IDIAP RESEARCH REPORT

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

Guillaume Lathoud [a,b] Julien Bourgeois [c]
Jürgen Freudenberger [c]

IDIAP–RR 04-67

December 2004

[a] IDIAP Research Institute, CH-1920 Martigny, Switzerland
[b] Swiss Federal Institute of Technology (EPFL), CH-1015 Lausanne, Switzerland
[c] DaimlerChrysler Research and Technology, Ulm, Germany

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 04-67

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

Guillaume Lathoud   Julien Bourgeois   Jürgen Freudenberger

December 2004

Abstract. Speech-based command interfaces are becoming more and more common in cars. Applications include automatic dialog systems for hands-free phone calls as well as more advanced features such as navigation systems. However, interferences, such as speech from the codriver, can greatly hamper the performance of the speech recognition component, which is crucial for those applications. This issue can be addressed with *adaptive* interference cancellation techniques such as the Generalized Sidelobe Canceller (GSC). In order to cancel the interference (codriver) while not cancelling the target (driver), adaptation must happen *only* when the interference is active and dominant. To that purpose, this paper proposes two efficient adaptation control methods called "implicit" and "explicit". The "implicit" method adapts the filter coefficients continuously, the speed of adaptation being determined by the energy of the filtered output signal. The "explicit" method decides whether to adapt or not in a binary fashion, depending on prior estimation of target and interference energies. A major contribution of this paper is a direct, robust method for such estimation, directly derived from sector-based detection and localization techniques. Experiments on real in-car data validate both adaptation methods, including a case with 100 km/h background road noise.

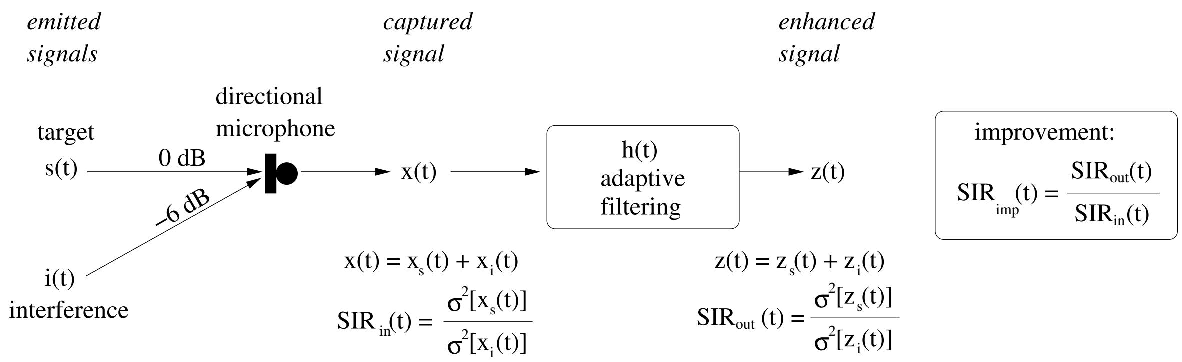


Figure 1: Schematic diagram of the entire acquisition process, from emitted signals to the enhanced signal. This paper focuses on the adaptive filtering block, so that the SIR improvement $\mathrm{SIR}_{\mathrm{imp}}(t)$ is maximized when the interference is active (interference cancellation). The s and t subscript designate contributions of target and interference, respectively. The whole process is supposed to be linear.

1 Introduction

Speech-based command interfaces are becoming more and more common in cars. Applications include automatic dialog systems for hands-free phone calls as well as more advanced features such as navigation systems. However, interferences, such as speech from the codriver, can greatly hamper the performance of the speech recognition component, which is crucial for those applications. This is of particular importance since spontaneous multi-party speech contains lots of overlaps between participants, as found in meetings [1].

An immediate enhancement can be directly provided by hardware, through the use of a directional microphone oriented towards the driver. The directional microphone lowers the energy level of the codriver interference, since the codriver is not placed in front of it. For example, in the Mercedes S320 setup used in this article, if the directional microphone receives sounds with equal energy from both directions, the signal coming out of the directional microphone is the sum of two signals, the driver's signal being 6 dB higher than the codriver's, in terms of energies. However, given the highly time-varying nature of speech, at a given time, low energy from the driver may coincide with energy from the codriver so that this improvement may not be enough. Therefore, there is a need for an additional *software* improvement. This issue can be addressed with *adaptive* beamforming techniques. In the rest of this section, first the problem is formalized and then adaptive beamforming techniques are reviewed. Finally, the proposed approaches are described.

Let us now formalize the problem. For a given short time-frame of typically 20 to 30 ms, speech signals can be modelled as ergodic, stationary processes. For a given signal $x(t)$, $\sigma^2\left[x(t)\right]$ is its variance (energy), which can be computed on a single realization of x, over the short time-frame around t, over which the ergodicity and stationarity assumptions hold. The Signal-to-Interference Ratio (SIR) is then defined as the ratio between two such variances, as detailed in Fig. 1. Each step of the acquisition process is modelled in a linear fashion. The focus of this paper is to maximize the SIR improvement from the *captured signal* $x(t)$ to the *enhanced signal* $z(t)$. To that purpose, a linear filter $h(t)$ is used, that can vary over time, thus the task is called *adaptive* filtering.

This task can be addressed with a class of solutions called beamforming: several microphones are placed close together, thus producing several simultaneous signals with slightly different characteristics, due to their different locations. "Beamforming" means recombining these various inputs to improve the quality of a target signal coming from a given direction, while cancelling interference signals coming from other directions. In our case this means maximizing the SIR improvement $\mathrm{SIR}_{\mathrm{imp}}(t)$.

Many beamforming algorithms have been proposed with various degrees of relevance in the car environment. The most simple kind, delay-and-sum beamforming, provides limited reduction of the

(a) Implicit approach (b) Explicit approach

Figure 2: Implicit and explicit adaptation control. $\mathbf{x}\,(t) = [x_1\,(t) \cdots x_M\,(t)]^T$ are the signals captured by the M microphones, and $\mathbf{h}\,(t) = [\mathbf{h}_1\,(t) \cdots \mathbf{h}_M\,(t)]^T$ are their associated filters.

background noise and the coherent interferers. Superdirective beamformers, see [2] and the references herein, are derived from the Mininum Variance Distortionless Response principle (MVDR). The original adaptive versions assume a-priori known and fixed acoustic propagation channel and exactly calibrated microphones. This allows for target-free interferer estimates in the generalized sidelobe canceller (GSC) structure. If this assumption is not met, as in most practical setups, not only the interferer but also the target is attenuated. This effect is called "target leakage", and can be addressed with two strongly related schemes. First, one detects periods of target activity, and stops adaptation during those periods [3]. Second, one tracks the acoustic channel, which, in turn, can only be done when the target is dominant [4, 5, 6]. Self-calibration algorithms [7] are closely related. Both ways show the necessity of a reliable estimation of the input SIR: $\hat{\mathrm{SIR}}_{\mathrm{in}}\,(t)$. In many implementations, $\hat{\mathrm{SIR}}_{\mathrm{in}}\,(t)$ is estimated based on adapted quantities such as $z\,(t)$ [7, 6], as shown by Fig. 2a. This approach is called "implicit" in the rest of this paper. Speaker detection errors are fed back into the adapted parts: a single detection error may have dramatical effects. Such techniques have good results when target and interference speech exhibit little overlap in time. For simultaneous speakers, it is more robust to decouple the detection from the the adaptation [8], as described by Fig. 2b. This approach is called "explicit" in the rest of this paper.

The contribution of this paper is twofold: first, an implicit method is proposed, where the speed of adaptation is determined in a manner that makes it very robust to target leakage problems. This robustness is both shown in theory and verified in experiments. There is no additional computational cost. The second contribution is an explicit method that decides whether to adapt or not in a binary fashion, depending on prior estimation of target and interference energies. A major contribution of this paper is a direct, robust method for such estimation from the received signals $\mathbf{x}\,(t)$ themselves, that extends a previously proposed sector-based, frequency-domain detection and localization technique [9]. We show that it is closely related to delay-sum beamforming, *averaged* over a sector of space. We also give a low-cost practical implementation. A topological interpretation is also included, introducing the concept of Phase Domain Metric (PDM).

