MICROPHONE ARRAY POST-FILTER FOR SEPARATION OF SIMULTANEOUS NON-STATIONARY SOURCES

Jean-Marc Valin, Jean Rouat and François Michaud

Department of Electrical Engineering and Computer Engineering, Université de Sherbrooke
2500 boul. Université, Sherbrooke, Quebec, Canada, J1K 2R1
`Jean-Marc.Valin@USherbrooke.ca`

ABSTRACT

Microphone array post-filters have demonstrated their ability to greatly reduce noise at the output of a beamformer. However, current techniques only consider a single source of interest, most of the time assuming stationary background noise. We propose a microphone array post-filter that enhances the signals produced by the separation of simultaneous sources using common source separation algorithms. Our method is based on a loudness-domain optimal spectral estimator and on the assumption that the noise can be described as the sum of a stationary component and of a transient component that is due to leakage between the channels of the initial source separation algorithm. The system is evaluated in the context of mobile robotics and is shown to produce better results than current post-filtering techniques, greatly reducing interference while causing little distortion to the signal of interest, even at very low SNR.

1. INTRODUCTION

Mobile robots with abilities to talk and listen should be able to discriminate and separate simultaneous sound sources while moving. For example, in the context of the cocktail party effect, the algorithms have to be robust and should allow the separation of simultaneous voices.

In the present work we first perform a crude linear separation of the sources and then use the proposed post-filter to further enhance the signals and suppress the contribution of the perturbating sources. Our post-filter is inspired by the original work of Cohen [1] who proposes a post-filter designed for a beamformer with one source of interest in the presence of stationary background and transient noises. In the present work we extend the principle to multiple localized sources of interest.

We assume that both the signal of interest and the interferences may be present at the same time and for the

same frequency bin. The novelty of our approach resides in the fact that, for each source of interest, we decompose the noise estimate into a stationary and a transient component assumed due to leakage between channels occuring during the initial separation stage.

For each output channel of the linear source separator, we adaptively estimate the interference parameters (variance and SNR) and use them 1) to compute the probability of targeted speech presence 2) in the suppression rule when both speech and interference are present.

We also propose the use of a Minimum Mean Square Estimation (MMSE) of the loudness – instead of the common log amplitude estimation – yielding a more efficient cleaning of the signal when targeted speech is not present in the channel of interest.

Section 2 gives an overview of the system and Section 3 describes the proposed post-filter. Results and discussion are then presented in Section 4 with the conclusion in Section 5.

2. SYSTEM OVERVIEW

The source separation system discussed here is composed of two subsystems: 1) a linear source separation (LSS) algorithm and 2) the proposed post-filter (Fig. 1). By *linear separation algorithm*, we mean any separation algorithm for which a channel output is the result of a linear transformation of the microphone signals. Most Blind Source Separation (BSS) algorithms fall in this category, as well as distortion-less beamformers and Geometric Source Separation (GSS) techniques [2].

The Linear Source Separation system used for our experiments is inspired from the second constrained (C2) method in [2] and comprises

1. The localization algorithm such as the one described in [3] – It is based on the Time Delay of Arrival (TDOA) estimation;
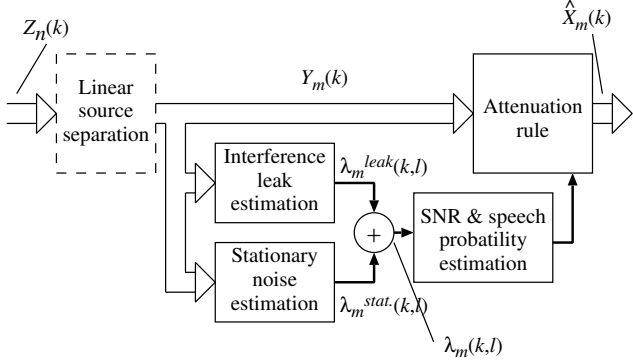
Fig. 1. Overview of the complete separation system. $Z_n(k,l), n = 0\ldots N-1$: Microphone inputs, $Y_m(k,l), m = 0\ldots M-1$: Inputs to the post-filter, $\hat{X}_m(k,l) = G_m(k,l)Y_m(k,l), m = 0\ldots M-1$: Post-filter outputs.

2. The estimated mixing matrix – Assuming unity gain for all microphones, while the phases are computed from the localization algorithm;

3. The pseudo-inverse of the estimated mixing matrix.

In the design of the proposed system, we already take into account that the final application is mobile robotics. As a consequence, our implementation of the LSS system does not include any iterative algorithm – by the time convergence is reached, the robot (or one source) has already moved. We are aware that the LSS algorithm is far from perfect (hence the need for a post-filter) because of localisation accuracy, reverberation and imperfect microphones (non-identical response). We design the post-filter in such a way that any source separation algorithm (including blind algorithms that do not require localization of the sources) can be used.

