Linear Combining of Audio Features for Signal Classification in Ad-hoc Microphone Arrays

Sebastian Gergen, Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, 44780 Bochum, Germany
Email: {sebastian.gergen, rainer.martin}@rub.de
Web: http://www.ruhr-uni-bochum.de/ika/

Abstract

Audio signals are often corrupted by signal contributions from competing sources and reverberation in the acoustic environment. In an audio signal classification task these effects introduce a mismatch between test and training data, which decreases the classification accuracy. When multiple sources are simultaneously active and captured by multiple ad-hoc distributed microphones in a room, it is of interest to determine the type of each source based on the captured signal mixtures. Obviously, the microphones closest to a particular source are most suitable for its classification. However, it is not clear how to combine signal features extracted from the microphone signals in an ad-hoc array in order to classify the source signals reliably. In this contribution different data combination strategies are introduced. The resulting classification performance is analyzed based on simulations and audio recordings. When information from microphones within the critical distance of a source is combined with information from the other microphones in the room, a high classification accuracy can be obtained.

1 Introduction

The classification of audio signals into predefined categories, e.g. speech, music and noise, is an important ingredient of many audio-related applications like hearing aids or mobile phones [1]. Other fields of application are, e.g. teleconference scenarios and automatic speech recognition systems [2]. The classification of clean and undistorted signals for general classes like speech/music/noise can be done quite accurately, already. However, in a real world scenario, environmental conditions like room reverberation and simultaneously active sound sources are often inevitable. These effects generate signal distortions which reduce the performance of classification algorithms as they introduce a mismatch between training and test conditions.

One approach to obtain robust classification results is to utilize the spatial distribution of microphones provided by ad-hoc microphone arrays. These arrays are composed of, e.g., mobile devices like smartphones, tablet computers and laptops which allow for audio signal capturing and processing and can be connected wirelessly. The position of a device in a room in this context is not known in general [3]. Ad-hoc arrays (not exclusively wireless) have been used so far in the context of blind-source separation, e.g., in [4]. Further, noise reduction and beamforming algorithms based on distributed sensors are introduced, e.g., in [5, 6]. For signal classification, setups with multiple microphones are used, e.g., for surveillance systems [7] or military purposes [8]. However, microphone arrays in the conventional sense with exactly known sensor positions are considered in these systems.

In [9], an algorithm is introduced which estimates clusters of microphones in ad-hoc arrays based on extracted audio features. Typically, a cluster comprises the microphones which are in the vicinity to a particular source in the acoustic environment. The feature extraction is considered to be performed on each device individually. By the use of a compact feature representation of the audio signals, just a small amount of data has to be transmitted over the network and synchronization needs can be relaxed. Signal classification is then performed for each cluster, e.g., in one central device.

In this contribution we focus on different feature vector combination strategies to merge data which is provided by the microphone network. To retain flexibility in practical applications, the training of the classifier is performed on single channel data and thus, the combination of the feature vectors is necessary. Unlike [9] where clusters of microphones are estimated by an fuzzy clustering algorithm, we here employ a manual assignment of microphones into clusters to be independent of potential errors in the microphone clustering step. Then, a least-squares based feature combination is applied with the aim of achieving a high classification accuracy of audio sources in simulations and real audio recordings.

The remainder of this paper is organized as follows. In Section 2 the algorithm and its key components are outlined. The combination of feature vectors is explained in Section 3. In Section 4 we introduce the evaluation setups and present the results in Section 5. Finally, we conclude the paper in Section 6.

2 Algorithm components

For our investigations, we consider a spatial distribution of simultaneously active sound sources and ad-hoc distributed sensors in a room. We assume that some microphones are relatively close to an active source, i.e. within the critical distance where the direct path sound energy is larger than the reverberated sound energy [10]. Other microphones are located somewhere in the room, and possibly outside of the critical distances of all sources. Based on this spatial distribution we assign microphones to the sources: we form one cluster of microphones for each source and one additional cluster with the microphones which pick up mostly reverberation (see Fig. 2 for an example). Note, that the assignment of microphones to sources can be solved by e.g. a fuzzy clustering algorithm as shown in [9] when the position of the microphones and sources is not known. For the classification of a source, we then want to utilize the data extracted from audio signals which are captured by multiple receivers. Thus, we consider the structure shown in Fig. 1, which comprises a feature combination step, which would not be necessary if we would use just one input feature vector in a single channel classification system. The focus of this contribution is to analyze the influence of the feature combination step onto the classification performance.