Experiments on real in-car data validate both contributions, testing two setups: either 2 or 4 directional microphones. In both cases, we show that the sector-based method can reliably estimate the input SIR ($\hat{\mathrm{SIR}}_{\mathrm{in}}\,(t)$). Application of the implicit and explicit control methods for speech enhancement shows that in both cases, the output SIR ($\mathrm{SIR}_{\mathrm{out}}\,(t)$) is improved in a robust manner. Tests in 100 km/h driving conditions are included. 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 details the sector-based activity detection approach recently proposed in [9], proposes improvements, a real-time implementation as well as physical and topological interpretations. 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. This technique is used in Section 5, which defines the adaptation control techniques and provides experimental results. Section 6 concludes. This paper constitutes a detailed version of an abstract presented in 2005 [10].

2 Sector-Based, Frequency-Domain Activity Detection

This section revisits and extends the SAM-SPARSE audio source detection and localization approach, previously proposed and tested on multi-party speech in the meeting room context [9]. The main idea is to divide the space around a microphone array 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 in phase domain, 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, an excellent explanation and review of past work around this idea is included in [11]. In brief: if two speech sources are active at a given time, "sparsity assumption" means that within each frequency bin, only one speech source is supposed to be active. This simplification is supported by direct evidence: statistical analysis of real speech signals 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. This is due to the highly-varying, non-stationary nature of speech.

We take advantage of this paper to generalize the SAM-SPARSE approach (Sections 2.1 and 2.2), simplify its implementation (Section 2.3), and give physical and topological interpretations (Section 2.4 and Annex, respectively). An extension is proposed to allow for a "soft" decision within each frequency bin, as opposed to the "hard decision" taken in [9].

2.1 A Phase Domain Metric

First, a few notations are defined. 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 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 ($1 \leq f \leq N_{\text{bins}}$), $\mathcal{R}e(\cdot)$ denotes the real part of a complex quantity, and $\hat{G}^{(p)}(f)$ is the estimated frequency domain cross-correlation for microphone pair p ($1 \leq p \leq P$):

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

where $(\cdot)^*$ denotes complex conjugate, i_p and j_p are indices of the 2 microphones: $1 \leq i_p < j_p \leq M$. Note that the total number of microphone pairs is $P = M(M-1)/2$. For a given frequency bin f, we denote the vector of *measured* phase values with:

$$\hat{\boldsymbol{\Theta}}(f) \stackrel{\text{def}}{=} \left[\hat{\theta}^{(1)}(f), \ldots, \hat{\theta}^{(p)}(f), \ldots, \hat{\theta}^{(P)}(f)\right]^T \quad \text{where} \quad \hat{\theta}^{(p)}(f) \stackrel{\text{def}}{=} \angle \hat{G}^{(p)}(f), \tag{2}$$

where $\angle(\cdot)$ designate the argument of a complex value.

The approach defined below (Section 2.2) is built upon the concept of distance between two vectors of phase values (angles in radians). The distance between two such vectors Θ_1 and Θ_2 in $\mathbb{R}^P$ is defined as follows:

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

$d(.,.)$ is defined very similarly to the Euclidean metric, except for the additional sine function, which is here to account for the "modulo 2π" definition of angles. The $1/P$ normalization factor simply ensures that $0 \leq d(.,.) \leq 1$.

We chose to use sine to compare two angle values rather than a piecewise linear function such as $\arg\min_k \left|\theta_1^{(p)} - \theta_2^{(p)} + k2\pi\right|$ for three reasons:

1. **Physical interpretation:** The first reason is that the use of sine is closely related to delay-sum beamforming, as shown in Section 2.4.

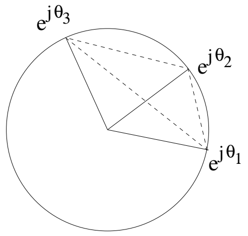


Figure 3: Illustration of the triangular inequality for the PDM in dimension 1: each point on the unit circle corresponds to an angle value modulo 2π. From the Euclidean metric: $\left|e^{j\theta_3} - e^{j\theta_1}\right| \leq \left|e^{j\theta_3} - e^{j\theta_2}\right| + \left|e^{j\theta_2} - e^{j\theta_1}\right|$.

2. **Smoothness:** The second reason involves optimization procedures (e.g. gradient descent): derivates of "$d(.,.)$" are simpler to manipulate when using the sine function, as it does not include the "$\arg\min$" explicitly. Moreover, $d(.,.,)$ is infinitely derivable in all points, which is not the case of "$\arg\min$". This is related to parameter optimization work not presented here.

3. **Topological interpretation:** Finally, the third reason is that $d(.,.)$ is a true Phase Domain Metric (PDM) (see Annex A.1 for a definition). This is straightforward for $P = 1$ by representing any angle θ with a point $e^{j\theta}$ on the unit circle. We then observe that for any two angle values θ_1 and θ_2:
$$\left|e^{j\theta_1} - e^{j\theta_2}\right| = 2\left|\sin\left(\frac{\theta_1-\theta_2}{2}\right)\right| = 2\ d\left(\theta_1, \theta_2\right) \tag{4}$$
so that the triangular inequality of the Euclidean metric directly translates into a triangular inequality for $d(.,.)$, as illustrated by Fig. 3. Hence $d(.,.)$ is a metric for $P = 1$. We have proved the triangular inequality for higher dimensions $P > 1$ (see Annex A.2 for a demonstration).

Note that the intuitive "argmin" alternative, i.e. to compare angle values themselves, is also a PDM, but does not have the "smoothness" and "physical interpretation" properties.

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, 9]. For example, the space around a horizontal planar microphone array can be partitioned in "vertical slices": for k=1 . . . N_{S}:

$$\mathbf{S}_k = \left\{(\rho, \theta_{\mathrm{az}}, \phi_{\mathrm{el}}) \in \mathbb{R}^3 \;\middle|\; \rho \geq \rho_0,\quad 2\pi\frac{k-1}{N_{\mathrm{S}}} \leq \theta_{\mathrm{az}} < 2\pi\frac{k}{N_{\mathrm{S}}},\quad 0 \leq \phi_{\mathrm{el}} \leq \frac{\pi}{2}\right\} \tag{5}$$

where here ρ, θ_{az}, ϕ_{el} designate radius, azimuth and elevation w.r.t. the microphone array center; microphones are all in the sphere $\rho < \rho_0$.

The SAM-SPARSE-MEAN approach treats each frequency bin separately. 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 source is active within that sector *and* that frequency bin. "SPARSE" stands for the sparsity assumption, as discussed in [11]: at most one sector is active per frequency bin. As mentioned earlier, this assumption is in fact a simplification that is strongly supported by statistical observations on concurrent speech of multiple speakers. It was shown in [9] 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 [13], 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, on high frequencies only, where the shadow cast by the head induces noticeable received power differences between the ears, of the order of 10 dB to 20 dB [14]. This is not applicable here, since there is no obstacle between the microphones.

As for computational complexity, it must be noted that since each frequency bin is processed independently, the SAM-SPARSE-MEAN method can be parallelized in a straightforward manner.

• The first step is to compute the root mean square distance ("MEAN") between the measured phase vector $\hat{\mathbf{\Theta}}(f)$ and theoretical phase vectors associated with *all* points within any given sector S_k, at any given frequency f, using the metric defined in Eq. 3:

$$\boxed{\overline{D}_{k,f} \stackrel{\text{def}}{=} \left[\int_{\mathbf{v} \in \mathbf{S}_k} d^2 \left(\hat{\mathbf{\Theta}}(f),\ \mathbf{\Gamma}\left(\mathbf{v}, f\right) \right) P_k\left(\mathbf{v}\right) d\mathbf{v} \right]^{\frac{1}{2}}} \tag{6}$$

where $\mathbf{\Gamma}\left(\mathbf{v}, f\right) = \left[\gamma^{(1)}(\mathbf{v}, f) \ldots \gamma^{(p)}(\mathbf{v}, f) \ldots \gamma^{(P)}(\mathbf{v}, f)\right]^T$ is the vector of theoretical phases associated with location $\mathbf{v}$ and frequency f and $P_k(\mathbf{v})$ is a weighting term: it is the prior distribution of active source location within sector S_k - it represents prior knowledge (e.g. uniform or Gaussian distribution). Note that $\mathbf{v}$ 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 $\mathbf{\Gamma}$ vector is given by:

$$\gamma^{(p)}(\mathbf{v}, f) = \pi \frac{f}{N_{\text{bins}}} \tau^{(p)}(\mathbf{v}), \tag{7}$$

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

$$\tau^{(p)}(\mathbf{v}) = \frac{f_s}{c} \left(||\mathbf{v} - \mathbf{m}_2^{(p)}|| - ||\mathbf{v} - \mathbf{m}_1^{(p)}|| \right), \tag{8}$$

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, $\mathbf{m}_1^{(p)}$ and $\mathbf{m}_2^{(p)} \in \mathbb{R}^3$ are spatial locations of microphone pair p.

