Hindawi Publishing Corporation
EURASIP Journal on Applied Signal Processing
Volume 2006, Article ID 20683, Pages 1–15
DOI 10.1155/ASP/2006/20683

Sector-Based Detection for Hands-Free Speech Enhancement in Cars

Guillaume Lathoud,[1,2] Julien Bourgeois,[3] and Jürgen Freudenberger[3]

[1] *IDIAP Research Institute, 1920 Martigny, Switzerland*
[2] *École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland*
[3] *DaimlerChrysler Research and Technology, 89014 Ulm, Germany*

Received 31 January 2005; Revised 20 July 2005; Accepted 22 August 2005

Adaptation control of beamforming interference cancellation techniques is investigated for in-car speech acquisition. Two efficient adaptation control methods are proposed that avoid target cancellation. The "implicit" method varies the step-size continuously, based on the filtered output signal. The "explicit" method decides in a binary manner whether to adapt or not, based on a novel estimate of target and interference energies. It estimates the average delay-sum power within a volume of space, for the same cost as the classical delay-sum. Experiments on real in-car data validate both methods, including a case with 100 km/h background road noise.

1. INTRODUCTION

Speech-based command interfaces are becoming more and more common in cars, for example in automatic dialog systems for hands-free phone calls and navigation assistance. The automatic speech recognition performance is crucial, and can be greatly hampered by interferences such as speech from a codriver. Unfortunately, spontaneous multiparty speech contains lots of overlaps between participants [1].

A directional microphone oriented towards the driver provides an immediate *hardware* enhancement by lowering the energy level of the codriver interference. In the Mercedes S320 setup used in this article, a 6 dB relative difference is achieved (value measured in the car). However, an additional *software* improvement is required to fully cancel the codriver's interference, for example, with *adaptive* techniques. They consist in a *time-varying* linear filter that enhances the signal-to-interference ratio (SIR), as depicted by Figure 1.

Many beamforming algorithms have been proposed, with various degrees of relevance in the car environment [2]. Apart from differential array designs, superdirective beamformers [3] derived from the minimum variance distortionless response principle (MVDR) apply well to our hardware setup, such as the generalized sidelobe canceller (GSC) structure. The original adaptive versions assume a fixed, known acoustic propagation channel. This is rarely the case in practice, so the target signal is reduced at the beamformer output. A solution is to adapt, *only when the interferer is dominant*, by varying the adaptation speed in a binary manner (explicit control), or in a continuous manner (implicit control).

Existing explicit methods detect when the target is dominant by thresholding an estimate of the input SIR, $\widehat{\mathrm{SIR}}_{\mathrm{in}}(t)$, or a related quantity. During those periods, adaptation is stopped [4] or the acoustic channel is tracked [5, 6] (and related self-calibration algorithms [7]). Typically, $\widehat{\mathrm{SIR}}_{\mathrm{in}}(t)$ can be the ratio of the delay-and-sum beamformer and the blocking matrix output powers [7–9]. If the blocking matrix is adapted, as in [8], speaker detection errors are fed back into the adapted parts and a single detection error may have dramatical effects. Especially for simultaneous speakers, it is more robust to decouple detection from adaptation [9, 10]. Most existing explicit methods rely on prior knowledge of the target location only. There are few implicit methods, such as [11], which varies the adaptation speed based on the input signal itself.

The contribution of this paper is twofold. First, an explicit method (Figure 2(a)) is proposed. It relies on a novel input SIR estimate, which extends a previously proposed sector-based frequency-domain detection and localization technique [12]. Similarly to some multispeaker segmentation works [13, 14], it uses phase information only. It introduces the concept of phase domain metric (PDM). It is closely related to delay-sum beamforming, *averaged* over a sector of space, for no additional cost. Few works investigated input

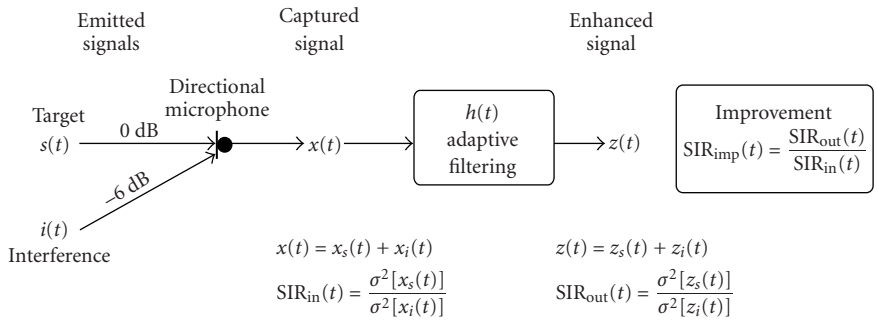
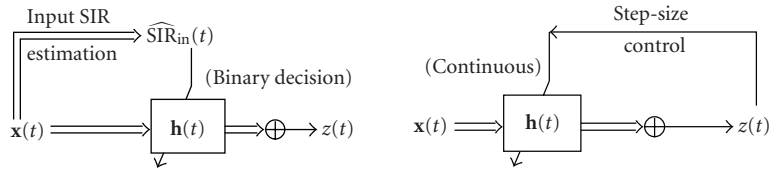


FIGURE 1: Entire acquisition process from emitted signals to the enhanced signal. This paper focuses on the adaptive filtering block $h(t)$, so that $\mathrm{SIR}_{\mathrm{imp}}(t)$ is maximized when the interference is active (interference cancellation). The s and t subscripts designate contributions of target and interference, respectively. The whole process is supposed to be linear. $\sigma^2[x(t)]$ is the variance or energy of a speech signal $x(t)$, estimated on a short-time frame (20 or 30 millisecond) around t, on which stationarity and ergodicity are assumed.



(a) Proposed explicit approach.

(b) Proposed implicit approach.

FIGURE 2: Proposed explicit and implicit adaptation control. $\mathbf{x}(t) = [x_1(t) \cdots x_M(t)]^{\mathrm{T}}$ are the signals captured by the M microphones, and $\mathbf{h}(t) = [\mathbf{h}_1(t) \cdots \mathbf{h}_M(t)]^{\mathrm{T}}$ are their associated filters. Double arrows denote multiple signals.

SIR estimation for nonstationary, wideband signals such as speech. In [9, 15], spatial information of the target only is used, represented as a single direction. On the contrary, the proposed approach (1) defines spatial locations in terms of sectors, (2) uses both target's and interference's spatial location information. This is particularly relevant in the car environment, where both locations are known, but only approximately.

The second contribution is an implicit adaptation method, where the speed of adaptation (step-size) is determined from the output signal $z(t)$ (Figure 2(b)), with theoretically-proven robustness to target cancellation issues. Estimation of the input SIR is not needed, and there is no additional computational cost.

Experiments on real in-car data validate both contributions on two setups: either 2 or 4 directional microphones. In both cases, the sector-based method reliably estimates the input SIR ($\widehat{\mathrm{SIR}}_{\mathrm{in}}(t)$). Both implicit and explicit approaches improve the output SIR ($\mathrm{SIR}_{\mathrm{out}}(t)$) in a robust manner, including in 100 km/h background noise. The explicit control yields the best results. Both adaptation methods are fit for real-time processing.

The rest of this paper is organized as follows. Section 2 summarizes, extends, and interprets the recently proposed [12] sector-based activity detection approach. Section 3 describes the two in-car setups and defines the sectors in each case. Section 4 derives a novel sector-based technique for input SIR estimation, based on Section 2, and validates it with experiments. Section 5 describes both implicit and explicit approaches and validates them with speech enhancement experiments. Section 6 concludes. This paper is a detailed version of an abstract presented in [16].

2. SECTOR-BASED FREQUENCY-DOMAIN ACTIVITY DETECTION

This section extends the SAM-SPARSE audio source detection and localization approach, previously proposed and tested on multiparty speech in the meeting room context [12]. The space around a microphone array is divided into volumes called "sectors." The frequency spectrum is also discretized into frequency bins. For each sector and each frequency bin, we determine whether or not there is at least one active audio source in the sector. This is done by comparing measured phases between the various microphone pairs (a vector of angle values) with a "centroid" for each sector (another vector). A central feature of this work is the sparsity assumption: within each frequency bin, at most one speech source is supposed to be active. This simplification is supported by statistical analysis of real two-speaker speech signals [17], which shows that most of the time, within a given frequency bin, one speech source is dominant in terms of energy and the other one is negligible.

Sections 2.1 and 2.2 generalize the SAM-SPARSE approach. An extension is proposed to allow for a "soft" decision within each frequency bin, as opposed to the "hard

decision" taken in [12]. Note that each time frame is processed fully independently, without any temporal integration over consecutive frames. Section 2.3 gives a low-cost implementation. Physical and topological interpretations are found in Section 2.4 and Appendix A, respectively.

2.1. A Phase domain metric

First, a few notations are defined. All frequency domain quantities are estimated through the discrete Fourier transform (DFT) on short finite windows of samples (20 to 30 millisecond), on which speech signals can be approximated as stationary.

M is the number of microphones. One time frame of N_{samples} multichannel samples is denoted by $\mathbf{x}_1, \ldots, \mathbf{x}_m, \ldots, \mathbf{x}_M$, with $\mathbf{x}_m \in \mathbb{R}^{N_{\text{samples}}}$. The corresponding positive frequency Fourier coefficients obtained through DFT are denoted by $\mathbf{X}_1, \ldots, \mathbf{X}_m, \ldots, \mathbf{X}_M$, with $\mathbf{X}_m \in \mathbb{C}^{N_{\text{bins}}}$.

