Optimal Bayesian Speech Enhancement by Parametric Joint Detection and Estimation

Van-Khanh Mai, Dominique Pastor, Abdeldjalil Aissa El Bey

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received December 23, 2019, accepted January 12, 2020, date of publication January 20, 2020, date of current version January 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2968132

Optimal Bayesian Speech Enhancement by Parametric Joint Detection and Estimation

VAN-KHANH MAI, DOMINIQUE PASTOR, (Member, IEEE), AND ABDELDJALIL AISSA-EL-BEY, (Senior Member, IEEE)
IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238 Brest, France
Corresponding author: Abdeldjalil Aissa-El-Bey (abdeldjalil.aissaelbey@imt-atlantique.fr)

This work was supported in part by the Brittany Region, and in part by the PRACOM.

ABSTRACT In this paper, we propose a general framework to estimate short-time spectral amplitudes (STSA) of speech signals in noise by joint speech detection and estimation to remove or reduce background noise, without increasing signal distortion. The approach is motivated by the fact that speech signals have sparse time-frequency representations and can reasonably be assumed not to be present in every time-frequency bin of the time-frequency domain. By combining parametric detection and estimation theories, the main idea is to take into consideration speech presence and absence in each time-frequency bin to improve the performance of Bayesian estimators. In this respect, for three Bayesian estimators, optimal Neyman-Pearson detectors are derived to decide on the absence or presence of speech in each given time-frequency bin. Decisions returned by such detectors are then used to improve the initial estimates. The resulting estimations have been assessed in two scenarios, namely, with and without reference noise power spectrum. The objective tests confirm the relevance of these approaches, both in terms of speech quality and intelligibility.

INDEX TERMS Unsupervised speech enhancement, parametric method, joint detection and estimation, Bayesian estimator, minimum mean square error (MMSE).

I. INTRODUCTION

A. CONTEXT AND MOTIVATION

In speech enhancement, one of the most important tasks is the removal or reduction of background noise from a noisy signal $y[n] = s[n] + x[n]$, where s and x are respectively the clean signal and independent noise in the time domain and $n \in \{0, 1, \ldots, T-1\}$ is the sampling time index. The observed signal is frequently segmented, windowed and transformed into the time-frequency domain. Then, the clean signal coefficients are usually retrieved by applying an enhancement algorithm to the noisy observations in this domain.

Despite good results obtained by machine learning approaches (see [1]–[3] for deep neural network or [4], [5] for dictionary-based methods), there is still room for unsupervised techniques, especially in applications where large enough databases are hardly available for all the types of noise and speech signals that can actually be encountered [6], [7]. This is the case in assisted listening for hearing aids, cochlear implants and voice communication applications.

In such applications, unsupervised techniques are expected to fulfill the following criteria, without resorting to any prior training, either for noise or for the signal of interest. Any such methods should achieve a good trade-off between intelligibility and quality. It should be robust to various stationary and non-stationary types of noise. Its complexity should be low so as to limit computational cost in real-time applications.

In this respect, many speech estimators, either parametric or non-parametric, have been designed over the last decades and have become standard. By parametric estimation, we mean a method assuming a prior model for the signal distribution, which makes it possible to resort to standard Bayesian and likelihood theory. In non-parametric inference, the signal distribution is unknown. In this paper we focus on parametric methods.

B. STATE-OF-THE-ART

Optimal Bayesian estimator algorithms aimed at removing or reducing background noise are frequently used in

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Kamrul Hasan.

speech enhancement. By assuming a statistical distribution for the signal of interest and the observation in the time-frequency domain, the estimator of the short-time spectral amplitude (STSA) is obtained by minimizing the statistical expectation of a cost function that measures the difference between the true amplitude and its estimate. These optimal estimators perform better than most unsupervised methods including the spectral-subtractive algorithms, the Wiener filtering and the subspace approach [8].

The first original optimal Bayesian STSA estimator was proposed in [9], where the cost function is the square error between the clean signal and its STSA estimate. A general STSA estimator was developed in [10], where the cost function of this method is defined by the square error of the β power amplitude. Based on the properties of auditory systems, a number of STSA Bayesian estimators are also derived by defining the cost function as the perceptual distortion metric [11], [12]. Taking advantage of the β-power and the auditory approaches, a weighted estimator is proposed in [13]. Similarly, instead of the Gaussian assumption as in the above methods, some Bayesian estimators are calculated or approximated by supposing the super-Gaussian or generalized Gamma distribution for the STSA [14]–[19].

Nevertheless, these algorithms implicitly suppose that speech is present in all time-frequency bins, which may degrade their performance. Hence, studies take into account speech presence uncertainty to estimate STSA for improving speech quality [20]–[22]. In those approaches, the gain function is simply multiplied by the speech presence probability, which induces much more noise attenuation. The speech presence probability is calculated by using the *a priori* probability of speech presence, which is assumed to be fixed or to vary with time and frequency [23], [24].

Since most algorithms do not improve speech intelligibility [25], recent research has tried to combine detection and estimation, as in the binary masking approach where, to improve speech intelligibility, spectral amplitudes in some time-frequency bins are retained, whereas others are discarded [26]. The gain function of these methods is defined as a generalized binary masking function. In this respect, since non-parametric approaches provide gain in intelligibility [25] and Bayesian approaches bring gain in speech quality [27], the authors in [28] propose a non-parametric joint detection/estimation approach combined with a Bayesian estimator. The non-parametric joint detection/estimation enhances speech intelligibility, whereas the Bayesian estimation improves speech quality by retrieving speech information in small coefficients returned by the non-parametric gain function.

In [29], the authors propose a Bayesian approach for both the detection and the estimation of speech in noise, without resorting to a logarithmic cost aimed at reflecting psycho-physics properties of the auditory system. Nevertheless, their theoretical framework requires the introduction of several parameters, among which the a priori probability of speech presence. It can thus be wondered whether a Neyman-Pearson test combined with a speech Bayesian estimation could not, at least, perform as well as the method exposed in [29]. Such a combination would avoid any prior on the speech probability of presence since, by construction, the Neyman-Pearson test would not require such knowledge. A preliminary answer to this question is proposed in [30]. By exploiting [31, Theorem 1] to derive a detector that feeds an estimator based on a non-continuous gain function, this solution is a continuation and extension of [26].

C. CONTRIBUTIONS

Instead of having a prior structure for the estimator as in [30], the present paper addresses the problem of deriving both the detector and the estimator from a given estimation risk. We restrict our attention to the single-channel case. In a nutshell, we exploit [31, Theorem 2] to derive a theoretical framework for joint detection/estimation of speech signals in noise, where the detection is performed by a Neyman-Pearson test and the denoising by a Bayesian estimator. In contrast to [30], the derivation requires no prior knowledge on the estimator. We investigate three Bayesian estimation risks chosen according to the underlying binary hypothesis testing problems. The resulting framework yields three different and optimal joint detection-estimation algorithms that are assessed experimentally. The framework involves no psycho-physics and the resulting algorithms induce no performance loss in intelligibility and quality, and even bring some improvement in comparison to [29] and [30], with a reduced number of parameters.

Within the theoretical framework outlined above, we further consider two types of binary hypothesis testing models are considered and the Bayesian estimation is performed by assigning an estimation cost to each correct and each erroneous decision. The first model is the well-known strict binary speech presence and absence model, where it is assumed that a given time-frequency bin pertains to either noise only or to the sum of speech and noise. In the second model, we assume that speech is always present with variable energies in each time-frequency bin. Specifically, we suppose that, under the null hypothesis, the observed signal is composed of noise and negligible speech while, in the alternative hypothesis, the observed signal is the sum of noise and speech of actual interest.

As can easily foreseen, the main difference between the two models is that the former will lead to a solution where no amplitude estimate is provided when the null hypothesis (*i.e* the absence of speech) is accepted, whereas the latter will always introduce a rough estimate of the speech amplitude, even when the null hypothesis (*i.e* speech of little interest is present) is accepted. It can therefore be expected that introducing some estimate, even when speech signals may have small amplitudes, makes it possible to improve speech intelligibility without affecting speech quality too much.

D. PAPER ORGANIZATION

The remainder of this article is organized as follows. Section II presents notation and assumptions about speech and noise. In Section III, the joint detection and estimation theory for speech enhancement is presented in its generality. Based on this, the generalized binary STSA combined estimator in the strict speech presence and absence model is derived in Section IV. Similarly, Section V addresses the second type of binary model where speech signal is supposed to be present with two different amplitude levels. Then, in Section VI, experimental results conducted on both synthetic and real-world noise emphasizes the gain brought by our methods. Finally, Section VII concludes this article.

II. SIGNAL MODEL IN THE TIME-FREQUENCY DOMAIN AFTER SHORT TIME FREQUENCY TRANSFORM (STFT)

As most methods in the literature, this article considers observations in the time-frequency domain after short time Fourier transform (STFT). The corrupted speech in the time-frequency domain is hereafter denoted by $Y[m,k] = S[m,k] + X[m,k]$, where m and k denote the time frame and frequency-bin indices, respectively, and $S[m,k]$ and $X[m,k]$ denote the STFT coefficients of the clean speech signal and noise, correspondingly. These STFT coefficients within the same time-frequency bin are assumed to have complex Gaussian distributions with zero-mean and to be uncorrelated [9]. For convenience, the m and k indices will be omitted in the sequel unless for clarification. In this respect, we often write Y, S and X instead of $Y[m,k]$, $S[m,k]$ and $X[m,k]$, respectively. Estimates are pointed by a wide hat symbol: *e.g.* $\widehat{\psi}$ is an estimate of ψ. In the sequel, random variables will be in capital letter, whereas their realizations will be denoted by lowercase letters. The complex noisy coefficients in polar form are written as $A_Y e^{\Phi_Y} = A_S e^{\Phi_S} + A_X e^{\Phi_X}$, where $\{A_Y, A_S, A_X\}$ and $\{\Phi_Y, \Phi_S, \Phi_X\}$ are the amplitudes and phases of the observed signal, clean speech and noise respectively. Since the clean speech and noise STFT coefficients are supposed to be uncorrelated and centered within a given time-frequency bin, we have $\mathbf{E}(A_Y^2) = \sigma_S^2 + \sigma_X^2$ where $\sigma_S^2 = \mathbf{E}(A_S^2)$, $\sigma_X^2 = \mathbf{E}(A_X^2)$ and $\mathbf{E}$ is the expectation. The *a priori* signal-to-noise ratio (SNR) ξ and the *a posteriori* SNR γ are defined as follows

$$\xi = \sigma_S^2/\sigma_X^2 \quad \text{and} \quad \gamma = A_Y^2/\sigma_X^2. \tag{1}$$

For the sake of simplicity, we simply denote the clean speech amplitude A_S by A.

III. JOINT DETECTION AND ESTIMATION APPROACH: GENERAL FRAMEWORK

In order to take into account the presence and absence of speech, the general framework involves a two-state model, specified by two hypotheses H_0 and H_1 for the absence and presence of speech signal, respectively. Specifically, H_0 models the case where speech is absent or present with little interest, whereas hypothesis H_1 models the case where speech is present. Under each hypothesis H_i ($i = 0, 1$), Y is supposed to follow a probability density function (pdf) denoted by $f_Y(y; H_i)$. The foregoing is summarized as:

$$\begin{aligned} H_0 &: Y \sim f_Y(y; H_0) \\ H_1 &: Y \sim f_Y(y; H_1), \end{aligned} \tag{2}$$

Given the observation Y, the decision $\mathbf{D}$ takes its value in $\{0, 1\}$ and thus returns the index of the so-called accepted hypothesis. In the sequel, for $i, j \in \{0, 1\}$, our convention is to use index i to designate the index of true hypothesis and to use index j to designate the outcome of $\mathbf{D}$ (index of the accepted hypothesis). In this respect, we use the following conventional notation:

- A_i denotes the clean speech amplitude under hypothesis H_i;
- $\widehat{A}_j$ denotes the speech amplitude estimate when the decision is $\mathbf{D} = j$;
- a_i designates a realization of the random variable A_i;
- $\widehat{a}_j$ designates a realization of the random variable $\widehat{A}_j$;

TABLE 1. Cost functions based on the two-state model.

