published in: *Detection and Classification of Acoustic Scenes and Events (DCASE) Workshop, 3rd September 2016, Budapest, Hungary*

BAG-OF-FEATURES ACOUSTIC EVENT DETECTION FOR SENSOR NETWORKS

Julian Kürby, Rene Grzeszick, Axel Plinge, and Gernot A. Fink

TU Dortmund University, Dortmund, Germany

ABSTRACT

In this paper a novel approach for acoustic event detection in sensor networks is presented. Improved and more robust recognition is achieved by making use of the signals from multiple sensors. To this end, various known fusion strategies are evaluated along with a novel method using classifier stacking. A comparative evaluation of these fusion strategies is performed on two different datasets: the ITC-Irst database, and a set of smart room recordings. In both datasets, 32 distributed microphones were used for recording. Furthermore, the effect of previously observed as well as unobserved locations is investigated. The proposed stacking yields a notable improvement. The performance of recognizing events at previously unobserved locations can be improved by sorting the channels according to their posterior probabilities.

***Index Terms*—** Bag-of-Features, Acoustic Event Detection, Sensor Arrays, Robustness, Acoustic Sensor Networks

1. INTRODUCTION

The detection and classification of acoustic events is important for many practical applications in various environments: The recognition of such events can be used for meeting and online lecture analysis and annotation [1]. Surveillance in cluttered scenes can be improved by an acoustic analysis in order to detect unexpected scenarios that are not easily visually recognizable (e. g. screams or glass breaking) [2]. In a slightly different field of research outdoor applications are addressed. These include mobile robots for security [3], urban planning [4], and the analysis of possible noise complaints [5]. It can also be used to improve the robustness of different real world applications, such as speech enhancement, speaker tracking, or the calibration of microphone arrays [6–8]. What makes this problem difficult is the vast diversity of the acoustic events.

Methods for online analysis of acoustic events are typically applied over short time windows and combined with a sliding window approach. One common approach stems from speaker identification [9]. A Gaussian mixture model (GMM) is trained for each class. The estimates of all GMMs are summed up over all frames and the class with the highest likelihood is chosen. These methods are sometimes termed 'Bag-of-Frames' [10,11]. Over the last years, methods that build on the Bag-of-Features (BoF) principle have emerged in the field of acoustic event detection [12,13]. There, features are clustered in order to obtain a histogram representation which is then classified. The BoF principle has been proven to generalize well with respect to the diversity of the acoustic events.

Many methods in acoustic event detection focus on a single signal. However, in many scenarios a sensor network with multiple microphones is available (cf. [8,14]). In [15] multiple channels are used to extract features describing spatial information. These features work well for classifying scenes where the sound sources occur at distinct locations. In [16] a multi-channel approach that uses Regression Forests is proposed. The confidence scores of different channels are accumulated and then the presence is predicted using a pre-defined threshold. In [17] different combination strategies including accumulation of log probabilities, the maximum rule, and majority voting are evaluated. Both works show that the combination of information obtained from multiple channels improves the robustness of the system and the detection results.

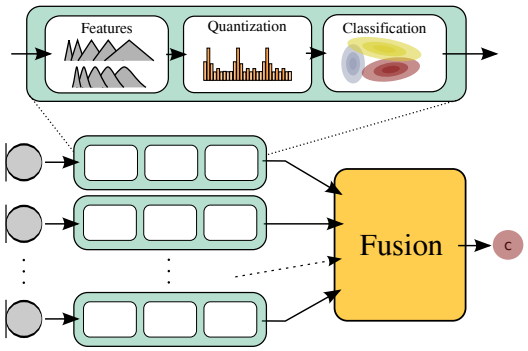

Figure 1: Overview of the proposed method. A three-step BoF approach for acoustic event detection is applied to every source in a sensor network comprised of many microphones, before the results are combined by a fourth fusion step.

This paper extends the BoF approach discussed in [13,18] and provides a thorough evaluation of different multi-channel fusion strategies in the context of acoustic event detection. The heuristic combination strategies presented in [17] are compared with a novel method based on classifier stacking. A comparative evaluation on two different datasets is given. Furthermore, different training and test setups are evaluated having a closer look at the prerequisites necessary for successfully exploiting information from multiple sources.

