Gaussian Model-Based Multichannel Speech Presence Probability

Mehrez Souden, *Student Member, IEEE*,
Jingdong Chen, *Senior Member, IEEE*,
Jacob Benesty, *Senior Member, IEEE*, and
Sofiène Affes, *Senior Member, IEEE*

***Abstract*—The knowledge of the target speech presence probability in a mixture of signals captured by a speech communication system is of paramount importance in several applications including reliable noise reduction algorithms. In this correspondence, we establish a new expression for speech presence probability when an array of microphones with an arbitrary geometry is used. Our study is based on the assumption of the Gaussian statistical model for all signals and involves the noise and noisy data statistics only. In comparison with the single-channel case, the new proposed multichannel approach can significantly increase the detection accuracy. In particular, when the additive noise is spatially coherent, perfect speech presence detection is theoretically possible, while when the noise is spatially white, a coherent summation of speech components is performed to allow for enhanced speech presence probability estimation.**

***Index Terms*—Microphone array, noise reduction, speech detection, speech presence probability.**

I. Introduction

Microphone array signal processing has attracted a significant amount of research attention over the last few decades. Indeed, the extra spatial dimension inherent to the array spatial aperture results in more degrees of freedom and additional key functions can be ensured as contrasted to the classical single-channel processing. Well known functions include source localization [1], [2], noise reduction with low or even no speech distortion [1], [3], [4], multichannel inverse filtering, and dereverberation [5]. This paper is concerned with the utilization of microphone arrays for accurate speech presence probability estimation in adverse conditions.

The estimation of the speech presence probability is one of the key components in noise reduction algorithms [6]. For example, with the exact knowledge of whether the speech is present or not in a certain frame, accurate updates of the noise power spectrum density (PSD) matrix can be properly performed, thereby mitigating uncontrolled artificial distortions of the speech signal due to its leakage into noise statistics estimates. A common trend in the literature has been to use a single channel to estimate the speech presence probability. In [7] and [8], the speech presence uncertainty in the noisy observation was taken into account while deriving the minimum mean-squared error (MMSE) estimator. This uncertainty is found by using a Gaussian statistical model. In [9], Cohen proposed a robust single-channel noise tracking method that relies on the signal presence probability. In [4], Gannot and Cohen proposed a log-spectral amplitude post-filtering for the multichannel generalized-sidelobe canceller that involves a single-channel-based speech presence probability. Another notable contribution was proposed in [10] where Potamitis developed a multichannel speech presence probability for a uniform linear array with the assumptions of far-field, known source location, and spatially white noise. However, these assumptions are very restrictive. In addition, reverberation and other types of noise (such as point source interference which is spatially coherent) are ubiquitous in realistic environments.

In this correspondence, we investigate the speech presence probability when an array with arbitrary geometry and arbitrary number of microphones is used. We propose a simplified treatment that jointly considers the overall microphone outputs in order to decide whether the speech is present in a mixture of observed signals or not. Our approach is based on the Gaussian statistical model and applies for a general noise type and array geometry. We show the advantage of using multiple microphones to increase the accuracy of speech detection. Particularly, we prove that a perfect speech detection is theoretically possible when the noise is fully coherent. In the case of non-coherent noise, a constructive summation of all speech components is performed to enhance the speech detectability.

The rest of this paper is organized as follows. Section II describes the investigated problem with an emphasis on the importance of accurate speech presence probability estimation. Section III provides all the steps leading to the new proposed expression of the multichannel speech presence probability. Section IV illustrates the effectiveness of the proposed multichannel speech presence probability expression through numerical examples. Finally, Section V contains some concluding remarks.

II. Problem Statement

Let $s(t)$ denote a speech signal impinging on an array of N microphones with an arbitrary geometry. The resulting observations are given by

$$\begin{aligned} y_n(t) &= g_n * s(t) + v_n(t) \\ &= x_n(t) + v_n(t),\ n = 1, 2, \ldots, N \end{aligned} \quad (1)$$

where $*$ is the convolution operator, g_n is the channel impulse response encountered by the source before impinging on the nth microphone, $x_n(t) \triangleq g_n * s(t)$ is the noise-free speech component, and $v_n(t)$ is the noise at microphone n [the noise can be either white or colored, but is uncorrelated with $s(t)$]. We assume that all the noise components and $s(t)$ are zero-mean random processes. The short-time Fourier transform (STFT) is commonly utilized instead of the discrete-time Fourier transform (DTFT) in several processing schemes, including the implementation of noise reduction filters [4], [11]. Specifically, all the microphone outputs are chopped into small frames, sufficiently zero-padded, and transformed to the frequency domain via a K-length STFT