	H_0	H_1
$\mathbf{D} = 0$	$c_{00}(\widehat{a}_0, a_0)$	$c_{01}(\widehat{a}_0, a_1)$
$\mathbf{D} = 1$	$c_{10}(\widehat{a}_1, a_0)$	$c_{11}(\widehat{a}_1, a_1)$

Following the Bayesian framework, we define four cost functions c_{ji} for $i, j \in \{0, 1\}$. These cost functions are shown in Table 1, where each cost function is defined in $[0, \infty) \times [0, \infty)$ and valued in $[0, \infty)$. In the sequel, we focus on non-randomized decisions, that is, decisions $\mathbf{D}$ for which exists a function or test δ defined on $\mathbb{C}$ and valued in $\{0, 1\}$ such that $\mathbf{D} = \delta(Y)$. Therefore, the weighted cost function under H_i can be defined by setting:

$$C_i(\widehat{A}_1, \widehat{A}_0, A_i) = c_{1i}(\widehat{A}_1, A_i)\,\delta(Y) + c_{0i}(\widehat{A}_0, A_i)\,(1 - \delta(Y)). \tag{3}$$

Below, as often, the estimates $\widehat{A}_j$ of the true amplitude when $j \in \{0, 1\}$ are sought in the form $\widehat{A}_j = \psi_j(Y)$ where ψ_0, ψ_1 are functions defined on $\mathbb{C}$ and valued in $[0, \infty)$. The Bayesian risk under hypothesis H_i with $i \in \{0, 1\}$ can be defined as :

$$\mathbf{R}_i(\psi_1, \psi_0, \delta) = \mathbf{E}_i\left[C_i(\widehat{A}_1, \widehat{A}_0, A_i)\right], \tag{4}$$

where $\mathbf{E}_i$ denotes the statistical expectation under H_i. We then follow [31] by calculating the maps ψ_1^*, ψ_0^* and the test δ^* such that $(\psi_1^*, \psi_0^*, \delta^*)$ is a solution to the following constrained optimization problem:

$$\begin{aligned} (\psi_1^*, \psi_0^*, \delta^*) = \underset{\psi_1, \psi_0, \delta}{\operatorname{argmin}}\ & \mathbf{R}_1(\psi_1, \psi_0, \delta) \\ \text{subject to } & \mathbf{R}_0(\psi_1, \psi_0, \delta) \le \alpha. \end{aligned} \tag{5}$$

where α is the level of the test and is chosen in $(0, 1)$. By so proceeding, we control the cost of erroneously estimating the signal amplitude under H_0, namely, when the signal is absent or present with low energy and there is no real need to estimate it accurately. So, we can be satisfied by upper-bounding the estimation cost under H_0. Of course, the upper-bound must be fixed to a small value α, which allows for

a trade-off between speech quality and intelligibility. In contrast, under H_1, the speech signal must be estimated as accurately as possible. In this case, we want to minimize the estimation cost.

According to [31, Theorem 2], there exists a solution $(\psi_1^*, \psi_0^*, \delta^*)$ to (5). To present this solution, we need to introduce the conditional risks

$$r_{ji}(y;\psi_j) = \mathbf{E}_i\left[c_{ji}(\psi_j(Y), A_i)|Y=y\right] = \int_{\mathbb{R}} c_{ji}(\psi_j(y), a_i) f_{A_i|Y=y}(a_i; H_i)\, da_i\ , \quad (6)$$

where $f_{A_i,Y}(a_i, y; H_i)$ is the joint pdf of A_i and Y under hypothesis H_i. The solution $(\psi_1^*, \psi_0^*, \delta^*)$ then obtains as follows. To begin with, for any $y \in \mathbb{C}$,

$$\psi_j^*(y) = \underset{\psi_j}{\operatorname{argmin}}\ \mathbb{D}_j(y;\psi_j), \quad j \in \{0,1\} \quad (7)$$

when ψ_j ranges in the set of all functions defined on $\mathbb{C}$, and valued in $[0,\infty)$ and where:

$$\mathbb{D}_j(y;\psi_j) = f_Y(y;H_1) r_{j1}(y;\psi_j) + \tau f_Y(y;H_0) r_{j0}(y;\psi_j). \quad (8)$$

In (8), τ is a Lagrange factor whose calculation is performed at the same time we compute the test δ^* as follows. Specifically, δ^* is defined for all $y \in \mathbb{C}$ as:

$$\delta^*(y) = \begin{cases} 1 & \text{if } p_{01}(y) \geq \tau p_{10}(y) \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

with

$$p_{ji}(y) = f_Y(y;H_i)\left(r_{ji}(y;\psi_j^*) - r_{ii}(y;\psi_i^*)\right). \quad (10)$$

The Lagrange factor τ is determined by solving

$$\mathbf{R}_0\left(\psi_1^*, \psi_0^*, \delta^*\right) = \alpha. \quad (11)$$

On the basis of the foregoing (see (7) and (9)), the general strategy is thus specified by the following three steps:

Step #1 (Prior estimation): Compute the close forms, if any, for the solutions ψ_0^* and ψ_1^* to (7). These close forms depend on τ.

Step #2 (Decision) : Take the decision via (9). The parameter τ is determined at this stage by solving (11).

Step #3 (Final estimation) : Estimate the amplitude by $\widehat{A}_j = \psi_j^*(Y)$ where $j = \delta^*(Y)$.

In the next sections, we apply these three steps to different cost functions c_{ji}. In each case, we specify the probability density function (pdf) involved in (2). We then choose the cost functions c_{ji} with respect to the possible decisions and hypotheses.

IV. STRICT PRESENCE/ABSENCE JOINT ESTIMATOR

In this section, the noisy speech signal is modeled as:

$$\begin{aligned} &H_0\,(\text{ speech is absent}) : Y = X \\ &H_1\,(\text{ speech is present}) : Y = S + X, \end{aligned} \quad (12)$$

where H_0 and H_1 are the null and alternative hypotheses denoting speech presence and speech absence in the given time-frequency bin, respectively. The two-state model (12) is henceforth called the strict model (SM) because, in contrast to the models in Section V, the two hypotheses H_0 and H_1 it involves do not cover the case of transient speech signals with feeble amplitudes of poor interest for the denoising. Accordingly, the solution $(\psi_1^*, \psi_0^*, \delta^*)$ to (5) is hereafter denoted by $(\psi_1^{\mathrm{SM}}, \psi_0^{\mathrm{SM}}, \delta^{\mathrm{SM}})$.

TABLE 2. Cost functions based on the strict presence/absence model.

	H_0	H_1
$\mathbf{D}=0$	$c_{00}(\widehat{a}_0, a_0) = 0$	$c_{01}(\widehat{a}_0, a_1) = a_1^2$
$\mathbf{D}=1$	$c_{10}(\widehat{a}_1, a_0) = 1$	$c_{11}(\widehat{a}_1, a_1) = (\widehat{a}_1 - a_1)^2$

We make a complex Gaussian assumption for the pdf of Y under each hypothesis H_i, $i \in \{0,1\}$:

$$f_Y(y;H_0) = \frac{1}{\pi\sigma_X^2}\exp\left(-\frac{|y|^2}{\sigma_X^2}\right), \quad (13)$$

$$f_Y(y;H_1) = \frac{1}{\pi\sigma_X^2(1+\xi)}\exp\left(-\frac{|y|^2}{\sigma_X^2(1+\xi)}\right). \quad (14)$$

We propose the cost functions defined in Table 2. In coherence with (12), the cost functions c_{00} and c_{01} are chosen to force the STSA estimate to 0 when H_0 is accepted. When the alternative hypothesis is true and accepted, it is natural to consider a quadratic cost. Now, according to (3) and (4), and since we choose $c_{00} = 0$, the Bayesian risk under H_0 is $\mathbf{R}_0(\psi_1, \psi_0, \delta) = \mathbf{E}_0\left[c_{10}(\widehat{A}_1, A_0)\delta(Y)\right]$. On the other hand, $\mathbf{E}_0[\delta(Y)]$ is the false alarm probability of test δ. Therefore, it turns out that choosing a constant function c_{10} makes it possible to cancel the impact of a false alarm on the STSA estimate under H_0 and to get a risk proportional to the false alarm probability of δ. In this section, we choose $c_{10} = 1$.

These choices for the cost functions make the approach similar to ideal binary masking [32] in the sense that, when the decision is that noise only is present, the estimated amplitude is set to 0. However, when the presence of speech is accepted, the joint detection/estimation approach provides a Bayesian estimate, in contrast to ideal binary masking that merely keeps the observed noisy amplitude.

Note that if we choose $c_{ii} = 0$ and $c_{ji} = 1$ for $i \neq j$, the optimization problem (5) is a binary hypothesis testing problem whose solution is provided by the Neyman-Pearson lemma.

Step #1: Prior estimation

It follows from Table 2, (6) and (8) that, given $\psi_0, \psi_1 : \mathbb{C} \to [0,\infty)$ and $y \in \mathbb{C}$,

$$\mathbb{D}_0(y;\psi_0) = f_Y(y;H_1)\int a_1^2 f_{A_1|Y=y}(a_1;H_1)\, da_1. \quad (15)$$

and

$$\mathbb{D}_1(y;\psi_1) = f_Y(y;H_1)\int (\psi_1(y) - a_1)^2 f_{A_1|Y=y}(a_1;H_1)\, da_1 + \tau f_Y(y;H_0)\int f_{A_0|Y=y}(a_0;H_0)\, da_0. \quad (16)$$

Since the right-hand side (rhs) in (15) does not depend on ψ_0, it follows that ψ_0^{SM} can be any function of $\mathbb{C}$ into $[0, \infty)$. However, to force to 0 the STSA when H_0 is accepted, we hereafter choose $\psi_0^{\mathrm{SM}} = 0$. With this choice, the estimated STSA under H_0 is:

$$\widehat{A}_0 = 0.$$

Since the minimization of $\mathbb{D}_1(y; \psi_1)$ does not depend on $f_Y(y; H_0)$, a solution ψ_1^{SM} to Eq. (7) is a function that minimizes

$$f_Y(y; H_1) \int (\psi_1(y) - a_1)^2 f_{A_1|Y=y}(a_1; H_1) da_1.$$

Developing the integral leads to a second-order equation in $\psi_1(y)$. The minimum of this second-order equation is found to be:

$$\psi_1^{\mathrm{SM}}(y) = \int_0^\infty a_1 f_{A_1|Y=y}(a_1; H_1)\, da_1.$$

A close form for $\psi_1^{\mathrm{SM}}(y)$ is given by [9, Eq. (7)]:

$$\psi_1^{\mathrm{SM}}(y) = G(\xi, \gamma)|y|, \tag{17}$$

with

$$G(\xi, \gamma) = \frac{\sqrt{\pi \nu}}{2\gamma} e^{-\nu/2} \left[(1+\nu) I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right) \right],$$
$$\nu = \frac{\gamma \xi}{1+\xi} \tag{18}$$

and where $I_0(\cdot)$ and $I_1(\cdot)$ are the modified Bessel functions of zero and first order, respectively. The gain (18) is a function of two variables: the *a priori* SNR ξ and the *a posteriori* SNR γ. As mentioned in [9], for high *a posteriori* SNR, this gain function is close to the Wiener gain function. In addition, the *a posteriori* SNR is directly given by the observed amplitude A_Y. In contrast, the *a priori* SNR is unknown. This variable ξ can be estimated via the decision directed approach [9]:

$$\xi[m,k] = \varpi \frac{\widehat{a}_1^2[m-1,k]}{\sigma_X^2[m-1,k]} + (1-\varpi)\max\left((\gamma[m,k]-1), 0\right), \tag{19}$$

where $0 < \varpi < 1$ is the smoothing parameter and $\widehat{A}_1[m-1,k]$ is the estimated STSA at the previous frame. Thus, the STSA estimate under hypothesis H_1 is obtained as:

$$\widehat{A}_1 = G(\xi, \gamma) A_Y.$$

Step #2: Decision

According to (6), Table 2 and the results of the preceding section, we calculate the conditional expectations $r_{ji}(y; \psi_j^{\mathrm{SM}})$, before injecting them into (10) and calculating the test (9).

