker (Dehak et al., 2009) as well as language identification (Hubeika et al., 2008) tasks incorporating only a few seconds of speech. Also, the subspace $\boldsymbol{U}$ is usually trained on hundreds of hours of speech.

5.9. System calibration/Fusion

It is advisable to calibrate the system outputs as the obtained scores for our systems do not represent proper posterior probabilities for the classes. A certain GMM may generally produce higher scores than the others in the set. Furthermore, a consequent step is to fuse several of the systems that incorporate partly complementary information, as we have created many different systems based on diverse features and modeling techniques. We have observed huge gains in performance using this technique (Brümmer et al., 2007) even for system configurations that differ only slightly (e.g. only different feature sets).

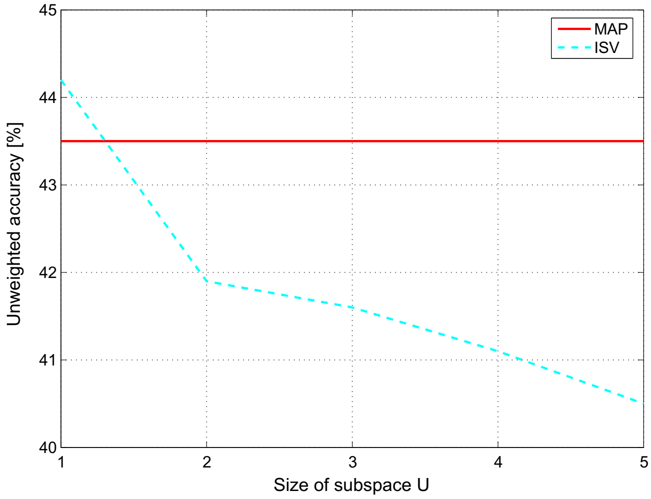

Fig. 5. Effect of eigenchannel subspace size on AIBO corpus. UA for RASTA-ΔΔ features and *ISV* model with 64 components.

For these purposes we use multi-class linear logistic regression (MLLR) (Brümmer and du Preez, 2006) to perform calibrated fusion of our system outputs. Posterior probabilities of class $\mathcal{C}_e$ given the score vector ϕ are then given by:

$$p(\mathcal{C}_e|\phi) = \frac{\exp(a_e)}{\sum_{j=1}^{E}\exp(a_j)} \tag{24}$$

with activations

$$a_e = \mathbf{w}_e^T\phi \tag{25}$$

and ϕ containing the concatenated scores from all systems to be fused. The fusion parameters $\mathbf{w}_e$ are trained on each split of the development set and are then averaged to ensure fair circumstances.

First we perform fusions of two systems that are using the same features but four different modeling techniques. Fusion results for all combinations are presented in Table 8 and are mostly better than the best single ISV system with 44.2%. Significant gains are achieved due to fusion of two heterogeneous systems, like one background model based (*GMM-UBM* or *ISV* system) and one standard GMM model (*ML* or *MMI* system). Fusion of systems where one is derived from the other, like *ML* and *MMI*, results only in a small improvement. Fusion of all 4 systems does not result in further improvement.

Furthermore, we evaluate the effect of fusing systems using different feature sets while keeping the modeling approach fixed (*ISV*). For this purpose we have selected 4 different feature sets that should be most complementary. We select the RASTA MFCCs without further temporal context (37.4%); the SDC features (41.9%); the simple prosodic DPE features (32%); and our standard RASTA-ΔΔ features (44.2%). Results in Table 9 show the same trend as our previous fusion experiments. All combinations are better than the best incorporated single system. Significant gains can be achieved and the best result of 45.9% is obtained for a fusion of RASTA-ΔΔ and SDC features. Again, we fuse all 4 systems without any further improvement.

To conclude these experiments we change both variables (features and modeling techniques) at once. We fuse different combinations but without any further improvement.

5.10. Emotion challenge 2009

This section shows the results for the systems we have selected to submit for the official Open Performance

Table 8
Results (UA) for fusion of 2 systems with same features (RASTA-ΔΔ) and different modeling approaches [%].

