Applause Sound Detection*

CHRISTIAN UHLE, *AES Member*
(Christian.Uhle@iis.fraunhofer.de)

Fraunhofer Institute for Integrated Circuits, DE-91058 Erlangen, Germany

A comprehensive investigation of the detection of applause sounds in audio signals is presented. It focuses on the processing of single-channel recordings in real time with low latency. Of particular concern are the intensity of the applause within the sound mixture and the influence of interfering sounds on the recognition performance, which is investigated experimentally. Various feature sets, feature processings, and classification methods are compared. Low-pass filtering of the feature time series leads to the concept of sigma features and yields further improvement of the detection result.

0 INTRODUCTION

The detection of applause is relevant to a number of applications. In the context of music information retrieval (MIR) it provides valuable information for the segmentation and classification of audio or video, which is useful for the discrimination between live recordings and studio recordings, for searching in archives of sound effects, for highlight spotting, and for content summarization.

The classification of audio content has been a vital research area for more than a decade, with influencing contributions by Wold et al. [1], Casey [2], and others. Various works relate to the analysis of audio with the aim to detect events in sports videos (see, for example, [3], [4]). A segmentation and classification of compressed audio data considering four classes (silence, speech, music, applause) has been described in [5]. Here the mean and variances of the subband spectral centroid within segments of 1-second length each are used as features for the classification. Cai et al. presented a framework for audio key sound detection for a variable number of sound classes using hidden Markov models [6]. Another method for the detection of laughter, cheer, and applause using hidden Markov models has been investigated earlier in [7], where signal segments of 1-second length each are classified. Applause sounds were also considered in approaches to audio segmentation and classification [8], [9].

These methods aim at the classification of multiple key sounds. However, the separate detection of each of the key sounds in scope is beneficial in cases when different key sounds occur simultaneously within a sound mixture. Jarina and Olajec [10] investigated the influence of feature selection on applause detection using genetic algorithms and simulated annealing from a feature set comprising Mel-frequency cepstral coefficients (MFCC) and delta features on the classification performance of Gaussian mixture models. Related to this work are the feature selection experiments for the discrimination between synchronous and asynchronous applause described in [11]. The detection of applause in recordings of meetings is addressed in [12] using a method based on information about the length of nonsilent segments (determined by means of voice activity detection) and estimates of the fundamental frequency.

Besides the applications in the context of MIR, applause detection can be applied successfully for spatial audio coding (SAC), such as MP3 Surround [13] and MPEG Surround [14], and for stereo coding with the MPEG Surround 2-1-2 mode as used in MPEG unified speech and audio coding (USAC) [15]. Since applause signals are critical items for today's spatial audio codecs [16], a dedicated processing of applause may result in higher audio quality [17]. This application requires a classification of short-signal segments (or fine granularity) with low latency, which differs from many of the previous works on applause sound detection [5], [7], [10], [11].

Applause is produced by the clapping of hands by many people. Neda et al. [18] reported that applause begins often as incoherent clapping but is frequently followed by a synchronization process between the individuals in the audience, leading to rhythmic applause. They observed that during synchronization the clapping period increases. This synchronization may disappear and reappear several times while the clapping period increases and decreases.

Another property of applause is that the signals of a multimicrophone recording of applause are in general highly decorrelated, especially when recorded with spaced microphones, because of the spatial distribution of the audience. However, this property is not considered for the feature extraction in this work for three reasons. First it is not featured in all audio recordings. Second a

*Presented at the 127th Convention of the Audio Engineering Society, New York, 2009 October 9–12, under the title "Applause Sound Detection with Low Latency"; revised 2010 December 20.

human listener is able to detect applause without this cue and so might do a machine. And finally the computational complexity is increased when processing multichannel recordings. Instead, the proposed method processes one-channel audio signals.

Applause may occur separately from other sounds (such as music, speech, laughter, and screaming) or together with interfering sounds. Its detection is assumably more reliable if the signal-to-noise ratio (SNR) is large, where noise relates to all nonapplause sound components. On the other hand, the applications mentioned put different demands on the classification of applause within a sound mixture. Examples are the discrimination between live and studio recordings (where applause needs to be detected even for low SNRs), the segmentation of a recording of a concert into songs (where only the detection of applause with high SNR, which marks the beginning or end of a song, is desired), and SAC (where even more than two classes resembling different SNRs may be desired to allow the selection of different parameters for the processing, depending on the SNR). This work investigates the influence of the SNR on the recognition performance.

The paper is organized as follows. Section 1 explains the method, namely, the feature extraction and processing and the classifiers under investigation. The experiments and results are described in Section 2. Section 3 describes an application example, and Section 4 presents the conclusions of this work.

1 METHOD DESCRIPTION

Human listeners can easily detect whether or not an audio recording features applause without special training. They do so as easily as they can solve a variety of tasks of auditory scene analysis, such as recognizing speech in noisy environments, tapping the beat of a musical piece with syncopated rhythm after listening to a couple of notes, or discriminating a trumpet from a saxophone once they are familiar with the sounds of both instruments.

A common approach to enable a computer to solve such tasks is supervised learning. The knowledge needed for the problem at hand is obtained by the machine from a set of training data, comprising examples (for example, for the task of classifying musical instruments such examples could be recordings of single-instrument sounds) and reference labels (such as instrument labels). The general approach to supervised learning is shown in Fig. 1.

It should be noted that by using this approach one does not tacitly assume that supervised learning is the key element of human perception and cognition. It does not explain the capability of the human auditory system and brain, which rely to a large extent on the ability to separate auditory streams within a mixture of sounds (cocktail party problem [19]), to generalize between different instances of the same class, and to update an internal model given a set of new observations. However, by finding a set of features which are robust with respect to variations of the sound to be detected and the background noise (that is, all other sounds), successful sound detection can be implemented using computers.

1.1 Feature Extraction

Various features have been proposed and applied to the classification and segmentation of audio signals. Many of them originate from applications of speech signal processing. Although this paper aims at a comprehensive investigation, we will restrict the set of features under investigation owing to the large number of possible combinations of features and parameters for feature extraction, processing, and classification. An overview of the features used here is given in Table 1. They comprise the MFCC used for applause sound detection in a previous work [10] and other features which similarly to

Table 1. Overview of features investigated and feature vector length n.