3. LOUDNESS-DOMAIN SPECTRAL ATTENUATION

We derive a frequency-domain post-filter that is based on the optimal estimator originally proposed by Ephraim and Malah [4, 5]. The novelty of our approach resides in the fact that, for a given channel output of the LSS, the transient components of the corrupting sources is assumed to be due to *leakage* from the other channels during the LSS process. Furthermore, for a given channel, the stationary and the transient components are combined into a single noise estimator used for noise suppression, as shown in Figure 1.

For this post-filter, we consider that all interferences (except the background noise) are localized (detected) sources and we assume that the leakage between channels is constant. This leakage is due to reverberation, localization error, differences in microphone frequency responses, near-field effects, etc.

The next subsection describes the estimation of noise variances that are used to compute the weighting function G_m by which the outputs Y_m of the LSS is multiplied to generate a cleaned signal which spectrum is denoted $\hat{X}_m$.

3.1. Noise estimation

The noise variance estimation $\lambda_m(k,l)$ is expressed as:

$$\lambda_m(k,l) = \lambda_m^{stat.}(k,l) + \lambda_m^{leak}(k,l) \qquad (1)$$

where $\lambda_m^{stat.}(k,l)$ is the estimate of the stationary component of the noise for source m, at frame l, for the kth frequency component and $\lambda_m^{leak}(k,l)$ is the estimate of source leakage.

We compute the stationary noise estimate $\lambda_m^{stat.}(k,l)$ using the Minima Controlled Recursive Average (MCRA) technique proposed by Cohen [6].

To estimate λ_m^{leak} we assume that the interference from other sources is reduced by a factor η (typically $-10\,\mathrm{dB} \leq \eta \leq -5\,\mathrm{dB}$) by the separation algorithm (LSS). The leakage estimate is thus expressed as:

$$\lambda_m^{leak}(k,l) = \eta \sum_{i=0, i\neq m}^{M-1} S_i(k,l) \qquad (2)$$

where $S_m(k,l)$ is the smoothed spectrum of the mth source, $Y_m(k)$, and is recursively defined (with $\alpha_s = 0.7$) as:

$$S_m(k,l) = \alpha_s S_m(k,l-1) + (1-\alpha_s)Y_m(k,l) \qquad (3)$$

3.2. Suppression rule in the presence of speech

We now derive the suppression rule under H_1, the hypothesis that speech is present. From here on, unless otherwise stated, the m index and the l arguments are omitted for clarity and the equations are given for each m and for each l.

The proposed noise suppression rule is based on MMSE estimation of the spectral amplitude in the loudness domain, $|X(k)|^{1/2}$. The choice of the loudness domain over the spectral amplitude [4] or log-spectral amplitude [5] is motivated by better results obtained using this technique, mostly when dealing with speech presence uncertainty (Section 3.3).

The loudness-domain amplitude estimator is defined by:

$$\hat{A}(k) = (E[|X(k)|^{\alpha}|Y(k)])^{\frac{1}{\alpha}} = G_{H_1}(k)|Y(k)| \qquad (4)$$

where $\alpha = 1/2$ for the loudness domain and $G_{H_1}(k)$ is the spectral gain assuming that speech is present.

The spectral gain for arbitrary α is derived from Equation 13 in [5]:

$$G_{H_1}(k) = \frac{\sqrt{\upsilon(k)}}{\gamma(k)} \left[\Gamma\left(1+\frac{\alpha}{2}\right) M\left(-\frac{\alpha}{2};1;-\upsilon(k)\right) \right]^{\frac{1}{\alpha}}$$
(5)

where $M(a;c;x)$ is the confluent hypergeometric function, $\gamma(k) \triangleq |Y(k)|^2 / \lambda(k)$ and $\xi(k) \triangleq E\left[|X(k)|^2\right] / \lambda(k)$ are respectively the *a posteriori* SNR and the *a priori* SNR. We also have $\upsilon(k) \triangleq \gamma(k)\xi(k) / (\xi(k)+1)$ [4].

The *a priori* SNR $\xi(k)$ is estimated recursively as:

$$\hat{\xi}(k,l) = \alpha_p G_{H_1}^2(k,l-1)\gamma(k,l-1) + (1-\alpha_p)\max\{\gamma(k,l)-1,0\}$$
(6)

using the modifications proposed in [6] to take into account speech presence uncertainty.