$$Y_n(k,l) = X_n(k,l) + V_n(k,l),\ n = 1, 2, \ldots, N \quad (2)$$

where $k \in \{0, \ldots, K-1\}$ is the frequency index and l is the time-frame index. We also have $X_n(k,l) = G_n(k,l)S(k,l)$ and $\mathbf{g}(k) \triangleq [G_1(k) \cdots G_N(k)]^T$ is the STFT of the transfer functions of the propagation channels between the source and all microphone locations, where ^{T} denotes the transpose operator. In addition, we use the following vector notations: $\mathbf{y}(k,l) \triangleq [Y_1(k,l) \cdots Y_N(k,l)]^T$, $\mathbf{x}(k,l) \triangleq [X_1(k,l) \cdots X_N(k,l)]^T$, and $\mathbf{v}(k,l) \triangleq [V_1(k,l) \cdots V_N(k,l)]^T$. We also define the noise and noisy data PSD matrices as $\mathbf{\Phi}_{vv}(k,l) \triangleq E\{\mathbf{v}(k,l)\mathbf{v}^H(k,l)\}$ and $\mathbf{\Phi}_{yy}(k,l) \triangleq E\{\mathbf{y}(k,l)\mathbf{y}^H(k,l)\}$, where ^{H} is the transpose-conjugate of a matrix. For a given random process $a(t)$, we denote its PSD at frequency k and instant l as $\phi_{aa}(k,l)$.

Manuscript received September 16, 2009; revised October 05, 2009. First published October 30, 2009; current version published June 16, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nakatani Tomohiro.

M. Souden, J. Benesty, and S. Affes are with the Université du Québec, INRS-EMT, Montréal, QC H5A 1K6, Canada (e-mail: souden@emt.inrs.ca).

J. Chen is with the Bell Laboratories, Alcatel-Lucent, Murray Hill, NJ 07974 USA.

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

Digital Object Identifier 10.1109/TASL.2009.2035150

Since noise and speech components are assumed to be uncorrelated, we can calculate the noise-free signals statistics as $\mathbf{\Phi}_{xx}(k,l) \triangleq E\{\mathbf{x}(k,l)\mathbf{x}^H(k,l)\} = \mathbf{\Phi}_{yy}(k,l) - \mathbf{\Phi}_{vv}(k,l)$. In practice, recursive smoothing is used to approximate the mathematical expectation involved in the previous PSD matrices. In other words, at time frame l, the noisy and noise data statistics are updated recursively as[1]

$$\begin{aligned}\mathbf{\Phi}_{yy}(k,l) = {} & [1-\alpha_y(k,l)]\mathbf{\Phi}_{yy}(k,l-1) \\ & + \alpha_y(k,l)\mathbf{y}(k,l)\mathbf{y}^H(k,l)\end{aligned} \tag{3}$$

and

$$\begin{aligned}\mathbf{\Phi}_{vv}(k,l) = {} & [1-\alpha_v(k,l)]\,\mathbf{\Phi}_{vv}(k,l-1) \\ & + \alpha_v(k,l)\mathbf{y}(k,l)\mathbf{y}^H(k,l)\end{aligned} \tag{4}$$

where $0 \leq \alpha_y(k,l) \leq 1$ and $0 \leq \alpha_v(k,l) \leq 1$ are two forgetting factors. The proper choice of these two parameters is very important in order to correctly update the noisy and noise PSD matrices. For instance, $\alpha_v(k,l)$ should be large enough when the speech is absent so that the estimate of the noise PSD matrix can follow the noise statistics, but when the speech is present, this parameter should be sufficiently small to avoid noise PSD matrix overestimation. Clearly, the parameter $\alpha_v(k,l)$ is closely related to the speech presence probability that has to be properly computed. This is the purpose of this work.[2]

III. Speech Presence Probability

The speech presence probability in the single-channel case has been exhaustively studied [7]–[9]. Here, we generalize the study to the multichannel scenario. Following the standard procedure, we distinguish between two hypotheses.

1) $H_1(k,l)$: in which case the speech is present, i.e.,

$$\mathbf{y}(k,l) = \mathbf{x}(k,l) + \mathbf{v}(k,l). \tag{5}$$

2) $H_0(k,l)$: in which case the speech is absent, i.e.,

$$\mathbf{y}(k,l) = \mathbf{v}(k,l). \tag{6}$$

Using the Bayes rule [7]–[9], we can show that the speech presence probability is given by

$$p(k,l) \triangleq P[H_1(k,l)|\mathbf{y}(k,l)] = \frac{\Lambda(k,l)}{1+\Lambda(k,l)} \tag{7}$$

where

$$\Lambda(k,l) = \frac{1-q(k,l)}{q(k,l)} \cdot \frac{p[\mathbf{y}(k,l)|H_1(k,l)]}{p[\mathbf{y}(k,l)|H_0(k,l)]}$$

is the generalized likelihood ratio (GLR) [7], [8] and

$$q(k,l) \triangleq P[H_0(k,l)]$$

is the *a priori* probability of speech absence. In practice, $q(k,l)$ has to be chosen such that it reflects our prior knowledge of whether speech sections are more probable than silence ones or not. In other words, we favor speech presence or absence by choosing $q(k,l) < 0.5$ or $q(k,l) \geq 0.5$, respectively. In the current work, we suppose that this probability is fixed *a priori* (see [8], for instance), even though it can be estimated online following the algorithm described in [9] for the single-channel case. In what follows, our purpose is to find a simplified expression of $p(k,l)$.