2.1 Feature Computation

For clustering and classification, audio data is transformed into a compact feature vector. For our investigation we utilize a cepstro-temporal representation of the signal as these audio features have shown good classification results before [11]. The computed feature set provides a high classification accuracy in experiments with anechoic and clean data [12]. The feature set is based on the MFCC coefficients $X_{\text{mfcc}}(\eta, b)$ of a signal with the cepstral coefficient index η and the time frame $b = 0, 1, \ldots, B-1$ which is computed as described in [9]. The short-time MFCC modulation spectrum $\hat{X}_{\text{mfcc}}(\nu, \eta, c)$ is derived using a sliding window DFT:

$$\hat{X}_{\text{mfcc}}(\nu, \eta, c) = \sum_{\ell=0}^{L-1} X_{\text{mfcc}}(\eta, cQ + \ell) e^{-j\frac{2\pi\ell\nu}{L}}, \qquad (1)$$

where starting at sub-frame index $b = cQ$, the sliding window considers L consecutive sub-frames. The modulation frequency bin index is specified by $\nu = 0, 1, \ldots, L/2$ and c and Q denote the modulation window index and shift, respectively, with $c =$

ISBN 978-3-8007-3640-9

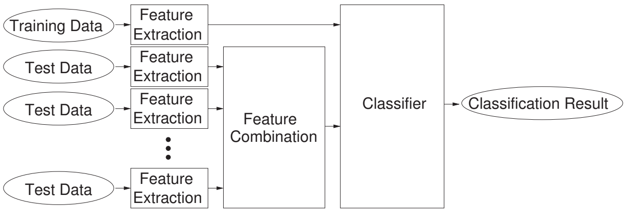

Figure 1: Structure of a multi-channel classification system which combines several feature vectors to obtain one classification result.

$0,1,\dots,C_T-1$ [12]. Then, the absolute values of these modulation spectra are averaged over all C_T frames (2) and cepstral modulation ratios (CMR) are computed to approximate the modulation spectrum (3):

$$\tilde{X}_{\mathrm{mfcc}}(\nu,\eta)=\frac{1}{C_T}\sum_{c=0}^{C_T-1}|\hat{X}_{\mathrm{mfcc}}(\nu,\eta,c)|, \tag{2}$$

$$r_{\nu_1|\nu_2}(\eta)=\frac{\sum_{\nu=\nu_1}^{\nu_2}\tilde{X}_{\mathrm{mfcc}}(\nu,\eta)}{(\nu_2-\nu_1+1)\tilde{X}_{\mathrm{mfcc}}(0,\eta)}. \tag{3}$$

Note, that in (3) the average of several modulation frequency bands within $\nu_1\leq\nu\leq\nu_2$ is normalized on the zeroth modulation frequency band. Finally, for our feature vector we stack two CMRs as well as $\bar{X}_{\mathrm{mfcc}}(\eta)$ into a vector, where

$$\bar{X}_{\mathrm{mfcc}}(\eta)=\frac{1}{L}\sum_{\nu'=0}^{L-1}\tilde{X}_{\mathrm{mfcc}}(\nu',\eta) \tag{4}$$

is the modulation spectrum averaged over all modulation frequencies ν for each MFCC bin η which serves as an estimate of the root mean square energy in each cepstral bin.

In nearly every realistic situation, room dependent reverberation contaminates the audio signals. The source-receiver transmission path determines the reverberation and can be represented by the room impulse response (RIR). A change in the source-receiver setup or other room properties causes changes on the RIR. Cepstral mean normalization (CMN) is one approach to reduce the effect of reverberation in audio signal processing. By averaging over a certain amount of time and subtracting this average from the cepstrum X_{mfcc}, the influence of reverberation can be reduced [13]. An improved classification performance using CMN was already observed in [9], therefore CMN is applied in this investigation, too.

2.2 Supervised Classification

The classification algorithm considered in our investigation is a linear discriminant analysis (LDA) classifier [14]: An observation $\mathbf{v}$ is assigned to that class ω_s which maximizes the posterior probability $P(\omega_s|\mathbf{v})$, where $s=1,\dots,S$ and S is the number of possible classes. This posterior probability can be expressed in terms of a-priori probabilities $P(\omega_1),P(\omega_2),\dots,P(\omega_S)$ and trained probability densities $p(\mathbf{v}|\omega_s)$ by the application of Bayes rule [14]. A multivariate normal distribution for the probability density and a pooled estimate of the feature covariance matrix across all classes is assumed.

