Applied Acoustics 83 (2014) 57–63

Contents lists available at ScienceDirect

Applied Acoustics

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

Classification of audio events using permutation transformation

S. Fagerlund *, U.K. Laine

Department of Signal Processing and Acoustics, School of Electrical Engineering, Aalto University, P.O. Box 13000, FI-00076 AALTO, Finland

ARTICLE INFO

Article history:
Received 19 April 2013
Received in revised form 6 March 2014
Accepted 10 March 2014
Available online 12 April 2014

Keywords:
Stop consonant recognition
Temporal fine structure
Permutation transformation
Feature extraction
Audio event detection
Pattern recognition

ABSTRACT

Automatic detection and classification of short and nonstationary events in noisy signals is widely considered to be a difficult task for traditional frequency domain and even time–frequency domain approaches. A novel method for audio signal classification is introduced. It is based on statistical properties of the temporal fine structure of audio events. Artificially generated random signals and unvoiced stop consonants of speech are used to evaluate the method. The results show improved recognition accuracy in comparison to traditional approaches.

1. Introduction

Most present audio signal analysis and classification methods operate in the frequency domain. Generally, attempts to analyze signals in the time domain have not attracted much interest. This is largely due to the possible sensitivity of these signals to the phase variations caused by distorting factors such as time varying channel parameters. Since variation in the waveform due to signal phase variation is often considered irrelevant, the analysis method should not react to this property. Another undesirable characteristic of time domain methods, such as autocorrelation, is that larger amplitude values have a greater impact on the result than smaller amplitude values. However, many signals, e.g. in biology, have scale invariant nature where the absolute amplitude values are not important. Instead, the temporal fine structure of the signal may be much more informative [1]. Our approach to signal analysis and representation is based on the relative amplitudes of the samples – on their temporal ordering.

The audio event is considered nonstationary if spectrum is changing as function of time. In audio signal classification feature extraction is typically applied to the short-term frames. Frame length is selected such that signals are sufficiently stationary within the frame, typical not less than 10 ms because shorter frame lengths are insufficient with traditional methods. However many audio signals indicate highly non-linear characteristics in very short time intervals [2]. Especially speech, music and many environmental audio signals are often considered nonstationary [2,3].

Accurate analysis in the frequency domain has requirements that cannot generally be fulfilled with nonstationary signals. In order to perform analysis in the frequency domain, the signal should remain sufficiently stationary within the analysis window. However, the stationarity assumption is not satisfied for many time series, e.g., when analyzing many natural audio signals or certain parts of speech [2–5]. Frequency domain analysis also requires signals of sufficient length. However, the length of some audio events can be considerably shorter than the analysis window that is often optimized according to the average time-scale of the signal structures, leading inaccuracy in signal analysis.

Studies on ordering of events or observations have received relatively little attention. Early studies concerning ordering of events have been focused on correlation of ordered observations made by two or more observers [6]. Recently, more attention has focused on signal analysis and pattern recognition in time series based on temporal structures [7–9]. These studies have resulted in a wide number of applications, including the detection of abnormalities in aircraft engines [10], analysis of EEG and EMG signals [11–13], and measurement of the complexity in time series [14]. One aspect common to these applications is that the signal events of interest are nonstationary and/or short in duration. Further Arroyo et al. [15] used successfully permutation entropy to detect transient and nonstationary dynamics in neuronal activity signals.

Amigó et al. [16] applied topological permutation entropy to discover deterministic signal structures hidden in white noise. They summarized that noisy signal with a weak deterministic component, that is wrongly classified as white noise (null hypotheses) by standard methods like autocorrelation or BDS algorithm, can be separated from the pure white noise signals by using ordinal

* Corresponding author. Tel.: +358 405210896.
E-mail address: Seppo.Fagerlund@aalto.fi (S. Fagerlund).

http://dx.doi.org/10.1016/j.apacoust.2014.03.006
0003-682X/

patterns and entropy measure. The reason to this effect is that the distribution of ordinal patterns of a white noise is more uniform than that of a noisy deterministic signal. The hidden deterministic component can be revealed based on comparison of the distributions.

In this paper, we introduce a novel time-domain method, permutation transformation, for the classification of nonstationary audio signals that can have diverse and noisy time structure. This type of signals are common in e.g. environmental audio signals, music or some parts of speech [2]. In this work we limit only to the audio signals but method could be applied to many other signals too. We assume that essential structural information of the auditory events can be identified in very short analysis windows, especially within nonstationary signals. Information carried in the temporal fine structure of the signals is coded and preserved during the transformation. Thus, for an analysis time window consisting of five samples, this coding would provide the space of different permutations (sample orderings) equivalent to $5! = 120$. A time domain signal can be transformed into a sequence of codes, (e.g. integers in the range of 1–120), with each of these representing a reference to the corresponding permutation. The sequence of codes can be statistically processed further. To accomplish this, a histogram of connected (subsequent) index pairs is applied for modeling the temporal fine structure of the original signal.