Abbreviation	Description	n
LLD	Set of low-level descriptors	20
LPC	Linear prediction coefficient	17
LSF	Line spectral frequency	17
MFCC	Mel-frequency cepstral coefficient	12
PLP	Perceptual linear prediction coefficient	12
RPLP	Relative spectra (RASTA) perceptual linear prediction	12

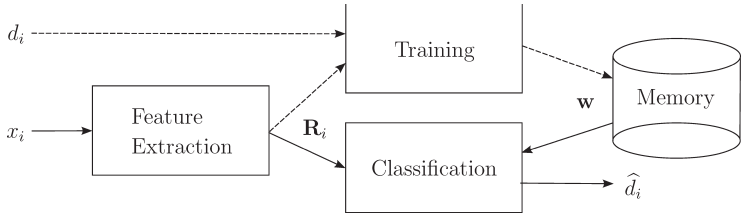

Fig. 1. General block diagram of supervised learning. Objects x_i are classified into classes labeled $\hat{d}_i$ using a set of features $\mathbf{R}_i$. Classifier is controlled by parameter set $\mathbf{w}$, which is obtained during a training process (dashed lines) from examples with known labels d_i and is stored in memory. Parameters are then used to classify patterns under test (solid lines).

the MFCC represent information on the gross spectral shape and originated from the speech recognition community. These features are complemented by a set of low-level descriptors (LLDs), which are frequently applied in automated analysis of music signals. The LLD comprises the spectral centroid, the spectral rolloff, the spectral flatness measure, the spectral flux, and the spectral spread. They are computed within four logarithmically spaced frequency bands in the range between 300 and 9600 Hz. The LPC and LSF are complemented by the power of the prediction error. The features are computed using a short-term Fourier transform with a frame size of 23 ms at a sampling rate of 44 100 Hz (STFT), 50% overlap, and a Hann window function. The resulting rate at which the features are extracted and processed is 86 Hz.

The algorithm processes input signals with one or more channels by processing each channel signal separately, or a downmix of them to reduce the computational load. In the following description of the features, X_k denotes a vector of K spectral coefficients.

1.1.1 Spectral Centroid

The spectral centroid SC equals the normalized frequency which determines the center of gravity of the spectrum. Here the centroid is computed from the magnitude spectrum of a subband signal,

$$\mathrm{SC} = \frac{1}{\sum_{k=0}^{K-1} |X_k|} \sum_{k=0}^{K-1} \frac{k+0.5}{K} |X_k|. \tag{1}$$

1.1.2 Spectral Spread

Considering a spectrum as a probability mass function with the frequencies being the values of the sample space and the coefficients relating to the probabilities, then the spectral centroid is the expectation value. Given this analogy the spectral spread SSp is defined as the variance of the distribution, that is, the spread of the spectrum around its centroid,

$$\mathrm{SSp} = \frac{1}{\sum_{k=0}^{K-1} |X_k|} \sum_{k=0}^{K-1} \left(\frac{k+0.5}{K} - \mathrm{SC} \right)^2 |X_k|. \tag{2}$$

1.1.3 Spectral Rolloff

The spectral rolloff SR determines the normalized frequency which divides the spectrum such that a determined percentage $\varrho \cdot 100\%$, with $0 \leq \varrho \leq 1$, of the signal energy is contained in the lower part,

$$\sum_{k=0}^{\mathrm{SR}(K-1)} |X_k^2| = \varrho \cdot \sum_{k=0}^{K-1} |X_k^2|. \tag{3}$$

1.1.4 Spectral Flux

The spectral flux SF is defined as the dissimilarity between spectra of successive frames [20] and is implemented by means of a distance metric. In this work the spectral flux is computed using the Euclidian distance according to Eq. (4), where B denotes the buffered spectrum of the preceeding frame,

$$\mathrm{SF} = \sqrt{\sum_{k=0}^{K-1} (|X_k| - |B_k|)^2}. \tag{4}$$

1.1.5 Spectral Flatness Measure

The spectral flatness measure SFM quantifies the deviation of a spectrum from a flat shape. Various definitions for the computation of the flatness of a vector or the tonality of a spectrum (which is related inversely to the flatness of a spectrum) exist [21], [22]. The spectral flatness measure SFM used here is computed as the ratio of the geometric mean to the arithmetic mean of the spectral coefficients of the subband signal,

$$\mathrm{SFM} = \frac{\exp\left\{ \left[\sum_{k=0}^{K-1} \log(|X_k|) \right] / K \right\}}{(1/K) \sum_{k=0}^{K-1} |X_k|}. \tag{5}$$

The geometric mean in Eq. (5) is computed from the arithmetic mean of the logarithmically transformed data.

1.1.6 Linear Prediction Coefficients

Linear prediction has been applied successfully for time series analysis and signal coding and is an important and efficient model for voiced speech in speech signal processing [23].

The LPCs used herein are the coefficients of an all-pole filter (or autoregressive model) which predicts the actual value $x(k)$ of a time series from the preceding values such that the squared error $E = \Sigma_k (\hat{x}_k - x_k)^2$ is minimized,

$$\hat{x}(k) = -\sum_{j=1}^{p} a_j x_{k-j}. \tag{6}$$

The LPCs are computed using the autocorrelation method [24]. Minimization of the error function by means of the gradient descent method leads to a set of normal equations whose coefficients are computed from the autocorrelation function of the signal. Due to the Toeplitz structure of the autocorrelation matrix, the set of equations can be solved efficiently using the Levinson–Durbin algorithm.

1.1.7 Mel-Frequency Cepstral Coefficients

MFCCs have been widely used as a front end for speech recognition and various audio content classification tasks. Their computation follows closely the implementation in [25]. The power spectra are projected according to the Mel scale with 13 linearly spaced bands and 27 logarithmically spaced bands using triangular weighting functions. The MFCCs are computed by taking the

logarithm and computing the discrete cosine transform for the first 13 basis functions.

1.1.8 Perceptual Linear Prediction Coefficients

Hermansky introduced PLP as a feature for speech analysis, which is more consistent with the human hearing than the conventional linear prediction analysis. The algorithm for computing the PLP extends the autoregressive model by several concepts of psychoacoustics [26]. The power spectra are projected according to a frequency-band resolution which resembles the human auditory system (in the original paper the Bark scale, which in this work is replaced by the Mel scale). The resulting spectra are then preemphasized according to an equal-loudness contour, which simulates the sensitivity of the human ear at different frequencies for a moderate sound level. The spectral coefficients are subsequently compressed by $y = x^{0.33}$ to approximate the power law of auditory intensity sensation.

The subsequent computations resemble the autocorrelation method of all-pole spectral modeling as used in linear prediction. An analogy to the autocorrelation function is computed using the inverse Fourier transform from which the autoregressive coefficients are computed using the Levinson–Durbin algorithm.

The resulting coefficients are further processed by applying the recursive formula, which is used for the conversion from LPC to cepstral coefficients [23]. Prior simulations (unpublished) have shown that the cepstrum conversion improves the classification results slightly.