• The second step is to determine, for each frequency bin f, the sector to which the measured phase vector is the closest: $k_{min}(f) \stackrel{\text{def}}{=} \arg\min_k \overline{D}_{k,f}$.

Finally, the posterior probability of having at least one active source in sector $S_{k_{min}(f)}$ and at frequency f is modeled with:

$$\boxed{P\left(\text{sector } S_{k_{min}(f)} \text{ active at frequency } f \,\middle|\, \hat{\mathbf{\Theta}}(f)\right) = e^{-\lambda\left(\overline{D}_{k_{\min}(f),f}\right)^2}} \tag{9}$$

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:

$$\boxed{\forall k \neq k_{min}(f) \quad P\left(\text{sector } S_k \text{ active at frequency } f \,\middle|\, \hat{\mathbf{\Theta}}(f)\right) = 0} \tag{10}$$

In previous work [9] only "hard" decisions were taken ($\lambda = 0$) and the entire spectrum was supposed to be active. In particular, this had the effect that even if a sector did not contain any active source at any frequency, it would still be attributed some random part of the spectrum. Eq. 10 represents a generalization ($\lambda > 0$) of previous work, that in practice allows to detect *inactivity* at a given frequency and therefore avoids the random effect. The λ parameter can be trained on any (small) amount of development data. Values around 10 or 20 were found to be reasonable on our development data. In fact the choice of a value for λ can be interpreted as follows, in the case of a single microphone pair $P = 1$:

$$d^2\left(\theta_1, \theta_2\right) = \sin^2\left(\frac{\theta_1 - \theta_2}{2}\right) \tag{11}$$

therefore, a "small" estimated probability

$$e^{-\lambda d^2(\theta_1,\theta_2)} = \epsilon \tag{12}$$

corresponds to an angle difference

$$\Delta\theta = \arg\min_k |\theta_1 - \theta_2 + k2\pi| = 2\sin^{-1}\sqrt{\frac{-\log\epsilon}{\lambda}} \tag{13}$$

which gives, for $\epsilon = 0.1$ and $\lambda = 10$ a phase angle $\Delta\theta \approx \frac{\pi}{3}$. For $\epsilon = 0.1$ and $\lambda = 20$, we have $\Delta\theta \approx \frac{\pi}{5}$. $\Delta\theta$ indicates the value of the phase angle beyond which the probability of activity is very small. This is equivalent to detect *inactivity* within a given frequency bin.

2.3 Practical Implementation

In general it is not possible to derive an analytical solution for Eq. 6. It is therefore approximated with a discrete summation:

$$\overline{D}_{k,f} \approx \hat{\overline{D}}_{k,f} \quad \text{where} \quad \hat{\overline{D}}_{k,f} \stackrel{\text{def}}{=} \sqrt{\frac{1}{N}\sum_{n=1}^{N} d^2\left(\hat{\mathbf{\Theta}}(f),\ \mathbf{\Gamma}\left(\mathbf{v}_{k,n}, f\right)\right)} \tag{14}$$

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. Note that the sampling is not necessarily random, for example it could be a regular grid in the case of a uniform distribution.

This approximation can be expressed in a manner that does not depend on the number of points N, as shown in the rest of this section.

$$\left(\hat{\overline{D}}_{k,f}\right)^2 = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{P}\sum_{p=1}^{P}\sin^2\left(\frac{\hat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}_{k,n},\ f)}{2}\right) \tag{15}$$

Using the relation $\sin^2 u = \frac{1}{2}\left(1 - \cos 2u\right)$ we can write:

$$\left(\hat{\overline{D}}_{k,f}\right)^2 = \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - \frac{1}{N}\sum_{n=1}^{N}\cos\left(\hat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}_{k,n},\ f)\right)\right\} \tag{16}$$

$$\left(\hat{\overline{D}}_{k,f}\right)^2 = \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - \mathcal{R}e\left[\frac{1}{N}\sum_{n=1}^{N} e^{j\left(\hat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}_{k,n},\ f)\right)}\right]\right\} \tag{17}$$

$$\left(\hat{\overline{D}}_{k,f}\right)^2 = \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - \mathcal{R}e\left[e^{j\hat{\theta}^{(p)}(f)}\frac{1}{N}\sum_{n=1}^{N} e^{-j\gamma^{(p)}(\mathbf{v}_{k,n},\ f)}\right]\right\} \tag{18}$$

$$\left(\hat{\overline{D}}_{k,f}\right)^2 = \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - \mathcal{R}e\left[e^{j\hat{\theta}^{(p)}(f)}\ A_k^{(p)}(f)\ e^{-jB_k^{(p)}(f)}\right]\right\} \tag{19}$$

$$\boxed{\left(\hat{\overline{D}}_{k,f}\right)^2 = \frac{1}{2P}\sum_{p=1}^{P}\left\{1 - A_k^{(p)}(f)\ \cos\left(\hat{\theta}^{(p)}(f) - B_k^{(p)}(f)\right)\right\}} \tag{20}$$

where $\mathcal{R}e(.)$ is the real part of a complex quantity, $A_k^{(p)}(f)$ and $B_k^{(p)}(f)$ are two values in $\mathbb{R}$ that do not depend on the measured phase $\hat{\theta}^{(p)}(f)$:

$$A_k^{(p)}(f) \stackrel{\text{def}}{=} \left|Z_k^{(p)}(f)\right|, \quad B_k^{(p)}(f) \stackrel{\text{def}}{=} \angle Z_k^{(p)}(f) \quad \text{and} \quad Z_k^{(p)}(f) \stackrel{\text{def}}{=} \frac{1}{N}\sum_{n=1}^{N} e^{j\gamma^{(p)}(\mathbf{v}_{k,n},\ f)} \tag{21}$$

This is particularly interesting from the point of view of computational complexity: the approximation has to be computed only once, in the form of A and B parameters. Any large number N can be used, therefore, the approximation $\hat{\overline{D}}_{k,f}$ can be as close to $\overline{D}_{k,f}$ as desired. During runtime, the cost of computing $\hat{\overline{D}}_{k,f}$ does not depend on N: it is directly proportional to $N_{\text{bins}} \cdot P \cdot N_{\text{S}}$.

In other terms, the proposed approach ($\overline{D}_{k,f}$) does not suffer from its practical implementation ($\hat{\overline{D}}_{k,f}$), 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, where "average" means average over all points $\mathbf{v}_{k,n}$ for $n = 1 \ldots N$.

Finally, we note that the SAM-SPARSE-C method defined in a previous work [9] is strictly equivalent to a modification of $\hat{\overline{D}}_{k,f}$ where all $A_k^{(p)}(f)$ parameters would be replaced with 1.

2.4 Physical Interpretation

In this section we show that the use of the proposed PDM, as described by Eq. 6, is closely related to the average delay-sum power over all points in a sector (weighted by the prior distribution). We show that for a given triplet (sector, frequency bin, pair of microphones), if we neglect the energy difference between microphones, there is equivalence with the delay-sum power, *averaged* over all points in the sector.

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

$$X_{i_p}(f) \stackrel{\text{def}}{=} \alpha_1^{(p)}(f)\, e^{j\beta_1^{(p)}(f)} \qquad \text{and} \qquad X_{j_p}(f) \stackrel{\text{def}}{=} \alpha_2^{(p)}(f)\, e^{j\beta_2^{(p)}(f)}, \tag{22}$$

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 $\hat{\theta}^{(p)}(f) \equiv \beta_1^{(p)}(f) - \beta_2^{(p)}(f)$, where the $\equiv$ symbol denotes congruence of angles (equality modulo 2π).

