EURASIP Journal on Audio,
Speech, and Music Processing

EDITORIAL

Open Access

Signal processing and machine learning for speech and audio in acoustic sensor networks

Walter Kellermann[1*], Rainer Martin[2] and Nobutaka Ono[3]

Nowadays, we are surrounded by a plethora of recording devices, including mobile phones, laptops, tablets, smartwatches, and camcorders, among others. However, conventional multichannel signal processing methods can usually not be applied to jointly process the signals recorded by multiple distributed devices because synchronous recording is essential. Thus, commercially available microphone array processing is currently limited to a single device where all microphones are mounted. The full exploitation of the spatial diversity offered by multiple audio devices without requiring wired networking is a major challenge, whose potential practical and commercial benefits prompted significant research efforts over the past decade.

Wireless acoustic sensor networks (WASNs) have become a new paradigm of acoustic sensing to overcome the limitations of individual devices. Along with wireless communications between microphone nodes and addressing new challenges in handling asynchronous channels, unknown microphone positions, and distributed computing, the WASN enables us to spatially distribute many recording devices. These may cover a wider area and utilize the nodes to form an extended microphone array. It promises to significantly improve the performance of various audio tasks such as speech enhancement, speech recognition, diarization, scene analysis, and anomalous acoustic event detection.

For this special issue, six papers were accepted which all address the above-mentioned fundamental challenges when using WASNs: First, the question of which sensors should be used for a specific signal processing task or extraction of a target source is addressed by the papers of Guenther et al. and Kindt et al. Given a set of sensors, a method for its synchronization on waveform level in dynamic scenarios is presented by Chinaev et al., and a localization method using both sensor signals and higher-level environmental information is discussed by Grinstein et al. Finally, robust speaker counting and source separation are addressed by Hsu and Bai and the task of removing specific interference from a single sensor signal is tackled by Kawamura et al.

The paper ‘Microphone utility estimation in acoustic sensor networks using single-channel signal features’ by Guenther et al. proposes a method to assess the utility of individual sensors of a WASN for coherence-based signal processing, e.g., beamforming or blind source separation, by using appropriate single-channel signal features as proxies for waveforms. Thereby, the need for transmitting waveforms for identifying suitable sensors for a synchronized cluster of sensors is avoided and the required amount of transmitted data can be reduced by several orders of magnitude. It is shown that both estimation-theoretic processing of single-channel features and deep learning-based identification of such features lead to measures of coherence in the feature space that reflect the suitability of distributed sensors for coherent processing.

*Correspondence:
Walter Kellermann
walter.kellermann@fau.de
[1] Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
[2] Ruhr-Universität Bochum, Bochum, Germany
[3] Tokyo Metropolitan University, Hino-shi, Japan

In the paper 'Robustness of ad hoc microphone clustering using speaker embeddings: Evaluation under realistic and challenging scenarios' by Kindt et al., the robustness of speaker embeddings learned from multiple microphone signals as a feature for identifying useful clusters for extracting a speech signal is studied with respect to several key aspects: The dependency on the distance metrics for clustering, the observation interval required for establishing robust clusters determining the stationarity requirements for the acoustic scenario, and the performance for increasingly challenging acoustic scenarios and multiple speakers. For evaluation, a source separation task in realistic noisy and reverberant environments is investigated using several separation techniques applied for the resulting clusters. The proposed speaker embeddings are also compared to established MFCC-based features with respect to multiple state-of-the-art criteria for signal enhancement.

The paper 'Online distributed waveform-synchronization for acoustic sensor networks with dynamic topology' by Chinaev et al. is dedicated to network-wide sample-level synchronization relying on a previously published acoustic waveform-based sampling rate-offset estimation and compensation for pairs of nodes. Assuming that the WASN is organized as a directed minimum spanning tree (MST), the proposed network-wide synchronization scheme propagates from a root node over the entire network. Additionally, a network protocol is proposed that ensures the synchronization even if the network topology changes, e.g., because of node failure, broken transmission links, or newly appearing nodes. The efficacy of the method is demonstrated for dynamic scenes in a simulated dynamic acoustic scenario in an apartment with several rooms.

In their paper 'Dual input neural networks for positional sound source localization', Grinstein et al. combine multiple microphone signals from a distributed microphone array with information describing the acoustic properties of the scene for improved sound source localization. This information includes the positions of microphones, the room size, and the reverberation time. They present a Dual Input Neural Network (DI-NN) as a straightforward and efficient technique to construct a neural network capable of processing two distinct data types. It is tested in different scenarios, comparing it to alternative models such as a traditional least-squares method and a convolutional recurrent neural network. Although the proposed DI-NN is not retraining for each new scenario, the authors' results demonstrate the superiority of the proposed DI-NN, achieving a substantial reduction in localization errors on synthetic data and a data set with real recordings.

The paper 'Learning-based robust speaker counting and separation with the aid of spatial coherence' by Hsu and Bai tackles speaker counting and speaker separation in noisy and reverberant environments. The authors combine traditional and learning-based methods to enhance these tasks and to achieve robustness to unseen room impulse responses (RIRs) and array configurations. They formulate a three-stage approach that entails the computation of a spatial coherence matrix (SCM) based on whitened relative transfer functions (wRTFs) as a spatial signature of directional sources. They evaluate the SCM and local coherence functions to detect the activity of the target speaker. Then, the eigenvalues of the SCM and the maximum similarity of inter-frame global activity distributions between two speakers are fed into a network for speaker counting (SCnet). To extract each independent speaker signal, a global and local activity-driven network (GLADnet) is employed. The authors demonstrate the benefits of the proposed approach on a data set of real meeting recordings.

The last paper, entitled 'Acoustic object canceller: removing a known signal from monaural recording using blind synchronization' by Kawamura et al., addresses the problem of removing undesired interference from individual microphone signals if a reference signal for the interference is available. The authors propose a method that treats the interference as an acoustic object whose signal is linearly filtered before arriving at the receiving microphone. Assuming that the signals of the acoustic object and the microphone exhibit different sampling rates, the signals are first synchronized, and then the frequency response of the propagation path from the object to the microphone is determined via maximum-likelihood estimation using the majorization-minimization algorithm, investigating and evaluating various statistical models for the desired signal that should be preserved.

We like to thank all authors for their excellent contributions to this special issue and hope that this collection will be a useful resource for research in WASNs in the years to come.

Published online: 17 December 2023

Publisher's Note