Under H_0, we first have $r_{00}(y; \psi_0^{\mathrm{SM}}) = 0$ and

$$r_{10}(y; \psi_1^{\mathrm{SM}}) = \int f_{A_0|Y=y}(a_0; H_0)\, da_0 = 1 \tag{20}$$

Now, under H_1, we obtain:

$$r_{01}(y; \psi_0^{\mathrm{SM}}) = \int_0^\infty a_1^2 f_{A_1|Y=y}(a_1; H_1) da_1. \tag{21}$$

and, similarly:

$$r_{11}(y; \psi_1^{\mathrm{SM}}) = \int_0^\infty (\psi_1^{\mathrm{SM}}(y) - a_1)^2 f_{A_1|Y=y}(a_1; H_1)\, da_1. \tag{22}$$

Expanding the square in the rhs of (22), we now get:

$$r_{11}(y; \psi_1^{\mathrm{SM}}) = r_{01}(y; \psi_0^{\mathrm{SM}}) - \left(\psi_1^{\mathrm{SM}}(y)\right)^2$$

Therefore, in the strict presence/absence model considered in this section for STSA estimation, the decision (9) becomes:

$$\delta^{\mathrm{SM}}(y) = \begin{cases} 1 & \text{if } \mathcal{D}^{\mathrm{SM}}(y) \geq \tau^{\mathrm{SM}} \\ 0 & \text{otherwise.} \end{cases} \tag{23}$$

where

$$\tau^{\mathrm{SM}} = -\sigma_X^2 \lambda(\xi, \gamma) G(\xi, \gamma)^2 \log(\alpha) \tag{24}$$

is calculated by solving (11) in Appendix A and

$$\mathcal{D}^{\mathrm{SM}}(y) = \lambda(\xi, \gamma) \left(\psi_1^{\mathrm{SM}}(y)\right)^2, \tag{25}$$

where, according to Eqs. (13) and (14):

$$\lambda(\xi, \gamma) = \frac{f_Y(y; H_1)}{f_Y(y; H_0)} = \frac{\exp(\nu)}{1+\xi}. \tag{26}$$

Step #3: Final estimation (SM-STSA)

In short, for each time-frequency bin, the proposed joint method estimates first the speech STSA by using the Bayesian estimator. Then, the detector is based on this estimate to detect the presence or absence of speech at each bin. If speech is absent, the SM-STSA joint estimator sets the speech STSA to 0. The STSA estimate thus returned by SM-STSA can be written as a binary masking:

$$\widehat{A} = G_{\mathrm{SM}}(\xi, \gamma) A_Y. \tag{27}$$

The gain function $G_{\mathrm{SM}}(\xi, \gamma)$ is:

$$G_{\mathrm{SM}}(\xi, \gamma) = \begin{cases} G(\xi, \gamma) & \text{if } \mathcal{D}^{\mathrm{SM}}(y) \geq \tau^{\mathrm{SM}} \\ 0 & \text{otherwise,} \end{cases} \tag{28}$$

where the threshold τ^{SM} is determined to guarantee a probability of false alarm (PFA) equal to α (see Appendix A).

According to the equations above and after some easy simplifications, the algorithm can be summarized as follows:

V. UNCERTAIN PRESENCE/ABSENCE JOINT ESTIMATORS

The proposed above method based on strict presence/absence hypotheses may introduce musical noise since the estimator can randomly generate isolated peaks in the time frequency domain. To overcome this issue, we proceed as in [33] by assuming that, under hypothesis H_0, speech is present with small amplitude. Under the alternative hypothesis H_1, the noisy signal remains the sum of speech and noise. Therefore, with these hypotheses, the two-state model is

$$H_0 : Y = S_0 + X,$$
$$H_1 : Y = S_1 + X, \tag{29}$$

Algorithm 1 SM-STSA

1: Input: $Y = A_Y e^{\Phi_Y}$, σ_X, PFA $= \alpha$, ξ and γ (cf. Eq. (1))

Step #1: Prior estimation

2: Compute $\nu = \gamma\xi/(1+\xi)$

3: Compute

$$G(\xi,\gamma) = \frac{\sqrt{\pi\nu}}{2\gamma} e^{-\nu/2}\left[(1+\nu)I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right)\right]$$

Step #2: Decision

4: Compute $\tau^{\mathrm{SM}} = -\sigma_X^2 \log(\alpha)$

Step #3: Final estimation

5: Calculate

$$\widehat{A}_S = \begin{cases} G(\xi,\gamma)A_Y & \text{if } A_Y^2 \geq \tau^{\mathrm{SM}} \\ 0 & \text{otherwise,} \end{cases}$$

6: Output: $\widehat{S} = \widehat{A}_S\, e^{\Phi_Y}$.

where the speech signal S is either S_0 or S_1, depending on the true hypothesis. Clearly, S_0 is key to distinguish between the two models summarized by (12) and (29). Therefore, we suppose that $S_0 = \sqrt{\beta}X$ where β ($0 < \beta \ll 1$) is a constant spectral floor parameter [34], which is empirically chosen. Under these assumptions, the conditional pdfs are now:

$$f_Y(y;H_0) = \frac{1}{\pi\sigma_X^2(1+\beta)} \exp\left(-\frac{|y|^2}{\sigma_X^2(1+\beta)}\right) \quad (30)$$

$$f_Y(y;H_1) = \frac{1}{\pi\sigma_X^2(1+\xi)} \exp\left(-\frac{|y|^2}{\sigma_X^2(1+\xi)}\right) \quad (31)$$

The main difference between the two pdfs above is that, under hypothesis H_0, the *a priori* SNR β is identical for all frequency bins since β is fixed once for all, whereas, under hypothesis H_1, the *a priori* SNR $\xi = \xi[m,k]$ varies in time and frequency.

The standard likelihood ratio $\Lambda(\xi,\gamma)$ is directly computed by using (30) and (31) and equals:

$$\Lambda(\xi,\gamma) = \frac{f_Y(y;H_1)}{f_Y(y;H_0)} = \frac{1+\beta}{1+\xi} \exp\left(\frac{\gamma(\xi-\beta)}{(1+\beta)(1+\xi)}\right). \quad (32)$$

A. INDEPENDENT STSA JOINT ESTIMATOR

In this section, we consider the same standard quadratic cost function for the four different decision cases. This cost is defined as:

$$c_{ji}(a,b) = (a-b)^2. \quad (33)$$

Therefore, the cost functions are those of Table 3.

TABLE 3. Cost functions based on the uncertain presence/absence model for independent STSA joint estimation.

	H_0	H_1
$\mathbf{D}=0$	$c_{00}(\widehat{a}_0, a_0) = (\widehat{a}_0 - a_0)^2$	$c_{01}(\widehat{a}_0, a_1) = (\widehat{a}_0 - a_1)^2$
$\mathbf{D}=1$	$c_{10}(\widehat{a}_1, a_0) = (\widehat{a}_1 - a_0)^2$	$c_{11}(\widehat{a}_1, a_1) = (\widehat{a}_1 - a_1)^2$

Step #1: Prior estimation

Given $\psi_j : \mathbb{C} \to [0,\infty)$ and $y \in \mathbb{C}$, Table 3 induces that $\mathbb{D}_j(y;\psi_1)$ can be rewritten as:

$$\mathbb{D}_j(y;\psi_j) = f_Y(y;H_1)\int (\psi_j(y)-a_1)^2 f_{A_1|Y=y}(a_1;H_1)da_1 + \tau f_Y(y;H_0)\int (\psi_j(y)-a_0)^2 f_{A_0|Y=y}(a_0;H_0)da_0$$

We have a convex function of $\psi_j(y)$ and by derivation with respect to $\psi_j(y)$, some routine algebra shows that the function ψ_j^* that minimizes $\mathbb{D}_j(y;\psi)$ does not depend on j and equals

$$\psi_j^*(y) = \psi^{\mathrm{IUM}}(y) = G^{\mathrm{IUM}}(\xi,\gamma)|y|, \quad (34)$$

where

$$G^{\mathrm{IUM}}(\xi,\gamma) = \frac{\Lambda(\xi,\gamma)\,G(\xi,\gamma) + \tau\, G(\beta,\gamma)}{\Lambda(\xi,\gamma)+\tau}. \quad (35)$$

In the equation above, $G(\xi,\gamma)$ is defined by (18) and τ can be any non-negative real value. According to Appendix B-A, we however choose

$$\tau = \tau^{\mathrm{IUM}} = \Lambda(\xi, -(1+\beta)\log(\alpha)). \quad (36)$$

In the notation above, IUM means "Independent" estimator in the "Uncertain Model".

Step #2: Decision

According to (6) and Table 3,

$$r_{ji}(y;\psi^{\mathrm{IUM}}) = \int_{\mathbb{R}} \left(\psi^{\mathrm{IUM}}(y) - a_i\right)^2 f_{A_i|Y=y}(a_i;H_i)\,da_i.$$

It follows from the foregoing equality and (10) that $p_{10} = p_{01} = 0$. Therefore, the presence of speech is always accepted, which is coherent with the model (29).

Step #3: Final estimation (IUM-STSA)

Since the presence of speech is always accepted, the estimated STSA is always:

$$\widehat{A}_1 = \psi^{\mathrm{IUM}}(Y) = G^{\mathrm{IUM}}(\xi,\gamma)A_Y, \quad (37)$$

where the gain function $G^{\mathrm{IUM}}(\xi,\gamma)$ is given by (35) and (36). Because we get the same STSA estimator under each hypothesis, we call it independent uncertain model STSA joint estimator (IUM-STSA).

B. JOINT STSA ESTIMATOR

For further taking the role of the presence and absence of speech into account, we consider the cost functions of Table 4.

Specifically, unlike Subsection V-A, the cost functions penalize both the estimation and the detection errors.

Algorithm 2 IUM-STSA

1: Input: $Y = A_Y\, e^{\Phi_Y}$, ξ, γ, PFA $= \alpha$, spectral floor $= \beta$, ξ and γ (cf. Eq. (1))

Step #1: Prior estimation

2: Compute $\Lambda(\xi, \gamma) = \frac{1+\beta}{1+\xi} \exp\left(\frac{\gamma(\xi-\beta)}{(1+\beta)(1+\xi)}\right)$

3: Compute $\tau = \Lambda(\xi, -(1+\beta)\log(\alpha))$

4: Compute

$$G(\xi, \gamma) = \frac{\sqrt{\pi\nu}}{2\gamma} e^{-\nu/2}\left[(1+\nu)I_0\left(\frac{\nu}{2}\right) + \nu I_1\left(\frac{\nu}{2}\right)\right]$$

5: $G^{\mathrm{IUM}}(\xi, \gamma) = \frac{\Lambda(\xi,\gamma)G(\xi,\gamma)+\tau G(\beta,\gamma)}{\Lambda(\xi,\gamma)+\tau}$

Step #2: Decision

The presence of speech is always accepted,

Step #3: Final estimation

6: Compute $\widehat{A}_S = G^{\mathrm{IUM}}(\xi, \gamma)A_Y$

7: Output: $\widehat{S} = \widehat{A}_S\, e^{\Phi_Y}$.