First, we need to express $p[\mathbf{y}(k,l)|H_1(k,l)]$ and $p[\mathbf{y}(k,l)|H_0(k,l)]$ into analytical forms. To this end, we assume that the speech and noise components are multivariate Gaussian and that the real and imaginary parts of all signals are uncorrelated and identically distributed [12]. Consequently, we obtain

$$\begin{aligned}& p[\mathbf{y}(k,l)|H_1(k,l)] \\ & = \frac{1}{\pi^N \det[\mathbf{\Phi}_{xx}(k,l)+\mathbf{\Phi}_{vv}(k,l)]} \\ & \quad \times \exp\{-\mathbf{y}^H(k,l)[\mathbf{\Phi}_{vv}(k,l) \\ & \quad + \mathbf{\Phi}_{xx}(k,l)]^{-1}\mathbf{y}(k,l)\}\end{aligned} \tag{8}$$

and

$$\begin{aligned}& p[\mathbf{y}(k,l)|H_0(k,l)] \\ & = \frac{1}{\pi^N \det[\mathbf{\Phi}_{vv}(k,l)]} \\ & \quad \times \exp\left\{-\mathbf{y}^H(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{y}(k,l)\right\}.\end{aligned} \tag{9}$$

Therefore, the GLR is given by

$$\begin{aligned}& \Lambda(k,l) \\ & = \frac{1-q(k,l)}{q(k,l)} \cdot \frac{\det[\mathbf{\Phi}_{vv}(k,l)]}{\det[\mathbf{\Phi}_{xx}(k,l)+\mathbf{\Phi}_{vv}(k,l)]} \\ & \quad \cdot \exp\left\{\mathbf{y}^H(k,l)\left[\mathbf{\Phi}_{vv}^{-1}(k,l)\right.\right. \\ & \quad \left.\left. -[\mathbf{\Phi}_{vv}(k,l)+\mathbf{\Phi}_{xx}(k,l)]^{-1}\right]\mathbf{y}(k,l)\right\}.\end{aligned} \tag{10}$$

Using the matrix inversion lemma, we can show the following

$$\begin{aligned}& \mathbf{\Phi}_{vv}^{-1}(k,l) - [\mathbf{\Phi}_{vv}(k,l)+\mathbf{\Phi}_{xx}(k,l)]^{-1} \\ & \qquad = \frac{\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)}{1+\operatorname{tr}\left[\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\right]},\end{aligned} \tag{11}$$

where $\operatorname{tr}[\cdot]$ denotes the trace of a square matrix. In addition, we have

$$\begin{aligned}& \det[\mathbf{\Phi}_{xx}(k,l)+\mathbf{\Phi}_{vv}(k,l)] \\ & \quad = \det\left\{\mathbf{\Phi}_{vv}(k,l)\left[\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)+\mathbf{I}\right]\right\} \\ & \quad = \det[\mathbf{\Phi}_{vv}(k,l)]\left\{1+\operatorname{tr}\left[\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\right]\right\}.\end{aligned} \tag{12}$$

Hence, the expression of the GLR in (10) can be written as

$$\begin{aligned}& \Lambda(k,l) \\ & = \frac{1-q(k,l)}{q(k,l)} \cdot \frac{1}{1+\operatorname{tr}\left[\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\right]} \\ & \quad \cdot \exp\left\{\frac{\mathbf{y}^H(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{y}(k,l)}{1+\operatorname{tr}\left[\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\right]}\right\}.\end{aligned} \tag{13}$$

In a similar fashion to the single-channel approach, let

$$\xi(k,l) \triangleq \operatorname{tr}\left[\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\right] \tag{14}$$

denote the multichannel *a priori* signal-to-noise ratio (SNR). Let us also define

$$\beta(k,l) \triangleq \mathbf{y}^H(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{\Phi}_{xx}(k,l)\mathbf{\Phi}_{vv}^{-1}(k,l)\mathbf{y}(k,l). \tag{15}$$

Now, the speech presence probability is given by

$$p(k,l) = \left\{1+\frac{1}{\Lambda(k,l)}\right\}^{-1} \tag{16}$$

[1] We do not distinguish between the estimate and the exact expression of the PSD matrices for notational convenience.

[2] The utilization of the speech presence probability for online estimation of the noise PSD matrix is currently under investigation.

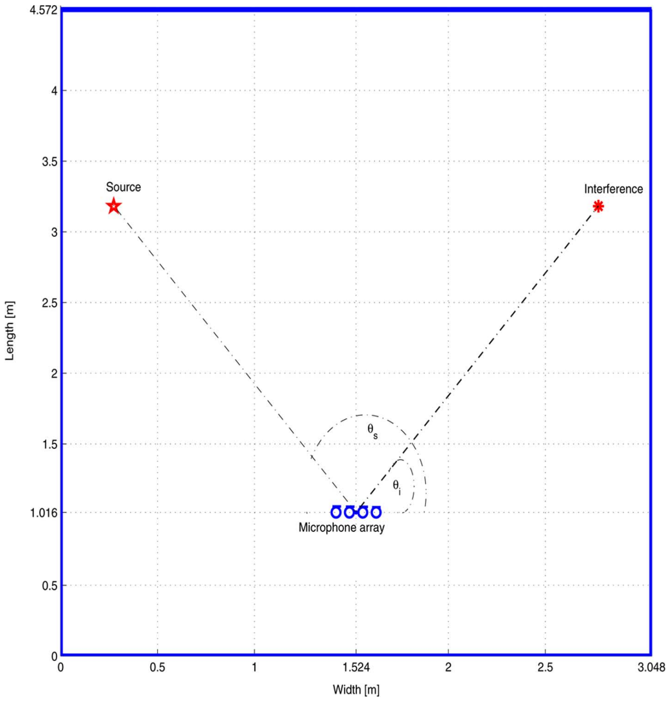


Fig. 1. Scheme of the reverberant enclosure and the locations of the source, the interference, and the microphone array.

