Signal Processing 107 (2015) 21–32

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Classification of reverberant audio signals using clustered ad hoc distributed microphones

Sebastian Gergen *[,1], Anil Nagathil [2], Rainer Martin

Institute of Communication Acoustics, Ruhr-Universität Bochum, Universitätsstraße 150, 44801 Bochum, Germany

ARTICLE INFO

Article history:
Received 1 February 2014
Received in revised form
17 April 2014
Accepted 27 April 2014
Available online 9 May 2014

Keywords:
Audio signal classification
Ad hoc microphone array
Microphone clustering

ABSTRACT

In a real world scenario, the automatic classification of audio signals constitutes a difficult problem. Often, reverberation and interfering sounds reduce the quality of a target source signal. This results in a mismatch between test and training data when a classifier is trained on clean and anechoic data. To classify disturbed signals more accurately we make use of the spatial distribution of microphones from ad hoc microphone arrays. In the proposed algorithm clusters of microphones that either are dominated by one of the sources in an acoustic scenario or contain mainly signal mixtures and reverberation are estimated in the audio feature domain. Information is shared within and in between these clusters to create one feature vector for each cluster to classify the source dominating this cluster. We evaluate the algorithm using simultaneously active sound sources and different ad hoc microphone arrays in simulated reverberant scenarios and multichannel recordings of an ad hoc microphone setup in a real environment. The cluster based classification accuracy is higher than the accuracy based on single microphone signals and allows for a robust classification of simultaneously active sources in reverberant environments.

1. Introduction

Automatic classification of audio signals into predefined categories, e.g., speech, music or noise, or sub-categories of these, e.g., gender of a speaker, music genre or noise type, is a useful step in many applications in which signal processing is used to improve the quality of a target signal and to reduce the contribution of other signals that are interfering and hence are degrading the target signal. Hearing aids may use audio signal classification to select and adjust appropriate signal processing algorithms, for example if the desired signal is a speech signal and the background is either a second speaker, babble noise or music [1,2]. Further fields of application are, for instance, hands-free telephony, teleconferencing, and automatic speech recognition [3], e.g., in Ambient Assisted Living (AAL) scenarios [4].

However, most of these previously mentioned applications face the challenge of an unknown acoustic environment that influences the received signals by introducing reverberation and background noise. In terms of the signal classification these effects will introduce a mismatch if the training is performed on anechoic and clean data. Since the amount of reverberation and noise among different environments may differ considerably, training a classification system on a specific environment does not solve the problem. An independent approach to obtain a robust classification result is required.

Recently, ad hoc microphone arrays, which are often called the Wireless Acoustic Sensor Networks (WASN),

* Corresponding author. Tel.: +49 234 32 26662; fax: +49 234 32 14165.
E-mail addresses: sebastian.gergen@rub.de (S. Gergen), anil.nagathil@rub.de (A. Nagathil), rainer.martin@rub.de (R. Martin).

[1] The work of Sebastian Gergen is supported by the Ministry of Economic Affairs and Energy of the State of North Rhine-Westphalia (Grant IV.5-43-02/2-005-WFBO-009), Germany.
[2] The work of Anil Nagathil is funded by the German Research Foundation (DFG), Collaborative Research Center 823, Subproject B3.

http://dx.doi.org/10.1016/j.sigpro.2014.04.034
0165-1684/

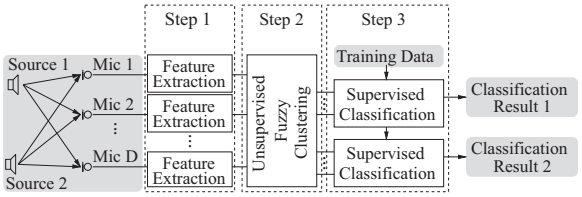

Fig. 1. Architecture of the proposed algorithm for the example of two source signals.

increasingly attract the attention of researchers in the acoustic signal processing community [5]. These arrays consist of devices that can be linked wirelessly and provide signal capturing and signal processing capabilities. Compared to conventional microphone array (MA) technology, there are some major differences which can be summarized as follows [5]:

- WASNs may provide microphones distributed all over the environment. Hence, at least one device might always be rather close to a potential target source. The geometry of the WASN, however, is generally not known.
- WASNs may consist of different devices, each of which having an own clock. Therefore, there is an unavoidable clock skew and so far no exact synchronization, e.g., for signal capturing and sampling is available.
- WASNs allow for distributed signal processing, as each device is considered to have autonomous signal processing capabilities. However, the signal processing and the wireless data transmission (the type and amount of data, data sinks and sources) between the devices have to be controlled, e.g., by a fusion center.

Much research effort is put into the field of (not exclusively wireless) ad hoc distributed microphones in the context of blind source separation (BSS) [6–9]. Further, noise reduction algorithms using WASNs have been introduced as well. E.g., in [10] adaptive noise reduction is performed based on a Multichannel Wiener Filter, and in [11] a Generalized Side-lobe Canceller (GSC) is introduced, which uses several devices in a room, and several microphones per device. The statistical performance of a Speech-Distortion-Weighted Multichannel Wiener Filter using randomly distributed microphones is evaluated in [12].

Some algorithms already use multiple microphones as a part of classification algorithms, e.g., for a surveillance system [13], in military sound classification applications [14], or to classify the type of the surrounding sound field characteristics [15] rather than the signal type. However, all these consider MAs in the conventional sense.

In this contribution we analyze the influence of reverberation and competing acoustical sources on the classification of audio signals which are captured by ad hoc distributed microphones. Distributed processing in those devices is foreseen for audio feature extraction without the requirement of exact synchronization of the audio signals. Based on the feature data, clusters of microphones are estimated where a cluster contains all microphones that are positioned closely to an active sound source. Further, one additional background cluster aggregates microphones that mainly receive diffuse signal mixtures and reverberation. Finally, the classification of the sources is carried out with respect to the estimated clusters of microphones. To obtain robust classification results, information in terms of the extracted audio features is exchanged between the clusters: a strategy to combine data of the cluster that is to be classified and the complementary clusters that provide information about interfering signals and reverberation is applied.

The remainder of this paper is structured as follows: in Section 2 the basic idea of the algorithm and its components are introduced. In Section 3 we explain the evaluation scenarios and in Section 4 we present experimental results. Finally, we draw conclusions in Section 5.

2. Algorithm concept and components

For our investigations we consider a random spatial distribution of sound sources and 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 higher than the reverberated sound energy [16]. Other devices are located somewhere in the room, possibly outside of the critical distances of all sources. The first step of our algorithm (Fig. 1) is the audio feature extraction. It is meant to be performed in each device of the WASN directly. A feature vector represents the captured data in a compact way and therefore simplifies the data exchange. For our algorithm, no full bandwidth audio signals have to be transmitted between the devices. Details on the feature extraction are given in Section 2.1.

For audio signal classification based on single channel training data in conjunction with test data from several audio capturing devices in a room, a data combination strategy is required. This strategy has to create compatibility between test and training data, and at the same time, exploit the diversity of the information obtained by the spatially distributed receivers. There are different domains in which such a strategy can be applied: the signal domain, the feature domain and the decision domain [17]. Based on the ad hoc microphone array setup with its properties, e.g., the lack of knowledge about the microphone position, the limited data transmission and difficult synchronization, strategies in the signal domain, for example beamforming, are less appropriate. Therefore, we introduce strategies, both, for the feature domain (Fig. 2) and the decision domain (Fig. 3). Fig. 2 shows a strategy in which at first extracted feature vectors are combined and then used as common test sample in the classification system, which we term as *combine-then-analyze* (CTA) strategy. The algorithm in Fig. 3 first classifies feature vectors from different devices independently, followed by a combination to finally result in one shared classification decision. Therefore we term this the *analyze-then-combine* (ATC) strategy.