In order to demonstrate the efficiency of the method, it is first applied to the classification of artificially generated noise signals having equal spectra but different temporal structure. The method is then evaluated in the classification of stop consonant bursts. In speech recognition, stop consonants are generally considered difficult to classify with current automatic speech recognition algorithms [17] due to their short duration and nonstationary temporal structure [17,26]. Most studies on consonant recognition are based on spectral information concerning the burst and formant information of the following vowel [18–22]. Furthermore, it is generally assumed that the phonetic context has a large impact on the stop consonant properties and that it provides even more information for stop consonant classification than the consonant burst itself. On the other hand, Nathan and Silverman [23] have argued that no clear consensus exists concerning what are the best features for recognizing unvoiced stop consonants. Similarly, Halle et al. [24] noted as early as 1957 that the burst part has an important role in the classification of stop consonants. One recent study shows that stop consonants can be recognized from the burst part only [25]. This study uses sub-band energy features of the burst part in addition to formant features. With sub-band features recognition accuracy of 82.2% was achieved and combined with formant features accuracy increased to 84.4%. Thasleema et al. proposed the time domain approach for classification of Malayalam consonants [26]. The method reconstructs a state space model from consonants. Classification results shows that the method can extract non-linear characteristic of consonants and encourage to use of time domain methods for classification of nonstationary signals.

This paper is organized as follows. Section 2 describes the permutation transformation method and a novel method for matrix smoothing that helps to alleviate the problem of data sparsity. An example with artificial noise signal detection and classification is presented in Section 3. Section 4 describes experiments with stop consonant classification, and Section 5 presents the results for stop consonant classification. Finally, Section 6 concludes the paper.

2. Methods

Permutation of a set is an ordered arrangement of its elements. A set consisting of n different elements can be ordered in $n!$ different ways and thus it has $n!$ different possible permutations. In this work the amplitude values of the signals are ranked in a short time window. Ranking of amplitude values in each window is coded by a permutation code index forming a symbolic sequence. Symbols or permutation codes can be seen as pointers to the corresponding ranking (permutation). In order to represent the structure of a given signal, a permutation pair frequency (PPF) matrix can be derived from a sequence of permutation codes. The PPF matrix is distribution of permutation code pairs in the signal. In many practical cases, the PPF matrix may be sparse. However, the sparseness of the PPF matrix can be alleviated by smoothing it with a novel spatial filtering method. It is shown in the experiments that the smoothing process improves the robustness of the permutation transformation in classification of audio events. Fig. 1 shows a block diagram to build smoothed PPF matrix of the auditory event.

2.1. Permutation transformation

Continuous amplitude values in the permutation window are thus quantized simultaneously with permutation transformation. A new permutation code index is created for each moment of time. Rank numbers depend on relative amplitude values of samples only, and thus permutation transformation is a *scale free* method to *perform local waveform* quantization of a signal. New signal consists of $n!$ different structural quantization levels, where n is the size of the permutation window in time samples.

The method starts by dividing the signal into permutation windows. Rank numbers replaces signal amplitude values in each window. Permutation of real valued signal $x(t)$ at the location t is denoted by π_n^τ where τ is time delay between the analyzed signal samples (not necessary equal to one) and n is size of the permutation window. Formally, permutation of a time series is defined as

$$\pi_n^\tau = \begin{pmatrix} 1 & 2 & 3 & \cdots & n \\ r_1 & r_2 & r_3 & \cdots & r_n \end{pmatrix} \tag{1}$$

satisfying

$$x_{t-r_1\tau} \geqslant x_{t-r_2\tau} \geqslant \cdots \geqslant x_{t-r_{n-1}\tau} \geqslant x_{t-r_n\tau} \tag{2}$$

where r is corresponding rank number of the sample value within the permutation window. When $\tau = 2$, the permutation window consists of every second sample in the original time domain signal up to the used window length. Equal amplitude values of the original signal within the permutation window are assumed to be very rare but if they occur we define $r_1 > r_{1-t}$. Fig. 2 illustrates a permutation window (black continuous line) of an arbitrary signal with time delay $\tau = 1$ and $\tau = 2$.

Each permutation pattern is coded with its symbolic reference (in this work one of the $n!$ permutation indices) and they form a

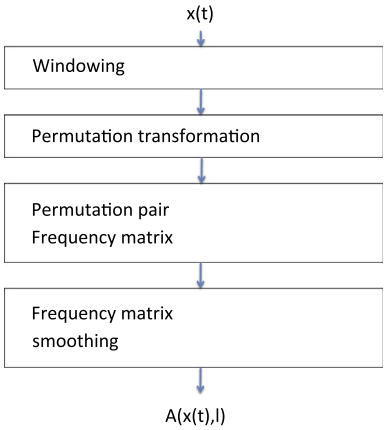

Fig. 1. Block diagram of the method.