CORE
Metadata, citation and similar papers at core.ac.uk
Provided by INRIA a CCSD electronic archive server

HAL
archives-ouvertes.fr

Gaussian framework for interference reduction in live recordings

Diego Di Carlo, Ken Déguernel, Antoine Liutkus

To cite this version:

Diego Di Carlo, Ken Déguernel, Antoine Liutkus. Gaussian framework for interference reduction in live recordings. AES International Conference on Semantic Audio, Jun 2017, Erlangen, Germany. hal-01515971

HAL Id: hal-01515971
https://hal.inria.fr/hal-01515971

Submitted on 28 Apr 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Audio Engineering Society

Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

Gaussian framework for interference reduction in live recordings

Diego Di Carlo[1], Ken Déguernel[1,2], and Antoine Liutkus[1]

[1]*Inria, Multispeech team, Villers-les-Nancy, F-54600, France*
[2]*IRCAM STMS Lab (CNRS, UPMC, Sorbonne Universités), Paris, France*

Correspondence should be addressed to Diego Di Carlo (diego.dicarlo89@gmail.com)

ABSTRACT

In live multitrack recordings, each voice is usually captured by dedicated close microphones. Unfortunately, it is also captured in practice by other microphones intended for other sources, leading to so-called "interferences". Reducing this interference is desirable because it opens new perspectives for the engineering of live recordings. Hence, it has been the topic of recent research in audio processing. In this paper, we show how a Gaussian probabilistic framework may be set up for obtaining good isolation of the target sources. Doing so, we extend several state-of-the art methods by fixing some heuristic parts of their algorithms. As we show in a perceptual evaluation on real-world multitrack live recordings, the resulting principled techniques yield improved quality.

1 Introduction

In typical studio conditions, instrumental voices are often recorded simultaneously because this promotes spontaneity and musical interaction between the musicians, but also because it optimizes studio time usage. For live musical performances, each musician from a band gets its dedicated microphones, so that the different voices may be optimized independently and on-demand by sound engineers.

In all these situations, having clean isolated recordings for all instrumental voices is desirable because it allows much flexibility for further processing, remixing and exploitation. However, it is inevitable that *interferences* will occur, so that some voices are captured by microphones intended to other voices. This classical fact is also called *leakage* or *bleeding* by sound engineers, who have a strong expertise in designing specific acoustic setups to minimize them. However, unless the musicians do not play in the same room, which is detrimental to musical spontaneity, interferences are bound to occur in practice.

In the last 10 years, research has been conducted on the topic of interference reduction [1, 2, 3, 4]. Its goal is to propose signal processing algorithms that may be used by sound engineers to reduce the amount of leakage in live multitrack recordings. Most of the time, these methods are applicable a posteriori and require important computing resources. However, some studies have focused on real-time alternatives for ad hoc situations [5] leading to the development of some dedicated commercial products[1]. We shortly review this line of research now.

[1]See, e.g http://accusonus.com/products/drumatom.

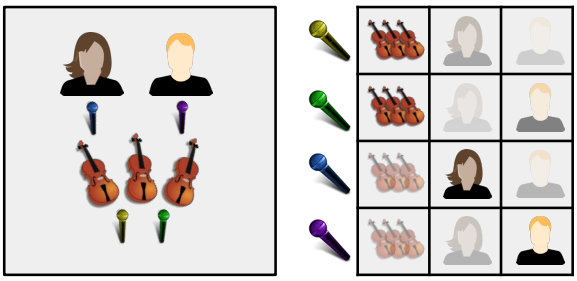

Fig. 1: Illustration of typical interferences found in multitrack live recordings. In the setup considered here: violin section, male singer, female singer, each voice gets its own dedicated microphones. However, the resulting signals all get leakage from all voices. The amount of interference is quantified in our model by the interference matrix, as proposed in [7] (courtesy of R. Bittner).

Although early research in interference removal has been focused in exploiting inter-microphone phase dependencies [1], the breakthrough brought in by [2, 4] made it clear that neglecting these dependencies and rather concentrating on energy redundancies over channels brings robustness and computational effectiveness. After identifying the Power Spectral Densities (PSD) of the sources, a simple Wiener filter is applied in each channel to recover the desired signals [6]. Therefore, the main challenge these methods face is the estimation of the PSD of the sources to achieve good performance [4]. Their main working hypothesis is that the close-microphones for a given voice already present good isolation properties and may be used as the PSD to use for Wiener filtering. This idea can be further improved by enforcing some prior information about what each voice should sound like in terms of spectral characteristics. This led to products specialized in the reduction of interferences for drum signals [5], as well as to recent developments able to concentrate on orchestral leakage reduction [7, 8].

