Low-Complexity Hybrid Time-Frequency Audio Signal Pattern Detection

Marco Martalò, Gianluigi Ferrari, and Claudio Santo Malavenda *Student Member, IEEE*

***Abstract*— In this paper, we present a low-complexity hybrid time-frequency approach for the detection of audio signal patterns by proper spectral signatures. The proposed detection algorithm evolves through two main processing phases, denoted as coarse and fine, respectively. The evolution through these two phases is described by a finite state machine model. The use of different processing phases is expedient to reduce the computational complexity and thus the energy consumption. Our results show that the proposed approach allows the efficient detection of the presence of signals of interest. The efficiency of the proposed detection algorithm is first investigated using "ideal" audio signals recovered from publicly available databases and then experimental audio signals acquired with a commercial microphone.**

***Index Terms*— Audio signal pattern detection, experimental validation, finite state machine (FSM), time-frequency processing.**

I. Introduction

In VARIOUS applications (such as, for example, surveillance, health care assistance, etc.) it is often of interest to determine the class which a specific audio signal belongs to. This problem is typically referred to as *sound recognition* [1]. Several approaches (often computationally intense) have been proposed in the literature and most of them rely on the analysis of the statistical properties of the audio signals [1]. In [2], the authors propose an audio detection and classification scheme based on machine learning techniques, which can outperform classical sound recognition schemes. In [3], the authors characterize the relevant spectral peaks of different audio patterns (for health care purposes) in order to perform the recognition task. In [4], an audio-based recognition system for gun shot detection is presented and its robustness against variable and adverse conditions is analyzed. Different time and frequency domain metrics for audio-based context recognition systems are analyzed in [5], comparing system performance with the accuracy of human listeners performing the same task.

The problem of *voice activity detection* (VAD) has also been widely studied. Unlike the previous problem of sound recognition, in this case one wants to detect the time intervals during which a (known) audio signal of interest (typically voice) appears, given that it will (sooner or later) appear for sure. A first possible strategy to detect the presence of an audio signal of interest, through a time domain-based analysis, consists of evaluating the energy of the audio signal samples, as in [6]. Another class of VAD algorithms is based on statistical analysis of the signal frames' spectra, obtained through the discrete Fourier transform (DFT) [7], as discussed in [8], [9].

In this paper, we focus on the following problem: detecting, with limited complexity, the presence *and* the pattern of an audio signal. In this sense, our problem is related to both sound recognition and VAD. As in sound recognition, we want to determine the type of a detected signal, *but* the processing complexity needs to be kept as low as possible. As in VAD, we want to determine the presence of a signal of interest, *but* its presence is not guaranteed. This problem is meaningful, for example, in wireless sensor networking surveillance scenarios, where nodes, typically battery-powered, need to recognize intruders and the node energy consumption is a critical issue [10]. The proposed algorithm has been considered in the implementation of the MasterZone wireless sensor network-based surveillance system produced by SELEX Sistemi Integrati [11]. Our approach relies on two main processing phases: (a) a *coarse* processing phase, carried out in the time domain, to detect the presence of an "atypical" signal and (b) a *fine* processing phase to verify, through a comparison with a predefined spectral signature, if the atypical signal is of the type of interest. To the best of our knowledge, the combination of time and frequency domain processing, typically kept separate in the literature (e.g., in [5]), is novel. The evolution between the two processing phases is described through a finite state machine (FSM) model. In particular, we first present simulation results, based on the use of "ideal" audio signals (available in public audio signal databases and acquired with sophisticated microphones), and then show the effectiveness of our approach with experimental audio signals acquired through a low-cost commercial microphone.

II. Training and Time Domain Energy Detection

Let us consider a discrete-time representation (with sampling frequency f_s) of the front-end of the audio sensor (i.e., the microphone), which can be represented by a proper linear filter. The output signal samples are denoted as $\{x_k\}$, with

$$x_k = \begin{cases} s_k + n_k & \text{in the presence of an atypical signal} \\ n_k & \text{in the absence of any atypical signal} \end{cases}$$

Manuscript received October 5, 2011; revised March 1, 2012; accepted September 6, 2012. Date of publication September 14, 2012; date of current version January 11, 2013. This work was supported by Elsag-Datamat S.p.A. (ED), Rome, Italy. The associate editor coordinating the review of this paper and approving it for publication was Prof. Julian W. Gardner.

M. Martalò and G. Ferrari are with the Department of Information Engineering, University of Parma, Parma 43124, Italy (e-mail: marco.martalo@unipr.it; gianluigi.ferrari@unipr.it).

C. S. Malavenda is with Selex Sistemi Integrati S.p.A., Rome 00012, Italy (e-mail: csmalavenda@selex-si.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSEN.2012.2219045

where s_k is the useful signal component and n_k is the noise sample. More precisely, n_k can be expressed as $n_k = n_{\text{mic},k} + n_{\text{env},k}$, where $n_{\text{mic},k}$ is the noise generated by the microphone (on the order of 100 nV/Hz$^{0.5}$ [12]) and $n_{\text{env},k}$ is the environmental audio noise. Typically, $n_{\text{mic},k} \ll n_{\text{env},k}$.

Our approach is based on per-frame processing, where a frame corresponds to a sequence of consecutive discrete-time samples. Denoting as K the number of samples per frame, the average per frame SNR can be defined as follows:

$$\text{SNR} \triangleq \frac{E_{\text{voice}}}{E_{\text{noise}}} = \frac{\dfrac{\sum_{i=1}^{K} |s_i|^2}{K}}{\dfrac{\sum_{i=1}^{K} |n_i|^2}{K}} = \frac{\sum_{i=1}^{K} |s_i|^2}{\sum_{i=1}^{K} |n_i|^2}. \tag{1}$$

Under the assumption that the noise is ergodic, its average energy E_{noise} at the denominator in (1) can be estimated during an initial training phase, when the background (noisy) audio signal is sensed but the system is still inactive for the purpose of pattern detection.

The presence of an audio signal (of interest) can therefore be identified by the "appearance" of an energy variation with respect to existing audio background noise. Therefore, one could first analyze, as in VAD approaches, the energies of consecutive audio signal frames in order to detect abrupt energy changes.[1] The basic principle consists in comparing the average energy of a frame with a proper threshold $E_{\text{th-initial}}$, which depends on the mean and variance of the background noise energy (denoted as μ_{low} and σ^2_{low}, respectively). Therefore, accurate estimation of the latter energy is fundamental and is the goal of a training phase.

