A SIMPLE AND EFFICIENT FADER ESTIMATOR FOR BROADCAST RADIO UNMIXING

Mathieu Ramona[*]

IRCAM, Centre Pompidou
1, place Igor Stravinsky, 75004 Paris, France
mathieu.ramona@ircam.fr

Gaël Richard

Institut Telecom, Telecom ParisTech
LTCI-CNRS
37-39 rue Dareau, 75014 Paris, France

ABSTRACT

This paper presents a framework for the estimation of the faders gain of a mixing console, in the context of broadcast radio production. The retrieval of the console state is generally only possible through a human-machine interface and does not permit the automatic processing of such information. A simple algorithm is provided to estimate the faders position from the different inputs and the output signal of the console. This method also allows the extraction of an additional unknown input, present in the mix output. An exhaustive study on the optimal parameter setting is then detailed, that shows good results on the estimation.

1. INTRODUCTION

The transition of the broadcast technical craft from analog to digital audio casting is an important imminent change for radio stations in France. Indeed, the forthcoming revolution of the radio media is the emission of associated interactive visual content that provides a live complement to the audio content. The automatic production of additional multimedia content implies an increased control on the whole media production process. Such feature requires the upstream knowledge of the audio media produced and emitted, which is not possible with the actual broadcast model state.

Interfacing with a mixing console is a typical example of this lack. Typically, several inputs of the console are active and dedicated to various audio streams (i.e. jingles, advertisements, liners...) but only a small part of them is actually present in the mix output emitted by the station. This type of material is highly proprietary and an open machine interface is rarely provided to check the state of the controls. However, knowing the exact content of the output is essential to be able to generate data associated.

A typical example of this issue is the displaying of the album covers of a musical playlist, on a multimedia stream coupled with the audio stream. Succeeding musical tracks are assigned to different channels of the console, and mix-faded. The track faders position determine the song that is actually heard. The blind identification of audio tracks is commonly proceeded through fingerprinting techniques. The contributions in the field are numerous, both from the industrial actors [1][2] the academic world [3]. However, most audio fingerprinting methods are inefficient in the presence of several mixed tracks, and these techniques could only detect the presence of the tracks, not their respective gains. This article shows how a simple signal-based method answers this problem.

Our scope of interest is widened by considering the eventual presence of an additional unknown input in the mix process. Indeed, the estimation of known sources mix logically allows the deduction of the unknown source contribution. We will see that the proposed system is able to extract this source from the output, and give an extensive study on the integrity of the signal extracted. This issue is indeed relevant in our use-case since the speakers microphones are usually directly connected to the mixing console with no possibility to retrieve the signal independently, while others pre-recorded sources are directly accessible to a program.

The definition of the mix estimation problem and the proposed algorithm are presented in Section 2, followed by the description of the experimental protocol of our study in Section 3. An analysis of the results and refining of the parameters will follow in section 4, and Section 5 concludes this work.

2. MIX ESTIMATION

2.1. Definition of the problem

The problem stated is the estimation of the fader gains of a mixing console from the known inputs and output. The inputs of the console are fed with pre-recorded sounds, e.g. jingles, liners or musical tracks. The output is directly retrieved from the mixing console. An important issue, is the effect of the track filters (modelled as Finite Impulse Response filters) applied on each input of the mixing console. The inputs considered in the estimation process are thus previously filtered with the corresponding impulse response, that is measured using Golay codes [4] on each input. The two objectives are the estimation of the fader gains in a dynamic context, and the estimation of an unknown additional input, that contains a signal that is not directly retrievable.

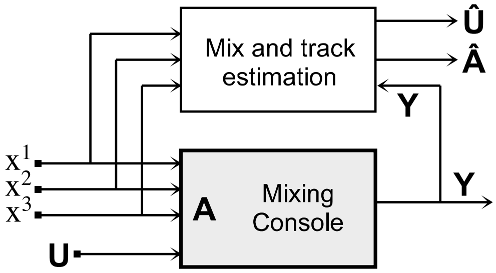


Figure 1: *Architecture of the system.*

The system architecture is summed up in figure 1. The following notations are used in the remainder of this article:
$\mathbf{X_n^i} = \left[x^i(n), \ldots, x^i(n-N+1)\right]^T$ is the N sample column vector for the ith input at instant n,
$\mathbf{X_n} = \left[\mathbf{X_n^1}, \ldots, \mathbf{X_n^I}\right]$ is the input matrix a scenario involving I known inputs, at instant n,
$\mathbf{Y_n} = [y(n) \ldots y(n-N+1)]^T$ is the output column vector.
$\mathbf{U_n}$ is the additional unknown voice input vector, at instant n,

[*] This work was done during my PhD period at the radio station RTL.

