Digital Signal Processing 63 (2017) 123–134

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Unsupervised detection of acoustic events using information bottleneck principle

Yanxiong Li *, Qin Wang, Xianku Li, Xue Zhang, Yuhan Zhang, Aiwu Chen, Qianhua He, Qian Huang

School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510640, China

ARTICLE INFO

Article history:
Available online 5 January 2017

Keywords:
Acoustic event detection
Information bottleneck principle
Audio signal processing

ABSTRACT

An unsupervised approach based on Information Bottleneck (IB) principle is proposed for detecting acoustic events from audio streams. In this paper, the IB principle is first concisely presented, and then the practical issues related to the application of IB principle to acoustic event detection are described in detail, including definitions of various variables, criterion for determining the number of acoustic events, tradeoff between amount of information preserved and compression of the initial representation, and detection steps. Further, we compare the proposed approach with both unsupervised and supervised approaches on four different types of audio files. Experimental results show that the proposed approach obtains lower detection errors and higher running speed compared to two state-of-the-art unsupervised approaches, and is little inferior to the state-of-the-art supervised approach in terms of both detection errors and runtime. The advantage of the proposed unsupervised approach over the supervised approach is that it does not need to pre-train classifiers and pre-know any prior information about audio streams.

1. Introduction

With the development of multimedia technologies, the volumes of audio files containing various types of acoustic events (e.g. music, applause, laughter, fighting) have been rapidly increasing. How to effectively discover these acoustic events in the huge mass of audio files has been received more and more attentions in the field of audio signal processing [1]. Currently, there are two kinds of approaches for discovering acoustic events in audio files: the supervised approach and the unsupervised approach. In the supervised approach, the features are first extracted from audio streams, such as Mel-Frequency Cepstral Coefficients (MFCCs). Then, the frames (or segments) of the features are classified as one of the pre-defined types after feeding the features to the pre-trained classifiers, such as Deep Neural Network (DNN) [2,3], Support Vector Machine (SVM) [3–9], regression forests [10], Gaussian Mixture Model (GMM) [11], Hidden Markov Model (HMM) [8,12,13], and discriminative binary classifiers [14]. In the unsupervised approach, the features are first extracted from the audio streams. Then, a clustering algorithm, e.g. spectral clustering [15], is used to merge the frames (or segments) of features belonging to the same audio type into one cluster and to assign one unique label to each cluster, without pre-training classifiers for each type of acoustic events and pre-knowing the types of acoustic events.

Both the types and numbers of acoustic events in complex audio files are generally unknown in practice. How to discover various types of acoustic events from a huge mass of audio files is an important issue for audio content analysis and retrieval. The supervised approach has to pre-know the types of acoustic events and has to pre-train a model (e.g. DNN, HMM, SVM, GMM) for each type of acoustic event. On the contrary, the unsupervised approach can find various types of acoustic events in complex audio files without pre-knowing the prior information and without pre-training complex models for each type of acoustic event in advance. Hence, the unsupervised approach is more flexible and universal than the supervised approach for content analysis of complex audio files. In this paper, we investigate a clustering technique based on the Information Bottleneck (IB) principle [16] for detecting acoustic events from audio streams. Compared to other clustering techniques, the IB-based clustering is based on preserving the relevant information specific to a given problem instead of arbitrarily introducing a distance function between elements. In addition, it tries to find the tradeoff between the most compact representation and the most informative representation of the data. The IB principle has been applied for processing text documents [17] and images [18], but these tasks are different from

* Corresponding author at: Room 223, Shaw Science Building, School of Electronic and Information Engineering, South China University of Technology, 381 Wushan Road, Guangzhou, 510640, China
E-mail address: yanxiongli@163.com (Y. Li).

http://dx.doi.org/10.1016/j.dsp.2016.12.012
1051-2004/

unsupervised detection of acoustic events in audio streams. To the best of our knowledge, the IB principle has never been proposed to detect different types of acoustic events up to now. The main contributions of this paper are to propose an unsupervised approach for acoustic event detection by using IB-based clustering and its comparison to state-of-the-art approaches. Practical issues for applying the IB principle to acoustic event detection are solved, including definitions of various variables, criterion for determining the number of acoustic events, tradeoff between amount of information preserved and compression of initial representation, and detection steps.

The remainder of the paper is organized as follows. Related works about acoustic event detection are introduced in Section 2. In Section 3, we briefly describe the IB principle. The proposed approach for acoustic event detection is discussed in detail in Section 4. Section 5 presents experimental results and discussions, and finally conclusions are drawn in Section 6.