Denoting as $N_{\text{tr-f}}$ the number of consecutive frames considered in the training phase and as $N^{\text{tr-s}}$ the number of samples per frame, the mean and the variance of the noise energy can be computed as follows:

$$\mu_{\text{low}} \triangleq \frac{1}{N_{\text{tr-f}}} \sum_{i=1}^{N_{\text{tr-f}}} \frac{1}{N^{\text{tr-s}}} \sum_{k=1}^{N^{\text{tr-s}}} \left| x_k^{(i)} \right|^2$$

$$\sigma^2_{\text{low}} \triangleq \frac{1}{N_{\text{tr-f}}} \sum_{i=1}^{N_{\text{tr-f}}} \frac{1}{\left(N^{\text{tr-s}} - 1\right)} \sum_{k=1}^{N^{\text{tr-s}}} \left(\left| x_k^{(i)} \right|^2 - \mu_{\text{low}} \right)^2.$$

Upon completion of the training phase, denoting as $\{x_k\}_{k=1}^{N^{\text{low-s}}}$ the $N^{\text{low-s}}$ samples in a generic collected frame, the following binary decision rule can be considered to determine the presence ($D_{\text{low}} = 1$) or absence ($D_{\text{low}} = 0$) of an "atypical" signal:

$$\frac{\sum_{k=1}^{N^{\text{low-s}}} |x_k|^2}{N^{\text{low-s}}} \underset{D_{\text{low}}=0}{\overset{D_{\text{low}}=1}{\gtrless}} E_{\text{th-initial}} \tag{2}$$

where $E_{\text{th-initial}} \triangleq \mu_{\text{low}} + \epsilon \sigma_{\text{low}}$, with the parameter $\epsilon > 0$ tuning the sensitivity in detecting the signals. Our results show that $\epsilon = 1$ leads to a good performance. Note that if $D_{\text{low}} = 0$ (i.e., no significant energy variation is detected), the average energy and the variance of the background noise can be adapted by taking into account the newly processed frame. In particular, the following adaptation rule can be used upon the reception of the ℓ-th frame ($\ell = 1, 2, \ldots$)

$$\begin{aligned} \mu_{\text{low}}(\ell + 1) &= \frac{\mu_{\text{low}}(\ell - 1) + \mu_{\text{low}}(\ell)}{2} \\ \sigma^2_{\text{low}}(\ell + 1) &= \frac{\sigma^2_{\text{low}}(\ell - 1) + \sigma^2_{\text{low}}(\ell)}{2} \end{aligned} \tag{3}$$

where $\mu_{\text{low}}(0) \triangleq \mu_{\text{low}}$ and $\sigma^2_{\text{low}}(0) \triangleq \sigma^2_{\text{low}}$. Note that the updates (3) are useful especially in the presence of noise characterized by a highly fluctuating variance, e.g., with pink or factory noise. In this case, when no signal of interest is identified in the "triggered" fine processing phase (described in Section III), the noise characteristics can be updated, to track the environmental changes.

Our approach can be extended to a more general scenario where the audio signal to be detected might be subject to filtering (i.e., to the presence of convolutional noise). Under the assumption of perfect estimation of the channel response, the proposed detection strategy can be then applied, as its training phase takes into account automatically the statistics of the filtered noise. In the case of unknown or time-varying channel impulse response, one should first consider channel estimation, but this goes beyond the scope of this paper.

By using the introduced energy-based processing (based on (2)), one can detect the presence of an atypical energy variation. However, our problem requires also to distinguish different audio signal patterns. Time-domain energy detection does not allow to do this, as the probability mass functions (PMFs) of the frame normalized energies[2] of very different signals tend to be very similar. This should be expected, since a VAD-inspired approach only allows to detect the time intervals where a signal of interest (e.g., the voice) is present, without giving any information about the "content" of the audio signal.

III. Spectral Signature Extraction and Audio Signal Pattern Recognition

Since VAD-inspired approaches are not sufficient to distinguish different audio signal patterns, we resort to speech recognition techniques [13]. More precisely, we apply the ideas behind these techniques, typically used to recognize different spoken words, to classify different audio signal patterns. In particular, our key idea is that of characterizing an audio signal frame with a spectral signature and then, through frequency domain processing, detect if the received audio signal matches with the signature.

Upon the collection of the sequence of the samples of a single frame, denoted as $\{x_k\}_{k=1}^{N^{\text{high-s}}}$, its DFT[3] $\{X(n)\}_{n=1}^{N^{\text{high-s}}}$ is computed

$$X(n) = \sum_{k=1}^{N^{\text{high-s}}} x_k \exp\left\{ -j \frac{2\pi}{N^{\text{high-s}}} kn \right\}.$$

[1] Note that one may also consider frequency-based VAD approaches [8], [9]. However, this approaches are not effective in the considered scenario, since they require the use of larger numbers of collected frames and, therefore, have higher computational complexity.

[2] Normalization is required to make the comparison meaningful.

[3] It is possible to efficiently compute the DFT through a fast Fourier transform (FFT) if $N^{\text{high-s}}$ is a power of 2 or by properly applying zero-padding. Note also that $N^{\text{high-s}}$ might, in general, be different from $N^{\text{low-s}}$ (introduced in Section II).

The sequence $\{|X(n)|^2\}$ is a particular instance of periodogram (in the absence of windowing between consecutive frames) associated with the sequence $\{x_k\}$ obtained by sampling the received audio signal with a sampling rate denoted as f_s^{high}. The sequence $\{|X(n)|^2\}$ thus represents an accurate estimate of the signal power spectral density [7]. Depending on the chosen technology, evaluating the periodogram mat be computationally heavy (because of the presence of the squares of the moduli of the DFT coefficients). In particular, this may happen in wireless sensor networking applications, where the computational load should be kept as low as possible in order to extend the battery's lifetime. Therefore, we simply consider, as a representative "spectral shape" of the audio signal frame, the sequence of the moduli of the DFT coefficients, i.e., $\{|X(n)|\}$. As the spectral shape depends on the particular SNR, in order to use the same spectral signature, regardless of the SNR, we consider the following heuristic "normalized" spectral shape:

$$|Y(n)| \triangleq \frac{|X(n)|}{\sum_{m=1}^{N^{\text{high-s}}} |X(m)|} \qquad n = 1, \ldots, N^{\text{high-s}} \tag{4}$$

such that the condition $\sum_{n=1}^{N^{\text{high-s}}} |Y(n)| = 1$ holds.

The key principle of the proposed approach is to compare the normalized spectral shape of the frames of the received audio signal with a proper reference spectral *signature* (with unitary energy) of a frame of the signal of interest: if there is a "good agreement" between them, then the detected signal is declared of interest. In order to implement this strategy, the spectral signature and the "agreement" criterion have to be properly identified. Note that the proposed spectral signature-based approach cannot be applied if the signature is not available. The identification of the spectral signature requires the availability of a sufficiently large number of frames of the audio signal pattern of interest. However, our results show that a "coarse" characterization of the spectral characteristics of the audio signal of interest (e.g., using a few frames) is sufficient to guarantee good performance.

In the presence of non-stationary audio signals (e.g., voices), our results have shown that the best choice is to emphasize the high energy frequency components of the audio signal of interest. To this end, the best spectral signature of an audio signal of interest is typically given by the *envelope* of the sequence of normalized spectral shapes of the available frames of the reference audio signal of interest. In the following, we propose an efficient (recursive) approach to the extraction of an envelope spectral signature.

