Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

An improved time–frequency noise reduction method using a psycho-acoustic Mel model

Samir Ouelha [a,*,1], Abdeljalil Aïssa-El-Bey [b], Boualem Boashash [a]

[a] Qatar University, Department of Electrical Engineering, Doha, Qatar
[b] IMT Atlantique, UMR CNRS 6285 Lab-STICC, Université Bretagne Loire, F-29238 Brest, France

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Time–frequency analysis
Psycho-acoustic model
Noise reduction
Signal enhancement
Wavelet thresholding
Mel filterbank

ABSTRACT

This paper addresses the problem of noise reduction in non-stationary signals. The paper first describes a human physiology based time–frequency (TF) representation (TFHP) using Mel filterbanks. It is then used to improve a noise reduction algorithm that does not require any *a priori* information about the signal of interest and the noise. This algorithm is efficiently implemented using an original wavelet shrinkage method. The overall method results in an original TF denoising procedure that yields a denoised TFHP (DTFHP). From this representation one can reconstruct a denoised time-domain signal and therefore define a new improved noise reduction algorithm, whose performance is evaluated and compared with other state-of-the-art methods. The performance assessment uses several criteria: (1) signal-to-noise-ratio (SNR), (2) segmental SNR (SSNR) and (3) mean square error (MSE). The results indicate an improvement of up to 4.72 dB with respect to SNR, 2.79 dB w.r.t. SSNR and 4.72 dB w.r.t. MSE for a speech database signals corrupted with four different noises. In addition, other applications such as EEG signal enhancement show promising results.

1. Introduction

Most real signals are non-stationary, however traditional time-domain or frequency-domain representations are inadequate to analyze such signals because they assume the signal as stationary [1]. Instead, one can use joint time–frequency (t, f) representations as they were found to be better to process such signals. Two family of time–frequency (TF) methods have been widely used in the state-of-the-art: (1) linear TF and (2) quadratic TF representations [1–3]. Linear methods such as short-time Fourier transform (STFT) are the most used in practice because they are cross-terms free (when components are spaced enough in the TF domain [1, Section 4.1]) and computationally efficient [4]. The main drawback of these types of representations are their poor resolution performance. Quadratic methods have shown improved resolution performance but generally they required the setting of several parameters to obtain a good trade-off between resolution performance and cross-terms suppression [1]. Therefore, it could be difficult for a non-expert to get the best TF representation; in addition optimal parameters are generally signal dependent, therefore such methods are not suitable for an automatic classification system (e.g. automatic speech recognition). To overcome the last limitation, signal dependent kernel methods have been developed with automatic parameters selection [5,6], however these methods are not computationally efficient for long duration signals (e.g. speech signals).

Another difficulty for the processing of real signals is that they are generally corrupted by noise. In many applications, such as geophysics [7,8], EEG abnormalities detection [9] or speech recognition [4,10], efficient signal enhancement techniques are required [11]. In the open literature, there are several methods available to suppress noise that depend on the knowledge of characteristics of the useful signal and/or the noise. Some algorithms require *a priori* knowledge about the signal and noise second order statistics [12], while others only require knowledge of the noise spectral density (e.g. Wiener filtering) [13]. Unfortunately, in real applications such information is not available and must be estimated [14]. Other studies made the assumption of the noise being Gaussian or sub-Gaussian in order to use wavelet based denoising approaches [15,16]. This is a rough assumption, as in real-life there are various noise sources [17]. Furthermore, in mobile communications, the signal of interest is speech and it often arises from conversations that take place in noisy and nonstationary environments such as inside a car, in the street, or inside airports. In such case, there is no justification to assume Gaussian noise. Therefore, noise reduction methods based on this *ideal* assumption may likely fail

[*] Corresponding author.
E-mail addresses: samir_ouelha@hotmail.fr (S. Ouelha), abdeljalil.aissaelbey@imt-atlantique.fr (A. Aïssa-El-Bey), boualem.boashash@gmail.com (B. Boashash).

[1] Please contact the first author to get the Matlab code for reproducing the experiments done in this paper.

https://doi.org/10.1016/j.dsp.2018.04.005
1051-2004/© 2018 Elsevier Inc. All rights reserved.

in real life applications [18]. Many authors proposed modeling the noise, but these techniques are application dependent and cannot be used in different situations [19–21].

This paper describes an improved denoised TF representation and blind noise reduction method that performs well without prior information about the signal and noise. The proposed TF representation is based on a psycho-acoustic TF model and it deals effectively with the non-stationarity of signals and noise. It is based on the finding that the basilar membrane inside the cochlea can be conceived as a bank of band-pass filters that have logarithmically increasing bandwidth [22]. In this study, a Mel filterbank is used to construct the resulting TF representation as it has shown promising results in modeling the human cochlea [22]. Some of the material presented in this paper is an extension and refinement of the findings in [23,24]; the main contribution of this study is to design an improved algorithm for noise variance estimation with performance supported by extensive experimental comparisons.

This paper is organized as follows; Section 2 reviews the main principles of the TF representation based on Mel filters called HPTF. Section 3 describes a method to reduce noise in the HPTF. After that, Section 4 discusses reconstructing the signal of interest from the denoised HPTF (DHPTF). Section 5 presents experiments and discusses the results. Finally, section 6 concludes the study and summarizes the main findings.