$f \in \mathbb{N}$ is a discrete frequency ($1 \le f \le N_{\text{bins}}$), $\mathcal{R}e(\cdot)$ denotes the real part of a complex quantity, and $\widehat{G}^{(p)}(f)$ is the estimated frequency-domain cross-correlation for microphone pair p ($1 \le p \le P$):

$$\widehat{G}^{(p)}(f) \stackrel{\text{def}}{=} X_{i_p}(f) \cdot X^*_{j_p}(f), \tag{1}$$

where $(\cdot)^*$ denotes complex conjugate and i_p and j_p are indices of the 2 microphones: $1 \le i_p < j_p \le M$. Note that the total number of microphone pairs is $P = M(M-1)/2$.

In all this work, the sector-based detection (and in particular, estimation of the cross-correlation $\widehat{G}^{(p)}(f)$) does *not* use any time averaging between consecutive frames: each frame is treated fully independently. This is consistent with the work that we are building on [12], and avoids smoothing parameters that would need to be tuned (e.g., forgetting factor). Experiments in Section 4.2 show that this is sufficient to obtain a decent SIR estimate.

Phase values *measured* at frequency f are denoted:

$$\begin{aligned} &\widehat{\boldsymbol{\Theta}}(f) \stackrel{\text{def}}{=} [\widehat{\theta}^{(1)}(f), \ldots, \widehat{\theta}^{(p)}(f), \ldots, \widehat{\theta}^{(P)}(f)]^{\mathrm{T}} \\ &\text{where } \widehat{\theta}^{(p)}(f) \stackrel{\text{def}}{=} \angle \widehat{G}^{(p)}(f), \end{aligned} \tag{2}$$

where $\angle(\cdot)$ designates the argument of a complex value. The distance between two such vectors, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ in $\mathbb{R}^P$, is defined as

$$d(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \stackrel{\text{def}}{=} \sqrt{\frac{1}{P} \sum_{p=1}^{P} \sin^2\left(\frac{\theta_1^{(p)} - \theta_2^{(p)}}{2}\right)}, \tag{3}$$

$d(\cdot, \cdot)$ is similar to the Euclidean metric, except for the sine, which accounts for the "modulo 2π" definition of angles. The $1/P$ normalization factor ensures that $0 \le d(\cdot, \cdot) \le 1$. Two reasons motivate the use of sine, as opposed to a piecewise linear function such as $\arg\min_k |\theta_1^{(p)} - \theta_2^{(p)} + k2\pi|$:

(i) the first reason is that $d(\cdot, \cdot)$ is closely related to delay-sum beamforming, as shown by Section 2.4;

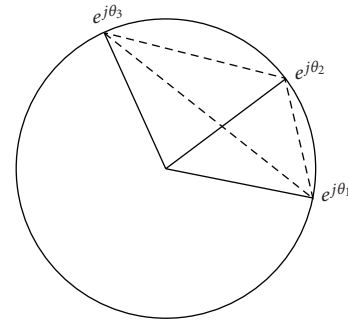


Figure 3: Illustration of the triangular inequality for the PDM in dimension 1: each point on the unit circle corresponds to an angle value modulo 2π. From the Euclidean metric $|e^{j\theta_3} - e^{j\theta_1}| \le |e^{j\theta_3} - e^{j\theta_2}| + |e^{j\theta_2} - e^{j\theta_1}|$.

(ii) the second reason is that $d^2(\cdot, \cdot)$ is infinitely derivable in all points, and its derivates are simple to express. This is not the case of "arg min." It is related to parameter optimization work not presented here.

Topological interpretation

$d(\cdot, \cdot)$ is a true PDM, as defined in Appendix A.1. This is straightforward for $P = 1$ by representing any angle θ with a point $e^{j\theta}$ on the unit circle, as in Figure 3, and observing that $|e^{j\theta_1} - e^{j\theta_2}| = 2|\sin((\theta_1 - \theta_2)/2)| = 2d(\theta_1, \theta_2)$. Appendix A.2 proves it for higher dimensions $P > 1$.

2.2. From metric to activity: SAM-SPARSE-MEAN

The search space around the microphone array is partitioned into N_S connected volumes called "sectors," as in [12, 18]. For example, the space around a horizontal circular microphone array can be partitioned in "pie slices." The SAM-SPARSE-MEAN approach treats each frequency bin separately. Thus, a parallel implementation is straightforward.

For each (sector, frequency bin), it defines and estimates a sector activity measure (SAM), which is a posterior probability that at least one audio source is active within that sector *and* that frequency bin. "SPARSE" stands for the sparsity assumption that was discussed above: at most one sector is active per frequency bin. It was shown in [12] to be both necessary and efficient to solve spatial leakage problems.

Note that only phase information is used, but not the magnitude information. This choice is inspired by (1) the GCC-PHAT weighting [19], which is well adapted to reverberant environments, and (2) the fact that interaural level difference (ILD) is in practice much less reliable than time-delays, as far as localization is concerned. In fact, ILD is mostly useful in the case of binaural analysis [20].

SAM-SPARSE-MEAN is composed of two steps.

(i) The first step is to compute the root mean-square distance ("MEAN") between the measured phase vector $\widehat{\boldsymbol{\Theta}}(f)$ and theoretical phase vectors associated with *all* points within a given sector S_k, at a given frequency f,

using the metric defined in (3):

$$\overline{D}_{k,f} \stackrel{\text{def}}{=} \left[\int_{\mathbf{v}\in\mathbf{S}_k} d^2\left(\widehat{\boldsymbol{\Theta}}(f), \boldsymbol{\Gamma}(\mathbf{v}, f)\right) P_k(\mathbf{v}) d\mathbf{v} \right]^{1/2}, \quad (4)$$

where

$$\boldsymbol{\Gamma}(\mathbf{v}, f) = [\gamma^{(1)}(\mathbf{v}, f), \ldots, \gamma^{(p)}(\mathbf{v}, f), \ldots, \gamma^{(P)}(\mathbf{v}, f)]^{\mathrm{T}} \quad (5)$$

is the vector of theoretical phases associated with location $\mathbf{v}$ and frequency f and $P_k(\mathbf{v})$ is a weighting term. $P_k(\mathbf{v})$ is the prior knowledge of the distribution of active source locations within sector S_k (e.g., uniform or Gaussian distribution). $\mathbf{v}$ can be expressed in any coordinate system (Euclidean or spherical) as long as the expression of $d\mathbf{v}$ is consistent with this choice. Each component of the $\boldsymbol{\Gamma}$ vector is given by

$$\gamma^{(p)}(\mathbf{v}, f) = \pi \frac{f}{N_{\text{bins}}} \tau^{(p)}(\mathbf{v}), \quad (6)$$

where $\tau^{(p)}(\mathbf{v})$ is the theoretical time-delay (in samples) associated with spatial location $\mathbf{v} \in \mathbb{R}^3$ and microphone pair p. $\tau^{(p)}(\mathbf{v})$ is given by

$$\tau^{(p)}(\mathbf{v}) = \frac{f_s}{c}\left(\left\|\mathbf{v} - \mathbf{m}_2^{(p)}\right\| - \left\|\mathbf{v} - \mathbf{m}_1^{(p)}\right\|\right), \quad (7)$$

where c is the speed of sound in the air (e.g., 342 m/s at 18 degrees Celsius), f_s is the sampling frequency in Hz and $\mathbf{m}_1^{(p)}$ and $\mathbf{m}_2^{(p)} \in \mathbb{R}^3$ are spatial locations of microphone pair p.

(ii) The second step is to determine, for each frequency bin f, the sector to which the measured phase vector is the closest:

$$k_{\min}(f) \stackrel{\text{def}}{=} \arg\min_k \overline{D}_{k,f}. \quad (8)$$

This decision does not require any threshold. Finally, the posterior probability of having at least one active source in sector $S_{k_{\min}(f)}$ and at frequency f is modeled with

$$P\left(\text{sector } S_{k_{\min}(f)} \text{ active at frequency } f \mid \widehat{\boldsymbol{\Theta}}(f)\right) = e^{-\lambda(\overline{D}_{k_{\min}(f),f})^2}, \quad (9)$$

where λ controls how "soft" or "hard" this decision should be. The sparsity assumption implies that all other sectors are attributed a zero posterior probability of containing activity at frequency f:

$$\forall k \neq k_{\min}(f) \quad P\left(\text{sector } S_k \text{ active at frequency } f \mid \widehat{\boldsymbol{\Theta}}(f)\right) = 0. \quad (10)$$

In previous work [12], only "hard" decisions were taken ($\lambda = 0$) and the entire spectrum was supposed to be active, which lead to attribution of inactive frequencies to random sectors. Equation (9) represents a generalization ($\lambda > 0$) that allows to detect *inactivity* at a given frequency and thus avoids the random effect. For example, in the case of a single microphone pair $P = 1$, for $\lambda = 10$, any phase difference between θ_1 and θ_2 larger than about $\pi/3$ gives a probability of activity $e^{-\lambda d^2(\theta_1,\theta_2)}$ less than 0.1. λ can be tuned on some (small) development data, as in Section 4.2. An alternative can be found in [21].

2.3. Practical implementation

In general, it is not possible to derive an analytical solution for (4). It is therefore approximated with a discrete summation:

$$\overline{D}_{k,f} \approx \widehat{\overline{D}}_{k,f}, \quad \text{where } \widehat{\overline{D}}_{k,f} \stackrel{\text{def}}{=} \sqrt{\frac{1}{N}\sum_{n=1}^{N} d^2\left(\widehat{\boldsymbol{\Theta}}(f), \boldsymbol{\Gamma}(\mathbf{v}_{k,n}, f)\right)}, \quad (11)$$

where $\mathbf{v}_{k,1}, \ldots, \mathbf{v}_{k,n}, \ldots, \mathbf{v}_{k,N}$ are locations in space ($\mathbb{R}^3$) drawn from the prior distribution $P_k(\mathbf{v})$ and N is the number of locations used to approximate this continuous distribution. The sampling is not necessarily random, for example, a regular grid for a uniform distribution.