Denote as $\{|Y_i(n)|\}_{i=1}^{N_{\text{frame}}}$ the normalized spectral shapes of the overall N_{frame} available frames of the reference signal of interest. For example, suppose that the reference signal of interest has a duration of 30 s and is acquired with a sampling frequency equal to $f_s^{\text{high}} = 4096$ Hz. If each frame has $N^{\text{high-s}} = 128$ samples, then $N_{\text{frame}} = 960$. The *partial* (spectral) envelope over m consecutive frames is defined as follows:

$$\mathscr{S}^{(m)}(n) \triangleq \max_{i=1,\ldots,m} |Y_i(n)| \qquad n = 1, \ldots, N^{\text{high-s}}$$

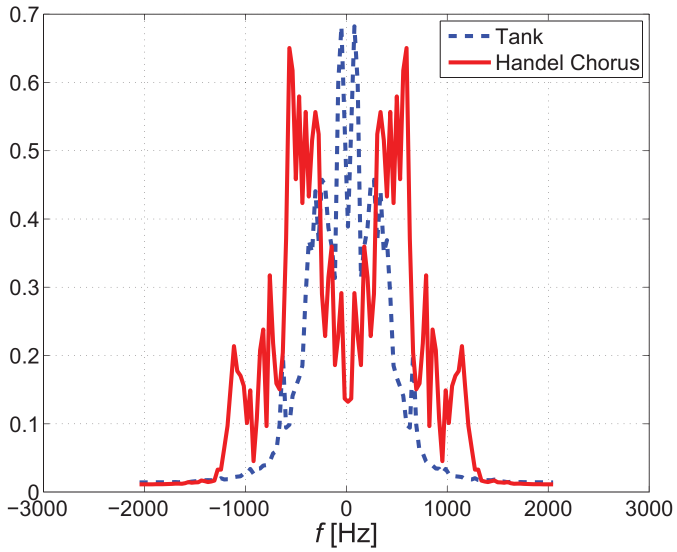


Fig. 1. Comparison between the spectral signatures of the tank and the Handel chorus signals.

where n corresponds to a frequency bin (for instance, when $f_s^{\text{high}} = 4$ kHz and $N^{\text{high-s}} = 128$, the bin width is approximately 32 Hz). In other words, the partial spectral envelope at the m-th step is obtained by considering the maximum, bin by bin, of the spectral shapes of all m collected frames. The *spectral signature*, denoted as $\mathscr{S}(n)$, is given by the partial envelope obtained considering all available frames, i.e., $\mathscr{S}(n) \triangleq \mathscr{S}^{(N_{\text{frame}})}(n)$, $n = 1, \ldots, N^{\text{high-s}}$. From an operative viewpoint, since the partial spectral envelope is a function which approaches (from below) the overall envelope (i.e., the signature), it is possible to recursively update it, upon reception of a new frame, as follows:

$$\begin{cases} \mathscr{S}^{(0)}(n) = 0 & \forall n \\ \mathscr{S}^{(m)}(n) = \max\left\{\mathscr{S}^{(m-1)}(n), |Y_m(n)|\right\} & \forall n \end{cases} \tag{5}$$

with $m = 1, \ldots, N_{\text{frame}}$. Since the spectral signature is normalized according to (4), the spectral envelope is an adimensional quantity. Moreover, the value of $\mathscr{S}(n)$ it can be interpreted as a percentage of the square root of the signal energy in the n-th bin.

As an illustrative example, we consider the following two discrete-time (ideal) audio signals: the audio signal emitted by a M109 vehicle (tank) moving at a constant speed of 30 km/h, with duration equal to 235 s and sampling frequency equal to 19.98 kHz, extracted from the NOISEX-92 database [14]; the audio signal of a choir singing the Handel's "Hallelujah Chorus," pre-loaded in MATLAB with sampling frequency equal to 8.192 kHz [15]. The sequences obtained with different sampling rates are downsampled to a common rate, denoted as f_s^{high}, so that they can be additively combined. The signatures are shown in Fig. 1 in the case with $N^{\text{high-s}} = 128$ samples per frame (using 128-point FFT). As one can see, the spectral envelopes are clearly different. This suggests that the two audio signal patterns may be successfully distinguished using the proposed spectral signature-based approach.

Once the spectral signature has been extracted, frequency domain audio pattern detection can be carried out as follows. Upon reception of a given number of frames of a (potentially of interest audio signal), its partial spectral envelope can be derived and compared with the signature. In particular, one can evaluate the mean linear error (MLE)

$$\mathrm{MLE}^{(m)} \triangleq \frac{\sum_{n=1}^{N^{\mathrm{high-s}}} \left| \mathscr{I}_{\mathrm{rx}}^{(m)}(n) - \mathscr{I}(n) \right|}{N^{\mathrm{high-s}}} \qquad (6)$$

where $\mathscr{I}_{\mathrm{rx}}^{(m)}(n)$ is the partial spectral envelope after processing m received frames. It can also be shown that, provided that the spectral signature $\mathscr{I}(n)$ is representative of all possible instances of the class of audio signals of interest (i.e., $\mathscr{I}(n) \geq \mathscr{I}_{\mathrm{rx}}^{(m)}(n)$, $\forall n, m$), $\mathrm{MLE}^{(m)}$ is a decreasing function of m. Therefore, the signal is declared of interest when the MLE becomes lower than a given threshold. The following (per frame) situations are then possible: correct detection (CD), if the MLE becomes lower than the threshold, given that there is the signal of interest; missed detection (MD), if the MLE does not become lower than the threshold, given that there is the signal of interest; false alarm (FA), if the MLE becomes lower than the threshold, given that there is not the signal of interest. The value of the MLE threshold can be chosen according to the behavior of $\{\mathrm{MLE}^{(m)}\}$, as will be discussed in more detail in Subsection V-A. This choice is crucial in order to optimize the performance of the proposed detection algorithm.

We remark that the complexity of the above frequency domain processing approach is comparable to other existing frequency domain-based algorithms (e.g., [5]).

IV. Low-Complexity Hybrid Time/Frequency Audio Signal Pattern Recognition

The frequency domain-based approach proposed in Section III does not take into account the energy content of the acquired signal, which can be evaluated in the time domain with a much lower computational complexity. In fact, when the signal of interest is not present in the acquired audio signal, the energy of the acquired signal coincides with that of the background noise. Therefore, one may exploit this idea to significantly reduce the computational complexity as follows. First, the *presence* of a possible signal of interest is detected, in terms of energy variation, by using the simple time domain processing described in Section II. Then, if an "atypical" signal is detected, its pattern is analyzed using the frequency domain processing technique described in Section III.

Taking into account possible correlations between consecutive frames, it is expedient to consider (as often done in VAD schemes [8]) a *hangover* FSM model, shown in Fig. 2, where the evolution between the state ($\mathrm{F}^{\mathrm{low}}$) associated with coarse processing and the state ($\mathrm{F}^{\mathrm{high}}$) associated with fine processing occurs through intermediate states. Every transition is a direct consequence of a single frame processing. In particular, the audio signal frames have fixed duration in each processing phase. Having fixed the frame duration, we denote as $N^{\mathrm{low-s}} = T_{\mathrm{frame}}^{\mathrm{low}} \cdot f_{\mathrm{s}}^{\mathrm{low}}$ and $N^{\mathrm{high-s}} = T_{\mathrm{frame}}^{\mathrm{high}} \cdot f_{\mathrm{s}}^{\mathrm{high}}$ the numbers of samples per frame in the coarse and fine processing phases, where $f_{\mathrm{s}}^{\mathrm{low}}$ and $f_{\mathrm{s}}^{\mathrm{high}}$ are the sampling rates in the two phases, respectively. The sampling frequency $f_{\mathrm{s}}^{\mathrm{low}}$ is low: namely, $f_{\mathrm{s}}^{\mathrm{low}} < 2 f_{\mathrm{Nyq}}$, where f_{Nyq} is the Nyquist frequency of the audio signal at hand. This choice is not critical, since in the coarse processing phase our goal is simply to detect abrupt energy changes, but not an accurate signal reconstruction. On the other hand, $f_{\mathrm{s}}^{\mathrm{high}}$ should be higher than $2 f_{\mathrm{Nyq}}$. However, in our experimental results we will consider a microphone with a slightly lower value of $2 f_{\mathrm{Nyq}}$. Our results show that this does not hinder the performance—recall that the pattern, rather than the specific signal, needs to be detected—yet allowing complexity reduction.

