Pattern Recognition Letters 34 (2013) 1085–1093

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

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

Overlapping sound event recognition using local spectrogram features and the generalised hough transform

J. Dennis [a,b,*], H.D. Tran [a], E.S. Chng [b]

[a] Institute for Infocomm Research, 1 Fusionopolis Way, #08-01 South Tower Connexis, Singapore 138632, Singapore
[b] School of Computer Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Singapore

ARTICLE INFO

Article history:
Received 30 August 2012
Available online 14 March 2013

Communicated by S. Sarkar

Keywords:
Overlapping sound event recognition
Local spectrogram features
Keypoint detection
Generalised Hough Transform

ABSTRACT

In this paper, we address the challenging task of simultaneous recognition of overlapping sound events from single channel audio. Conventional frame-based methods are not well suited to the problem, as each time frame contains a mixture of information from multiple sources. Missing feature masks are able to improve the recognition in such cases, but are limited by the accuracy of the mask, which is a non-trivial problem. In this paper, we propose an approach based on Local Spectrogram Features (LSFs) which represent local spectral information that is extracted from the two-dimensional region surrounding "keypoints" detected in the spectrogram. The keypoints are designed to locate the sparse, discriminative peaks in the spectrogram, such that we can model sound events through a set of representative LSF clusters and their occurrences in the spectrogram. To recognise overlapping sound events, we use a Generalised Hough Transform (GHT) voting system, which sums the information over many independent keypoints to produce onset hypotheses, that can detect any arbitrary combination of sound events in the spectrogram. Each hypothesis is then scored against the class distribution models to recognise the existence of the sound in the spectrogram. Experiments on a set of five overlapping sound events, in the presence of non-stationary background noise, demonstrate the potential of our approach.

1. Introduction

The topic of sound event recognition (SER) covers the detection and classification of sound events in unstructured environments, which may contain multiple overlapping sound sources and non-stationary background noise. Many sounds contribute to the understanding and context of the surrounding environment, and therefore should not be regarded simply as noise, as is common in automatic speech recognition (ASR). Instead, such sounds are useful in many applications, such as security surveillance (Gerosa et al., 2007), bioacoustic monitoring (Bardeli et al., 2010), meeting room transcription (Temko and Nadeu, 2009; Zhuang et al., 2010), and ultimately "machine hearing" (Lyon, 2010).

Although a variety of techniques has been developed for SER (Cowling and Sitte, 2003), the most popular approaches are often based on frame-based features, such as Mel-frequency cepstral coefficients (MFCCs) from ASR, or MPEG-7 descriptors (Casey, 2001). These can then be modelled with Gaussian Mixture Models (GMMs) and combined with Hidden Markov Models (HMMs) for recognition, or used to train a Support Vector Machine (SVM) for discriminative classification. While these methods are effective in ASR for clean single-source speech recognition (O'Shaughnessy, 2008), such systems may not perform well in the challenging mismatched conditions present in many SER tasks. Missing feature recognition systems can overcome the problem to an extent (Raj and Stern, 2005), however a major challenge is estimating the mask to separate the signal from the background noise (Wang, 2005), and in practise the performance of such systems is highly dependent on the quality of the mask. Such frame-based techniques are also not well suited to recognition of overlapping sounds, as each feature contains information from multiple sources.

Research into the human understanding of speech (Allen, 1994) shows that there is little biological evidence for frame-based features, and that the human auditory system may be based on the partial recognition of features that are local and uncoupled across frequency. This enables the human recognition system to be robust to noise and distortion occurring across separate regions of the spectrum. Therefore, in this paper we develop an SER system based on local spectrogram features (LSFs), which provide a significant departure from conventional frame-based features. This work extends our initial presentation of the idea in (Dennis et al., 2012), where here we introduce a more complete model of the sound

[*] Corresponding author at: Institute for Infocomm Research, 1 Fusionopolis Way, #08-01 South Tower Connexis, Singapore 138632, Singapore. Tel.: +65 84842036.
E-mail addresses: stujwd@i2r.a-star.edu.sg (J. Dennis), hdtran@i2r.a-star.edu.sg (H.D. Tran), ASESChng@ntu.edu.sg (E.S. Chng).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.02.015

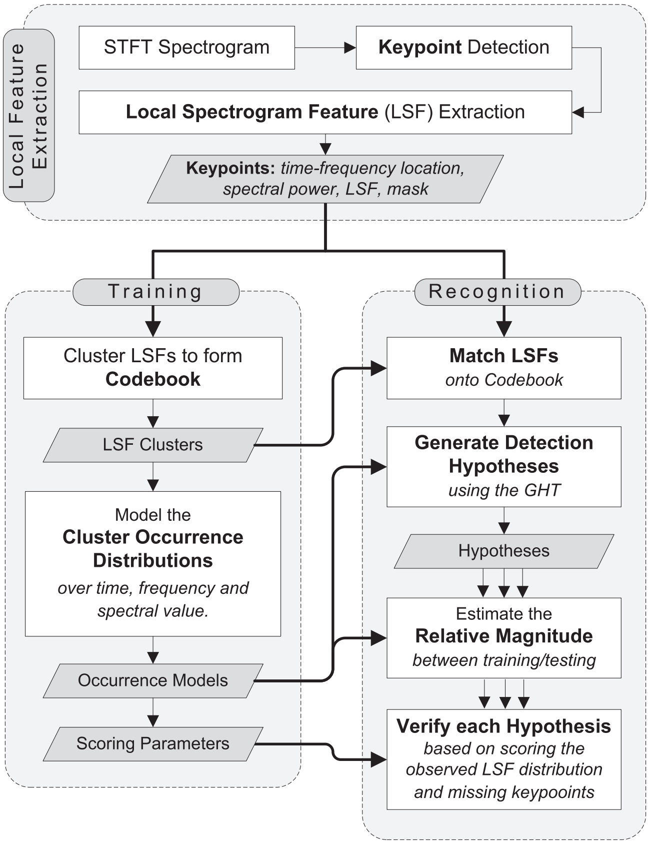
Local Feature Extraction
STFT Spectrogram
Keypoint Detection
Local Spectrogram Feature (LSF) Extraction
Keypoints: time-frequency location, spectral power, LSF, mask
Training
Recognition
Cluster LSFs to form Codebook
LSF Clusters
Model the Cluster Occurrence Distributions over time, frequency and spectral value.
Occurrence Models
Scoring Parameters
Match LSFs onto Codebook
Generate Detection Hypotheses using the GHT
Hypotheses
Estimate the Relative Magnitude between training/testing
Verify each Hypothesis based on scoring the observed LSF distribution and missing keypooints