	GMM-UBM	ISV	ML	MMI
GMM-UBM	–	44.1	**45.5**	45.3
ISV		–	**45.5**	45.1
ML			–	44.3

Table 9
Results (UA) for fusion of 2 systems with same modeling technique (*ISV*) but different feature sets [%].

	SDC	RASTA-ΔΔ	DPE
RASTA	43.9	44.5	39.3
SDC	–	**45.9**	44.5
RASTA-ΔΔ		–	44.2

Sub-Challenge (Schuller et al., 2009) of the Interspeech Emotion Challenge 2009. Results are presented with the official metric on 5-class tasks, similar to our results on the development set. As classes are highly unbalanced, the rules stipulated the use of the unweighted average recall (UA) as the primary measure and the weighted average recall (WA) as the secondary measure.

We have selected the four different modeling approaches (*ML, MMI, GMM-UBM, ISV*) we used in the system development for the best performing features on the test set. They are based on MFCCs generated with RASTA filter, double-deltas, CMS and VAD (RASTA-ΔΔ). Note, that the scores computed on the test set could only be uploaded up to 25 times. So we had to select the most promising configurations. In Schuller et al. (2009) baseline recognition results on the test set are provided for two different baseline systems. A dynamic modeling approach using frame-based features and a Hidden-Markov-Model (HMM) as a classifier; and the second static approach uses high-dimensional chunk based features fed to a Support Vector Machine classifier. The best baseline results on the proposed primary measure are 35.9% for the HMM baseline and 38.2% for the SVM baseline.

Table 10 shows the results for the 5-class task for the 4 submitted models. We achieve the best results for the *ISV* system with 41.3%. Surprisingly, the *ML* and the *MMI* system perform significantly worse with only about 38.5%, unlike than on the development set. This might indicate that even the ML trained model is already over-adapted to the training data and does not generalize well. The simple GMM-UBM system performs significantly better than the ML/MMI approaches. We get improvement (not significant) from the intersession variability compensation. On the UA we achieve a 15%/8% relative improvement to the HMM and SVM modeling, respectively, which was provided as a baseline.

As proposed in the last section, we want to combine several complementary systems to achieve the best results. We select the most promising fusion of two systems as evaluated in Table 9. We fuse 2 systems using the same ISV model with 64 Gaussians, one with RASTA-ΔΔ features and one with SDC features. The fusion parameters are the same as used in our system development.

Table 10
Submitted systems for the 5-class task [%]. All using RASTA-ΔΔ features.

Feature	UA	WA
GMM-UBM	40.8	41.0
JFA	**41.3**	43.9
ML	38.5	45.4
MMI	38.7	46.0

We get another improvement and achieve an unweighted average recall of 41.7%. This is the highest recognition rate achieved in the Interspeech 2009 Emotion Challenge for the 5-class task. Still, our result was not significantly better than that of some other participants. The organizers (Schuller et al., 2009) could show that further fusion of the (completely independent) participating systems could significantly increase the recognition rate to over 44%.

6. Berlin database of emotional speech

In this section we will present some additional experiments mainly to further investigate the effect of intersession compensation for emotion recognition. As test utterances are extremely short on the FAU Aibo corpus we selected a database with longer test utterances. The Berlin Database of Emotional Speech (Burkhardt et al., 2005) consists entirely of whole sentences that are several seconds long.

6.1. Database

This database contains acted emotional speech. Ten actors (5 male and 5 female) simulated seven different emotions on ten German utterances (5 short and 5 long). Emotion classes are **A**nger, **F**ear, **N**eutral, **J**oy, **S**adness, **D**isgust and **B**oredom. The recordings are studio-quality and the whole database contains 535 sentences. It should be noted, that although the single utterances are longer than for the AIBO corpus, the overall amount of speech data is much smaller (less than one hour).

6.2. Development set

Similar to the AIBO database we use a full jackknifing approach for the whole training set. Ten splits are created out of the training set, each excluding one speaker. The actual number of sentences available to train the classifiers are depicted in Table 11. Similar to Section 5.2, results are presented in terms of unweighted accuracy (UA). It should be noted, as the amount of speech data for class **D** is extremely low and preliminary testing fails completely in this class, we discard class **D** from our development set and take only 6-classes into account.