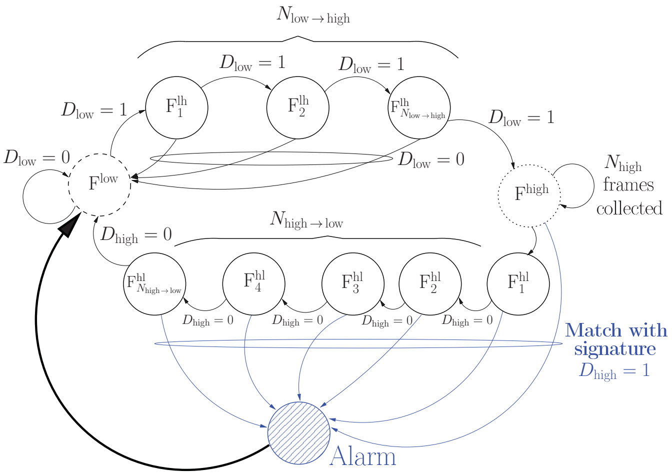

$N_{\mathrm{low}\to\mathrm{high}}$
$D_{\mathrm{low}} = 1$
$D_{\mathrm{low}} = 1$
$D_{\mathrm{low}} = 1$
$D_{\mathrm{low}} = 1$
$\mathrm{F}_1^{\mathrm{lh}}$
$\mathrm{F}_2^{\mathrm{lh}}$
$\mathrm{F}_{N_{\mathrm{low}\to\mathrm{high}}}^{\mathrm{lh}}$
$D_{\mathrm{low}} = 0$
$D_{\mathrm{low}} = 0$
$\mathrm{F}^{\mathrm{low}}$
$\mathrm{F}^{\mathrm{high}}$
N_{high} frames collected
$D_{\mathrm{high}} = 0$
$N_{\mathrm{high}\to\mathrm{low}}$
$\mathrm{F}_{N_{\mathrm{high}\to\mathrm{low}}}^{\mathrm{hl}}$
$\mathrm{F}_4^{\mathrm{hl}}$
$\mathrm{F}_3^{\mathrm{hl}}$
$\mathrm{F}_2^{\mathrm{hl}}$
$\mathrm{F}_1^{\mathrm{hl}}$
$D_{\mathrm{high}} = 0$
$D_{\mathrm{high}} = 0$
$D_{\mathrm{high}} = 0$
$D_{\mathrm{high}} = 0$
Match with signature $D_{\mathrm{high}} = 1$
Alarm

Fig. 2. FSM model for the proposed hybrid time-frequency audio signal pattern detection scheme.

The evolution of the proposed processing algorithm over the FSM can be described as follows. Most of time, the state of the FSM is $\mathrm{F}^{\mathrm{low}}$. After low-complexity processing (in time domain) of an $N^{\mathrm{low-s}}$-sample frame, a binary decision D_{low} on the presence of an atypical signal is taken: if $D_{\mathrm{low}} = 0$ (no atypical signal), the algorithm remains in $\mathrm{F}^{\mathrm{low}}$; if $D_{\mathrm{low}} = 1$, the algorithm evolves to the next intermediate state, denoted as $\mathrm{F}_1^{\mathrm{lh}}$, where low-complexity processing is considered. In general, one can consider $N_{\mathrm{low}\to\mathrm{high}}$ intermediate states $(\mathrm{F}_1^{\mathrm{lh}}, \ldots, \mathrm{F}_{N_{\mathrm{low}\to\mathrm{high}}}^{\mathrm{lh}})$ to evolve from $\mathrm{F}^{\mathrm{low}}$ to $\mathrm{F}^{\mathrm{high}}$. The use of the intermediate states is expedient to avoid useless and computationally intensive fine processing in the presence of impulsive noise, which may lead to short significant energy variations but, obviously, is not of interest. In the illustrative FSM model in Fig. 2, $N_{\mathrm{low}\to\mathrm{high}}$ is set to 3.

If for $N_{\mathrm{low}\to\mathrm{high}} + 1$ consecutive frames the presence of an atypical signal is verified (i.e., $D_{\mathrm{low}} = 1$), then the algorithm moves to $\mathrm{F}^{\mathrm{high}}$. In this state, a *fixed* number $N_{\mathrm{high-f}}$ (to be properly selected, as discussed in Subsection V-A) of frames, with $N^{\mathrm{high-s}}$ samples each, is collected. After processing the $N_{\mathrm{high-f}}$ frames in the frequency domain, as described in Section III, a binary decision D_{high} is taken: if $D_{\mathrm{high}} = 1$ (i.e., $\mathrm{MLE}^{(N_{\mathrm{high-f}})}$ is below threshold), then the signal pattern is declared of interest, a proper alarm is emitted, and the algorithm moves back to $\mathrm{F}^{\mathrm{low}}$; if $D_{\mathrm{high}} = 0$, then the algorithm moves to an intermediate state $\mathrm{F}_1^{\mathrm{hl}}$ and processes one more frame. At this point, if $D_{\mathrm{high}} = 1$, then the algorithm moves to $\mathrm{F}^{\mathrm{low}}$ and an alarm is emitted; otherwise, it moves to the next intermediate state $\mathrm{F}_2^{\mathrm{hl}}$. Eventually, if $D_{\mathrm{high}} = 0$ for $N_{\mathrm{high}\rightarrow\mathrm{low}} + 1$ consecutive frames (after exiting $\mathrm{F}^{\mathrm{high}}$), then the algorithm comes back to $\mathrm{F}^{\mathrm{low}}$ and no alarm is emitted: in other words, the atypical signal detected in the coarse processing phase is declared of no interest. The intermediate states $\{\mathrm{F}_1^{\mathrm{hl}}, \ldots, \mathrm{F}_{N_{\mathrm{low}\rightarrow\mathrm{high}}}^{\mathrm{hl}}\}$ from $\mathrm{F}^{\mathrm{high}}$ to $\mathrm{F}^{\mathrm{low}}$ can be interpreted as "back-up" states used to collect a larger number of frames to be fine processed, in order to improve the reliability of the decision on the presence of a signal of interest. In the illustrative example in Fig. 2, it holds that[4] $N_{\mathrm{high}\rightarrow\mathrm{low}} = 5$.