1.1.9 Relative Spectra Perceptual Linear Prediction Coefficients

Relative spectra perceptual prediction coefficients, in short RASTA-PLP (in this paper abbreviated RPLP for convenience) were developed originally as a robust representation for speech recognition and enhancement in noisy environments [27]. The algorithm for computing the RPLP supplements the derivation of the PLP with an attenuation of "the spectral components that change more slowly or quickly than the typical range of change of speech" [27]. This attenuation is carried out using bandpass filtering of the logarithmically compressed subband envelope signals and subsequent exponential expansion.

1.2 Feature Processing

The features are further processed prior to feeding them into the classifier by means of filtering and normalization, as described in the following.

1.2.1 Delta Features

Delta features (or dynamic or derivative features) [28] have been applied successfully to speech recognition and audio content classification in the past. They are an estimate of the time derivative of each feature and are often computed by convolving the time sequence of a feature with a linear slope. The length of the FIR filter was set to 100 ms. Delta–delta features (or double-delta features) are obtained by applying the delta operation to the delta features. A computational more efficient method using first-order IIR high-pass filters is additionally used and compared in this work. A Butterworth filter with a cutoff frequency of 5.2 Hz is used here. Throughout the paper the different variations of the delta features are designated by Δ with superscripts f for FIR filter and i for IIR filter, and subscripts 1 for delta features, 2 for delta–delta features, and 12 for a combination of delta and delta–delta features. (For example, Δ_1^{i} designates delta features computed using an IIR filter.)

1.2.2 Sigma Features

The concept of sigma features is introduced here analogously to the delta features. The basic idea of sigma features is to complement the short-term observation by information derived from preceding signal frames. This is motivated by the fact that the characteristics of applause signals are varying slowly.

Sigma features are computed by integrating (or FIR low-pass filtering) the feature time series. They are especially useful when classifying short frames of audio data. Another approach, which has been used for applause detection (and other content analysis tasks), is compared to the new approach in the following. The classification of a single signal frame (as used for the STFT) may be prone to errors due to the short observation length and can be improved by classifying a larger signal segment (or a couple of frames). On the other hand it is important to use a short-term analysis for the feature extraction because audio signals are in general short-term stationary. One possibility to do so is to compute features on a frame-by-frame basis and to compute statistical moments (such as mean and variance) from a couple of subsequent frames (the observation window) and feed the classifier with it. This approach is not appropriate for the detection of the beginning and end of applause (segmentation) if the observation window is large. Furthermore it is very demanding with respect to memory. Shortening the observation window may lead to less robust classification results.

The basic idea of sigma features is to complement the short-term features with the integrated feature rather than to replace them. A computationally efficient method derived by analogy to delta features is to low-pass filter the feature time series. The filters used here have cutoff frequencies of 1.4 and 2.4 Hz for the IIR and FIR filters, respectively.

The sigma features are designated by Σ with superscripts f for FIR filter or i for IIR filter, and subscripts 1 for sigma features, 2 for sigma–sigma features, and 12 for a combination of sigma and sigma–sigma features. (For example, Σ_{12}^{i} designates sigma features computed using an IIR filter with additional sigma–sigma features.)

1.2.3 Centering and Variance Normalization

All features are processed by means of centering, that is, removing the mean of the feature, and variance

normalization, namely, dividing the features by their standard deviation and thereby normalizing them to have unit variance. The feature means and standard deviations are computed from the set of training data and stored for application of the classifier together with the model parameters.

1.3 Classification

The task of detecting a key sound such as applause or speech relates to the classification into two classes: the key sound is present or not present. However, a finer distinction is achieved by defining six classes resembling different SNRs. Thereby the influence of the SNR on the performance can be investigated. For the task of binary classification, the original classes can be assigned to two classes, as described in Section 3.1.

The classification is computed using a multilayer perceptron (MLP, also known as feedforward neural network) [29], which is probably the most widely used architecture of neural networks for practical applications. The networks used here have one hidden layer with 10 neurons with logistic activation functions. The scaled conjugate gradients algorithm from the NetLab toolbox [30] is used for training of the MLP. The maximum number of iterations is set to 100. The number of hidden neurons and the maximum number of iterations were determined in prior test simulations.

For comparison a linear classifier based on the Fischer linear discriminant (FLD) [31] and a support vector machine (SVM) [32] with two different kernel functions (a linear kernel and a radial basis function) are used for the binary classifications. The FLD has been chosen because of its simplicity as a computationally efficient alternative. It works by projecting the sample space onto a line, which yields an efficient discrimination of the two classes. The discriminant function maximizes the ratio of the difference of the class means (the between-class scatter) and a measure for the dispersion of the samples within each class (the within-class scatter) and can be computed in a closed form.

The SVM is a binary classifier which became popular in recent years. It can be viewed as a linear classifier which works in a higher dimensional feature space. By projecting the data in a sufficiently high-dimensional space a separation by a hyperplane can be achieved ideally. The optimal linear separation maximizes the distance between the hyperplane and the nearest training vectors. The parameters of the discrimination function are computed as the solution of a quadratic programming problem. For an introduction to SVM the reader is referred to [33].

2 EXPERIMENTS AND RESULTS

2.1 Data Sets

The data set for training and test comprises 210 excerpts of commercial recordings of between 9- and 30-second length each. It contains live recordings of different musical genres and stand-up comedy (speech). The total number of classified feature vectors (the number of STFT frames) is about 730 000. The audio signals were labeled manually by one listener with labels R_n, with $n = 1, ..., 6$. The labels were obtained by manually setting markers in an audio editor and converting the markers into a reference signal using a MATLAB script. A description of the reference labels is given in Table 2. Please note that it is not discriminated between synchronous and asynchronous applause as in [11].

The labels are fuzzy and it can be argued that a single labeling listener is not able to provide consistent labels for the intermediate states $\{R_2, ..., R_5\}$ for each item. However, there are two alternatives with comparable drawbacks. Averaging the labels provided by many listeners is very demanding with respect to workforce. More reliable and even continuous labels (with a finer resolution than six steps which in addition allow for the application of regression methods instead of classifiers) are obtained by using separated applause and background sounds, mixing them at varying levels and computing the labels from the ratios of power or loudness of both signals. However, such items are artificial (in any case a drawback) and the labels do not necessarily relate to human perception (which concededly is not a drawback for all applications mentioned). Fig. 2 shows a histogram of the class labels in the data set. It is shown that

- The nonapplause signals (labeled R_1) are predominant
- If applause is present, it is often superimposed by other sounds, as in the signals with labels $R_2, ..., R_5$.