A crucial part of our combination strategy is the estimation of a neighborhood of devices in the array of

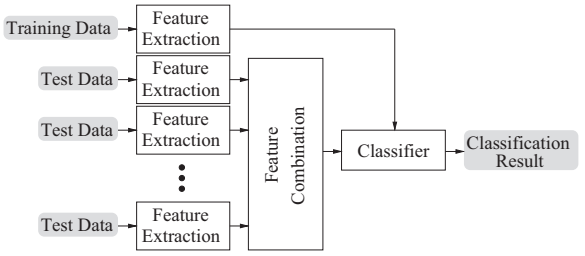

Fig. 2. Structure of a CTA classification system that combines several feature vectors to obtain a single classification result.

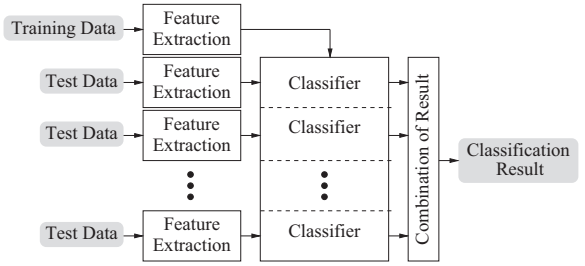

Fig. 3. Structure of an ATC classification system that combines single channel classification results, e.g., via majority decision.

multiple audio capturing devices in a room and an estimation about the homogeneity of the information that is captured and can be shared. Therefore, a measure for the similarity or relationship between microphone data is required that is, however, not based on information about the devices position. Data from devices that are positioned closely together is probably more similar than data from devices placed at the other end of a room, and therefore probably close to a different audio source. Moreover, data from devices that receive balanced mixtures of source signals should induce an intermediate value of similarity to the neighborhoods of the sources. To solve this neighborhood estimation, one central device in the network, the fusion center, performs a fuzzy-clustering procedure and finds clusters of devices in the feature space (Step 2 in Fig. 1). In the following we use the term cluster for microphones that are in one neighborhood and are dominated by one type of audio signal, e.g., a specific source signal or reverberation. The unsupervised clustering procedure that is used is explained in Section 2.2.

After this clustering step we are able to apply a combination strategy and perform the signal classification. The supervised classification step in Fig. 1 allows for both of the mentioned approaches (ATC and CTA). In our implementation we use the CTA strategy as we want to exploit the clustering for the improvement of the feature vectors. Section 2.3 presents details on the combination of the feature vectors amongst clusters of devices.

2.1. Signal feature extraction

For clustering and classification, captured audio data is transformed into a parametric representation, which allows to distinguish between different classes while reducing the amount of necessary data. Several different types of audio features were introduced and evaluated in various classification experiments, e.g., for clean speech, speech in noise, noise and music classification [2,18–20] or music genre classification [21,22]. For our investigation we consider a cepstro-temporal representation of the signal [23] which we call the Modulation Mel-Frequency Cepstral Coefficients (Mod-MFCCs) features.

To compute the feature set, first a captured audio signal $x(t)$ is sampled with the sampling rate f_s. The time-discrete representation $x(l)$, where l is the discrete time index, is segmented into B possibly overlapping frames of length Y using a window function $\mathcal{W}(y)$, e.g., the Hann window, with $y=1,2,\ldots,Y$ and a frame shift P. The discrete Fourier transform (DFT) of the weighted frames is calculated as

$$X(k,b)=\sum_{y=0}^{Y-1} x(bP+y)\mathcal{W}(y)e^{-j2\pi yk/Y}, \tag{1}$$

where $b=0,1,\ldots,B-1$ and $k=0,1,\ldots,Y-1$ denote the frame index and the frequency bin, respectively. Now, the squared magnitude spectrum $|X(k,b)|^2$ is mapped onto the mel scale using overlapping triangular windows $\mathcal{V}_{k'}(k)$, which define band-pass filters [24]. This results in a mel-spectrum:

$$X_{\mathrm{mel}}(k',b)=\sum_{k=0}^{Y/2}|X(k,b)|^2\mathcal{V}_{k'}(k), \tag{2}$$

where $k'=0,1,\ldots,K'-1$ is the index of the mel scale frequency bin. Then, the MFCCs are calculated by computing the discrete cosine transform (DCT) of the logarithm of the mapped power spectrum

$$X_{\mathrm{mfcc}}(\eta,b)=\sum_{k'=0}^{K'-1}\log(|X_{\mathrm{mel}}(k',b)|)\ \cos\left(\frac{\pi}{K'}(k'+0.5)\eta\right), \tag{3}$$

with the cepstral coefficient index η. It has already been shown that the time evolution of short-time features can be exploited to improve classification results [21,23]. To consider this we compute the short-time MFCC modulation spectrum $\hat{X}_{\mathrm{mfcc}}(\nu,\eta,c)$ using a sliding window DFT:

$$\hat{X}_{\mathrm{mfcc}}(\nu,\eta,c)=\sum_{\ell=0}^{L-1}X_{\mathrm{mfcc}}(\eta,cQ+\ell)e^{-j2\pi\ell\nu/L}, \tag{4}$$

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 $c=0,1,\ldots,C_T-1$ [23]. A time averaging of these modulation spectra increases the robustness of the feature set against synchronization errors in the WASN when we assume that, in contrast to short-time audio features, the averaged temporal behavior of a signal is rather stationary. Therefore, the absolute values of the modulation spectra are averaged over all C_T frames (5) and cepstral modulation ratios (CMR) $r_{\nu_1|\nu_2}(\eta)$ are computed to approximate the modulation spectrum (6):

$$\tilde{X}_{\mathrm{mfcc}}(\nu,\eta)=\frac{1}{C_T}\sum_{c=0}^{C_T-1}\left|\hat{X}_{\mathrm{mfcc}}(\nu,\eta,c)\right| \tag{5}$$

$$r_{\nu_1|\nu_2}(\eta)=\frac{\sum_{\nu=\nu_1}^{\nu_2}\tilde{X}_{\mathrm{mfcc}}(\nu,\eta)}{(\nu_2-\nu_1+1)\tilde{X}_{\mathrm{mfcc}}(0,\eta)} \tag{6}$$

Note that in (6) the average of several modulation frequency bands within $\nu_1 \le \nu \le \nu_2$ is normalized on the zeroth modulation frequency band. If $\nu_1 = \nu_2 \neq 0$, this reduces to a single modulation frequency band $\nu_1 \ge 1$ that is normalized by the zeroth band. Finally, for our feature vector we stack two CMRs as well as $\overline{X}_{\mathrm{mfcc}}(\eta)$ into a vector, where

$$\overline{X}_{\mathrm{mfcc}}(\eta) = \frac{1}{L}\sum_{\nu'=0}^{L-1} \tilde{X}_{\mathrm{mfcc}}(\nu', \eta) \tag{7}$$

is the modulation spectrum averaged over all modulation frequencies ν for each MFCC bin η.

In nearly every realistic situation room dependent reverberation and additive noise contaminate audio signals. The source–receiver transmission path determines the reverberation and can be represented by the room impulse response (RIR). Cepstral mean normalization (CMN) is one approach to reduce the effect of reverberation for audio signal processing. The convolutional distortion in the time domain corresponds to an additive term in the cepstral domain. By averaging over a certain amount of time, in which the RIR can be assumed to be constant, and subtracting this average from the cepstrum X_{mfcc}, the influence of reverberation can be reduced [25,26].

2.2. Unsupervised signal clustering

Cluster analysis is an unsupervised learning procedure that is used to subdivide a data set into clusters. The elements of the data set should have properties that allow for a reasonable allocation into subgroups, with small intra-cluster differences and large inter-cluster differences. Cluster analysis finds application in many fields in which analytical or observational data is available, e.g., medicine, geology, business, engineering and signal processing [27].