The rest of this section expresses this approximation in a manner that does not depend on the number of points N.

$$\left(\widehat{\overline{D}}_{k,f}\right)^2 = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{P}\sum_{p=1}^{P}\sin^2\left(\frac{\hat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}_{k,n}, f)}{2}\right). \quad (12)$$

Using the relation $\sin^2 u = (1/2)(1 - \cos 2u)$, we can write

$$\begin{aligned}
\left(\widehat{\overline{D}}_{k,f}\right)^2 &= \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - \frac{1}{N}\sum_{n=1}^{N}\cos\left(\hat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}_{k,n}, f)\right)\right\},\\
\left(\widehat{\overline{D}}_{k,f}\right)^2 &= \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - \mathcal{R}e\left[\frac{1}{N}\sum_{n=1}^{N} e^{j\left(\hat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}_{k,n}, f)\right)}\right]\right\},\\
\left(\widehat{\overline{D}}_{k,f}\right)^2 &= \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - \mathcal{R}e\left[e^{j\hat{\theta}^{(p)}(f)}\frac{1}{N}\sum_{n=1}^{N} e^{-j\gamma^{(p)}(\mathbf{v}_{k,n}, f)}\right]\right\},\\
\left(\widehat{\overline{D}}_{k,f}\right)^2 &= \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - \mathcal{R}e\left[e^{j\hat{\theta}^{(p)}(f)} A_k^{(p)}(f) e^{-jB_k^{(p)}(f)}\right]\right\},\\
\left(\widehat{\overline{D}}_{k,f}\right)^2 &= \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - A_k^{(p)}(f)\cos\left(\hat{\theta}^{(p)}(f) - B_k^{(p)}(f)\right)\right\},
\end{aligned} \quad (13)$$

where $A_k^{(p)}(f)$ and $B_k^{(p)}(f)$ are two values in $\mathbb{R}$ that do not depend on the measured phase $\hat{\theta}^{(p)}(f)$:

$$A_k^{(p)}(f) \stackrel{\text{def}}{=} \left|Z_k^{(p)}(f)\right|, \qquad B_k^{(p)}(f) \stackrel{\text{def}}{=} \angle Z_k^{(p)}(f),$$
$$Z_k^{(p)}(f) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} e^{j\gamma^{(p)}(\mathbf{v}_{k,n}, f)}. \quad (14)$$

Hence, the approximation is wholly contained in the A and B parameters, which need to be computed only once. Any large number N can be used, so the approximation $\widehat{\overline{D}}_{k,f}$ can be as close to $\overline{D}_{k,f}$ as desired. During runtime, the cost of computing $\widehat{\overline{D}}_{k,f}$ does not depend on N: it is directly proportional to P, which is the same cost as for a point-based measure $d(\cdot,\cdot)$. Thus, the proposed approach ($\overline{D}_{k,f}$) does not suffer from its practical implementation ($\widehat{\overline{D}}_{k,f}$) concerning both numerical precision and computational complexity. Note that each $Z_k^{(p)}(f)$ value is nothing but a component of the *average* theoretical cross-correlation matrix

over all points $\mathbf{v}_{k,n}$ for $n = 1, \ldots, N$. A complete Matlab implementation can be downloaded at: http://mmm.idiap.ch/lathoud/2005-SAM-SPARSE-MEAN.

The SAM-SPARSE-C method defined in a previous work [12] is strictly equivalent to a modification of $\widehat{D}_{k,f}$, where all $A_k^{(p)}(f)$ parameters would be replaced with 1.

2.4. Physical interpretation

This section shows that for a given triplet (sector, frequency bin, pair of microphones), if we neglect the energy difference between microphones, the PDM proposed by (4) is equivalent to the delay-sum power *averaged* over all points in the sector.

First, let us consider a point location $\mathbf{v} \in \mathbb{R}^3$, a pair of microphones $(\mathbf{m}_1^{(p)}, \mathbf{m}_2^{(p)})$, and a frequency f. In frequency domain, the received signals are:

$$X_{i_p}(f) \stackrel{\text{def}}{=} \alpha_1^{(p)}(f) e^{j\beta_1^{(p)}(f)}, \qquad X_{j_p}(f) \stackrel{\text{def}}{=} \alpha_2^{(p)}(f) e^{j\beta_2^{(p)}(f)}, \tag{15}$$

where for each microphone $m = 1, \ldots, M$, $\alpha_m(f)$ and $\beta_m(f)$ are real-valued, respectively, magnitude and phase of the received signal $X_m(f)$. The observed phase is

$$\widehat{\theta}^{(p)}(f) \equiv \beta_1^{(p)}(f) - \beta_2^{(p)}(f), \tag{16}$$

where the $\equiv$ symbol denotes congruence of angles (equality modulo 2π).

The delay-sum energy for location $\mathbf{v}$, microphone pair p and frequency f, is defined by aligning the two signals, with respect to the theoretical phase $\gamma^{(p)}(\mathbf{v}, f)$:

$$E_{\text{ds}}^{(p)}(\mathbf{v}, f) \stackrel{\text{def}}{=} \left| X_{i_p}(f) + X_{j_p}(f) e^{j\gamma^{(p)}(\mathbf{v},f)} \right|^2. \tag{17}$$

Assuming the received magnitudes to be the same $\alpha_{i_p} \approx \alpha_{j_p} \approx \alpha$, (17) can be rewritten:

$$\begin{aligned} E_{\text{ds}}^{(p)}(\mathbf{v}, f) &= \left| \alpha e^{j\beta_1^{(p)}(f)} \left(1 + e^{j(-\widehat{\theta}^{(p)}(f) + \gamma^{(p)}(\mathbf{v},f))}\right) \right|^2 \\ &= \alpha^2 \left[\left(1 + \cos\left(- \widehat{\theta}^{(p)}(f) + \gamma^{(p)}(\mathbf{v}, f)\right)\right)^2 \right. \\ &\qquad \left. + \sin^2\left(- \widehat{\theta}^{(p)}(f) + \gamma^{(p)}(\mathbf{v}, f)\right)\right] \\ &= \alpha^2 \left[2 + 2\cos\left(- \widehat{\theta}^{(p)}(f) + \gamma^{(p)}(\mathbf{v}, f)\right)\right]. \end{aligned} \tag{18}$$

On the other hand, the square distance between observed phase and theoretical phase, as defined by (3), is expressed as

$$d^2\left(\widehat{\theta}^{(p)}(f), \gamma^{(p)}(\mathbf{v}, f)\right) \stackrel{\text{def}}{=} \sin^2\left(\frac{\widehat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}, f)}{2}\right) \tag{19}$$

$$= \frac{1}{2}\left(1 - \cos\left(\widehat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}, f)\right)\right). \tag{20}$$

From (18) and (20),

$$\frac{1}{4\alpha^2} E_{\text{ds}}^{(p)}(\mathbf{v}, f) = 1 - d^2\left(\widehat{\theta}^{(p)}(f), \gamma^{(p)}(\mathbf{v}, f)\right). \tag{21}$$

Thus, for a given microphone pair, (1) maximizing the delay-sum power is strictly equivalent to minimizing the PDM, (2) comparing delay-sum powers is strictly equivalent to comparing PDMs. This equivalence still holds when averaging over an entire sector, as in (4). Averaging across microphone pairs, as in (3), exploits the redundancy of the signals in order to deal with noisy measurements and get around spatial aliasing effects.

The proposed approach is thus equivalent to an average delay-sum over a sector, which differs from a classical approach that would compute the delay-sum only at a point in the middle of the sector. For sector-based detection, the former is intuitively more sound because it incorporates the prior knowledge that the audio source may be *anywhere* within a sector. On the contrary, the classical point-based approach tries to address a sector-based task without this knowledge; thus, errors can be expected when an audio source is located far from any of the middle points. The advantage of the sector-based approach was confirmed by tests on more than one hour of real meeting room data [12]. The computational cost is the same, as shown by Section 2.3.

The assumption $\alpha_{i_p} \approx \alpha_{j_p}$ is reasonable for most setups, where microphones are close to each other and, if directional, oriented to the same direction. Nevertheless, in practice, the proposed method can also be applied to other cases, as in Setup I, described in Section 3.1.

3. PHYSICAL SETUPS, RECORDINGS, AND SECTOR DEFINITION

The rest of this paper considers two setups for acquisition of the driver's speech in a car. The general problem is to separate speech of the driver from interferences such as codriver speech.

3.1. Physical setups

Figure 4 depicts the two setups, denoted I and II.

Setup I has 2 directional microphones on the ceiling, separated by 17 cm. They point to different directions: driver and codriver, respectively.

Setup II has 4 directional microphones in the rear-view mirror, placed on the same line with an interval of 5 cm. All of them point towards the driver.

3.2. Recordings

Data was not simulated, we opted for real data instead. Three 10-seconds long recordings sampled at 16 kHz, made in a Mercedes S320 vehicle, are used in experiments reported in Sections 4.2, 5.5, and 5.6

Train: mannequins playing prerecorded speech. Parameter values are selected on this data.

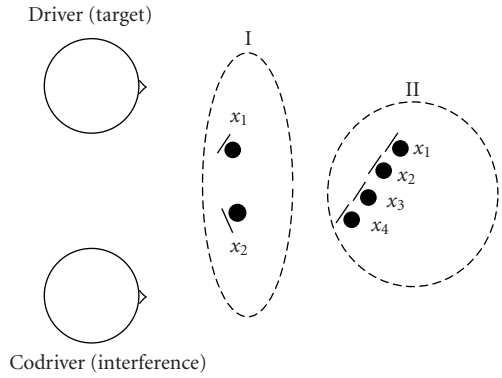


Figure 4: Physical Setups I (2 mics) and II (4 mics).