2. METHOD

For the acoustic event detection in sensor networks, a single channel Bag-of-Features (BoF) approach is extended to multiple channels by adding an additional fusion step that combines the information from different microphones. A sliding window approach is used for detection. For each window, four basic processing steps are applied, as shown in Fig. 1:

1. Given an input signal and a short time window, a set of feature vectors is calculated for all frames in this window.
2. The feature vectors of all frames in the training set are clustered in a supervised manner using a GMM for each class. The features within one window are assigned to the clusters using soft assignment. These are accumulated in a histogram, the BoF representation.

3. These representations are then used for classification, applying maximum likelihood classification.
4. The results from multiple channels are fused in order to get a more robust classification. A novel fusion strategy based on classifier stacking is proposed.

2.1. Single-channel BoF acoustic event classification

For the single-channel BoF based acoustic event classification, a single microphone or beamformed signal is processed in short time windows. The processing steps are explained in more detail in the following.

Features Given an input signal and a time window n of w milliseconds, a set of feature vectors $Y_n = (y_1 \dots y_K)$ is calculated. For sound and especially speech processing, the mel frequency cepstral coefficients (MFCCs) are one of the most widely used features. The input signal is filtered by a mel frequency filter bank, from the logarithm of its magnitude the discrete cosine transform (DCT) is computed and its second to 13^{th} coefficient is used. From that the gammatone frequency cepstral coefficients (GFCCs) were derived in [19]. Here, the filterbank of the MFCCs is replaced by linear phase gammatone filters. As for the MFCCs, the second to 13^{th} GFCC coefficients are used. In addition, a single loudness filter is evaluated. In total the feature vector has a dimensionality of 27. A whitening transformation is computed on the training data which is applied to all feature vectors.

Feature Representation A BoF approach is used for building a codebook of *acoustic words* from the training set. While the classical BoF uses hard quantization via the k-Means algorithm, soft quantization by GMMs has been shown to improve the performance [13,20]. The basic principle also employs a globally estimated codebook which can lead to mitigation of significant differences. A remedy for this effect is to build codebooks of size I for all C classes Ω_c separately and then concatenating them into a large super-codebook [13]. Here, the expectation maximization (EM) algorithm is applied to all feature vectors $\boldsymbol{y}_k$ for each class Ω_c in order to estimate I means and standard deviations $\mu_{i,c}, \sigma_{i,c}$ for all C classes. All means and deviations are concatenated into a super-codebook $\boldsymbol{v}$ with $V = I \cdot C$ elements

$$v_{j=(I \cdot c+i)} = (\mu_{i,c}, \sigma_{i,c}) \tag{1}$$

where the index j is computed from the class index c and the Gaussian index i as $j = I \cdot c + i$. Using this codebook, a soft quantization of a feature vector $\boldsymbol{y}_k$ can be computed as

$$q(\boldsymbol{y}_k, v_j) = \mathcal{N}(\boldsymbol{y}_k | \mu_j, \sigma_j) \; / \; \sum_{j'} \mathcal{N}(\boldsymbol{y}_k | \mu_{j'}, \sigma_{j'}) \; . \tag{2}$$

Then, a histogram $\boldsymbol{b}$ can be computed over all K frames of the input window by

$$b(Y_n, v_j) = \frac{1}{K} \sum_k q(\boldsymbol{y}_k, v_j) \; . \tag{3}$$

Classification The probability $P(v_j|\Omega_c)$ of an acoustic word v_j given class Ω_c is estimated using a set of training samples $Y_n \in \Omega_c$ for each class c by Lidstone smoothing:

$$P(v_j|\Omega_c) = \frac{\alpha + \sum_{Y_n \in \Omega_c} b(Y_n, v_j)}{\alpha V + \sum_{m=1}^{V} \sum_{Y_n \in \Omega_c} b(Y_n, v_m)} \tag{4}$$

A typical choice for the smoothing factor α is in the range of $[0, 1]$. Here, α is set to 0.5. Since all classes are assumed to be equally likely and have the same prior, maximum likelihood classification is used. The posterior is estimated using the relative frequency of all acoustic words

$$P(Y_n|\Omega_c) = \prod_{v_j \in \boldsymbol{v}} P(v_j|\Omega_c)^{b(Y_n, v_j)} \; . \tag{5}$$

2.2. Multi-channel fusion

In a sensor network containing M microphones the approach can be evaluated for each microphone m individually. It is assumed, that all microphones are synchronized at least at a frame level. The results can then be combined in order to obtain a more robust classification. In the following three traditional heuristic fusion strategies (cf. [17]) will be reviewed and a novel approach based on classifier stacking will be introduced.

