The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Detection of Acoustic Patterns by Stochastic Matched Filtering

Julien Bonnal, Patrick Danès and Marc Renaud

Abstract— **The detection of a pattern in an audio sequence is considered. An approach relying on the Stochastic Matched Filtering theory is proposed. It consists in first defining offline a basis from the statistics of the pattern and of the noise, then in isolating the pattern by means of a likelihood ratio test involving the online decomposition of the audio sequence on this basis. A simulated case study is proposed, which provides some guidelines to the tuning of the algorithm. Then, experimental results concerning the application of the method to voice activity detection are presented.**

I. INTRODUCTION

One of the important functions in Robot Audition is Voice Activity Detection (VAD). In a Human Robot Interaction context, this is a necessary preprocessing stage, which, when used upstream an Automatic Speech Recognition (ASR), improves the recognition performance. Exploiting the power of the sensed acoustic signal for VAD is unsuited to dynamic environments. This is the reason why other algorithms rely on the coupling of acoustic features and spatial selectivity [1][2], on detection algorithms dedicated to human voice [3][4]—which require a supervised learning stage from a basis of speech sequences—or on the fusion of sound cues with visual-based motion tracking of lips [5].

This paper aims at detecting an acoustic pattern—*e.g.*, though not necessarily, a voiced speech—in a noisy audio signal, by merging information on the power of this signal with its projection on a subspace synthesized offline from the statistics of the pattern. Its theoretical roots are formalized into the Stochastic Matched Filtering (SMF), first proposed in [6][7] for oceanography applications. It is proved well-suited to robotics because of its relatively low computational cost, and because the tuning of the parameters underlying its detection stage can be systematized.

The paper is organized as follows. In Section II, the stochastic matched filter (SMF) theory is presented, which constitutes the framework to detection. Section III proposes a study based on simulated random signals, whose theoretical statistics can be perfectly known. This way, some guidelines to the selection of the single free parameter of the algorithm, namely the detection threshold, are established. Then, strategies to the estimation of the pattern and noise statistics, which are required by the robotics context, are proposed in Section IV. The whole method is illustrated on real VAD experiments. A conclusion and open issues end the paper.

This work was supported by the French ANR AMORCES and the EU FP6-STREP CommRob projects.

J. Bonnal, P. Danès and M. Renaud are with CNRS; LAAS; 7 avenue du colonel Roche, F-31077 Toulouse, France; and with Université de Toulouse; UPS, INSA, INP, ISAE; LAAS; F-31077 Toulouse, France {patrick.danes,julien.bonnal,marc.renaud}@laas.fr

II. THE STOCHASTIC MATCHED FILTERING (SMF) THEORY

This section outlines the theory of stochastic matched filtering, which supports the detection algorithm.

A. Decomposition of Continuous-Time Random Signals

Consider a zero-mean random signal $Z(t) \in \mathbb{R}$ defined on a domain $\mathbb{D}_Z$, onto which it is stationary and ergodic. An infinite deterministic basis $\{\Psi_i(t)\}_{i\in\mathbb{N}^\star}$ can be defined in such a way that the coefficients $\{z_i\}_{i\in\mathbb{N}^\star}$ of the decomposition of $Z(t)$ form a zero-mean white discrete random sequence, *i.e.*

$$\forall t \in \mathbb{D}_Z,\ Z(t) = \sum_{i\in\mathbb{N}^\star} z_i \Psi_i(t) \tag{1}$$

$$\text{where } \forall i,j \in \mathbb{N}^\star,\ \mathsf{E}\{z_i\} = 0,\ \mathsf{E}\{z_i z_j\} = \delta_{i,j}\mathsf{E}\{z_i^2\}, \tag{2}$$

and $\delta_{i,j}$ stands for the Kronecker symbol. There exists an infinite number of $\{\Psi_i(t), z_i\}_{i\in\mathbb{N}^\star}$ such that the series (1) satisfies (2). They are united by

$$\forall\ t \in \mathbb{D}_Z,\ \forall i \in \mathbb{N}^\star, \Psi_i(t) = \frac{\mathsf{E}\{z_i Z(t)\}}{\mathsf{E}\{z_i^2\}}. \tag{3}$$

An infinite set of functions $\{\Phi_i\}_{i\in\mathbb{N}^\star}$ can also be introduced in order to express the random variables $\{z_i\}_{i\in\mathbb{N}^\star}$ as

$$z_i \triangleq \int_{\tau\in\mathbb{D}_Z} Z(\tau)\Phi_i(\tau)\mathrm{d}\tau. \tag{4}$$

Then, by defining $\Gamma_Z(t_1,t_2)$ as the autocovariance of $Z(t)$ at times t_1 and t_2—which, because of the stationarity of $Z(t)$, depends only on (t_2-t_1)—(3) can be turned into

$$\forall t,i,\ \Psi_i(t) = \frac{1}{\mathsf{E}\{z_i^2\}} \int_{\tau\in\mathbb{D}_Z} \Gamma_Z(t,\tau)\Phi_i(\tau)\mathrm{d}\tau. \tag{5}$$