2. Related works

Many studies have been done by using the supervised approaches to detect or to classify acoustic events. Gencoglu et al. [2] used a DNN classifier for recognizing sixty-one distinct classes of acoustic events. The DNN classifier obtained classification accuracy of 60.3%, whereas the conventional HMM classifier yielded classification accuracy of 54.8%. McLoughlin et al. [3] presented an acoustic events classification framework that compared auditory image front-end features with spectrogram image-based front-end features, using SVM and DNN classifiers. The results showed that the DNN-based classifier was superior to the SVM-based classifier under multiple SNR (Signal Noise Ratio) conditions. In the work of Kucukbay et al. [4], a framework was introduced for detecting sixteen distinct acoustic events. Acoustic events were recognized using MFCCs features along with SVM classifiers. The results showed that 55% F-measure score was attained after optimizing the parameters of MFCCs and SVM, and 7% improvement of the F-measure score was obtained compared with the standard method. Lu et al. [5] represented the temporal-frequency structures of acoustic events as a bag of spectral patch exemplars, and applied K-means clustering based vector quantization to the whitened spectral patches. A sparse feature representation was extracted based on the similarity measurement of the learned spectral exemplars. A SVM classifier was built on the sparse representation for detecting acoustic events. The results showed that the sparse representation significantly outperformed the traditional frame based representation for discriminating nine different acoustic events. The work of Tran et al. [6] reported a classification method based on probabilistic distance SVM. They studied a parametric approach for characterizing sound signals using the distribution of the sub-band temporal envelope, and kernel techniques for the sub-band probabilistic distance under the framework of SVM. Evaluated on a database containing ten types of acoustic events, the results showed that their approach significantly outperformed conventional SVM classifiers with MFCCs. Lu et al. [7] proposed a SVM based method to detect five types of acoustic events in audio files. They discussed the complementarity of various kinds of features for detection of acoustic events. Evaluated on 8-hour TV data, an average F value of 79.71% was obtained. The study carried out by Portelo et al. [8] focused on detecting fifteen semantic acoustic events by using SVM and HMM-based classifiers with multiple features, different kernels and several analysis windows. Evaluated on documentaries and films, they yielded promising results, in spite of the difficulties posed by mixtures of acoustic events. For classifying multi-class acoustic events, Peng et al. [9] designed an entropy-based binary hierarchical classifier with selected feature subsets for individual component classifier. They used SVM as the component classifier for classifying seven acoustic events for elder-care application. Their approach obtained competitive performance compared to the traditional one-against-one approach while the number of training and testing SVMs was much less in their hierarchical scheme. In addition, features selection facilitated the training of the component classifier by filtering out possible redundant and irrelevant feature components. Phan et al. [10] proposed an approach for detecting acoustic events based on regression forests. Using the concept of acoustic super-frames, two classifiers were trained for identifying the super-frames of background and different acoustic events of interest. Cluster-specific regressors were learned by using the super-frames of acoustic events. Evaluated on two different databases, their approach significantly exceeded those of three baseline systems. Zhang et al. [11] extracted time-frequency features by using tensor-based sparse approximation for acoustic event classification. In their method, the observed data encoded as a higher-order tensor and discriminative features were extracted in spectrotemporal domain. When used to classify thirteen unique acoustic events extracted from two corpora, the results showed that the new features with GMM classifier yielded average accuracy improvement by 9.7% and 12.5% compared with matching pursuit and MFCCs features with the same classifiers, respectively. Niessen et al. [12] presented a two-layer hierarchical HMM for recognizing acoustic events in which one layer corresponded to acoustic events and another layer corresponded to sub-event clusters. When evaluated on the experimental data consisting of sixteen types of acoustic events, the results showed that the hierarchical HMM achieved an average frame-based F-measure score of 45.5%, and obtained better performance than the traditional GMM and HMM classifiers. Cai et al. [13] proposed a flexible framework for detecting acoustic events in audio streams. HMMs were used to model acoustic events, and a grammar network was designed to connect various HMMs to fully explore the transitions among them. The framework was convenient to add or delete target acoustic events. Evaluations on 12 h audio data showed that the method achieved a F score of 82.6% for detecting ten types of acoustic events. Kumar et al. [14] developed a technique for detecting acoustic events based on identifying patterns of occurrences of automatically learned sound atomic units. The results showed that their approach worked well for detecting ten types of acoustic events in complex audio files.

From the aforementioned introductions, it can be seen that many studies have been done for detecting or classifying acoustic event by using supervised approaches. On the contrary, the reports concerning unsupervised approaches for acoustic event detection are very few. Lu et al. [15] presented an unsupervised approach to discover acoustic events in audio files. They used a spectral clustering algorithm with context-dependent scaling factors to cluster audio segments, with the aim to obtain the type number of acoustic events and to merge audio segments of the same type into one cluster. Another state-of-the-art unsupervised approach is Agglomerative Hierarchical Clustering (AHC) algorithm with using Bayesian Information Criterion (BIC) as stopping criterion, i.e. the AHC + BIC algorithm [19], which is predominantly used for speaker clustering instead of acoustic event clustering. In the AHC + BIC algorithm, BIC is obtained as an approximation of marginal log-likelihood of the data given a model, and is only valid in the large data limit [20]. In addition, it is difficult to set a proper value for tuning the BIC penalty, which significantly affects the performance of the AHC + BIC algorithm [21].

3. Information bottleneck principle

The IB principle is a distributional clustering approach inspired from the rate-distortion theory [22,23]. Let $\boldsymbol{Y}$ be a set of variables related to input data $\boldsymbol{X}$, and $\boldsymbol{C}$ be a compressed representation