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

$$E_{\text{ds}}^{(p)}(\mathbf{v}, f) \stackrel{\text{def}}{=} \left| X_{i_p}(f) + X_{j_p}(f)\, e^{j\gamma^{(p)}(\mathbf{v},f)} \right|^2 . \tag{23}$$

Assuming the received magnitudes to be the same $\alpha_{i_p} \approx \alpha_{j_p} \approx \alpha$, Eq. 23 can be rewritten:

$$\begin{aligned} E_{\text{ds}}^{(p)}(\mathbf{v}, f) &= \left| \alpha e^{j\beta_1^{(p)}(f)} \left(1 + e^{j\left(-\hat{\theta}^{(p)}(f) + \gamma^{(p)}(\mathbf{v},f)\right)} \right) \right|^2 \\ &= \alpha^2 \left[\left(1 + \cos\left(-\hat{\theta}^{(p)}(f) + \gamma^{(p)}(\mathbf{v}, f) \right) \right)^2 + \sin^2\left(-\hat{\theta}^{(p)}(f) + \gamma^{(p)}(\mathbf{v}, f) \right) \right] \\ &= \alpha^2 \left[2 + 2\cos\left(-\hat{\theta}^{(p)}(f) + \gamma^{(p)}(\mathbf{v}, f) \right) \right] \end{aligned} \tag{24}$$

On the other hand, the square distance between observed phase and theoretical phase, as defined by Eq. 3, is expressed as:

$$d^2\left(\hat{\theta}^{(p)}(f), \gamma^{(p)}(\mathbf{v}, f) \right) = \sin^2\left(\frac{\hat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}, f)}{2} \right), \tag{25}$$

which can be rewritten as:

$$d^2\left(\hat{\theta}^{(p)}(f), \gamma^{(p)}(\mathbf{v}, f) \right) = \frac{1}{2} \left(1 - \cos\left(\hat{\theta}^{(p)}(f) - \gamma^{(p)}(\mathbf{v}, f) \right) \right). \tag{26}$$

From Eqs. 24 and 26:

$$\boxed{\frac{1}{4\alpha^2} E_{\text{ds}}^{(p)}(\mathbf{v}, f) = 1 - d^2\left(\hat{\theta}^{(p)}(f), \gamma^{(p)}(\mathbf{v}, f) \right)} \tag{27}$$

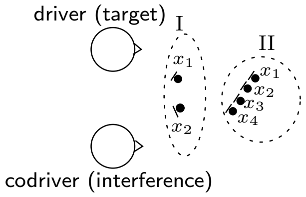


Figure 4: Physical setups I (2 mics) and II (4 mics).

Therefore we can see that 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 Eq. 6. Averaging across microphone pairs, as in Eq. 3, exploits the redundancy of the signals in order to deal with noisy measurements and get around spatial aliasing effects.

This equivalence explains the difference between the proposed approach - equivalent to an average delay-sum over a sector - and a classical approach that would compute the delay-sum only at a point in the middle of the sector. It is intuitively more sound to give equal importance to all points in a sector given that the task is sector-based detection. Moreover, tests on more than one hour of real meeting room data have confirmed the advantage of the proposed approach [9]. The computational cost is the same, as shown in Section 2.3.

We note that the assumption $\alpha_{i_p} \approx \alpha_{j_p}$ is reasonable for most setups where microphones are close to each other, and oriented to the same direction if directional microphones. Nevertheless, in practice we found that 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

The two setups are denoted I and II, and are depicted by Fig. 4:

- 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.

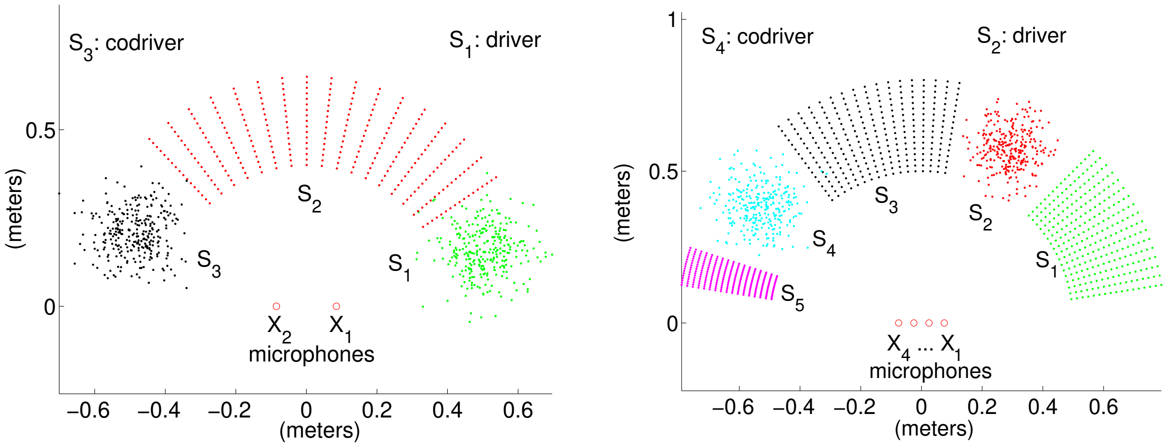
3.2 Recordings

Data was not simulated, we opted for real data instead. Three 10-second 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` **:** (training) Mannequins playing pre-recorded speech. It is used as a training data to select parameter values.

`test` **:** (testing) Real human speakers. It is used for testing only: all parameters determined on `train` were "frozen".

`noise` **:** (testing) Both persons silent, the car running at 100 km/h.



(a) Setup I

(b) Setup II

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

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 Signal-to-Interference Ratio (SIR) at microphone x_1. This will be useful to evaluate the input SIR estimation technique presented in Section 4, as well as the SIR improvement in Section 5.

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

3.3 Sector Definition

Figs. 5a and 5b depict the way we defined sectors for each setup. We used prior knowledge on 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 reasoning behind the other sectors having uniform distributions is that reverberations may come from any of those directions.

4 Input SIR Estimation

This section describes a method to estimate the *input* Signal-to-Interference Ratio (SIR), which in our case is the ratio between the energy of the speech received in $x_1(t)$ from the driver, and the energy of the speech received in $x_1(t)$ from the codriver. This estimate of the input SIR is used by the "explicit" adaptation control method described in Section 5.2.

4.1 Method

From a given frame of samples at microphone 1:

$$\mathbf{x}_1(t) = [x_1(t - N_{\text{samples}}),\ x_1(t - N_{\text{samples}} + 1),\ \ldots,\ x_1(t)]^T\ , \tag{28}$$

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

In order to estimate the input SIR at $x_1(t)$, 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:

$$\hat{\text{SIR}}_1 = \frac{\sum_f E_1(f) \cdot P\left(\text{sector } S_{\text{driver}} \text{ active at frequency } f \;\middle|\; \hat{\mathbf{\Theta}}(f)\right)}{\sum_f E_1(f) \cdot P\left(\text{sector } S_{\text{codriver}} \text{ active at frequency } f \;\middle|\; \hat{\mathbf{\Theta}}(f)\right)}, \tag{29}$$

and

$$\hat{\text{SIR}}_2 = \frac{\sum_f P\left(\text{sector } S_{\text{driver}} \text{ active at frequency } f \;\middle|\; \hat{\mathbf{\Theta}}(f)\right)}{\sum_f P\left(\text{sector } S_{\text{codriver}} \text{ active at frequency } f \;\middle|\; \hat{\mathbf{\Theta}}(f)\right)}, \tag{30}$$

where $P(.|\mathbf{\Theta}(f))$ is the posterior probability given by Eqs. 9 and 10. This ratio can be seen as the ratio between two mathematical expectations. $\hat{\text{SIR}}_1$ weights each frequency with its energy, while $\hat{\text{SIR}}_2$ 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 $\hat{\text{SIR}}_1$ gives more weights to the low frequencies, while $\hat{\text{SIR}}_2$ gives equal weights to low and high frequencies. It can be expected that $\hat{\text{SIR}}_2$ provides better results as long as microphones are close enough to make proper use of high frequencies.

4.2 Experiments

On the entire recording `train` , we ran the source detection algorithm described in Section 2 and compared the estimates $\hat{\text{SIR}}_1$ or $\hat{\text{SIR}}_2$ with the true input SIR.

First, we noted that an additional affine scaling in log domain (1st order polynomial) was needed. A possible interpretation is that it compensates for the simplicity of the function chosen for probability estimation (Eq. 9). This affine scaling is the only post-processing that we used, temporal filtering (smoothing) was not used.

For each setup and each method, we tuned the parameters (λ and the two parameters of the affine scaling) on `train` in order to minimize an objective criterion: the RMS error of input SIR estimation, in log domain (dB). Results are reported in Table 1. In all cases a RMS error of about 10 dB is obtained, and soft decision ($\lambda > 0$) is beneficial. In setup I, $\hat{\text{SIR}}_1$ gives the best results. In setup II, $\hat{\text{SIR}}_2$ gives the best results. This confirms the above-mentioned expectation that $\hat{\text{SIR}}_2$ yields better results when microphones are close enough. We note that 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: Figs. 6a and 6b. We can see that 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,

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

Table 1: RMS error of input SIR estimation on `train` , calculated in log domain (dB). The best result for each setup is in bold figures. 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.