Besides, in order to ensure the non-correlation property (2), the following conditions must be put on $\{\Phi_i\}_{i\in\mathbb{N}^\star}$:

$$\forall i,j \in \mathbb{N}^\star,\ \delta_{i,j}\mathsf{E}\{z_i^2\} = \iint_{t_1,t_2\in\mathbb{D}_Z} \Gamma_Z(t_1,t_2)\Phi_i(t_1)\Phi_j(t_2)\mathrm{d}t_1\mathrm{d}t_2. \tag{6}$$

Because of (5), the following holds,

$$\int_{t\in\mathbb{D}_Z} \Psi_i(t)\Phi_j(t)\mathrm{d}t = \frac{1}{\mathsf{E}\{z_i^2\}} \iint_{\tau,t\in\mathbb{D}_Z} \Gamma_Z(t,\tau)\Phi_i(\tau)\Phi_j(t)\mathrm{d}\tau\mathrm{d}t, \tag{7}$$

which, from (6), leads to the bi-orthogonality relationship

$$\int_{t\in\mathbb{D}_Z} \Psi_i(t)\Phi_j(t)\mathrm{d}t = \delta_{i,j}. \tag{8}$$

Importantly, the above equations are valid whatever the infinite pair $\{\Psi_i(t), z_i\}_{i\in\mathbb{N}^\star}$.

The Karhunen-Loève (K-L) decomposition [8] is a special case of (1)–(2)–(4) for which $\{\Phi_j(t)\}_{j\in\mathbb{N}^\star}$ are the eigenfunctions of the $\Gamma_Z(.,.)$ kernel, *i.e.*

$$\forall j \in \mathbb{N}^\star,\ \int_{\tau\in\mathbb{D}_Z} \Gamma_Z(t,\tau)\Phi_j(\tau)\mathrm{d}\tau = \lambda_j\Phi_j(t) \tag{9}$$

with $\{\lambda_j\}_{j\in\mathbb{N}^\star}$ the associated eigenvalues. In addition, $\{\Phi_j(t)\}_{j\in\mathbb{N}^\star}$ form an orthogonal basis w.r.t. the scalar product entailed in (4). Because of (5), the LHS term of (9) is equal to $\mathsf{E}\{z_j^2\}\Psi_j(t)$. So, for the special case of the K-L decomposition, $\Psi_j(t)$ and $\Phi_j(t)$ are equal up to a constant α.

B. The Discrete-Time Case

Consider a discrete-time stationary random sequence $Z\in\mathbb{R}^M$. Similarly to (1), it can be expanded as

$$Z = \sum_{i=1}^{M} z_i\Psi_i, \tag{10}$$

where $\{\Psi_i\}_{i\in\{1,\ldots,M\}}$ forms a deterministic basis of $\mathbb{R}^M$ and the random real scalar sequence $\{z_i\}_{i\in\{1,\ldots,M\}}$ is zero-mean and white. The discrete counterpart of (3) writes as

$$\forall i \in \{1,\ldots,M\},\ \Psi_i = \frac{\mathsf{E}\{z_i.Z\}}{\mathsf{E}\{z_i^2\}}. \tag{11}$$

A set of M functions $\{\Phi_i\}_{i\in\{1,\ldots,M\}}$ is introduced such that the sequence $\{z_i\}_{i\in\{1,\ldots,M\}}$ is obtained by the scalar product

$$\forall i \in \{1,\ldots,M\},\ z_i \triangleq Z^T\Phi_i. \tag{12}$$

The non-correlation property of $\{z_i\}_{i\in\{1,\ldots,M\}}$ leads to

$$\forall i,j \in \{1,\ldots,M\},\ \Phi_i^T\Gamma_Z\Phi_j = \mathsf{E}\{z_i^2\}\,\delta_{i,j}, \tag{13}$$

with $\Gamma_Z \triangleq \mathsf{E}\{ZZ^T\}$ the autocovariance matrix of Z. Eq. (5) and the bi-orthogonality property (8) are turned into

$$\forall i,j \in \{1,\ldots,M\},\ \Psi_i = \frac{1}{\mathsf{E}\{z_i^2\}}\Gamma_Z\Phi_i \text{ and } \Phi_i^T\Psi_j = \delta_{i,j}. \tag{14}$$

This stochastic decomposition will be shown to be useful in the case of a signal corrupted by noise.