While early methods based on Wiener filter are straightforward to implement [2], they suffer from one important drawback: the voice models are initialized using their close-mic recordings and are assumed to have the same energy within all tracks. Extending the weighting coefficients model [4] as a way to quantify how much each voice is present in each track, Prätzlich in [7] introduce the *interference matrix*. The concept is illustrated in figure 1. While [7] then concentrates on a grounded way to learn this interference matrix automatically, the spectral models are updated in a somewhat ad hoc fashion, leading to clear sub-optimality of the estimation algorithm.

In this study, we show how a rigorous probabilistic Gaussian framework [6, 9, 10] may be used to yield provably optimal algorithms to learn all the parameters required for good interference reduction. We present and detail four alternative algorithms to this end and provide an open-source Python implementation. The discussed methods are compared with state of the art in a perceptual study led on real legacy multitrack recordings from the Montreux Jazz Festival[2], one of the most important musical events in Europe for more than 50 years.

2 Model and Methods

2.1 Notation and probabilistic model

First, we detail our notations for referring to the signals. Let J be the number of voices and I be the number of microphones. For $i = 1 \dots I$, x_i is the signal recorded by the ith microphone, called a *mixture*. In full generality and because of interferences, this i^{th} mixture captures sound from all the voices. Hence, for $j = 1 \dots J$, we define the *image* y_{ij} as the contribution of voice j in mixture i, so that we have $x_i = \sum_{j=1}^{J} y_{ij}$.
Let $X_i(f,t)$ be the STFT of mixture x_i and similarly for Y_{ij} with y_{ij}. They are all complex matrices of dimension $F \times T$, where F is the number of frequency bands and T the number of frames. We have:

$$X_i(f,t) = \sum_{j=1}^{J} Y_{ij}(f,t). \tag{1}$$

An entry (f,t) of any such matrix is referred to as a Time-Frequency (TF) bin. Now, let finally the power *spectrogram* of x_i be the $F \times T$ matrix V_i with nonnegative entries defined as:

$$V_i(f,t) \triangleq |X_i(f,t)|^2. \tag{2}$$

where $\triangleq$ denotes a definition. The goal of interference reduction is to compute an estimate $\hat{Y}_{ij}$ of the images Y_{ij}, for all i and j.

[2] www.montreuxjazzfestival.com

Second, we now briefly present our probabilistic model. To begin with, we assume that the signals originating from different voices $j = 1 \dots J$ are independent. Then, for each voice j, we assume that its contributions Y_{ij} in the different mixtures i are independent. This means we do not take the phase dependencies between the different channels into account. That arguable assumption proves important in practice for both robustness to real-world scenarios and computational complexity. Finally, for a given Y_{ij}, we model it through the Local Gaussian Model (LGM, [11, 9]), a popular model accounting for the local stationarity of audio. All the entries of Y_{ij} are taken independent and distributed with respect to a complex isotropic Gaussian distribution:

$$Y_{ij}(f,t) \sim \mathcal{N}_c(0,\, P_{ij}(f,t)), \tag{3}$$

where $P_{ij}(f,t) \geq 0$ is the *Power Spectral Density* (PSD) of y_{ij} and stands for its time-frequency energy.

Third, we detail the core idea we use for interference reduction, presented in [7]. Although phase dependencies between channels are neglected, the PSDs P_{ij} of a voice image in all channels are assumed to be the same up to channel-dependent scaling factors $\lambda_{ij}(f)$:

$$P_{ij}(f,t) = \lambda_{ij}(f) P_j(f,t), \tag{4}$$

where $P_j(f,t) \geq 0$ is called the latent PSD of voice j and is independent of the channel i. The scalar $\lambda_{ij}(f) \geq 0$ specifies the amount of interference of voice j into microphone i at frequency band f. They are gathered into $I \times J$ matrices $\Lambda(f)$ called *interference matrices*.