3 Feature Combination Strategies

We utilize an ad-hoc array of D microphones. An A-dimensional feature vector $\mathbf{v}_d$ is extracted from the captured signal of each microphone m_d, with $d=1,\dots,D$. The estimation of N clusters $n=1,\dots,N$ results in an assignment of microphone m_d to one of the sources which can be written as m_{d_n}. The number of microphones in a cluster n is specified by D_n and all microphones in cluster n are collected in the set Ω_n. Further, the microphones $m_{d_{\tilde{n}}}$ are those microphones which are not elements of cluster n, so that $D=D_n+D_{\tilde{n}}$ and $\Omega_n\cap\Omega_{\tilde{n}}=\emptyset$, with $\tilde{n}=1,\dots,N$. Thus, an estimation of cluster n produces a complementary cluster $\tilde{n}$.

D_n, the number of microphones in a cluster, is not known a-priori and may vary between the clusters. It cannot be used as information in the training step to perform a 'multi channel' training for the classifier. Thus, the training is performed on clean and anechoic single channel data. For this reason, we want to compute a single feature vector $\bar{\mathbf{v}}_n$ for the classification of audio sources related to microphone clusters $n=1,\dots,N$. One approach which can be applied to reduce the $A\times D$ data matrix $\mathbf{V}=[\mathbf{v}_1\mathbf{v}_2\dots\mathbf{v}_D]$ to an $A\times 1$ vector for each of the clusters n is the linear combination

$$\bar{\mathbf{v}}_n=\mathbf{V}\theta_n, \tag{5}$$

where $\theta_n=[\Theta_{n,1}\Theta_{n,2}\dots\Theta_{n,D}]^t$ is a $D\times 1$ weighting vector which determines the contribution of each microphone to the feature vector $\bar{\mathbf{v}}_n$. It is considered for the case when only the extracted feature vectors are available, e.g., at a fusion center in an ad-hoc array (Subsections 3.1-3.3). In Subsection 3.4, a different approach computes $\bar{\mathbf{v}}_n$ based on a MFCC modulation spectrum in which data from several microphones is combined, and the consecutive feature computation using (3) and (4). To do so, $\tilde{X}_{\mathrm{mfcc}}(\nu,\eta)$ from all microphones needs to be available at the fusion center.

3.1 Single Microphone

We reduce the multichannel system (Fig. 1) to a single channel system by picking the feature vector of just one microphone m_{d_n} for cluster n. Thus, $\theta_{n,\hat{d}}$ at element $\hat{d}=1,\dots,D$ is

$$\theta_{n,\hat{d}}=\begin{cases}1; & \text{for one } m_{\hat{d}}\in\Omega_n;\\ 0; & \text{otherwise;}\end{cases} \tag{6}$$

for $n=1,2,\dots,N$. To choose one of the microphones of Ω_n, we could use information delivered by the (fuzzy) clustering algorithm [9]. In this investigation, however, we consider the microphone positions as known. Thus, we choose that microphone which is closest to the respective audio sources for a cluster with an active sound source and the microphone with the largest distance to all sources for the reverberation cluster.

3.2 Simple Cluster Average

The number of microphones probably varies between the clusters. A cluster averaged feature vector might be useful to exploit information from all microphones within a cluster. Thus, independent of the number of microphones in a cluster, a representative feature vector can be created. E.g., the simple averaged feature vector which contains only information from microphones in cluster n can be computed for $n=1,2,\dots,N$ using

$$\theta_{n,\hat{d}}=\begin{cases}\frac{1}{D_n}; & \text{if } m_{\hat{d}}\in\Omega_n;\\ 0; & \text{otherwise.}\end{cases} \tag{7}$$

