Expert Systems With Applications 117 (2019) 29–41

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation

Pierre Laffitte[a,*], Yun Wang[b], David Sodoyer[a], Laurent Girin[c]

[a] Univ Lille Nord de France, F-59000 Lille, IFSTTAR, COSYS, LEOST, Villeneuve Ascq, F-59650, France
[b] Carnegie Mellon Department of Computer Science, Language Technologies Institute, Pittsburgh, PA, USA
[c] Grenoble-INP, Gipsa-lab, Grenoble, France

ARTICLE INFO

Article history:
Received 28 March 2018
Revised 31 July 2018
Accepted 29 August 2018
Available online 20 September 2018

Keywords:
Audio surveillance
Acoustic event detection
Transportation
Classification
Neural networks

ABSTRACT

As intelligent transportation systems are becoming more and more prevalent, the relevance of automatic surveillance systems grows larger. While such systems rely heavily on video signals, other types of signals can be used as well to monitor the security of passengers. The present article proposes an audio-based intelligent system for surveillance in public transportation, investigating the use of some state-of-the-art artificial intelligence methods for the automatic detection of screams and shouts. We present test results produced on a database of sounds occurring in subway trains in real working conditions, by classifying sounds into screams, shouts and other categories using different Neural Network architectures. The relevance of these architectures in the analysis of audio signals is analyzed. We report encouraging results, given the difficulty of the task, especially when a high level of surrounding noise is present.

1. Introduction

The proposed study is an attempt to build an intelligent surveillance system capable of automatically detecting abnormal situations in public transportation environments, such as underground subways or metros, based on the analysis of audio signals. Recently, Neural Networks have rose to prominence in most intelligent and expert systems, in applications as varied as image classification, customer behavior prediction, or medical diagnosis (Affonso, Rossi, Vieira, & de Leon Ferreira de Carvalho, 2017; Eshtay, Faris, & Obeid, 2018; Vanneschi, Horn, Castelli, & Popovič, 2018). In the literature, intelligent systems for automatic surveillance generally use video signals (Baran, Rusc, & Fornalski, 2016; Foggia, Petkov, Saggese, Strisciuglio, & Vento, 2016; Hata, Kuwahara, Nozawa, Schwenke, & Vetro, 2005; Orhan Bulan, 2013; Velastin, Boghossian, & Vicencio-Silva, 2006), but the use of audio can be an interesting complement as it helps circumvent issues inherent to video signals such as vision field obstruction or lighting changes. In this paper, we focus on a particular context for surveillance systems; namely, that of public transportation. While most research in this specific application also use video signals (He, Chen, Jiang, Lu, & Yuan, 2017; Orhan Bulan, 2013; Velastin et al., 2006), we tackle the issue from a different angle by using audio signals instead. This approach has been introduced within the framework of public transportation surveillance; in (Rouas, Louradour, & Ambellouis, 2006) to detect screams, in (Ganansia et al., 2011) to detect and localize shouts and graffiti sprays, in (Zouaoui et al., 2015) to detect abnormal sounds, and within the framework of general surveillance systems in (Valenzise, Gerosa, Tagliasacchi, Antonacci, & Sarti, 2007) to detect screams and gunshot sounds. The contribution of this paper is mostly experimental and applicative, not methodological. It shows that using state of the art neural networks can serve the purpose of audio surveillance in public transportation, and reveals some interesting characteristics of those methods, as well as how a good understanding of them can lead to an improvement in performances.

From a more general perspective, the task we address is referred to as acoustic event detection (AED), which is a research topic of growing interest in the audio signal processing community (Chu, Narayanan, & Kuo, 2009; Crocco, Cristani, Trucco, & Murino, 2016; Dennis, Tran, & Chang, 2013; Diment, Cakir, Heittola, & Virtanen, 2015; Fernández-Delgado, Cernadas, Barro, & Amorim, 2014; McLoughlin, Zhang, Xie, Song, & Xiao, 2015). The DCASE challenge (Virtanen et al., 2016) attests to the popularity of this task, which has many applications ranging from smart houses involving automatic systems for domestic events detection using audio and video data streams (Wang, Lin, Chen, & Tsai, 2014; Wu, Gong, Chen, Zhong, & Xu, 2009), to humanoid robotics where an

* Principal Corresponding Author.
E-mail addresses: pl.laffitte@gmail.com (P. Laffitte), yun.wang@cs-cmu.edu (Y. Wang), david.sodoyer@ifsttar.fr (D. Sodoyer), laurent.girin@gipsa-lab.grenoble-inp.fr (L. Girin).