6.3. ISV model

We perform experiments on a similar system as used for the Interspeech Emotion Challenge. We create MFCC

Table 11
Number of utterances in the Berlin Database of Emotional Speech to train each classifier.

A	B	D	F	J	S	N	∑
127	81	46	69	71	62	79	535

features, apply the RASTA filter and CMS and augment the features with delta and double-deltas. Afterwards, speech frames are selected using voice activity detection (RASTA-ΔΔ).

These spectral features are first used to train a UBM with 64 Gaussians. We again use a class-balanced data set for background model training. A single UBM for each split consists of approximately 300 sentences, 50 for each class. After UBM training we train the intersession variability subspace $\boldsymbol{U}$. We use the same recipe for subspace estimation as in the previous experiments: PCA initialization of $\boldsymbol{U}$ with successive ML-training. We again concatenate all utterances per speaker to train the intersession variability subspace. The whole database was recorded in an anechoic chamber using high-quality equipment so channel effects are minimal. Effects of speaker normalization might be even more meaningful than for the AIBO corpus as the database consists of adult speech.

Experiments are carried out to investigate the effect of intersession compensation for this database. For this purpose we train and evaluate *GMM-UBM* and *ISV* models as described in the previous sections. In Fig. 6 an interesting trend can be observed which is different from the experiments on the AIBO corpus. While we reach an unweighted accuracy of 57% using relevance MAP, we achieve a significant improvement by using the same system incorporating intersession variability compensation. As depicted by the dashed line in Fig. 6 we reach a recognition rate of 63% with the use of one subspace direction. The use of a larger subspace further increases the performance and the best unweighted accuracy of 67% is achieved with a subspace size of 5. This is a significant improvement of an absolute 10% UA over the *GMM-UBM* baseline.

We are aware that better recognition rates have been reported on this database. In (Schuller et al., 2006) accuracies of over 80% are reached but only by using much more complex large-scale feature sets. For these studio-quality recordings, features like pitch and voice quality will be of high accuracy and might explain the huge difference in recognition performance. In (Gaurav, 2008), frame-based MFCCs using GMMs are also evaluated and performed in a similar way to our baseline system. Furthermore, GMMs are outperformed by SVM approaches in that work. Our conclusion for the performance gap to the state-of-the art SVM systems is that SVMs might be better suitable to handle the general small amount of training data in this database.

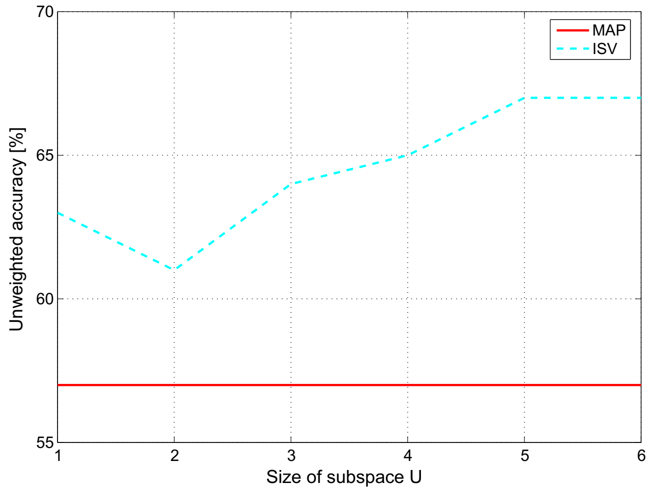

Fig. 6. Effect of eigenchannel subspace size on Berlin Database of Emotional Speech corpus. UA for RASTA-ΔΔ features and *ISV* model with 64 components.

Nevertheless, our experiments show the capability of intersession compensation techniques for emotion recognition.

7. Conclusions

We show that feature extraction and statistical modeling methods that are usually used in speaker and language recognition can be successfully used for emotion recognition as well.

We could achieve the best results for the 5-class task in the Interspeech Emotion Challenge 2009 and significantly outperformed the provided state-of-the-art baseline systems.

