A SYSTEM FOR AUDIO SUMMARIZATION IN ACOUSTIC MONITORING SCENARIOS

David Damm, Dirk von Zeddelmann, Marc Oispuu, Miriam Häge, Frank Kurth

Fraunhofer-FKIE, KOM Department
Neuenahrer Str. 20, 53343 Wachtberg, Germany
phone: + (49) 228-9435868, fax: + (49) 228-856277, email: frank.kurth@fkie.fraunhofer.de
web: www.fkie.fraunhofer.de

ABSTRACT

A system for the acoustic monitoring of relevant audio events in real-life outdoor recording environments is presented. Targeting such realistic scenarios relies on the usage of robust and effective audio signal processing methods. In this paper, we address the key aspects of a corresponding audio monitoring system comprising an array of innovative acoustic sensors providing bearing information, robust, unsupervised methods for both detecting and localizing audio events, the integration of the acquired data streams, and a novel user interface for the audio-visual browsing of recorded audio scenes and detected audio events therein.

Index Terms— Audio monitoring system, acoustic event detection, acoustic localization, acoustic sensors.

1. INTRODUCTION

Heterogeneous sensor networks play an increasingly important role for surveillance and protection of critical infrastructures, major events, etc. In such monitoring scenarios, acoustic sensors yield important information on out-of-sight events or under conditions of bad visibility. A system for audio summarization typically has to report relevant acoustic events such as sounds produced by humans, vehicles, machines, or animals. Furthermore, the detection of conspicious events such as suddenly occuring, *anomalous*, noises or repeated sounds is of interest.

Acoustic event detection in realistic monitoring scenarios, generally outdoor recordings with uncontrolled noise conditions, results in particular challenges. Usually, such recordings are a mixture of different sound sources and are affected by heavy background noises that change over time. Using traditional methods for speech and audio processing, such as voice activity detection (VAD) and keyword spotting, in audio monitoring scenarios is problematic. One main reason is that those methods are usually developed for clean audio recordings and then adapted to particular types of distortion such as car-, babble-, or GSM-induced noise. Furthermore, those approaches usually require extensive amounts of training data to adapt integrated statistical classifiers to the properties of the noise environment or the expected speech. In monitoring scenarios, such an extensive training is not possible in most cases as noises are rarely predictable and individual speakers are a priori unknown.

In this paper, we address how acoustic event detection in realistic scenarios can be performed despite of the latter limitations. We first make some crucial observations: in contrast to classical speech processing approaches such as VAD, many audio monitoring applications do not require a precise temporal localization but only a rough estimate of the period of audio activity. In contrast to automatic speech recognition or keyword spotting, no exact transcription of speech components may be necessary, but only the identification of speech fragments contained in the recording with some probability. We therefore propose to model the aspect of roughness directly in the feature extraction process. By additionally adapting suitable technology from audio retrieval, we obtain a set of detection and matching techniques. Those allow for roughly detecting speech activity and spotting sequences of words which is sufficient for many monitoring scenarios. A method for extracting repeated acoustic events turns out to be particularly powerful in the monitoring context. To overcome the lack of training data, the proposed techniques work largely unsupervised and require only a limited amount of adaptation to environmental conditions. On the sensor side, an innovative type of acoustic sensor [1] is employed that allows us to incorporate localization information, hence further improving the accuracy of acoustical monitoring. To create a holistic system for audio summarization in monitoring scenarios, we propose a graphical audio event browser which allows for audio-visual navigation and playback of the recorded audio streams and detected audio events therein.

Several approaches of multimodal signal fusion for audio classification and localization have been proposed in recent years which, however, were mostly designed or tested for audio scenes with low or uniform noise levels. In [2], fusion of multimodal signals using canonical correlation analysis (CCA), and copula theory to detect the presence of a human using footstep signals from seismic and acoustic sensors, is presented. The proposed method is general and can be ex-

© EURASIP, 2012 - ISSN 2076-1465

tended to data obtained from other sensing modalities. Butko et al. [3] present a multimodal monitoring system for meeting room scenarios based on a feature-level fusion approach to improve the recognition rate of acoustic event detection using information from auditive and visual modalities. In [4], different types of classification and localization techniques for handling different indoor acoustic events in well-controlled office environments are evaluated. A sensor network-based acoustic source localization strategy is described which can cope a wide variety of sounds.