TABLE 4. Cost functions based on the uncertain presence/absence model for joint estimation.

	H_0	H_1
$\mathbf{D}=0$	$c_{00}(\widehat{a}_0, a_0) = (\widehat{a}_0 - a_0)^2$	$c_{01}(\widehat{a}_0, a_1) = a_1^2$
$\mathbf{D}=1$	$c_{10}(\widehat{a}_1, a_0) = a_0^2$	$c_{11}(\widehat{a}_1, a_1) = (\widehat{a}_1 - a_1)^2$

For instance, in case of a missed detection, the cost is the square of the true amplitude.

Step #1: Prior estimation

Similar to the above subsection, given $\psi_0, \psi_1 : \mathbb{C} \to [0, \infty)$ and $y \in \mathbb{C}$, Table 4 induces that:

$$\mathbb{D}_1(y; \psi_1) = f_Y(y; H_1) \int (\psi_1(y) - a_1)^2 f_{A_1|Y=y}(a_1; H_1) da_1 + \tau f_Y(y; H_0) \int a_0^2 f_{A_0|Y=y}(a_0; H_0) da_0 \quad (38)$$

$$\mathbb{D}_0(y; \psi_0) = f_Y(y; H_1) \int a_1^2 f_{A_1|Y=y}(a_1; H_1) da_1 + \tau f_Y(y; H_0) \int (\psi_0(y) - a_0)^2 f_{A_0|Y=y}(a_0; H_0) da_0 \quad (39)$$

By derivation with respect to $\psi_j(y)$ of each $\mathbb{D}_j(y; \psi_j)$, the function $\psi_j^*(y)$ that minimizes $\mathbb{D}_j(y; \psi_j)$ is the function ψ_j^{JUM} defined by:

$$\psi_j^{\mathrm{JUM}}(y) = \int_0^\infty a_j f_{A_j|Y=y}(a_j; H_j) da_j = G(\theta_j, \gamma)|y|. \quad (40)$$

where $\theta_1 = \xi$ as in the standard gain function $G(\xi, \gamma)$ whereas $\theta_0 = \beta$. The notation JUM means "Joint" estimation in the "Uncertain Model". According to the foregoing, the estimated $\widehat{A}_i$ is given:

$$\widehat{A}_j = \psi_j^{\mathrm{JUM}}(Y) = G(\theta_j, \gamma)A_Y, \quad (41)$$

Step #2: Decision

According to (6) and Table 4, the Bayesian risk r_{ji} for $j \neq i$ is given by:

$$r_{ji}(y; \psi_j^{\mathrm{JUM}}) = \int_0^\infty a_i^2 f_{A_i|Y=y}(a_i; H_i) da_i \quad (42)$$

Moreover, the Bayesian risk r_{ii} is computed by using (6) and (40) and equals:

$$r_{ii}(y; \psi_i^{\mathrm{JUM}}) = \int_0^\infty \left(\psi_i^{\mathrm{JUM}}(y) - a_i\right)^2 f_{A_i|Y=y}(a_i; H_i) da_i = r_{ji}(y; \psi_j^{\mathrm{JUM}}) - \left(\psi_i^{\mathrm{JUM}}(y)\right)^2, \quad (43)$$

with $j \neq i$. Injecting (43) and (32) into (9), we obtain the decision rule as:

$$\delta^{\mathrm{JUM}}(y) = \begin{cases} 1 & \text{if } \mathcal{D}^{\mathrm{JUM}}(y) \geq \tau^{\mathrm{JUM}} \\ 0 & \text{otherwise} \end{cases} \quad (44)$$

where $\mathcal{D}^{\mathrm{JUM}}$ is given by:

$$\mathcal{D}^{\mathrm{JUM}}(y) = \Lambda(\xi, \gamma)\left(\frac{\psi_1^{\mathrm{JUM}}(y)}{\psi_0^{\mathrm{JUM}}(y)}\right)^2. \quad (45)$$

and τ^{JUM} can be estimated as detailed in Appendix B-B.

Step #3: Final estimation (JUM-STSA)

The gain function of the JUM-STSA estimator deriving from the foregoing is written as

$$G^{\mathrm{JUM}}(\xi, \gamma) = \begin{cases} G(\xi, \gamma) & \text{if } \mathcal{D}^{\mathrm{JUM}}(y) \geq \tau^{\mathrm{JUM}}, \\ G(\beta, \gamma) & \text{otherwise}, \end{cases} \quad (46)$$

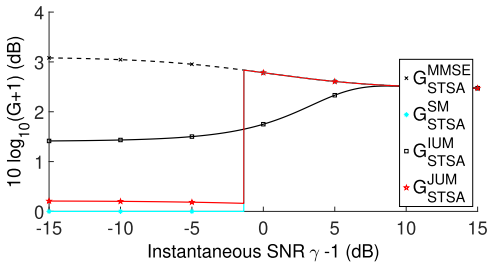

FIGURE 1. Attenuation curves of all joint detection/estimations in comparison with the standard MMSE-STSA method at *a piori* SNR level $\xi = 5$dB. The detector thresholds were calculated with $\alpha = 0.05$ and $\beta = -25$ dB.

VI. EXPERIMENTAL RESULTS

1) PURPOSE

We have proposed three different methods, namely, SM-STSA, IUM-STSA and JUM-STSA. These methods rely on different estimation cost functions and different models for speech absence and presence. For instance, the gain functions of all the methods described above are displayed in Figure 1. Compared to the standard MMSE-STSA method [9], the joint estimators seemingly provide more impact at low instantaneous SNR. We recall that the instantaneous SNR is defined

Algorithm 3 JUM-STSA

1: Input: $Y = A_Y\, e^{\Phi_Y}$, PFA $= \alpha$, spectral floor β, ξ and γ (cf. Eq. (1))

Step #1: Prior estimation

2: Compute $\nu = \gamma\xi/(1+\xi)$

3: Compute

$$G(\xi,\gamma)=\frac{\sqrt{\pi\nu}}{2\gamma}e^{-\nu/2}\left[(1+\nu)I_0\left(\frac{\nu}{2}\right)+\nu I_1\left(\frac{\nu}{2}\right)\right]$$

4: Compute $\widehat{A}_0 = G(\beta,\gamma)A_Y$ and $\widehat{A}_1 = G(\xi,\gamma)A_Y$

Step #2: Decision

5: Compute $\Lambda(\xi,\gamma) = \frac{1+\beta}{1+\xi}\exp\left(\frac{\gamma(\xi-\beta)}{(1+\beta)(1+\xi)}\right)$

6: Compute $\mathcal{D}^{\mathrm{JUM}}(A_Y) = \Lambda(\xi,\gamma)\left(\frac{\widehat{A}_1}{\widehat{A}_0}\right)^2$

Step #3: Final estimation

7: Compute

$$\widehat{A}_S = \begin{cases}\widehat{A}_1 & \text{if } \mathcal{D}^{\mathrm{JUM}}(A_Y) \geq \tau^{\mathrm{JUM}} \\ \widehat{A}_0 & \text{otherwise,}\end{cases}$$

8: Output: $\widehat{S} = \widehat{A}_S\, e^{\Phi_Y}$.

by $(\gamma - 1)$ [35]. This figure illustrates the fact that, when noise or low level speech signals in noise can be detected, the joint estimators will reduce more the background noise than MMSE-STSA and thus, are expected to improve speech intelligibility and quality.

It follows that the three methods are expected to perform differently, depending on the type of noise and the criterion, either objective or subjective, to assess them. Therefore, the purpose of this section is to provide the reader with all possible information making it possible to choose the most appropriate method with regard to a given context and criteria relevant to it. In this respect, we also benchmark SM-STSA, IUM-STSA and JUM-STSA with methods drawn from the literature.

More precisely, we experimentally assessed the contribution in speech denoising of our algorithms based on optimal joint Neyman-Pearson decision and STSA Bayesian estimation, especially in comparison to the optimal fully Bayesian approach for speech joint detection/estimation [29, Eqs. (20) & (23)] and hereafter termed SDE-STSA (SDE for Simultaneous Detection and Estimation). The MMSE-STSA was also involved in the assessment as the standard baseline within the class of STSA Bayesian estimators with quadratic costs.

In order to carry out a fair assessment, methods involving psycho-physics in the definition of the Bayesian cost were not considered. More generally, an exhaustive assessment involving most of the numerous speech enhancement methods available from the literature — and not necessarily based on joint detection/estimation — is beyond the scope of the paper.

We also bench-marked our algorithms to SDE-NCG-Ab introduced and recommended in [30]. However, we do not display the results obtained with SDE-NCG-Ab for two reasons. On the one hand, SDE-NCG-Ab performs similarly to SDE-STSA in terms of SSNR and does not outperform it with respect to composite criteria. On the other hand, displaying the results obtained with SDE-NCG-Ab in addition to SM-STSA, IUM-STSA, JUM-STSA, SDE-STSA and MMSE-STSA would be detrimental to the readiness of the subsequent figures.

2) SPEECH AND NOISE MATERIAL

The assessment involved the whole NOIZEUS database [8], [35]. This database involves 30 different clean speech sentences produced by three male and three female speakers. For each of these clean signals, the database also provides 9 noisy versions: 3 noisy signals respectively obtained by adding three different types of quasi-stationary noise (car, train and station) and 6 noisy signals respectively obtained by adding six different types of non-stationary noise (airport, exhibition, restaurant, street, modulated WGN and babble). All these types of noise are from the AURORA database. In addition, we corrupted the 30 clean speech sentences by two other types on additive noise: synthetic white Gaussian noise and 2nd-order auto-regressive (AR) noise. The tests involved four SNR levels, namely 0, 5, 10 and 15 dB. These SNR levels are of practical interest since, in the single-channel case, all the methods considered in the paper fail below 0 dB in presence of quasi- and non-stationary noises.

3) STFT PARAMETERS

In our experiments, speech signals were sampled at 8 kHz, segmented into frames of 256 samples each, transformed by STFT with 50% overlapped Hamming windows. The smoothing parameter ϖ in (19) was, as often, set to 0.98. The false alarm probability α was fixed to 0.05 for all noise levels and the spectral floor parameter β was set to 0.002 [35].