Given is a set of D observations $\Omega = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D\}$ in the A-dimensional Euclidean Space $\mathbb{R}^A$. $\mathbf{v}_d$ is the d-th observation (feature) vector and v_{dj} the j-th feature of $\mathbf{v}_d$, with $j \in \{1, 2, \ldots, A\}$.

A hard partition of Ω into N clusters $\Omega_1, \Omega_2, \ldots, \Omega_N$, with $2 \le N \le D$, satisfies the following three conditions: each of the clusters is a non-empty set and therefore contains some cluster members. An intersecting set of two clusters is the empty set and the union of all N clusters yields the observation set Ω. These conditions can be presented in matrix notation [27]. Let $\nabla = [\mu_{nd}]$ be a matrix of dimensions $N \times D$ that contains the elements μ_{nd}. This matrix represents a hard partition of Ω when

$$\mu_{nd} = \begin{cases} 1 & \text{if } \mathbf{v}_d \in \Omega_n; \\ 0 & \text{otherwise;} \end{cases} \tag{8a}$$

$$\sum_{d=1}^{D} \mu_{nd} > 0 \quad \text{for all } n; \tag{8b}$$

$$\sum_{n=1}^{N} \mu_{nd} = 1 \quad \text{for all } d. \tag{8c}$$

Each element μ_{nd} in ∇ represents a membership value {0,1}, depending on whether the observation d is a part of cluster n (8a) or not.

Corresponding to the definition of Fuzzy Sets [28], one may generalize ∇ to a fuzzy partition of Ω by replacing the selection of $\mu_{nd} \in \{0, 1\}$ by $\mu_{nd} \in [0, 1]$, which now can be interpreted as a continuous membership grade of $\mathbf{v}_d$ in the fuzzy clusters of Ω. Eqs. (8b) and (8c) are still valid and the hard partitions of Ω are special cases of fuzzy ones [27].

Several algorithms have been proposed to estimate an optimal fuzzy partition of Ω. The most studied and most popular method is the Fuzzy c-Means Algorithm (FCM). It evaluates a least-squared error functional, given as [29]

$$J_m(\nabla, \mathbf{u}) = \sum_{d=1}^{D}\sum_{n=1}^{N} (\mu_{nd})^{\alpha} \| \mathbf{v}_d - \mathbf{u}_n \|_{\beta}^2 \tag{9}$$

and estimates ∇ and the cluster centers $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_N$ in an iterative manner, where $\mathbf{u}_n = (u_{n,1}, u_{n,2}, \ldots, u_{n,A})^T$ is the center of cluster n. In addition to the data Ω and the number of clusters N, a weighting exponent α with $1 \le \alpha \le \infty$ and an distance norm $\| \|_{\beta}$ on $\mathbb{R}^A$ are used in (9). The squared distance between an observation and a cluster center is computed as

$$\delta_{nd}^2 = \| \mathbf{v}_d - \mathbf{u}_n \|_{\beta}^2 = (\mathbf{v}_d - \mathbf{u}_n)^T \boldsymbol{\beta} (\mathbf{v}_d - \mathbf{u}_n), \tag{10}$$

where different types of weighting matrices $\boldsymbol{\beta}$ can be applied, resulting in, e.g., the squared Euclidean norm, diagonal norm or Mahalanobis norm [27]. The choice of the weighting matrix determines the shape of the estimated clusters. The parameter α controls the relative weight of each element δ_{nd}^2. $\alpha = 1$ results in a hard partitioning and $\alpha \to \infty$ leads towards $\mu_{nd} = 1/N$, which implies a higher fuzziness.

2.3. Feature vector combination

In the considered scenario, D distributed devices with microphones capture audio data in a room. Based on the extracted audio feature data, an unsupervised clustering is performed to group the microphones into N clusters that are dominated by either one of the $N-1$ sources or by reverberation components. Obviously, the microphones that are located close to a source deliver the most valuable information to classify this source correctly. To apply the CTA combination strategy, the cluster estimation and the fuzzy cluster membership information μ_{nd} provided by the fuzzy-clustering algorithm are used in order to combine the feature vectors from devices in the vicinity of a source. This combined feature vector is then used in the supervised classification step as test data.

The microphone of an signal capturing device m_d, with $d = 1, 2, \ldots, D$, in this context is represented by the A-dimensional feature vector $\mathbf{v}_d$, which is extracted from the captured audio data of m_d. Because of the fuzzy clustering step, we have $\mu_{nd} > 0$ for all microphones m_d and clusters n and in this algorithm a microphone m_d is assigned to that cluster n for which μ_{nd} attains the largest value. All microphones in a cluster $n = 1, \ldots, N$ are collected in the set Ω_n. To formulate the CTA strategy we collect the feature vectors extracted from signals of all microphones into an $A \times D$ matrix $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_D]$. Then, for each estimated cluster this matrix has to be reduced to an

A-dimensional vector. This can be expressed as

$$\overline{\mathbf{v}}_n = \mathbf{V}\boldsymbol{\theta}_n, \tag{11}$$

where $\overline{\mathbf{v}}_n$ is a CTA-feature vector of cluster n and $\boldsymbol{\theta}_n = (\theta_{n,1}, \theta_{n,2}, \ldots, \theta_{n,D})^T$ is a D-dimensional weighting vector that determines the contribution of each microphone m_d to the feature vector $\overline{\mathbf{v}}_n$. If the number of sensors in a cluster n is given as D_n, $\overline{\mathbf{v}}_n$ may be obtained as a simple average of the respective D_n feature vectors, thus

$$\theta_{n,\hat{d}} = \begin{cases} \dfrac{1}{D_n} & \text{if } m_{\hat{d}} \in \Omega_n; \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

In a more refined approach we further define a complementary cluster $\tilde{n}$ with $\tilde{n} = 1, \ldots, N$ which includes those microphones which are not assigned to cluster n, so that $D_n + D_{\tilde{n}} = D$ and $\Omega_n \cup \Omega_{\tilde{n}} = \Omega$. Hence, each cluster n defines a complementary cluster $\tilde{n}$. It was shown in [30] that the sum of weighted averages

$$\overline{\mathbf{v}}_n = \mathbf{V}\boldsymbol{\theta}_n + \mathbf{V}\boldsymbol{\theta}_{\tilde{n}} \tag{13}$$

of the feature vectors of cluster n and of cluster $\tilde{n}$ can lead to an improved classification accuracy if $\theta_{n,\hat{d}}$ and $\theta_{\tilde{n},\hat{d}}$ are computed as shown in (14) and (15), where optimized weights w and $\tilde{w}$ are used:

$$\theta_{n,\hat{d}} = \begin{cases} \dfrac{w}{D_n} & \text{if } m_{\hat{d}} \in \Omega_n; \\ 0 & \text{otherwise;} \end{cases} \tag{14}$$

$$\theta_{\tilde{n},\hat{d}} = \begin{cases} \dfrac{\tilde{w}}{D_{\tilde{n}}} & \text{if } m_{\hat{d}} \in \Omega_{\tilde{n}}; \\ 0 & \text{otherwise.} \end{cases} \tag{15}$$

The classification performance is increased if data from $\tilde{n}$ is used in a compensational way in the construction of $\overline{\mathbf{v}}_n$. Therefore, we here choose $w=2$ and $\tilde{w} = -1$, which allows for an easy and intuitive interpretation: the averaged information from cluster n is modified by the difference of the averaged feature vectors of n and $\tilde{n}$. This leads to an improvement of contrast between the averaged feature vectors of n and $\tilde{n}$. Feature components that are larger (smaller) in $\overline{\mathbf{v}}_n$ than in $\overline{\mathbf{v}}_{\tilde{n}}$ are increased (decreased) by the difference.