$\mathbf{A_n} = \left[a_n^1, \ldots, a_n^I\right]^T$ models the fader gain values at instant n. The mixing console effect is modeled by $\mathbf{Y_n} = \mathbf{X_n}\,\mathbf{A_n} + \mathbf{U_n}$.

2.2. Algorithm

The mix estimation is solved with least mean squares. $\mathbf{A_n}$ is considered as the projection of the output $\mathbf{Y_n}$ on the space generated by the input matrix $\mathbf{X_n}$. Let $\mathbf{X_n}^\dagger$ denote the pseudo-inverse of $\mathbf{X_n}$, then:

$$\mathbf{\hat{A}_n} = \mathbf{X_n}^\dagger\,\mathbf{Y_n} = \left(\mathbf{X_n}^T\,\mathbf{X_n}\right)^{-1}\mathbf{X_n}^T\,\mathbf{Y_n} \tag{1}$$

The faders gain vector $\mathbf{\hat{A}_n}$ is estimated on frames of N samples, with a hop size of R samples between frames. The delay induced by the mixing process in the output can be rendered by the RIF filters applied to each input, and is thus ignored in our formalization.

The fader gain for each track i is thus described by a sequence $\mathbf{A^i} = \left[a_0^i\,a_R^i \ldots a_{k\cdot R}^i\right]$, sampled at $1/R$. The upper Figure 2 shows an example of estimated gain sequences $(\mathbf{\hat{A}^i})_{i=1,\ldots,I}$ for an added noise of 20 dB Signal to Noise Ratio, that models the $\mathbf{U_n}$ signal. To reduce the distortion induced by the added noise, a post-process consisting of a median filter on F samples, is applied. Median filtering is a robust, fast, and very common way to smoothen estimation curves [5]. The lower Figure 2 illustrates the drastic effect in the estimates. The choice of the filter size, fixed to $F = 20$ samples in the figure, must meet a compromise in the reduction of distortions between static and transient parts.

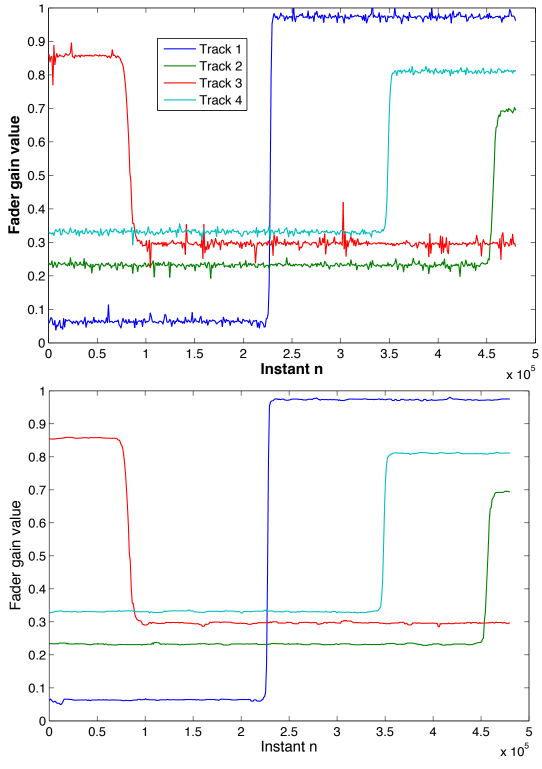


Figure 2: Estimated $\mathbf{\hat{A}^i}$ sequences with added noise 20 dB SNR, with $N = 2000$ and $R = 500$. (up) no post-processing (down) median filter, 20 sample window.

The unknown track is then estimated, with the estimated mix gain:

$$\mathbf{\hat{U}_n} = \mathbf{Y_n} - \mathbf{X_n}\,\mathbf{\hat{A}_n} \tag{2}$$

The remainder of this article focuses on the influence of parameters N, R and F on the estimation, for different Mix to Noise Ratios (MNR) or Mix to Added voice Ratio (MAR), where Mix denotes the mix of the known inputs : $\mathbf{X_n}\,\mathbf{A_n}$.

