Applied Acoustics 135 (2018) 124–135

Contents lists available at ScienceDirect

Applied Acoustics

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

Audio-cough event detection based on moment theory

Jesús Monge-Álvarez[a,*], Carlos Hoyos-Barceló[a], Keshav Dahal[a,c], Pablo Casaseca-de-la-Higuera[a,b]

[a] Centre for Artificial Intelligence, Visual Communications, and Networks, School of Engineering and Computing, University of the West of Scotland, United Kingdom
[b] Laboratorio de Procesado de Imagen, ETSI Telecomunicación, Universidad de Valladolid, Spain
[c] Nanjing University of Information Science and Technology (NUIST), China

ARTICLE INFO

Keywords:
Audio event detection
Cough segmentation
Moment theory
k-Nearest neighbours
Time-frequency analysis

ABSTRACT

Cough detection has recently been identified as of paramount importance to fully exploit the potential of telemedicine in respiratory conditions and release some of the economic burden of respiratory care in national health systems. Current audio-based cough detection systems are either uncomfortable or not suitable for continuous patient monitoring, since the audio processing methods implemented therein fail to cope with noisy environments such as those where the acquiring device is carried in the pocket (e.g. smartphone). Moment theory has been widely applied in a number of complex problems involving image processing, computer vision, and pattern recognition. Their invariance properties and noise robustness make them especially suitable as "signature" features enabling character recognition or texture analysis. A natural extension of moment theory to one-dimensional signals is the identification of meaningful patterns in audio signals. However, to the best of our knowledge only marginal attempts have been made in this direction. This paper applies moment theory to perform cough detection in noisy audio signals. Our proposal adopts the first steps used to extract Mel frequency cepstral coefficients (time-frequency decomposition and application of a filter bank defined in the Mel scale) while the innovation is introduced in the second step, where energy patterns defined for specific temporal frames and frequency bands are characterised using moment theory. Our results show the feasibility of using moment theory to solve such a problem in a variety of noise conditions, with sensitivity and specificity values around 90%, significantly outperforming popular state-of-the-art feature sets.

1. Introduction

Cough is a symptom associated with over one hundred medical conditions like respiratory diseases (e.g., asthma, bronchiectasis, or chronic obstructive pulmonary disease), generic pathologies such as cold or allergies or even lifestyle (smokers) [1]. Respiratory conditions constitute a significant burden for national health systems and economies [2,3], with clear potential to be released if objective continuous monitoring of symptoms such as cough was made possible. Consequently, cough detection has recently been identified as of paramount importance to fully exploit the potential of telemedicine in respiratory conditions and thus decrease their economic burden [4].

Audio cough events are non-stationary signals presenting a sparse spectrum that exhibits a high-energy peak around 400 Hz and a secondary peak between 1 and 1.5 kHz. Detecting and properly characterising them is hindered by a lack of a clear pitch structure [5]. Moreover, there exist other events produced by the human body such as throat clearing, gasping breath or laugh whose acoustic properties are very similar. Also, a continuous monitoring environment (e.g. when carrying a smartphone in the pocket) can prevent accurate detection due to the presence of noise with diverse spectral content.

Audio event detection (AED) was originally posed as a binary classification problem to differentiate speech from non-speech events. Such systems are commonly known as voice activity detectors [6]. Later, the generalisation of other information sources such as audio/video streaming, musical repositories or online video-games [7–9] led to other applications involving non-speech signals: query-by-humming, recommender systems or automatic music transcription [10]. The signal processing and machine learning techniques applied within this new context are often referred to as *machine hearing* [11]. Moreover, the emergence of new devices – e.g. smartphones, tablets or *wearables* – with their increasing computational capabilities diversified the variety of applications requiring AED [6]. These new applications were initially focused on content-based audio classification and retrieval [10,12]. However, nowadays there is an increasing number of applications such as medical telemonitoring [13], ambient sound recognition [14], or audio surveillance (e.g. monitoring of wildlife areas [15] or classification of aircraft noise [16]).

* Corresponding author at: School of Engineering and Computing, Paisley Campus, University of the West of Scotland, High St, Paisley PA1 2BE, United Kingdom.
E-mail addresses: jesus.monge@uws.ac.uk, jsmonge@outlook.es (J. Monge-Álvarez).

https://doi.org/10.1016/j.apacoust.2018.02.001
Received 25 July 2017; Received in revised form 21 December 2017; Accepted 2 February 2018

Despite the fact that automatic detection and analysis of speech are still active research areas [17], their methods are not always directly applicable to other AED problems [6] due to two main reasons. First, speech or music repositories are in general larger than other databases that are more difficult to record or whose production is less frequent [18]. This point often leads to suboptimal or unfeasible applications of speech/music-specific methods to these smaller-sized datasets. Second, the differences in acoustic (e.g., formant frequencies or pitch contour) and spectral (e.g. distribution or spread) properties among different audio signals play an important role. Many of these methods are specifically designed on the basis of speech properties [17]. Thus, their application to other types of audio events such as acoustic biomedical signals or environmental sounds does not always produce satisfactory results [6].