Test: real human speakers, used for testing only: all parameters determined on train were "frozen."
Noise: both persons silent, the car running at 100 km/h.

For both train and test, we first recorded the driver, then the codriver, and added the two waveforms. Having separate recordings for driver and codriver permits to compute the *true* input SIR at microphone x_1, as the ratio between the instantaneous frame energies of each signal. The true input SIR is the reference for evaluations presented in Sections 4 and 5.

The noise waveform is then added to repeat speech enhancement experiments in a noisy environment, as reported in Section 5.6.

3.3. Sector definition

Figures 5(a) and 5(b) depict the way we defined sectors for each setup. We used prior knowledge of the locations of the driver and the codriver with respect to the microphones. The prior distribution $P_k(\mathbf{v})$ (defined in Section 2.2) was chosen to be a Gaussian in Euclidean coordinates, for the 2 sectors where the people are, and uniform in polar coordinates for the other sectors ($P_k(\mathbf{v}) \propto \|\mathbf{v}\|^{-1}$). Each distribution was approximated with $N = 400$ points.

The motivation for using Gaussian distributions is that we know where the people are on average, and we allow slight motion around the average location. The other sectors have uniform distributions because reverberations may come from any of those directions.

4. INPUT SIR ESTIMATION

This section describes a method to estimate the *input* SIR $\mathrm{SIR}_{\mathrm{in}}(t)$, which is the ratio between driver and codriver energies in signal $x_1(t)$ (see Figure 1). It relies on SAM-SPARSE-MEAN, defined in Section 2.2, and it is used by the "explicit" adaptation control method described in Section 5.2. As discussed in introduction, it is novel, and a priori well adapted to the car environment, as it uses approximate knowledge of both driver and codriver locations.

4.1. Method

From a given frame of samples at microphone 1,

$$\mathbf{x}_1(t) = \left[x_1(t - N_{\mathrm{samples}}), x_1(t - N_{\mathrm{samples}} + 1), \ldots, x_1(t)\right]^{\mathrm{T}}. \tag{22}$$

DFT is applied to estimate the local spectral representation $\mathbf{X}_1 \in \mathbb{C}^{N_{\mathrm{bins}}}$. The energy spectrum for this frame is then defined by $E_1(f) = |X_1(f)|^2$, for $1 \le f \le N_{\mathrm{bins}}$.

In order to estimate the input SIR, we propose to estimate the proportion of the overall frame energy $\sum_f E_1(f)$ that belongs to the driver and to the codriver, respectively. Then the input SIR is estimated as the ratio between the two. Within the sparsity assumption context of Section 2, the following two estimates are proposed:

$$\widehat{\mathrm{SIR}}_1 \stackrel{\mathrm{def}}{=} \frac{\sum_f E_1(f) \cdot P\left(\text{sector } S_{\mathrm{driver}} \text{ active at frequency } f \mid \hat{\boldsymbol{\Theta}}(f)\right)}{\sum_f E_1(f) \cdot P\left(\text{sector } S_{\mathrm{codriver}} \text{ active at frequency } f \mid \hat{\boldsymbol{\Theta}}(f)\right)},$$

$$\widehat{\mathrm{SIR}}_2 \stackrel{\mathrm{def}}{=} \frac{\sum_f P\left(\text{sector } S_{\mathrm{driver}} \text{ active at frequency } f \mid \hat{\boldsymbol{\Theta}}(f)\right)}{\sum_f P\left(\text{sector } S_{\mathrm{codriver}} \text{ active at frequency } f \mid \hat{\boldsymbol{\Theta}}(f)\right)}, \tag{23}$$

where $P(\cdot \mid \boldsymbol{\Theta}(f))$ is the posterior probability given by (9) and (10). Both $\widehat{\mathrm{SIR}}_1$ and $\widehat{\mathrm{SIR}}_2$ are a ratio between two mathematical expectations over the whole spectrum. $\widehat{\mathrm{SIR}}_1$ weights each frequency with its energy, while $\widehat{\mathrm{SIR}}_2$ weights all frequencies equally. In the case of a speech spectrum, which is wideband but has most of its energy in low frequencies, this means that $\widehat{\mathrm{SIR}}_1$ gives more weights to the low frequencies, while $\widehat{\mathrm{SIR}}_2$ gives equal weights to low and high frequencies. From this point of view, it can be expected that $\widehat{\mathrm{SIR}}_2$ provides better results as long as microphones are close enough to avoid spatial aliasing effects.

Note that $\widehat{\mathrm{SIR}}_2$ seems less adequate than $\widehat{\mathrm{SIR}}_1$ in theory: it is a ratio of numbers of frequency bins, while the quantity to estimate is a ratio of energies. However, in practice, it follows the same trend as the input SIR: due to the wideband nature of speech, whenever the target is louder than the interference, there will be more frequency bins where it is dominant, and vice-versa. This is supported by experimental evidence in the meeting room domain [12]. To conclude, we can expect a biased relationship between $\widehat{\mathrm{SIR}}_2$ and the true input SIR, that needs to be compensated (see the next section).

4.2. Experiments

On the entire recording train, we ran the source detection algorithm described in Section 2 and compared the estimates $\widehat{\mathrm{SIR}}_1$ or $\widehat{\mathrm{SIR}}_2$ with the true input SIR, which is defined in Section 3.2.

First, we noted that an additional affine scaling in log domain (fit of a first order polynomial) was needed. It consists in choosing two parameters Q_0, Q_1 that are used to correct

Table 1: RMS error of input SIR estimation calculated in log domain (dB). Percentages indicate the ratio between RMS error and the dynamic range of the true input SIR (max-min). Values in brackets indicate the correlation between true and estimated input SIR.

(a) Results on train. The best result for each setup is in bold face.

Setup	Dynamic range	Method	Hard decision ($\lambda = 0$)	Soft decision ($\lambda > 0$)
I (2 mics)	87.8 dB	$\widehat{\text{SIR}_1}$	10.5% (0.90)	$\lambda = 12.8$: **10.2%** (**0.91**)
		$\widehat{\text{SIR}_2}$	16.0% (0.75)	$\lambda = 22.7$: 12.5% (0.86)
II (4 mics)	88.0 dB	$\widehat{\text{SIR}_1}$	12.0% (0.86)	($\lambda = 0$)
		$\widehat{\text{SIR}_2}$	13.1% (0.83)	$\lambda = 10.7$: **11.2**% (**0.89**)

(b) Results on test and test + noise. Methods and parameters were selected on train.

Setup	Dynamic range	Method		Results on test	
				clean	test+ noise
I	71.6 dB	$\widehat{\text{SIR}_1}$, soft	All frames	14.0% (0.77)	15.1% (0.73)
			True input SIR > 6 dB	16.1% (0.25)	17.8% (0.27)
			True input SIR < −6 dB	12.4% (0.71)	16.3% (0.63)
II	70.2 dB	$\widehat{\text{SIR}_2}$, soft	All frames	9.3% (0.90)	11.4% (0.84)

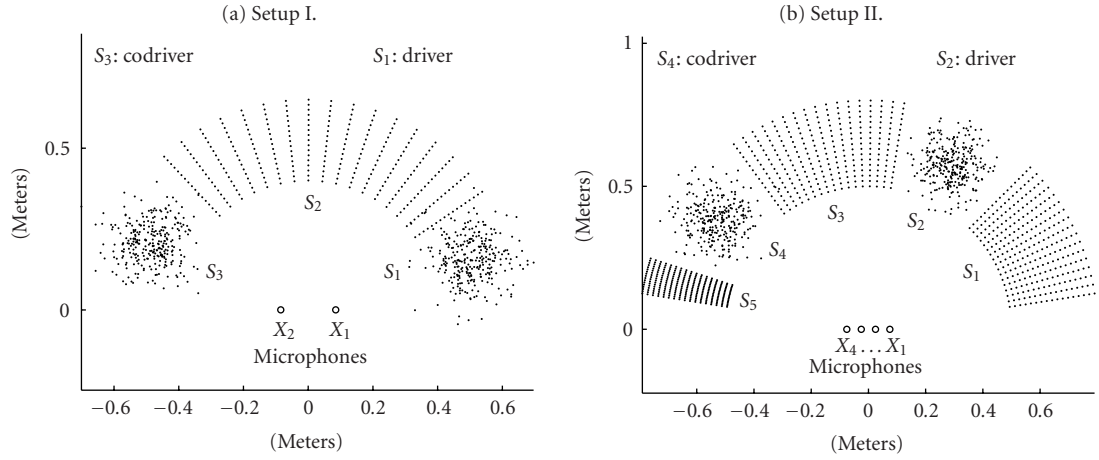


Figure 5: Sector definition. Each dot corresponds to a $\mathbf{v}_{k,n}$ location, as defined in Section 2.3.

the SIR estimate: $Q_1 \cdot \log \widehat{\text{SIR}} + Q_0$. It compensates for the simplicity of the function chosen for probability estimation (9), as well as a bias in the case of $\widehat{\text{SIR}_2}$. This affine scaling is the only post-processing that we used: temporal filtering (smoothing), as well as calibration of the average signal levels, were not used. For each setup and each method, we tuned the 3 parameters (λ, Q_0, Q_1) on train in order to minimize the RMS error of input SIR estimation, in log domain (dB). Results are reported in Table 1a. In all cases, an RMS error of about 10 dB is obtained, and soft decision ($\lambda > 0$) is beneficial. In Setup I, $\widehat{\text{SIR}_1}$ gives the best results. In Setup II, $\widehat{\text{SIR}_2}$ gives the best results. This confirms the above-mentioned expectation that $\widehat{\text{SIR}_2}$ yields better results when microphones are close enough. For both setups, the correlation between true SIR and estimated SIR is about 0.9.