Majority voting A straightforward fusion approach is evaluating each channel separately so that a set of class labels

$$\hat{c}_{(m)} = \underset{c}{\operatorname{argmax}} \; P_m(Y_n|\Omega_c) \tag{6}$$

is estimated. Then, a majority voting over all decisions $\hat{c}_{(m)}$ is performed. This assumes that most microphones are able to detect the correct event. However, it discards the posterior probabilities which might carry important information about the confidence of the single channels.

Maximum rule The maximum rule is a fusion strategy that considers the posterior probabilities of each channel instead of the labels. It chooses the class with the overall highest posterior probability. For each class the maximum over all channels is computed and then the class with the highest probability in the complete sensor network is chosen:

$$\hat{c} = \underset{c}{\operatorname{argmax}} \; \max_m \; P_m(Y_n|\Omega_c) \; . \tag{7}$$

This approach can be highly influenced by positive outliers. It is assumed that at least one microphone is positioned well with respect to the acoustic event.

Product rule Alternatively the product of the posterior probabilities is used. For each of the classes the product of the posteriors of all channels is computed. Then, the class with the highest probability product in the complete sensor network is chosen:

$$\hat{c} = \underset{c}{\operatorname{argmax}} \; \prod_m \; P_m(Y_n|\Omega_c) \; . \tag{8}$$

In contrast to taking the highest probability this strategy is strongly influenced by negative outliers.

Classifier stacking While the previous approaches are mere heuristic approaches that decide on a fusion strategy, it is also possible to learn a combination strategy from the training data. A second classifier is trained that uses the posterior probabilities from all microphones in the sensor network as input features. The learned classification function $\mathcal{F}$ is then used for predicting the class:

$$\hat{c} = \mathcal{F}\left((P_m(Y_n|\Omega_c))_{(c,m)} \right) \; . \tag{9}$$

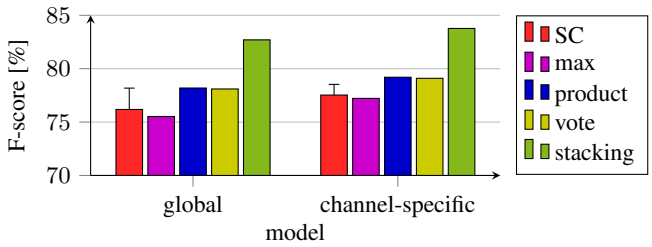

Figure 2: Frame-wise F-score [%] comparing fusion strategies with the single-channel (SC) baseline on the ITC-Irst dataset. For the single-channel results the mean and standard deviation are plotted.

Two thirds of the training data are used for training the single-channel BoF models and the single-channel classifiers and the last third is used in order to train a Random Forest classifier on the posterior probabilities of the single-channel evaluations.

Note that the classifier learns the probabilities based on their ordering. Therefore, it implicitly learns the position of the microphones and also the locations at which the different acoustic events occur. This can be an advantage for events or especially noise sources with a fixed location (e. g. doors or windows). However, it can be a limitation for events that can occur at arbitrary locations such as speech. A remedy for this effect is ordering the M channels according to the highest posterior probability. Sorting the channels descending by probability provides a new ordering:

$$\mathcal{M} = \left[\underset{m}{\operatorname{argsort}} \max_{c} \; P_m(Y_n|\Omega_c) \right] \tag{10}$$

After re-ordering of the channels, the posterior probabilities are again used as input for the classifier $\mathcal{F}$. Thus, the indices (c, m) in eq. 9 are replaced by $(c, \mathcal{M}_m)$.

2.3. Detection

Due to its simplicity and rapid computation, the BoF approach can easily be adapted to event detection, where a sequence of acoustic events is given. It currently runs in approx. 20% real time on a single core i7 cpu. The classification window is moved forward in a sliding window approach by one frame k at a time. The recognition result is used for the frame that is centered in the window so that context information is available for a short time before and after the frame. As the window has a length of w milliseconds, there is a processing delay of only $w/2$ milliseconds.

