Expert Systems With Applications 46 (2016) 77–86

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Acoustic Event Classification using spectral band selection and Non-Negative Matrix Factorization-based features

Jimmy Ludeña-Choez [a,b], Ascensión Gallardo-Antolín [a,*]

[a] Department of Signal Theory and Communications, Universidad Carlos III de Madrid, Avda. de la Universidad, 30, Leganés Madrid, 28911, Spain
[b] Facultad de Ingeniería y Computación, Universidad Católica San Pablo, Campus Campiña Paisajista s/n Quinta Vivanco, Barrio de San Lázaro, Arequipa, Peru

ARTICLE INFO

Keywords:
Acoustic Event Classification
Feature extraction
Temporal feature integration
Feature selection
Mutual information
Non-Negative Matrix Factorization

ABSTRACT

Feature extraction methods for sound events have been traditionally based on parametric representations specifically developed for speech signals, such as the well-known Mel Frequency Cepstrum Coefficients (MFCC). However, the discrimination capabilities of these features for Acoustic Event Classification (AEC) tasks could be enhanced by taking into account the spectro-temporal structure of acoustic event signals. In this paper, a new front-end for AEC which incorporates this specific information is proposed. It consists of two different stages: short-time feature extraction and temporal feature integration. The first module aims at providing a better spectral representation of the different acoustic events on a frame-by-frame basis, by means of the automatic selection of the optimal set of frequency bands from which cepstral-like features are extracted. The second stage is designed for capturing the most relevant temporal information in the short-time features, through the application of Non-Negative Matrix Factorization (NMF) on their periodograms computed over long audio segments. The whole front-end has been evaluated in clean and noisy conditions. Experiments show that the removal of certain frequency bands (which are mainly located in the medium region of the spectrum for clean conditions and in low frequencies for noisy environments) in the short-time feature computation process in conjunction with the NMF technique for temporal feature integration improves significantly the performance of a Support Vector Machine (SVM) based AEC system with respect to the use of conventional MFCCs.

1. Introduction

In recent years, the problem of automatically detecting and classifying acoustic non-speech events has attracted the attention of numerous researchers. Although speech is the most informative acoustic event, other kind of sounds (such as laughs, coughs, keyboard typing, etc.) can give relevant cues about the human presence and activity in a certain scenario (for example, in an office room). This information could be used in different applications, mainly in those with perceptually aware interfaces such as smart-rooms (Temko & Nadeu, 2006), automotive applications (Muller, Biel, Kim, & Rosario, 2008), mobile robots working in diverse environments (Chu, Narayanan, Kuo, & Mataric, 2006) or surveillance systems (Principi, Squartini, Bonfigli, Ferroni, & Piazza, 2015).

Acoustic Event Classification (AEC) systems can be formulated as a machine learning problem consisting in two main stages: feature extraction (or front-end) and classification (or back-end). The first one obtains a parametric and compact representation of the audio signals more appropriate for classification. The purpose of the second one is to determine which Acoustic Event (AE) has been produced through a certain decision process. Several front-ends and classifiers have been proposed and compared in the literature for this task. Nevertheless, the high correlation between the performance of different classifiers suggests that the main problem is not the choice of the classification technique, but a design of a suitable feature extraction process for AEC (Kons & Toledo, 2013). This paper, precisely, focuses on this issue.

Many state-of-the art front-ends are composed of two modules: *short-time feature extraction*, in which acoustic coefficients are computed on a frame-by-frame basis (typically, the frame period used for speech/audio analysis is about 10–20 ms) from analysis windows of 20–40 ms, and *temporal feature integration* (Meng, Ahrendt, & Larsen, 2007), in which features at larger time scales are extracted by combining somehow the short-time characteristics information over a longer time-frame composed of several consecutive frames. The resulting characteristics are often called *segmental features* (Ludeña-Choez & Gallardo-Antolín, 2013a, 2015; Zhang & Schuller, 2012).

* Corresponding author. Tel.: +34 916246250; fax: +34 916248749.
E-mail addresses: jimmy@tsc.uc3m.es, jludenac@ucsp.edu.pe (J. Ludeña-Choez), gallardo@tsc.uc3m.es (A. Gallardo-Antolín).

http://dx.doi.org/10.1016/j.eswa.2015.10.018
0957-4174/

In this paper, two techniques which improve the performance of each of these modules by taking into account the specific spectro-temporal structure of acoustic events are presented. For short-time feature extraction, an automatic spectral band selection method is applied in order to emphasize the more relevant frequencies (and less redundant) of the acoustic events in the parameterization procedure, whereas for temporal feature integration, Non-Negative Matrix Factorization (NMF) (Lee & Seung, 1999) is used for obtaining a set of segmental features which better summarizes the temporal information contained in the frame-based acoustic characteristics.

This paper is organized as follows: Section 2 introduces related work on feature extraction of acoustic event signals. Section 3 describes the short-time feature extraction process based on spectral band selection. Section 4 presents the application of NMF for the design of the temporal feature integration module. Section 5 presents the experiments and results to end with some conclusions in Section 6.