The submitted system incorporated quite simple acoustic features. We did not make use of excessive spectral, prosodic or lexical features. Eventually, we used two different feature sets both derivatives of MFCC features. Several experiments on our development set indicated that MFCC features using RASTA filter and augmented with first and second order derivatives performed the best for this task. It should be noted, that this feature set is very close to those used in automatic speech recognition. As a complementary feature set we use Shifted Delta Cepstra with an even broader temporal context.

Our prosodic feature set showed bad performance compared to the spectral features. While this is a common effect also observed in other fields of speech based pattern recognition tasks, we can conclude that in this case the given test utterances are really too short to exploit a syllable based long-temporal span feature extraction. Future work should consider exploiting a simpler prosodic feature set like frame based pitch values or functionals computed on shorter fixed size windows.

The proposed GMM based modeling approaches generally perform very well. However, the more sophisticated approaches, namely discriminative training and intersession variability compensation, were not convincing on the FAU AIBO corpus. While both approaches have proven their potential in terms of language identification we could only reach marginal improvements. Our conclusion is that this effect is mainly due to the short test utterances and the general small amount of training data per class. In the mentioned NIST evaluations for language identification the core condition consists of test utterances with durations of 30 s. In this task MMI as well as intersession variability compensation has shown up to 50% relative improvement, while on a 3 s task the gain degrades to approximately 10%

relative improvement, both for MMI and intersession variability compensation.

That there is indeed a capability for intersession variability compensation for emotion recognition is shown in the Berlin Database of Emotional Speech. Here we can obtain significant gains through the use of the *ISV* model. Still, it should be mentioned that in both cases we used *ISV* mainly to reduce the effects of intersession variability representing speaker characteristics instead of channel characteristics as is usually done.

Large-scale feature SVM modeling still seems to be superior on acted non-spontaneous studio-quality recordings, unlike that on real-world data. Our impression is that prosodic and voice-quality features are very accurate on this type of recordings and yield the high accuracies. Still, SVMs seem to be a good choice to handle very small amounts of training data while generative statistical models like GMMs get simply overtrained.

Furthermore, we could show that system combinations by score level fusion can significantly enhance performance. In conclusion, in this way diverse modeling techniques (like SVM or GMMs) and feature sets (acoustic, prosodic, chunk or frame based, etc.) can be exploited for high accuracy in emotion recognition tasks.

References

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., 2006. Combining efforts for improving automatic classification of emotional user states. In: Proceedings of IS-LTC, pp. 240–245.

Bishop, C., 2006. Pattern recognition and machine learning.

Brümmer, N., 2004. Spescom DataVoice NIST 2004 system description. In: Proceedings NIST Speaker Recognition Evaluation 2004, Toledo, Spain, June. 2004.

Brümmer, N., Burget, L., Cernocky, J., Glembek, O., Grezl, F., Karafiat, M., van Leeuwen, D.A., Matejka, P., Schwarz, P., Strasheim, A., 2007. Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006. IEEE Trans. Audio, Speech Lang. Process. 15 (7), 2072–2084.

Brümmer, N., du Preez, J., 2006. Application-independent evaluation of speaker detection. Comput. Speech Lang. 20 (2–3), 230–275.

Burget, L., Matejka, P., Schwarz, P., Glembek, O., Cernocky, J., 2007. Analysis of feature extraction and channel compensation in a GMM speaker recognition system. IEEE Trans. Audio, Speech, Lang. Process. 15 (7), 1979–1986.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A database of german emotional speech. In: Ninth European Conference on Speech Communication and Technology.

Cohen, J., Kamm, T., Andreou, A., 1995. Vocal tract normalization in speech recognition: Compensating for systematic speaker variability. J. Acoust. Soc. Amer. 97, 3246.

Davis, S., Mermelstein, P., 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Audio, Speech Lang Process. 28 (pp. 1–4), 357–366.

Dehak, N., Kenny, P., eda Dehak, R., Dumouchel, P., Ouellet, P., 2009. Front-end factor analysis for speaker verification. IEEE Trans. Audio, Speech Lang. Process., 1–23.