As a consequence of our assumptions (1) and (4), the observations $X_i(f,t)$ also follow the LGM as in (3) but with PSDs written $P_i(f,t)$. We have:

$$X_i(f,t) \sim \mathcal{N}_c(0, P_i(f,t)), \text{with } P_i(f,t) = \sum_{j=1}^{J} P_{ij}(f,t). \tag{5}$$

The free parameters of our model are written

$$\Theta = \left\{ \Lambda(f), \{P_j(f,t)\}_j \right\}. \tag{6}$$

Then, if the parameters are known, the model readily permits effective filtering to recover the voice images. Indeed, according to the Gaussian theory, it is easy to compute the posterior distribution of a voice image Y_{ij} given X_i and the parameters Θ [9]:

$$Y_{ij} \mid X_i, \Theta \sim \mathcal{N}_c\left(\frac{P_{ij}}{P_i} X_i,\, \left(1 - \frac{P_{ij}}{P_i}\right) P_{ij} \right), \tag{7}$$

where we drop the dependence in (f,t) of all quantities for readability. From a Bayesian perspective, this distribution encapsulates everything we know about Y_{ij} once the mixtures and the parameters are known.
Following (7), the maximum a posteriori (MAP) estimate of Y_{ij} is given by:

$$\hat{Y}_{ij} \triangleq \mathbb{E}[Y_{ij} \mid X_i, \Theta] = W_{ij} X_i \triangleq \frac{P_{ij}}{P_i} X_i. \tag{8}$$

In the Gaussian case, this estimate also happens to be the Minimum Mean Squared Error (MMSE) and the Best Linear Unbiased Estimate (BLUE). In any case, the coefficient $W_{ij}(f,t)$ is usually called the *Wiener gain*. The time-domain signals of the estimated images can be obtain from (8) via inverse STFT.

For a given voice j, we are usually not interested in estimating Y_{ij} for all recordings i, but rather only for some, that we call the *close-mics* for voice j, as in [7]. They are given by the channel selection function for voice j, $\varphi(j) \subseteq \{1, \dots, I\}$. It indicates which microphones were positioned to capture voice j and is assumed known.

2.2 Parameter estimation

As discussed in the previous section, if the parameters are known, excellent separation performance can be obtained using the simple Wiener filter (8). The challenge to be overcome is hence to estimate those parameters from the observation of the mixture signals X_i only.
In this section we describe two procedures to perform parameter estimation. They both take as input the STFTs X_i of the recorded signals and the channel selection function φ. Then, they return estimates $\hat{\Theta}$ for the parameters, to be used for separation. A summary can be found in the Algorithm 1 box.

2.2.1 Marginal Modeling

According to [9], a way to estimate our parameters is to maximize the likelihood of the observations, that is find the Θ such that $\mathbb{P}[X \mid \Theta]$ is maximum.
According to our probabilistic framework, all entries $\{X_i(f,t)\}_{i,f,t}$ of the STFTs of the observed microphone signals are independent and distributed according to (5). It follows that we can compute the negative

log-likelihood $\mathscr{L}(\Theta)$ of the parameters Θ as:

$$\begin{aligned}\mathscr{L}(\Theta) &= -\log \mathbb{P}\left[\{X_i(f,t)\}_{i,f,t} \mid \Theta\right] \\ &= -\sum_{f,t,i} \log \mathbb{P}\left[X_i(f,t) \mid \Theta\right].\end{aligned} \tag{9}$$

Maximum Likelihood Estimation (MLE) of the parameters Θ then simply amounts to minimize (9):

$$\hat{\Theta} \leftarrow \arg\min_{\Theta} \mathscr{L}(\Theta). \tag{10}$$

It can be shown equivalent to:

$$\hat{\Theta} \leftarrow \arg\min_{\Theta} \sum_{f,t,i} d_0\left(V_i(f,t) \| \sum_j \lambda_{ij}(f) P_j(f,t)\right) \tag{11}$$

where d_0 is the Itakura-Saito divergence [3], presented as "a measure of the goodness of fit between two spectra"[12].

Whereas [7] used the cost function (11) only for optimizing over Λ, we use it now for all Θ. This is done using the classical Non-negative Matrix Factorization (NMF) methodology, where both $\Lambda(f)$ and $\{P_j\}_j$ are updated alternatively, in a multiplicative fashion. As can be seen, the procedure can simply be understood as fitting the power spectrograms V_i of the recordings to their model P_i. This is done by exploiting the marginal distribution of the mixtures.