4) METHODOLOGY

The performance of all the methods were evaluated in two scenarios. In the first scenario, denoising is performed by using the reference noise power spectrum. If noise is stationary, the reference noise power spectrum is simply the theoretical power spectrum. Otherwise, the reference noise power spectrum of frame m in a given bin k is estimated as in [36] by:

$$\sigma_X^2[m,k] = \mu\sigma_X^2[m-1,k] + (1-\mu)|X[m,k]|^2, \quad (47)$$

where $\mu = 0.9$. This iterative estimation is initialized by setting $\sigma_X^2[0,k] = |X[0,k]|^2$. The purpose of this scenario is to assess the performance of the denoising in itself, as much as possible. In the second scenario, for all the methods, the noise power spectrum was estimated using the B-E-DATE algorithm introduced in [37]. This scenario makes it possible

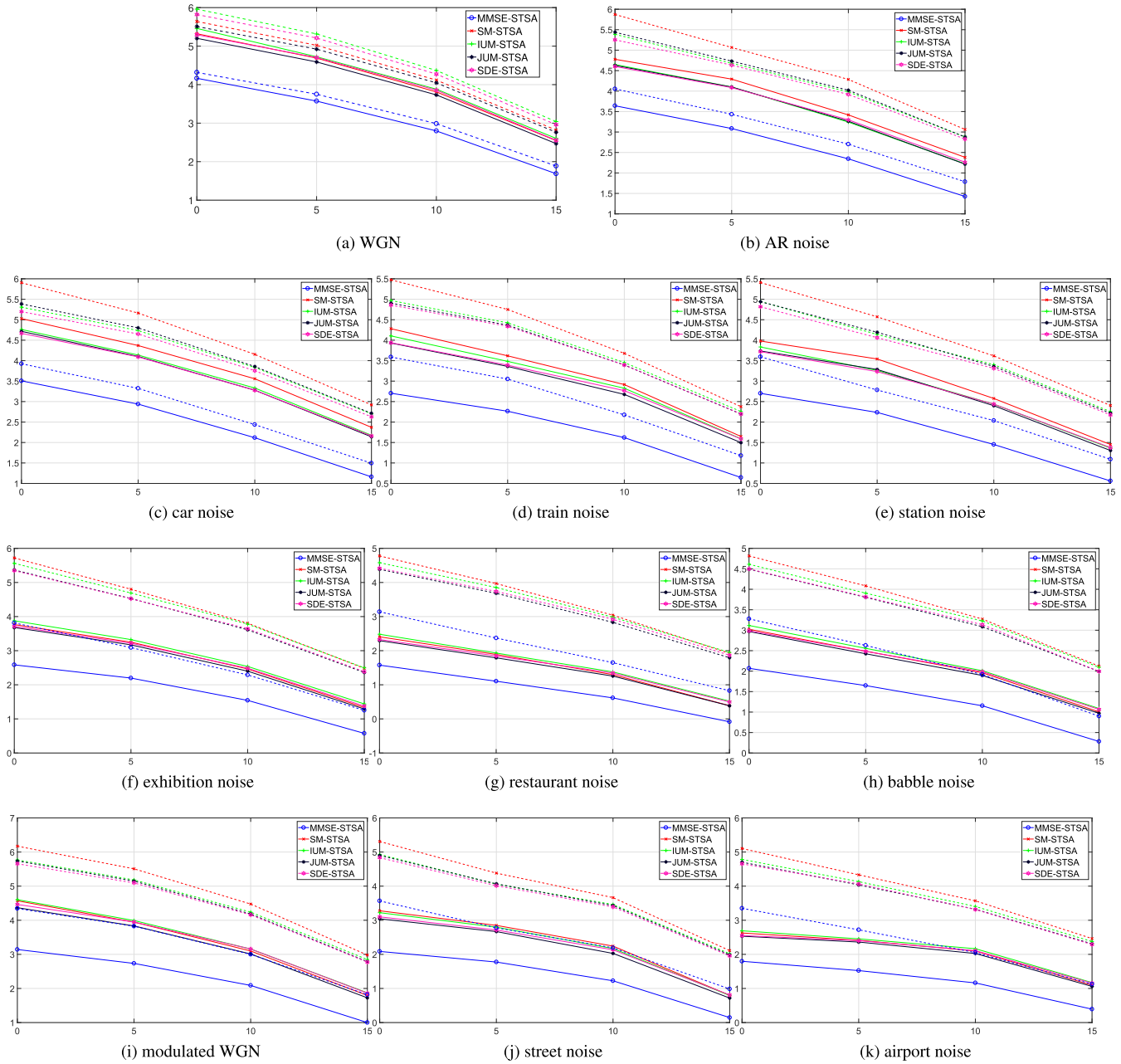

FIGURE 2. **Speech quality evaluation by SSNR improvement after speech denoising using STSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. Performance measurements of MMSE-STSA, SM-STSA, IUM-STSA, JUM-STSA and SDE-STSA with reference noise power spectrum are represented by dashed solid lines. The solid lines correspond to the performance measurements obtained by these same algorithms when the B-E-DATE is used to estimate the noise spectrum.**

to estimate the performance loss in denoising incurred by integrating an up-to-date noise estimator.

First, speech quality and intelligibility were evaluated via objective quality and intelligibility criteria. Speech quality was assessed using the standard segmental SNR (SSNR) (see Figure 2) [38, Eq. (2.22)], the Signal To Noise Ratio Improvement (SNRI) (see Figure 3) [38, Eq. (2.30)] and the overall speech quality criterion. SSNR values were trimmed so as to remain within the range $[-10,35$ dB$]$ and avoid the use of a silence/speech detector [35]. The overall speech quality was measured by the multivariate adaptive regression spline (MARSovrl) criterion (see Figure 4) [38, Eq. (2.31)]. This metric combines the Itakura-Saito distance (IS) [38, Eq. (2.23)] and the perceptual evaluation of speech quality (PESQ) [39, Eq. (3)]. It has been shown to strongly correlate with subjective assessments [39]. Speech intelligibility was estimated by the short-time objective intelligibility (STOI) criterion (see Figure 5) [38, Eqs. (2.39)]. Basically, the STOI criterion computes the mean correlation between the clean and the estimated speech [40]. It is known to be

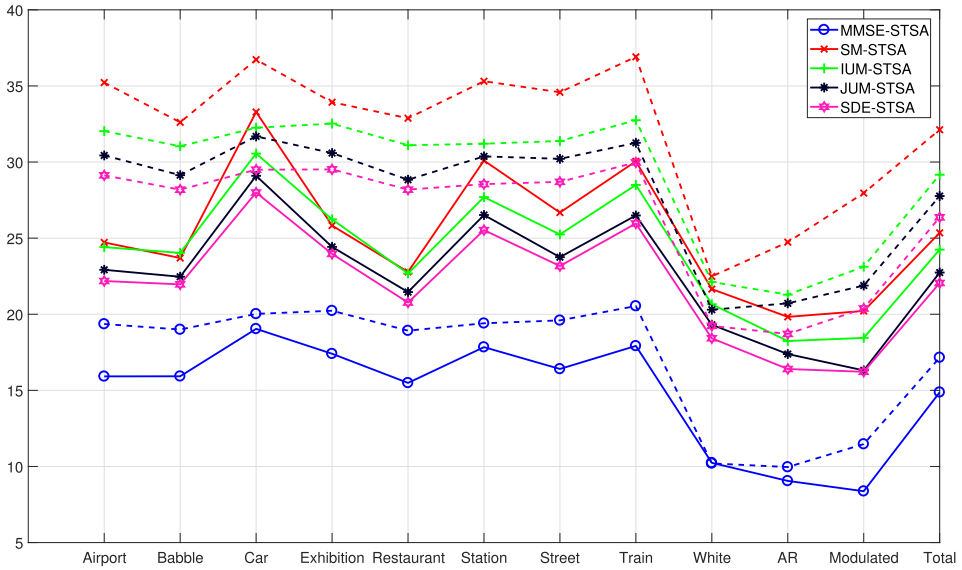

FIGURE 3. SNRI with various noise types for all STSA-based methods with and without the reference noise power spectrum. Dashed lines correspond to measurements obtained with reference noise power spectrum. The solid lines represent the performance measurements returned by the algorithms combined to B-E-DATE for noise spectrum estimation.

highly correlated with intelligibility scores obtained by listening tests. We applied the logistic function [38, Eq. (2.40)], as usual [40, Eq. (8)], to map the STOI measure to a meaningful intelligibility score. The interested reader can download the Matlab routines from [41] to compute these criteria. Second, in complement to these objective measurements, informal subjective tests were performed as well to assess speech intelligibility and quality.

A. SSNR IMPROVEMENT AND SNRI

Figure 2 displays the average SSNR improvement for different noise types and SNR levels and under the two scenarios mentioned above. In the ideal situation where noise is Gaussian and known, IUM-STSA yields the best score at all SNR levels shown (see Figure 2a). More specifically, JUM-STSA and SDE-STSA provide almost the same results, whereas IUM-STSA and SM-STSA perform slightly better, with an improvement of 0.25dB only. Compared to MMSE-STSA, the gain of the joint estimators is from 1dB to 1.8dB. In the more realistic case where noise power spectrum is estimated by B-E-DATE, the SSNR improvements obtained by the joint estimators are even closer. Their gain with respect to MMSE-STSA is now around 1dB. Not-so-good estimates of the noise spectrum can generate undesirable effects both at the detection and estimation steps of a speech joint estimator.

For AR noise and slowly-changing non-stationary (car, train and station) noise as in Figures 2b-2e, all joint estimators yield the same measure (SM-STSA slightly performs better) and outperform MMSE-STSA with a gain of around 1.5dB in the first scenario and a gain of around 1dB in the second scenario.

For fast-changing and speech-like non-stationary (modulated, street, airport, exhibition, restaurant and babble) noise, the joint estimators perfor similarly again, with a slight improvement gain achieved by IUM-STSA (see Figures 2f-2k). Globally, the gain brought by the joint estimators is around 1.5dB in the first scenario and 1dB in the second scenario in comparison to MMSE-STSA.

The second SNR-based criterion of interest is the SNRI. The results obtained by the methods considered in this paper are displayed in Figure 3. IUM-STSA provides the best overall SNRI in the two scenarios. For fast-changing non-stationary noise, the improvement obtained by the join estimators with respect to MMSE-STSA (resp. SDE-STSA) is from 6dB to 10dB (resp. 1dB to 4dB) when using B-E-DATE and from 8dB to 16dB (resp. 1dB to 4dB) when using the reference noise power. For stationary and slowly-changing non-stationary noise, in comparison to MMSE-STSA (resp. SDE-STSA) the gain is around 10dB (resp. 1dB to 4dB) when using noise power spectrum estimated by B-E-DATE and 11.5dB (resp. 1dB to 8dB) when using the reference noise power spectrum.

In summary, the joint estimators generally outperform MMSE-STSA in terms of SSNR improvement and SNRI in all situations, with an overall gain from 6dB to 10dB, which is emphasized by label "Total" in Figure 3. In comparison to SDE-STSA, IUM-STSA and JUM-STSA perform slightly better, whereas SM-STSA provides a gain from 3dB to 5dB.