Gaurav, M., 2008. Performance analysis of spectral and prosodic features and their fusion for emotion recognition in speech. In: Spoken Language Technology Workshop, 2008. SLT 2008. IEEE, pp. 313–316.

Hermansky, H., Morgan, N., 1994. Rasta processing of speech. IEEE Trans. Speech Audio Process. 2 (pp. 1–4), 578–589.

Hubeika, V., Burget, L., Matejka, P., Schwarz, P., 2008. Discriminative training and channel compensation for acoustic language recognition. In: Proceedings of Interspeech, 1990–9772.

Kenny, P., Ouellet, P., Dehak, N., Gupta, V., Dumouchel, P., 2008. A study of inter-speaker variability in speaker verification. IEEE Trans. Audio, Speech, Lang. Process. 16 (5), 980–988.

Kinnunen, T., Li, H., 2010. An overview of text-independent speaker recognition: From features to supervectors. Speech Commun. 52 (1), 12–40.

Kockmann, M., Burget, L., 2008. Contour modeling of prosodic and acoustic features for speaker recognition. In: Spoken Language Technology Workshop, SLT 2008. IEEE, pp. 45–48.

Kockmann, M., Burget, L., Cernocky, J., 2009. Brno University of technology system for interspeech 2009 emotion challenge. In: Proceedings of Interspeech, Brighton, pp. 348–351.

Matejka, P., Burget, L., Glembek, O., Schwarz, P., Hubeika, V., Fapso, M., Mikolov, T., Plchot, O., Cernocky, J., 2008. But language recognition system for NIST 2007 evaluations. In: Proceedings of Interspeech.

Matejka, P., Burget, L., Schwarz, P., Cernocky, J., 2006. Brno University of Technology system for NIST 2005 language recognition evaluation. In: Proceedings of Odyssey.

NIST, 2005. The 2005 NIST language recognition evaluation plan, pp. 1–6.

Povey, D., 2003. Discriminative Training for Large Vocabulary Speech Recognition. Ph.D thesis, Cambridge University Engineering Department, 2003, pp. 1–172.

Reynolds, D., Quatieri, T., Dunn, R., 2000. Speaker verification using adapted Gaussian Mixture Models. Digital Signal Process. 10 (1–3), 19–41.

Schuller, B., Arsic, D., Wallhoff, F., Rigoll, G., 2006. Emotion recognition in the noise applying large acoustic feature sets. Speech Prosody, Dresden.

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. INTERSPEECH 2007, 1–4, June.

Schuller, B., Steidl, S., Batliner, A., Feb 2009. The INTERSPEECH 2009 Emotion Challenge. In: Proceedings of Interspeech, Brighton, pp. 1–4.

Schwarz, P., Matejka, P., Cernocky, J., 2006. Hierarchical structures of neural networks for phoneme recognition. In: Proceedings of ICASSP 2006, Toulouse, pp. 325–328.

Seppi, D., Batliner, A., Schuller, B., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Aharonson, V., 2008. Patterns, prototypes, performance: classifying emotional user states. In: Proceedings of Interspeech.

Shriberg, E., Kajarekar, S., Scheffer, N., 2009. Does session variability compensation in speaker recognition model intrinsic variation under mismatched conditions? Interspeech Brighton.

Steidl, S., 2009. Automatic classification of emotion-related user states in spontaneous children's speech. Studien zur Mustererkennung, Bd. 28, ISBN 978-3-8325-2145-5, 1–260 (January).

Torres-Carrasquillo, P., Singer, E., Kohler, M., Greene, R., Reynolds, D., Jr, J.D., 2002. Approaches to language identification using Gaussian Mixture Models and shifted delta cepstral features. In: Seventh International Conference on Spoken Language Processing.

Ververidis, D., Kotropoulos, C., 2003. A state of the art review on emotional speech databases. In: Proceedings of 1st Richmedia Conference, pp. 109–119.

Vlasenko, B., Schuller, B., Wendemuth, A., Rigoll, G., 2007. Combining frame and turn-level information for robust recognition of emotions within speech. In: Proceedings of Interspeech, pp. 2249–2252.

Young, S., Evermann, G., Gales, M., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The htk book version 3.4.