3.3 Weighted Cluster Averages

Cluster n obviously contains information about the signal which is dominant in this cluster, whereas cluster $\tilde{n}$ holds information about the concurring signals and reverberation. Therefore we use information from both clusters n and $\tilde{n}$ for the classification of n. One strategy to do so is to use a weighted sum of the averaged feature vectors from the microphones in Ω_n and $\Omega_{\tilde{n}}$. When the vectors θ_n and $\theta_{\tilde{n}}$ are computed for $n=1,2,\dots,N$ with

$$\theta_{n,\hat{d}}=\begin{cases}\frac{w}{D_n}; & \text{if } m_{\hat{d}}\in\Omega_n;\\ 0; & \text{otherwise;}\end{cases} \tag{8}$$

$$\theta_{\tilde{n},\hat{d}}=\begin{cases}\frac{\tilde{w}}{D_{\tilde{n}}}; & \text{if } m_{\hat{d}}\in\Omega_{\tilde{n}};\\ 0; & \text{otherwise;}\end{cases} \tag{9}$$

we may compute the feature vector for cluster n as

$$\bar{\mathbf{v}}_n=\mathbf{V}\theta_n+\mathbf{V}\theta_{\tilde{n}}. \tag{10}$$

The estimation of w and $\tilde{w}$ is obtained in a training step by a least-squares (LS) optimization which is introduced in Subsection 3.5.

Table 1: Room sizes and information about reverb. time and critical distance r_H of the simulated rooms 1-3 and the laboratory.

	Size [m^3]	T60 [ms]	r_H [m]
Room 1 (Simulation)	$4.7 \times 3.4 \times 2.4$	340	0.6
Room 2 (Simulation)	$6.7 \times 4.9 \times 3.5$	490	0.9
Room 3 (Simulation)	$9.3 \times 6.9 \times 4.9$	630	1.3
Laboratory (Recording)	$7.5 \times 6.3 \times 3.3$	300	1.3

3.4 Feature Dependent Weighting of Modulation Spectra for Feature Computation

The previously mentioned approach uses one weight for the averaged feature vector of cluster n and one for the averaged feature vector of cluster $\tilde{n}$. Therefore, all elements of the feature vector are weighted equally. The contribution of each feature to a good classification result however may be varying. It is straightforward to define a feature dependent weighting. Thus, we estimate new optimal weights

$$\phi^{\langle\eta\rangle}_{n,\hat{d}} = \begin{cases} \frac{w^{\langle\eta\rangle}}{D_n}; & \text{if } m_{\hat{d}} \in \Omega_n; \\ 0; & \text{otherwise}; \end{cases} \tag{11}$$

$$\phi^{\langle\eta\rangle}_{\tilde{n},\hat{d}} = \begin{cases} \frac{\tilde{w}^{\langle\eta\rangle}}{D_{\tilde{n}}}; & \text{if } m_{\hat{d}} \in \Omega_{\tilde{n}}; \\ 0; & \text{otherwise}; \end{cases}. \tag{12}$$

The superscript $\langle\eta\rangle$ indicates the dependency on the cepstral bin η. Then, the optimized weights are used for the computation of optimized averaged modulation spectra for cluster n and $\tilde{n}$: for each cepstral bin η, we average the modulation spectra $\tilde{X}_{\text{mfcc}}$ of microphones in cluster n and weight the result with the respective weight $\phi^{\langle\eta\rangle}_{n,d}$ in all modulation frequencies ν. Similarly, we average the modulation spectra $\tilde{X}_{\text{mfcc}}$ of microphones in cluster $\tilde{n}$ and weight the result with the respective weight $\phi^{\langle\eta\rangle}_{\tilde{n},d}$. Finally, similar to (10), we add both weighted modulation spectra and derive $\bar{\mathbf{v}}_n$ using (3) and (4). The estimation of $w^{\langle\eta\rangle}$ and $\tilde{w}^{\langle\eta\rangle}$ is an extension of the previously mentioned LS solution and is introduced in Subsection 3.5.

