Int. J. Electron. Commun. (AEÜ) 69 (2015) 1715–1723

Contents lists available at ScienceDirect

International Journal of Electronics and Communications (AEÜ)

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

Microphone array based classification for security monitoring in unstructured environments

Simone Scardapane[a,*], Michele Scarpiniti[a], Marta Bucciarelli[a], Fabiola Colone[a], Marcello Vincenzo Mansueto[b], Raffaele Parisi[a]

[a] Department of Information Engineering, Electronics and Telecommunications (DIET), "Sapienza" University of Rome, Via Eudossiana 18, 00184 Rome, Italy
[b] Intecs S.p.A., Salita del Poggio Laurentino 7, 00144 Rome, Italy

ARTICLE INFO

Article history:
Received 27 October 2014
Accepted 11 August 2015

Keywords:
Security monitoring
Beamforming
Computational scene analysis
Classification
Multi-level learning

ABSTRACT

The aim of this paper is to describe a novel security system able to localize and classify audio sources in an outdoor environment. Its primary intended use is for security monitoring in severe scenarios, and it has been designed to cope with a large set of heterogeneous objects, including weapons, human speakers and vehicles. The system is the result of a research project sponsored by the Italian Ministry of Defense. It is composed of a large squared array of 864 microphones arranged in a rectangular lattice, whose input is processed using a classical delay-and-sum beamformer. The result of this localization process is elaborated by a complex multi-level classification system designed in a modular fashion. In this paper, after presenting the details of the system's design, with a particular emphasis on the innovative aspects that are introduced with respect to the state-of-the-art, we provide an extensive set of simulations showing the effectiveness of the proposed architecture. We conclude by describing the current limits of the system, and the projected further developments.

1. Introduction

In the last decade, the wide availability of cheap sensor instrumentation has made automatic surveillance an economical and technical possibility. A notable example in this sense is the Secure Perimeter Awareness Network (SPAN) in use at the J.F.K. International Airport [1], an integrated system of sensors which is used, between others, for automatic intrusion detection in the perimeter of the airport. Of particular interest for their flexibility and cheapness are the systems based on acoustic sensors [2]. When we consider generic outdoor scenarios, an equivalent automatic security monitoring system based on a microphone array would be an invaluable tool in assessing and controlling any type of situation occurring in them [3]. This includes, but is not limited to, handling large civil events, or increasing the awareness of a terrain in military contexts. Moreover, a sensor-based system possesses an intrinsic degree of security, being by design a completely passive device. However, implementing automatic outdoor security systems able to work with noisy, realistic and diversified data in an unstructured environment is a challenging and largely unexplored area. This is particularly due to noise, air distortion, low signal-to-noise ratio and presence of multiple, possibly conflicting sources.

In this context, last year the Italian Ministry of Defense funded the SMART-OPTIGRID project, carried on by Intecs S.p.A. in collaboration with the DIET Dept. of "Sapienza" University of Rome. The SMART-OPTIGRID project is aimed at a feasibility study of a microphone array based acoustic antenna for the detection, localization and classification of sound sources in severe outdoor scenarios. In particular, we are interested in classifying a set of sources acquired from a large set of sensors, and locate the concurrent presence of weapons, vehicles and/or spoken sources. Therefore the conceived system should be highly reliable, reasonably limited in size so that it can be moved if necessary, and extremely adaptive to different operative conditions. To this purpose particular attention has been devoted to the design of the microphone array geometry, to the definition of proper detection and localization strategies, and to the development of innovative classification techniques to be effective in the considered scenarios.

From an algorithmic point of view, the first innovative aspect of our system is the investigation of an array composed of a large number of microphones, which are integrated into a small surface of limited size. The second innovative aspect, instead, is the

* Corresponding author. Tel.: +39 06 44585495; fax: +39 06 4873300.
E-mail addresses: simone.scardapane@uniroma1.it (S. Scardapane), michele.scarpiniti@uniroma1.it (M. Scarpiniti), marta.bucciarelli@uniroma1.it (M. Bucciarelli), fabiola.colone@uniroma1.it (F. Colone), marcello.mansueto@intecs.it (M.V. Mansueto), raffaele.parisi@uniroma1.it (R. Parisi).

http://dx.doi.org/10.1016/j.aeue.2015.08.007
1434-8411/© 2015 Elsevier GmbH. All rights reserved.

introduction of a modular four-level classification stage, designed to cope with the large number of possible sources that can be present in an environment of interest.