B. MARSOVL

The measurements of the composite speech quality overall (MARSovrl) criterion are displayed in Figure 4. Consider first stationary (white and AR) noise. In the two scenarios,

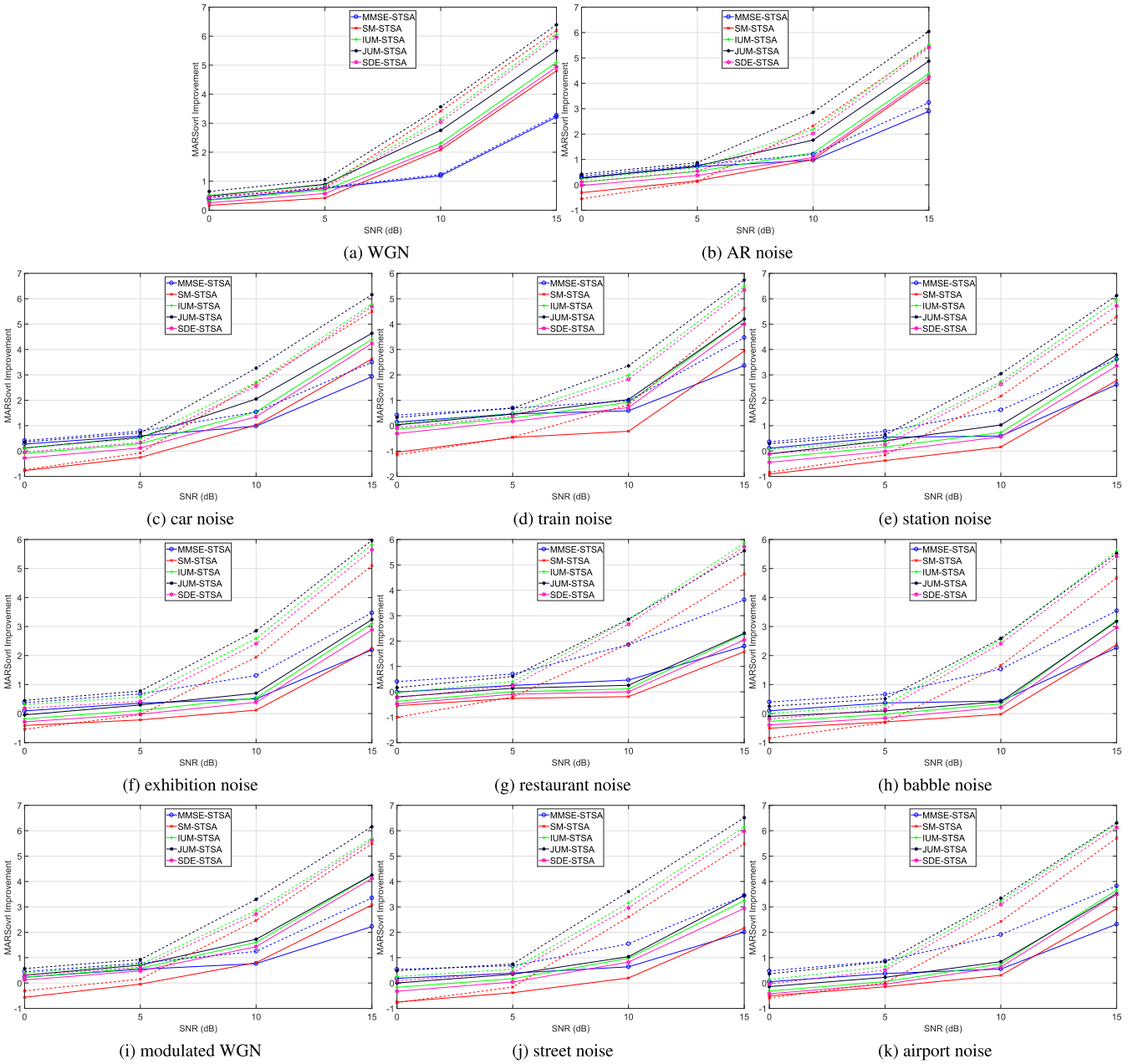

FIGURE 4. **Speech quality evaluation by MARSovrl improvement after speech denoising using STSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. As above, the dashed solid lines correspond to the performance measurements with known reference noise power spectrum, whereas the solid lines display the results obtained when the noise spectrum is estimated by B-E-DATE.**

at low SNR levels (0dB and 5dB), IUM-STSA and JUM-STSA yield the same score and outperform SM-STSA, MMSE-STSA and SDE-STSA (see Figures 4a-4b). However, the gain is not significant. At high SNR levels, still for the two scenarios, the proposed joint estimators outperform MMSE-STSA and JUM-STSA yields the best score.

For slowly-changing non-stationary noise, yet in the two scenarios, at low SNR levels, JUM-STSA and MMSE-STSA yield the same measures and perform slightly better than SM-STSA, IUM-STSA and SDE-STSA. At high SNR levels, all joint estimators outperform MMSE-STSA, except at 10dB for train noise when the noise power spectrum is estimated by B-E-DATE (see Figure 4d). In this case, JUM-STSA and IUM-STSA achieve the best results with a small gain compared to SDE-STSA, whereas SM-STSA and MMSE-STSA perform similarly.

In the case of fast-changing and speech-like non-stationary noise, when the estimators are combined with B-E-DATE, all the methods provide similar scores at low SNR levels, even at 10dB, except for modulated WGN where JUM-STSA returns significantly better results. The relevance of joint

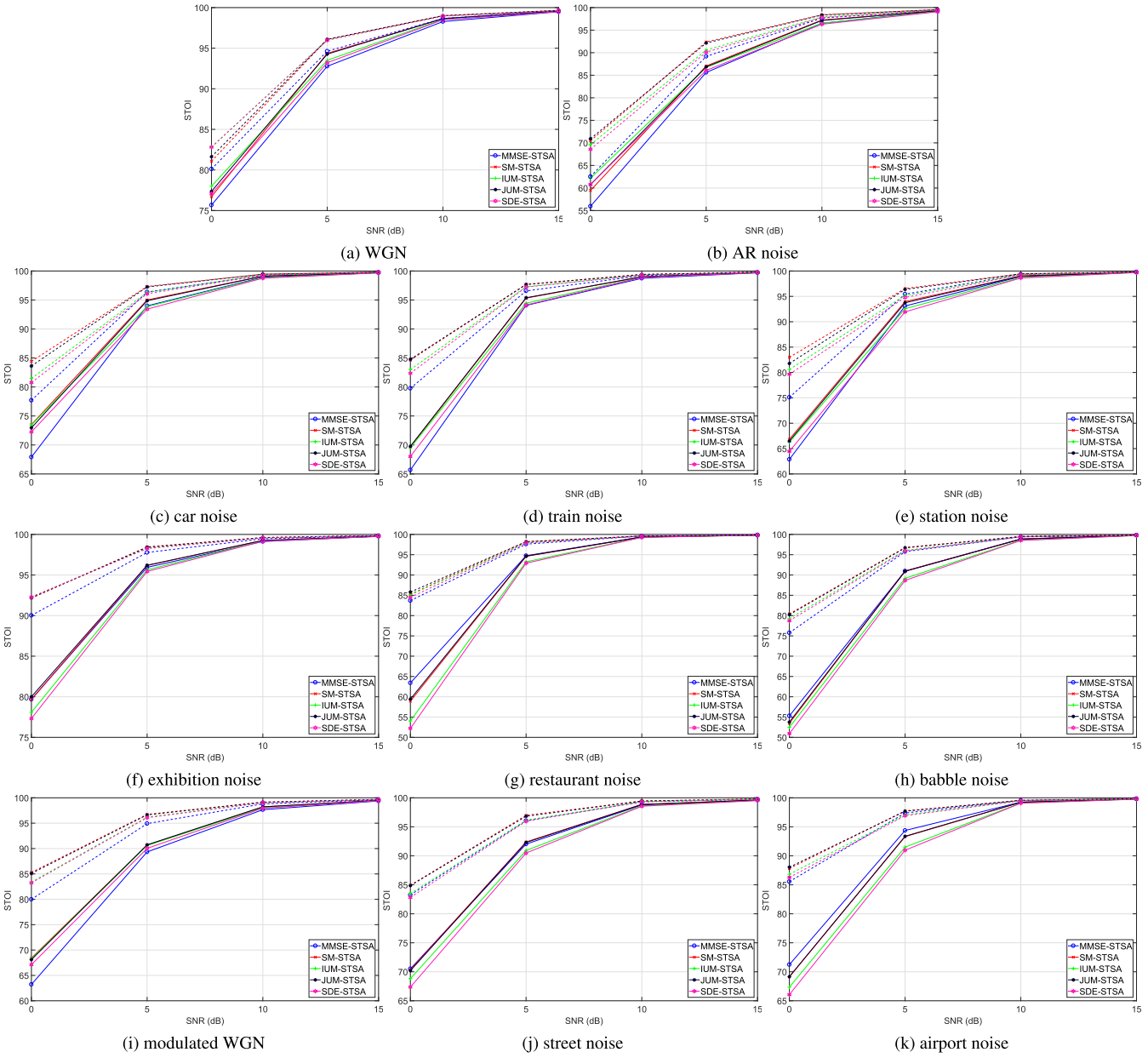

FIGURE 5. Speech intelligibility evaluation by STOI after speech denoising using STSA-based methods for stationary, slowly-changing,speech-like and fast-changing non-stationary noise. The legend is the same as that of Figures 2 and 4.

detection/estimation is confirmed at 15dB (see Figures 4f-4k) only. However, when using the reference noise power spectrum, for 10dB and 15dB, a significant gain is yielded by joint estimators, in comparison to MMSE-STSA for almost all types of noise. This emphasizes that inaccurate noise estimates may induce erroneous decisions. At low SNR levels, JUM-STSA and MMSE-STSA perform slightly better than the other methods. Note that JUM-STSA provides always the best score in this scenario.

In terms of overall speech quality, the joint estimators outperform MMSE-STSA in almost all cases. JUM-STSA performs generally better than IUM-STSA, SM-STSA and SDE-STSA. It thus turns out when the null hypothesis is accepted, providing an estimation of the speech signal better improves speech quality than forcing the estimated amplitude to 0.

C. STOI

Finally, the intelligibility score (IS) obtained by mapping the STOI measure is shown in Figure 5. For 10dB and 15dB, all the methods yield the same performance in the two scenarios. At 5dB, the differences in performance are not really significant. Therefore, we hereafter focus on the results at 0dB.

For stationary (white and AR) and slowly-changing non-stationary (car, train and station) noises, the proposed SM-STSA, JUM-STSA and IUM-STSA give a small gain compared to SDE-STSA and yield a better score than MMSE-STSA. The IS scores of these methods exceed by 5% to 10% that of MMSE-STSA, whatever the scenario for noise spectrum estimation.

For the fast-changing airport and speech-like non-stationary noises, SM-STSA, JUM-STSA and MMSE-STSA perform equally and better than IUM-STSA and SDE-STSA in the second scenario. In the first scenario, the joint estimators (including SDE-STSA) yield the same performance and outperform MMSE-STSA. For modulated WGN and street noises, in the two scenarios, the gain in IS is from 4% to 7% when using SM-STSA and JUM-STSA in comparison to MMSE-STSA (see Figures 5f-5k).

In conclusion, as in the preceding subsection, SM-STSA, JUM-STSA and IUM-STSA yield better results than MMSE-STSA and slightly outperform SDE-STSA. Note however that JUM-STSA tends to outperform the others.

D. INFORMAL SUBJECTIVE ASSESSMENT

The authors have conducted informal subjective experiments on the same database, by involving the 5 methods considered in the paper (IUM-STSA,JUM-STSA,SM-STSA, SDE-STSA, MMSE-STSA). In these experiments, we considered 4 types of noise (white, babble, street and train) with thus a strong emphasis on non-stationary noise since it is the case of main practical interest. We tested 4 SNRs (0, 5, 10, 15 dB) and evaluated the methods when the noise spectrum is known versus the case where this spectrum is unknown and estimated. In our subjective and informal evaluations, we empirically assessed the intelligibility and quality of the denoised sentences.

These experiments confirm that SM-STSA, JUM-STSA, IUM-STSA generally outperform MMSE-STSA. By and large, JUM-STSA, IUM-STSA, SM-STSA and SDE-STSA perform similarly and, in some cases, SM-STSA provides some slight improvement. In comparison with MMSE-STSA, the gain brought by the methods proposed in this paper are especially noticeable at low and medium SNRs. This subjective assessment is thus consistent with the attenuation curves displayed in Figure 1 and the tendency emphasized by the objective SNRI measurements of Section VI-A. The main difference in performance between MMSE-STSA and the group of methods (JUM-STSA, IUM-STSA, SM-STSA) is that these latter, in contrast to MMSE-STSA, do not introduce audible distortions, even for high SNRs. The loss induced by estimating the noise spectrum remains limited, does not modify the conclusions above and, in any case, remain imperceptible for SNR $\geq$ 10 dB.