2. Related work

In first works on Acoustic Event Classification and detection, the parametric representations of audio signals used were strongly based on those previously developed for speech processing and related tasks, such as speech and speaker recognition. As these acoustic parameters are usually extracted on a frame-by-frame basis, they are commonly known as short-time features. Good examples are the conventional Mel-Frequency Cepstral Coefficients (MFCC) (Kwangyoun & Hanseok, 2011; Temko & Nadeu, 2006; Zhuang, Zhou, Hasegawa-Johnson, & Huang, 2010; Zieger, 2008), log filter bank energies (Zhuang et al., 2010), Perceptual Linear Prediction (PLP) (Portelo et al., 2009), log-energy, spectral flux, entropy and zero-crossing rate (Perperis et al., 2011; Temko & Nadeu, 2006). The combination of some of these short-time features into high-dimensional acoustic vectors has also been studied, as well as the application of feature selection algorithms over these large pools of characteristics, in order to precisely reduce their dimensionality (Butko & Nadeu, 2010; Kiktova-Vozarikova, Juhar, & Cizmar, 2015; Zhuang et al., 2010; Zhuang, Zhou, Huang, & Hasegawa-Johnson, 2008).

Nevertheless, as pointed in Zhuang et al. (2010), many of these conventional acoustic features are not necessarily the more appropriate for AEC tasks because most of them have been designed according to the spectral characteristics of speech which are quite different from the spectral structure of acoustic events. In addition, some types of acoustic events present a typical temporal structure (for example, the periodic pattern of phone rings) that should be somehow exploited in order to improve feature representation and discrimination capabilities. For these two reasons, recent research is being focused on finding a set of features that adequately represents the acoustic events.

To deal with the first problem, new acoustic parameters such as Power Normalised Cepstral Coefficients (PNCC) (Principi et al., 2015) and those derived from Gammatone (Plinge, Grzeszick, & Fink, 2014) or Gammachirp filter banks (Alam, Kenny, & O'Shaughnessy, 2014) have been proposed. Other works try to discover the hidden structure of the acoustic data by means of the application of Non-Negative Matrix Factorization (NMF) or K-Singular Value Decomposition (KSVD) on audio spectrograms (Choi, Park, Han, & Ko, 2015). In an alternative approach (Ludeña-Choez & Gallardo-Antolín, 2013a), from the analysis of the AE spectral characteristics, it was concluded the importance of medium and high frequencies for discriminating between different acoustic events, yielding to the design of a new front-end based on the high pass filtering of the audio signals, which achieves good results in clean and noisy conditions (Ludeña-Choez & Gallardo-Antolín, 2015). Note that all these approaches can be seen as different modifications of the conventional mel-scaled auditory filter bank which is applied on the audio spectrograms in the short-time feature extraction process.

Following the idea that some frequency bands may be more useful for distinguishing between different sounds than others, in this paper, a modified mel-scaled filter bank is proposed in which only a selected set of spectral bands are considered in the computation of the short-time characteristics. In contrast to the already mentioned approaches, in this work, an automatic method is used to find this optimal set of frequency bands from which cepstral-like coefficients are derived, as explained in Section 3. In particular, several Feature Selection (FS) techniques based on Mutual Information (MI) measures have been evaluated and compared for this purpose. Note that, in comparison with previous works about FS for tasks related to acoustic events, in this paper it is not intended to use FS for dimensionality reduction but to provide a better spectral representation of the AEs through the selection of the more relevant and less redundant spectral bands.

In order to cope with the second problem, the idea of simultaneously performing temporal and spectral analysis to yield so-called spectro-temporal features has lately emerged, e.g. high-level features (also called audio banks) (Sandhan, Sonowal, & Choi, 2014), spectrogram patch modeling using Restricted Boltzman Machines (RBM) (Espi, Fujimoto, Kubo, & Nakatani, 2014) and 2D Gabor-based biologically inspired features (Schroder, Goetze, & Anemuller, 2015). As these methods are usually very computational demanding, temporal feature integration techniques, in which features at larger time scales are extracted by combining the short-time parameters contained in long audio segments, have become an interesting alternative. Among these techniques, the approach based on Filter Bank Coefficients (FC), which was initially proposed for general audio and music genre classification (Arenas-García, Larsen, Hansen, & Meng, 2006; McKinney & Breebaart, 2003; Meng et al., 2007), has been experimented for AEC with promising results (Mejía-Navarrete, Gallardo-Antolín, Peláez, & Valverde, 2011). Its main advantage is that it allows to capture the dynamic structure in the short-time features. The idea behind FC is to summarize the periodogram of each short-time feature dimension by computing the power in several predefined frequency bands using a filter bank, which is usually the one proposed in McKinney and Breebaart (2003). However, as pointed in Arenas-García et al. (2006), this fixed filter bank is not general enough since the relevance of the dynamics in the short-time features for classification can be expected to be task-dependent.

Based on this premise, in Ludeña-Choez and Gallardo-Antolín (2013b) a method based on Non-Negative Matrix Factorization (NMF) for the design of a filter bank for the computation of FC-based features more suitable for AEC has been proposed by the authors and successfully tested in clean conditions. In comparison with similar works (Arenas-García et al., 2006), the approach described in Ludeña-Choez and Gallardo-Antolín (2013b), which is described in Section 4, is unsupervised and general enough to be applied to any sound signals.

In summary, in view of the main limitations of the audio feature extraction methods existing in the literature, in this paper, a novel front-end for AEC tasks is proposed. The major contributions of this work are the following: the development of a new short-time parameterization based on the automatic selection of spectral bands which better reflects the spectral characteristics of audio events, its combination with a feature integration technique based on NMF which aims to improve the modeling of the temporal behavior of short-time features; and the evaluation of the complete front-end in both, clean and noisy conditions.

Fig. 1 represents the block diagram of the whole audio feature extraction process. As mentioned before, it can be observed that it consists of two main stages: short-time feature extraction and temporal feature integration. Next sections are devoted to the description of both modules.