From a computational point of view, suppose that the signal of interest is present only in a fraction α of the N_{frame} collected frames (typically, $\alpha \ll 1$). If only frequency-based processing is performed, the total computational complexity can be quantified as follows:

$$\mathscr{C}_{\mathrm{tot}}^{\mathrm{F}} = N_{\mathrm{frame}}\,\mathscr{C}_{\mathrm{freq}}$$

where $\mathscr{C}_{\mathrm{freq}}$ is the computational complexity of frequency domain-based processing of a single frame. When, instead, the hybrid time-frequency approach is considered, the overall computational complexity becomes

$$\mathscr{C}_{\mathrm{tot}}^{\mathrm{T-F}} = \alpha N_{\mathrm{frame}}\mathscr{C}_{\mathrm{freq}} + (1-\alpha) N_{\mathrm{frame}}\mathscr{C}_{\mathrm{time}}$$

where $\mathscr{C}_{\mathrm{time}}$ is the computational complexity of time-based processing of a frame.

Since time domain processing consists in computing the per-frame average energy, its computational complexity (in terms of basic operations, e.g., summations and multiplications) is a linear function of the number $N^{\mathrm{low-s}}$ of per-frame samples. The complexity of the frequency-based processing, instead, is dominated by the signature comparison, which is quadratic in the number $N^{\mathrm{high-s}}$ of per-frame samples. Therefore, one obtains

$$\begin{aligned}\mathscr{C}_{\mathrm{tot}}^{\mathrm{F}} &= N_{\mathrm{frame}}(N^{\mathrm{high-s}})^2\\ \mathscr{C}_{\mathrm{tot}}^{\mathrm{T-F}} &= \alpha N_{\mathrm{frame}}(N^{\mathrm{high-s}})^2 + (1-\alpha)N_{\mathrm{frame}}N^{\mathrm{low-s}}\\ &\simeq \alpha N_{\mathrm{frame}}(N^{\mathrm{high-s}})^2\end{aligned}$$

where we have used the fact that, typically, $N^{\mathrm{low-s}} \ll N^{\mathrm{high-s}}$. After a few simple manipulations, the complexity reduction brought by the use of the hybrid time-frequency pattern detection algorithm is on the order of

$$\frac{\mathscr{C}_{\mathrm{tot}}^{\mathrm{F}}}{\mathscr{C}_{\mathrm{tot}}^{\mathrm{T-F}}} \simeq \frac{1}{\alpha} \gg 1.$$

This is intuitively expected, since the hybrid approach concentrates the complexity only in the presence of an atypical signal. If the atypical signal is of interest and appears for a fraction α of the time, then the complexity reduction is on the order of $1/\alpha$.

V. Performance Analysis

In the following, we assume that the human voice (e.g., the Handel chorus or experimentally acquired voices) is the signal of interest, whereas non-human voice audio signals (e.g., tank or car) are not. The results in a scenario where the tank/car signals are of interest and the human voice signal is not are not reported here for lack of space.

Just to summarize the obtained results, when the considered signals are acquired by quasi perfect microphones (i.e., ideal signals), the performance with the tank as signal of interest is very similar to that (shown in Subsection V-A) obtained with the human voice. This is due to the fact that, in the presence of ideal signals, a very good signature can be constructed in both cases (e.g., tank or human voice). On the other hand, when a non-ideal microphone commercial microphone is considered, the performance with the car as signal of interest worsens and false alarm increases with respect to considering human voices (as in Subsection V-B). This is due to the fact that the considered non-speech signal is more repetitive and, therefore, very similar to the background noise.

A. Ideal Audio Signals

The following set-up is considered. A "slice" of the audio signal of interest is 8 s long and is randomly additively combined with a background noisy audio signal of duration equal to 235 s and sampling frequency equal to 19.98 kHz, extracted from the NOISEX-92 database [14]. The following background noises (with the same sampling frequency) are extracted from the NOISEX-92 database [14]: (i) white noise, (ii) pink noise, typical of acoustic applications (e.g., durability or heat tests on loudspeakers or power amplifiers), and (iii) factory noise, typical of industrial environments. The spectral envelopes of pink and factory noises have strong components at low frequencies, whereas they rapidly decrease for higher frequencies. The filter that will first be considered, in our simulations, to process ideal audio signals is derived from a low-pass filter (LPF) of the commercial microphone and its matching circuit which will be used to collect realistic audio signals [12], as described in more detail in Subsection V-B. In order to approximate this LPF, we use a second-order Butterworth infinite impulse response (IIR) filter with a 3 dB bandwidth approximately equal to 2.2 kHz.

On top of the background noisy signal, a slice of another audio signal (with a spectral signature different from that of the signal of interest) is inserted. The two slices do not overlap: otherwise, our system would not be able to detect any of them. The training phase is carried out considering $N_{\mathrm{tr-f}} = 20$ frames of the background noisy signal. The sampling frequencies for the coarse and fine processing phases are $f_{\mathrm{s}}^{\mathrm{low}} = 1024$ Hz and $f_{\mathrm{s}}^{\mathrm{high}} = 4096$ Hz, respectively. As anticipated in Section II, the audio sequences (tank, Handel chorus, and background noise) obtained with different sampling rates are downsampled to $f_{\mathrm{s}}^{\mathrm{low}} = 1024$ Hz in the coarse processing phase[5] and to $f_{\mathrm{s}}^{\mathrm{high}} = 4096$ Hz in the fine processing phase. The numbers of samples per frame analyzed in the coarse and in the fine processing phases are $N^{\mathrm{low-s}} = 16$ and $N^{\mathrm{high-s}} = 128$, so that the frame durations in the two phases are equal to $T_{\mathrm{frame}}^{\mathrm{low}} \simeq 16$ ms and

[4] Typically, in the VAD literature $N_{\mathrm{high}\rightarrow\mathrm{low}} > N_{\mathrm{low}\rightarrow\mathrm{high}}$ [8].

[5] We consider a very small value of $f_{\mathrm{s}}^{\mathrm{low}}$ in order to reduce the computational complexity, thus saving as much battery energy as possible in wireless sensor network-based applications.