3. Experiments

For the evaluation of the algorithm we conduct experiments with simulated and recorded microphone signals. We consider a scenario of two audio sources (speech and music) and several microphones in a reverberant room. For the simulations, rooms of three different sizes and reverberation times, each containing 15 distributed microphones are simulated (Table 1). For all experiments beside the baseline experiments, which are based on clean and anechoic audio signals, we simulate several source–microphone setups. The RIRs are created using the method in [31], which is an extended version of the image source method [32]. To generate the microphone signals, the clean source signals are convolved with the respective RIRs and summed up. In this way each simulated microphone picks up signals from Source 1 and Source 2. The recording of the audio data is performed in an acoustic laboratory with a 16 microphone setup (Fig. 4) for three different reverberation times T_{60} (Table 2). We record the audio signals of the two sources consecutively. This allows us to create a high amount of different speech–music combinations as we are able to randomly select one of the speech and one of the music signals and sum up the respective recorded microphone signals to obtain the mixed microphone signal.

For all investigations, we sample microphone signals of $T=4$ s duration at $f_s = 16$ kHz and extract Mod-MFCC features with CMN. For the spectral and cepstral analysis, the frame length is $Y=512$ and frame shift is $P=256$ samples. For the cepstral modulation analysis the frame

Table 1
Sizes, reverberation times and respective critical distances of the simulated rooms.

Attribute	Room 1 (small)	Room 2 (medium)	Room 3 (large)
Size (m^3)	$4.7 \times 3.4 \times 2.4$	$6.7 \times 4.9 \times 3.5$	$9.3 \times 6.9 \times 4.9$
T_{60} (ms)	340	490	630
r_H (m)	0.6	0.9	1.3

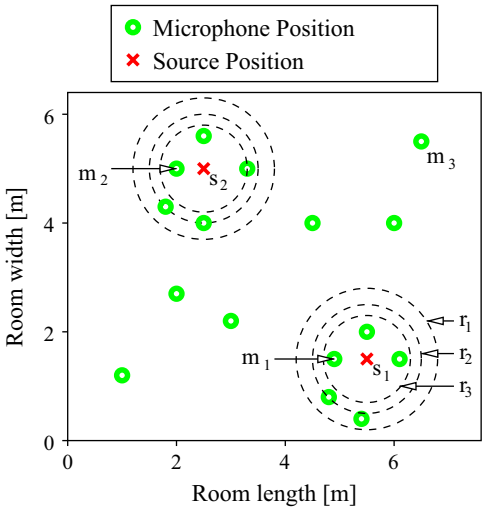

Fig. 4. Setup of microphones and sources in the acoustic laboratory room. The critical distances for the three reverberation times (low reverb.: r_1, medium reverb.: r_2, high reverb.: r_3) are indicated in dashed circles.

Table 2
Size, reverberation times and respective critical distances of the acoustic laboratory.

Attribute	Low reverb.	Medium reverb.	High reverb.
Size (m^3)		$7.6 \times 6.4 \times 3.4$	
T_{60} (ms)	300	500	700
r_H (m)	1.3	1	0.85

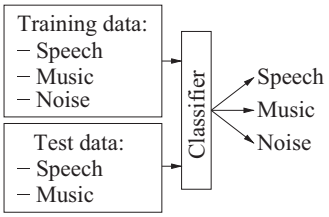

Fig. 5. Overview of data used for the supervised classification.

length and shift is set to $M=16$ and $Q=8$, respectively. The normalized averaged modulation content is computed for the first 13 MFCC features. The modulation ratios $r_{\nu_1|\nu_2}(\eta)$ for $\nu_1=1,\nu_2=1$ as well as for $\nu_1=2,\nu_2=8$ and the averaged modulation spectrum $\overline{X}_{\mathrm{mfcc}}(\eta)$ are rewritten in vector notation $\mathbf{r}_{1|1}$, $\mathbf{r}_{2|8}$ and $\overline{\mathbf{X}}_{\mathrm{mfcc}}$ and finally stacked into one feature vector:

$$\mathbf{v}=(\overline{\mathbf{X}}_{\mathrm{mfcc}}^T,\mathbf{r}_{1|1}^T,\mathbf{r}_{2|8}^T)^T. \tag{16}$$

This results in 39 coefficients to summarize 4 s of data. For the classification we use a linear discriminant analysis (LDA) [33].

To conduct a realistic classification experiment an additional background noise class is added in the training step. However, noise signals are not added to the microphone signals and thus are not part of the test data (Fig. 5). This noise class allows us to test for a misclassification of speech or music as noise in the evaluation. All speech signals are taken from the TIMIT database [34]. Music data covers different songs of several genres and the noise class includes typical indoor noise sounds (e.g., vacuum cleaner, dish washer).

The positioning information is used in the following investigations only for evaluation purposes, but is not given to the algorithm as *a priori* information. The clustering and classification algorithms process the feature vectors only.

3.1. Baseline experiment

As a baseline experiment we carry out a speech/music/noise classification based on clean and anechoic data. In this way, the classification accuracy of the feature set and the LDA classifier can be analyzed independent of the influence of concurring signals and reverberation effects. We consider 100 audio files from the databases for each type of audio data. We randomly select 75% as training data for the LDA classifier and 25% as test samples. The classification accuracy is averaged over 50 cross-validation iterations in which the allocation of test and training data is randomized.

3.2. Single microphone signal classification in reverberant environments

To analyze the influence of interference and reverberation onto the single channel classification performance we simulate RIRs of a two source and five microphone scenario in Room 3 (Fig. 6). The left source (Source 1, close to Mic. 2) represents a speaker and the right source (Source 2, close to Mic. 4) represents a music playing device. Further, we use three of the recorded microphone signals for the analysis of the single channel classification. These

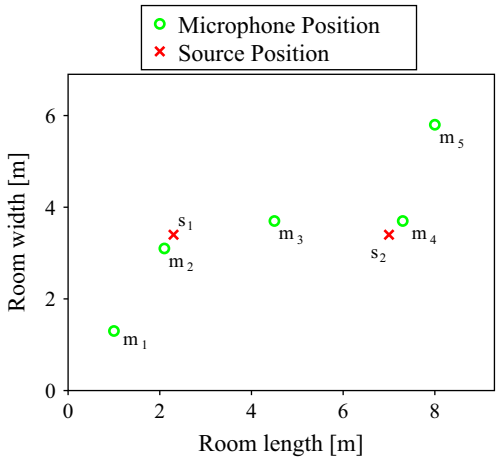

Fig. 6. Simulation setup for the classification of single channel microphone signals.

are the two microphones that are the closest to one or the other of the sources in the room (Mic. 1 for the speech source and Mic. 2 for the music source, Fig. 4) and a third microphone (Mic. 3, Fig. 4) that is positioned farthest away from both sources and therefore receives a balanced mixture of both source signals and a rather high amount of reverberation.

The training of the LDA classifier is performed again on 75% of 100 audio files of each category, speech, music and noise for both experiments. The remaining 25% of the speech and music data are used to generate the reverberant signals. For the simulations, the respective test data is convolved with the RIRs and added to obtain simulated microphone signals. If just one active source is simulated, the RIRs from the other source are set to zero.

For the recorded audio data the respective 25 of the 100 audio signals from speech or music are picked from the recordings if just one source is considered as active. When both sources should be active simultaneously, the microphone signals from the separate recordings of speech and music signals are summed up. The classification accuracy is averaged over 10 cross-validation iterations in which the allocation of test and training data is randomized.

Reverberation introduces a temporal smearing of the signal [35]. In particular, the temporal structure of speech with its distinct pauses is modified by this smearing and the classification performance may be reduced when training is performed on anechoic data. As a counter measure, we use a low amount of additive pink noise (SNR = 15 dB) in the training of the speech data in an additional experiment. This noise is meant to fill the gaps caused by the speech pauses similarly as it occurs for reverberant speech and thus to reduce the mismatch between training and test data in the feature domain, which is based on the effects of reverberation. The corresponding evaluation of the single channel classification performance with pink noise in the speech training data is performed using the same settings as described before.