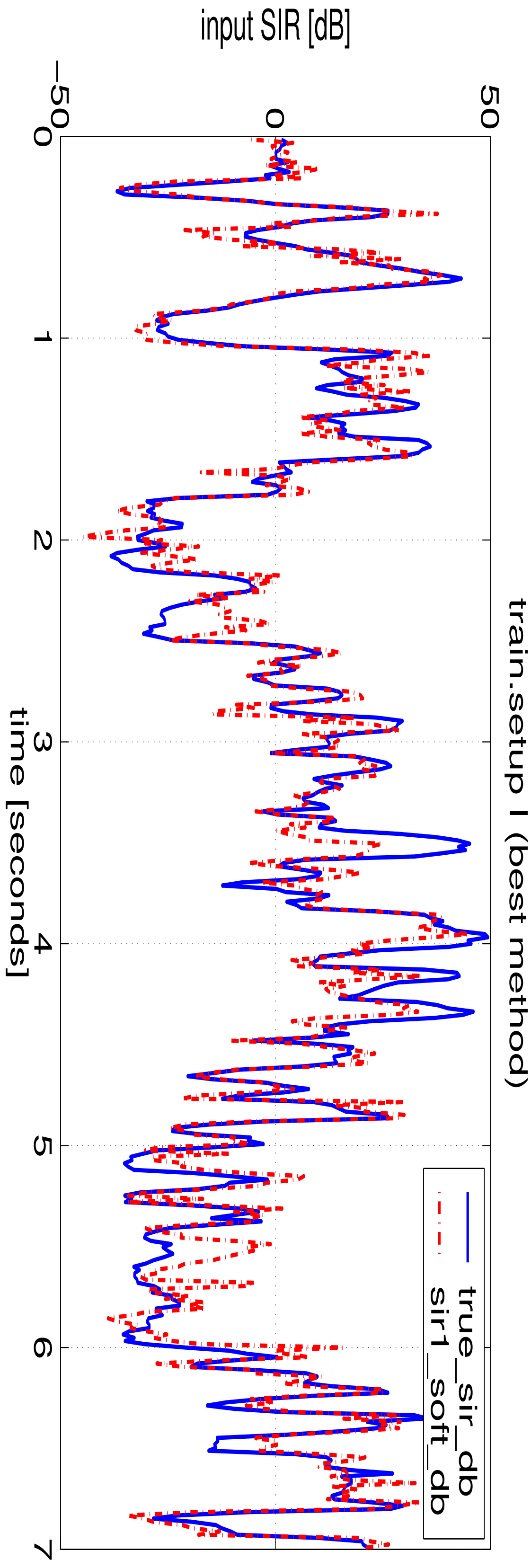
Setup	Dynamic range	Method	Result
I	71.6 dB	$\hat{\mathrm{SIR}}_1$, soft	All frames: 14.0% (0.77)
			True input SIR > 6 dB: 16.1% (0.25)
			True input SIR < -6 dB: 12.4% (0.71)
II	70.2 dB	$\hat{\mathrm{SIR}}_2$, soft	All frames: 9.3% (0.90)

Table 2: RMS error of input SIR estimation on `test` , calculated in log domain (dB). Methods and parameters were selected on `train` . 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.

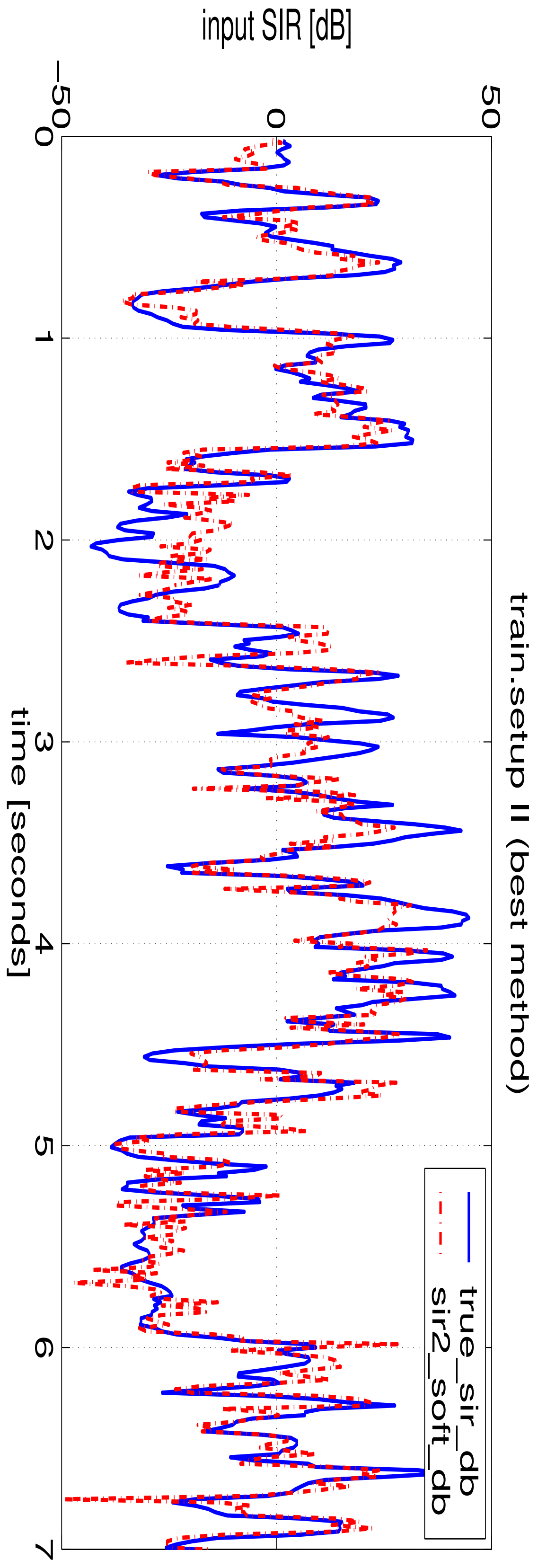
which is not exploited by the sector-based detection algorithm. Also the sector-based detection gives equal role to all microphones, while we are mostly interested in $x_1(t)$. In spite of these limitations, we can safely state that the obtained SIR curve is very satisfying for triggering the adaptation, as verified in Section 5.

As it is not sufficient to evaluate results on the same data that was used to tune the parameters, results on the `test` recording are also reported in Tab. 2 and Figs. 7a nd 7b. 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. A possible interpretation is that the method does not take into account the low coherence between the two directional microphones, due to their very different orientations. However, in an interference cancellation application with Setup I, we are mostly interested in accurate detection of periods of negative input SIR, rather than positive input SIR. On those periods the RMS error is lower (12.4%). We will therefore see in Section 5 that this result can still be used in a speech enhancement application. For Setup II, the results are quite similar to those of `train` .

To conclude, the proposed methodology for estimation of input SIR gives acceptable results. Despite a RMS error of about 10 dB, the estimated input SIR curve follows the true curve accurately enough for detecting 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 has no impact on results presented in Section 5. This method is particularly robust since it does not involve any temporal integration or thresholding.