3. EVALUATION

The experiments are conducted on two different datasets for acoustic event detection, the ITC-Irst dataset [14] as well as a set of recordings conducted in a smart conference room at TU Dortmund University. On these datasets the detection performance of the presented multi-channel approaches are evaluated and compared to a single-channel baseline. All channels were synchronized with a global clock in both datasets.

3.1. ITC-Irst Dataset

The ITC-Irst dataset is comprised of 16 different acoustic events, including *door knock, door slam, steps, chair moving, spoon (cup jingle), paper wrapping, key jingle, keyboard typing, phone ring, applause, cough, laugh, door open, phone vibration, mimo pen buzz, falling object,* and *unknown/background*. The recording room was equipped with 32 microphones, 28 of which were located in seven T-shaped arrays on the walls and four were table microphones. The experiments consist of twelve recording session on three different days. The first three sessions of each day are considered as training and the fourth session is used for testing.

The first experiments were conducted using all sounds except silence and unknown as classes of interest. Then, in order to allow for comparability with existing experiments [14,16], only the first twelve classes were considered as foreground and the remaining ones as background.

evaluation	method	channels	error	F-score
event-based	RF [16]	mean (4)	15.4%	91.8%
	RF [16]	fusion (4)	**13.0%**	**93.3%**
	HMM2 [14]	SC (1)	23.6%	-
	HMM1 [14]	SC (1)	45.2%	-
	SVM [14]	SC (1)	64.4%	-
frame-based	RF [16]	fusion (4)	30.7%	82.8%
	proposed	mean (32)	39.0%	77.4%
	proposed	stacking (32)	**25.6%**	**84.2%**

Table 1: Results on the ITC-Irst dataset using the CLEAR evaluation protocol with the first 12 classes as foreground in comparison to literature results. The methods use either a single channel or different fusion approaches (number of channels in parentheses).

Baseline For the evaluation, two different setups were considered. First, the BoF model is trained on the events of all microphones yielding a global model. Second, a separate model is trained for each microphone in the sensor network. In both cases, the BoF model is computed using a codebook size of $I = 30$ centroids for each class and a window size of $w = 600$ ms based on the results in [18]. For the baseline each channel is evaluated separately and the average over all microphones is reported (single channel is denoted as SC).

Fusion experiments In the following the multi-channel fusion strategies are compared with each other and to the baseline of single-channel results. The first two sessions of each day are used for training the base classifier, the third session for training the stacking classifier. Since the positions of the acoustic events were changed for each of the three recording days, the stacking classifier is able to learn different acoustic locations. The fourth session is used for testing so that it contains the different locations from all three days. Note that the single-channel and heuristic approaches are trained on the complete training set. The frame-wise F-scores are shown in Fig. 2. The models that are trained for every channel separately perform much better than a single global model. Furthermore, it can be seen that the classifier stacking that learns a fusion strategy from the training data outperforms the heuristic approaches.

Literature comparison For comparison with the literature, only the first twelve classes are used as foreground (cf. [14,16]). Regression Forest (RF) were evaluated in combination with a multi-channel fusion approach using this setup [16]. Note that only four channels were used for evaluating the RF while the proposed approach is able to incorporate all 32 microphones. In contrast to [16], a frame-based evaluation protocol is used in this paper. The

set	model	SC	max	prod.	vote	stacking	sorted (32)	sorted (5)
separate	global	$10.7 \pm 4.2\%$	$8.5 \pm 3.5\%$	$9.3 \pm 3.3\%$	$9.5 \pm 3.2\%$	$7.6 \pm 3.0\%$	$7.2 \pm 2.3\%$	$\mathbf{7.1 \pm 2.4\%}$
	channel-wise	$12.2 \pm 4.0\%$	$9.2 \pm 3.8\%$	$9.4 \pm 3.3\%$	$9.8 \pm 3.1\%$	$10.0 \pm 2.9\%$	$9.5 \pm 2.7\%$	$\mathbf{8.4 \pm 2.7\%}$
mixed	global	$7.6 \pm 3.2\%$	$5.1 \pm 1.3\%$	$6.2 \pm 1.7\%$	$6.5 \pm 1.7\%$	$\mathbf{3.2 \pm 0.9\%}$	$3.4 \pm 1.0\%$	$3.8 \pm 1.1\%$
	channel-wise	$6.3 \pm 2.6\%$	$4.6 \pm 1.6\%$	$5.3 \pm 1.8\%$	$5.5 \pm 1.7\%$	$\mathbf{2.7 \pm 1.0\%}$	$2.7 \pm 0.9\%$	$2.8 \pm 0.8\%$

Table 2: Mean frame-wise classification error and its standard deviation over five splits of the position experiments. In "separate", the events of the training and test set occur on different sides of the smart room, in "mixed" on both.