C. Expansion of a Discrete-Time Signal into Noise

Consider a discrete-time M-length random signal S of interest, corrupted by an additive colored noise N. Both are assumed mutually independent, stationary, zero-mean and of respective standard deviations σ_S and σ_N. Define Z as

$$Z = S+N = \sigma_S S_0 + \sigma_N N_0, \tag{15}$$

with $\mathsf{E}\{S_0^2\} = \mathsf{E}\{N_0^2\} = 1$. Similarly to Sections II-A and II-B, it can be shown that there exist M-dimensional sets $\{\Psi_i\}_{i\in\{1,\ldots,M\}}$, $\{\Phi_i\}_{i\in\{1,\ldots,M\}}$ such that S_0, N_0, Z satisfy the following equations, with $\{s_i\}_{i\in\{1,\ldots,M\}}$, $\{n_i\}_{i\in\{1,\ldots,M\}}$, $\{z_i\}_{i\in\{1,\ldots,M\}}$ zero-mean scalar random sequences, the latter being white:

$$Z=\sum_{i=1}^{M} z_i\Psi_i,\ z_i=Z^T\Phi_i,\ \Psi_i=\frac{\mathsf{E}\{z_iZ\}}{\mathsf{E}\{z_i^2\}},\ \Psi_i^T\Phi_j=\delta_{i,j}, \tag{16}$$

$$S_0 = \sum_{i=1}^{M} s_i\Psi_i,\ N_0 = \sum_{i=1}^{M} n_i\Psi_i,\ s_i = S_0^T\Phi_i,\ n_i = N_0^T\Phi_i, \tag{17}$$

$$z_i=\sigma_S s_i+\sigma_N n_i,\ \mathsf{E}\{z_iZ\}=\Gamma_Z\Phi_i,\ \Gamma_Z=(\sigma_S^2\Gamma_{S_0}+\sigma_N^2\Gamma_{N_0}), \tag{18}$$

$$\mathsf{E}\{s_iS_0\} = \Gamma_{S_0}\Phi_i,\ \mathsf{E}\{n_iN_0\} = \Gamma_{N_0}\Phi_i. \tag{19}$$

Denote $\{\lambda_i,\Phi_i\}_{i\in\{1,\ldots,M\}}$ the generalized eigenvalues and eigenvectors of $(\Gamma_{S_0},\Gamma_{N_0})$, *i.e.* such that

$$\Gamma_{S_0}\Phi_i = \lambda_i\Gamma_{N_0}\Phi_i,\ i\in\{1,\ldots,M\}, \tag{20}$$

and assume that $\{\lambda_i\}_{i\in\{1,\ldots,M\}}$ are sorted in decreasing order and $\{\Phi_i\}_{i\in\{1,\ldots,M\}}$ are normalized so that $\Phi_i^T\Gamma_{N_0}\Phi_j = \delta_{i,j}$. Let Φ be the impulse response of a FIR filter, and $z = Z^T\Phi$ the result of the filtering of Z. The expected average signal-to-noise ratio (SNR) $\rho \triangleq \frac{\mathsf{E}\{|S^T\Phi|^2\}}{\mathsf{E}\{|N^T\Phi|^2\}} = \frac{\sigma_s^2}{\sigma_N^2}\frac{\mathsf{E}\{|S_0^T\Phi|^2\}}{\mathsf{E}\{|N_0^T\Phi|^2\}} = \frac{\sigma_s^2}{\sigma_N^2}\frac{\Phi^T\Gamma_{S_0}\Phi}{\Phi^T\Gamma_{N_0}\Phi}$ is maximum iff $\Phi=\Phi_1$. From (20) and (17), the following holds:

$$\frac{\Phi_i^T\Gamma_{S_0}\Phi_i}{\Phi_i^T\Gamma_{N_0}\Phi_i} = \frac{\mathsf{E}\{s_i^2\}}{\mathsf{E}\{n_i^2\}} = \mathsf{E}\{s_i^2\} = \lambda_i. \tag{21}$$

Then, truncating the expansion of Z as

$$Z_\Phi \triangleq \sum_{i=1}^{Q} z_i\Psi_i, \tag{22}$$

with Q such that $\lambda_1 \geq \ldots \geq \lambda_Q > 1 \geq \lambda_{Q+1} \geq \ldots \lambda_M$

defines a projection Z_Φ of Z which focuses on the signal and lowers the effect of noise. In addition, $\Psi_i = \Gamma_{N_0}\Phi_i$.

D. The SMF Strategy for Pattern Detection

Consider the discrete-time signal Z defined in (15) and its projection (22) on the Q-dimensional subspace generated by $\{\Psi_i\}_{i=1,\ldots,Q}$. The $(Q\times Q)$ covariance matrix of $z \triangleq [z_1,\ldots,z_Q]^T$ can be readily shown to write as

$$\Gamma_{z_\Phi} = \begin{bmatrix} \sigma_s^2\lambda_1+\sigma_N^2 & & 0 \\ & \ddots & \\ 0 & & \sigma_s^2\lambda_Q+\sigma_N^2 \end{bmatrix}. \tag{23}$$