3.3. Optimal gain modification under speech presence uncertainty

In order to take into account the probability of speech presence, we derive the estimator for the loudness domain:

$$\hat{A}(k) = (E\left[A^\alpha(k)\,|Y(k)]\right)^{\frac{1}{\alpha}}$$
(7)

Considering H_1, the hypothesis of speech presence for source m, and H_0, the hypothesis of speech absence, we obtain:

$$\begin{aligned} E\left[A^\alpha(k)|Y(k)\right] &= p(k)E\left[A^\alpha(k)|\,H_1,Y(k)\right] \\ &+ [1-p(k)]E[A^\alpha(k)|H_0,Y(k)] \end{aligned}$$
(8)

where $p(k)$ is the probability of speech at frequency k.

The optimally modified gain is thus given by:

$$G(k) = \left[p(k)G_{H_1}^\alpha(k) + (1-p(k))G_{min}^\alpha\right]^{\frac{1}{\alpha}}$$
(9)

where $G_{H_1}(k)$ is defined in Eq. 5, and G_{min} is the minimum gain allowed when speech is absent. Unlike the log-amplitude case it is possible to set $G_{min} = 0$ without running into problems. For $\alpha = 1/2$, this leads to:

$$G(k) = p^2(k)G_{H_1}(k)$$
(10)

Setting $G_{min} = 0$ means that there is no arbitrary limit on attenuation. Therefore, when the signal is certain to be non-speech, the gain can tend toward zero. This is especially important when the interference is also speech since, unlike stationary noise, residual babble noise always results in musical noise.

The probability of speech presence is computed as:

$$p(k) = \left\{1 + \frac{\hat{q}(k)}{1-\hat{q}(k)}\left(1+\xi(k)\right)\exp\left(-\upsilon(k)\right)\right\}^{-1}$$
(11)

Table 1. Log spectral distortion and segmental SNR for each of the 3 separated sources.

LSD/SegSNR (dB)	female 1	female 2	male 1
Mic. input	23.4/-5	21.6/-5	21.6/-6.2
LSS output	19.2/2.5	17.1/4.1	17.5/1.6
1-ch. post-filter	10.4/6.1	9.4/7.7	9.9/4.1
Cohen p-f	8.9/6.4	9.7/4.7	10.3/4.5
Proposed p-f	6.5/7.6	6.7/8.1	7/7.1

where $\hat{q}(k)$ is the *a priori* probability of speech presence for frequency k and is defined as:

$$\hat{q}(k) = 1 - P_{local}(k)P_{global}(k)P_{frame}$$
(12)

where $P_{local}(k)$, $P_{global}(k)$ and P_{frame} are defined in [6] and correspond respectively to a speech measurement on the current frame for a local frequency window, a larger frequency and for the whole frame.

4. RESULTS

The system is evaluated in a context of mobile robotics, where an array of 8 microphones is mounted on a mobile robot. In order to test the system, 3 voices (2 female, 1 male) were recorded separately, in a quiet environment. The background noise was recorded on a mobile robot and is comprised of room ventilation and internal robot fans. All four signals were recorded using the same microphone array and subsequently mixed together to allow SNR and distance mesures.

In evaluating our post-filter, we use both the segmental SNR and the log spectral distortion (LSD), which is defined as:

$$LSD = \frac{1}{L}\sum_{l=0}^{L-1}\left[\frac{1}{K}\sum_{k=0}^{K-1}\left(20\log_{10}\frac{|X(k,l)|+\epsilon}{\left|\hat{X}(k,l)\right|+\epsilon}\right)^2\right]^{\frac{1}{2}}$$
(13)

where L is the number of frames, K is the number of frequency bins and ϵ is meant to prevent extreme values.

Table 1 compares the results for separation of each of the 3 original sources with the single-channel and the multi-channel Cohen post-filters, both described in [1]. The Cohen post-filter is adapted to uses the other sources as reference noise signals. The improvement of our post-filter in terms of LSD and SegSNR are confirmed by informal listening.