For each setup, a time plot of the results of the best method is available, see Figures 6(a) and 6(b). The estimate follows the true value very accurately most of the time. Errors happen sometimes when the true input SIR is high. One possible explanation is the directionality of the microphones, which is not exploited by the sector-based detection algorithm. Also the sector-based detection gives equal role to all microphones, while we are mostly interested in $x_1(t)$. In spite of these limitations, we can safely state that the obtained SIR curve is very satisfying for triggering the adaptation, as verified in Section 5.

As it is not sufficient to evaluate results on the same data that was used to tune the 3 parameters (λ, Q_0, Q_1), results on the test recording are also reported in Table 1b and Figures 6(c) and 6(d). Overall, all conclusions made on train still hold on test, which tends to prove that the proposed approach is not too dependent on the training data. However, for Setup I, a degradation is observed, mostly on regions with high input SIR, possibly because of the low coherence

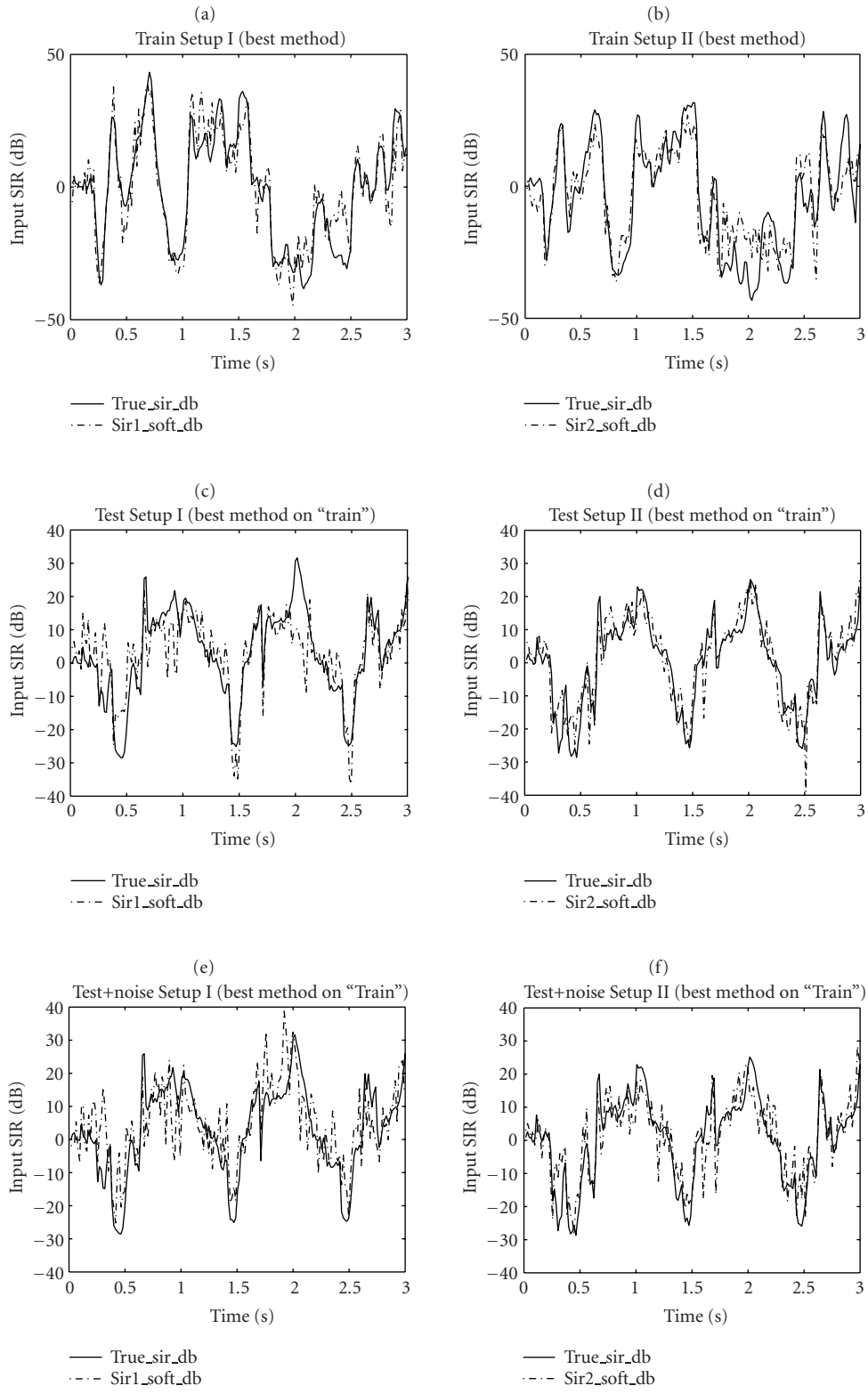
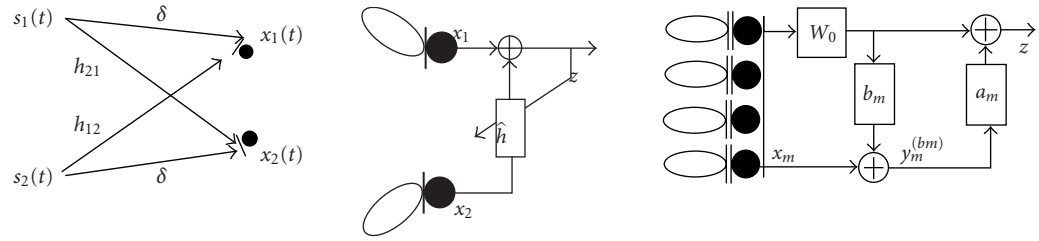


Figure 6: Estimation of the input SIR for Setups I (left column) and II (right column). Beginning of recordings train (top row), test (middle row), and test + noise (bottom row).



(a) Setup I: mixing channels.

(b) Setup I: noise canceller.

(c) Setup II: GSC.

Figure 7: Linear models for the acoustic channels and the adaptive filtering.

between the two directional microphones, due to their very different orientations. However, an interference cancellation application with Setup I mostly needs accurate detection of periods, of negative input SIR rather than positive input SIR. On those periods the RMS error is lower (12.4%). Section 5 confirms the effectiveness of this approach in a speech enhancement application. For Setup II, the results are quite similar to those of train.

Results in 100 km/h noise (*test* + *noise*) are also reported in Table 1b and Figures 6(e) and 6(f). The parameter values are the same as in the clean case. The curves and the relative RMS error values show that the resulting estimate is more noisy, but still follows the true input SIR quite closely in average, and correlation is still high. The estimated ratio still seems accurate enough for adaptation control in noise, as confirmed by Section 5.6. This can be contrasted with the fact that car noise violates the sparsity assumption with respect to speech. A possible explanation is that in (23), numerator and denominator are equally affected, so that the ratio is not biased too much by the presence of noise.

To conclude, the proposed methodology for input SIR estimation gives acceptable results, including in noise. The estimated input SIR curve follows the true curve accurately enough to detect periods of activity and inactivity of the driver and codriver. With respect to that application, only one parameter is used: λ, and the affine scaling (Q_0, Q_1) has no impact on results presented in Section 5. This method is particularly robust since it does not need any thresholding or temporal integration over consecutive frames.

5. SPEECH ENHANCEMENT

5.1. Adaptive interference cancellation algorithms

Setup I provides an input SIR of about 6 dB in the driver's microphone signal $x_1(t)$. An estimate of the interference signal is given by $x_2(t)$. Interference removal is attempted with the linear filter $\widehat{\mathbf{h}}$ of length L depicted by Figure 7(b), which is adapted to minimize the output power $\mathbf{E}\{z^2(t)\}$, using the NLMS algorithm [22] with step size μ:

$$\widehat{\mathbf{h}}(t+1) = \widehat{\mathbf{h}}(t) - \mu \frac{\mathbf{E}\{z(t)\mathbf{x}_2(t)\}}{||\mathbf{x}_2(t)||^2}, \tag{24}$$

where $\mathbf{x}_2(t) = [x_2(t), x_2(t-1), \ldots, x_2(t-L+1)]^{\mathrm{T}}$, $\widehat{\mathbf{h}}(t) = [\widehat{h}_0(t), \widehat{h}_1(t), \ldots, \widehat{h}_{L-1}(t)]^{\mathrm{T}}$, $\|\mathbf{x}\|^2 = \sum_{i=1}^{L} x^2(i)$, and $\mathbf{E}\{\cdot\}$ denotes expectation, taken over realizations of stochastic processes (see Section 5.3 for its implementation).

To prevent instability, adaptation of $\widehat{\mathbf{h}}$ must happen only when the interference is active: $\|\mathbf{x}_2(t)\|^2 \neq 0$, which is assumed true in the rest of this section. In practice, a fixed threshold on the variance of $x_2(t)$ can be used.

To prevent target cancellation, adaptation of $\widehat{\mathbf{h}}$ must happen *only* when the interference is active and dominant.

In Setup II, $M = 4$ directional microphones are in the rear-view mirror, all pointing at the target. It is therefore not possible to use any of them as an estimate of the codriver interference signal. A suitable approach is the linearly constrained minimum variance beamforming [23] and its robust GSC implementation [24]. It consists of two filters b_m and a_m for each input signal $x_m(t)$, with $m = 1, \ldots, M$, as depicted by Figure 7(c). Each filter b_m (resp., a_m) is adapted to minimize the output power of $y_m^{(b_m)}(t)$ (resp., $z(t)$), as in (24). To prevent leakage problems, the b_m (resp., a_m) filters must be adapted *only* when the target (resp., interference) is active and dominant.