Using classical NMF derivations, we can show that optimizing (11) over both Λ and P_j amounts in alternating between the two following updates:

$$P_j(f,t) \leftarrow P_j(f,t) \cdot \frac{\sum_{i=1}^{I} P_i(f,t)^{-2} V_i(f,t) \lambda_{ij}(f)}{\sum_{i=1}^{I} P_i(f,t)^{-1} \lambda_{ij}(f)} \tag{12}$$

$$\lambda_{ij}(f) \leftarrow \lambda_{ij}(f) \cdot \frac{\sum_{t=1}^{T} P_i(f,t)^{-2} V_i(f,t) P_j(f,t)}{\sum_{t=1}^{T} P_i(f,t)^{-1} P_j(f,t)} \tag{13}$$

2.2.2 Expectation Maximization

The second strategy involves the *Expectation-Maximization* iterative algorithm (EM, [13]). Instead of fitting the model directly using the marginal distribution of the observations, the EM methodology introduces the images Y_{ij} as latent variables and each EM iteration alternates between separation and re-estimation of the parameters [11].

[3] a particular case of β-divergence, d_β, with $\beta = 0$

In the so-called *E-step*, exploiting the posterior distribution $\mathbb{P}[Y_{ij} \mid X_i, \Theta]$ of the images, we can compute the posterior total variance $Z_{ij}(f,t)$ as:

$$Z_{ij} \leftarrow \mathbb{E}\left[\left|Y_{ij}\right|^2 \mid X_i, \Theta\right] = W_{ij}^2 V_i + \left(1 - \frac{P_{ij}}{P_i}\right) P_{ij}. \tag{14}$$

In the *M-step*, the parameters are re-estimated so that the image PSDs P_{ij} fit the posterior total variances (14):

$$\Theta \leftarrow \arg\min_{\Theta} \sum_{f,t,i,j} d_0\left(Z_{ij}(f,t) \| P_{ij}(f,t)\right). \tag{15}$$

As in the section 2.2.1, we derive the corresponding updating rule for P_j and $\lambda_{ij}(f)$:

$$P_j(f,t) \leftarrow P_j(f,t) \cdot \frac{\sum_{i=1}^{I} P_{ij}(f,t)^{-2} Z_{ij}(f,t) \lambda_{ij}(f)}{\sum_{i=1}^{I} P_{ij}(f,t)^{-1} \lambda_{ij}(f)} \tag{16}$$

$$\lambda_{ij}(f) \leftarrow \lambda_{ij}(f) \cdot \frac{\sum_{t=1}^{T} P_{ij}(f,t)^{-2} Z_{ij}(f,t) P_j(f,t)}{\sum_{t=1}^{T} P_{ij}(f,t)^{-1} P_j(f,t)} \tag{17}$$

It should be emphasized that the computation of P_{ij} always involves the latest version available of the parameters P_j and Λ. It can be shown that iterating over this EM procedure is guaranteed to lead the parameters to a local optimum for the optimization problem (10) [13].

2.3 Enforcing W-disjoint orthogonality

In the previous section, we presented two alternative methods to estimate our parameters under a maximum likelihood criterion. In both cases, the parameters are refined iteratively so as to best match the observations. We highlight here that the overall optimization problem (10) is non-convex, so that both optimization methods we proposed are sensitive to initialization.

As already advocated in [7], initializing the voice PSD P_j using $\varphi(j)$ already provides a very good efficiency for the algorithm. The rationale of this procedure is that close-mics should already provide a good guess of what each voice should sound like, taking us close to the desired solution. Pioneering work in the field [2] can actually be understood as directly separating the mixtures with this initialization and $\lambda_{ij}(f) = 1$, through the Wiener filter (8).