3.3. Fuzzy clustering of microphones in reverberant environments with two sources

The performance of the fuzzy clustering of the microphones is analyzed using simulated and recorded microphone signals. We simulate microphone–source-setups in three different rooms (Table 1). Source 1 is randomly positioned in the left half of each room and Source 2 is randomly positioned in the right half of each room. For each of the sources, a random number of microphones D_n with $2 \leq D_n \leq 4$ is placed within its critical distance r_H. Then, further microphones are placed randomly in the room such that the total number of microphones in a room is 15. For each room size we simulate ten different scenarios based on this definition. The microphone signals are again simulated by convolving clean signals with the RIRs and adding contributions from both sources for each simulated microphone.

The recording setup is shown in Fig. 4. Two loudspeakers are positioned in the room, pointing towards each other. The microphones are distributed and serve as an ad hoc array. Again, some microphones are within the critical distance of one or the other source and some microphones are located outside of the critical distances of both sources.

As in this investigation on unsupervised clustering no training data is necessary, we use all the 100 audio examples, for both speech (Source 1) and music (Source 2) to analyze the algorithm. In this experiment both sources are considered as active. The clustering is performed on the extracted audio features for all 100 combinations of speech and music. We use a freely available *MATLAB* implementation of the FCM algorithm [36] setting the number of clusters to be estimated to $N=3$, the weighting exponent to $\alpha=2$ and β as the identity matrix which results in the Euclidean metric.

For the evaluation of the estimation of microphone clusters using the feature based fuzzy clustering we have to consider the knowledge of the microphone and source positions. If the median d_{median} of the distances between a source and all microphones in one cluster meets the condition $d_{\text{median}} \leq r_H$ the cluster is assigned to this source. The median is used for robustness reasons. A single microphone that is not located within the critical distance of a source but might be misassigned to the respective cluster will not modify the cluster-to-source mapping. Here, we defined cluster 1 as the cluster related to the speech source and cluster 2 as the cluster related to the music source. If $d_{\text{median}} \leq r_H$ is not true for both sources, then the cluster represents reverberation.

To get a good impression of the acoustic situation in a room, information is needed about which of the estimated clusters is dominated by a source in a room, and which is the cluster of microphones that mainly receives reverberation components and signal mixtures. The clusters that are related to one active source should have a higher fuzzy membership value for the mixed reverberation cluster than for an cluster that is dominated by another source. Therefore, for each cluster we average the fuzzy membership values of all microphones that are assigned to a cluster. Then, we search for those two of all three estimated clusters which accordingly have the medium membership value related to the same (third) cluster. This experiment is performed again on simulated and recorded microphone data which is created as described before in this section.

3.4. Combined experiment: fuzzy clustering, feature combination, and classification

Finally, we analyze the complete performance of the algorithm, including fuzzy clustering and supervised classification. For this, we again use 75% of the 100 data samples of each class to train the speech/music/noise LDA classifier. The remaining 25% of the speech and music data are used in the room simulation and from the audio recordings as described before. Based on the microphone signals, the audio features are extracted and clustered. The feature vectors that are related to one cluster are combined using the CTA strategy as described and used as test sample for the LDA classification. To allow for a better interpretation of the results, additionally we perform the classification based on feature data extracted from single microphone signals. First, as optimal reference, the microphone which is closest to the actual source in the source dominated clusters and, therefore, has a high SNR with respect to the signal of this close source [37] is picked for clusters 1 and 2, and the microphone which is farthest away from both sources in the room is selected for cluster 3. For this, we exploit the knowledge about the setup of microphones and sources. Second, as this information about the setup is not available for the algorithm, we perform a classification based on the signals of the microphones that have the highest fuzzy membership value in each of the three clusters. The classification accuracy is averaged over 10 cross-validation iterations for each room scenario, simulated as well as recorded, in which the allocation of test data and training data is randomized.

4. Results

4.1. Baseline experiment

Table 3 presents the averaged classification performance based on the LDA classifier for speech/music/noise discrimination for clean and anechoic training and test data. The performance exceeds 90% for all three classes. In particular, the classes of speech and music, which are of interest in the following evaluations, are classified correctly in more than 96% of all cases.

Table 3
Averaged result of speech/music/noise classification (in %) based on clean and anechoic data.

	Classified as		
Real class	Speech	Music	Noise
Speech	99.5	0.5	0.0
Music	1.7	96.2	2.1
Noise	0.0	8.2	91.8

Table 4
Classification (in %) based on simulated microphone signals of 5 microphones that are spatially distributed as shown in Fig. 6, Training on speech/music/noise, Source 1 is speech and Source 2 is music.

Active sources	Mic.-No.	1	2	3	4	5
Source 1 (speech)	Speech	63.0	100.0	66.2	53.6	59.2
	Music	37.0	0.0	33.8	46.4	40.8
	Noise	0.0	0.0	0.0	0.0	0.0
Source 2 (music)	Speech	0.4	0.0	0.0	1.4	0.0
	Music	99.4	100.0	100.0	98.6	99.0
	Noise	0.2	0.0	0.0	0.0	1.0
Sources 1+2	Speech	0.4	29.8	0.2	0.6	0.2
	Music	99.6	70.2	99.8	99.4	99.8
	Noise	0.0	0.0	0.0	0.0	0.0

Table 5
Classification (in %) based on simulated microphone signals of 5 microphones that are spatially distributed as shown in Fig. 6, Training on speech(+pink noise)/music/noise, Source 1 is speech and Source 2 is music.

Active sources	Mic.-No.	1	2	3	4	5
Source 1 (speech)	Speech	93.6	100.0	95.8	92.6	92.8
	Music	6.4	0.0	4.2	7.4	7.2
	Noise	0.0	0.0	0.0	0.0	0.0
Source 2 (music)	Speech	0.8	0.2	0.6	5.0	0.8
	Music	99.2	99.8	99.4	94.4	99.2
	Noise	0.0	0.0	0.0	0.6	0.0
Sources 1+2	Speech	8.2	93.0	7.2	5.6	3.6
	Music	91.8	7.0	92.8	94.4	96.4
	Noise	0.0	0.0	0.0	0.0	0.0

Table 6
Classification (in %) based on recorded microphone signals of 3 microphones that are spatial distributed as shown in Fig. 4, Training on speech/music/noise, Source 1 is speech and Source 2 is music.

		$T_{60} = 300$ ms			$T_{60} = 500$ ms			$T_{60} = 700$ ms		
Active sources	Mic.-No.	1	2	3	1	2	3	1	2	3
Source 1 (speech)	Speech	92.0	74.0	78.3	74.3	50.3	53.2	75.1	50.2	54.4
	Music	8.0	26.0	21.7	25.7	49.7	64.8	24.9	49.8	46.6
	Noise	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Source 2 (music)	Speech	2.0	2.0	1.7	1.0	1.0	0.8	0.7	0.1	1.6
	Music	98.0	98.0	98.3	99.0	99.0	99.2	99.3	99.3	98.4
	Noise	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.0	0.0
Sources 1+2	Speech	10.0	0.0	0.0	5.0	0.0	0.0	4.4	0.0	0.0
	Music	90.0	100.0	100.0	95.0	100.0	100.0	95.6	100.7	100.0
	Noise	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

4.2. Single microphone signal classification in reverberant environments

The influence of reverberation on the classification performance is shown in Tables 4–7. The results are very similar for the simulated (Tables 4 and 5) and for the recorded data (Tables 6 and 7). Music in general is classified very reliably (e.g., Tables 4 and 5). However, even when no music source is active the reverberation causes problems for the correct classification of speech signals (Tables 4 and 6). Signals with a low amount of direct speech, e.g., from microphones 4 and 5 in the simulated data or microphones 2 and 3 in the recorded setup, are often classified as music. This can be explained by taking into account the temporal envelopes of speech and music signals. For most music genres, music signals generally have a rather smooth and slowly varying temporal envelope, whereas for speech it changes more

Table 7
Classification (in %) based on recorded microphone signals of 3 microphones that are spatially distributed as shown in Fig. 4, Training on speech(+pink noise)/music/noise, Source 1 is speech and Source 2 is music.