5.2. Implicit and explicit adaptation control

For both setups, an adaptation control is required that slows down or stops the adaptation according to target and interference activity. Two methods are proposed: "implicit" and "explicit." The implicit method introduces a continuous, adaptive step-size $\mu(t)$, whereas the explicit method relies on a binary decision, whether to adapt or not.

Implicit method

We present the method in details for Setup I. They also apply to Setup II, as described in Section 5.3. The goal is to increase the adaptation step-size whenever possible, while not turning (24) into an unstable divergent process. With respect to existing implicit approaches, the novelty is a well-grounded mechanism to prevent instability while using the filtered output.

For Setup I, as depicted by Figure 7(a), the acoustic mixing channels are modelled as

$$\begin{aligned} x_1(t) &= s_1(t) + h_{12}(t) * s_2(t), \\ x_2(t) &= h_{21}(t) * s_1(t) + s_2(t), \end{aligned} \tag{25}$$

where $*$ denotes the convolution operator.

As depicted by Figure 7(b), the enhanced signal is $z(t) = x_1(t) + \widehat{h}(t) * x_2(t)$, therefore,

$$\begin{aligned} z(t) &= \underbrace{\left(\delta(t) + \widehat{h}(t) * h_{21}(t)\right)}_{} * s_1(t) + \underbrace{\left(h_{12}(t) + \widehat{h}(t)\right)}_{} * s_2(t) \\ &= \Omega(t) * s_1(t) + \Pi(t) * s_2(t). \end{aligned} \tag{26}$$

The goal is to minimize $\mathbf{E}\{\varepsilon^2(t)\}$, where $\varepsilon(t) = \Pi(t) * s_2(t)$. It can be shown [25] that when $s_1(t) = 0$, an optimal step-size is given by $\mu_{\text{impl}}(t) = \mathbf{E}\{\varepsilon^2(t)\}/\mathbf{E}\{z^2(t)\}$.

We assume s_2 to be a white excitation signal, then,

$$\mu_{\text{impl}}(t) = \mathbf{E}\{\Pi^2(t)\}\frac{\mathbf{E}\{x_2^2(t)\}}{\mathbf{E}\{z^2(t)\}} = \mathbf{E}\{\Pi^2(t)\}\frac{||\mathbf{x}_2||^2}{\|\mathbf{z}\|^2}. \tag{27}$$

Note

Under stationarity and ergodicity assumptions, $\mathbf{E}\{\cdot\}$ is implemented by averaging on a short time-frame:

$$\mathbf{E}\{x^2(t)\} = (1/L)\|\mathbf{x}\|^2. \tag{28}$$

As $\mathbf{E}\{\Pi(t)^2\}$ is unknown, we approximate it with a very small positive constant ($0 < \mu_0 \ll 1$) close to the system mismatch expected when close to convergence:

$$\mu_{\text{impl}}(t) \approx \mu_0 \frac{||\mathbf{x}_2||^2}{\|\mathbf{z}\|^2}, \tag{29}$$

and (24) becomes

$$\widehat{\mathbf{h}}(t+1) = \widehat{\mathbf{h}}(t) - \mu_0 \frac{\mathbf{E}\{z(t)\mathbf{x}_2(t)\}}{||\mathbf{z}(t)||^2}. \tag{30}$$

The domain of stability of the NLMS algorithm [22] is defined by $\mu_{\text{impl}}(t) < 2$, therefore (30) can only be applied when $\mu_0(\|\mathbf{x}_2\|^2/\|\mathbf{z}\|^2) < 2$. In other cases, a fixed step-size adaptation must be used as in (24). The proposed implicit adaptive step-size is therefore

$$\mu(t) = \begin{cases} \mu_{\text{impl}}(t) & \text{if } \mu_{\text{impl}}(t) < 2 \text{ (stable case)}, \\ \mu_0 & \text{otherwise (unstable case)}, \end{cases} \tag{31}$$

$0 < \mu_0 \ll 1$ is a small constant.

This effectively reduces the step-size when the current target power estimate is large and conversely it adapts faster in absence of the target.

Physical interpretation

Let us assume that $s_1(t)$ and $s_2(t)$ are uncorrelated blockwise stationary white sources of powers σ_1^2 and σ_2^2, respectively. From (25) and (26), we can expand (29) into

$$\mu_{\text{impl}}(t) = \mu_0 \frac{||h_{21}||^2\sigma_1^2 + \sigma_2^2}{||\mathbf{\Omega}(t)||^2\sigma_1^2 + ||\mathbf{\Pi}(t)||^2\sigma_2^2}. \tag{32}$$

In a car, the driver is closer to x_1 than to x_2. Thus, given the definition of the mixing channels depicted by Figure 7(a), it is reasonable to assume that $\|h_{21}\| < 1$, h_{21} is causal, and $h_{21}(0) = 0$. Therefore $\|\mathbf{\Omega}(t)\| \geq 1$.

Case 1. The power received at microphone 2, from the target, is *greater* than the power received from the interference: $\|h_{21}\|^2\sigma_1^2 > \sigma_2^2$. In this case, (32) yields

$$\mu_{\text{impl}}(t) < \mu_0 \frac{2||h_{21}||^2\sigma_1^2}{||\mathbf{\Omega}(t)||^2\sigma_1^2 + ||\mathbf{\Pi}(t)||^2\sigma_2^2} < 2\mu_0 \frac{||h_{21}||^2}{||\mathbf{\Omega}(t)||^2} < 2, \tag{33}$$

which falls in the "stable case" of (31).

Case 2. The power received at microphone 2, from the target, is *less* than the power received from the interference: $\|h_{21}\|^2\sigma_1^2 \leq \sigma_2^2$. In this case, (32) yields

$$\mu_{\text{impl}}(t) \leq \mu_0 \frac{2\sigma_2^2}{||\mathbf{\Omega}(t)||^2\sigma_1^2 + ||\mathbf{\Pi}(t)||^2\sigma_2^2}, \tag{34}$$

therefore,

$$||\mathbf{\Omega}(t)||^2\frac{\sigma_1^2}{\sigma_2^2} + ||\mathbf{\Pi}(t)||^2 \leq 2\frac{\mu_0}{\mu_{\text{impl}}(t)}. \tag{35}$$

Thus, in the "unstable case" of (31), we have

$$\begin{aligned} ||\mathbf{\Pi}(t)||^2 &\leq \mu_0, \\ \frac{\sigma_1^2}{\sigma_2^2} &\leq \frac{\mu_0}{||\mathbf{\Omega}(t)||^2} \leq \mu_0. \end{aligned} \tag{36}$$

The first line of (36) means that the adaptation is close to convergence. The second line of (36) means that the input SIR is very close to zero, that is, the interference is largely dominant. Overall, this is the only "unstable case," that is, when we fall back on $\mu_{\text{impl}}(t) = \mu_0$ (31).

Explicit method

For both setups, the sector-based method described in Section 4 is used to directly estimate the input SIR at $x_1(t)$. Two thresholds are set to detect when the target (resp., the interference) is dominant, which determines whether or not the fixed step-size adaptation of (24) should be applied.

5.3. Implementation details

In Setup I, the $\widehat{h}$ filter has length $L = 256$. In Setup II, the b_m filters have length $L = 64$ and the a_m filters have length $L = 128$.

For all methods, the filters are initialized as follows. In Setup I, filter $\widehat{h}$ is initialized to zeros. In Setup II, filters b_m are initialized to cancel signals coming from driver's direction of arrival [23], and the filters a_m are initialized to zeros.

Adaptation is implemented as follows

(i) No control: a baseline method that adapts all the time, with a constant step size, as in (24). In Setup II, filters a_m are adapted all the time and filters b_m are not adapted.

(ii) Implicit method: in both setups, all filters are adapted all the time, with the adaptive step-size of (31). In Setup II, the tunable constant parameter μ_0 was found to be larger for a_m (0.01) than for b_m (0.0001).

(iii) Explicit method: all filters are adapted with (24). In Setup I, filter $\widehat{\mathbf{h}}$ is adapted only when the estimated input SIR is below a threshold. In Setup II, filter a_m (resp., b_m) is adapted only when the estimated input SIR is below (resp., above) a threshold.

Note on (24) and (30): in the original NLMS algorithm [22], the instantaneous estimate $\mathbf{E}\{z(t)\mathbf{x}_2(t)\} \approx z(t)\mathbf{x}_2(t)$ is used and filter coefficients are updated every sample. In this work, in order to reduce computational load, filter coefficients are updated only once every K sample, and $\mathbf{E}\{z(t)\mathbf{x}_2(t)\}$ is estimated by averaging the K instantaneous estimates ($K = 64$, 4 millisecond for $f_s = 16$ kHz). The underlying assumption is that signals are stationary and ergodic within the current block. See [26] for a sample-by-sample study.

5.4. Performance evaluation

For both setups, we measured the instantaneous SIR improvement on the real 16 kHz recordings, with respect to the output when no adaptation is performed. Thus, the reference in Setup I is the true input SIR at microphone x_1, and the reference in Setup II is the SIR at the output of the delay-and-sum beamformer W_0. "Instantaneous" means on half-overlapping short time-frames—that is, where speech can be safely considered as stationary. We used 32 millisecond-long time-frames. Section 3.2 describes the recordings and the method of computation of the true input SIR.

Five seconds of the train recording were used to tune all parameters. Then the entire test recording (real human speakers, 10 seconds) was used to test the methods. It contains a significant degree of overlap between the two speakers (56% of speech frames).

Based on the instantaneous SIR improvement, the segmental SIR improvement is computed in three cases: the true input SIR is low, close to 1, or high. "Segmental" means that only frames containing speech from either driver or co-driver or both are considered. This in turns assumes a reliable marking of speech frames and silence frames in the recording of each person.