The aim is to detect whether either the hypothesis [H_0:*Z is considered as noise*] or [H_1:*Z is the signal of interest S corrupted by noise*] holds. So, we introduce the probability density functions of z conditioned on H_0 and H_1. By denoting N_Φ the projection of the noise onto $\{\Psi_i\}_{i\in\{1,\ldots,Q\}}$, and

$$\Gamma_{n_\Phi} \triangleq \sigma_N^2\mathbb{I}_{Q\times Q} \tag{24}$$

the covariance of $\sigma_N[n_1,\ldots,n_Q]^T$, one gets

$$H_0:\ \begin{cases} Z=N \ \Leftrightarrow\ \mathsf{E}\{z_i^2\} = \sigma_N^2 \ \text{ and} \\ p(z|H_0) = \dfrac{1}{(2\pi)^{Q/2}\sqrt{|\Gamma_{n_\Phi}|}}\exp\left[-\dfrac{1}{2}z^T\Gamma_{n_\Phi}^{-1}z\right] \end{cases}$$
and
$$H_1:\ \begin{cases} Z=S+N \ \Leftrightarrow\ \mathsf{E}\{z_i^2\} = \sigma_S^2\lambda_i+\sigma_N^2 \ \text{ and} \\ p(z|H_1) = \dfrac{1}{(2\pi)^{Q/2}\sqrt{|\Gamma_{z_\Phi}|}}\exp\left[-\dfrac{1}{2}z^T\Gamma_{z_\Phi}^{-1}z\right]. \end{cases} \tag{25}$$

The detection itself is based on the likelihood ratio test (LRT) [8]

$$\Lambda(z) = \frac{p(z|H_1)}{p(z|H_0)} \underset{D_1}{\overset{D_0}{\lessgtr}} \xi, \tag{26}$$

where the threshold ξ can be theoretically related to the probability of false alarm. A trade-off in order to set ξ is explained in Section III. Noticeably, (26) is equivalent to

$$\sum_{i=1}^{Q} z_i^2 \frac{\sigma_S^2 \lambda_i}{\sigma_N^2(\sigma_S^2\lambda_i+\sigma_N^2)} \underset{D_1}{\overset{D_0}{\lessgtr}} \ln(\xi) + \sum_{i=1}^{Q} \ln\left(\frac{\sigma_S^2}{\sigma_N^2}\lambda_i + 1\right) \quad (27)$$

where the LHS term, depending on z_i, is assessed online, while the RHS term is computed offline.

The essentials of Stochastic Matched Filtering have been introduced. We now adopt a more practical point of view, by underlining the offline initialization and online stages.

E. The SMF Algorithm in Practice

The online part of the SMF is summarized into Algorithm 1. It essentially consists in computing the SMF basis, and in determining the LRT threshold ξ.

Algorithm 1: Detection based on Stochastic Matched Filtering - Part I (offline)

OFFLINE, **do**;
begin
1. estimate the signal and noise parameters $(\sigma_S^2, \Gamma_{S_0}, \sigma_N^2, \Gamma_{N_0})$;
2. compute the generalized eigenvalues and eigenvectors $\{\lambda_i, \Phi_i\}_{i\in\{1,\ldots,M\}}$ of the matrix pencil $(\Gamma_{S_0}, \Gamma_{N_0})$ along (20);
3. select Q such that $\lambda_1 \geq \ldots \geq \lambda_Q > 1 \geq \lambda_{Q+1} \geq \ldots \lambda_M$;
4. normalize each eigenvector Φ_i by setting $\widetilde{\Phi}_i \leftarrow \frac{\Phi_i}{\sqrt{\Phi_i^T \Gamma_{N_0} \Phi_i}}$;
5. determine a relevant value for ξ;
6. compute the right-hand side of (27) for further use;
end

The online steps are listed into Algorithm 2.

Algorithm 2: Detection based on Stochastic Matched Filtering - Part II (online)

ONLINE, **do**;
begin
7. compute the random variables $z_i = Z^T \widetilde{\Phi}_i$, then stack them into vector z;
8. deduce the left-hand side of (27);
9. compute the decision output according to (27);
end

Importantly, the parameters which condition the performance of the detection strategy are the characteristics of the signal of interest S (average power σ_S^2 and covariance matrix Γ_{S_0}), the characteristics of the noise (average power σ_N^2 and covariance matrix Γ_{N_0}), and the LRT threshold ξ.

III. A SIMULATED CASE STUDY

The purpose of this section is to evaluate the influence of the LRT threshold ξ entailed in Algorithm 1. First, we introduce a case study and its involved signals. Then, we define tools which can provide some guidelines to the selection of an effective value for ξ. Finally, we return to the practical aspect of the suggested method, and propose a comprehensive process well-suited to robotics.

A. The Considered Problem

As aforementioned, the parameters which influence the SMF are the statistics $\sigma_S, \Gamma_{S_0}, \sigma_N, \Gamma_{N_0}$ of the speech pattern and of the noise, as well as the decision threshold ξ. To study how to set ξ and how this affects the performances of the decision stage, we first consider simulated signals. This way, the exact analytical expressions of σ_S, σ_N, Γ_{S_0} and Γ_{N_0} are available, and errors due to their practical estimation in step 1 of Algorithm 1 are eliminated.