meaning that

$$p(k,l) = \left\{1 + \frac{q(k,l)}{1-q(k,l)}[1+\xi(k,l)]\ \exp\left[-\frac{\beta(k,l)}{1+\xi(k,l)}\right]\right\}^{-1}. \tag{17}$$

Equation (17) represents the new proposed expression for the speech presence probability in the multichannel case. This expression is a generalization of the single-channel speech presence probability when the Gaussian statistical model is used as it will be shown next. It describes how the information contained in multiple observations has to be properly combined. Moreover, it only involves the estimates of the noise and noisy data PSD matrices in addition to the current (at time instant l) data samples vector. In contrast to [10], the proposed expression is valid for a general type of noise (e.g., mixture of interferers and spatially white noise) and does not depend on the array geometry. Below, we prove that great accuracy of speech signal detection can be achieved using multiple microphones, particularly when the additive noise is spatially coherent (point source of interference) or white, which is the case in several practical scenarios [1], [4], [11].

A. Single-Channel Case

If only one microphone is used (say microphone 1, for instance), (17) can be written in a degenerated form as

$$\begin{aligned} p_1(k,l) &\triangleq P\left[H_1(k,l)|Y_1(k,l)\right] \\ &= \left\{1 + \frac{q(k,l)}{1-q(k,l)}[1+\xi_1(k,l)]\right. \\ &\qquad \left.\times \exp\left[-\frac{\gamma_1(k,l)\xi_1(k,l)}{1+\xi_1(k,l)}\right]\right\}^{-1} \end{aligned} \tag{18}$$

where

$$\gamma_1(k,l) \triangleq \frac{|Y_1(k,l)|^2}{\phi_{v_1v_1}(k,l)} \tag{19}$$

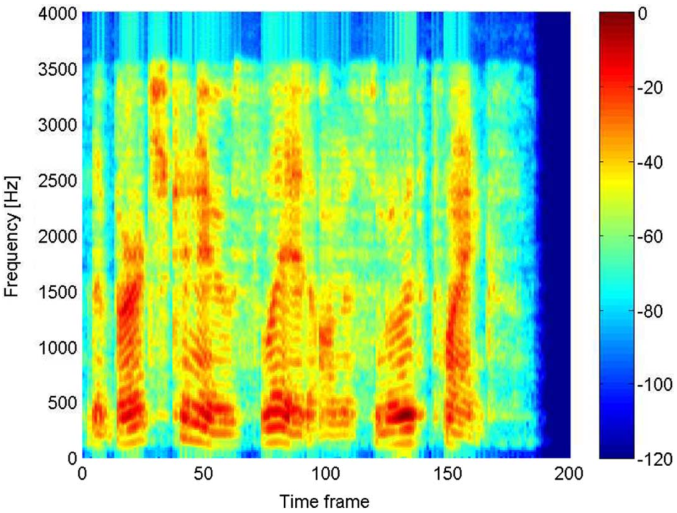


Fig. 2. Distribution of the clean speech energy on the time-frequency plane.

and

$$\xi_1(k,l) \triangleq \frac{\phi_{x_1x_1}(k,l)}{\phi_{v_1v_1}(k,l)}. \tag{20}$$

The result in (18)–(20) is identical to the single-channel speech presence probability given in [7]–[9]. In other words, the single-channel speech presence probability is a particular case of (17). Nevertheless, the great advantage of (17) is that, when multiple microphones are used, it can capture the mutual information among all the sensors in an optimal fashion.

B. Additive Coherent Plus Incoherent Noise Effects

When the noise is generated by a point source (interference), its PSD matrix is given by $\mathbf{\Phi}_{vv}(k,l) = \phi_{ii}(k,l)\mathbf{d}(k)\mathbf{d}^H(k)$, where $\phi_{ii}(k,l)$ and $\mathbf{d}(k)$ are the PSD and the propagation vector of the interference, respectively. Note that in this case $\mathbf{\Phi}_{vv}(k,l)$ is not invertible. However, in practice an additive non-coherent noise is generally present as well. Thus, we have instead

$$\mathbf{\Phi}_{vv}(k,l) = \delta\mathbf{I}_{N\times N} + \phi_{ii}(k,l)\mathbf{d}(k)\mathbf{d}^H(k) \tag{21}$$

where δ is the power of a spatially white (incoherent) noise with independent and identically distributed (i.i.d.) components. By applying the following matrix inversion