3. EXPERIMENTS

3.1. Corpus

The evaluation corpus consists of excerpts of radio broadcast news shows, and thus mainly filled with speech, with a possible background liner. The audio tracks are monophonic with 16 bits quantization, sampled at 16 kHz. Each result is computed on a 20 minutes mix simulation.

Four tracks are used for the known inputs $\mathbf{X}^i$ ($I = 4$). The additional signal $\mathbf{U}$ is either a Gaussian noise or another excerpt of the broadcast news. Since different speech signals are more correlated than music and speech signals, our experiment is more constrained than the original requisites. We have also tested the unknown track extraction with the musical known inputs from the RWC music genre database [6], but this brings no significant improvement.

3.2. Fade simulation

As stated earlier, the mix estimation process behaves differently on static and transient parts of the fader gain sequences $(\mathbf{A^i})$. Indeed, the correct estimation of the transient is only done through a linear interpolation between successive frame values. The gain values are thus prone to more distortions on fadings.

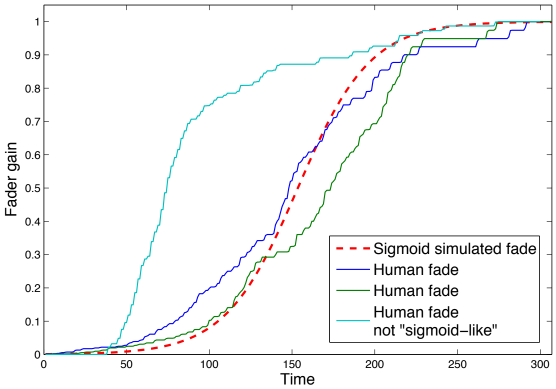


Figure 3: Example of measured fade curves (solid) and modeling by a sigmoid curve (dashed)

The way humans move faders is quite variable, as shown in the solid curve examples of Figure 3, acquired from a mixing console. However, the first two human fadings are quite similar to a sigmoid curve, defined by $S(t) = 1/(1 + e^{-\alpha t})$. This model is used here for the artificiel fading transitions. Because of the fast convergence on the edges, it helps modelling a fast and continuous transition between two values.

In this experiment, the four tracks gain are changed alternatively at random intervals (around 15 s) and follow a sigmoid fade curve during a random interval around 0.5 s. The mean duration of the total fade intervals on each 20 minute mix test is 1 minute.

3.3. Evaluation

For the evaluation of the mix estimation process, the criterion is the mean gain distortion on all tracks (expressed in dB) :

$$G_{dist} = 10 \log_{10} \frac{1}{I} \sum_i \frac{\|\hat{\mathbf{A}}^i - \mathbf{A}^i\|^2}{\|\mathbf{A}^i\|^2} \tag{3}$$

Since this problem is also a source separation problem (with high prior knowledge), the criteria presented in [7] for Blind Source Separation scoring are also relevant in this context, especially for the unknown track estimation. They give a more specific measure for separation than the usual Signal to Noise Ratio.

Let $\mathbf{M}$ be the mixed signal without the unknown track $\mathbf{U}$, the estimate $\hat{\mathbf{U}}$ can be projected on $\mathbf{U}$ and the mixed signal $\mathbf{M}$, with ϵ_{artif} the residual of the projection:

$$\hat{\mathbf{U}} = \langle\hat{\mathbf{U}}, \mathbf{U}\rangle\mathbf{U} + \langle\hat{\mathbf{U}}, \mathbf{M}\rangle\mathbf{M} + \epsilon_{artif}, \tag{4}$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. The Signal to Distortion Ratio (SDR) is a global measure of the separation quality, while the Signal to Interference (SIR) and Signal to Artifacts (SAR) ratios respectively measure the amount of unknown track and artefacts remaining in the separated mixed signal. They are defined as:

$$SDR = 10\log_{10} \frac{\|\langle\hat{\mathbf{U}}, \mathbf{U}\rangle\mathbf{U}\|^2}{\|\langle\hat{\mathbf{U}}, \mathbf{M}\rangle\mathbf{M} + \epsilon_{artif}\|^2} \tag{5}$$

$$SIR = 10\log_{10} \frac{\|\langle\hat{\mathbf{U}}, \mathbf{U}\rangle\mathbf{U}\|^2}{\|\langle\hat{\mathbf{U}}, \mathbf{M}\rangle\mathbf{M}\|^2} \tag{6}$$

$$SAR = 10\log_{10} \frac{\|\langle\hat{\mathbf{U}}, \mathbf{U}\rangle\mathbf{U} + \langle\hat{\mathbf{U}}, \mathbf{M}\rangle\mathbf{M}\|^2}{\|\epsilon_{artif}\|^2} \tag{7}$$

The SIR helps particularly in measuring the proportion of mixed signal kept in the estimation of the unknown track.

