Pattern Recognition Letters 31 (2010) 1543–1551

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Real-world acoustic event detection

Xiaodan Zhuang *, Xi Zhou, Mark A. Hasegawa-Johnson, Thomas S. Huang

Beckman Institute of Advanced Science and Technology, Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

ARTICLE INFO

Article history:
Available online 19 February 2010

Keywords:
Acoustic Event Detection
Feature selection
Hidden markov model
Artificial neural network
Tandem model
Gaussian mixture model supervector

ABSTRACT

Acoustic Event Detection (AED) aims to identify both timestamps and types of events in an audio stream. This becomes very challenging when going beyond restricted highlight events and well controlled recordings. We propose extracting discriminative features for AED using a boosting approach, which outperform classical speech perceptual features, such as Mel-frequency Cepstral Coefficients and log frequency filterbank parameters. We propose leveraging statistical models better fitting the task. First, a tandem connectionist-HMM approach combines the sequence modeling capabilities of the HMM with the high-accuracy context-dependent discriminative capabilities of an artificial neural network trained using the minimum cross entropy criterion. Second, an SVM–GMM-supervector approach uses noise-adaptive kernels better approximating the KL divergence between feature distributions in different audio segments. Experiments on the CLEAR 2007 AED Evaluation set-up demonstrate that the presented features and models lead to over 45% relative performance improvement, and also outperform the best system in the CLEAR AED Evaluation, on detection of twelve general acoustic events in a real seminar environment.

1. Introduction

Much research in audio content analysis has typically addressed the problem of segregating a few audio sources (Brown and Cooke, 1994; Ellis, 1996) or segmenting an audio stream into a small number of acoustically compact categories (Pinquier, 2002; Zhang and Kuo, 2001). Acoustic Event Detection (AED) aims to detect specified acoustic events such as gunshots (Clavel et al., 2005), explosions (Naphade, 2001; Cui et al., 2003a), speech/music transitions (Pinquier, 2002), cough events (Smith et al., 2006), or audience cheering at a sports event (Baillie and Jose, 2003). The existence and timestamps of many non-speech sounds, i.e. (non-speech) acoustic events, reveal human and social activities. Such information is very helpful in applications such as surveillance, multimedia information retrieval and intelligent conference rooms.

While most of the work in acoustic event detection focuses on a few highlight acoustic events, the 2007 AED Evaluation sponsored by the project "Classification of Events, Activities and Relationships (CLEAR)" (Temko et al., 2006; Temko, 2007) was performed on a continuous audio database recorded in real seminars (Temko and Nadeu, 2005). Systems attempted to identify both the temporal boundaries and labels of twelve acoustic events (door slam, paper wrapping/rustling, foot steps, knocking, chair moving, phone ringing, spoon/cup jingle, key jingle, keyboard typing, applause, cough, and laughter). Instead of being exclusively highlight events, many of the acoustic events in the CLEAR Evaluations were either subtle (low SNR, e.g. steps, paper wrapping/rustling, and keyboard typing), or/and overlapping with speech, making the task particularly challenging. The real environment factor added to the variation of the events as well as the difficulty of segmenting the audio stream. Although different system architectures and feature sets have been explored (Temko et al., 2006; Temko, 2007), even the top rated AED system (around 30% accuracy) left much space for improvement (Zhou et al., 2007). By contrast, classification of performed isolated events in silent rooms saw very good performance achieved by some of the same research teams (Temko et al., 2006). The evaluation highlighted the challenges in the detection of a large set of ordinary acoustic events in a real world environment.

To tackle AED in such a realistic setting, we believe further improvement is possible with features and statistical models better fitting the task, drawing lessons from the CLEAR 2007 AED Evaluation. A small part of this work was previously reported (Zhou et al., 2007; Zhuang et al., 2008).

Analysis of the spectral structure of acoustic events and design of a suitable feature set are important for AED. Various audio perceptual features have been proposed for different analysis tasks (Brown and Cooke, 1994; Scheirer, 1999; Cui et al., 2003b). In the recent CLEAR Evaluations for AED, the most popular features are speech perception features (Temko et al., 2006; Atrey et al., 2006), such as Mel-Frequency Cepstral Coefficients (MFCC) and log frequency filter bank parameters, which have been proven to represent speech spectral structure well. However, these features

* Corresponding author. Tel.: +1 217 898 6732; fax: +1 217 244 9233.
E-mail addresses: xiaodan.zhuang@gmail.com (X. Zhuang).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.02.005

are not necessarily suitable for AED for the following reasons. First, limited work has been done in studying the spectral structure of acoustic events. The speech features designed according to the spectral structure of speech might be far from optimal for AED. Second, the Signal-to-Noise Ratio (SNR) is low for AED especially when the overlapping speech can be seen as noise.

In this study we propose a new front-end feature analysis and selection approach for AED. Considering the varying discriminative capabilities of each feature component for the AED task, we propose a boosting approach to construct a discriminative feature set from a large feature pool.