The data set is also used for the classification of two or three classes by assigning the original class labels R_n, with $n = 1, ..., 6$, to class labels B_n, $n = 1, ..., 2$ (for the binary classification) and T_n, $n = 1, ..., 3$ (for the ternary classification), respectively. The class mappings M used here are shown in Table 3 for the binary classification and in Table 4 for the ternary classification. Samples with no or less applause are labeled with small indices and samples with more applause are labeled with larger indices. For example, the class mapping M_{b2} with $B_1 =$

Table 2. Description of reference labels.

Label	Description
R_1	No applause
R_2	Spurious applause
R_3	Less applause than other sounds
R_4	Applause is perceived equally loud compared to other sounds
R_5	Recording contains more applause than other signals
R_6	Only applause

$\{R_1, R_2\}$, $B_2 = \{R_3, ..., R_6\}$ relates to the task of discriminating signals with no or little applause from signals with more applause.

2.2 Evaluation Procedure and Metrics

The experimental results presented in the following are derived using a tenfold cross validation, where 90% of the data set are used for training of the classifier and the remaining 10% are used for the test. The classification results are evaluated using the mean error, that is, the mean absolute differences between reference and predicted class labels. The confusion matrix is presented for selected experiments, and the references are shown in rows and the estimates in columns.

2.3 Feature Sets and Delta Features

The first experiment compares the classification performance of all feature sets under investigation and the different delta features. The classification result is shown in Fig. 3 for the six-class problem. Please note that in the following figures the range of the y axis is adapted for better visibility.

Table 3. Binary class mapping.

	B_1	B_2
M_{b1}	1	2, ..., 6
M_{b2}	1, 2	3, ..., 6
M_{b3}	1, ..., 3	4, ..., 6
M_{b4}	1, ..., 4	5, 6
M_{b5}	1, ..., 5	6

It is shown that

- The classification performance for MFCC, PLP, and LLD is comparable.
- The RPLPs perform distinctly worse.
- Delta features yield on average an improvement.
- The Δ^{i} perform better than the Δ^{f} in most cases.
- The additional use of the Δ^{i}_{12} yields slight improvements compared to the Δ^{i}_{1} alone.
- Using Δ^{i}_{2} without Δ^{i}_{1} impairs the performance.

The different classification results obtained with PLP and RPLP can be attributed to the band-pass filtering of the subband trajectories in the RPLP computation. This processing is beneficial for the original application to speech processing in noisy environments but for the application at hand it removes two important properties of subband signals of applause sound, namely, the occurrence of transients and slowly varying steady components.

2.4 Performance of Sigma Features

The next experiment compares the classification results obtained with sigma features. The results are shown in Fig. 4 and indicate the following.

Table 4. Ternary class mapping.

	T_1	T_2	T_3
M_{t1}	1	2, ..., 5	6
M_{t2}	1, 2	3, ..., 5	6
M_{t3}	1	2, ..., 4	5, 6
M_{t4}	1, 2	3, 4	5, 6

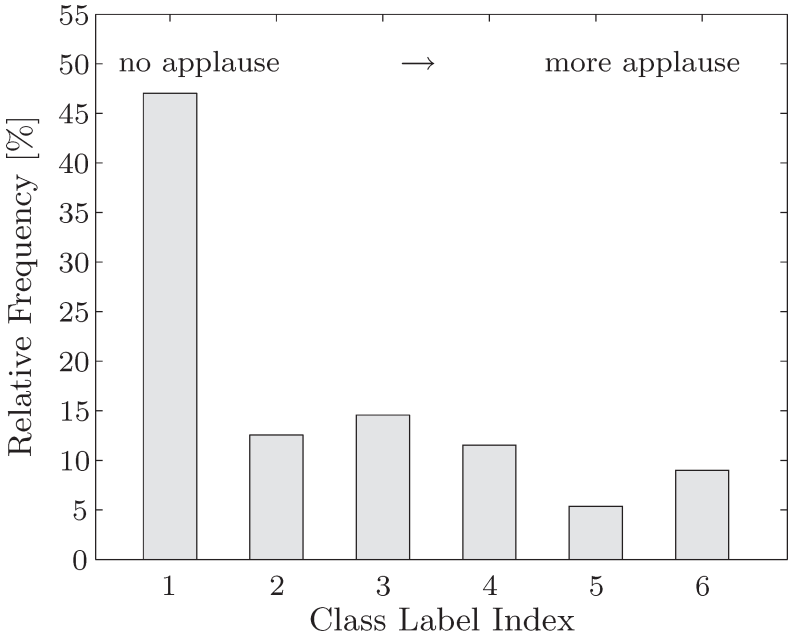

Fig. 2. Histogram of class labels in data set.

- The use of sigma features improves the classification result. The benefit is larger than when using delta features.
- The differences between Σ^f and Σ^i are small compared to the improvement over using no sigma features.
- The differences between Σ_1, Σ_2, and Σ_{12} are small.

Due to the small differences between the sigma variations the Σ_1^i are chosen for the next experiments. In comparison to sigma–sigma features, this leads to a smaller feature space and thus reduces both the computational complexity and the risk of overfitting the classifier.

2.5 Combining Delta Features with Sigma Features

A consistent continuation of the experiments is to combine delta and sigma features, but only the four best feature sets (LSF, MFCC, PLP, and LLD) are investigated further. Fig. 5 illustrates the experimental results for delta and sigma features and combinations of both.

Here the outcome of the evaluation is not as clear as in the preceding experiments. It is shown that a combination of delta and sigma features improves the results. However, there is no combination that outperforms the others consistently for all feature sets. It should be noted that these combinations lead to an increased dimensionality of the feature space, and care has to be taken with respect to overfitting. Consequently the experiments are continued with feature pruning to reduce the dimensionality of the feature space.

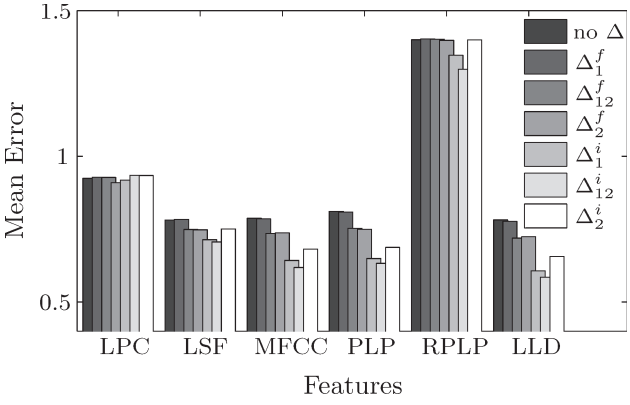

Fig. 3. Classification results for different feature sets and delta features; MLP with six classes.

