Expert Systems with Applications 42 (2015) 5668–5683

Contents lists available at ScienceDirect

Expert Systems with Applications

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

An integrated system for voice command recognition and emergency detection based on audio signals

Emanuele Principi *, Stefano Squartini, Roberto Bonfigli, Giacomo Ferroni, Francesco Piazza

Università Politecnica delle Marche, Via Brecce Bianche, 60131 Ancona, Italy

ARTICLE INFO

Article history:
Available online 6 March 2015

Keywords:
Automatic speech recognition
Novelty detection
Emergency detection
Ambient assisted living

ABSTRACT

The recent reports on population ageing in the most advanced countries are driving governments and the scientific community to focus on technologies for providing assistance to people in their own homes. Particular attention has been devoted to solutions based on acoustic signals since they provide a convenient way to monitor people activities and they enable hands-free human–machine interfaces. In this context, this paper presents a complete solution for voice command recognition and emergency detection based on audio signals entirely integrated in a low-consuming embedded platform. The system combines an active operation mode were distress calls are captured and a vocal interface is enabled for controlling the home automation subsystem, and a pro-active mode, were a novelty detection algorithm detects abnormal acoustic events to alert the user of a possible emergency. In the first operation mode, a Voice Activity Detector captures voice segments of the audio signal, and a speech recogniser detects commands and distress calls. In the pro-active mode, an acoustic novelty detector is employed in order to be able to deal with unknown sounds, thus not requiring an explicit modelling of emergency sounds. In addition, the system integrates a VoIP infrastructure so that emergencies can be communicated to relatives or care centres. The monitoring unit is equipped with multiple microphones and it is connected to the home local area network to communicate with the home automation subsystem. The algorithms have been implemented in a low-consuming embedded platform based on a ARM Cortex-A8 CPU. The effectiveness of the adopted algorithms has been tested on two different databases: ITAAL and A3Novelty. The obtained results show that the adopted solutions are suitable for speech and audio event monitoring in a realistic scenario.

1. Introduction

In the recent years, the interest in smart homes and ambient intelligence in general has been constantly increasing. Indeed, according to independent studies (Markets & Markets, 2013; Sorrell, 2014), the smart home market will significantly grow by 2018, with estimated revenues that will more than double by that date. Generally, a smart home should provide the technologies for improving the comfort of its inhabitants, however the recent estimates on population ageing in the most advanced countries (Giannakouris, 2008, 2010) drove the scientific community to concentrate its efforts on in-home healthcare technologies (Acampora, Cook, Rashidi, & Vasilakos, 2013). Such studies are also encouraged by governments, which are aware that social healthcare systems will face serious challenges as the demand for health services increases.

A closer look to in-home healthcare technologies points out that they can be distinguished based on their application scenario. Recalling the classification proposed in Acampora et al. (2013), they can be divided in monitoring applications (e.g., for vital parameters, inhabitants behaviour, or emergency detection), ambient assisted living (i.e., for supporting people in their everyday lives), therapy and rehabilitation, persuasive and emotional well being, and smart hospitals. A common approach in all the scenarios is to adopt one or more sensor technology to capture ambient and/or vital parameters, intelligently process the acquired signals to extract semantically meaningful information and exploit it in reasoning algorithms to make appropriate decisions. A communication infrastructure is also present to allow an information exchange among the entities inside the home (e.g., sensors, processing devices, etc.), and outside it (e.g., remote healthcare services, a relative, etc.).

* Corresponding author.
E-mail addresses: e.principi@univpm.it (E. Principi), s.squartini@univpm.it (S. Squartini), r.bonfigli@univpm.it (R. Bonfigli), g.ferroni@univpm.it (G. Ferroni), f.piazza@univpm.it (F. Piazza).

http://dx.doi.org/10.1016/j.eswa.2015.02.036
0957-4174/

In this context, the focus of this paper is on assistive and monitoring applications for emergency detection. A background on the recent literature on the topic is provided in the following section.

1.1. Background

The approaches followed by recent works in the literature differ from the sensors employed, the type of emergency and how it is detected. In Chernbumroong, Cang, Atkins, and Yu (2013) emergencies are detected by means of a wrist-worn device equipped with an accelerometer, a temperature sensor and an altimeter. Signals are then employed for detecting activities of daily living (ADLs) that are classified using a Support Vector Machine (SVM). In Charlon, Fourty, Bourennane, and Campo (2013), Paoli et al. (2012), a wearable device operates together with ambient sensors. In particular, Paoli et al. (2012) propose a wearable sensor node equipped with a 3-axis accelerometer, and an ambient sensor network consisting in infra-red, magnetic, and pressure sensors. The system is able to detect falls using mainly the wearable device signals, and algorithms operate on a miniPC platform. Similarly, in Charlon et al. (2013), infra-red motion sensors detect people's activity and a wearable electronic patch equipped with an accelerometer is employed for identifying and locating them, as well as to detect falls. The system also comprises a local and remote computer, and it is mainly tailored for hospitals. In other works, videocameras are employed as ambient sensors (Weinland, Ronfard, & Boyer, 2011). In particular, the system in Bosch-Jorge, Sánchez-Salmerón, Valera, and Ricolfe-Viala (2014) is based on a single wide-angle camera. The algorithm is able to detect falls by means of features based on the gravity vector and an SVM. In Botia, Villa, and Palma (2012), ADLs are analysed to detect abnormal patterns that can be hints for emergencies. ADLs are recognised by employing the signals coming from multiple sensors. In the paper, the authors simulate the behaviour of the ambient sensor, thus not specifying kind of sensors employed.