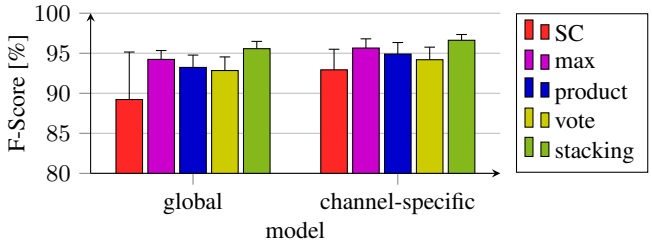

Figure 3: Mean F-score [%] and the standard deviation over the five splits of smart room recordings using different fusion strategies on the dataset.

Acoustic Frame Error Rate (AFER) is calculated analogously to the Acoustic Event Error Rate (AEER) [14], but with respect to the frames[1]. The frame-wise results of [16] are calculated on the resulting sequences, that were kindly provided by the authors. The results are shown in Tab. 1, reporting F-Scores, AEER (event-based error) and AFER (frame-based error). Additionally the event-based results of the RF and the CLEAR evaluation [14] are shown in Tab. 1. The CLEAR evaluation compared two different HMM and an SVM approach for acoustic event detection on this setup using a single microphone. The event-based results show that RFs achieve state of the art results.

On a frame level the proposed approach yields similar performance to the RF. The proposed classifier stacking shows the best results with 25.6% AFER and 84.2% F-score. The performance which is obtained using the stacking improves the results by a margin compared to the single-channel performance.

3.2. Smart room recordings

An additional set of acoustic events has been recorded in a smart room at TU Dortmund University.[2] The room is equipped with 32 microphones of which 16 are located at the table and the remaining 16 are mounted at the ceiling. The acoustic events were located at multiple positions in the room without any overlap. There is a lot of structure-borne noise changing the characteristics of sounds based on the microphones location. 19 sound categories have been recorded: *applause, chairs, cups, door, doorbell, doorknock, keyboard, knock, music, paper, phonering, phonevibration, pouring, screen, speech, steps, streetnoise, touching, ventilator,* and *silence*. Following the approach proposed in [5], acoustic events that are longer than five seconds were split into blocks of up to four seconds.

[1] A similar evaluation has been proposed for the DCASE2016 challenge referred to as a segment based metric.

[2] The dataset is publicly available as *Multi-channel acoustic event dataset* at http://patrec.cs.tu-dortmund.de/cms/en/home/Resources/

General experiments For generating different training and test sets five random splits were performed. Each split randomly selects two thirds of the data from each class for training and the remaining third for testing. For the stacking experiments the training data is randomly divided in two thirds for training the single-channel models and classifiers and the other third for training the stacking classifier. The sliding window is evaluated within the annotated four second blocks. All classes are considered as foreground events, using *silence* as background. The results are shown in Fig. 3. As for the ITC-Irst experiments, the stacking approach outperforms the heuristic fusion strategies. The best results are obtained using channel-specific models and classifier stacking which yields a frame-wise F-Score of $96.6 \pm 0.7\%$. This is an improvement of 3.7% compared to the mean results of the single-channel evaluation.

Position experiments The dataset contains a set of nine classes occurring on multiple positions (*applause, door, doorknock, keyboard, music, phonering, phonevibration, speech,* and *ventilator*). Here, the data has been recorded on different sides of the room (left & right respectively). Again five splits have been computed. Each split randomly selects the data of an acoustic event from the left or right side for training and the other side for testing, and vice versa. Hence, the robustness of the stacking classifier toward location changes can be investigated. The classification results are reported as the frame-wise classification error in Tab. 2. The classification error is used, because in this experiment all classes are considered as foreground. For comparison, five mixed sets using data from both sides for testing and training are also shown. As expected, the proposed stacking approach works well if a diverse set of training samples is provided. However, there is a drop in the performance when the locations in the test differ from the training set. This limitation can be overcome by sorting the input for the stacking classifier. In Tab. 2 the results for all 32 and the first 5 microphones of the sorted set $\mathcal{M}$ (denoted as sorted (32) and sorted (5) respectively) are shown (see eq. 10). Interestingly, the global model seems more robust toward reducing the information covered by the training set. This is probably due to the fact that multiple event locations and all microphones at different postions are used for training the model.