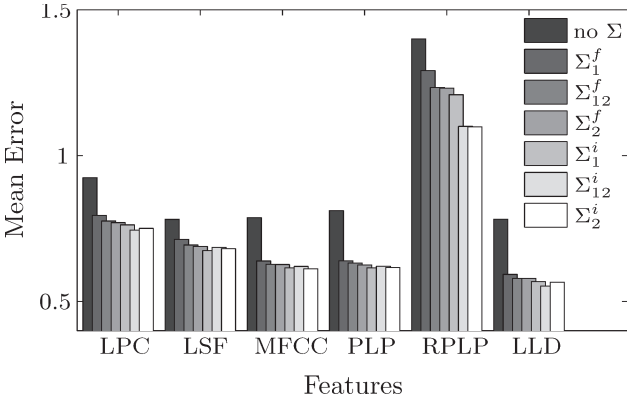

Fig. 4. Classification results for different feature sets and sigma features; MLP with six classes.

2.6 Pruning of Features

The set of LLDs has shown the best classification performance in the preceding experiments. The LLDs have been selected from a large set of possible descriptors to limit the number of features. The selection was admittedly based on heuristics. It is now investigated experimentally which of the features do not contribute to the classification by means of backward selection, that is, removing single features from the set until the performance declines.

Fig. 6 illustrates the classification results as a function of the pruned features with and without delta features. It is shown that discarding either SR or SC does not have an impact on the performance, whereas discarding both has. Further investigation reveals high correlations between both features computed within the same subband (larger than 0.93). Discarding other features than SR or SC leads to a decrease in the recognition rate.

The backward selection is applied to investigate the influence of the frequency band used for the feature extraction. The results obtained when discarding the features corresponding to one of the four subbands are shown in Fig. 7 (again with and without delta features) and indicate the following.

- The pruning of subband features increases the error.
- The first and third subbands have the largest influence on the classification performance.

At this point we consider combining the feature sets tested but restrict the number of experiments. The preceding experiments indicate that MFCC and PLP yield good and comparable results. It is therefore of interest to investigate combinations of LLD with MFCC, which are preferred because of the lower computational complexity compared to PLP.

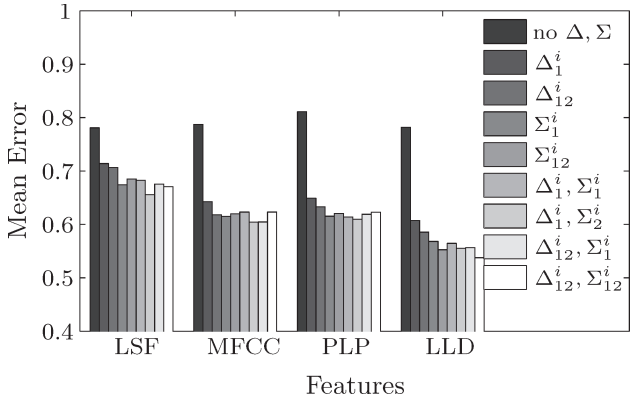

Fig. 5. Classification results for delta and sigma features using MLP and six classes.

In order to further reduce the dimensionality of the feature space, pruning experiments are carried out for the MFCC, with the results shown in Fig. 8. In this test the higher coefficients are discarded one after another. The results indicate that the higher order coefficients have no significant influence on the classification result.

2.7 Combining Different Feature Sets

The experiment illustrated in Fig. 9 compares the feature sets with

- Nine MFCCs
- The subset of LLDs comprising SC, SF, SFM, and SSp
- A combination of the LLD subset and the nine MFCCs.

It can be seen that the combination of MFCC and LLD leads to the best results.

The confusion matrix is shown in Table 5. The number of misclassifications for the items labeled R_1 and R_6 (that is, where no applause or only applause is present) is rather small. Misclassifications occur more often in mixtures of applause and nonapplause sounds.

2.8 Binary and Ternary Classification with Class Mapping

With the reference data at hand classifying the problems with coarser classes can be investigated by mapping the original classes to two or three classes, as described in Section 2.1. The results for the binary classification are illustrated in Fig. 10 for the combination of LLD and MFCC. They indicate the following.

- The classification of two classes with class mappings $M_{b3}...M_{b5}$ yields good results, with the smallest error rates below 5%.
- It is difficult to detect applause if the recording is dominated by other sounds.
- By using delta and sigma features the results are improved.
- Sigma features yield more improvement than delta features.

The binary classification can be computed alternatively from the classification results using six classes and subsequently mapping of the estimated classes to two

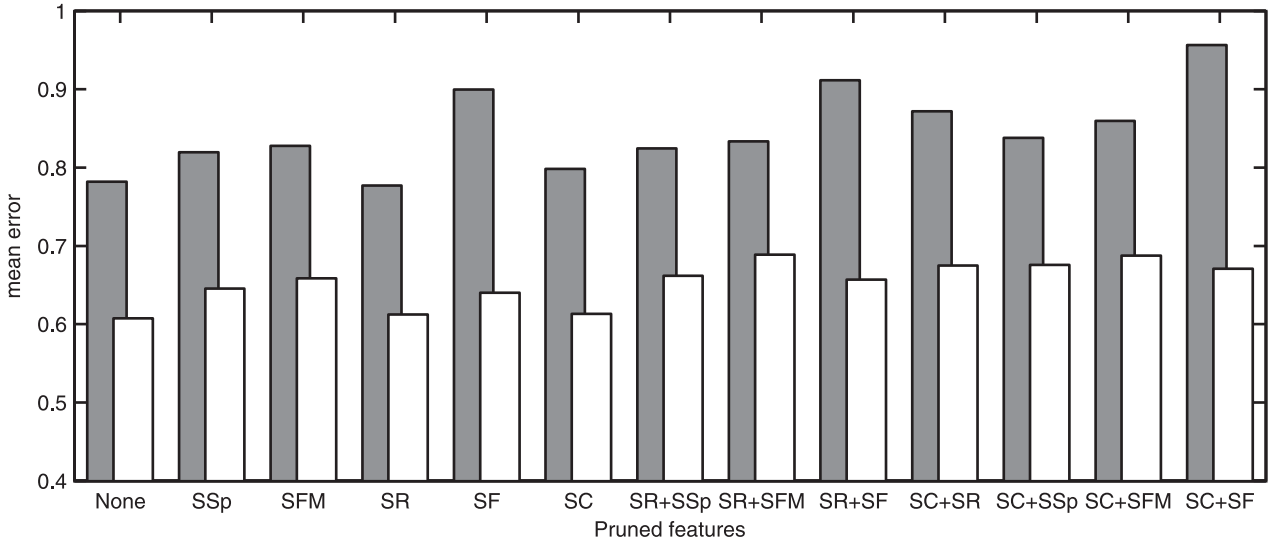

Fig. 6. Classification results of subsets of LLD without (dark) and with Δ_1^i (white).