Our paper is organized as follows. Starting with the design of general audio features, Sect. 2 describes how to transfer methods for unsupervised audio retrieval to the audio monitoring context. In Sect. 3, aspects of the overall monitoring system are presented, covering the acoustic sensors, localization, data processing issues, and the novel multimodal user interface. Sect. 4 briefly describes the findings of a first case study for an outdoor monitoring scenario.

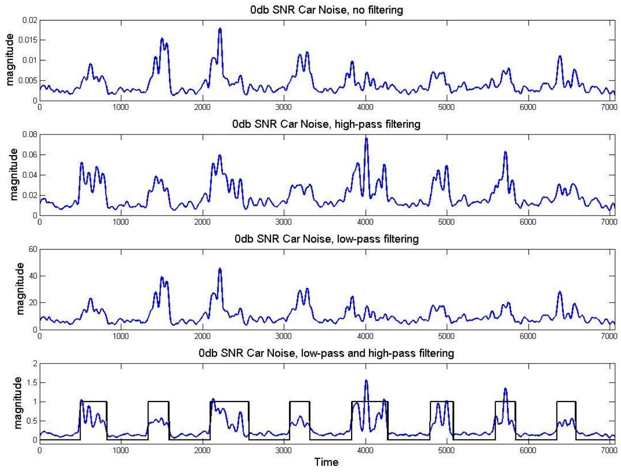


Fig. 1. Energy envelopes (top down): unfiltered, lowpass, highpass, combined high- and lowpass filtering.

2. ROBUST AUDIO MONITORING: METHODS

2.1. Feature Extraction

As a starting point to construct robust features, we follow an approach by Skowronski et al. [5] to generalize the well-known MFCC (Mel-Frequency Cepstral Coefficients) features by introducing an additional degree of freedom in the underlying filter bank. Whereas the filters of the classical MFCC-filterbank have bandwidths determined by the center frequencies of the adjacent bands, they propose to choose the bandwidth of the mel-spaced filters according to the bark scale of human perception [5]. The resulting features, called HFCCs (Human Factor Cepstral Coefficients), appear to better represent the phoneme progression in human speech independently of the speaker than MFCCs do. Subsequently, HFCC-ENS features derived from HFCCs were successfully employed to the unsupervised detection of short sequences of words in a given signal of recorded speech [6]. That approach extends HFCCs by computing short time energy normalized statistics (ENS), hence adapting the feature resolution from a standard of 100 Hz (or a step size of 20 ms) to a coarser resolution which is better suited to represent the typical phoneme resolution.

To allow adaptation of the audio features to the particular demands of our monitoring scenario, we propose to generalize this feature extraction process even more: In the latter, frontend (frame-based spectral analysis) and backend processing (decorrelating DCT) coincide with the well-known MFCC feature extraction. The mel-filterbank is however replaced by a general filterbank (FB) which is specified by (i) the total frequency range, (ii) the number of filters in this range, (iii) the spacing of the center frequencies, and (iv) the bandwidths of the filters. For MFCCs, a common choice are (i) a frequency range of 6500 Hz with (ii) 40 filters which are (iii) spaced according to the mel-scale, where (iv) the bandwidth of the i-th filter extends from the center frequency of filter $(i-1)$ to that of filter $(i+1)$. The latter *frequency* parameters allow us to control the spectral feature resolution. In the subsequent ENS-process, we first perform an energy normalization followed by a feature-based (component-wise) quantization. The quantization basically generalizes the log-scale compression performed in the MFCC feature extraction. Normalization and quantization gives us a set of *amplitude* or *energy* parameters. Afterwards, smoothing and downsampling of the resulting feature sequence allows us to adapt the temporal resolution of the features by choosing both the smoothing window size (in ms) and the target feature resolution (in Hz) as *temporal* parameters. The general features produced after the final DCT step are called FBCC-ENS.

2.2. Voice Activity Detection

For the proposed VAD application, FBCC-ENS with 220 filters in the range of 1–10 kHz and a smoothing window of 800 ms were used. The resulting features provide a good representation of the local spectral properties of a signal. Formant regions are clearly visible, discriminating voiced speech from unvoiced speech and noise.