In regard to audio-based systems, as demonstrated by the recent projects and works (Brutti, Cristoforetti, Matassoni, Svaizer, & Omologo, 2013; Gemmeke et al., 2013; Vacher, Lecouteux, & Portet, 2014), they significantly increased their popularity in the recent years. The motivation resides in their versatility, since audio signals allow capturing people's activity, monitoring the acoustic environment and they enable speech-based user interfaces. In addition, people perceive microphones as less invasive sensors respect to video cameras (Goetze, Schroder, & Gerlach, 2012), and they do not risk to forget them as with wearable sensors.

The related contributions in this field come from recent important projects. DIRHA[1] (Brutti et al., 2013) is focused on natural spontaneous speech interaction with distant microphones in a home environment. Several microphone arrays are placed in the smart home so that multichannel algorithms increase the robustness against acoustic distortions and track the user location. Information is extracted from audio signals with a speech recogniser and with a sound classifier. The results are then employed by the understanding and dialogue management module that undertakes appropriate decisions. Acoustic events are employed mainly to increase the robustness of the speech recogniser. Similarly, in the Sweet-Home project (Vacher et al., 2013, 2014), the authors developed PATSH, a framework for processing multichannel audio signals and recognising both speech commands, distress calls and sound events. The framework combines algorithms for detecting and extracting sound events and for discriminating between speech and general sounds. Then, based on the type of signal, a speech recogniser captures commands and distress calls, or a sound classifier determines the type of the sound class. Users' location is determined exploiting both audio and other sensors information, and the decision logic is based on a Markov Logic Network. In Karpov et al. (2014) both audio and video signals are employed, thus improving the detection accuracy at the cost of a more invasive solution compared to audio-only ones.

Recent works specifically address the issue of vocal interfaces for people with speech disorders. This is motivated by the difficulty of general purpose speech recognisers to give results of sufficient accuracy with persons affected by this impairment (Mustafa, Rosdi, Salim, & Mughal, 2015). For people affected by dysarthria, a common approach is to propose speaker adaptive algorithms able to capture this kind of speech variation (Rudzicz, 2011; Sharma & Hasegawa-Johnson, 2013). The solution proposed by Christensen, Casanueva, Cunningham, Green, and Hain (2013) consists in a vocal user interface composed of a local in-home part, and of a remote part, with the former employed to issue vocal commands an for interacting with the devices, and the latter that performs the actual speech recognition. The approach followed in the ALADIN project (Gemmeke et al., 2013) consists in providing a framework for allowing the user to continuously adapt the vocabulary and the grammar of the vocal user interface. The continuous adaptation is achieved by learning vocal commands with the support of a direct intervention of the user that executes the action corresponding to the vocal command concurrently to the speech emission. Words in a command are then analysed for building a new grammar describing the relation of complex commands, and a new vocabulary.

Other contributions in the literature are more focused on the recognition of acoustic events. The system presented in Hollosi, Schroder, Goetze, and Appell (2010), Goetze et al. (2012) is based on microphones placed on the ceiling and on floor lamps. The authors developed a framework for the detection and localisation of acoustic events, comprising both speech (e.g., coughing) and non-speech sounds. Events classes are modelled by means of Gaussian Mixture Models and they are located with the general cross-correlation phase-transform (GCC-PHAT) algorithm. This information is then used for emergency detection. Similarly, the work by Ntalampiras, Potamitis, and Fakotakis (2010) addresses the classification of sound events using generative models and multidomain features (Mel-Frequency Cepstral Coefficients, MPEG-7 and features based on wavelet packets). The work is part of the Prometheus project (Ahlberg et al., 2008), where the goal is integrating the information coming from heterogeneous sensors to capture the behaviour of humans.

1.2. Contribution of this work

The system presented in this paper addresses the automatic detection of emergencies and the recognition of home automation commands. An emergency here is represented by a situation of distress for the user, where he/she intentionally asks for help, or by an abnormal acoustic event. Recalling the classification proposed in Acampora et al. (2013), the application scenario is thus monitoring for emergency detection and ambient assisted living. The system operates in two modalities that are chosen by the user to monitor different situations. The first modality, speech monitoring, is enabled when the user is inside the home and consists in recognising home automation commands and distress calls. Commands are automatically interpreted to control the appliances and the devices connected to the home automation subsystem. Distress calls are employed to provide tele-assistance to the users. In particular, a distress call triggers an automatic phone call to a relative or a care centre that then can provide assistance to the user. The acoustic environment is constantly monitored to detect speech signals by means of a Voice Activity Detector (VAD), and a speech recogniser based on PocketSphinx (Huggins-Daines et al., 2006) captures

[1] <http://dirha.fbk.eu/>.