3.5 LS estimation of weighting factors

To estimate the weights w and $\tilde{w}$ in (8) and (9), and the feature dependent weights $w^{\langle\eta\rangle}$ and $\tilde{w}^{\langle\eta\rangle}$ in (11) and (12) we simulate audio data for ten different scenarios with $D = 15$ randomly placed microphones, two randomly placed sources and 100 different combinations of source files for each combination of differing source types (speech, music and noise) for Room 1-3 (Tab. 1). This results in $K = 9000$ different audio scenarios for the 15 microphones. Details about the simulation setup are given in Section 4. Using the knowledge about the randomly placed microphones we assign them to each source by selecting the microphones within its critical distance. We extract the feature vectors of the microphones m_{d_n} and $m_{d_{\tilde{n}}}$ for $n = \tilde{n} = 1,2$ and compute the cluster-averaged feature vectors $\bar{\mathbf{v}}_n$ and $\bar{\mathbf{v}}_{\tilde{n}}$ using (5) and (7) which we stack into a matrix $\bar{\mathbf{V}}_n = [\bar{\mathbf{v}}_n \bar{\mathbf{v}}_{\tilde{n}}]$. Then, we create a matrix $\mathbf{\Phi} = [\bar{\mathbf{V}}_1^t \bar{\mathbf{V}}_2^t]^t$ with the information about both clusters $n = 1,2$. To estimate $\mathbf{w} = [w, \tilde{w}]^t$ we can utilize the information about the type of source which is dominant for cluster n and formulate an error criterion based on the squared Euclidean distance

$$e(\mathbf{w}) = ||\bar{\mathbf{y}} - \mathbf{\Phi}\mathbf{w}||_2^2 \tag{13}$$

which we want to minimize by computing the LS-solution as

$$\mathbf{w} = (\mathbf{\Phi}^t\mathbf{\Phi})^{-1}\mathbf{\Phi}^t\bar{\mathbf{y}}, \tag{14}$$

where $\bar{\mathbf{y}} = [\mathbf{y}_1^t \mathbf{y}_2^t]^t$ is an averaged feature vector extracted from training data of the type of the respective sources (speech/music, speech/noise or music/noise). To obtain an averaged $\mathbf{w}$ over all scenarios in all rooms, we create a matrix which contains $\mathbf{\Phi}$ of all simulations $\mathbf{\Phi}_{\text{all}} = [\mathbf{\Phi}_1^t \mathbf{\Phi}_2^t \dots \mathbf{\Phi}_K^t]^t$ and in a similar fashion $\bar{\mathbf{y}}_{\text{all}} = [\bar{\mathbf{y}}_1^t \bar{\mathbf{y}}_2^t \cdots \bar{\mathbf{y}}_K^t]^t$. Then, (14) can be applied accordingly. The resulting weightings are $w = 2.79$ and $\tilde{w} = -1.51$. To compute $\mathbf{w}^{\langle\eta\rangle} = [w^{\langle\eta\rangle} \tilde{w}^{\langle\eta\rangle}]^t$ we solve the respective LS-equation similar to (14) for the components $1, \dots, \eta_{\max}$ of the feature vector, using the complete dataset of all K simulated audio scenarios. Fig. 3 presents the result for $w^{\langle\eta\rangle}$ and $\tilde{w}^{\langle\eta\rangle}$, where $\eta_{\max} = 13$. For both, the feature independent weights w and $\tilde{w}$ and the cepstral bin dependent weights $w^{\langle\eta\rangle}$ and $\tilde{w}^{\langle\eta\rangle}$ (Fig. 3), we observe that there is a compensational behavior between n and $\tilde{n}$. n is amplified and then normalized by the contribution of $\tilde{n}$. This leads to a contrast improvement in the feature vector. The coefficients of $w^{\langle\eta\rangle}$ have a high positive weighting with a decreasing trend. The coarse spectral structure experiences the strongest compensational effects when the cepstral bin dependent weighting is applied.

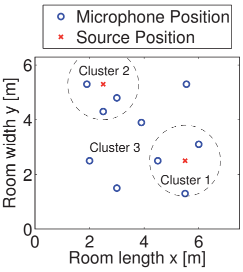

Figure 2: Setup for the recordings in the laboratory with $N = 3$ clusters and $D = 15$ microphones.

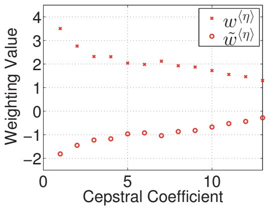

Figure 3: Weightings for the computation of weighted averaged MFCC modulation spectra of cluster n and $\tilde{n}$.