VII. CONCLUSION AND PROSPECTS

This paper has proposed a unified framework for speech enhancement based on the optimal combination of Neyman-Pearson detectors and Bayesian speech estimators of speech in noise. The key idea is to take the presence and absence of speech in each time-frequency bin into account so as to improve speech quality and intelligibility in noisy environments. In contrast to the optimal Bayesian joint detection/estimation [29], the Neyman-Pearson test performs the detection without prior knowledge of the speech presence probability.

Several joint estimators resulting from this combination have been derived for speech STSA estimation. When absence of speech is decided, they force the estimate to zero or to a small spectral floor for avoiding musical noise. These algorithms require the false alarm probability α specified for the Neyman-Pearson test. This parameter can easily be set empirically, once for all and without much effort, via preliminary and limited experiments, or by simply resorting to the value proposed in Section VI. The constant spectral floor β needed to perform IUM-STSA and JUM-STSA can be set on the basis of experimental results exposed in the literature [35].

The objective performance evaluation was conducted in two scenarios, one when the reference noise power spectrum is used and one when noise is estimated by an up-to-date method. The experimental results show the relevance of the approach. Without prior knowledge and estimation of the speech presence probability and thus, with a reduced number of parameters, the optimal joint estimators proposed in this paper do not induce any performance loss and even yield some improvement with respect to [29], [30]. Actually, although they are not stricto sensu parameter-free, they need one single parameter only, namely, the false alarm probability α. We have not conducted an exhaustive analysis of the possible range for this parameter, but a value of $\alpha = 0.05$ provides a correct order of magnitude for it. Therefore, the results presented in this paper, in addition to [29], [30], emphasize the interest of joint detection/estimation in speech enhancement. In a nutshell, although it is well-known that a deep understanding of the signals under consideration leads to very good results and even optimality by Bayesian approaches, the Bayesian joint detection/estimation methods considered above are generic and provide a good trade-off between performance and robustness with relatively coarse models for the speech signals and noise. In particular, these joint estimators are of practical interest for their performance and easy tuning. The choice between SM-STSA, IUM-STSA and JUM-STSA can then be ruled by the type of criterion the practitioner wishes to optimize. The informal subjective assessment is consistent with SNRI measurements and confirms the interest of joint detection/estimation procedures.

At this stage, it can be wondered to what extent other costs, such as logarithm functions, could also be considered. Elements of answer to this question can be given along the following lines, within a larger perspective on the prospects opened by this work.

To begin with, the approach followed in this paper is very generic. Although we have dealt with standard quadratic costs on the amplitudes of the speech STFT coefficients, our

theoretical framework in the vein of [31] can be adapted to other time-frequency representations, such as those based on the windowed Discrete Cosine Transform (DCT). It can also be applied to time-scale representations after wavelet transformations, either continuous or discrete. In this vein, although we experimentally used standard values for the STFT parameters, it would be desirable to study to what extent transform parameters can be optimized with respect to joint speech detection/estimation performance.

Because it is generic, the approach can probably be applied to signals other than speech, the crux being to have a reasonable statistical model for the signal and the noise coefficients returned by the transformation. In this respect, it may be profitable to look for probability distributions other than Gaussian. In particular, the Gamma distribution seems to be a good candidate for modeling speech DFT coefficients [15].

The generic nature of the approach allows for costs other than those considered above. In particular, as mentioned by a reviewer, the Log-MMSE-STSA [12], aimed at taking psycho-physical properties of the auditory system into account, improves by one dB the standard MMSE-STSA. Therefore, according to Section VI, the joint estimators considered in this paper yield the same performance order as the Log-MMSE-STSA. Consequently, joint detection/estimation with costs on the speech logarithmic spectral amplitudes (LSA) can be expected to perform better than SM-, IUM- and JUM-STSA. A paper dedicated to this topic is in-progress. The interested reader can however refer to [38, Chapter 4] to note that the approach can actually be profitably extended to quadratic costs on LSA.

APPENDIXES

APPENDIX A
STRICT MODEL

In the strict presence/absence model, the threshold τ^{SM} is chosen by fixing the false alarm probability to a specified value α and by solving (11), which becomes in our case $\mathbf{R}_0\left(\psi_1^{\mathrm{SM}}, \psi_0^{\mathrm{SM}}, \delta^{\mathrm{SM}}\right) = \alpha$. Since $\psi_0^{\mathrm{SM}} = 0$, it follows from (3), (4), (17), (23) and Table 2 that:

$$\begin{aligned}\mathbf{R}_0\left(\psi_1^{\mathrm{SM}}, \psi_0^{*}, \delta^{\mathrm{SM}}\right) &= \mathbf{E}_0\left[\delta^{\mathrm{SM}}(Y)\right] \\ &= \mathbb{P}\left[\mathcal{D}^{\mathrm{SM}}(Y) \geqslant \tau^{\mathrm{SM}}\right] \\ &= \mathbb{P}\left[A_Y^2 \geqslant \tau^{\mathrm{SM}}/(\lambda(\xi,\gamma)G^2(\xi,\gamma))\right]\end{aligned}$$

According to the Gaussian assumption (13), the pdf of A_Y under H_0 is Rayleigh [35, p.212] and given by:

$$f_{A_Y}(a; H_0) = \frac{2a}{\sigma_X^2}\exp\left(-\frac{a^2}{\sigma_X^2}\right). \tag{48}$$

By injecting this density into the value of $\mathbf{R}_0(\psi_1^{\mathrm{SM}}, \psi_0^{\mathrm{SM}}, \delta^{\mathrm{SM}})$ above, some routine algebra leads to (24).

APPENDIX B
UNCERTAIN MODEL

As a preliminary result, it follows from (30) that the pdf of A_Y under H_0 is given by:

$$f_{A_Y}(a; H_0) = \frac{2a}{\sigma_X^2(1+\beta)}\exp\left(-\frac{a^2}{\sigma_X^2(1+\beta)}\right). \tag{49}$$

A. INDEPENDENT ESTIMATORS

With regard to (35) and by referring to the likelihood ratio (32), the impact of τ on the gain function G^{IUM} is described as follows. First, when $\Lambda(\xi,\gamma) \gg \tau$, we can approximate the gain function by $G^{\mathrm{IUM}}(\xi,\gamma) \simeq G(\xi,\gamma)$, which amounts to considering that the speech is significantly present. Second, when $\Lambda(\xi,\gamma) \ll \tau$, the gain function is approximated by $G^{\mathrm{IUM}}(\xi,\gamma) \simeq G(\beta,\gamma)$, which corresponds to the case where the observation pertains to speech with low amplitude.

The foregoing can thus be interpreted as a decision rule to discriminate high from low speech energies. In this respect, τ is a threshold on the likelihood ratio to make such a decision and can therefore be calculated so as to guarantee a specified false alarm probability α. According to (49), the false alarm probability equals α if $\mathbb{P}\left[\Lambda(\xi,\gamma) \geq \tau\right] = \alpha$. We then derive from (32) and the definition of γ that the foregoing equality is equivalent to:

$$\mathbb{P}\left[A_Y^2 \geq \log\left(\tau\frac{1+\xi}{1+\beta}\right)\left(\sigma_X^2\frac{(1+\beta)(1+\xi)}{(\xi-\beta)}\right)\right] = \alpha$$

A computation similar to that carried out in Appendix A leads to $\tau = \Lambda(\xi,\gamma_0)$ with $\gamma_0 = -(1+\beta)\log(\alpha)$.

B. JOINT ESTIMATOR

As in the preceeding subsection, we must solve

$$\mathbf{R}_0(\psi_1^{\mathrm{JUM}}, \psi_0^{\mathrm{JUM}}, \delta^{\mathrm{JUM}}) = \alpha,$$

where ψ_1^{JUM} and ψ_0^{JUM} are given by (40). According to (3), (4), (41), (44) and Table 4, we have:

$$\begin{aligned}&\mathbf{R}_0\left(\psi_1^{\mathrm{JUM}}, \psi_0^{\mathrm{JUM}}, \delta^{\mathrm{JUM}}\right) \\ &= \mathbf{E}_0\left[A_0^2\delta^{\mathrm{JUM}}(Y)\right] + \mathbf{E}_0\left[\left(\widehat{A}_0 - A_0\right)^2(1-\delta^{\mathrm{JUM}}(Y))\right] \\ &= \mathbf{E}_0\left[A_0^2\delta^{\mathrm{JUM}}(Y)\right] + \mathbf{E}_0\left[\left(G(\beta,\gamma)A_Y - A_0\right)^2(1-\delta^{\mathrm{JUM}}(Y))\right]\end{aligned}$$

We begin by computing the second term to the rhs in the equality above. The standard chain rule yields:

$$\begin{aligned}&\mathbf{E}_0\left[\left(G(\beta,\gamma)A_Y - A_0\right)^2\left(1-\delta^{\mathrm{JUM}}(Y)\right)\right] \\ &= \int (1-\delta^{\mathrm{JUM}}(y))\mathbf{E}\left[\left(G(\beta,\gamma)|y| - A_0\right)^2 | Y = y\right] f_Y(y; H_0)dy\end{aligned} \tag{50}$$

We have:

$$\begin{aligned}&\mathbf{E}\left[\left(G(\beta,\gamma)|y| - A_0\right)^2 | Y = y\right] \\ &= \int \left(G(\beta,\gamma)|y| - a_0\right)^2 f_{A_0|Y=y}(a_0)da_0 \qquad (51) \\ &= \left(G_0^2 - G^2(\beta,\gamma)\right)|y|^2 \qquad (52)\end{aligned}$$

where (52) is obtained by expanding the integrand in (51), using (40) and (41) and injecting the equality [35, Eq.(7.94)]:

$$\mathbf{E}\left[A_0^2|Y=y\right]=\frac{\beta}{1+\beta}\left(\frac{1+\nu_\beta}{\gamma}\right)|y|^2=G_0^2|y|^2, \quad (53)$$

with $G_0^2=\frac{\beta}{1+\beta}\left(\frac{1+\nu_\beta}{\gamma}\right)$ and $\nu_\beta=\gamma\beta/(1+\beta)$. It now follows from (53) that:

$$\mathbf{E}_0\left[A_0^2\delta^{\mathrm{JUM}}(Y)\right]=\int G_0^2|y|^2\delta^{\mathrm{JUM}}(y)f_Y(y;H_0)dy \quad (54)$$

According to (50), (52) and (54), we obtain:

$$\begin{aligned}&\mathbf{R}_0(\psi_1^{\mathrm{JUM}},\psi_0^{\mathrm{JUM}},\delta^{\mathrm{JUM}})\\&=\int\delta^{\mathrm{JUM}}(y)G_0^2|y|^2f_Y(y;H_0)dy\\&\quad+\int(1-\delta^{\mathrm{JUM}}(y))(G_0^2-G^2(\beta,\gamma))|y|^2f_Y(y;H_0)dy \quad (55)\end{aligned}$$

This risk could be numerically calculated so as to determine τ^{JUM}. However, we can resort to the following approximations to get a close form for an estimate of this threshold.