In this study, we go further than just hoping our initialization will be close enough for the algorithms to obtain good results. On top of our datafit criterion embodied

Algorithm 1: Gaussian Interference Reduction

1. **Input:**
   - $X_i(f,t)$ for each channel x_i;
   - Channel selection function $\varphi(j)$ for each voice j;
   - Minimal interference ρ;
   - Number N_{iter} of iterations;
   - Number N'_{iter} of *inner* iterations (*only for EM*).
   - Sparsity coefficient γ;
2. **Initialization:**
   (a) For each f, i, j, $\lambda_{ij}(f) = \begin{cases} 1 & : i \in \varphi(j) \\ \rho & : \text{otherwise} \end{cases}$
   (b) $P_j(f,t) \leftarrow \frac{1}{|\varphi(j)|} \sum_{i \in \varphi(j)} \frac{1}{\lambda_{ij}(f)} V_i(f,t)$
3. **Parameter Fitting:**
   Marginal Modeling algorithm (*MM*):
   (a) Update all $P_j(f,t)$ with (12), including (21) and (22) to numerator and denominator, respectively
   (b) Update all $\lambda_{ij}(f)$ as in (13)

   Expectation-Maximization algorithm (*EM*):
   (a) Compute Z_{ij} as in (14)
   (b) Update all $P_j(f,t)$ with (16), including (21) and (22) to numerator and denominator, respectively
   (c) Update all $\lambda_{ij}(f)$ as in (17)
   (d) For another *inner* iteration, return to step 3b
4. For another iteration, return to step 3
5. **Separation and output:**
   $\forall j, \forall i \in \varphi(j)$: compute $\hat{Y}_{ij}(f,t)$ as in (8)

by the negative log-likelihood in (9), we propose to also enforce *W-disjoint orthogonality* of the different sources PSDs, as formalized in [14].
W-disjoint orthogonality means that the voices will mostly have energy in different TF bins. Equivalently, it says that for any TF bin, only a few voices should have a significant energy. This phenomenon is often observed in practice and has been exploited for the separation of audio. One contribution of this study is to notice that W-disjoint orthogonality can be understood in terms of sparsity of the vectors $P(f,t)$, defined as the concatenation of the voice PSDs:

$$P(f,t) \triangleq [P_1(f,t), \ldots, P_J(f,t)]. \tag{18}$$

We propose to estimate the parameters by using a new regularized criterion, as:

$$\hat{\Theta} \leftarrow \arg\min_{\Theta} \mathscr{L}(\Theta) + \gamma \sum_{f,t} \Psi(P(f,t)), \tag{19}$$

where $\gamma \geq 0$ indicates the strength of the regularization, while Ψ is a *regularizing function* or *sparsity criterion* that is small whenever its argument is sparse (see [15] for a review). In this study, we considered the Wiener Entropy as a sparsity regularization. For a vector p of length J, it is given by:

$$\Psi(P(f,t)) = \frac{\left(\prod_{j=1}^{J} P_j(f,t)\right)^{\frac{1}{J}}}{\frac{1}{J}\left(\sum_{j=1}^{J} P_j(f,t)\right)}. \tag{20}$$

Since Ψ is independent of Λ, the updates (13) and (17) for Λ are unchanged. Concerning the updates of P_j, as in [15] the formulas (12) and (16) are modified adding the quantities $\nabla^{-}_{\Psi,j}(f,t)$ to their numerator and $\nabla^{+}_{\Psi,j}(f,t)$ to their denominator, as defined by:

$$\nabla^{-}_{\Psi,j}(f,t) = \gamma \frac{J\left(\prod_{j=1}^{J} P_j(f,t)\right)^{\frac{1}{J}}}{\left(\sum_{j=1}^{J} P_j(f,t)\right)^{2}}. \tag{21}$$

$$\nabla^{+}_{\Psi,j}(f,t) = \gamma \frac{\left(\prod_{j=1}^{J} P_j(f,t)\right)^{\frac{1}{J}}}{P_j(f,t)\left(\sum_{j=1}^{J} P_j(f,t)\right)} \tag{22}$$

3 Evaluation

In order to evaluate the proposed algorithms, we conducted an online listening test. The algorithms were applied on a whole pop rock live recording session of 'Huey Lewis and the News' *Hip to Be Square* at the Montreux Jazz Festival 2000 (length: 4'40"). This recording features 23 microphones recording 20 voices. It has a sample-rate of 48 kHZ and a depth of 16 bits/sample. The multitrack recording was provided by the Montreux Jazz Digital Project and EPFL. From this full-length processed recording, a set of two 10 seconds excerpts was extracted for perceptual evaluation.
Because of the live setup, all the microphone signals contain interferences, so that the standard evaluation metrics for blind source separation [16] were not applicable, since they require a clean reference signal against which to compare the results. Instead, we performed a perceptual audio evaluation inspired by the ITU-BS.1534-2 protocol, a.k.a. MUltiple Stimuli with Hidden Reference and Anchor (MUSHRA, [17]), with some modifications and simplifications based on [18].