4 Experimental Setup

For the evaluation we use simulations and recorded audio data. We simulate 15 microphones and two active sound sources in three different rooms (see Tab. 1). For each room, we create ten different scenarios of source-microphone setups. In each setup, $2 \leq D_n \leq 4$ microphones for cluster $n = 1,2$ are randomly located within the critical distance of the respective source. Additional $15 - D_1 - D_2$ microphones for cluster $n = 3$ are placed randomly all over the room. The position of each of the sources is randomized in one or the other half of each room. We create RIRs using the method in [15]. To generate microphone signals which contain contributions from both sources, speech (clean and anechoic, male and female, English [16]) and music data (different genres, private database) are convolved with the respective RIRs and summed up. The audio recordings are conducted in an auditory laboratory (Tab. 1). We use two loudspeakers as speech and music sources. The sound is recorded at ten different locations in the room (Fig. 2) with a standard voice recording application of a smartphone. The evaluation then is performed offline. For the classification experiment we use 100 files for each: speech, music and noise (different types of indoor noise, private database). The noise class is used as a training class to carry out a speech/music/noise discrimination, but is not used as input data in the audio simulation or the recording. Therefore, this third class is used only to test for mis-classification of speech or music signals as noise. We use 75% of this data to train our LDA. 25% of the speech and music is used in the simulation of the microphone signals or as speaker signals in the recording setup, respectively. We average our classification results over 10 cross-validation iterations in the simulations and 5 cross-validation iterations in the recordings in which the allocation of training and test data is randomized. For all analyzed scenarios, we know the position of the microphones and therefore we can assign all microphones to clusters $n = 1,2,3$ as follows: cluster 1 combines all microphones which are located within the critical distance of the speech source, cluster 2 contains all microphones which are located within the critical distance of the music source and cluster 3 combines all microphones which are located outside of the critical distance of both sources.

Based on the microphone data we extract the audio features from signals of 4 seconds duration, sampled with 16 kHz. The spectral and cepstral analysis is carried out with a frame length

Table 2: Averaged classification results in % for LDA classification of cluster based feature vectors averaged over all simulations. Cluster 1 should be classified as speech, Cluster 2 as music.

		Cluster 1	Cluster 2	Cluster 3
Strategy 1: Feature vector of single microphone	Speech	29.1	1.4	2.2
	Music	70.9	98.6	97.8
	Noise	0.0	0.0	0.0
Strategy 2: Averaged feature vector of cluster *n*	Speech	13.4	1.2	2.2
	Music	86.6	98.8	97.8
	Noise	0.0	0.0	0.0
Strategy 3: Weighted averaged feature vectors of cluster *n* and $\tilde{n}$	Speech	99.5	12.6	38.6
	Music	0.5	80.8	61.4
	Noise	0.0	6.6	0.0
Strategy 4: Feature Dependent Weighting of Modulation Spectra	Speech	84.3	0.8	30.4
	Music	15.7	96.3	69.6
	Noise	0.0	2.9	0.0

of 512 samples and a frame shift of 256 samples. The modulation analysis of 13 MFCC coefficients is computed using the frame length and shift $L = 16$ and $Q = 8$. We compute $\bar{X}_{\text{mfcc}}(\eta)$ and the CMRs $r_{1|1}(\eta)$ and $r_{2|8}(\eta)$ to create a feature vector $\mathbf{v} = [\bar{\mathbf{X}}^t_{\text{mfcc}}\mathbf{r}^t_{1|1}\mathbf{r}^t_{2|8}]^t$. Note, that the dependency on η of all components of $\mathbf{v}$ is expressed in vector notation. Our final feature vector represents 4 seconds of data with 39 coefficients.

5 Results

Tab. 2 presents the classification accuracy of the simulated signals for the different feature combination strategies averaged over all rooms and all cross-validation iterations. Feature vectors created with strategies 1 and 2 use information purely from microphones within the critical distance of an active source for $n = 1,2$. Based on these, speech is misclassified as music very often. This was already observed in [9] and is a consequence of the modified spectral structure of speech due to reverberation and the presence of music signals during low energy phones and speech pauses. Cluster 3 is classified as music as well. However, when information from the additional microphones in the room is introduced with strategy 3 and 4, the classification of speech becomes much more accurate. The classification performance of music classification is reduced when strategy 3 is used, but the overall performance is better compared to strategy 1 and 2. The application of a feature dependent weighting in strategy 4 performs best with a high classification accuracy for both, speech classification in cluster 1 and music classification in cluster 2. Cluster 3 is classified as music in most of the cases where again reverberation effects and music are dominant.

Tab. 3 presents the results of the cluster based classification of the recorded audio data for the different feature vector combination strategies. The results are similar to the results based on the simulations (Tab. 2). Strategies 1 and 2 show a shift towards the classification result music. However the introduction of information from all microphones in the room results in a high classification accuracy, especially for the feature dependent weighting.