Because of the severe noise conditions in real acoustic channels, we suggest to apply additional postprocessing steps in the feature domain in order to enhance speech components and attenuate noise-like parts. Those steps consist of (i) highpass filtering, (ii) lowpass filtering, (iii) envelope estimation, and (iv) feature averaging. When the SNR is rather low, voiced speech regions begin to melt with enclosing noise, while fricatives such as /s/, /sh/, /ch/, /f/, or /z/ may be lost completely. As described in Subsect. 2.1, first a noise dependent quantizer tries to overcome this problem. For our postprocessing, we employ a highpass FIR filter of first order

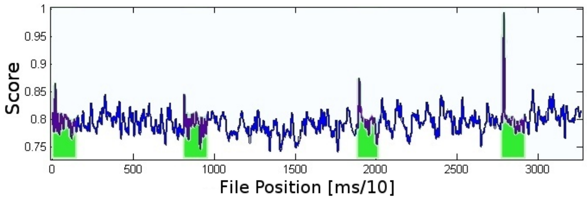


Fig. 2. Similarity function Δ obtained for a particular scenario of keyphrase matching.

as a typical form of a pre-emphasis filter. By applying this filter directly to the feature subbands, the desired effect of accentuating high frequency sounds such as unvoiced fricatives is obtained. By an additional subsequent lowpass filtering of the features using a de-emphasis filter, the influence of low frequency noise such as vehicular noise is partly rejected. The resulting energy content of frequency regions is our final discriminant parameter used in the VAD decision. For energy estimation, we proceed by first calculating the Hilbert envelopes of the filtered FBCC-ENS bands, resulting in a coarse temporal shape of the 2D feature surface. By finally taking the mean of all resulting Hilbert envelopes, inter-band variations are compensated. After median filtering the final envelope, remaining energy fluctuations are smoothed out over time. Fig. 1 illustrates the benefits of the proposed filtering stages and shows the resulting averaged temporal envelope. In the bottom-most graph, the ground truth is marked by a black line. Finally, VAD is performed by applying a thresholding procedure to the resulting mean envelope.

2.3. Keyword Spotting

Our basic approach for automatically detecting short sequences of words – in this context called *phrases* – in audio monitoring signals combines the technique of *audio matching*, known from domain of music retrieval [7] with HFCC-ENS features [6]. To this end, both the phrase (given in form of a short audio signal) and the monitoring signal are converted to feature sequences $q = (q_1, \dots, q_M)$ and $d = (d_1, \dots, d_N)$, where each of the q_i respectively d_j are feature vectors. Matching is then performed using a cross-correlation-like approach, where a similarity function $\Delta(n) := \frac{1}{M}\sum_{\ell=1}^{M}\langle q_\ell, d_{n-1+\ell}\rangle$ gives the similarity of phrase and monitoring signal at position n. Using normalized feature vectors, values of Δ in a range of $[0, 1]$ can be enforced.

Fig. 2 shows a similarity function $\Delta(n)$ obtained by an example where the German phrase *“Heute ist schönes Frühlingswetter”* was matched to a long audio recording composed of 40 phrases spoken by different speakers. Among those are four versions of the query phrase, each by a different speaker. All of them are identified as matches (indicated in green) by applying a suitable peak-picking strategy. In order to be more flexible with respect to the typical nonlinear variations in speaking tempo, in our monitoring scenario we replace the above correlation-based approach to calculate a cost function by a variant of subsequence DTW (dynamic time warping) [6].

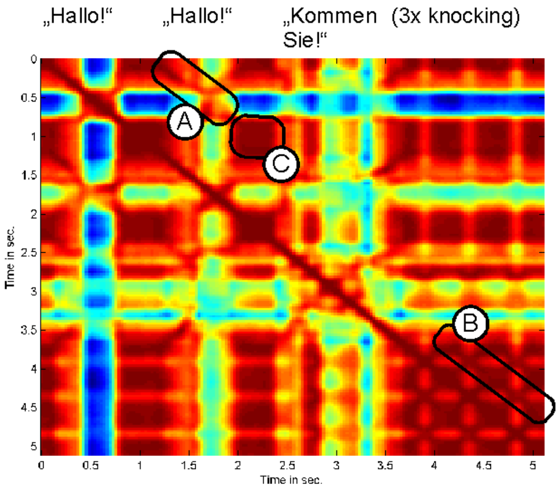


Fig. 3. Self-similarity matrix of a five second recording. Element (A) indicates a repeated cry for help, (B) a three times knocking, while rectangular regions (C) indicate silence.