MUSHRA is a standard methodology for subjective evaluation of audio with "intermediate impairments" (i.e. significant degradation noticeable in most listening environment), such as in source separation and in interference reduction.
However in our context MUSHRA protocols can not be strictly applied: the reference sound is not hidden and not able to be evaluated and there are not any anchors, that are very bad sounds. We therefore adapted it by following the guidelines found in [18].

3.1 Listeners, data and procedure

There were 28 participants (24 men and 4 women), including the authors, aged between 23 and 57 yr (mean=32.9 yr). Web listening evaluations must take hearing abilities and listening environments of the participants into account. Thus, some preliminary questions about gear and musical background were asked.
The participants were asked 9 questions on the two different 10 seconds excerpts. Each question corresponded to a couple comprising one particular voice instrument and one quality scale. Each question was formulated as a MURSHA-like trial: given a question, it was asked to rate different stimuli on a 100-based quality scale in comparison to a reference. There were 6 sounds to evaluate per question, corresponding to the different algorithms. The instrument selected were the voice of Huey Lewis, the bass guitar and the drums.
The presented scales are a modification of the ones presented in [18] to fit the interference reduction problem:

1. *Acoustic quality of the target sound*: how does the target sound.
   Here is the exact wording of its explanation: "only pay attention to the target sound and do not consider the background, such as other instruments. Provide bad ratings if the target sound is highly distorted, highly unnatural, badly equalized, or misses some parts."
2. *Suppression of background sounds*: how much the background sounds have been suppressed from the recording.
   "Only pay attention to the background (e.g. other instruments or the audience) and do not consider the target sounds. Provide good ratings if background is silent and bad ratings for loud artificial or loud original background sound."
3. *Acoustic quality of background sounds*: how does the background sound.
   "Only pay attention to the background sounds and do not consider the target one. Provide bad ratings if the backgrounds sounds (e.g. other instruments or the audience) are highly distorted, badly equalized, present loud bleeps, rumbles, pops that are not included in the mixture."

3.2 Considered algorithms

With this perceptual evaluation we want to compare the performance of the proposed 4 alternative algorithms and the KAMIR algorithm and its fast approximation presented in [7]. Methods in [2, 4] are not taken into account in this work because they have been already compared to KAMIR in [7]. So that the comparison considered methods are the followings:

K:	KAMIR algorithm
$\tilde{\mathbf{K}}$**:**	Approximation to KAMIR
EM:	Expectation Maximization
$\mathbf{EM}+\mathbf{S}$**:**	Expectation Maximization with sparsity
MM:	Marginal Modeling
$\mathbf{MM}+\mathbf{S}$**:**	Marginal Modeling with sparsity

For all the tests, we chose an FFT size of 4096 samples with 75% overlap, an initial floor interference parameter $\rho = 0.1$, $N_{\text{iter}} = 5$ iterations for the algorithm and $N'_{\text{iter}} = 5$ inner iterations for the EM variants. For the sparse variants, we picked a sparsity weight $\gamma = 1000$.

3.3 Results

In order to conduct a statistical analysis on the collected subjective data, the assessments for each participant are converted linearly to the range 0 to 100. Using some data-visualization tool, we could detect outliers: 3 incomplete and 1 totally-inconsistent evaluations have been legitimately removed. Moreover, dividing the participants according to a self-declared musical expertise significantly changed the results. For instance, background quality ratings are significantly different between non-experts and experts: p-value($\mathbf{EM}+\mathbf{S}$) = 0.0084, p-value($\mathbf{MM}+\mathbf{S}$) = 0.009). Moreover, the outliers mentioned before all belong to the non-experts group. We believe that non-expert participants introduced a big bias in the evaluation and were discarded for analysis, leaving 24 sets of results in total.