In order to synthesize mutually independent zero-mean stationary signals S and N, these are generated as the outputs from two separate discrete-time linear dynamic systems with independent initial conditions and excited by mutually independent white noises. Such systems are described by equations of the form —with $x \in \{S,N\}$—

$$x(k+1) = a_x x(k) + K_x(1-a_x)w_x(k),\ a_x \in [0;1[. \quad (28)$$

The independent—Gaussian—input noises and initial conditions are sampled from the distributions

$$w_x(k) \sim \mathcal{N}(0,1) \text{ and } x(0) \sim \mathcal{N}\left(0, \frac{K_x^2(1-a_x)^2}{1-a_x^2}\right). \quad (29)$$

So, the processes are stationary, zero-mean, and their variance and autocovariance write as

$$\sigma_x^2 \triangleq \mathsf{E}\left\{x(k)^2\right\} = \frac{K_x^2(1-a_x)^2}{1-a_x^2} \quad (30)$$

$$\Gamma_x(k+n) \triangleq \mathsf{E}\left\{x(k)x(k+n)\right\} = a_x^n \frac{K_x^2(1-a_x)^2}{1-a_x^2}. \quad (31)$$

Then, the signal Z is readily obtained by adding the signal of interest S and the colored noise N. Admittedly, these signals are not supposed to represent auditive signals in a Robotics context. Nevertheless, the degrees of freedom in the selection of a_S, a_N, K_S and K_N can be exploited so as to obtain a signal more correlated than the noise, with various signal to noise ratios (SNRs).

B. Towards Guidelines to the Tuning of the LRT Threshold

As aforementioned, the aim is to assess the impact of the LRT threshold ξ on the performances of the pattern detection. The probabilities of false alarm P_{FA} and of non-detection—or miss—P_M constitute basic classical metrics. P_{FA} is the probability that the algorithm detects the occurrence of the signal (H_1) while there is none, *i.e.* while H_0 is actually in effect. Contrarily, P_M is the probability to detect no pattern (H_0) although the signal of interest is in fact present (H_1). These can be summarized as

$$P_{FA} = p(\text{DECISION} = H_1 | Z = N), \quad (32)$$

$$P_M = p(\text{DECISION} = H_0 | Z = S+N). \quad (33)$$

Refs. [9][10] propose analytical expressions of P_{FA} and P_M for many Gaussian LRTs. However, in the considered case, depending on whether H_0 or H_1 is in effect, the expression of $\mathsf{E}\left\{z_i^2\right\}$ is either equal to σ_N^2 or to $\lambda_i\sigma_S^2 + \sigma_N^2$, respectively. As the $\{\lambda_i\}_{i=1,\ldots,Q}$ may be distinct, there seems not to exist any general closed-form expression of P_{FA} and P_M, and one

may be reduced to establish approximations. The accustomed reader can refer to [9][10] regarding this point. However, due to the complexity of analytical calculations, simulation based estimation of P_{FA} and P_M will henceforth be preferred.

From the signal Z described above, ξ is successively set to various values in order to get an experimental receiver operating characteristic (ROC) curve (see Figure 1), describing the performance of the LRT. This process is repeated for -6 dB and -3 dB SNRs so as to evaluate the influence of the environmental noise.

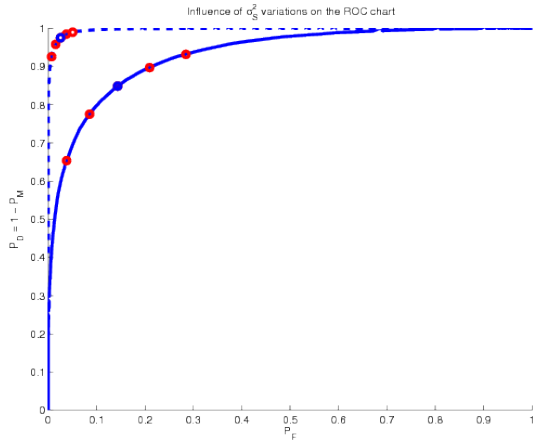

Fig. 1. ROC diagram for two simulated signals Z with -3 dB (dash line) and -6 dB (plain line) SNR. For the two signals, ξ is selected so as to obtain the better trade-off between P_{FA} and P_M (marked with blue dots). The effect of variations of $+/-20\%$ around the true σ_S^2 is pointed by red dots.

C. The Approach in Practice

Returning to Algorithm II-E and assuming that stages 1-4—estimation of the signal and noise parameters and computation of the SMF basis—are carried out offline, the aim is to select "the best" ξ in practice. The method we suggest consists in requiring from the user to record a "reference" sequence so as to tune the LRT. From this measurement of Z, the user can vary ξ in order to visually exhibit the zones of detection and non-detection of the pattern S out of Z. Indeed, in order to compute the probabilities P_{FA} and P_M, S should be segmented by hand from Z. This operation, though more rigorous, seems us quite tiresome and unsuitable for a quick setup/fast initialization, so that we prefer a representation such as Figures 3-4 with the detection zones that vary online depending on the value of ξ selected by the user.