Compared to classical keyword spotting [8, 9], the proposed approach is particularly beneficial when the target phrase consists of at least 3–4 syllables [6]. Advantages inherited from using the proposed HFCC-ENS features for this task are speaker and also gender independence.

2.4. Detection of Repeated Acoustic Events

To obtain the similarity of a short feature sequence q and a particular position of a longer sequence d, the similarity function Δ averages M local comparisons $\langle q_i, d_j\rangle$ of feature vectors q_i and d_j. In general, the similarity between two feature sequences $a := (a_1, \dots, a_K)$ and $b := (b_1, \dots, b_L)$ can be characterized by calculating a *similarity matrix* $S_{a,b} := (\langle a_i, b_j\rangle)_{1\le i\le K, 1\le j\le L}$ consisting of all pair-wise comparisons. To analyze the structure of an audio signal, the *self-similarity matrix* $S_a := S_{a,a}$ of the corresponding feature sequence a can be employed [10]. Fig. 3 shows a self-similarity matrix for an outdoor monitoring recording of five seconds duration. Color coding is chosen in a way such that dark red regions indicate high local similarity and blue regions correspond to low local similarity. Diagonal-like trajectories indicate the presence of *repeated* audio events within the analyzed signal. In Fig. 3, the diagonal-like elements correspond to (A) a repeated human exclamation and (B) a repeated knocking on some metal plate. Rectangular regions (C) usually indicate silence. In our monitoring context, we compute a list of all repeated events following the heuristic approach proposed in [10].

3. AN AUDIO MONITORING SYSTEM

To solve the considered acoustic localization problem, an array of three ground-located acoustic vector sensors (AVS) is used that simultaneously acquires time difference of arrival (TDOA)- as well as bearing measurements. Each AVS jointly measures the acoustic sound pressure and the acoustic particle velocity [1]. Acoustic localization can be realized by traditional triangulation, TDOA localization, or a combination of both. For triangulation, we may directly exploit the bearing measurements provided by the AVS. The intersection of the bearing lines determines the desired source location. TDOA measurements can be obtained by using two time-synchronized acoustic sensors. A single TDOA measurement can be geometrically interpreted as a hyperbola that determines the possible source locations. Using at least three TDOA sensors at dislodged locations, the corresponding hyperbolas can be intersected in order to localize the acoustic sound source. Exploiting time-sychronization of our sensors, we are able to use a combination of TDOA and triangulation-based localization.

In a first approach, the overall data processing chain including detection, classification, localization, and reporting of acoutic events is done in a cascaded fashion on multiple levels. Let S^1, S^2, and S^3 denote the three AVS. Each S^i records signals in four channels $S^i_j, j \in \{1:4\}$, with a sampling rate of currently 44.1 kHz (and 16 bit). The first three channels $S^i_j, j \in \{1:3\}$, are used for both acoustic classification and bearing of acoustic events. Each channel captures the air flow of different spatial directions, providing an azimuth- and elevation angle. The fourth channel S^i_4 essentially provides GPS timestamps, thus allowing for time synchronization. The first level of data processing operates on the signals S^i_j. Let $S^i_{j,c} := \{(t,\kappa)\} := \{(t,\kappa)_{t,\kappa}\}$ denote a set of recognized acoustic events of class c at time t with confidence κ. Furthermore, let $\Omega^i := \left\{(t, \omega^i_a, \omega^i_e, \kappa)\right\}$ denote a set of bearing results, where ω^i_a and ω^i_e refers to an azimuth- and elevation angle, respectively, measured at time t with confidence κ at S^i. Still operating on the data streams, on the second level all three individual level-1 classification results $S^i_{j,c}, j \in \{1:3\}$, are merged into a single overall classification result $S^i_c := \{(t,\kappa)\}$. On the third level, all single overall classification results are merged into the set $S^{\{1,2,3\}}_c := \{(t,\kappa)\}$ of compound classification results. Individual level-1 bearing results are merged into the set $S^{\{1,2,3\}}_p := \{(t,(p_{lat},p_{lon}),\kappa)\}$, denoting crossing bearing, where (p_{lat},p_{lon}) refers to a GPS coordinate. Finally, the level-3 classification and localization results are merged into a set $S^{\{1,2,3\}} := \{(t, c, \kappa_c(p_{lat},p_{lon}), \kappa_p)\}$ of detected, classified, and localized acoustic events.