As a first analysis, we performed a non-parametric *Friedman test* to compare the results of each pair of algorithms along the three proposed scales. These results indicate that the **MM** and **EM** algorithms performs significantly better than $\tilde{\mathbf{K}}$ and **K** in terms of quality for

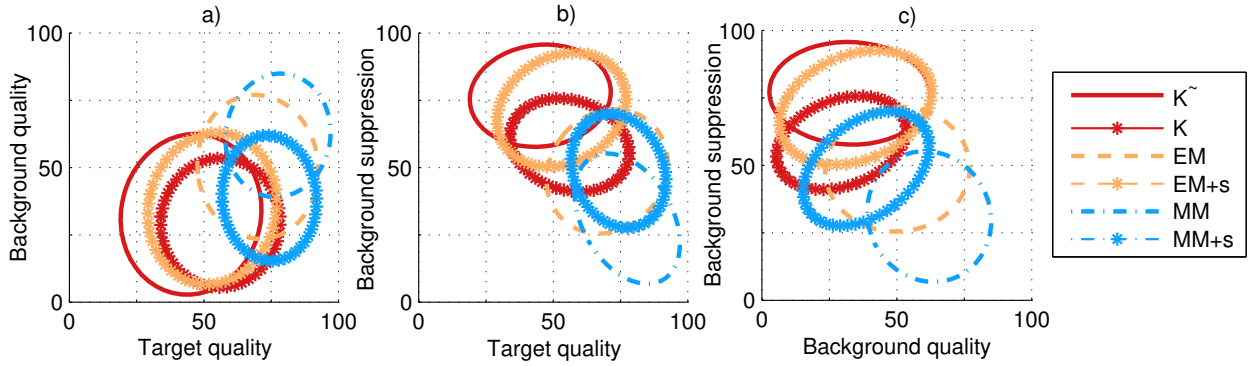

Fig. 2: Listening test result as confidence ellipse

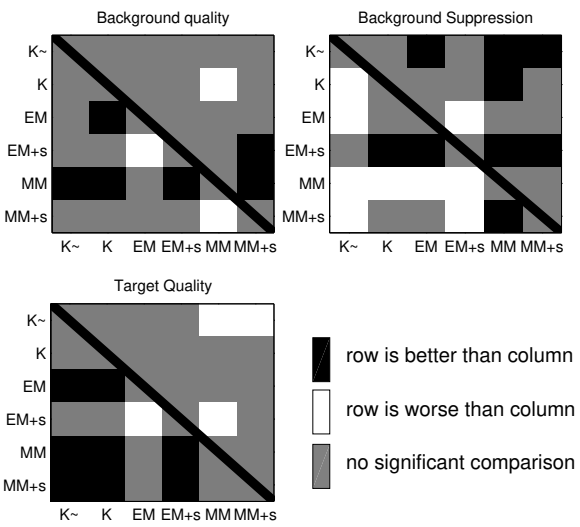

Fig. 3: Pair-wise test for each scale. Lower triangles are for all instruments, upper triangles for vocals only.

both background sounds and target sounds, but worse in terms of suppression. This indicates that these proposed modifications lead to better acoustic quality at the expense of less isolation. However, including a sparsity penalty term to both **EM** and **MM**, improves the suppression capability of the algorithms, suggesting that γ acts as a trade-off between isolation and target quality. Considering now the upper parts of the matrices, we see that the results for vocals only are slightly different, in any case in favor of the proposed modifications.

Figure 2 shows the confidence ellipse of the scores obtained by each algorithm on each pair of scales. It shows how the **EM** and **MM** perform slightly better than KAMIR in both of its fashions. As in Figure 3, we see the benefits of the sparsity penalty as improving background suppression at the cost of introducing some artifacts. An interesting observation is that $\mathbf{EM}+\mathbf{S}$ and $\mathbf{MM}+\mathbf{S}$ appear closer to $\mathbf{K}$ and $\tilde{\mathbf{K}}$ than **EM** and **MM**.

Regardless of the amount of noise that may affect the evaluation results, the EM method presented in this paper leads to slightly better results than state of the art. Close investigation reveals that its main difference with KAMIR lies in handling the uncertainty of the model through the posterior variance in (7). Then, the W-disjoint orthogonality penalty γ in (19) is seen as controlling the trade-off between isolation and distortion. The MM approach does not seem to perform significantly better than KAMIR algorithms, especially for the suppression of background. Still, adding a penalty γ brings it closer to EM, while having a significantly smaller computational complexity.