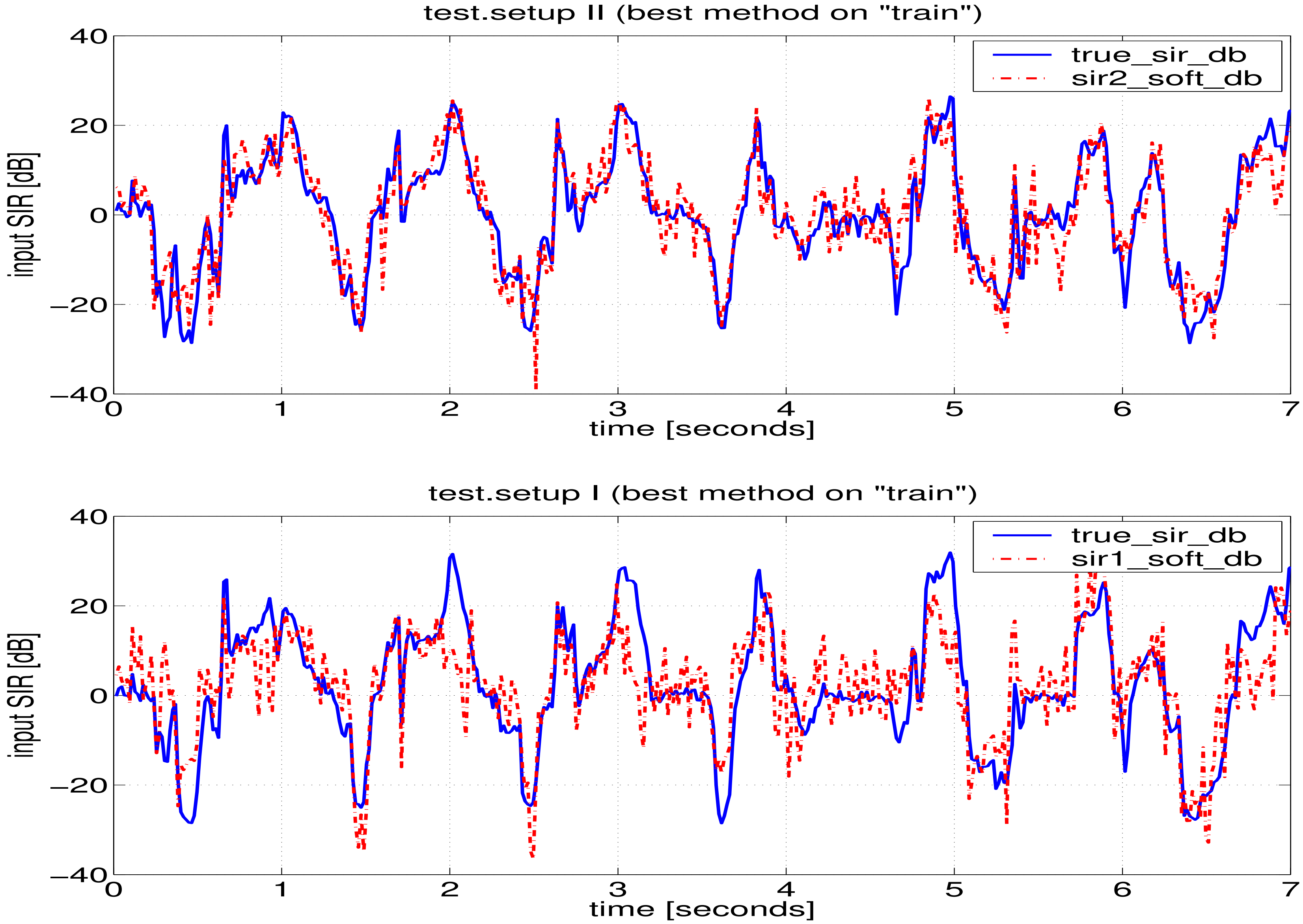


(a) Setup I



(b) Setup II

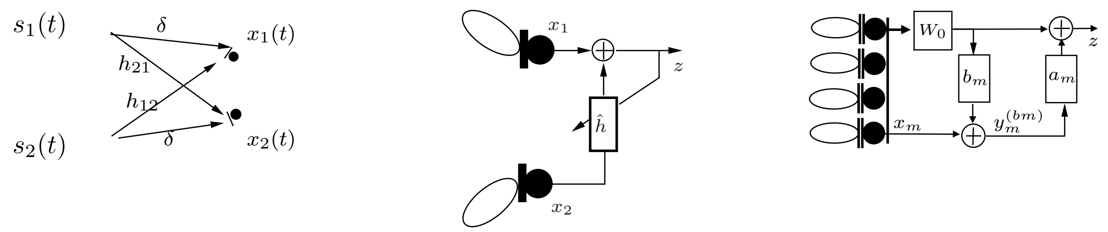
Figure 6: Estimation of the input SIR for setups I and II (beginning of recording `train`).



(b) Setup II

(a) Setup I

Figure 7: Estimation of the input SIR for setups I and II (beginning of recording `test`).



(a) Setup I: mixing channels. (b) Setup I: noise canceller. (c) Setup II: GSC.

Figure 8: Linear models for the acoustic channels and the adaptive filtering.

5 Speech Enhancement

5.1 Adaptive Interference Cancellation Algorithms

Setup I provides an input Signal-to-Interference Ratio (SIR) of about 6 dB at the driver's microphone signal $x_1(t)$. The other signal $x_2(t)$ is used as a reference, i.e. an estimate of the interference signal. In order to remove the interference from $x_1(t)$, the linear filter depicted by Fig. 8b is used. The filter $\hat{\mathbf{h}}$ of length L is adapted to minimize the output power $\mathbf{E}\{z^2(t)\}$, using the NLMS algorithm [15] with step size μ:

$$\boxed{\hat{\mathbf{h}}(t+1) = \hat{\mathbf{h}}(t) - \mu \frac{\mathbf{E}\{z(t)\mathbf{x}_2(t)\}}{\|\mathbf{x}_2(t)\|^2}} \tag{31}$$

where $\mathbf{x}_2(t) = [x_2(t),\ x_2(t-1),\ \ldots,\ x_2(t-L+1)]^T$, $\hat{\mathbf{h}} = \left[\hat{h}_0(t),\ \hat{h}_1(t),\ \ldots,\ \hat{h}_{L-1}(t)\right]^T$, and $\|\mathbf{x}\|^2 = \sum_{i=1}^{L} x^2(i)$. Note: $\mathbf{E}\{.\}$ is the expectation operator, taken over realizations of stochastic processes. Under stationarity and ergodicity assumptions, we implement it by averaging on a short time-frame. For example:

$$\mathbf{E}\{x^2\,(t)\} \approx \frac{1}{L}\|\mathbf{x}\|^2 \tag{32}$$

To prevent instability, adaptation 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, the filter $\hat{\mathbf{h}}$ of length L must be adapted *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 [16] and its robust GSC implementation [17]. 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 Fig. 8c. 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 Eq. 31. 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.

With respect to existing implicit approaches, the novelty of the implicit method proposed here is a well-grounded mechanism to increase its robustness by preventing instability.

A major novelty resides in the explicit method, in so far as it estimates the input SIR directly from the captured signals, using the method described in Section 4.

5.2.1 Implicit method

We present the method in details for Setup I, then briefly give the corresponding results for Setup II. The goal is to increase the adaptation step-size whenever possible, while not turning Eq. 31 into an unstable, divergent process.

For Setup I, as depicted by Fig. 8a, the acoustic mixing channels are modelled as:

$$\begin{cases} x_1(t) = s_1(t) + h_{12}(t) * s_2(t), \\ x_2(t) = h_{21}(t) * s_1(t) + s_2(t), \end{cases} \tag{33}$$

where $*$ denotes the convolution operator.

As depicted by Fig. 8b, the enhanced signal is $z(t) = x_1(t) + \hat{h}(t) * x_2(t)$ therefore:

$$\begin{aligned} z(t) &= \underbrace{\Big(\delta(t) + \hat{h}(t) * h_{21}(t)\Big)} * s_1(t) + \underbrace{\Big(h_{12}(t) + \hat{h}(t)\Big)} * s_2(t) \\ &= \quad\Omega(t) \quad * s_1(t) + \quad \Pi(t) \quad * s_2(t) \end{aligned} \tag{34}$$

The goal is to minimize $\mathbf{E}\{\varepsilon^2(t)\}$ where $\varepsilon(t) = \Pi(t) * s_2(t)$. It can be shown [18] 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:

$$\mu_{\text{impl}}(t) = \mathbf{E}\{\Pi^2(t)\}\frac{\mathbf{E}\{x_2^2(t)\}}{\mathbf{E}\{z^2(t)\}} = \mathbf{E}\{\Pi^2(t)\}\frac{\|\mathbf{x_2}\|^2}{\|\mathbf{z}\|^2}. \tag{35}$$

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:

$$\mu_{\text{impl}}(t) \approx \mu_0 \frac{\|\mathbf{x_2}\|^2}{\|\mathbf{z}\|^2}, \tag{36}$$

and Eq. 31 becomes:

$$\hat{\mathbf{h}}(t+1) = \hat{\mathbf{h}}(t) - \mu_0 \frac{\mathbf{E}\{z(t)\mathbf{x}_2(t)\}}{\|\mathbf{z}(t)\|^2}. \tag{37}$$

The domain of stability of the NLMS algorithm [15] is defined by $\mu_{\text{impl}}(t) < 2$ therefore Eq. 37 can only be applied when $\mu_0 \frac{\|\mathbf{x_2}\|^2}{\|\mathbf{z}\|^2} < 2$. In other cases, a fixed step-size adaptation must be used, as in Eq. 31. The proposed implicit adaptive step-size is therefore:

$$\boxed{\mu(t) = \begin{cases} \mu_{\text{impl}}(t) & \text{if} \quad \mu_{\text{impl}}(t) < 2 & \text{(stable case)} \\ \mu_0 & \text{otherwise} & \text{(unstable case).} \end{cases}} \tag{38}$$

where μ_0 is a very small positive constant ($0 < \mu_0 \ll 1$). 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 Eqs. 33 and 34, we can expand Eq. 36 into:

$$\mu_{\text{impl}}(t) = \mu_0 \frac{\|h_{21}\|^2\sigma_1^2 + \sigma_2^2}{\|\mathbf{\Omega}(t)\|^2\sigma_1^2 + \|\mathbf{\Pi}(t)\|^2\sigma_2^2}. \tag{39}$$

In a car, the driver is closer to x_1 than to x_2. Thus, given the definition of the mixing channels depicted by Fig. 8a, 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 Eq. 39 yields:

$$\mu_{\text{impl}}(t) \;<\; \mu_0 \,\frac{2\;\|h_{21}\|^2\sigma_1^2}{\|\mathbf{\Omega}(t)\|^2\sigma_1^2+\|\mathbf{\Pi}(t)\|^2\sigma_2^2} \;<\; 2\;\mu_0\frac{\|h_{21}\|^2}{\|\mathbf{\Omega}(t)\|^2} \;<\; 2, \tag{40}$$

which falls in the "stable case" of Eq. 38.