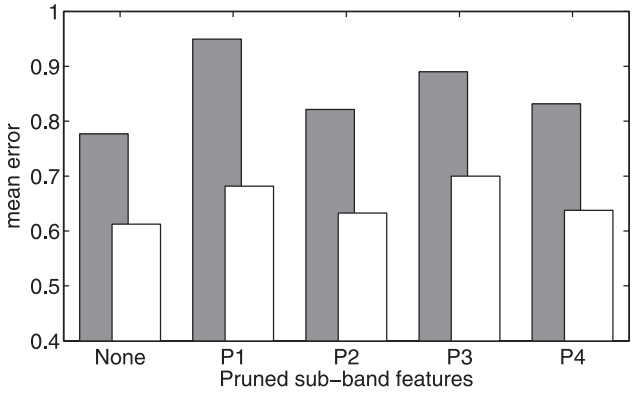

Fig. 7. Pruning of subband LLD features without (dark) and with Δ_1^i (white).

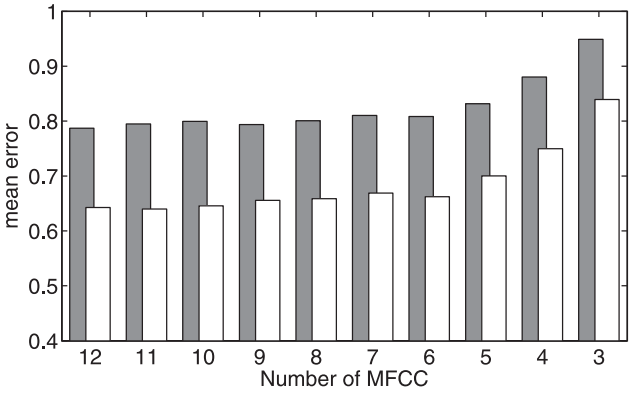

Fig. 8. Classification results of subsets of MFCC without (dark) and with Δ_1^i (white).

classes. Interestingly enough this leads to similar results. This is exemplified by the confusion matrix in Table 6 for class mapping M_{b4} with Δ_{12}^{i} and Σ_{1}^{i}. The similarity between the results of both approaches also shows that the feature selection using a larger number of classes is valid even for a classification task with fewer classes.

In order to investigate the influence of the interfering sounds on the classification result the confusion matrix in Table 7 is computed with the original class labels R_1, ..., R_6. It is shown that the number of misclassifications for samples with R_1, R_2, and R_6 (samples dominated by either applause or nonapplause sounds) are below 1 percent. Misclassifications occur more often if applause and nonapplause sounds occur simultaneously with similar intensity (R_3 and R_4) and at the boundary of the class mapping (R_4 and R_5). An example for the binary classification with class mapping M_{b3} is illustrated in Fig. 11.

A ternary classification problem can be approached in a similar way, with the results shown in Fig. 12. As expected, the recognition performance decreases with an increasing number of classes. The conclusions regarding the delta and sigma features are the same as for the binary classification experiment.

2.9 Comparison of Classifiers

The performances of different classifiers (FLD, MLP, and SVM with linear kernel and SVM with radial basis functions) are evaluated and compared in the following for binary classification with nine MFCC coefficients and with LLD separately, both without delta or sigma features. Delta or sigma features have been discarded in order to compare the classifiers without relying on a feature set that is selected based on experiments with one of the classifiers under test (namely, the MLP).

As shown in Fig. 13, the FLD performs distinctly worse compared to the other classifiers. The differences between

Table 5. Confusion matrix for multiclass classification using MFCC and LLD and Δ_{12}^{i} with Σ_{1}^{i}.

	E_1	E_2	E_3	E_4	E_5	E_6
R_1	44.8	0.8	2.6	0.3	0.0	0.0
R_2	7.5	0.8	3.3	0.8	0.1	0.2
R_3	6.3	1.0	5.8	1.1	0.3	0.3
R_4	0.9	0.3	1.3	5.4	0.3	0.6
R_5	0.3	0.1	0.7	1.5	0.8	2.5
R_6	0.08	0.1	0.1	0.3	0.5	8.1

Table 6. Confusion matrix for binary classification using M_{b4}, with MFCC, LLD, Δ_{1}^{i}, and Σ_{1}^{i}.

	E_1	E_2
B_1	83.0 / 83.0	2.0 / 2.0
B_2	3.0 / 3.1	12.0 / 11.9

*Matrix element (B_k, E_l) indicates percentage of objects with reference class k classified as class l of total number of classifications. Left number shows result obtained by training two classes; right number is result obtained by classifying into six classes and subsequently assigning estimated class labels to two classes.

Table 7. Confusion matrix for binary classification using M_{b4} and original reference labels.

	E_1	E_2
R_1	47.8	0.1
R_2	12.6	0.2
R_3	14.3	0.7
R_4	8.2	1.1
R_5	2.3	3.3
R_6	0.6	8.7

*Matrix element (R_k, E_l) indicates percentage of objects with reference class k classified as class l of total number of classifications.

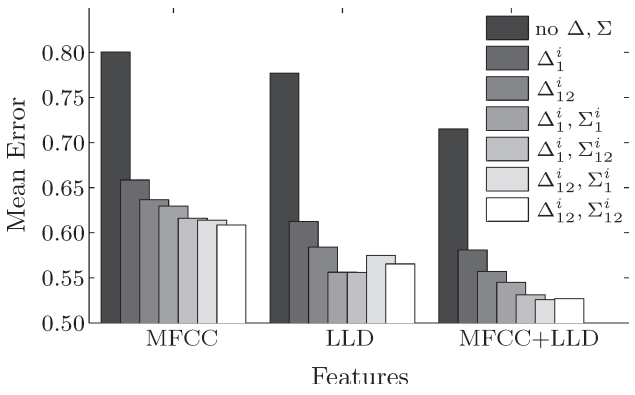

Fig. 9. Classification results for MFCC, LLD, and combinations of both.