4 Conclusion

In this paper, we showed how a Gaussian probabilistic model for multitrack signals is useful in designing effective interference reduction algorithms. The core ideas of the model are twofold: neglecting the overly-complex phase dependencies between channels and rather focusing on energy relationships. In contrast to previous studies, we derived estimation procedures for all parameters of the model, leading to provably optimal methods for leakage reduction with this model. In a perceptual evaluation on real-world live recordings

from the Montreux Jazz Festival, we showed that the proposed method behave well when compared with state-of-the-art.

Acknowledgment

This work is made with the support of the French National Research Agency, in the framework of the project DYCI2 "Creative Dynamics of Improvised Interaction" (ANR-14-CE24-0002-01). Access to the Montreux Jazz Festival Database provided by EPFL in the context of this project.

References

[1] Uhle, C. and Reiss, J., "Determined Source Separation for Microphone Recordings using IIR Filters," in *Proceedings of the Audio Engineering Society Convention(AES)*, 2010.

[2] Kokkinis, E. K. and Mourjopoulos, J., "Unmixing Acoustic Sources in Real Reverberant Environments for Close-Microphone Applications," *Journal of the Audio Engineering Society*, 58(11), pp. 907–922, 2010.

[3] Clifford, A. and Reiss, J., "Microphone interference reduction in live sound," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)*, 2011.

[4] Kokkinis, E. K., Reiss, J. D., and Mourjopoulos, J., "A Wiener Filter Approach to Microphone Leakage Reduction in Close-Microphone Applications," *IEEE Transactions on Audio, Speech & Language Processing*, 20(3), pp. 767–779, 2012.

[5] Kokkinis, E., Tsilfidis, A., Kostis, T., and Karamitas, K., "A New DSP Tool for Drum Leakage Suppression," in A. E. S. (AES), editor, *Audio Engineering Society Convention*, 2013.

[6] Benaroya, L., Bimbot, F., and Gribonval, R., "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1), pp. 191–199, 2006.

[7] Prätzlich, T., Bittner, R. M., Liutkus, A., and Müller, M., "Kernel additive modeling for interference reduction in multi-channel music recordings," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 584–588, IEEE, 2015.

[8] Prätzlich, T., Müller, M., Bohl, B. W., Veit, J., and Seminar, M., "Freischütz Digital: Demos of Audio-related Contributions," in *Demos and Late Breaking News of the International Society for Music Information Retrieval Conference (ISMIR), Málaga, Spain*, 2015.

[9] Liutkus, A., Badeau, R., and Richard, G., "Gaussian Processes for Underdetermined Source Separation," *IEEE Transactions on Signal Processing*, 59(7), pp. 3155 –3167, 2011, ISSN 1053-587X, doi:10.1109/TSP.2011.2119315.

[10] Souviraà-Labastie, N., Olivero, A., Vincent, E., and Bimbot, F., "Multi-channel audio source separation using multiple deformed references," *IEEE Transactions on Audio, Speech, and Language Processing*, 23(11), pp. 1775–1787, 2015.

[11] Duong, N., Vincent, E., and Gribonval, R., "Under-Determined Reverberant Audio Source Separation Using a Full-Rank Spatial Covariance Model," *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(7), pp. 1830 –1840, 2010, ISSN 1558-7916, doi:10.1109/TASL.2010.2050716.

[12] Févotte, C. and Idier, J., "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, 23(9), pp. 2421–2456, 2011.

[13] Feder, M. and Weinstein, E., "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Transactions on acoustics, speech, and signal processing*, 36(4), pp. 477–489, 1988.

[14] Jourjine, A., Rickard, S., and Yilmaz, O., "Blind separation of disjoint orthogonal signals: Demixing N sources from 2 mixtures," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on*, volume 5, pp. 2985–2988, IEEE, 2000.

[15] Joder, C., Weninger, F., Virette, D., and Schuller, B., "A comparative study on sparsity penalties for NMF-based speech separation: Beyond LP-norms," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pp. 858–862, IEEE, 2013.

[16] Vincent, E., Gribonval, R., and Févotte, C., "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, 14(4), pp. 1462–1469, 2006.

[17] Series, B., "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[18] Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M., "Fast and easy crowdsourced perceptual audio evaluation," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pp. 619–623, IEEE, 2016.