6 Conclusion

We considered scenarios of two active sound sources and multiple ad-hoc distributed microphones in a room. When clusters of microphones are formed that are dominated by one of the sound sources, an accurate cluster based signal classification can be performed. Information in form of extracted feature vectors from the microphone signals in the cluster can be combined with information from the other microphones in the room which pick up mainly concurring signals and reverberation. The improvement which is achieved with a feature dependent weighting implies a varying importance of the different elements of the feature vector for a correct classification result. This is not taken into account when all elements of the feature vector are weighted equally. However, a simple feature independent weighting already delivers good classification results and may be seen as more class independent. All captured audio data for this evaluation was recorded using a commercially available smartphone. This demonstrates the applicability of the algorithm to a scenario with ad-hoc distributed mobile devices.

Table 3: Averaged classification results in % for LDA classification of cluster based feature vectors for the recorded audio data. Cluster 1 should be classified as speech, Cluster 2 as music.

		Cluster 1	Cluster 2	Cluster 3
Strategy 1: Feature vector of single microphone	Speech	21.6	0.0	0.0
	Music	78.4	100.0	100.0
	Noise	0.0	0.0	0.0
Strategy 2: Averaged feature vector of cluster *n*	Speech	2.4	0.0	0.0
	Music	97.6	100.0	100.0
	Noise	0.0	0.0	0.0
Strategy 3: Weighted averaged feature vectors of cluster *n* and $\tilde{n}$	Speech	99.2	22.4	83.2
	Music	0.8	76.0	16.8
	Noise	0.0	1.6	0.0
Strategy 4: Feature Dependent Weighting of Modulation Spectra	Speech	80.8	1.6	28.8
	Music	19.2	98.4	71.2
	Noise	0.0	0.0	0.0

References

[1] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, and U. Rass, "Signal processing in high-end hearing aids: State of the art, challenges, and future trends," *EURASIP Journal on Applied Signal Processing*, vol. 18, pp. 2915 – 2929, 2005.

[2] H. Feng, C. Jiang, and X. Yang, "An audio classification and speech recognition system for video content analysis," in *International Conference on Multimedia Technology (ICMT)*, 2011.

[3] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *IEEE Symposium on Communications and Vehicular Technology (SCVT)*, 2011.

[4] Y. Hioka and B. Kleijn, "Distributed blind source separation with an application to audio signals," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2011.

[5] A. Bertrand, J. Callebaut, and M. Moonen, "Adaptive distributed noise reduction for speech enhancement in wireless acoustic sensor networks," in *International Workshop on Acoustic Echo and Noise Control (IWAENC)*, 2010.

[6] S. Markovich-Golan, S. Gannot, and I. Cohen, "Performance of the SDW-MWF with randomly located microphones in a reverberant enclosure," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 21, pp. 1513–1523, July 2013.

[7] A. R. Abu-El-Quran, R. A. Goubran, and A. D. C. Chan, "Security monitoring using microphone arrays and audio classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 55, no. 4, pp. 1025–1032, 2006.

[8] U. Srinivas, N. Nasrabadi, and V. Monga, "Graph-based multisensor fusion for acoustic signal classification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[9] S. Gergen, A. Nagathil, and R. Martin, "Audio signal classification in reverberant environments based on fuzzy-clustered ad-hoc microphone arrays," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[10] H. Kuttruff, *Room Acoustics*. London: Applied Science Publishers Ltd, 1979.

[11] M. McKinney and J. Breebaart, "Features for audio and music classification," in *International Conferece on Music Information Retrieval (ISMIR)*, 2003.

[12] R. Martin and A. Nagathil, "Cepstral modulation ratio regression (CMRARE) parameters for audio signal analysis and classification," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2009.

[13] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. O.-G. D., Petrovska-Delacrétaz, and D. A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 430–451, Jan. 2004.

[14] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: Wiley, 2. ed., 2001.

[15] S. Gergen, C. Borß, N. Madhu, and R. Martin, "An optimized parametric model for the simulation of reverberant microphone signals," in *Proc. of the International Conference on Signal Processing, Communications and Computing (ICSPCC 2012)*, 2012.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," 1993. Linguistic Data Consortium, Philadelphia.