IV. SMF BASED VOICE ACTIVITY DETECTION FOR ROBOTICS APPLICATIONS

A critical step of the algorithm is to estimate the statistics σ_S^2, σ_N^2, Γ_{S_0} and Γ_{N_0}, which condition the performances of the detection. In this section we attempt to propose a viable experimental procedure to be embedded on a robot. As σ_S^2 may vary in the audio sequence, its influence is taken into account, and detection results on robotics signals are shown.

A. About the Stationarity of the Signals

The assessment of the stationarity of a random signal, though studied for a long time, constitutes a complex problem. Indeed, experimentally, this property is tightly linked to the length of the window over which the signal is considered. In our case, this duration is set to M/f_s with $f_s = 15024$Hz the sampling frequency of the "Embedded Audition for Robotics" (EAR) sensor [11]. The purpose of this paragraph is to define the number of consecutive samples M for which the signals S and N can be considered as stationary.

In the contexts of speaker or speech recognition, the typical length of a time window used to study a voiced signal is about 20ms, which is related to the phoneme articulation by the vocal tract [12][13]. However, in a context of voice activity detection (VAD), much shorter signals can intervene. This can be compared with some methods of voiced signal segmentation where the window length is about $5-7$ms [14]. Based on these studies, we set $M = 100$.

Though M has been set based on some work in the field of audio signal processing, its value has been checked against the underlying assumption that the autocovariance function computed on an M-length sliding window should be constant all along the time sequence. Another representation, based on spectral components of the windowed signal has been derived from the log-spectra equation described in [15]. Other contributions to the field characterize the stationarity of a signal using tools originally used to study transitory signals, such as multiscale representations [16]. A further study on the subject is planned, which consists in testing the stationarity of the pattern and the noise under a wider window of study, and check its impact on the performance of SMF based detection.

B. Estimation of Γ_{S_0}

In addition to ensuring the signals stationarity, it is necessary to estimate a covariance matrix which is most representative of the pattern to detect. However, in the case of VAD, the focus is put on voiced speech. Many syllables can be portrayed as voiced [17] and it would be interesting to consider several different vowels, which have their own autocovariance function, see Fig. 2. However our algorithm relies on a single signal covariance matrix Γ_{S_0}. Future work will study how to detect voiced speech through a single SMF, rather than running several instances of the algorithm in parallel, each one being dedicated to a specific syllable.

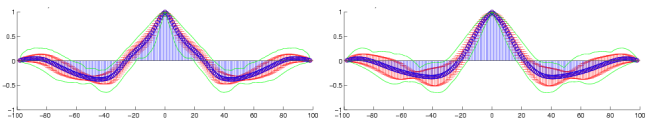

Fig. 2. Covariance matrix estimate $\widehat{\Gamma}_{S_0}$ for the two vowels [a] and [e] at the same pitch, together with standard deviations (in red).

In practice, the signal covariance matrix cannot be estimated online. Indeed, the training sequence S must be recorded in advance, in a quiet environment so that the quality of the estimate $\widehat{\Gamma}_{S_0}$ is not corrupted by noise. As for

the noise covariance matrix estimation, using a long sequence Z will tend to cancel the presence of intermittent non desired voiced speech during the acquisition. Indeed, $\widehat{\Gamma}_{N_0}$ is obtained by averaging the autocovariance matrix of several M-samples non-overlapped segments.

C. The Importance of σ_S

Up to now, estimates of the covariance matrices Γ_{S_0} and Γ_{N_0} were discussed. This paragraph concerns the practical estimations of σ_S and σ_N. Despite the expression Γ_{S_0} requires to be learned beforehand in order not to be corrupted by noise, the variance σ_S^2 can be estimated during the initialization phase. By assuming that the occurrences of the signal S and these of the noise N in the signal Z are conveniently segmented, one gets

$$Z_{init\ noise} = N \text{ and } Z_{init\ signal} = S+N, \tag{34}$$

which leads to the basic estimates

$$\widehat{\sigma_N}^2 = \frac{1}{T}\sum_{t=1}^{T} Z_{init\ noise}(t)^2 \tag{35}$$

$$\text{and } \widehat{\sigma_S}^2 \simeq \frac{1}{T}\sum_{t=1}^{T} Z_{init\ signal}(t)^2 - \widehat{\sigma_N}^2, \tag{36}$$

where T stands for the length of the signals $Z_{init\ noise}$ and $Z_{init\ signal}$.

To assess this basic approximation, a simulation was conducted in the aforementioned theoretical context, and the influence on the ROC chart of errors on $\widehat{\sigma_S}^2$ was measured. In this framework, the uncertainties on the noise and signal variances can be reduced by an appropriate choice of the LRT threshold. Indeed, although the ROC graphs are similar, the performance of the algorithm are very different with a same threshold ξ and different values of $\widehat{\sigma_S}^2$. On Figure 1, ξ is set as a good trade-off between P_{FA} and P_M, and marked with a blue dot. Setting the estimate $\widehat{\sigma_S}^2 = \beta\sigma_S^2$ with $\beta = [0.8, 0.9, 1.1, 1.2]$ and considering the same ξ, significantly distinct performances are obtained, pointed by red dots.