For a given person, marking speech frames by hand is questionable, as it may well introduce a bias in the evaluation (silence marked as speech and vice-versa). Another possibility was to set a fixed threshold on the frame energy, but then again, it is not clear how to select a value for the threshold without introducing a bias in the evaluation.

Finally, we opted for an unsupervised approach: for each person, a bi-Gaussian model was fitted on the log energy, using the EM algorithm [27]. The Gaussian with the lowest (resp., highest) mean is expected to capture the silent (resp., speech) frames. The resulting posterior probability of speech is an almost binary value, so that a threshold can be easily set (e.g., 0.5 or 0.9) without much impact on the resulting classification into speech frames and silent frames. This way, we attempt to minimize the bias of the performance evaluation.

Below is a description of the 3 cases that were evaluated.

(i) True input SIR < -6 dB: when the energy of the co-driver is dominant in signal x_1. This quantifies how much of the interference signal is cancelled during silences of the driver: a significantly positive value. All three methods can be expected to perform well in this case.

(ii) True input SIR in $[-6 + 6]$ dB: when both driver and codriver are comparatively active. This quantifies how much of the interference signal is cancelled during overlap periods (both persons speaking): a positive value. We can expect a slight degradation in the case of the baseline method, because of leakage issues.

(iii) True input SIR $> +6$ dB: when the energy of the driver is dominant in signal x_1. No improvement is expected here: a value around zero. If this value is markedly negative, it means that a given method is suffering from leakage issues—as expected for the baseline method.

5.5. Experiments: clean data

The first 3 seconds of test are depicted by Figure 8(b). The periods where SIR improvement is consistently close to 0 dB correspond to silences of both speakers. Average SIR improvement over the entire recording is given in Table 2a. The result of the "no control" baseline method highlights the target cancellation problem and confirms the necessity of adaptation control. In both setups, both "implicit" and "explicit" methods are robust against this problem, and the explicit method provides the best results. Although the implicit method does not give the best results (first two rows of the table), we note that it successfully avoids leakage problems (last row of the table). Note that in the case of Setup II, both implicit and explicit approaches give better results than the delay-sum W_0. Overall, all expectations given in Section 5.4 are verified.

5.6. Experiments with 100 km/h noise

The same experiments as in Section 5.5 were conducted again after adding the background road noise waveform noise. The resulting wave files have an average segmental SNR of 11.6 dB in Setup I, and 9.6 dB in Setup II. In the case of the explicit control, the same detection threshold and the same parameters (λ, Q_0, Q_1) were used as those obtained in experiments on clean data. Only the step-size was lowered

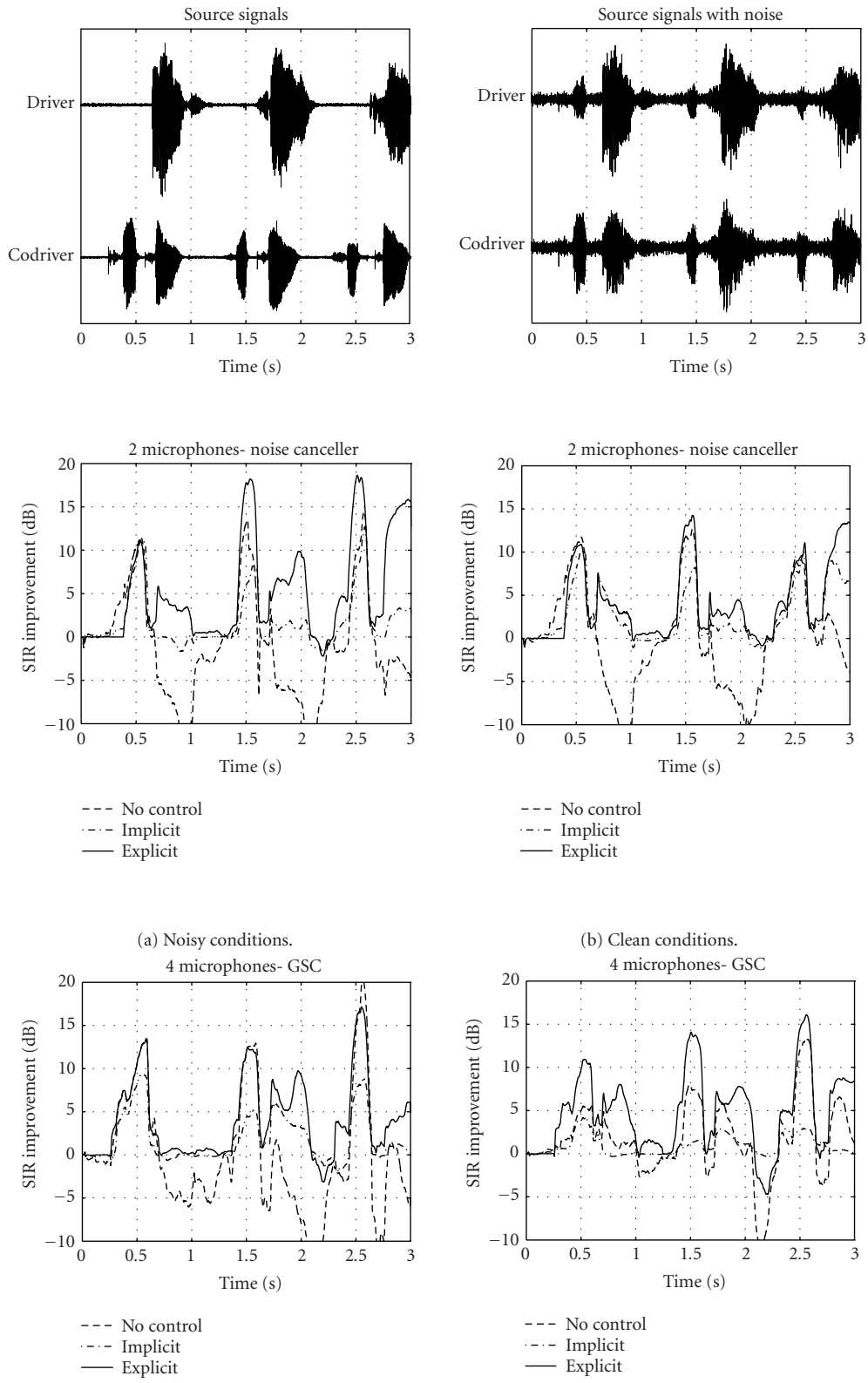


Figure 8: Improvement over input SIR (100 millisecond moving average, first 3 seconds shown). (a) Shows results on clean data (test), whereas (b) shows results on noisy data (test + noise: 100 km/h background road noise).

Table 2: Average segmental SIR improvement in dB. In Setup I, the reference is the output x_1 of microphone 1. In Setup II, the reference is the output of the delay-sum W_0. (W_0 brings an SIR improvement over x_1 of 0.1, 1.6, 2.2 dB, resp., in the "codriver," "both," and "driver" cases.)

(a) test (clean data).

	Setup I (2 mics) reference: x_1			Setup II (4 mics) reference: W_0		
Range of the true input SIR	No control (baseline)	Implicit	Explicit	No control (baseline)	Implicit	Explicit
< -6 (codriver)	6.5	5.9	10.7	10.4	6.1	10.5
$[-6, +6]$ (both)	−0.6	1.2	5.8	0.6	2.3	3.3
$> +6$ (driver)	−7.7	−0.2	2.6	−10.0	0.0	−0.8

(b) test + noise.

	Setup I (2 mics) reference: x_1			Setup II (4 mics) reference: W_0		
Range of the true input SIR	No control (baseline)	Implicit	Explicit	No control (baseline)	Implicit	Explicit
< -6 (codriver)	6.4	7.1	7.4	7.9	3.8	10.3
$[-6, +6]$ (both)	1.0	2.7	3.5	1.2	1.6	3.2
$> +6$ (driver)	−4.7	0.4	1.9	−6.3	0.2	−2.4

to take into account the lower quality of the incoming signal due to noise.

The goal of this experiment is to determine whether the proposed approaches can cope with background noise. It is not obvious, since they do not explicitly model background noise, which may be incoherent or localized outside of the defined sectors. The hope is that reducing the adaptation step is enough, while keeping all other parameters unchanged.

The result is given in Figure 8(a) and Table 2(b). The behaviour in terms of SIR improvement, both over time and in average, is very similar to the clean case. The only negative result is "explicit" in the "driver" case, which is still no degradation compared to the input SIR at x_1. This is interesting, given that the threshold of the "explicit" method was not changed. Thus, we can state that both implicit and explicit approaches also work in a realistic case of a moving car.

6. CONCLUSION

Two adaptation control methods were proposed to cancel the codriver interference from the driver's speech signal: implicit and explicit control. At no additional cost, the implicit adaptation method provides robustness against leakage, but slower convergence. On the other hand, the explicit adaptation method relies on estimation of target and interference energies. A novel, robust method for such estimation was derived from sector-based detection and localization techniques. It relies on integration of the delay-sum energy over a volume of space, for the same cost as the classical delay-sum. In the end, the explicit control method provides both robustness and good performance. Both implicit and explicit methods are suitable for real-time implementation. One direction for future work is to investigate modelling of the microphone directionality for further enhancement of the sector-based detection framework. A second direction is to test on other noise cases, including other passengers.

APPENDIX

A.

Section A.1 defines a phase domain metric (PDM), similarly to the classical metric definition. Section A.2 proves that any 1-dimensional PDM can be composed into a multidimensional function which is also a PDM.