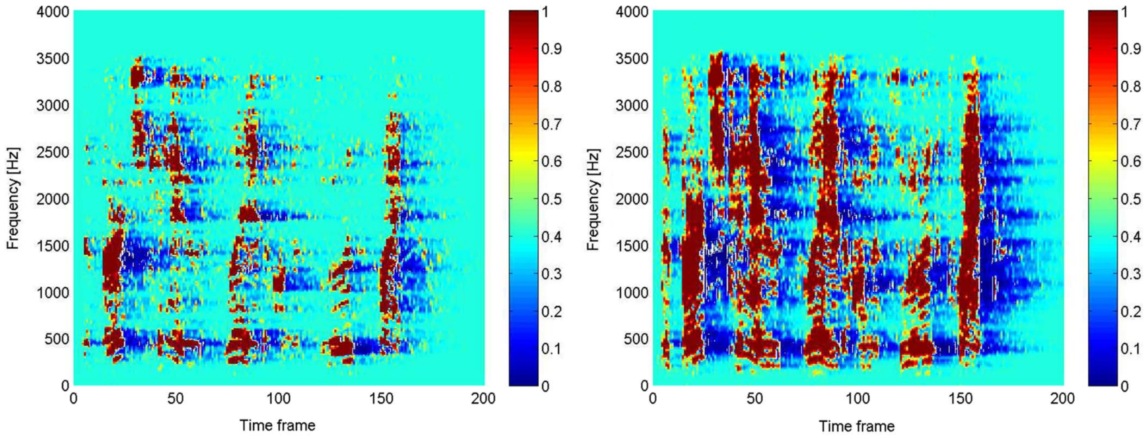
$$\mathbf{\Phi}_{vv}^{-1}(k,l) = \frac{1}{\delta}\left[\mathbf{I}_{N\times N} - \frac{\phi_{ii}(k,l)\mathbf{d}(k)\mathbf{d}^H(k)}{\delta + \phi_{ii}(k,l)\|\mathbf{d}(k)\|^2}\right] \tag{22}$$

we show that the *a priori* SNR is given by

$$\xi(k,l) = \frac{\phi_{ss}(k,l)}{\delta}\left[\|\mathbf{g}(k)\|^2 - \frac{|\mathbf{d}^H(k)\mathbf{g}(k)|^2}{\frac{\delta}{\phi_{ii}(k,l)} + \|\mathbf{d}(k)\|^2}\right]. \tag{23}$$

Also,

$$\beta(k,l) = \frac{\phi_{ss}(k,l)}{\delta^2}\left|\mathbf{y}^H(k,l)\mathbf{g}(k) - \frac{\mathbf{y}^H(k,l)\mathbf{d}(k)\mathbf{d}^H(k)\mathbf{g}(k)}{\frac{\delta}{\phi_{ii}(k,l)} + \|\mathbf{d}(k)\|^2}\right|^2. \tag{24}$$



(a) (b)

Fig. 3. Distribution of the speech presence probability on the time–frequency plane, SIR $= 0$ dB and SNR $= 10$ dB: (a) single channel; (b) multichannel $(N = 4)$.

Therefore,

$$\frac{\beta(k,l)}{1+\xi(k,l)} = \frac{1}{\delta}$$
$$\cdot \frac{\phi_{ss}(k,l)\left|\mathbf{y}^H(k,l)\mathbf{g}(k) - \frac{\mathbf{y}^H(k,l)\mathbf{d}(k)\mathbf{d}^H(k)\mathbf{g}(k)}{\frac{\delta}{\phi_{ii}(k,l)}+\|\mathbf{d}(k)\|^2}\right|^2}{\delta + \phi_{ss}(k,l)\left[\|\mathbf{g}(k)\|^2 - \frac{|\mathbf{d}^H(k)\mathbf{g}(k)|^2}{\frac{\delta}{\phi_{ii}(k,l)}+\|\mathbf{d}(k)\|^2}\right]}. \quad (25)$$

1) Effect of the Coherent Noise: When the speech signal is present, i.e., $\phi_{ss}(k,l)\|\mathbf{g}(k)\|^2 \neq 0$, and the speech and interference originate from different locations, i.e., $\mathbf{d}(k) \neq \mathbf{g}(k)$, we conclude from (23) and (25) that when $\delta \to 0$, we have

$$\xi(k,l) \propto \frac{1}{\delta}$$

and

$$\frac{\beta(k,l)}{1+\xi(k,l)} \propto \frac{1}{\delta}$$

meaning that

$$\lim_{\delta \to 0} p(k,l) = 1 \quad (26)$$

regardless of the level of interference. Hence, perfect speech detection is theoretically possible regardless of the level of the spatially coherent noise. This demonstrates the effectiveness of the multichannel speech presence probability expression developed here in dealing with a point interference source, which cannot be achieved using the traditional single-channel approach.