Regarding the state-of-the-art, automatic security monitoring by means of collected audio signals falls under the broader field of *Computational Auditory Scene Analysis* (CASA) [4], whose aim is to successfully analyze a stream of continuous audio to identify and isolate the sources of interest contained in it. The audio can be acquired either (i) using large acoustic antennas [2,5,6] (which is our design choice), or (ii) using distributed sensors (e.g., [7]). The subsequent analysis is typically performed by applying state-of-the-art machine learning techniques [8] to recognize the presence of specific objects. This last problem is a notable example of *Automatic Audio Classification* (AAC) [9], the task of automatically labeling a given audio signal in a set of predefined classes. Generally an AAC system works by subdividing the audio signal in small, overlapping frames, extracting some statistical features, and finally classifying them using a standard machine learning tool. Due to the reasons detailed above, AAC has been studied primarily in the context of single-level applications, where the classes are restricted to a very specific domain. For example, there exists a vast literature regarding speech discrimination [10–13], vehicle recognition [14–16] and weapon classification [17,18,7]. In addition, due to the maturity of the field there exist several commercial and open-source products that perform these tasks, such as the Halo system[1] and the Sphinx toolkit.[2]

If we consider a system with the need of performing more than one of the aforementioned tasks, however, their combination is not as straightforward as it may appear. A complex, realistic classification system has the need of being highly modular, flexible, and hierarchical, topics that were only marginally considered in the learning literature until the last decade [19]. Regarding AAC for security monitoring, a small number of systems were proposed recently that undertake this direction, particularly by first separating the speech detection problem from the non-speech detection. Atrey et al. [20] presented a four-level system for event detection. However, they used a single sensor to gather information, and considered only binary classification tasks. In our work, instead, we consider an array composed of a large number of microphones, and we are interested in classifying a wide range of possible sources of interest. Abu-El-Quran [21] and Zhao et al. [22] detail two systems that, starting from a microphone array, perform at the same time speech and non-speech recognition. Although their works bear some resemblance to the system we detail in this paper, they were primarily meant for use in an indoor application, and the non-speech classification was performed in a single step. An early application of the idea of multi-stage classification to an audio stream of data is described instead in [23].

In this paper, we detail the steps taken to design the various components and select the appropriate learning tools. Moreover, we show how we explicitly take into account the presence of air distortion by the use of virtual examples [24], to make our system robust to them. The result of this is a modular and flexible classification system that efficiently combines several small classifiers by virtue of its own structure. Some preliminary simulation results show the effectiveness of the proposed architecture.

The rest of the paper is organized as follows. In Section 2 we describe the general architecture of our system. Then, in Sections 3 and 4 we go into more detail with respect to the beamforming and classification operations, respectively. A brief analysis of the computational cost of the proposed architecture is given in Section 5. Some empirical evaluations are presented in Section 6 and finally Section 7 concludes the paper.

2. General system architecture

The surveillance system described in this paper is based on a microphone array acoustic antenna to be employed for detecting, localizing and classifying heterogeneous sources in severe outdoor scenarios.

The desired system specifications include the capability to operate in a wide search volume that spans the angular interval [-45°, 45°] in both the azimuth and the elevation directions. Moreover, this should be accomplished by using narrow listening beams with -3 dB aperture of few degrees. Both impulsive and continuous acoustic emissions should be taken into account with a coverage that might reach $8 \div 10$ km for high power sources. In particular the considered sources should include vehicles, aircrafts, weapons and spoken sources.

The conceived system architecture is described schematically in Fig. 1. The input of the system is provided by a square array of 864 microphones mounted in a triangular lattice. The microphones are considered to be omnidirectional and have a flat frequency response in the acoustic band. The design choices pertaining the acoustic array sub-system are detailed in the following section.

A set of properly steered listening beams are extracted from the microphones' raw output by performing a standard delay-and-sum beamforming operation [25,26]. At this stage, particular attention has been devoted to the design of appropriate strategies able to guarantee the required angular coverage even in the presence of sources with an impulsive nature (see Section 3). Once the active sources have been detected and localized at some listening beams, the corresponding signals are fed to the input of the processing stages responsible for the sources classification. This represents a very challenging task owing to the need to classify highly heterogeneous sources, possibly active at the same time in the environment.

To this purpose each of the designated beams' outputs is split in small audio frames, from which a collection of 42 statistical descriptors are extracted. The features are then passed as input to a hierarchical classification system that categorize them on each of the classes that we considered. The overall classifier is composed of four, modular stages requiring smaller binary classifiers. This allows to subdivide the original learning tasks into a series of smaller and simpler tasks that are efficiently solved by each of the classifier's modules. The detailed description of the innovative classification technique proposed in this paper are reported in Sections 4 and 6.

3. Beamforming

In its resting position, the microphone array consists of N_{mic} microphones arranged on a planar surface (2D array). As it is well known, by jointly processing the signals collected at the available microphones it is possible to synthesize a listening beam steered towards the selected search angular quantum. The electronic steering is obtained by summing the received signals after proper compensation of the different delays induced at each microphone, so that the contributions from a given direction are coherently integrated (so-called *delay-and-sum beamforming* [27,26]). This results in a significant improvement with respect to the single microphone since the listening beam shows a much higher gain and increased angular discrimination capability due to the extremely narrower beam-width.

With reference to the SMART-OPTIGRID project, aiming at limiting the final array dimensions, the beamformer has been designed to guarantee a listening beam with -3 dB aperture of

[1] http://www.roke.co.uk/halo/

[2] http://cmusphinx.sourceforge.net/