Fig. 1. Overview of the proposed LSF recognition system.

information in the spectrogram, and enhance the recognition and scoring process to achieve more robust recognition across a range of experimental conditions.

Our proposed method takes inspiration from works in the field of object detection from image processing by Lowe (2004) and Lehmann et al. (2011), where finding objects in a cluttered real-world scene can be seen as having many parallels with that of overlapping SER. The central idea is to characterise a spectrogram by a set of independent local features, where each feature represents a glimpse (Cooke, 2006) of the local spectral information. Fig. 1 gives an overview of the idea, which can be broken down into the following three steps:

1. *Local feature extraction*: We first detect "keypoints" in the spectrogram to locate characteristic spectral peaks and ridges. For each keypoint, we extract an LSF and local missing feature mask to represent the local spectral region.
2. *Training*: The extracted LSFs are first clustered to generate a codebook. Each sound event is then modelled through the keypoint-occurrence distribution of the codebook clusters in the training spectrograms, and scoring parameters are extracted for verification during testing.
3. *Recognition*: The LSFs are first matched onto the codebook. We then generate sound onset hypotheses using the Generalised Hough Transform (GHT) (Ballard, 1981), which is a voting system that sums the keypoint-cluster distribution information in the Hough accumulator space. Finally, we verify each hypothesis by estimating the relative magnitude of the sound event between training and testing, and scoring it against the trained model.

The key advantage of our method is the use of local features combined with the GHT, which was successful in object detection in image processing (Lowe, 2004). The local features allow for independent local glimpses of the sound to be extracted, and a local missing feature mask to be estimated, which makes the system more robust to non-stationary noise. In addition, as the Hough accumulator is a summation of local evidence, a sound can still be recognised even when a proportion of features is missing or corrupted due to noise or overlapping sounds. Also, the representation of each sound in the Hough accumulator space is sparse and separable, such that overlapping sounds will produce distinct spikes in the accumulator that can be detected. This is an advantage over conventional HMM recognition systems, where the likelihoods are multiplicative, such that noise or overlapping sounds affecting one part of the feature has an adverse affect on the whole recognition.

Previous work on recognition of overlapping sounds can be separated into two distinct methodologies. The first is blind source separation, where factorisation is commonly used to decompose the input signal. For example, Heittola et al. (2011) use unsupervised non-negative matrix factorisation (NMF) to process the input audio into four component streams, where different sound events may be separated into different streams for recognition. Dessein et al. (2012) apply additional constraints such as sparsity on the NMF to improve the decomposition. Experiments show that both systems can separate overlapping sounds to some extent, although it is noted by Heittola et al. (2011) that the problem of controlling the outcome of the factorisation is one of the major difficulties with the NMF approach.

The second group of methods are based on direct classification. One approach developed for ASR is Factorial HMMs (FHMMs) (Roweis, 2003), based on the MixMax model of source interaction (Nádas et al., 1989), where the best combination of hidden states is found among the trained models to explain the observed feature. A more recent approach by Temko and Nadeu (2009) uses hierarchical SVM, where the first SVM classifies the input as either isolated events or a combined "overlapped" class, and the second SVM then identifies the overlapped combination. Tran and Li (2011) use a different approach that transforms the probabilistic distribution of the subband information to a new domain, where SVM can be used to detect sounds within a confidence interval.

Previous works have also used local spectral information, although not in the context of overlapping sounds. For example, Kleinschmidt and Gelbart (2002) use local Gabor filters to approximate the spectro-temporal response field (STRF) of the human auditory cortex. More recently, Heckmann et al. (2011) learn a set of local features from the data, and combine this in a hierarchical framework to obtain features spanning a larger frequency and time region. Other works have applied image-processing techniques to the spectrogram for speech, sound and music environments (Schutte, 2009; Dennis et al., 2011; Matsui et al., 2011). However, these approaches often just extend frame-based techniques, apply methods directly from image processing, or combine block-based techniques with SVM to classify the whole spectrogram. The Hough transform has also been used previously for tasks such as localisation (Marchand et al., 2009) and word spotting (Barnwal et al., 2012), although in such works it is typically used to detect straight lines, and not general shapes as we perform here with the GHT.

The rest of the paper is organised as follows. Section 2 describes the LSF extraction. Section 3 details the clustering and modelling to train the system. Section 4 describes the recognition approach based on the GHT. Section 5 then details our experiments, before Section 6 concludes the work.

2. Local Spectrogram Feature Extraction

In this section, we describe the extraction of Local Spectrogram Features (LSFs), from the two-dimensional spectrogram representation of the sound. Here we use the log-power Short-Time Fourier