First, G_0 and $G(\beta,\gamma)$ are actual functions of y. However, $0\leq\beta\ll 1$ so that G_0 and $G(\beta,\gamma)$ both tend to 0. Therefore, G_0^2 and $|G_0^2-G^2(\beta,\gamma)|$ can both be upper-bounded by small constants so that we approximate:

$$\begin{aligned}&\mathbf{R}_0(\psi_0^{\mathrm{JUM}},\psi_1^{\mathrm{JUM}},\delta^{\mathrm{JUM}})\\&\approx G_0^2\int\delta^{\mathrm{JUM}}(y)|y|^2f_Y(y;H_0)dy\\&\quad+(G_0^2-G^2(\beta,\gamma))\int(1-\delta^{\mathrm{JUM}}(y))|y|^2f_Y(y;H_0)dy \quad (56)\end{aligned}$$

Second, it can be numerically verified that $\mathcal{D}^{\mathrm{JUM}}$ does not decrease with $|y|$ when $\beta\ll\xi$, which is the case of practical interest. As a second approximation, we consider that the testing performed by δ^{JUM} amounts to comparing $|y|$ to a threshold τ^*. Therefore, using (30), the Bayesian risk $\mathbf{R}_0(\psi_0^{\mathrm{JUM}},\psi_1^{\mathrm{JUM}},\delta^{\mathrm{JUM}})$ can further be estimated by:

$$\begin{aligned}&\widehat{\mathbf{R}}_0(\psi_0^{\mathrm{JUM}},\psi_1^{\mathrm{JUM}},\delta^{\mathrm{JUM}})\\&=G_0\int_{\tau_*}^{\infty}\frac{2r^3}{\sigma_X^2(1+\beta)}\exp\left(-\frac{r^2}{\sigma_X^2(1+\beta)}\right)dr\\&\quad+\Big(G_0-G^2(\beta,\gamma)\Big)\int_0^{\tau_*}\frac{2r^3}{\sigma_X^2(1+\beta)}\exp\left(-\frac{r^2}{\sigma_X^2(1+\beta)}\right)dr.\end{aligned}$$

After a change of variable and an integration by parts, some routine algebra leads to:

$$\begin{aligned}&\widehat{\mathbf{R}}_0(\psi_0^{\mathrm{JUM}},\psi_1^{\mathrm{JUM}},\delta^{\mathrm{JUM}})\\&=\sigma_X^2(1+\beta)(G_0^2-G^2(\beta,\gamma))\\&\quad+\Big(G^2(\beta,\gamma)\tau_*^2+\sigma_X^2(1+\beta)\Big)\exp\left(-\frac{\tau_*^2}{\sigma_X^2(1+\beta)}\right).\end{aligned}$$

We can seek a numerical solution $\tau_*(\alpha)$ to

$$\widehat{\mathbf{R}}_0(\psi_0^{\mathrm{JUM}},\psi_1^{\mathrm{JUM}},\delta^{\mathrm{JUM}})=\alpha\sigma_X^2.$$

Alternatively, since $G(\beta,\gamma)\approx 0$, we suppose $G^2(\beta,\gamma)=0$. This leads to

$$\tau_*=\begin{cases}\sigma_X\sqrt{\log\left(\frac{1}{\alpha-G_0^2(1+\beta)}\right)(1+\beta)} & \text{if } \gamma>\gamma_0\\ \tau_\infty & \text{otherwise}\end{cases} \quad (57)$$

with $\gamma_0=\frac{\beta(1+\beta)}{\alpha(1+\beta)-\beta^2}$. Since β is small, γ_0 is itself small. For instance, if $\beta=0.002$ and $\alpha=0.05$, $\gamma_0\approx 0.04$, which corresponds to an SNR of -14dB. This means that the first case in (57) embraces most of the situations encountered in practice. The second case corresponds to the presence of noise only or the presence of speech with low energy in noise. In this case, τ_* is fixed to a large value τ_∞. The detection threshold $\tau^{\mathrm{JUM}}(\alpha)$ is then approximated by:

$$\tau^{\mathrm{JUM}}(\alpha)=\mathcal{D}^{\mathrm{JUM}}(\tau_*) \quad (58)$$

REFERENCES

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.

[2] T. T. Vu, B. Bigot, and E. S. Chng, "Combining non-negative matrix factorization and deep neural networks for speech enhancement and automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, Mar. 2016, pp. 499–503.

[3] T. G. Kang, J. W. Shin, and N. S. Kim, "DNN-based monaural speech enhancement with temporal and spectral variations equalization," *Digit. Signal Process.*, vol. 74, pp. 102–110, Mar. 2018.

[4] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process. (ICASSP)*, Mar. 2008, pp. 4029–4032.

[5] S. Mavaddaty, S. Seyedin, and S. M. Ahadi, "Modified coherence-based dictionary learning method for speech enhancement," *IET Signal Process.*, vol. 9, no. 7, pp. 537–545, Sep. 2015.

[6] K. M. Jeon and H. K. Kim, "Audio enhancement using local SNR-based sparse binary mask estimation and spectral imputation," *Digit. Signal Process.*, vol. 68, pp. 138–151, Sep. 2017.

[7] C. K. A. Reddy, N. Shankar, G. S. Bhat, R. Charan, and I. Panahi, "An individualized super-Gaussian single microphone speech enhancement for hearing aid users with smartphone as an assistive device," *IEEE Signal Process. Lett.*, vol. 24, no. 11, pp. 1601–1605, Nov. 2017.

[8] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, nos. 7–8, pp. 588–601, Jul. 2007.

[9] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[10] C. H. You, S. N. Koh, and S. Rahardja, "β-order MMSE spectral amplitude estimation for speech enhancement," *IEEE Trans. Speech, Audio, Process.*, vol. 13, no. 4, pp. 475–486, Jul. 2005.

[11] P. Loizou, "Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 857–869, Sep. 2005.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[13] E. Plourde and B. Champagne, "Auditory-based spectral amplitude estimators for speech enhancement," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 8, pp. 1614–1623, Nov. 2008.

[14] P. Mowlaee, J. Stahl, and J. Kulmer, "Iterative joint MAP single-channel speech enhancement given non-uniform phase prior," *Speech Commun.*, vol. 86, pp. 85–96, Feb. 2017.

[15] I. Andrianakis and P. White, "Speech spectral amplitude estimators using optimally shaped Gamma and Chi priors," *Speech Commun.*, vol. 51, no. 1, pp. 1–14, Jan. 2009.

[16] B. Chen and P. C. Loizou, "A Laplacian-based MMSE estimator for speech enhancement," *Speech Commun.*, vol. 49, no. 2, pp. 134–143, Feb. 2007.

[17] J. S. Erkelens, R. C. Hendriks, R. Heusdens, and J. Jensen, "Minimum mean-square error estimation of discrete Fourier coefficients with generalized gamma priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 6, pp. 1741–1752, Aug. 2007.
[18] B. J. Borgström and A. Alwan, "Log-spectral amplitude estimation with generalized gamma distributions for speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, May 2011, pp. 4756–4759.
[19] I. Andrianakis and P. R. White, "MMSE speech spectral amplitude estimators with chi and gamma speech priors," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, vol. 3, May 2006, pp. 1068–1071.
[20] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal, Process.*, vol. 2, Mar. 1999, pp. 789–792.
[21] N. Soo Kim and J.-H. Chang, "Spectral enhancement based on global soft decision," *IEEE Signal Process. Lett.*, vol. 7, no. 5, pp. 108–110, May 2000.
[22] H. Taşmaz and E. Erçelebi, "Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE–STSA estimation in various noise environments," *Digit. Signal Process.*, vol. 18, no. 5, pp. 797–812, 2008.
[23] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, no. 4, pp. 113–116, Apr. 2002.
[24] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.
[25] P. C. Loizou and G. Kim, "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 1, pp. 47–56, Jan. 2011.
[26] J. Jensen and R. C. Hendriks, "Spectral magnitude minimum mean-square error estimation using binary and continuous gain functions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 92–102, Jan. 2012.
[27] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Interspeech*, 2006, pp. 1447–1450.
[28] V.-K. Mai, D. Pastor, A. Aïssa-El-Bey, and R. Le Bidan, "Semi-parametric joint detection and estimation for speech enhancement based on minimum mean square error," *Speech Commun.*, vol. 102, pp. 27–38, Sep. 2018.
[29] A. Abramson and I. Cohen, "Simultaneous detection and estimation approach for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 8, pp. 2348–2359, Nov. 2007.
[30] H. Momeni, H. R. Abutalebi, and A. Tadaion, "Joint detection and estimation of speech spectral amplitude using noncontinuous gain functions," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 8, pp. 1249–1258, Aug. 2015.
[31] G. V. Moustakides, G. H. Jajamovich, A. Tajer, and X. Wang, "Joint detection and estimation: Optimum tests and applications," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4215–4229, Jul. 2012.
[32] Y. Hu and P. C. Loizou, "Techniques for estimating the ideal binary mask," in *Proc. 11th Int. Workshop Acoust. Echo Noise Control*, 2008, pp. 154–157.
[33] A. Aziz-Sbaï, S. M. Aïssa-El-Bey, and D. Pastor, "Contribution of statistical tests to sparseness-based blind source separation," *EURASIP J. Adv. Signal, Process.*, vol. 2012, no. 1, pp. 1–15, 2012.
[34] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 4, Mar. 1979, pp. 208–211.
[35] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2013.
[36] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 3, pp. 541–553, Mar. 2008.
[37] V.-K. Mai, D. Pastor, A. Aïssa-El-Bey, and R. Le-Bidan, "Robust estimation of non-stationary noise power spectrum for speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 670–682, Apr. 2015.
[38] V. K. Mai, "Advanced methods of speech processing and noise reduction for mobile devices," Ph.D. dissertation, IMT Atlantique, Nantes, France, 2017. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01810623
[39] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[40] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 7, pp. 2125–2136, Feb. 2011.
[41] *Objective Measures for Predicting Speech Intelligibility*. Accessed: Jan. 2020. [Online]. Available: https://ecs.utdallas.edu/loizou/speech/software.htm

VAN-KHANH MAI was born in Vietnam, in 1987. He received the Engineering degree in electronic and information from the Hanoi University of Technology, Hanoi, Vietnam, the Research Master's degree in electronics and telecommunications from Rennes I University, Rennes, France, in 2013, and the Ph.D. degree in signal processing from IMT Atlantique, France, in 2017. He is currently a DSP and an Audio Innovation Engineer with Arkamys. His research interests include audio signal processing, noise reduction, and speech enhancement.

DOMINIQUE PASTOR (Member, IEEE) was born in Cahors, France, in 1963. He received the degree from Telecom Bretagne, Brest, France, in 1986, and the Ph.D. degree from the University of Rennes, France, in 1997. From 1987 to 2000, he was with Thales. From 1990 to 1998, he was with Thales Avionics, where his research concerned speech processing for applications to speech recognition systems embedded in military fast jet cockpits and, from 1998 to 2000, he was with Thales, The Netherlands, where he worked on the detection of radar targets in sea clutter. In September 2000, he joined Altran Technologies, The Netherlands, as a Senior Consultant. Since September 2002, he has been with Institut Telecom, where he is currently a Professor at IMT Atlantique (Telecom Bretagne). His current research interests include statistical signal processing and sparse transforms with applications to physiological signals including speech.

ABDELDJALIL AISSA-EL-BEY (Senior Member, IEEE) was born in Algiers, Algeria, in 1981. He received the State Engineering degree from Ecole Nationale Polytechnique (ENP), Algiers, in 2003, the M.S. degree in signal processing from Supelec and Paris XI University, Orsay, France, in 2004, and the Ph.D. degree in signal and image processing from Telecom Paris, France, in 2007. In 2007, he joined the Signal and Communications Department, IMT Atlantique (Telecom Bretagne), Brest, France, as an Associate Professor, and then a Full Professor, since 2015. He was a Visiting Researcher at Fujitsu Laboratories, Japan, and the Department of Electrical and Electronic Engineering, The University of Melbourne, Australia, in 2010 and 2015, respectively. His research interests include blind source separation, blind system identification and equalization, compressed sensing, sparse signal processing, statistical signal processing, wireless communications, and adaptive filtering.