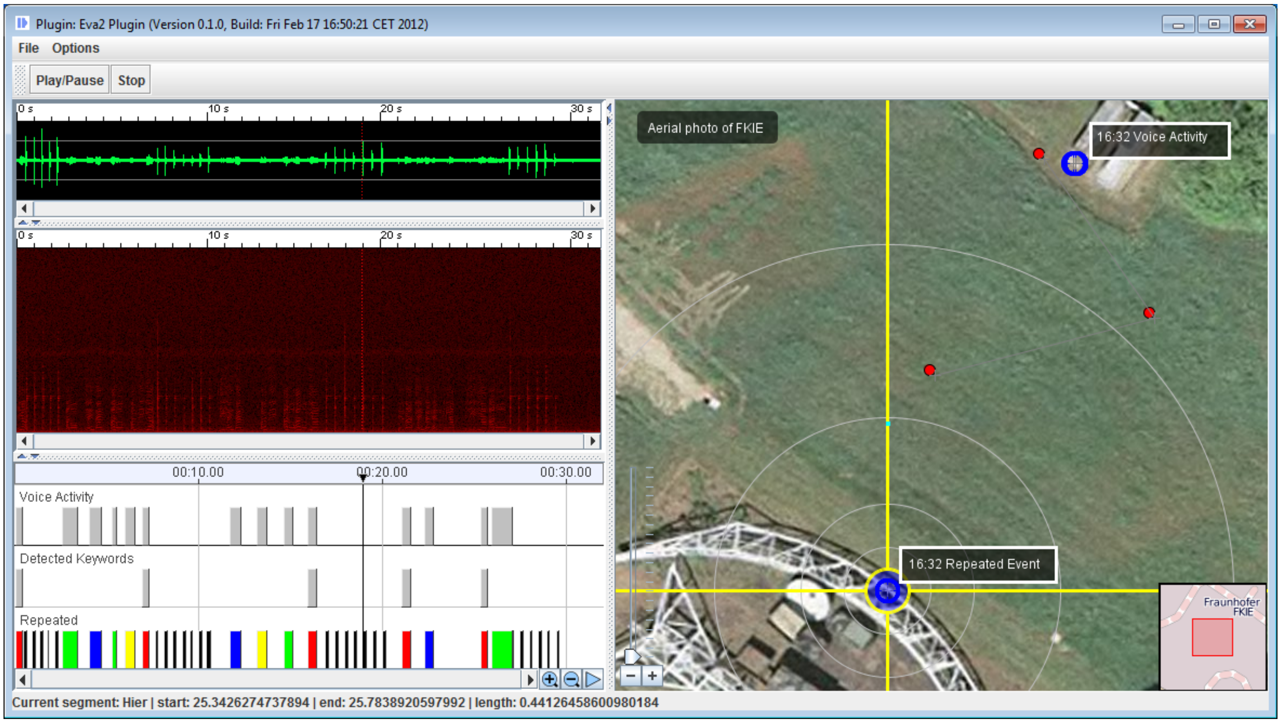
To present the detection and localization results to the user in an intuitive and easy-to-browse way, a multimodal user interface has been developed. Fig. 4 shows an overview of the main components for user interaction. The system basically acts as an interactive audio player where selected audio streams recorded by the sensors may be played back. On the left side, the system displays the signal and spectrogram of the selected audio channel (top left) and a timetable-like overview of the detected audio events (bottom left). In this example, voice activity, detected keywords (here cries for help shouted in German), and repeated audio events are displayed. Repeated instances of the same sounds are displayed in the same color; in this example the repeated sound of knocking on a metal plate is shown in black color, repeated exclamations of the same sequences of words are shown in red, green, blue, and yellow colors respectively. All of the visualizations are synchronized, indicated by a sliding cursor that moves during playback. By clicking on the detected audio events, playback can be continued from the corresponding temporal position. Localization information is presented in the right part of the user interface, showing an areal map of the monitored area, where the positions of the acoustic sensors are indicated by red bullets. Detected and localized events are indicated by small blue crosshairs during playback. By using short intervals of display time, temporal trajectories like detected footsteps can be visualized as well. Additionally, small textboxes indicate labels specifying individual events. The whole user interface works for real-time playback as well as in an offline-mode, the latter meaning that detected events can be selected in the event-list and localization information visualized in the map without playback being active.