A.1. Definition of a PDM

Similarly to the classical metric definition, we define a PDM on $\mathbb{R}^P$ as a function $g(\mathbf{x}, \mathbf{y})$ verifying all of the following conditions for all $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathbb{R}^P)^3$:

$$g(\mathbf{x}, \mathbf{y}) \geq 0, \tag{A.1}$$

$$g(\mathbf{x}, \mathbf{y}) = g(\mathbf{y}, \mathbf{x}), \tag{A.2}$$

$$g(\mathbf{x}, \mathbf{y}) = 0 \quad \text{iff } \forall\, p = 1, \ldots, P,\ \exists k_p \in \mathbb{Z},\ x_p = y_p + k2\pi, \tag{A.3}$$

$$g(\mathbf{x}, \mathbf{z}) \leq g(\mathbf{x}, \mathbf{y}) + g(\mathbf{y}, \mathbf{z}). \tag{A.4}$$

It is basically the same as a classical metric, except for (A.3) which reflects the "modulo 2π" definition of angles.

A.2. Property

Let G_1 be a 1-dimensional PDM, that is a PDM on $\mathbb{R}$. For any $P \in \mathbb{N}^*$, let G_P be the following function on $\mathbb{R}^P$:

$$G_P(\mathbf{x}, \mathbf{y}) \stackrel{\text{def}}{=} \sqrt{\frac{1}{P} \sum_{p=1}^{P} G_1(x_p, y_p)^2}. \tag{A.5}$$

The rest of this Section shows that all G_P functions are also PDMs. Equations (A.1), (A.2), and (A.3) are trivial to demonstrate. Equation (A.4) is demonstrated for G_P in the following.

Since G_1 is a PDM, it verifies (A.4) on $\mathbb{R}$. Therefore, for any $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in (\mathbb{R}^P)^3$,

$$G_P(\mathbf{x}, \mathbf{z}) \leq \sqrt{\frac{1}{P} \sum_{p=1}^{P} \left[G_1(x_p, y_p) + G_1(y_p, z_p) \right]^2}. \tag{A.6}$$

Now let us recall the Minkowski inequality [28]. For any $\beta > 1$ and $a_p > 0$, $b_p > 0$,

$$\left[\sum_{p=1}^{P} (a_p + b_p)^\beta \right]^{1/\beta} \leq \left[\sum_{p=1}^{P} {a_p}^\beta \right]^{1/\beta} + \left[\sum_{p=1}^{P} {b_p}^\beta \right]^{1/\beta}. \tag{A.7}$$

By applying the Minkowski inequality to the right-hand side of (A.6), with $\beta = 2$, $a_p = G_1(x_p, y_p)$, and $b_p = G_1(y_p, z_p)$, and dividing by $\sqrt{P}$, we obtain

$$G_P(\mathbf{x}, \mathbf{z}) \leq \sqrt{\frac{1}{P} \sum_{p=1}^{P} G_1(x_p, y_p)^2} + \sqrt{\frac{1}{P} \sum_{p=1}^{P} G_1(y_p, z_p)^2}, \tag{A.8}$$

$$G_P(\mathbf{x}, \mathbf{z}) \leq G_P(\mathbf{x}, \mathbf{y}) + G_P(\mathbf{y}, \mathbf{z}). \tag{A.9}$$

ACKNOWLEDGMENTS

The authors acknowledge the support of the European Union through the HOARSE project. This work was also carried out in the framework of the Swiss National Center of Competence in Research (NCCR) on interactive multimodal information management (IM)2. The authors would like to thank Dr Iain McCowan, Dr Mathew Magimai.-Doss, and Bertrand Mesot for helpful comments and suggestions.

REFERENCES

[1] E. Shriberg, A. Stolcke, and D. Baron, "Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation, disfluencies, and overlapping speech," in *Proceedings of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pp. 139–146, Red Bank, NJ, USA, October 2001.

[2] S. Affes and Y. Grenier, "Test of adaptive beamformers for speech acquisition in cars," in *Proceedings of 5th International Conference on Signal Processing Applications and Technology (ICSPAT '94)*, vol. 1, pp. 154–159, Dallas, Tex, USA, October 1994.

[3] B. D. Van Veen and K. M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[4] D. Van Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '90)*, vol. 2, pp. 833–836, Albuquerque, NM, USA, April 1990.

[5] S. Affes and Y. Grenier, "A signal subspace tracking algorithm for microphone array processing of speech," *IEEE Transactions Speech Audio Processing*, vol. 5, no. 5, pp. 425–437, 1997.

[6] O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 2, pp. 925–828, Atlanta, Ga, USA, May 1996.

[7] M. Buck and T. Haulick, "Robust adaptive beamformers for automotive applications," in *Proceedings of DAGA*, Strasbourg, France, March 2004.

[8] O. Hoshuyama, B. Begasse, A. Sugiyama, and A. Hirano, "A real time robust adaptive microphone array controlled by an SNR estimate," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '98)*, vol. 6, pp. 3605–3608, Seattle, Wash, USA, May 1998.

[9] W. Herbordt, T. Trini, and W. Kellermann, "Robust spatial estimation of the signal-to-interference ratio for non-stationary mixtures," in *Proceedings of International Workshop on Acoustic Echo and Noise Control (IWAENC '03)*, pp. 247–250, Kyoto, Japan, September 2003.

[10] O. Hoshuyama, A. Sugiyama, and A. Hirano, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," *IEEE Transactions Signal Processing*, vol. 47, no. 10, pp. 2677–2684, 1999.

[11] S. Gannot, D. Burshtein, and E. Weinstein, "Signal enhancement using beamforming and nonstationarity with applications to speech," *IEEE Transactions Signal Processing*, vol. 49, no. 8, pp. 1614–1626, 2001.

[12] G. Lathoud and M. Magimai.-Doss, "A sector-based, frequency-domain approach to detection and localization of multiple speakers," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, vol. 3, pp. 265–268, Philadelphia, Pa, USA, March 2005.

[13] D. Ellis and J. Liu, "Speaker turn segmentation based on between-channel differences," in *Proceedings of ICASSP-NIST Meeting Recognition Workshop*, pp. 112–117, Montreal, Quebec, Canada, May 2004.

[14] G. Lathoud, I. A. McCowan, and J.-M. Odobez, "Unsupervised location-based segmentation of multi-party speech," in *Proceedings of ICASSP-NIST Meeting Recognition Workshop*, Montreal, Quebec, Canada, May 2004.

[15] W. Herbordt, W. Kellermann, and S. Nakamura, "Joint optimization of LCMV beamforming and acoustic echo cancellation," in *Proceedings of 12th European Signal Processing Conference (EUSIPCO '04)*, pp. 2003–2006, Vienna, Austria, September 2004.

[16] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Multichannel speech enhancement in cars: explicit vs. implicit adaptation control," in *Proceedings of Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA '05)*, Piscataway, NJ, USA, March 2005.

[17] S. T. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proceedings of 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 1009–1012, Geneva, Switzerland, September 2003.

[18] G. Lathoud and I. A. McCowan, "A sector-based approach for localization of multiple speakers with microphone arrays," in *Proceedings of ISCA Tutorial and Research Workshop on Statistical and Perceptual Audio Processing (SAPA '04)*, Jeju, Korea, October 2004.

[19] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions Acoustics, Speech, Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[20] B. C. J. Moore, *An Introduction to the Psychology of Hearing*, Academic Press, London, UK, 4th edition, 1997.

[21] G. Lathoud, M. Magimai.-Doss, and B. Mesot, "A spectrogram model for enhanced source localization and noise-robust ASR," in *Proceedings of 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, Lisbon, Portugal, September 2005.

[22] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, USA, 1985.

[23] L. J. Griffiths and C. W. Jim, "An alternative approach to linearly constrained adaptive beamforming," *IEEE Transactions on Antennas and Propagation*, vol. 30, no. 1, pp. 27–34, 1982.

[24] O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '96)*, vol. 2, pp. 925–928, Atlanta, Ga, USA, May 1996.

[25] A. Mader, H. Puder, and G. U. Schmidt, "Step-size control for acoustic echo cancellation filters—an overview," *Signal Processing*, vol. 80, no. 9, pp. 1697–1719, 2000.

[26] J. Bourgeois, J. Freudenberger, and G. Lathoud, "Implicit control of noise canceller for speech enhancement," in *Proceedings of 9th European Conference on Speech Communication and Technology (INTERSPEECH '05)*, Lisbon, Portugal, September 2005.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[28] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*, Prentice-Hall, Upper Saddle River, NJ, USA, 2000.

Guillaume Lathoud received his M.S. in computer science and telecommunications in 1999 at the Institut National des Telecommunications (INT), France. He then spent more than 2 years as a Member of the Digital Television Team at the National Institute of Standards and Technology (NIST), USA, participating to terrestrial DTV standardization and implementation efforts. He joined IDIAP Research Institute, Switzerland, in 2002 as a Ph.D. student. His interests include microphone array processing, audio source localization, speaker tracking, multimodal processing, and noise-robust speech recognition.

Julien Bourgeois received the M.S. degree from the ESIEE Paris (Ecole Superieure d'Ingenieurs en Electronique et Electrotechnique de Paris) in 2001. He received a B.S. in mathematics concurrently from the University de Marne-La-Vallée in 2000. He joined DaimlerChrysler Research and Technology in 2002 as a Ph.D. student. His current research interests include multichannel signal processing and blind source separation with application to speech enhancement.

Jüergen Freudenberger received his Diplom-Ingenieur and Dr.-Ing. degrees in electrical engineering from the University of Ulm, Germany, in 1999 and 2004, respectively. After completing his dissertation, he joined DaimlerChrysler Research and Technology. Since July 2005, he is with Harman/Becker Automotive Systems. His research interests include information and coding theory, in particular transmission over channels with feedback, and signal processing for speech signals. He received a Villigst scholarship and is the recipient of the "ITG Frderpreis 2005" Award of the German Information Technology Society (ITG).