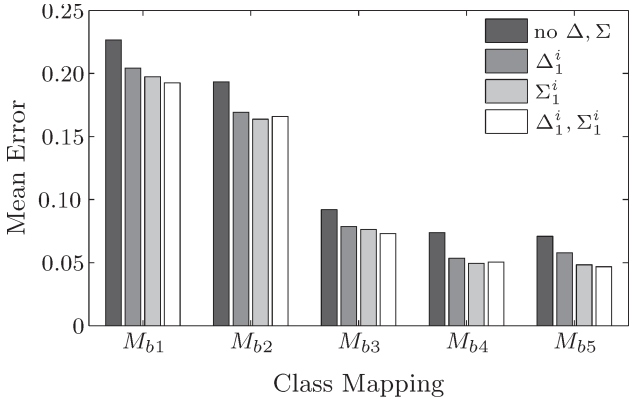

Fig. 10. Classification result for binary class mapping using a combination of LLD and MFCC.

MLP and SVM with radial basis functions are rather small, but in view of the computational complexity and memory requirements the MLP might be more appropriate than the SVM in some practical applications.

3 APPLICATION EXAMPLE

The method presented is implemented in the signal classification module of the reference quality encoder as currently used for the MPEG standardization effort of USAC. The feature set comprises the combination of LLD and nine MFCCs as described in Section 2.7 with delta features Δ_1^i and sigma features Σ_1^i. The MLP as described in Section 1.3 is applied for the classification using six classes. Its result is mapped to a binary decision with class mapping M_{b4}, as described in Table 3. The use of multiple classes allows for more flexibility in the tuning of the algorithm and for possible postprocessing of the classification result.

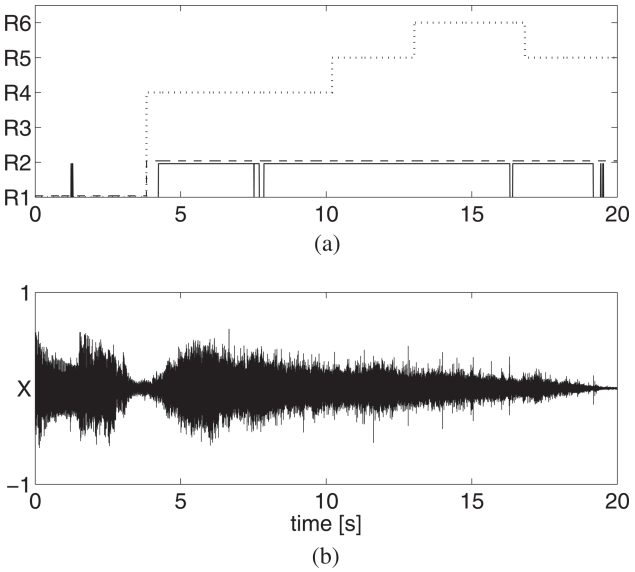

Fig. 11. (a) Classification result. (b) Example signal. ··· original reference; – – – mapped reference for binary classification; —— classification result.

The purpose of classification is to activate an applause processing mode within the parametric coding of stereo content, which uses a more appropriate time–frequency resolution and a dedicated transient processing.

4 CONCLUSIONS

A comprehensive investigation of the detection of applause sounds in audio signals in real time with low latency has been presented. Various features used in state-of-the-art audio content classification methods have been compared and combined. Delta features were evaluated and sigma features have been proposed by analogy to the delta features. In the experiments a combination of MFCC and LLD yielded the best classification performance. These results were obtained with the MLP and the SVM with radial basis functions as classifiers. It was shown that the sigma features improve the classification results more than the delta features. The computationally efficient recursive algorithm for delta and sigma features leads to better, or at least similar, results (depending on the feature set and classification task) compared to their nonrecursive counterparts. The impact of the intensity of the applause compared to nonapplause sounds within the sound mixture on the recognition performance has been investigated.

For the binary classification task it was shown that misclassifications occur more often if applause and nonapplause sounds occur simultaneously with similar intensity.

The results show that the method presented for detecting applause within a sound mixture is prone to errors, but still recognizes more than 95% of applause correctly if a discrimination of dominant applause versus no or spurious applause is desired. This final result indicates that the method presented can be beneficial for various applications, including audio coding and other applications where low latency is indispensable, but this

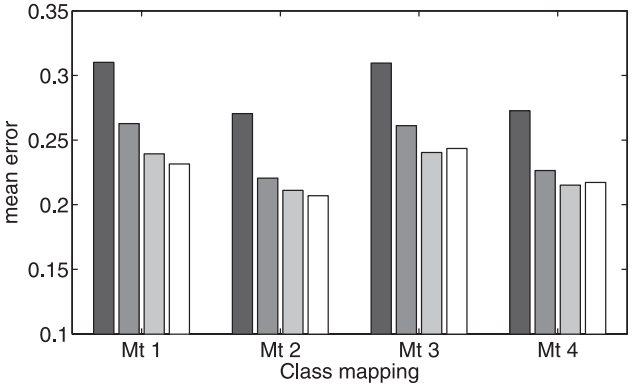

Fig. 12. Classification result for ternary class mapping using a combination of LLD and MFCC. Same conditions for delta and sigma features as in Fig. 10.

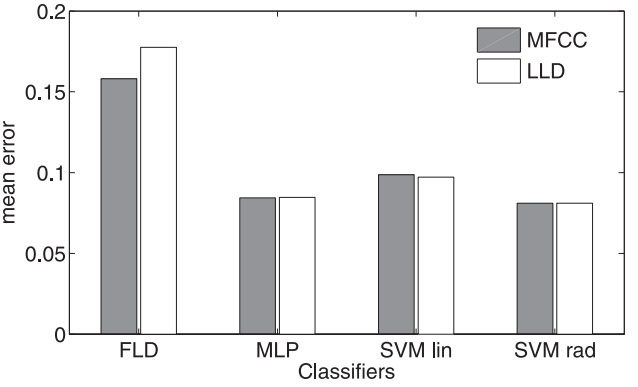

Fig. 13. Comparison of classifiers FLD, MLP, SVM with linear kernel and SVM with radial basis function for class mapping M_{b4} for two feature sets.

needs to be investigated further in combination with the respective application.

5 ACKNOWLEDGMENT

The author wishes to thank Achim Kuntz for proofreading.

6 REFERENCES

[1] E. Wold, T. Blum, D. Keislar, and J. Wheaton, "Content-Based Classification, Search, and Retrieval of Audio," *IEEE Multimedia*, vol. 3, pp. 27–36 (1996).

[2] M. Casey, "MPEG-7 Sound-Recognition Tools," *IEEE Trans. Circ. Sys. Video Technol*., vol. 11, pp. 737–747 (2001).

[3] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, "Audio Events Detection Based Highlights Extraction from Baseball, Golf and Soccer Games in a Unified Framework," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (2003).