• *Case 2:* The power received at microphone 2, from the target, is *less* than the power received from the interference: $\|h_{21}\|^2\sigma_1^2 \le \sigma_2^2$. In this case Eq. 39 yields:

$$\mu_{\text{impl}}(t) \;\le\; \mu_0 \,\frac{2\;\sigma_2^2}{\|\mathbf{\Omega}(t)\|^2\sigma_1^2+\|\mathbf{\Pi}(t)\|^2\sigma_2^2}, \tag{41}$$

therefore:

$$\|\mathbf{\Omega}(t)\|^2\frac{\sigma_1^2}{\sigma_2^2}+\|\mathbf{\Pi}(t)\|^2 \;\le\; 2\,\frac{\mu_0}{\mu_{\text{impl}}(t)}. \tag{42}$$

Thus, in the "unstable case" of Eq. 38, we have:

$$\begin{cases} \|\mathbf{\Pi}(t)\|^2 \le \mu_0 \\ \frac{\sigma_1^2}{\sigma_2^2} \le \frac{\mu_0}{\|\mathbf{\Omega}(t)\|^2} \le \mu_0 \end{cases} \tag{43}$$

The first line of Eq. 43 means that the adaptation is close to convergence. The second line of Eq. 43 means that the input SIR is very close to zero, i.e. the interference is largely dominant. Overall, this is the only "unstable case", i.e. when we fall back on $\mu_{\text{impl}}(t) = \mu_0$ (Eq. 38).

5.2.2 Explicit method

For both setups, a sector-based audio source detection method can be used to directly estimate the input SIR at $x_1(t)$, as detailed in Section 4. Two thresholds are set on the input SIR to detect when the target (resp. the interference) is dominant. This decision determines whether or not the fixed step-size adaptation of Eq. 31 should be applied.

5.3 Implementation

In Setup I, the $\hat{h}$ 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 $\hat{h}$ is initialized to zeros. In setup II, filters b_m are initialized to cancel signals coming from driver's direction of arrival [19], and the filters a_m are initialized to zeros.

Adaptation is implemented as follows:

- **No control:** A baseline method that adapts all the time, with a constant step size, as in Eq. 31. In setup II, filters a_m are adapted all the time, and filters b_m are not adapted.

- **Implicit method:** In both setups, all filters are adapted all the time, with the adaptive step-size of Eq. 38. 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).

- **Explicit method:** All filters are adapted with Eq. 31. In setup I, filter $\hat{h}$ 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.

5.4 Performance Evaluation

For both setups, we measured the instantaneous SIR improvement over the true input SIR at microphone x_1, on the real recordings described in Section 3.2. "Instantaneous" means on half-overlapping, short time-frames – i.e. where speech can be safely considered as stationary. We used 32 ms-long time-frames.

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).

The instantaneous SIR improvement is plotted over time in log-domain (dB), to directly compare the behaviour of the various methods, depending on the speech signals of each person.

Based on the instantaneous SIR improvement, the segmental SIR is computed in three cases. "Segmental SIR" means that only frames containing speech from either driver or codriver or both are considered.

1. True input SIR $<$ -6 dB: when the energy of the codriver 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.

2. 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.

3. 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 problems. We can expect the baseline method to yield a very negative value here, because of leakage issues.

Note: determining whether a given person is active or not in its individual signal (see Section 3.2) is done by fitting a bi-Gaussian model on the energy in log domain, using the (unsupervised) EM algorithm [20]. 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). This way, we avoid introducing bias into the performance evaluation, as could be the case for example by setting a manual threshold on energy, for each signal.

5.5 Experiments: clean data

The first 3 seconds are depicted in Fig. 9a. The periods where SIR improvement is consistently close to 0 dB correspond to silences of both speakers. The result of the "no control" baseline method highlights the target cancellation problem and confirms the necessity of adaptation control. Both "implicit" and "explicit" methods are robust against this problem, and the explicit method provides the best results.

This analysis is valid for both setups, and is confirmed by average SIR improvement values over the entire recording shown in Tab. 3. All expectations given in 5.4 are verified. 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). Again, best average results are yielded by the explicit method.

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 thresholds were used for detecting driver or codriver activity. Only one parameter was changed: the adaptation step μ_0 was lowered to take into account

	Setup I (2 mics)			Setup II (4 mics)		
Range of the true input SIR	No control (baseline)	Implicit	Explicit	No control	Implicit	Explicit
$<$ -6: (codriver)	6.5	5.9	10.7	10.0	6.0	10.1
$[-6, +6]$: (both)	-0.6	1.2	5.8	1.2	3.6	4.2
$>$ +6: (driver)	-7.7	-0.2	2.6	-8.3	2.2	1.3

Table 3: Average segmental SIR improvement in dB on `test` (clean data).

	Setup I (2 mics)			Setup II (4 mics)		
Range of the true input SIR	No control (baseline)	Implicit	Explicit	No control	Implicit	Explicit
$<$ -6: (codriver)	6.4	7.1	7.4	7.5	3.9	9.9
$[-6, +6]$: (both)	1.0	2.7	3.5	2.4	2.9	4.2
$>$ +6: (driver)	-4.7	0.4	1.9	-4.5	2.4	-0.4

Table 4: Average segmental SIR improvement in dB on `test` + `noise` .

the lower quality of the incoming signal due to noise. The goal of this experiment is to determine whether the proposed approach can cope with background noise. We note that it is not a given, since the proposed systems (implicit and explicit) do not explicitly model background noise: the noise source may be incoherent, or localized outside of the defined sectors. The hope is that reducing the adaptation step is enough, while keeping the exact same system w.r.t. all other parameters.

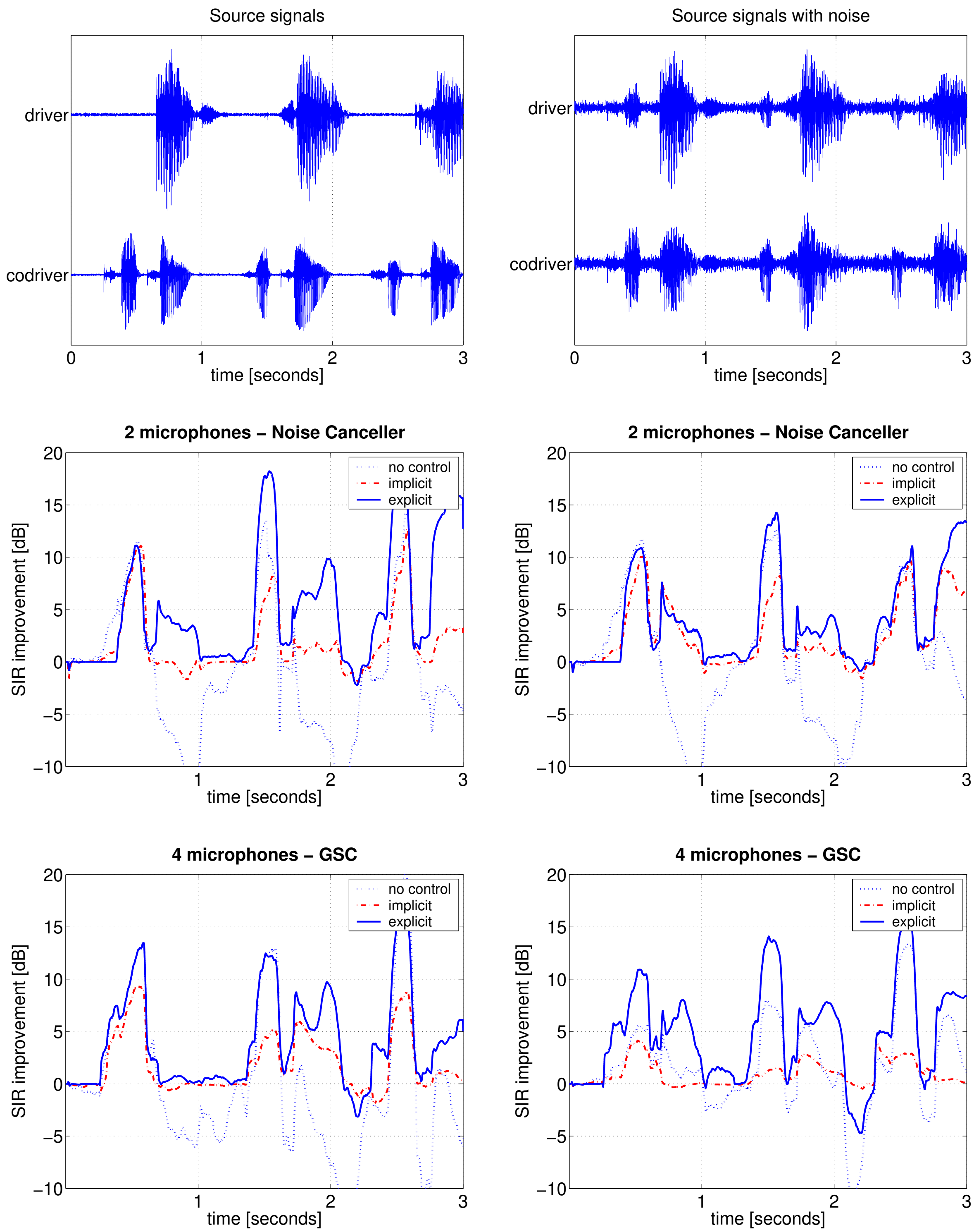
The result is given in Fig. 9b and Tab. 4. We can see that the behaviour in terms of SIR improvement, both over time and in average, is very similar to the clean case. Thus, we can state that the system also works in a realistic case of a moving car.