2) Effect of the Incoherent Noise: Here we assume that $\phi i_{ii}(\omega) = 0$. In this case

$$\xi(k,l) = [1 + R(k)]\xi_1(k,l), \quad (27)$$
$$\beta(k,l) = \xi_1(k,l)\frac{|\tilde{\mathbf{g}}^H(k)\mathbf{y}(k,l)|^2}{\delta} \quad (28)$$

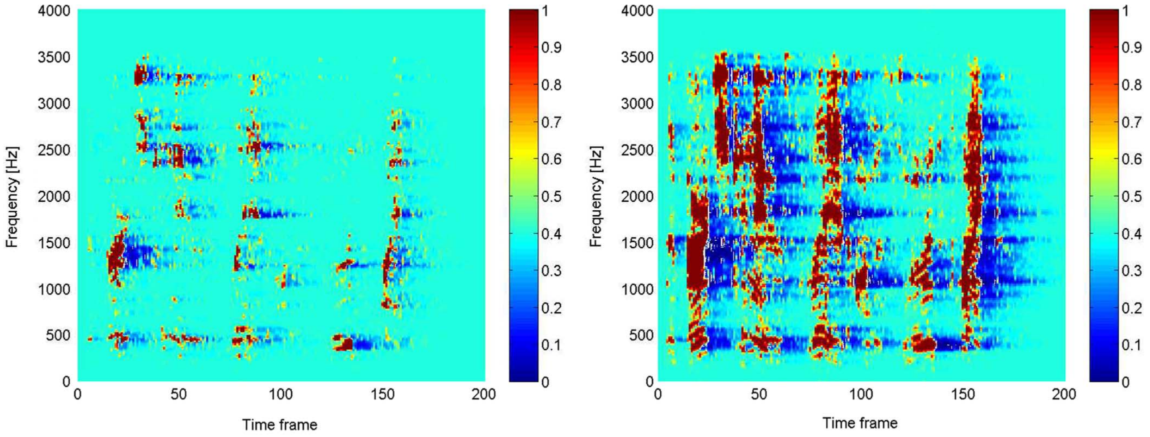
where $R(k) \triangleq \sum_{n=2}^{N}(|G_n(k)|^2/|G_1(k)|^2), \tilde{\mathbf{g}}(k) = (\mathbf{g}(k)/G_1(k))$, and $\xi_1(k,l)$ is defined in (20). A sort of matched beamforming is performed in both terms (27) and (28). The *a priori* SNR is increased by the diversity factor $R(k)$, while in (28), $(|\tilde{\mathbf{g}}^H(k)\mathbf{y}(k,l)|^2/\delta)$ involves a coherent summation of the desired signal part that necessarily leads to the enhancement of the effect of the signal components and the incoherent summation of the noise terms. These two facts result in better speech signal detection, especially its low energy components as compared to the single-channel case.

IV. Numerical Examples

We consider a simulation setup where a target speech signal taken from the IEEE sentences [13] (as described in [6, Ch. 11]) and sampled at 8-kHz rate is located in a reverberant enclosure (modeling a vehicle interior, teleconferencing room, office, etc.) with dimensions 3.048 m × 4.572 m × 3.81 m. The image method [14], [15] was used to generate the impulse responses. Without loss of generality, we consider, for illustration purpose only, a planar configuration where the target source, an interference (a tank noise taken from NOISEX-92 database [16]), and a set of microphones are located on a single plane as depicted in Fig. 1. Several other combinations of interference signals from NOISEX-92 database [16] and speech signals from the IEEE sentences [6], [13] were also tested and results similar to the ones shown here were obtained. In our setup, we consider a uniform linear array (ULA) of $N = 4$ microphones with $r = 0.069$ m being the inter-microphone spacing. The target source and the interferer have azimuthal angles of $\theta_{\rm s} = 120^\circ$ and $\theta_{\rm i} = 60^\circ$, which are measured counterclockwise from the array axis. The microphone array elements are placed on the axis ($y_0 = 1.016$ m, $z_0 = 1.016$ m) with the center of the microphone being at ($x_0 = 1.524$ m, y_0, z_0) and the nth one at $(x_0 - ((N-2n+1)/2)r, y_0, z_0)$ with $n = 1,\ldots,N$. The interferer and the target source are located at a distance of 2.50 m away from the center of the microphone array. It is important to note that the provided details about the system configuration (array geometry and locations of microphones, source, and interference) are not utilized as prior information in our simulations since the proposed speech presence probability expression in (17) depends on the desired speech and noise statistics only. The walls, ceiling, and floor reflection coefficients are set to achieve a reverberation time[3] $T_{60} = 130$ ms measured using the backward integration method (see [17, Ch. 2] for more details). The interfering source sound and computer generated, spatially uncorrelated Gaussian noise are added to the noise-free microphone signals such that the signal-to-interference ratio is SIR $= -10$ and 0 dB, while the signal-to-noise (white) ratio is SNR $=$10 dB and 0 dB in the scenarios investigated below.