2. HPTF representation

2.1. Principle

Previous studies observed that the human ear acts like filters, which are concentrated only on certain frequencies [25]. Mel filterbank is a psychoacoustic model which represents how humans perceives the sound [22]. With respect to bandwidth, these Mel filters are non-uniformly spaced on the frequency axis, with more filters in the low frequency regions and less in high frequency regions. With respect to shape, the magnitude of Mel filters transfer functions $H_m(f)$ are triangular shaped filters with respect to the Mel scale. This scale is given by the following formula for a given frequency f in Hz [22]:

$$\text{mel}\,(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

Thus, the Mel frequency scale is almost linear below 1000 Hz and logarithmic above. If we consider M Mel filters, $H_m(f)$, each of them is centered on a frequency f_m, for $m = 2, 3, \ldots, M-1$, and has a bandwidth $B(m)$ defined as follows:

$$B(m) = f_{m+1} - f_{m-1}, \quad \forall m = 2, 3, \ldots, M-1. \tag{2}$$

The center frequency f_m is calculated from its corresponding center frequency on the Mel scale using the following inverse formula obtained from Eq. (1):

$$f_m = 700\left(10^{\frac{\text{mel}\,(f_m)}{2595}} - 1\right), \tag{3}$$

where:

$$\text{mel}\,(f_m) = \frac{m}{M+1}\left(\text{mel}\,(f_{max}) - \text{mel}\,(f_{min})\right), \; \forall m = 1, 2, \ldots, M, \tag{4}$$

where f_{max} and f_{min} correspond respectively to the highest and the lowest frequencies of the input signal (generally $f_{min} = 0$ and $f_{max} = \frac{F_s}{2}$, where F_s is the sampling frequency). The impulse response $h_m(t)$ that corresponds the Mel filter $H_m(f)$ can then be expressed as:

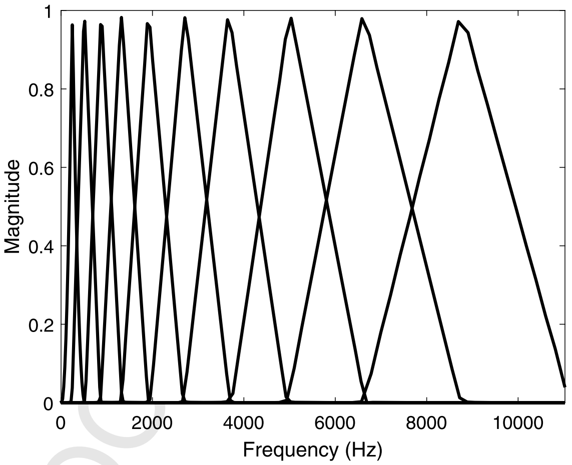

Fig. 1. Representation of the magnitude transfer functions of Mel filterbank $H_m(f)$ $\forall m = 1\ldots 10$ with $M = 10$.

$$\begin{aligned} h_m(t) &= \int_{-\infty}^{\infty} H_m(f)\, e^{j2\pi\, ft}\, df \\ &= \frac{1}{2\pi^2 t^2}\left(\frac{\cos(2\pi t f_{m-1}) - \cos(2\pi t f_m)}{f_{m-1} - f_m}\right. \\ &\quad \left. + \frac{\cos(2\pi t f_{m+1}) - \cos(2\pi t f_m)}{f_m - f_{m+1}}\right). \end{aligned} \tag{5}$$

Fig. 1 shows an example of Mel filter bank amplitude transfer functions for $M = 10$, $f_{min} = 0$ Hz and $f_{max} = 11025$ Hz, while Fig. 2 presents the impulse responses corresponding to $h_2(t)$ and $h_8(t)$ respectively.

2.2. HPTF construction

Let $\boldsymbol{z} \in \mathbb{R}^N$ be a vector of N samples containing data, obtained from an analog signal recorded by sensors and sampled at frequency F_s. This observation is a superposition of signal of interest $\boldsymbol{s} \in \mathbb{R}^N$ and noise $\boldsymbol{\epsilon} \in \mathbb{R}^N$:

$$\boldsymbol{z} = \boldsymbol{s} + \boldsymbol{\epsilon}. \tag{6}$$

The mth row of the HPTF shown in Fig. 3, denoted by $\boldsymbol{z}_m$, is the convolution between observation $\boldsymbol{z}$ and the sampled impulse response $\boldsymbol{h}_m$, $\forall\{m = 1\ldots M\}$ such that:

$$\boldsymbol{z}_m = \boldsymbol{z} * \boldsymbol{h}_m. \tag{7}$$

By using the linear property of the convolution, $\boldsymbol{z}_m$ equals the sum of the filtered signal of interest and the filtered noise, such that:

$$\boldsymbol{z}_m = \boldsymbol{s} * \boldsymbol{h}_m + \boldsymbol{\epsilon} * \boldsymbol{h}_m = \boldsymbol{s}_m + \boldsymbol{\epsilon}_m. \tag{8}$$

Eq. (7) and Eq. (8) correspond to a filtering process in the $H_m(f)$ bandwidth, where $H_m(f)$ is the Mel filter centered on the f_m frequency, according to Mel's scale (see Fig. 1). Therefore, $\boldsymbol{z}_m$ contains the spectral information of the input signal $\boldsymbol{z}$ around the frequency f_m, here expressed, for convenience, in the time-domain.

One can notice that the number of samples used to describe the impulse response $\boldsymbol{h}_m$ depends on the frequency f_m. Fig. 1 shows that $H_m(f)$ bandwidth is small for low frequencies, and conversely. As a consequence, the impulse response time support is smaller for high frequencies than for low frequencies; this is in accordance with the Heisenberg uncertainty principle [1, Chapter 2]. Hence, to