		$T_{60}=300$ ms			$T_{60}=500$ ms			$T_{60}=700$ ms		
Active sources	Mic.-No.	1	2	3	1	2	3	1	2	3
Source 1 (speech)	Speech	100.0	100.0	100.0	100.0	98.3	97.0	100.0	98.9	97.3
	Music	0.0	0.0	0.0	0.0	1.7	3.0	0.0	1.1	1.7
	Noise	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Source 2 (music)	Speech	2.0	4.0	1.7	2.7	3.7	1.3	1.8	3.6	0.9
	Music	96.0	94.0	98.3	94.7	95.3	98.2	95.3	95.8	98.8
	Noise	2.0	2.0	0.0	2.7	1.0	0.5	2.9	0.7	0.3
Sources 1+2	Speech	68.0	8.0	23.3	60.7	7.3	14.7	63.8	6.0	17.1
	Music	32.0	92.0	76.7	39.3	92.7	85.3	36.2	94.0	82.9
	Noise	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

abruptly. If speech is reverberated its envelope is temporally smoothed and thus resembles the envelope of music. The introduction of background noise in the speech training data increases the classification performance for speech signals immensely (Tables 5 and 7). When only Source 1 is active, all microphone signals are classified as speech with more than 90% accuracy. When both sources are active, the signals captured near the speech source (Mic. 2 in Table 5 and Mic. 1 in Table 7) are classified as speech with an increased performance. However, the classification accuracy of speech for this case still is only at around 60–70% when the recorded data is classified.

4.3. Fuzzy clustering of microphones in reverberant environments with two sources

Fig. 7 presents one of the ten scenarios of microphone distributions for one of the three simulated room setups and the result of a feature based fuzzy clustering of the devices when simulating an active speech and music source. The membership value estimated for the devices is indicated by the coloration of the dots. The three columns indicate the different clusters 1–3. The clustering works very well for the simulated data. The critical distance is indicated as black circles around the sources and delivers a good boundary for assigning a microphone to a source related cluster or the diffuse reverberation cluster. The chosen feature representation allows for a good separation between dominant direct speech, dominant direct music and mixed and reverberant signals. Similar results are obtained by the application of fuzzy clustering on the feature data that is extracted from the recorded audio data. Fig. 8 shows the result of the clustering for one of the three reverberation setups.

Table 8 presents averaged results for the fuzzy clustering over all ten scenarios for each simulated Rooms 1–3. The mean membership value $\overline{\mu}$ of all members of an estimated cluster n is larger than 0.7 for all rooms and grows with the room size. This indicates a better separation when the distance between the clusters is increasing. Further, the normalized distance $\tilde{d}_{s,n}$ of the position of an estimated cluster $n=1,2,3$ (represented by the averaged positions of the members of the cluster) to Source 1 $\tilde{d}_{1,n}$ and to Source 2 $\tilde{d}_{2,n}$ is presented in Table 8 for the simulated room with a speech and a music source. The normalization is performed on the actual source-to-source distance and is necessary, as the source-to-source distance and therefore the desired distance between the clusters 1 and 2 is different for each simulated scenario. For all simulated microphone–source setups, clusters 1 and 2 are located near their respective sources 1 and 2, indicated by small values $\tilde{d}_{1,1}$ and $\tilde{d}_{2,2}$ respectively. Both the normalized distances $\tilde{d}_{1,2}$ and $\tilde{d}_{2,1}$ are close to one. Therefore, the estimated clusters represent the distribution of sources in a room very well. Further, the third 'diffuse' cluster ($n=3$) is located with distances $0.38 \leq \tilde{d}_{s,3} \leq 0.7$ which can be interpreted as a good result for a microphone distribution all over the room. When the recorded data is used, the mean membership value $\overline{\mu}$ of all members of an estimated cluster n is larger than 0.6 and relatively constant for all reverberation times (Table 9). The evaluation of the normalized distances shows that the estimation of the speech cluster works well as $\tilde{d}_{1,1} \leq 0.15$ for all the reverberation setups. The normalized distance $\tilde{d}_{2,2}$ is a bit larger which might be related to some cases in which microphones outside of the critical distance are assigned to cluster 2. The values for the normalized inter-cluster distances $\tilde{d}_{1,2}$ and $\tilde{d}_{2,1}$ and for the normalized distances $\tilde{d}_{1,3}$ and $\tilde{d}_{2,3}$ confirm the good results from the simulated setups.

To automatically estimate which two of the three clusters are those that are dominated by an active source and which one is the reverberation and mixture dominated cluster we evaluate the order of the fuzzy membership value for each cluster. Table 10 shows that cluster 3 is detected as the mixed reverberation cluster in more than 89% of the cases based on simulated data and in more than 85% of the cases based on recorded data.

4.4. Combined experiment: fuzzy clustering, feature combination, and classification

Table 11 presents the results for the classification performance of the complete algorithm, including the cluster estimation, the feature vector combination and the training of speech data including pink noise. For the scenario of an active speech and music source, speech in

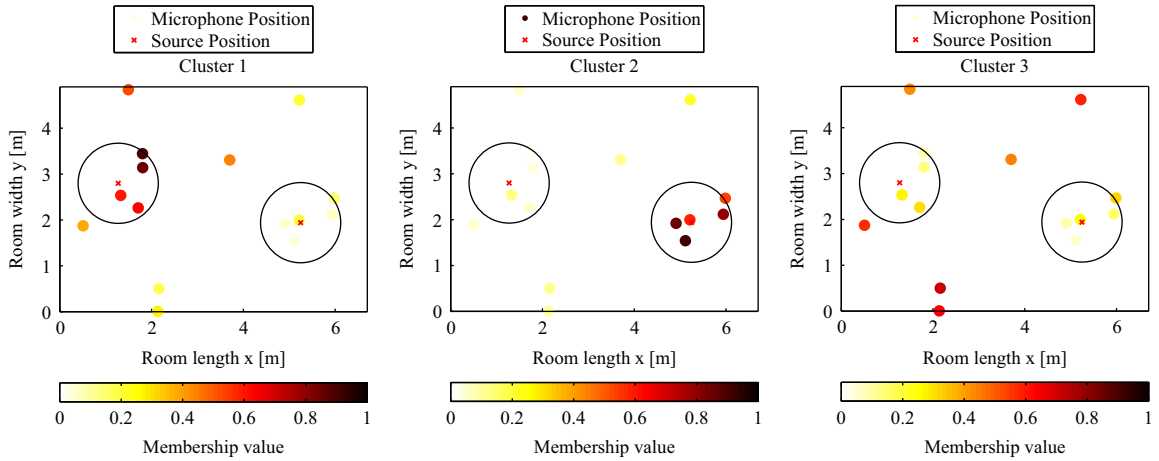

Fig. 7. Result of fuzzy clustering based on the extracted audio features, for the simulated Room 2. The critical distance is indicated.