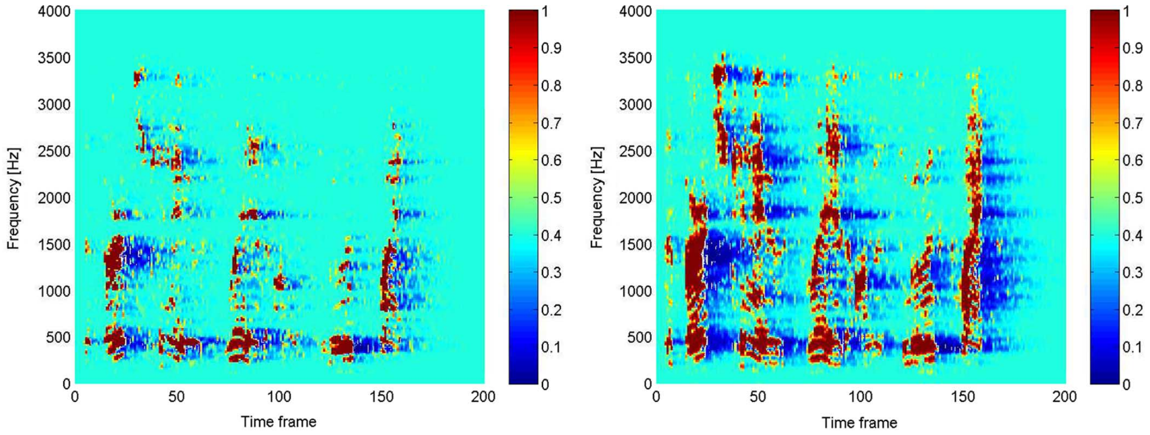
In order to show the advantage of using the proposed multichannel speech presence probability, we compare its performance to its single-channel counterpart. The noisy data and noise PSD matrices are estimated recursively using (3) and (4) and we assumed the knowledge of the noise samples. The update factors for both PSD matrices are set to $\alpha_v = \alpha_y = 0.85$. The *a priori* speech absence probability

[3] Other reverberation conditions were also tested and the results are similar to the investigated setup.



(a) (b)

Fig. 4. Distribution of the speech presence probability on the time–frequency plane, SIR $= -10$ dB and SNR $= 10$ dB: (a) single channel; (b) multichannel $(N = 4)$.



(a) (b)

Fig. 5. Distribution of the speech presence probability on the time–frequency plane, SIR $= 0$ dB and SNR $= 0$ dB: (a) single channel; (b) multichannel $(N = 4)$.

in (17) is empirically fixed as $q(k,l) = 0.6$ (note that one can also estimate it recursively as in [9]). The received signals are chopped into frames of 32 ms with 50% overlapping, sufficiently zero-padded and transformed to the frequency domain to calculate all the required terms. Fig. 2 provides the average PSD of the noise-free signals, i.e., $\mathrm{tr}[\mathbf{\Phi}_{xx}(k,l)]$, in the time-frequency plane. Fig. 3 compares the performance of the single- and multiple-channel speech presence probability in the time–frequency plane in the first simulation scenario where SNR $= 10$ dB and SIR $= 0$ dB. It is clear from Fig. 3(b) that the multichannel approach gives more accurate speech presence detection (high probability values whenever some speech energy exists and low probability values in the absence of speech energy). In contrast, the single-channel case is sensitive to the decay of the speech energy and fails to provide high speech presence probability especially for relatively low speech-energy components. Precisely, we can notice from Fig. 3(b) that the utilization of multiple microphones allows for better detection of the most significant part of the speech components, even the very weak ones having normalized energy of around -40 dB. The single-channel approach fails to detect these low energy components as can be seen in Fig. 3(a). Fig. 4 illustrates the performance of the two speech presence probability expressions (single- and multi-channel) when the SIR is chosen as -10 dB while the SNR is maintained at the same level. A remarkable degradation of the single-channel based processing is observed. Most of the time, speech is not detected even though it is present (see energy components of less than -20 dB). In practice, this would translate into total suppression of these components when a filter is deployed since it is not possible to detect the speech and distinguish it from the noise. The performance of the multichannel-based approach is also slightly deteriorated when we compare Figs. 3(b) to 4(b). However, most of the speech components were properly detected. In the last scenario, we set the SNR to 0 dB and the SIR to 0 dB. The results are given in Fig. 5. Again, by comparing the results achieved by the single and multichannel processing, we notice the clear advantage of the latter. It is also remarkable in Fig. 5(b) that the low-frequency components are better detected than the high-frequency ones [compared to Figs. 3 and 4(b)] which is justified by the speech energy distribution shown in Fig. 2. The overall results clearly demonstrate the advantage of using the multichannel speech presence probability developed in this paper.

V. Conclusion

In this paper, a multichannel speech presence probability was developed. We assumed a Gaussian statistical model for the signals and elaborated a new simplified closed-form expression for this probability when an array of an arbitrary number of microphones with arbitrary placements and reverberation condition are considered. The utilization of the multichannel speech presence probability is advantageous as illustrated by theory and numerical evaluations. From the theoretical point of view, we showed that by using the new formulation, a perfect detection of the speech components is possible if the noise originates from a point interference source, which can never be achieved with the single-channel case. In the case of incoherent noise, a coherent summation of the noise-free speech components is performed to allow for better speech detection, especially of low speech energy components as compared to the single-channel approaches. The proposed method applies for the general situation where the observed microphone signals are mixtures of a desired speech plus noise signals. The latter can be composed of interferences and other types of undesired signals (e.g., white noise).

References

[1] J. Benesty, J. Chen, and Y. Huang*, Microphone Array Signal Processing*. Berlin, Germany: Springer-Verlag, 2008.

[2] J. Dmochowski, J. Benesty, and S. Affes, "Direction of arrival estimation using the parameterized spatial correlation matrix," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1327–1339, May 2007.