AED in real seminars differs from classification of isolated events in a silent environment, calling for different statistical models. While SVMs were shown to be optimal for the latter (Schòlkopf and Smola, 2002), the former saw most leading CLEAR participants using dynamic Bayesian networks (Temko et al., 2006; Temko, 2007), in particular, hidden markov models (HMMs). HMMs owe their success to the Viterbi algorithm (Forney, 1972), which allows them to compute simultaneously optimal segmentation and classification of the audio stream: noise in individual frames is alleviated by the HMM's learned hysteresis, i.e., its typical learned preference for self-transitions rather than non-self-transitions in the hidden finite state machine.

To take advantage of this proven approach, we leverage a framework in which HMMs are used to achieve audio segmentation and event classification simultaneously. To alleviate HMM's problem that each hidden state models only local observations, we propose to use the tandem connectionist-HMM approach (Hermansky et al., 2000), where an artificial neural network (ANN) outputs posterior probabilities of event types based on very-long-duration, temporally overlapping observation vectors, leading to better contextual modeling and event discrimination. To further refine the event detection result, we propose using vectors of the per-segment adapted means of a Gaussian mixture model (GMM), so-called GMM supervectors (Campbell et al., 2006), to abstract the noisy features in the training audio segments and the hypothesized segments obtained by the tandem model. An SVM with kernels built on these GMM supervectors, namely the SVM–GMM-supervector classifier, is used to replace the labels proposed by the first-pass tandem model, when such replacement is desirable according to held-out development data.

We perform acoustic event detection experiments on the same setup as the AED Evaluation in CLEAR 2007. It is demonstrated that the discriminative feature set constructed by the boosted feature selection approach, the tandem connectionist-HMM approach and the SVM–GMM-supervector approach for refining the result jointly contribute to performance improvement from 28.2% to 41.2% absolute. This also outperforms our submission in the CLEAR 2007 AED Evaluation, which was the best ranked in the challenging AED task.

2. Discriminative features for AED

2.1. Spectral correlates of acoustic events

Over the past decades, a lot of research has been done on speech perceptual features (Hermansky, 1999; Reynolds and Rose, 1995). Currently, the speech features are designed mainly based on properties of speech production and perception. Based on knowledge of the human auditory system, the envelope of the spectrogram (formant structure) instead of the fine structure of the spectrogram (harmonic structure) is believed to hold most information for speech. Both log frequency filter bank parameters and Mel Frequency Cepstral Coefficients (MFCC) (Hermansky, 1999) use triangular band pass filters to smooth out the fine structure of the spectrogram. Moreover, to simulate the non-uniform frequency resolution observed in human auditory perception, these speech feature sets use bandwidths based on the perceptual critical band, e.g., they have higher resolution in the low frequency part of the spectrum. These features have been successfully used to characterize speech signal as well as other signal perceived by human audition, e.g., music (Logan, 2000).

The spectral structure of acoustic events is different from that of speech, as shown in Fig. 1, therefore speech feature sets designed according to the spectral structure of speech might be far from optimal for AED. For example, they might neglect frequency ranges that contain little speech discriminative information, but which may contain much discriminative information for acoustic events.

To analyze the spectral structure of acoustic events for AED, we carry out Kullback–Leibler Divergence (KLD) based feature discriminative capability analysis. This helps us to understand the relevance of different feature components (in a speech feature set) for the AED task, compared to speech recognition. The distance between the distributions associated with an acoustic event label and the other audio labels reveals the discriminative capability of the feature for that acoustic event.

KL Divergence (KLD), denoted by $D(p||q)$, is a measure (a "distance" in a heuristic sense) between two distributions, p and q, and is defined as the cross entropy between p and q minus the self entropy of p.

$$D(p||q) = \int p(x) \log \frac{p(x)}{q(x)}. \tag{1}$$

We use KLD to measure the discriminative capability of each feature component for each acoustic event. Let $d_{ij} = D(p_{ij}||q_i)$ denote the divergence between the distribution of the ith feature component for the jth acoustic event and the global distribution of the ith feature component for all the audio.

The global discrimative capability of the ith feature component is defined by

$$d_i = \sum_j P_j d_{ij}, \tag{2}$$

where P_j is the prior probability for the jth acoustic event.

To calculate the KLD without prior knowledge of each feature component's distribution, we use nonparametric density estimation, in particular, Parzen window density estimation (Duda et al., 2001) with Gaussian kernels to estimate the distribution of each feature component for each event.

The global discriminative capabilities for different log frequency filter bank parameters are estimated for AED and digit classification. The AED data used is the training data used in the detection experiments, as detailed in Section 7. The task of speech digit classification uses digit speech data in TIDIGITS dataset (Tidigits, 1993). In these preliminary experiments, we observe that the tasks of spoken digit recognition and acoustic event detection assign different relative levels of importance to each of the feature components.

2.2. Boosted feature selection

As discussed in the above subsection, the sum of the KLD between every event-specific distribution and the global distribution characterizes the discriminative capability of the concerned feature component. The goal of feature selection, however, is to find the most discriminative feature set instead of finding a set of individually most discriminative feature components.

A few algorithms exist for feature selection. In particular, a floating search approach was proposed in (Pudil et al., 1994), and an extended and more complicated version was later reported in