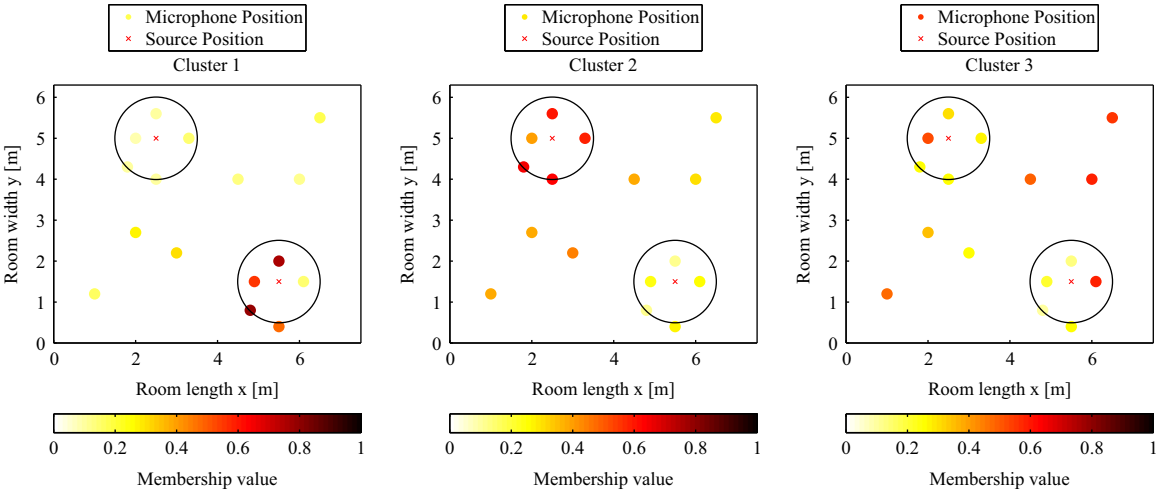

Fig. 8. Result of fuzzy clustering based on the extracted audio features of recorded signals, $T_{60} = 500$ ms. The critical distance is indicated.

Table 8
Averaged fuzzy weighting coefficient $\bar{\mu}$ for devices in clusters $n = 1, 2, 3$ and normalized distance $\tilde{d}_{1,n}$ to Source 1 (speech) and $\tilde{d}_{2,n}$ to Source 2 (music) based on simulated audio data.

	Room 1			Room 2			Room 3		
	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
$\bar{\mu}$	0.72	0.72	0.73	0.75	0.74	0.75	0.78	0.78	0.80
$\tilde{d}_{1,n}$	**0.09**	1.02	0.67	**0.11**	1.00	0.56	**0.08**	0.99	0.50
$\tilde{d}_{2,n}$	1.01	**0.12**	0.58	1.03	**0.10**	0.70	1.02	**0.14**	0.68

Table 9
Averaged fuzzy weighting coefficient $\bar{\mu}$ for devices in clusters $n = 1, 2, 3$ and normalized distance $\tilde{d}_{1,n}$ to Source 1 (speech) and $\tilde{d}_{2,n}$ to Source 2 (music) based on the recorded signals.

	$T_{60} = 300$ ms			$T_{60} = 500$ ms			$T_{60} = 700$ ms		
	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
$\bar{\mu}$	0.65	0.65	0.64	0.65	0.67	0.65	0.63	0.63	0.65
$\tilde{d}_{1,n}$	**0.11**	0.88	0.45	**0.12**	0.83	0.46	**0.15**	0.84	0.47
$\tilde{d}_{2,n}$	0.94	**0.20**	0.62	0.92	**0.26**	0.64	0.90	**0.24**	0.62

cluster $n=1$ and music in cluster $n=2$ are detected with a very good accuracy. The classification based on manually selected high SNR reference microphones (Table 12) is less accurate for the speech cluster. When the algorithm selects a single microphone automatically based on the fuzzy membership value, the performance of correct speech classification in cluster $n=1$ decreases further which indicates a suboptimal selection of a single microphone in terms of the SNR at the microphone for this cluster. Cluster $n=3$ is mainly classified as music when the combination strategy is used as well as when single microphone signals are classified. This is a plausible result, as firstly, for example in speech pauses, music is present at all sensors in the room, and secondly the influence of reverberation causes a shift towards the classification result music as shown in the previous experiments (Section 4.2). There is no distinct class to identify the third cluster as the diffuse one with a high amount of

Table 10
Detection rate (in %) of cluster n ($n=1$: close to speech source, $n=2$: close to music source, $n=3$: reverberation and mixture dominated) to be the mixed cluster. The evaluation is based on the averaged fuzzy membership values of all microphones within a cluster. The results of the detection are averaged over all three reverberation times of the simulation or recording setups.

Simulations			Recordings		
$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
7.6	3.3	89.1	8.7	6.0	85.3

Table 11
Combined clustering into clusters $n=1,2,3$ and cluster based classification (in %) in a speech/music scenario for the three different simulated Rooms 1–3.

	Room 1			Room 2			Room 3		
Class	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
Speech	97.7	0.9	16.4	93.9	0.4	7.9	88.3	0.4	4.2
Music	2.3	91.7	83.6	6.1	96.7	92.1	11.7	96.7	95.8
Noise	0.0	7.4	0.0	0.0	2.9	0.0	0.0	2.9	0.0

Table 12
Classification (in %) based on single microphone signals in a speech/music scenario averaged for the three different simulated Rooms 1–3. Two results are presented. First, we manually pick the reference microphones closest to the sources in clusters $n=1$ and $n=2$ and the microphone farthest away from both sources for cluster $n=3$. Second, the microphones with the highest estimated fuzzy membership value for each estimated cluster are selected.

	Reference microphone			Estimated microphone		
Class	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
Speech	75.1	2.3	9.8	57.5	2.6	9.6
Music	24.8	97.3	90.2	42.5	97.1	90.4
Noise	0.0	0.4	0.0	0.0	0.3	0.0

Table 13
Combined clustering into clusters $n=1,2,3$ and cluster based classification (in %) in a speech/music scenario based on the recorded audio data for three different reverberation times.

	$T_{60}=300$ ms			$T_{60}=500$ ms			$T_{60}=700$ ms		
Class	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
Speech	98.0	8.0	16.8	88.0	4.0	8.4	89.3	2.7	10.9
Music	2.0	92.0	83.2	12.0	96.0	91.6	10.7	97.3	89.1
Noise	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

reverberation and signal mixtures. However, it was shown that the information provided by the fuzzy clustering can be used to estimate which of the three clusters is this diffuse cluster. Therefore, the algorithm gives a good indication about which cluster of microphones is in the vicinity of the speech or of the music source and which receives mainly reverberated signal mixtures. The evaluation of the performance of the algorithm based on the recorded audio data is shown in Table 13. Again, the speech training data of the LDA classifier incorporates a low amount of pink noise. Speech in cluster 1 and music in cluster 2 are identified with a high accuracy of more than 88%. The classification results for cluster 3 are similar to the results from the simulated data. Compared to classification based on single microphone signals in Table 14, we see an improvement of correct speech classification in cluster $n=1$ when the CTA strategy is used. Here, the classification performance of the automatically selected microphone for each cluster is similar to the classification performance based on the manually selected high SNR reference microphones.

Table 14
Classification (in %) based on single microphone signals in a speech/music scenario averaged for the three different recording scenarios. Two results are presented. First, we manually pick the reference microphones closest to the sources in clusters $n=1$ and $n=2$ and the microphone farthest away from both sources for cluster $n=3$. Second, the microphones with the highest estimated fuzzy membership value for each estimated cluster are selected.