[3] M. Souden, J. Benesty, and S. Affes, "New insights into non-causal multichannel linear filtering for noise reduction," in *Proc. IEEE ICASSP*, 2009, pp. 141–144.

[4] S. Gannot and I. Cohen, "Adaptive beamforming and postfitering," in *Springer Handbook of Speech Processing*, J. Benesty, M. M. Sondhi, and Y. Huang, Eds. New York: Springer-Verlag, 2007, ch. 47, pp. 945–978.

[5] J. Benesty, J. Chen, Y. Huang, and J. Dmochowski, "On microphone-array beamforming from a MIMO acoustic signal processing perspective," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 1053–1065, Mar. 2007.

[6] P. C. Loizou*, Speech Enhancement: Theory and Practice*. New York: CRC Press, 2007.

[7] D. Middleton and R. Esposito, "Simultaneous optimum detection and estimation of signals in noise," *IEEE Trans. Inf. Theory*, vol. IT-14, no. 3, pp. 434–444, May 1968.

[8] 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.

[9] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *IEEE Trans. Speech, Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[10] I. Potamitis, "Estimation of speech presence probability in the field of microphone array," *IEEE Signal Process. Lett.*, vol. 11, no. 12, pp. 956–959, Dec. 2004.

[11] G. Reuven, S. Gannot, and I. Cohen, "Dual source transfer-function generalized sidelobe canceller," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 4, pp. 711–726, May 2008.

[12] A. V. D. Bos, "The multivariate complex normal distribution—A generalization," *IEEE Trans. Inf. Theory*, vol. 41, no. 2, pp. 537–539, Mar. 1995.

[13] IEEE Subcommittee, "IEEE recommended practice for speech quality measurements," *IEEE Trans. Audio Electroacoust.*, vol. AE-17, no. 3, pp. 225–246, Sep. 1969.

[14] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, pp. 943–950, Apr. 1979.

[15] P. Peterson, "Simulating the response of multiple microphones to a single acoustic source in a reverberant room," *J. Acoust. Soc. Amer.*, vol. 80, pp. 1527–1529, Nov. 1986.

[16] A. P. Varga, H. J. M. Steenekan, M. Tomlinson, and D. Jones, "The Noisex-92 study on the effect of additive noise on automatic speech recognition," *Tech. Rep., DRA Speech Research Unit*, 1992.

[17] Y. Huang, J. Benesty, and J. Chen*, Acoustic MIMO Signal Processing.*. Berlin, Germany: Springer-Verlag, 2006.

Statistical Text-to-Speech Synthesis Based on Segment-Wise Representation With a Norm Constraint

Stas Tiomkin, David Malah*, Life Fellow, IEEE*, and Slava Shechtman

***Abstract*—In statistical HMM-based text-to-speech systems (STTS), speech feature dynamics is modeled by first- and second-order feature frame differences, which, typically, do not satisfactorily represent frame to frame feature dynamics present in natural speech. The reduced dynamics results in over-smoothing of speech features, often sounding as muffled synthesized speech. In this correspondence, we propose a method to enhance a baseline STTS system by introducing a segment-wise model representation with a norm constraint. The segment-wise representation provides additional degrees of freedom in speech feature determination. We exploit these degrees of freedom for increasing the speech feature vector norm to match a norm constraint. As a result, statistically generated speech features are less over-smoothed, resulting in more natural sounding speech, as judged by listening tests.**

***Index Terms*—Segment-wise model representation, speech feature dynamics, statistical TTS, text-to-speech (TTS) synthesis.**

I. Introduction

Statistical TTS (STTS) systems employ statistical models for speech production, and speech is generated from previously learned statistical models. Contrary to concatenative TTS (CTTS), which may include discontinuities, particulary when small databases are used, STTS smoothly connects adjacent phonetic units.

However, STTS-generated speech is often over-smoothed, resulting in degraded speech quality in the form of muffled speech. A thorough review of STTS systems is provided in [1].

In this correspondence, we improve a baseline HMM-based STTS system by introducing 1) A robust model representation, based on a segment-wise representation, instead of the conventional frame-wise representation; and 2) A norm-regulated statistical speech feature vector that meets a norm constraint. These concepts are utilized in an iterative algorithm, proposed in this correspondence. This algorithm generates speech features with enhanced dynamics, resulting in improved generated speech naturalness, as compared to the conventional generating scheme, and verified by listening tests.

This correspondence is organized as follows. In Section II, we provide the essentials of the baseline STTS methodology used in this research. In Section III, we present the segment-wise model representation. In Section IV, we present the norm-regulated constraint, applied to the synthesized speech feature vector, and an iterative algorithm that generates speech features having enhanced dynamics. In Section V, we examine the performance of the enhanced statistical TTS system, and in Section VI we summarize this work.

Manuscript received August 25, 2009; revised December 14, 2009. First published January 19, 2010; current version published June 16, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gaël Richard.

S. Tiomkin and D. Malah are with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel (e-mail: stast@tx.technion.ac.il; malah@ee.technion.ac.il).

S. Shechtman is with the Speech Technologies Group, IBM Haifa Research Lab, Haifa University Campus, Mount Carmel, Haifa 31905, Israel (e-mail: slava@il.ibm.com).

Digital Object Identifier 10.1109/TASL.2010.2040795