https://doi.org/10.1016/j.eswa.2018.08.052
0957-4174/

audition model is a prerequisite for natural human-robot interaction (Janvier, Alameda-Pineda, Girin, & Horaud, 2012; Nakadai, Matsuura, Okuno, & Tsujino, 2004; Noda, Yamaguchi, Nakadai, Okuno, & Ogata, 2015; Wu et al., 2009), including automatic surveillance applications (Foggia et al., 2016; Velastin et al., 2006). AED can be thought of as a combination of automatic audio segmentation and audio event classification (Janvier, Alameda-Pineda, Girin, & Horaud, 2014), hence adding to the classification task the difficulty of identifying the temporal location of the audio events (Phan, Maas, Mazur, & Mertins, 2015), as it does not rely on prior segmentation of the data. Audio event classification techniques in the state of the art are diverse, with many different combinations of features and classifiers: Mel-Frequency Cepstral Coefficients (MFCCs) classified with Gaussian Mixture Models (GMMs) (Pohjalainen, Raitio, & Alku, 2011), with Support Vector Machines (SVMs) (Huang, Chiew, Li, Kok, & Biswas, 2010; Lei & Mak, 2014; Wu et al., 2009), with Hidden Markov Models (HMMs) (Ntalampiras, Potamitis, & Fakotakis, 2009); MFCCs and other spectral features classified with GMMs (Chu et al., 2009; Gerosa, Valenzise, Tagliasacchi, Antonacci, & Sarti, 2007), with the *k*-nearest-neighbors algorithm (kNNs) (Chu et al., 2009), and more recently with random forests (RFs) (Phan et al., 2015); Gabor features classified with GMMs (Geiger & Helwani, 2015; Gerosa et al., 2007) and with SVMs (Wang et al., 2014); all-pole group delay features classified with Deep Neural Networks (DNNs) (Diment et al., 2015); Gammatone-Wavelet features classified with SVMs (Valero & AlÃas, 2012); spectrogram image features classified with kNNs (Dennis et al., 2013), SVMs and DNNs (Diment et al., 2015; McLoughlin et al., 2015; Wei, Li, Pham, Das, & Qu, 2016); MFCCs and deep scattering features classified with RFs and the k-means algorithm (Salamon & Bello, 2015). However, the use of Neural Networks is becoming more and more prominent, be it Deep (Sharan & Moir, 2017), Convolutional (Hershey et al., 2017; Lee, Kim, Park, & Nam, 2017) or Recurrent Neural Networks (RNNs) (Wang & Metze, 2017). The task defined in this study is to detect violent events in the subway via automatic detection of screams and shouts emitted by the people involved in the events. Such events span different cases, such as people in physical difficulty, people quarreling, panic situations, calls for help, etc. Shouts and screams are here defined as loud vocal sounds with and without explicit semantic content, respectively. Since it turns out that scream occurrences are outnumbered by shout occurrences in our database, in the following we employ the general term “shout” to characterize the overall set of abnormally loud sounds generated by people subject to or witnesses of violent events. Although such alert signals are quite specific, this task remains challenging since there generally exists a large variability between different realizations of screams and shouts, depending on the causing events, a large variability of “speakers”, number of persons involved, their emotional state, etc. The first specific aspect of the present work is the rarity of the data: in order to design a realistic AED system, a dedicated database was recorded, consisting of real signals recorded in the Paris subway (called ’Metropolitain’, or simply Metro). A whole Metro train was booked for the recording sessions, thanks to the Paris public transportation authority (the RATP) being a partner of the research project which frames this study. Abnormal situations were enacted by actors, including many extra participants representing the crowd of passengers. As a consequence all recordings used in the present study are real and not derived from synthetic signals or simulated acoustic mixes, and the size of the corpus cannot match that of handcrafted synthetic data such as in (Lafay, Lagrange, Rossignol, Benetos, & Roebel, 2016) and (Wang & Metze, 2017). The second specific aspect concerns the characteristics of this environment which is very noisy and variable. It contains many objects that can act as sound sources and filters, shaping the acoustic scene; noise from the vehicle itself (e.g., motor noise, boogie-rails frictions), noise coming from the surrounding environment (e.g., railway traffic, station noise, loud-speaker announcements), and noise produced by passengers. Within such an environment, the choice of target classes used to define the acoustic landscape is crucial, especially within the framework of audio event detection. The classification techniques used here are different architectures of neural networks applied on acoustic MFCC features, namely (feedforward) deep neural networks (DNNs), convolutional neural networks (CNNs) and recurrent neural networks (RNNs; in particular we used Long Short Term Memory (LSTM) cells). We set the main task as a 3-way classification task, with target classes defined as shouts (comprising shouts and screams), conversational speech and background noise. Besides this main task, we created another set of 14 classes (by dividing each of the previous 3 classes into 5 smaller sub-classes) on which we performed a 14-way classification. We report results of an extensive benchmark made using the 3 types of neural networks, for both the 3-class problem and the 14-class problem. The remaining of this paper is organized as follows: Section 2.1 gives a detailed description of the database used for the experiments. Section 3 presents the methodological background of neural networks. Section 4 presents the parameters and settings of the experiments we carried out, while Section 5 reports the results. Finally, Section 6 draws some conclusions and perspectives.

2. Database and analysis of environment impact

As stated in the introduction, the present task is to detect dangerous situations occurring in the metro by analyzing the audio environment. In an effort to account for the likelihood that the situation exhibits a potentially dangerous/violent aspect, we devised the following three classes;

- **Shouts** (includes screams and all overlapping background sound),
- **Speech** (includes all overlapping background sound),
- **Background sounds** (all sounds not pertaining to the previous two classes; no speech nor shout sounds are assumed to be present).

Additionally we define some more classes to describe the acoustic environment related to the metro’s trajectory during its course:

- **Stand-by** (acoustic scene when the train is idle, in the station),
- **Compressor** (noise from compressor, very specific),
- **Departure** (acoustic scene during departure, when speed increases from zero to full-speed),
- **Cruise** (acoustic scene during the period of time when the train is at full speed),
- **Arrival** (acoustic scene during arrival in station, when speed decreases to zero).

Those classes will be used in order to help classify the main three classes.

2.1. Data acquisition

Within the framework of research project DéGIV (Zouaoui et al., 2015), a subway train from the automatic line 14 of the Paris Metropolitain was reserved for the recording sessions. An Omnidirectional microphone (C224 6v cell from ELNO brand) was placed on the ceiling of the metro car, and a low-latency JACK server audio interface, made specifically for this project, was used to record the signal, producing 16-bits/48-kHz PCM signals. Several sessions took place between 10 am and 4 pm while the train was running its usual course, among other trains from the same subway line, running between different stations and stopping at all of them.