	Reference microphone			Estimated microphone		
Class	$n=1$	$n=2$	$n=3$	$n=1$	$n=2$	$n=3$
Speech	61.7	7.1	18.4	63.3	8.5	10.3
Music	38.3	92.9	81.6	36.7	91.5	89.6
Noise	0.0	0.0	0.0	0.0	0.0	0.1

5. Conclusions

The automatic classification of audio signals constitutes a difficult task when the signals are modified and disturbed by the environment. In this paper, we introduce a three-stage algorithm for the classification of reverberant audio sources based on the signals that are captured by an ad hoc microphone array. In the first stage, Mod-MFCC audio features are extracted. Based on these, clusters of microphones are estimated in the second stage, each dominated by either an audio source or by reverberation. In the third stage, information is shared and combined within and in between the clusters such that enhanced features are obtained, and the combined information is used to perform a cluster related classification. It is shown that the cluster estimation works well using the extracted feature data. The information of the fuzzy clustering further allows for a detection of the cluster of microphones that receive mainly signal mixtures and reverberation components with an accuracy of more than 85%. Given a correct cluster assignment, a classification of the sources in their respective clusters with an accuracy of at least 88% can be performed when the feature combination strategy is applied. This outperforms the classification based on a single microphone for each cluster. The classifier is trained on clean and anechoic data and no training is performed on specific reverberation conditions. However, the performance of speech classification is increased by the addition of pink noise to the speech data in the training step, which reduces the mismatch between clean and anechoic training and reverberated test data in the feature domain. The evaluation for speech and music classification with recorded audio data from a multi-microphone setup

reinforces the findings obtained with simulated data. The algorithm still has potential for improvement: evaluations showed that the selection of one microphone per cluster based on the fuzzy membership value does not result in an optimal microphone selection, e.g., as reference microphone for a cluster. Here, a strategy to select the microphone that is probably closest to an active source in a room is desirable. Furthermore, if information about a specific reverberation scenario is available for an application, this could be used to create more customized training data. The conducted experiments with two sound sources in a room represents a realistic everyday scenario. However, research on the effect of a higher number of sources in a room on the performance of the algorithm is of interest. For this, an estimation of the number of sound sources in a room should be incorporated to adjust the number of clusters automatically.

References

[1] V. Hamacher, J. Chalupper, J. Eggers, E. Fischer, U. Kornagel, H. Puder, U. Rass, Signal processing in high-end hearing aids: state of the art, challenges, and future trends, EURASIP J. Appl. Signal Process. 18 (2005) 2915–2929.

[2] M. Büchler, S. Allegro, S. Launer, N. Dillier, Sound classification in hearing aids inspired by auditory scene analysis, EURASIP J. Adv. Signal Process. 18 (2005) 2991–3002.

[3] H. Feng, C. Jiang, X. Yang, An audio classification and speech recognition system for video content analysis, in: Proceedings of the International Conference on Multimedia Technology (ICMT), 2011, pp. 5272–5276.

[4] M. Vacher, F. Portet, A. Fleury, N. Noury, Challenges in the processing of audio channels for ambient assisted living, in: Proceedings of the 12th International Conference on E-health Networking, Application & Services (HealthCom), 2010.

[5] A. Bertrand, Applications and trends in wireless acoustic sensor networks: a signal processing perspective, in: IEEE Symposium on Communications and Vehicular Technology (SCVT), 2011, pp. 1–6.

[6] R. Lienhart, I. Kozintsev, S. Wehr, M. Yeung, On the importance of exact synchronization for distributed audio signal processing, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2003, pp. 139–144.

[7] J. Dmochowski, L. Zicheng, P. Chou, Blind source separation in a distributed microphone meeting environment for improved teleconferencing, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2008, pp. 89–92.

[8] L. Zicheng, Sound source separation with distributed microphone arrays in the presence of clock synchronization errors, in: Proceedings of the International Workshop on Acoustic Echo and Noise Control (IWAENC), 2008.

[9] Y. Hioka, B. Kleijn, Distributed blind source separation with an application to audio signals, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 233–236.

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

[11] S. Markovich-Golan, S. Gannot, I. Cohen, Distributed GSC beamforming using the relative transfer function, in: Proceedings of the 20th European Signal Processing Conference (EUSIPCO), 2012, pp. 1274–1278.

[12] S. Markovich-Golan, S. Gannot, I. Cohen, Performance of the SDW–MWF with randomly located microphones in a reverberant enclosure, IEEE Trans. Audio Speech Lang. Process. 21 (2013) 1513–1523.

[13] A.R. Abu-El-Quran, R.A. Goubran, A.D.C. Chan, Security monitoring using microphone arrays and audio classification, IEEE Trans. Instrum. Measur. 55 (2006) 1025–1032.

[14] U. Srinivas, N. Nasrabadi, V. Monga, Graph-based multi-sensor fusion for acoustic signal classification, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2013, pp. 261–265.

[15] R. Scharrer, M. Vorländer, Sound field classification in small microphone arrays using spatial coherences, IEEE Trans. Audio Speech Lang. Process. 21 (2013) 1891–1899.

[16] H. Kuttruff, Room Acoustics, Applied Science Publishers Ltd, London, 1979.

[17] D.L. Hall, J. Llinas, Handbook of Multisensor Data Fusion, CRC Press, Boca Raton, 2001.

[18] E. Scheirer, M. Slaney, Construction and evaluation of a robust multifeature speech/music discriminator, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1997, pp. 1331–1334.

[19] M. McKinney, J. Breebaart, Features for audio and music classification, in: Proceedings of the International Conference on Music Information Retrieval (ISMIR), 2003.

[20] A. Nagathil, P. Gottel, R. Martin, Hierarchical audio classification using cepstral modulation ratio regressions based on legendre polynomials, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 2216–2219.

[21] A. Meng, P. Ahrendt, J. Larsen, L. Hansen, Temporal feature integration for music genre classification, IEEE Trans. Audio Speech Lang. Process. 15 (2007) 1654–1664.

[22] A. Holzapfel, Y. Stylianou, Musical genre classification using non-negative matrix factorization-based features, IEEE Trans. Audio Speech Lang.Process. 16 (2008) 424–434.

[23] R. Martin, A. Nagathil, Cepstral modulation ratio regression (CMRARE) parameters for audio signal analysis and classification, in: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2009, pp. 321–324.

[24] S. Furui, Digital Speech Processing, Synthesis, and Recognition, Marcel Dekker, Inc., New York, 2000.

[25] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacrétaz, D.A. Reynolds, A tutorial on text-independent speaker verification, EURASIP J. Appl. Signal Process. 2004 (2004) 430–451.

[26] P.N. Garner, Cepstral normalisation and the signal to noise ratio spectrum in automatic speech recognition, Speech Commun. 53 (2011) 991–1001.

[27] J. Bezdek, R. Ehrlich, W. Full, FCM: the fuzzy c-means clustering algorithm, Comput. Geosci. 10 (1984) 191–203.

[28] L. Zadeh, Fuzzy sets, Inf. Control 8 (1965) 338–353.

[29] M.S. Yang, A survey of fuzzy clustering, Math. Comput. Model. 18 (1993) 1–16.

[30] S. Gergen, R. Martin, Linear combining of audio features for signal classification in ad-hoc microphone arrays, Unpublished Results.

[31] S. Gergen, C. Borß, N. Madhu, R. Martin, An optimized parametric model for the simulation of reverberant microphone signals, in: Proceedings of the International Conference on Signal Processing, Communications and Computing (ICSPCC), 2012, pp. 154–157.

[32] J. Allen, D.A. Berkley, Image method for efficiently simulating small-room acoustics, J. Acoust. Soc. Am. 65 (1979) 943–950.

[33] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, 2nd ed. Wiley, New York, 2001.

[34] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, V. Zue, TIMIT Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium, Philadelphia, 1993.

[35] E. Hänsler, G. Schmidt, Speech and Audio Processing in Adverse Environments, Springer, Berlin, Germany, 2008.

[36] J. Abonyi, Fuzzy Clustering and Data Analysis Toolbox, 2005. URL ⟨http://www.abonyilab.com/software-and-data/fclusttoolbox⟩.

[37] T.C. Lawin-Ore, S. Doclo, Reference microphone selection for MWF-based noise reduction using distributed microphone arrays, in: Proceedings of the ITG Conference on Speech Communication, 2012, pp. 154–157.