The spectrograms for the first source (female) is shown in Figure 2. Even though the task involves non-stationary interference with the same frequency content as the signal of interest, we observe that our method (unlike the single-channel post-filter) is able to remove most of the interference, while not causing excessive distortion to the signal of
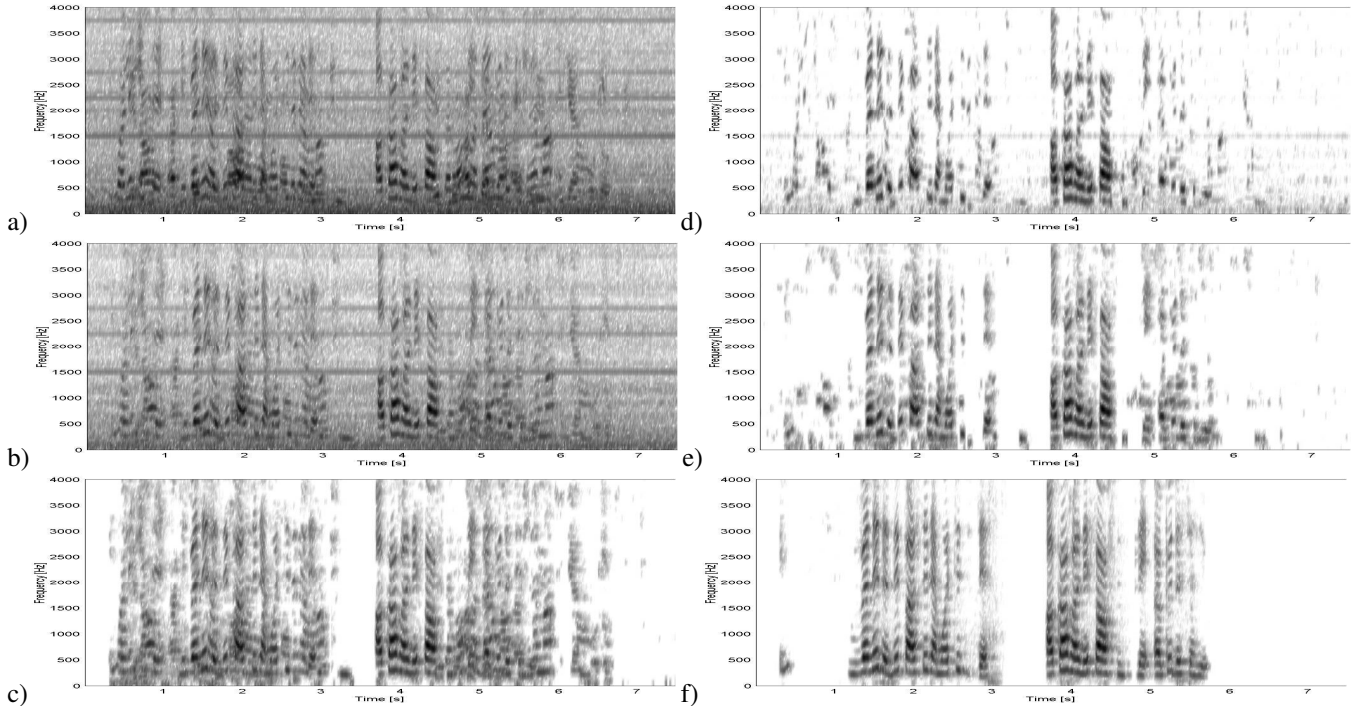
Fig. 2. Spectrogram for separation of first source (female voice) (a) Average of microphone inputs (b) Linear separation output (c) Single-channel post-filtering (d) Adaptation of Cohen post-filter (e) Proposed post-filter (f) Reference signal

interest. Also, for this task, we explain the improvement of our post-filter over the Cohen multi-channel post-filter by the fact that the interference is adaptively estimated even in the presence of the source of interest. This is not the case with the Cohen post-filter, for which the noise estimator (for both stationary and transient noise) is only adapted when the source of interest is absent.

5. CONCLUSION

We proposed a microphone array post-filter designed in the context of separation of multiple simultaneous sources. It is based on a loudness-domain MMSE estimator in the frequency domain with a noise estimate that is computed as the sum of a stationary noise estimate and an estimation of leakage due to the linear source separation (LSS) algorithm. Experimental results show a reduction in log spectral distortion of up to $12\,\mathrm{dB}$ compared to the output of the LSS and up to $4\,\mathrm{dB}$ over the single-channel post-filter.

The proposed post-filter is general enough to be applicable to most source separation algorithms. A possible improvement to the algorithm would be to derive a method that automatically adapts the leakage factor η to track the leakage of an adaptive LSS algorithm.

6. REFERENCES

[1] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *Proc. ICASSP*, 2002, pp. 901–904.

[2] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. SAP*, vol. 10, no. 6, pp. 352–362, 2002.

[3] J.-M. Valin, F. Michaud, J. Rouat, and D. Létourneau, "Robust sound source localization using a microphone array on a mobile robot," in *Proc. IROS*, 2003.

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

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

[6] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing (Elsevier)*, vol. 81, no. 2, pp. 2403–2418, 2001.