This experimentation, though carried out in simulation, shows the influence of errors in the estimation of σ_S^2. Moreover, the weaker the SNR, the more the performances of the LRT are affected. So, the estimate of σ_S^2 proposed above may be a limiting factor at low SNRs.

D. Experimentations

The following experiments rely on real speech signals acquired by the EAR sensor [11]. The noise is produced by a fan located in the vicinity of the speaker. Various SNRs are emulated.

Concerning the estimated parameters, Γ_{S_0} is estimated from learned occurrences of the pattern and Γ_{N_0} is estimated from the whole noise sequence. A particular effort was made to fit with the constraints related to robotics. Therefore, the estimation stage follows the aforementioned guidelines.

The forthcoming figures show the results of the VAD for various SNRs. We can see from Figure 3 the influence of ξ on the miss probability. For the first experimentation, although the threshold has been set to reduce P_M, some parts of the speech are not detected. As speech is not only composed of voiced signals [17], we conjecture that using a voiced pattern to estimate Γ_{S_0} affects the miss probability. Figure 4 exhibits a 5-segments sequence in which the noise remains constant and the SNR ranges between -6dB and 6dB. For this last experiment, the variance σ_S^2 has been estimated for a 0dB-SNR (middle segment). We can see that values of $\ln\{\Lambda\}$ (orange curve) vary significantly with the SNR, but remain very low when Z is essentially made of noise, hence the non-detection.

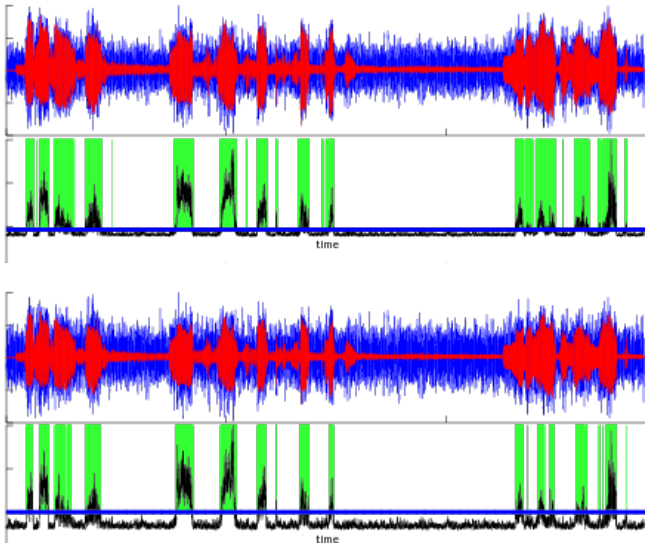

Fig. 3. Detection using the SMF algorithm for 3dB (top) and 0dB (bottom) SNRs. Blue and red plots respectively correspond to signals Z and S. The green areas correspond to the detection of [H_1:*Z is the signal of interest S corrupted by noise*], when $\Lambda(z)$ (black curve) is higher than ξ (blue line).

V. CONCLUSION

A new strategy to VAD, based on the Stochastic Matched Filtering, has been proposed. After its theoretical description, we assessed the influence of estimation errors of the considered signals statistics on the performances of the likelihood ratio test entailed in the detection. For an important range of uncertainties, the algorithm was able to detect an auditive pattern with a -6dB SNR. The downside is that its performances seem to fall quickly in an environment with dynamically changing SNR. So, a planned issue concerns the online re-estimation of some statistics, so as to enhance the robustness of the algorithm to environmental acoustic changes.

In addition, based on [14], the duration of signal analysis presented here was set to 6.6ms. A higher window size would increase the differences between Γ_{S_0} and Γ_{N_0}, and thus the discrimination of the projection of the genuine signal on the SMF basis. So, a more efficient stationarity test will be envisaged.

Finally, some portions of the pattern of interest were shown to be missed in experiments of Section IV. So, a work will be conducted in order to perform a better offline characterization of the pattern statistics.