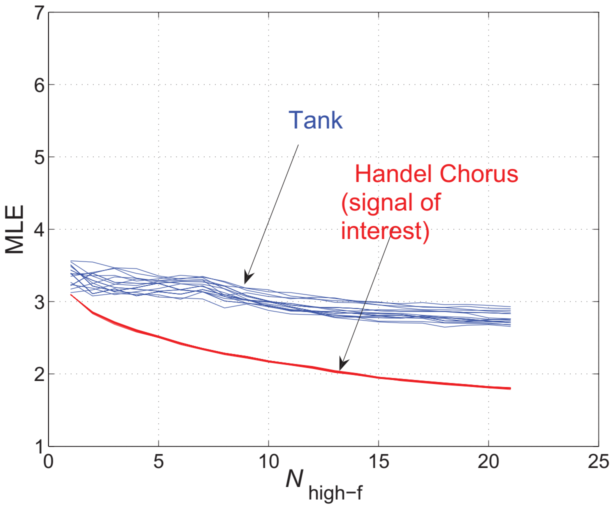


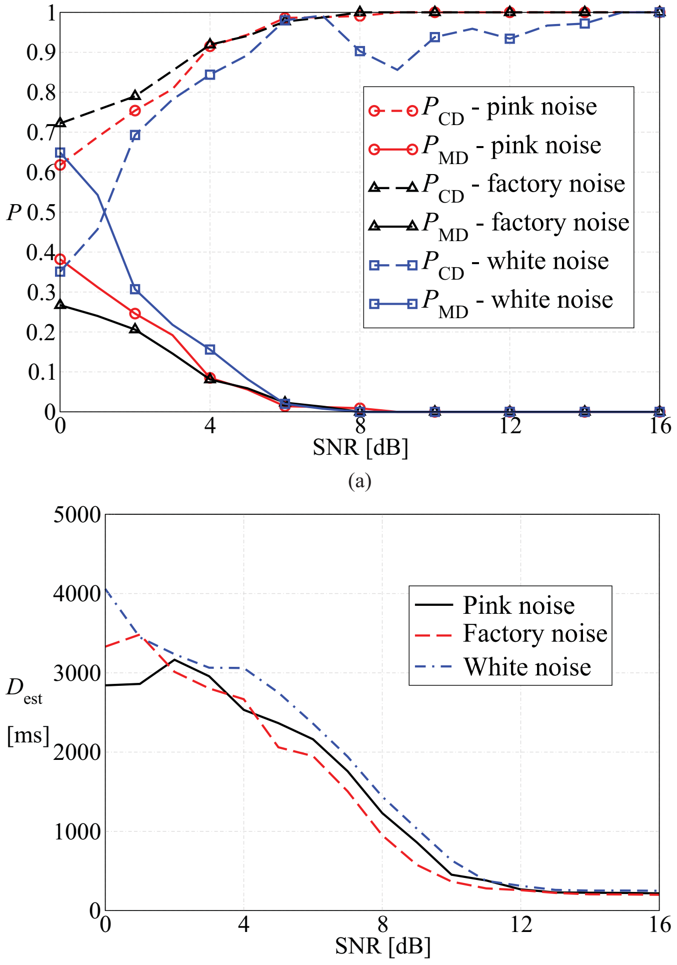
Fig. 3. MLE, as a function of the number of frames, in a scenario with AWGN background noise.

$T_{\text{frame}}^{\text{high}} \simeq 31$ ms, respectively. The numbers of intermediate states in the FSM have been heuristically set in the following way. More precisely, the value of $N_{\text{low}\to\text{high}}$ is set so that useless and computationally intensive fine processing in the presence of impulsive noise is avoided. The value of $N_{\text{high}\to\text{low}}$ is set as a tradeoff between the minimization of the false alarm probability and the detection delay. In particular, our results show that $N_{\text{low}\to\text{high}} = 3$ and $N_{\text{high}\to\text{low}} = 5$ are effective values.

As a first analysis step, we investigate the behavior of the MLE (between the partial envelope of the received signal and the spectral signature), as a function of number of collected frames, in a scenario with AWGN background noise. To this purpose, the SNR is set to 20 dB. In order to evaluate the system performance, we have performed 20 independent simulation runs (with random generation of disjoint initial time instants of the Handel chorus and tank audio signals). In Fig. 3, the MLE is shown as a function of the number of processed frames.

It is possible to observe that the set of curves associated with the signal of interest is lower than that associated with the signal of no interest. This behavior is pronounced also for a small number of frames: for instance, after 3 frames, the signals are easily separable. In other words, the proposed spectral signature-based detection approach is effective also when a few frames are collected and analyzed. From the results in Fig. 3, one can determine the number of frames $N_{\text{high}-\text{f}}$ which need to be processed, in the state F^{high} of the FSM, in order to reliably recognize the signal of interest. Simultaneously, the corresponding value of the threshold τ_{high} can be determined as a function of the selected value of $N_{\text{high}-\text{f}}$. For instance, if $N_{\text{high}-\text{f}} = 5$, then $\tau_{\text{high}} \simeq 2.7$. Reducing $N_{\text{high}-\text{f}}$, τ_{high} should increase. However, our results show that a higher value of τ_{high} makes the probability of FA increase dramatically. Therefore, for $N_{\text{high}-\text{f}} < 5$, the best performance is obtained with a "conservative" value of τ_{high} equal to 2.7.

In Fig. 4, (a) the probabilities of MD and CD and (b) the delay, in the presence of CD, are shown, as functions of the SNR, in the presence of white, pink, or factory background noise. For every value of SNR, 1000 independent simulation runs are performed, in order to eliminate possible statistical fluctuations. In all cases, $N_{\text{high}-\text{f}}$ and τ_{high} are set to 5 and 2.7, respectively. In Fig. 4(a), no result for the probability of FA is shown, since in all cases it is very small (on the order of 10^{-3} to 10^{-4}). One can note that approximately the same performance can be observed for both signals of interest. In particular, for sufficiently high values of the SNR (around 8 dB) the probability of MD goes to zero, whereas the probability of CD goes to 1. One may argue that this value is too high; however, it is possible to observe that for SNR $\geq$ 4 dB the probability of MD is already below 10%. Moreover, as it will be shown in the next subsection, the penalty with respect to "classical" frequency-based algorithms is limited. In general terms, one can consider that the detection algorithm is properly working when the probability of CD becomes significantly higher than the probability of MD. For instance, in the scenario considered in Fig. 4 the proposed detection algorithm becomes effective for SNR $\geq$ 7 dB. Note also the performance with pink or factory noise is better than that with white noise. This is intuitive, as the spectral signature of the Handel chorus does not have relevant peaks in correspondence to those of pink and factory noises.



(b)

Fig. 4. Performance of the proposed algorithm in scenarios with white, pink, or factory background noises. (a) Probabilities of MD and CD. (b) Delay, in the presence of CD, as functions of the SNR.

In Fig. 4(b), the delay (dimension: [ms]) is considered. The delay is evaluated only when there is CD, since, otherwise, it would not be meaningful. In fact, when the signal of interest is not detected, the state of the system continuously iterates, in the FSM, between the coarse and fine processing states.[6] For small values of the SNR, the delay is around 4 s and it would not be possible to detect signals with duration shorter than this maximum delay—recall that the entire duration of the signal "slice" of interest is 8 s. This is due to the large number of frames which are processed before the presence of an atypical signal is declared in the coarse processing phase. For large values of the SNR, instead, in 0.5 s the signal of interest if correctly detected, thus making the proposed algorithm almost real-time. One can observe that the delay depends only slightly on the number of processed frames.