A number of papers have addressed the problem of automatic cough detection from different perspectives. Commercial cough detectors achieve sensitivity values in the 80% range by employing features extracted not only from cough sounds but also from chest movement [19,20]. Matos et al. employed a keyword-spotting approach based on a hidden Markov model. Their average detection rate was 82% [21]. Drugman used mutual information-based measures and feature synchronisation to perform feature selection and classification for cough segmentation. Sensitivity and specificity values above 90% were reported [22].

Other authors have designed specific methods for cough segmentation. You et al. employed non-negative matrix factorisation, reaching sensitivity and specificity values around 85% [23]. They also proposed an ensemble multiple frequency subband features approach where recall values around 74% with an overall 82% performance were reported [24]. Finally, deep learning methods based on convolutional neural networks (CNN) and recurrent neural networks (RNN) have recently been used as well. Amoh and Odame achieved 83% sensitivity using this approach, although the CNN was superior (93%) in terms of specificity where the RNN only achieved 75% [25].

A number of approaches aiming at robust identification of audio events rely on interesting principles that could be adopted for cough detection. These approaches as such could only be applied to cough identification and not to cough detection, since they all work on previously segmented events of interest. Foggia et al. employed a *bag-of-audio-word* approach aimed at improving the discriminative power while the classification scheme is kept simple [26]. Dennis et al. developed spectrogram image features (SIF) for sound event classification. The spectrogram is normalised into grey-scale, and its dynamic range is quantised into regions before partitioning it into blocks whose distribution statistics are extracted to build a feature set for classification. The main disadvantage of this approach is the large dimension of the feature set (486) [27]. This drawback was partially solved by Sharan and Moir who improved the basic SIF approach for robust audio surveillance. They reduced the feature space dimension to 216 by computing the mean and standard deviation of the distribution statistics across rows and columns [28]. Other time-frequency representations have also been employed. *Cochleagram* image-based feature computation has found usage in speech recognition and audio separation applications [6]. The Wavelet transform, has also been used for speech and music discrimination since it provides better time and frequency localisation [6]. Finally, unsupervised classification approaches for environmental noise signal classification have recently been proposed [29].

Although extensively applied in image processing and computer vision, moment-based methods are still marginal in one-dimensional signal processing. These methods hold a number of features that make them suitable for cough detection due to their 2D nature. As image processing methods, they can be applied to windows including a time-frequency representation of the signal (e.g., spectrogram, *cochleagram*, as in the robust methods described above) and exploit this higher dimensionality to achieve robust cough detection. Recently, Sun et al. employed features based on local Hu moments (HUm) for speech emotion recognition [30]. Our previous work [13] showed that a similar approach could be successfully used to perform robust detection of audio-cough events. To the best of our knowledge, only the extensions of HUm in [30] and [13] have been applied so far to audio signal processing. The two examples described above show the promising applicability of moment theory for cough detection in particular and more generally for audio processing.

This paper proposes a novel methodology to extend moment theory to audio signal processing with a specific application to cough detection in noisy environments.[1] The individual pattern discrimination capability and robustness against noise of different moment families are studied and a discussion on the hyper parameter settings and design decisions in the methodology is presented. Our results show that using audio features based on moment theory significantly outperforms popular state-of-the-art feature sets such as Mel frequency cepstral coefficients (MFCC) [6] and linear predictive cepstral coefficients (LPCC) [31], especially in low Signal to Noise Ratio (SNR) scenarios. We also show that our method overcomes more noise-robust feature sets such as spectral subband centroid histograms (SSCH) [32] and power normalised cepstral coefficients (PNCC) [17]. It is worth highlighting at this point that in our context, cough detection is understood as a continuous process carried out while the signal is recorded in a *soft* real-time manner. We do not hold the assumption that the events are pre-segmented in different classes for further automatic classification as Audio Event Identification methods require.

The paper is structured as follows: Section 2 introduces the proposed methodology, including a description of the taxonomy of the different moment families selected for the study. Section 3 presents the experimental setup, including the design of the employed cough database, and the performance measurements used to evaluate the proposal. The experimental results are presented in Section 4 and discussed in Section 5. Both sections validate the proposal by justifying the adopted methodology against previous approaches and studying its sensitivity with respect to different parameter configurations and design choices. Section 6 finalises the paper with some conclusions and future directions.

2. Proposed methodology

2.1. Extension of moment theory to audio event detection

Many audio processing features are based on the spectral energy distribution of the acquired signals. For non-stationary signals, some type of Short Time Fourier Transform (STFT) computation provides a time-frequency decomposition to account for this spectral distribution along time. To obtain such representation, a filter bank is built to characterise the spectrum in several frequency bands. As an example, MFCC, one of the most widely used feature sets, employs the Mel frequency scale to set the limits of each filter in the filter bank. Once the filter bank is applied, the logarithm is computed for all energy values to obtain a representation close to the response of the human cochlea. This is the starting point of the proposed methodology, presented in the following paragraphs.

In order to build a time-frequency distribution, the one-sided normalised power spectral density ($PSD_k[f]$, $k = 1, \ldots, K$) is first estimated for each window as the Fourier transform of the autocorrelation function according to the Wiener-Khinchin-Einstein theorem [33]. Secondly, the logarithm of the energies is computed for every window in a series of bands defined by a filter bank in the Mel scale:

[1] We will employ the broad term “noisy” to refer to conditions in which unwanted signals overlap with the audio event of interest, regardless of their random or deterministic nature.