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. 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.

7 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 Multi-modal Information Management (IM)2. The authors would like to thank Dr Iain McCowan, Mathew Magimai.-Doss and Bertrand Mésot for helpful comments and suggestions.



(a) Clean conditions.

(b) Noisy conditions.

Figure 9: Improvement over input SIR (100 ms moving average, first 3 seconds shown). Column (a) shows results on clean data (`test`), whereas column (b) shows results on noisy data (`test` + `noise` : 100km/h background road noise).

Annex 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 $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \left(\mathbb{R}^P\right)^3$:

$$g(\mathbf{x}, \mathbf{y}) \geq 0 \tag{44}$$

$$g(\mathbf{x}, \mathbf{y}) = g(\mathbf{y}, \mathbf{x}) \tag{45}$$

$$g(\mathbf{x}, \mathbf{y}) = 0 \quad \text{iff} \quad \forall p = 1 \dots P \quad \exists k_p \in \mathbb{Z} \quad x_p = y_p + k2\pi \tag{46}$$

$$g(\mathbf{x}, \mathbf{z}) \leq g(\mathbf{x}, \mathbf{y}) + g(\mathbf{y}, \mathbf{z}) \tag{47}$$

It is basically the same as a classical metric, except for Eq. 46 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 $\mathbb{R}^P$:

$$G_P\left(\mathbf{x}, \mathbf{y}\right) \stackrel{\text{def}}{=} \sqrt{\frac{1}{P} \sum_{p=1}^{P} G_1\left(x_p, y_p\right)^2} \tag{48}$$

The rest of this Section shows that all G_P functions are also PDMs. Eqs. 44, 45, 46 are trivial to demonstrate. Eq. 47 is demonstrated for G_P in the following.

Since G_1 is a PDM, it verifies Eq. 47 on $\mathbb{R}$. Therefore, for any $(\mathbf{x}, \mathbf{y}, \mathbf{z}) \in \left(\mathbb{R}^P\right)^3$:

$$G_P\left(\mathbf{x}, \mathbf{z}\right) \leq \sqrt{\frac{1}{P} \sum_{p=1}^{P} \left[G_1\left(x_p, y_p\right) + G_1\left(y_p, z_p\right)\right]^2} \tag{49}$$

Now let us recall the Minkowski inequality [20]. For any $\beta > 1$ and $a_p > 0$, $b_p > 0$:

$$\left[\sum_{p=1}^{P} \left(a_p + b_p\right)^\beta\right]^{\frac{1}{\beta}} \leq \left[\sum_{p=1}^{P} {a_p}^\beta\right]^{\frac{1}{\beta}} + \left[\sum_{p=1}^{P} {b_p}^\beta\right]^{\frac{1}{\beta}} \tag{50}$$

By applying the Minkowski inequality to the right-hand of Eq. 49, 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:

$$G_P\left(\mathbf{x}, \mathbf{z}\right) \leq \sqrt{\tfrac{1}{P} \textstyle\sum_{p=1}^{P} G_1\left(x_p, y_p\right)^2} + \sqrt{\tfrac{1}{P} \textstyle\sum_{p=1}^{P} G_1\left(y_p, z_p\right)^2} \tag{51}$$

$$G_P\left(\mathbf{x}, \mathbf{z}\right) \leq G_P\left(\mathbf{x}, \mathbf{y}\right) + G_P\left(\mathbf{y}, \mathbf{z}\right) \tag{52}$$

References

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

[2] B. D. V. Veen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Signal Processing Magazine*, vol. 5, pp. 4–24, 1988.

[3] D. Van-Compernolle, "Switching adaptive filters for enhancing noisy and reverberant speech from microphone array recordings," in *Proceedings of the 1990 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-90)*, 1990.

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

[5] S. Gannot, D. Burshtein, and E. Weinstein, "Beamforming methods using nonstationarity with application to speech processing," *IEEE Transactions on Signal Processing*, vol. 49, no. 8, pp. 1614–1626, August 2001.

[6] O. Hoshuyama and A. Sugiyama, "A robust adaptive beamformer for microphone arrays with a blocking matrix using constrained adaptive filters," in *Proceedings of ICASSP '96*, vol. 2, May 1996, pp. 925–928.

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

[8] W. Herbordt, T. Trini, and W. Kellermann, "Robust spatial estimation of the signal-to-interference ratio for non-stationary mixtures," in *Proceedings of the Int. Conf. on Acoustic Echo and Noise Control*, 2003.

[9] G. Lathoud and M. Magimai.-Doss, "A Sector-Based, Frequency-Domain Approach to Detection and Localization of Multiple Speakers," in *Proceedings of the 2005 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-05)*, 2005.

[10] G. Lathoud, J. Bourgeois, and J. Freudenberger, "Multichannel Speech Enhancement in Cars: Implicit vs. Explicit Control," in *Proceedings of the 2005 Joint Workshop on Hands-Free Speech Communication and Microphone Arrays (HSCMA-05)*, 2005.

[11] S. Roweis, "Factorial Models and Refiltering for Speech Separation and Denoising," in *Proc. Eurospeech*, 2003.

[12] G. Lathoud and I. McCowan, "A Sector-Based Approach for Localization of Multiple Speakers with Microphone Arrays," in *Proc. SAPA 2004*, Oct. 2004.

[13] C. Knapp and G. Carter, "The Generalized Correlation Method for Estimation of Time Delay," *IEEE Transactions on Acoustic, Speech and Signal Processing*, vol. ASSP-24, no. 4, pp. 320–327, August 1976.

[14] B. C. J. Moore, *An Introduction to the Psychology of Hearing, fourth edition.* Academic Press, 1997.

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

[16] L. Griffiths and C. Jim, “An Alternative Approach to Linearly Constrained Adaptive Beamforming,” *IEEE Transactions on Antennas and Propagations*, vol. 30, no. 1, pp. 27–34, January 1982.

[17] O. Hoshuyama and A. Sugiyama, “A Robust Adaptive Beamformer for Microphone Arrays with a Blocking Matrix using Constrained Adaptive Filters,” in *Proceedings of the IEEE International Conferance on Acoustics, Speech and Signal Processing, Atlanta, GA*, 1996.

[18] A. Mader, H. Puder, and G. Schmidt, “Step-Size Control for Acousic Echo Cancellation Filters - An Overview,” *Signal Processing*, vol. 80, no. 9, pp. 1697–1719, 2000.

[19] L. Griffiths and C. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. on Antennas and Propagation*, vol. 30(1), pp. 27–34, January 1982.

[20] T. K. Moon and W. C. Stirling, *Mathematical Methods and Algorithms for Signal Processing*. Prentice Hall, 2000.