The limiting lower value of the delay for large values of the SNR is due to the fact that, even in the presence of correct detection of the signal of interest, $N_{\text{low}\to\text{high}}+1$ frames (with low sampling frequency) and $N_{\text{high}-\text{f}}$ frames (with high sampling frequency) need to be processed, thus leading to the following minimum achievable delay:

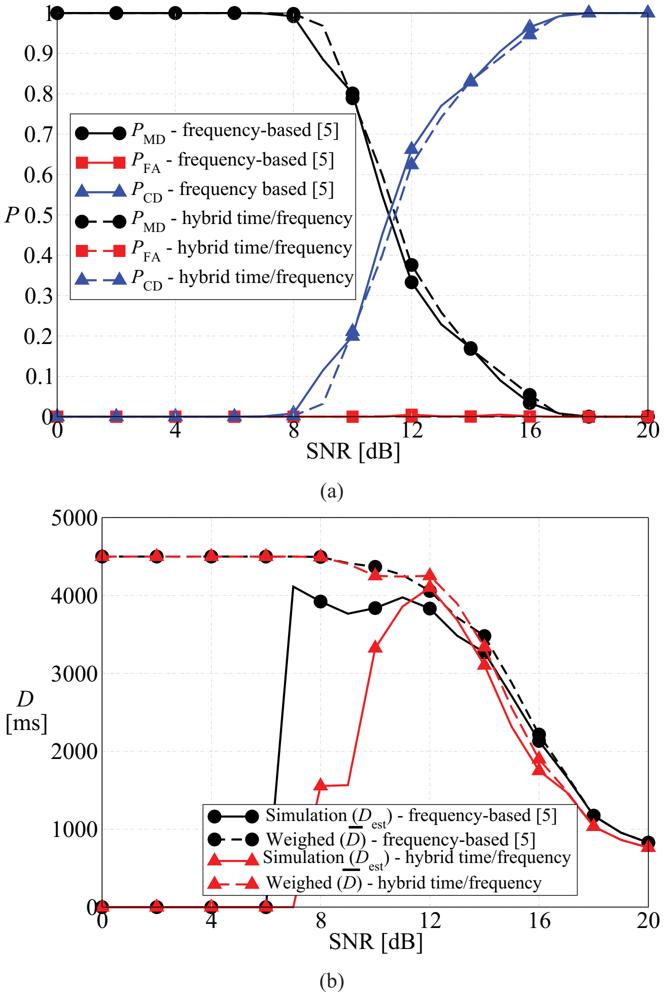
$$D_{\min} = (N_{\text{low}\to\text{high}} + 1)T_{\text{frame}}^{\text{low}} + N_{\text{high}-\text{f}}T_{\text{frame}}^{\text{high}}. \qquad (7)$$

Expression (7) holds for sufficiently large values of $N_{\text{high}-\text{f}}$ (e.g., $N_{\text{high}-\text{f}} = 5$). For $N_{\text{high}-\text{f}} = 1$, instead, the delay is slightly large, since more backup frames need to be processed before a spectral match is declared (i.e., till the MLE lowers below threshold). In other words, a single frame is not sufficient in F^{high} and it may happen that the system state starts moving back towards F^{low} before declaring a match.

B. Experimentally Acquired Audio Signals

We now analyze the performance of the proposed audio pattern detection algorithm in the presence of signals acquired through a realistic microphone characterized by a flat frequency response and with a sampling frequency equal to 3450 Hz [12], which has been used as the sampling frequency $f_{\text{s}}^{\text{high}}$ in the fine processing phase for all audio sequences. Moreover, the acquired audio sequences are further downsampled to $f_{\text{s}}^{\text{low}} = 1024$ Hz in the coarse processing phase and $N^{\text{high}-\text{s}} = 8$ to reduce the overall complexity. The other simulation parameters are set as described at the beginning of Subsection V-A. The use of proper filtering techniques at the output of the microphone, in order to improve the system performance, is investigated in [16].

The acquired speech signals used to derive the spectral signature correspond to the voices of 5 males (with ages between 23 and 35, at the University of Parma) reading some texts, both in Italian and in English. Since the manufacturer provides the microphone characterization only for frequencies higher than 100 Hz, the microphone behavior is unpredictable for frequencies below this threshold, although the matching circuit performance is known in this band [12]. Therefore, the signal components are highly distorted and in our analysis, with "realistic" acquired signals, we neglect the signal contributions below 100 Hz. To this end, in the fine processing phase the FFT of each frame is set to zero in the [0, 100] Hz band. As non-speech signal, we have acquired the sound emitted by the engine of a non-moving car (FIAT Punto, 1900 cc, turbo-diesel [17]) running at 3000 rpm. The duration of all acquired signals is around 30 s.



(a)

(b)

Fig. 5. Performance. (a) Probabilities of FA, MD, and CD. (b) Delay, in the presence of CD, as functions of the SNR. The frequency-based algorithm in [5] is compared to the proposed hybrid time-frequency algorithm.

The spectral signature for the human voice is obtained by computing the arithmetic average of the spectral envelopes associated with all persons (speaking in Italian or English). This choice is motivated by the fact that the system should be robust against possible variations of the signals to be detected, i.e., we ideally want to be able to detect all audio signals belonging to the same class (e.g., human voice) with a single spectral signature. The 8 s speech signal slices, randomly additively combined with a background white noise signal in the simulator, correspond to various people reading different scripts. By performing a preliminary MLE analysis, as in Fig. 3, it is possible to derive that $N_{\text{high}-\text{f}} = 15$ frames are needed to distinguish between different audio patterns. In this case, the MLE threshold τ_{high} (to be used in F^{high}) is set to 3.9.

In Fig. 5, (a) the probabilities of FA, MD, and CD and (b) the delay (in the presence of CD) are shown as functions

[6] One may consider a maximum number of iterations after which the system is reset.

of the SNR. Let us first comment on the performance of the proposed hybrid time/frequency algorithm. Comparing the results in Fig. 5(a) with those in Fig. 4, one can observe that the trends are similar, although the probability of CD reaches 1 for larger values of the SNR. Comparing the results in Fig. 5(b) with those in Fig. 4(b), the following observations can be carried out. Unlike in Fig. 4(b), in Fig. 5(b), for small values of the SNR, the delay decreases. This is due to the fact that the number of correct detections also reduces (according to the behavior of the probability of CD in Fig. 5(a)). In this case, a weighed average delay between the estimated delay (in the presence of CD) and a *pre-determined* maximum delay $D_{\max}$ (in the absence of CD), defined as $\overline{D} \triangleq D_{\max}(1 - P_{\text{CD}}) + D_{\text{est}} P_{\text{CD}}$, is more meaningful. In Fig. 5(b), a maximum delay $D_{\max} = 4$ s is considered. As expected, the $\overline{D}$ curves compare favorably with the delay curves in Fig. 4(b).

The proposed hybrid time/frequency algorithm is compared to the frequency-based detection algorithm in [5], with spectrum quantization in 8 sub-bands. It is possible to observe that there is a performance degradation with respect to the scenario presented in Fig. 5, due to the fact that the spectrum is quantized with a smaller number of points. One can note that the performance loss, in terms of probabilities of FA, MD, and CD, incurred by the hybrid approach is very limited (on the order of a fraction of dB). However, the delay with the proposed hybrid approach is lower than that with the frequency-based approach. This is due to the fact that in the latter case the system spends more time processing frames without energy atypicalities and this might delay the recognition of the pattern of interest.

VI. Conclusion

In this paper, we have presented a novel hybrid time-frequency approach to audio signal pattern detection. The algorithm relies on two main processing phases (coarse and fine) and the evolution between them is described by a hangover FSM model. The use of different processing phases is expedient to obtain a limited computational complexity. In fact, spectral signature-based fine processing (in the frequency domain and computationally heavy) is carried out a few times, namely only when an atypical (thus possibly of interest) signal is detected through coarse processing (in the time domain and computationally light). The performance of the proposed detection algorithm has been investigated, in terms of MLE, probabilities of CD/MD/FA, and delay, through simulations. Our results show that the proposed approach allows to efficiently detect the presence of an audio signal of interest, with significant complexity reduction and negligible performance loss with respect to frequency domain-based approaches. While this analysis has been first carried out using ideal audio signals (extracted from publicly available databases), we have then validated it through the use of realistic audio signals acquired with a commercial microphone.