4. A CASE STUDY

In a case study, the proposed system was set up on the grounds of Fraunhofer FKIE, see Fig. 4 for an areal map including sensor positions. In an experiment, different persons located at different distances to the sensor network were producing different types of verbal exclamations as well as various types of knocking sounds. Evaluations show that VAD, keyword spotting, and detection of repeated sounds work very robustly. However, the limited sensor range due to filter effects caused by the windshields will have to be considered next. Event localization is also possible for strong, isolated events. Separation and detection of simultaneous events was not possible in the current setup and will be subject of a subsequent study. In a second experiment, two persons walking in opposite directions along different surfaces, while talking to each other, were recorded. While separate experiments under controlled conditions were already performed for calibrating sensor equipment and estimating noise levels, further, more comprehensive and formal evaluations have to be performed.

5. CONCLUSIONS AND FUTURE WORK

A main focus of this paper was on devising robust audio features and adapting unsupervised audio retrieval techniques to the scenario of acoustical monitoring. Secondly, a system



Current segment: Hier | start: 25.3426274737894 | end: 25.7838920597992 | length: 0.44126458600980184

Fig. 4. User Interface of the proposed monitoring system. Selected audio channel along with spectrogram (top left), extracted audio events (bottom left), and area map showing localization information of detected events (right).

for multimodal presentation and navigation of monitoring results was presented. First case studies show how the proposed system benefits from the suitable integration of detection and localization results obtained from data recorded by an innovative acoustic sensor. Besides more comprehensive evaluations, future work has to deal with mechanisms for more systematically fusing the outputs of the different detectors. In this context, another issue will be how to combine results obtained from a larger, spatially distributed sensor network.

6. REFERENCES

[1] H.-E. de Bree et al., "The μ-flown: the microflown: a novel device measuring acoustical flows," *Sensors and Actuators A*, vol. 54, pp. 552–557, 1996.

[2] S.G. Iyengar, P.K. Varshney, and T. Damarla, "On the detection of footsteps based on acoustic and seismic sensing," in *in Proc. ASSC*, 2007, pp. 2248 – 2252.

[3] Taras Butko, Cristian Canton-Ferrer, Carlos Segura, Xavier Giró, Climent Nadeu, Javier Hernando, and Josep R. Casas, "Acoustic Event Detection Based on Feature-Level Fusion of Audio and Video Modalities," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. Article ID 485738, pp. 11 pages, 2011.

[4] Yukang Guo and Mike Hazas, "Localising speech, footsteps and other sounds using resource-constrained devices," in *Proc. IPSN*, May 2011, pp. 330 – 341.

[5] Mark D. Skowronski and John G. Harris, "Exploiting independent filter bandwidth of human factor cepstral coefficients in automatic speech recognition," *JASA*, vol. 116, no. 3, pp. 1774–1780, 2004.

[6] Dirk von Zeddelmann, Frank Kurth, and Meinard Müller, "Perceptual Audio Features for Unsupervised Key-Phrase Detection," in *Proc. IEEE ICASSP*, Dallas, TX, USA, Mar. 2010.

[7] M. Müller, F. Kurth, and M. Clausen, "Audio matching via chroma-based statistical features," in *Proc. ISMIR, London, GB*, 2005.

[8] J.G. Wilpon, L.R. Rabiner, C.-H. Lee, and E.R. Goldman, "Automatic recognition of keywords in unconstrained speech using hidden Markov models," *IEEE TASSP*, vol. 38, no. 11, pp. 1870–1878, 1990.

[9] Joseph Keshet, David Grangier, and Samy Bengio, "Discriminative keyword spotting," *Speech Communication*, vol. 51, pp. 317–329, 2009.

[10] M. Müller and F. Kurth, "Towards Structural Analysis of Audio Recordings in the Presence of Musical Variations," *EURASIP Journal on Applied Signal Processing*, vol. 2007, no. Article ID 89686, pp. 18 pages, 2007.