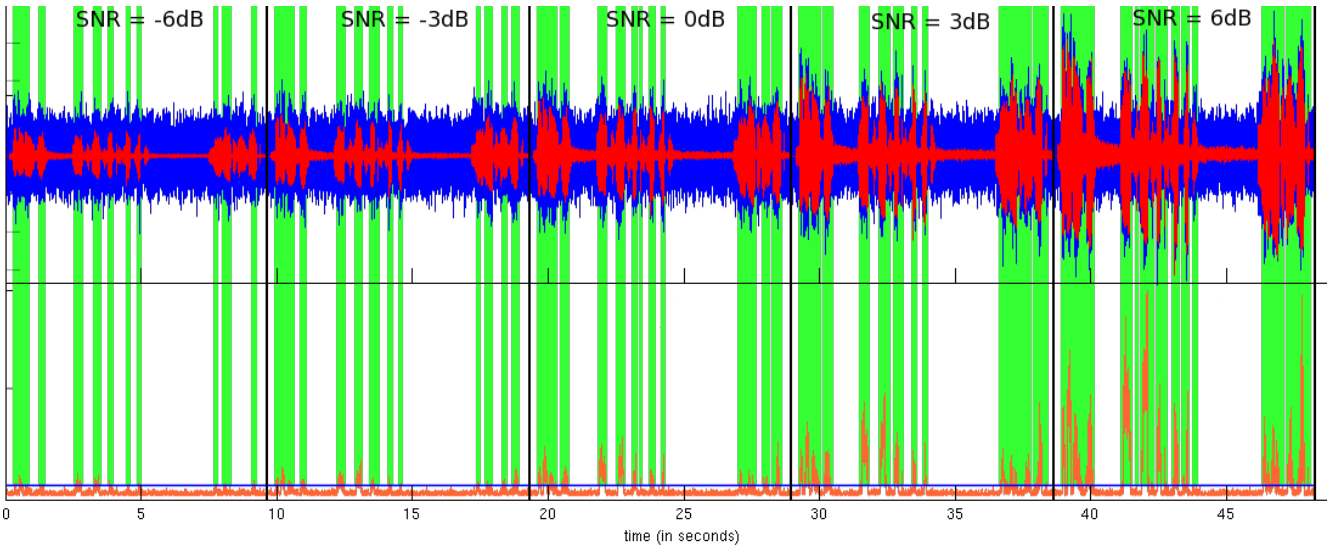

Fig. 4. Detection using the SMF algorithm for a 5-segments signal with $-6\,\text{dB} < SNR < 6\,\text{dB}$. Blue and red plots respectively correspond to signals Z and S. The green areas correspond to the detection of [H_1:*Z is the signal of interest S corrupted by noise*], when $\Lambda(z)$ (orange) is higher than ξ (blue line).

REFERENCES

[1] H. Kim, J. Kim, K. Komatani, T. Ogata, and H. Okuno, "Target speech detection and separation for communication with humanoid robots in noisy home environments," *VSP and Robotics Society of Japan*, vol. 23, no. 15, pp. 2093–2111, 2009.

[2] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J. Valin, K. Komatani, T. Ogata, and H. Okuno, "Real-time robot audition system that recognizes simultaneous speech in the real world," in *IEEE/RSJ IROS'2006*, Beijing, China, pp. 5333–5338.

[3] R. Brueckmann, A. Scheidig, and H.-M. Gross, "Adaptive noise reduction and voice activity detection for improved verbal human-robot interaction using binaural data," in *IEEE ICRA'2007*, Roma, Italy, pp. 782–1787.

[4] Y. Kida and T. Kawahara, "Evaluation of voice activity detection by combining multiple features with weight adaptation," in *INTERSPEECH'2006*, Pittsburgh, PA, pp. 1966–1969.

[5] T. Yoshida, K. Nakadai, and H. Okuno, "Automatic speech recognition improved by two-layered audio-visual, proceedings of," in *IEEE/RAS HUMANOIDS'2009*, Paris, France, pp. 604–609.

[6] J. Cavassilas, "Stochastic matched filter," in *Proc. of the Institute of Acoustics*, vol. 3, part 9, pp. 194–199, 1991.

[7] F. Chaillan, C. Fraschini, and P. Courmontagne, "Speckle noise reduction in SAS imagery," *Signal Processing*, vol. 87, no. 4, pp. 762–781, 2007.

[8] T. Kailath and H. Poor, "Detection of stochastic processes," *IEEE Trans. on Information Theory*, vol. 44, no. 6, pp. 2230–2259, 1998.

[9] H. Poor, *An Introduction to Signal Detection and Estimation*. Springer, 1994.

[10] H. Van Trees, *Detection, Estimation, and Modulation Theory - Part I*. Wiley, 1968.

[11] J. Bonnal, S. Argentieri, P. Danès, and J. Manhès, "Speaker localization and speech extraction with the EAR sensor," in *IEEE/RSJ IROS'2009*, Saint-Louis, MO, pp. 670–675.

[12] S. Schacht, J. C. Koreman, C. Lauer, A. Morris, D. Wu, and D. Klakow, "Frame based features," in *Speaker Classification (1)*, LNCS, Springer, C. Müller, Ed., vol. 4343, pp. 226–240, 2007.

[13] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, and A. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP J. on Applied Signal Processing*, vol. 4, pp. 430–451, 2004.

[14] M. Davy and S. Godsill, "Detection of abrupt spectral changes using support vector machines an application to audio signal segmentation," *IEEE ICASSP'2002*, pp. 1313–1316.

[15] V. Madisetti and D. Williams, *The Digital Signal Processing Handbook*. CRC Press - IEEE Press, 1998.

[16] J. Xiao, P. Borgnat, and P. Flandrin, "Testing stationarity with time-frequency surrogates,", *EUSIPCO'2007*, Poznán, Poland.

[17] V. Dellwo, M. Huckvale, and M. Ashby, "How is individuality expressed in voice? an introduction to speech production and description for speaker classification," in *Speaker Classification (1)*, LNCS, Springer, C. Müller, Ed., vol. 4343, pp. 1–20, 2007.