[4] M. Xu, N. Maddage, C. Xu, M. Kankanhalli, and Q. Tian, "Creating Audio Keywords for Event Detection in Soccer Video," in *Proc. IEEE Int. Conf. on Multimedia and Expo* (2003).

[5] Y. Nakajima, Y. Lu, M. Sugano, A. Yoneyama, H. Yanagihara, and A. Kurematsu, "A Fast Audio Classification from MPEG Coded Data," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing* (1999).

[6] R. Cai, A. Hanjalic, H. J. Zhang, and L. H. Cai, "A Flexible Framework for Key Audio Effects Detection and Auditory Context Interface," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 14, pp. 1026–1039 (2006).

[7] R. Cai, L. Lu, H. J. Zhang, and L. H. Cai, "Highlight Sound Effects Detection in Audio Stream," in *Proc. IEEE Int. Conf. on Multimedia and Expo* (2003).

[8] K. Biatov, "The Method of an Audio Data Classification and Segmentation," in *NATO Advanced Study Institute—Computative Noncommutative Algebra and Applications* (2003).

[9] J. Loeffler, K. Biatov, and J. Koehler, "Automatic Extraction of MPEG-7 Audio Metadata Using the Media Asset Management System iFinder," in *Proc. AES 25th Int. Conf. on Metadata for Audio* (London, UK, 2004 June 17–19).

[10] R. Jarina and J. Olajec, "Discriminative Feature Selection for Applause Sound Detection," in *Proc. 8th Int. Workshop on Image Analysis for Multimedia Interactive Services* (2007).

[11] J. Olajec, C. Erkut, and R. Jarina, "GA-Based Feature Selection for Synchronous and Asynchronous Applause Detection," in *Proc. Finnish Signal Processing Symp*. (2007).

[12] Y. X. Li, Q. H. He, S. Kwong, T. Li, and J. C. Yang, "Characteristics-Based Effective Applause Detection for Meeting Speech," *Signal Process*., vol. 89, pp. 1625–1633 (2009).

[13] J. Herre, C. Faller, C. Ertel, J. Hilpert, A. Hoelzer, and C. Spenger, "MP3 Surround: Efficient and Compatible Coding of Multichannel Audio," presented at the 116th Convention of the Audio Engineering Society, *J. Audio Eng. Soc*. (*Abstracts*), vol. 52, p. 793 (2004 July/Aug.), convention paper 6049.

[14] J. Herre, K. Kjörling, J. Breebaart, C. Faller, S. Disch, H. Purnhagen, J. Koppens, J. Hilpert, J. Rödén, W. Oomen, K. Linzmeier, and K. S. Chong, "MPEG Surround—The ISO/MPEG Standard for Efficient and Compatible Multichannel Audio Coding," *J. Audio Eng. Soc*., vol. 56, pp. 932–955 (2008 Nov.).

[15] Max Neuendorf et. al., "A Novel Scheme for Low Bitrate Unified Speech and Audio Coding MPEG RM0," presented at the 126th Convention of the Audio Engineering Society, (*Abstracts*) www.aes.org/events/126/126thWrapUp.pdf, (2009 May), convention paper 7713.

[16] C. Faller, "Parametric Coding of Spatial Audio," in *Proc. 7th Int. Conf. on Digital Audio Effects (DAFx)* (2004).

[17] G. Hotho, S. van de Par, and J. Breebart, "Multichannel Coding of Applause Signals," *EURASIP J. Adv. Signal Process*., pp. 1–9 (2008).

[18] Z. Neda, E. Ravasz, Y. Brechet, T. Vicsek, and A. L. Barabasi, "The Sound of Many Hands Clapping," *Nature*, vol. 403, pp. 849–850 (2000).

[19] E. C. Cherry, "Some Experiments on the Recognition of Speech, with One and with Two Ears," *J. Acoust. Soc. Am*., vol. 25, pp. 975–979 (1953).

[20] P. Masri, "Computer Modelling of Sound for Transformation and Synthesis of Musical Signals," Ph.D. thesis, University of Bristol, Bristol, UK (1996).

[21] ISO/IEC JTC1/SC29/WG11 Moving Pictures Expert Group, "MPEG-7, Information Technology—Multimedia Content Description Interface—Part 4: Audio," final draft, Int. Std. 15938-4, International Standards Organization, Geneva, Switzerland (2002).

[22] ISO/IEC JTC1/SC29/WG11 Moving Pictures Expert Group, "MPEG-1, Coding of Moving Pictures and Associated Audio for Digital Storage Media," final draft, 11172-3, International Standards Organization, Geneva, Switzerland (1993).

[23] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ, 1978).

[24] J. Makhoul, "Linear Prediction: A Tutorial Review," *Proc. IEEE*, vol. 63, pp. 561–580 (1975).

[25] M. Slaney, "Auditory Toolbox: A Matlab Toolbox for Auditory Modeling Work, version 2," Tech. Rep., Interval Research Corp. (1998).

[26] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis for Speech," *J. Acoust. Soc. Am*., vol. 87, pp. 1738–1752 (1990).

[27] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. Speech Audio Process*., vol. 2, pp. 578–589 (1994).

[28] S. Furui, "Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spec-

trum," *IEEE Trans. Acoust., Speech, Signal Process*., vol. 34, pp. 52–59 (1986).

[29] C. M. Bishop, *Pattern Recognition and Machine Learning* (Springer, New York, 2006).

[30] I. T. Nabney, *NetLab—Algorithms for Pattern Recognition* (Springer, New York, 2002).

[31] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. (Wiley, New York, 2000).

[32] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining Knowledge Discov*., vol. 2, pp. 121–167 (1998).

[33] B. Schoelkopf and A. J. Smola, *Learning with Kernels* (MIT Press, Cambridge, MA, 2002).

THE AUTHOR

Christian Uhle received Dipl.-Ing. and Ph.D. degrees from the Technical University of Ilmenau, Germany, in 1997and 2008, respectively.

His professional career began in the field of image processing and pattern recognition at the Technical University of Ilmenau and the Technical School of Chania, Greece. From 1998 to 2000 he worked in the Computer Architecture Department of the Technical University of Ilmenau on the development of real-time operating systems for digital signal processors. In 2000 he joined the Fraunhofer Institute for Digital Media Technology, Ilmenau, and was engaged in research on semantic analysis of audio signals. Since 2006 he has been with the Fraunhofer Institute for Integrated Circuits (IIS), Erlangen, Germany, and is working in research and development. His research interests include semantic analysis and processing of audio signals, blind source separation, and digital audio effects.

Dr. Uhle is a member of the Audio Engineering Society.