Acknowledgment

The author would like to thank S. Mattiacci, L. Di Donato, and P. Proietti, who were with ED and are now with Selex Sistemi Integrati, for useful discussions on audio signal processing.

References

[1] A. Dufaux, "Detection and recognition of impulsive sound signals," Ph.D. dissertation, Inst. Microtechnology, Univ. Neuchatel, Neuchatel, Switzerland, 2001.

[2] A. Rabaoui, H. Kadri, Z. Lachiri, and N. Ellouze, "One-class SVMs challenges in audio detection and classification applications," *EURASIP J. Adv. Signal Process.*, vol. 2008, p. 14, May 2008.

[3] I.-C. Yoo and D. Yook, "Automatic sound recognition for the hearing impaired," *IEEE Trans. Consum. Electron.*, vol. 54, no. 4, pp. 2029–2036, Nov. 2008.

[4] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2005. pp. 1306–1309.

[5] A. J. Eronen, V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition," *IEEE Trans. Acoust. Speech Lang. Process.*, vol. 14, no. 1, pp. 321–329, Jan. 2006.

[6] S. G. Tanyer and H. Ozer, "Voice activity detection in nonstationary noise," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 4, pp. 478–482, Jul. 2000.

[7] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall, 1989.

[8] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Acoust. Speech Lang. Process.*, vol. 14, no. 2, pp. 412–424, Mar. 2006.

[9] J. W. Shin, J.-H. Chang, and N. S. Kim, "Voice activity detection based on a family of parametric distributions," *Pattern Recognit. Lett.*, vol. 28, no. 11, pp. 1295–1299, Aug. 2007.

[10] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Caryirci, "A survey on sensor networks," *IEEE Commun. Mag.*, vol. 40, no. 8, pp. 102–114, Aug. 2002.

[11] *MasterZone* [Online]. Available: http://www.selex-si-uk.com/pdf/Masterzone.pdf

[12] *Infineon SM310* [Online]. Available: http://www.infineon.com/

[13] J.-C. Junqua and J.-P. Haton, *Robustness in Automatic Speech Recognition: Fundamentals and Applications*. New York: Academic, 1995.

[14] *NOISEX-92, Noise Database* [Online]. Available: http://www.speech.cs.cmu.edu/comp.speech/Section1/Data/noisex.html/

[15] *The MathWorks — MATLAB* [Online]. Available: URL: http://www.mathworks.com/

[16] M. Martalò, G. Ferrari, and C. Malavenda, "Low-complexity in-sensor audio detection with experimental validation," in *Proc. IEEE Int. Symp. Ind. Electron.*, Bari, Italy, Jul. 2010, pp. 1674–1679.

[17] *Fiat Punto* [Online]. Available: http://en.wikipedia.org/wiki/FiatPunto

Marco Martalò was born in Galatina (LE), Italy, in 1981. He received the Bachelor's, Master's (*summa cum laude*), and Ph.D. degrees from the University of Parma, Parma, Italy, in 2003, 2005, and 2009, respectively.

He has been an Assistant Professor with the E-Campus University, Novedrate (CO), Italy, since 2012. He is a Research Associate with the Wireless Ad-hoc and Sensor Networks Laboratory, University of Parma. He was a Visiting Scholar with EPFL, Lausanne, Switzerland, from 2007 to 2008. He is involved in several research projects funded by public bodies and private companies. He is co-author of the book *Sensor Networks with IEEE 802.15.4 Systems: Distributed Processing, MAC, and Connectivity* (Germany: Springer, 2011). His current research interests include digital communication systems analysis and design and, in particular, wireless sensor networking.

Dr. Martalò was a co-recipient of the Best Student Paper Award at IWWAN in 2006 and the first prize award together with the WASNLab team at the First BSN Contest organized in conjunction with the BSN 2011 conference. He was a TPC member of WSNPERF 2009, SPACOMM 2009–2010, and GLOBECOM 2011 Communication Theory Symposium. He is a reviewer of many international journals and conferences.

Gianluigi Ferrari was born in Parma, Italy, in 1974. He received the Laurea and Ph.D. degrees from the University of Parma, Parma, in 1998 and 2002, respectively.

He has been with the University of Parma since 2002, where he is currently an Associate Professor of Telecommunications. He was a Visiting Researcher with the University of Southern California, Los Angeles, from 2000 to 2001, Carnegie Mellon University, Pittsburgh, PA, from 2002 to 2004, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, in 2007, and the Universitė libre de Bruxelles, Brussels, Belgium, in 2010. Since 2006, he has been the Coordinator of the Wireless Ad-hoc and Sensor Networks Laboratory, Department of Information Engineering, University of Parma. He has authored or co-authored more than 170 papers in leading international journals and conferences, and more than 10 book chapters, co-authored seven books, including *Detection Algorithms for Wireless Communications, with Applications to Wired and Storage Systems* (Wiley, 2004), *Ad Hoc Wireless Networks: A Communication-Theoretic Perspective* (Wiley, 2006-technical best seller), *LDPC Coded Modulations* (Springer, 2009), and *Sensor Networks with IEEE 802.15.4 Systems: Distributed Processing, MAC, and Connectivity* (Springer, 2011). He edited the book *Sensor Networks: Where Theory Meets Practice* (Springer, 2010). His current research interests include digital communication systems analysis and design, wireless ad hoc and sensor networking, adaptive digital signal processing. He participates in several research projects funded by public and private bodies.

Prof. Ferrari is a co-recipient of the Best Student Paper Award at IWWAN'06, the Best Paper Award at EMERGING'10, the Award for the Outstanding Technical Contributions at ITST-2011, and the Best Paper Award at SENSORNETS 2012. The WASNLab team won the first Body Sensor Network (BSN) contest, held in conjunction with BSN 2011. He is a frequent Reviewer of many international journals and conferences and a Technical Program Member of many international conferences. He is currently on the Editorial Boards of several international journals. He was a Guest Editor of the 2010 EURASIP JWCN Special Issue on "Dynamic Spectrum Access: From the Concept to the Implementation."

Claudio Santo Malavenda (S'12) was born in Fiesole, Italy. He received the Bachelor's and Master's degrees in electronic engineering (*cum laude*), specializing in automation systems and robotics, from the Sapienza University of Rome, Rome, Italy, in 2003 and 2006, respectively. He received the Master of Business Administration diploma in 2008. He is currently pursuing the Ph.D. degree in electronic engineering.

He was an Intern with ST Microelectronics, Catania, Italy. He joined Elsag-Datamat, Defense Division in 2007, which became Selex Sistemi Integrati in 2010. He is currently with the Tactical Command and Control Delivery Group. His current research interests include radio frequency hardware design to firmware and software design for scaling and self-organizing network protocols and secure embedded device development.

Dr. Malavenda was a recipient of the Best Patent Award in 2009 and the Special Mention in Selex Innovation Award in 2011, respectively. He was a Project Leader or the Chief Scientist of several European FP7 projects.