4. CONCLUSION

In this paper a multi-channel approach for acoustic event detection in sensor networks that builds on the Bag-of-Features principle has been presented. It was shown that combining the information from different channels allows for improving the performance of the recognition system. A novel fusion strategy that uses classifier stacking has been introduced which yields state-of-the-art results. Sorting the ordering of the microphones according to the posterior probability can overcome the requirement of having all locations in the training set.

5. REFERENCES

[1] I. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard, and D. Zhang, "Automatic analysis of multimodal group actions in meetings." *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 305–17, Mar. 2005.

[2] V. Carletti, P. Foggia, G. Percannella, A. Saggese, N. Strisciuglio, and M. Vento, "Audio surveillance using a bag of aural words classifier," in *IEEE Int. Conf. on Advanced Video and Signal Based Surveillance*, Aug. 2013, pp. 81–86.

[3] S. H. Young and M. V. Scanlon, "Robotic vehicle uses acoustic array for detection and localization in urban environments," *SPIE Proc. Mobile Robot Perception*, vol. 4364, pp. 264–273, Sept. 2001.

[4] D. Steele, J. D. Krijnders, and C. Guastavino, "The sensor city initiative: Cognitive sensors for soundscape transformations," GIS Ostrava, 2013.

[5] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *ACM Int. Conf. on Multimedia*, 2014.

[6] A. Plinge and S. Gannot, "Multi-microphone speech enhancement informed by auditory scene analysis," in *Sensor Array and Multichannel Signal Process. Workshop*, Rio de Janeiro, Brazil, 2016.

[7] A. Plinge and G. A. Fink, "Multi-Speaker tracking using multiple distributed microphone arrays," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Proces.*, May 2014.

[8] A. Plinge and G. A. Fink, "Geometry calibration of multiple microphone arrays in highly reverberant environments," in *Int. Workshop on Acoustic Signal Enhancement*, Sept. 2014.

[9] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011.

[10] J.-J. Aucouturier, B. Defreville, and F. Pachet, "The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music," *The Journal of the Acoustical Society of America*, vol. 122, no. 2, pp. 881–891, 2007.

[11] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events: An IEEE AASP challenge," in *IEEE Workshop on Applications of Signal Process. to Audio and Acoustics*, 2013.

[12] S. Pancoast and M. Akbacak, "Bag-of-audio-words approach for multimedia event classification," in *Interspeech*, 2012.

[13] A. Plinge, R. Grzeszick, and G. Fink, "A Bag-of-Features approach to acoustic event detection," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Proces.*, May 2014.

[14] A. Temko, R. Malkin, C. Zieger, D. Macho, C. Nadeu, and M. Omologo, "Clear evaluation of acoustic event detection and classification systems," in *Multimodal Technologies for Perception of Humans*, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds. Springer Berlin Heidelberg, 2007, vol. 4122, pp. 311–322.

[15] K. Imoto and N. Ono, "Spatial-feature-based acoustic scene analysis using distributed microphone array," in *European Signal Process. Conf.* IEEE, 2015, pp. 734–738.

[16] H. Phan, M. Maass, L. Hertel, R. Mazur, and A. Mertins, "A multi-channel fusion framework for audio event detection," in *IEEE Workshop on Applications of Signal Process. to Audio and Acoustics*, 2015.

[17] P. Giannoulis, G. Potamianos, A. Katsamanis, and P. Maragos, "Multi-microphone fusion for detection of speech and acoustic events in smart spaces," in *European Signal Process. Conf.* IEEE, 2014, pp. 2375–2379.

[18] R. Grzeszick, A. Plinge, and G. A. Fink, "Temporal acoustic words for online acoustic event detection," in *German Conf. on Pattern Recognition*, 2015.

[19] Y. Shao, S. Srinivasan, and D. Wang, "Incorporating auditory feature uncertainties in robust speaker identification," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Proces.*, 2007, pp. 277–280.

[20] S. Pancoast and M. Akbacak, "Softening Quatization in Bag-of-Audio-Words," in *IEEE Int. Conf. on Acoustics, Speech, and Signal Proces.*, 2014, pp. 1384–1388.