The same criteria are also defined on the restriction to the parts containing fades (see section 3.2): $G^{\mathrm{F}}_{\mathrm{dist}}$, SDR$^{\mathrm{F}}$, SIR$^{\mathrm{F}}$, SAR$^{\mathrm{F}}$.

4. RESULTS

4.1. Mix without unknown input

Table 1 shows the gain distortion G_{dist} in the unnoised situation (i.e. $\mathbf{U_n} = 0\ \forall n$) for different median filter length (F) and window size (N) values. Not surprisingly, the mix estimation is more accurate when the median filter is longer and the window more narrow. A negligible gain distortion of -62 dB is measured for the best case ($F = 100$ and $N = 50$). When restricted to fade intervals, $G^{\mathrm{F}}_{\mathrm{dist}}$ is a few dB higher for all values of F and N but still remains very low in the best case ($G^{\mathrm{F}}_{\mathrm{dist}}$ = -58 dB).

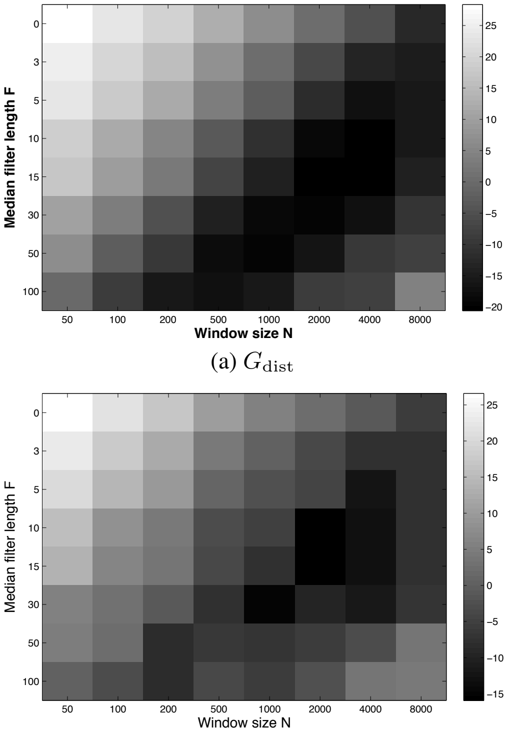
4.2. Robustness to distortions

Naturally, the gain distortion increases when noise is introduced in the mixed signal, and the parameters effect is different. The upper Figure 4 shows the gain distortion measured with F varying from 0 (no filtering) to 100, and N between 50 and 8000, for a SNR of 10dB. The figure clearly shows the correlation between the two parameters optimal values: the minimal gain distortion G_dist remains stable when the product $F \cdot N$ is constant. Indeed, Eq. 1 gets more over-determined when N increases, and F must consequently be lowered to avoid over-smoothing of the gain curves. A global minimum is observed around $N = 2000$ and $F = 15$, with $G_{\mathrm{dist}} = -20.0$ dB.

On the contrary, the gain distortion on the sole transitions (Figure 4) show a much more localized minimum. The minimum peak is also reached for $N = 2000$ and $F = 15$ with $G^F_{\mathrm{dist}} = -15.9$ dB, but decreases strongly outside these values, even when keeping $F \cdot N$ constant. This shows the higher sensitiveness of the gain estimation on fadings.

filt / N	50	200	500	2000	8000
0	-39.5	-34.3	-28.8	-27.9	-15.3
5	-45.1	-41.8	-34.1	-29.8	-15.3
15	-51.8	-52.1	-44.0	-29.8	-14.4
50	-58.9	-54.0	-44.0	-29.2	0.4
100	-62.5	-54.0	-44.0	-7.9	7.3

Table 1: Gain distortion G_{dist} for different configurations of F and N, without unknown input.



(a) G_{dist}

(b) on the sole fading intervals G^F_{dist}

Figure 4: Gain distortion for different F and N values, with an added noise of 10dB SNR.

The same experiment is followed for 20 dB and 5 dB SNRs. Figure 5 compares G_{dist} (solid) and $G^{\mathrm{F}}_{\mathrm{dist}}$ (dashed) for these three SNR values, with different F and a window length of $N = 2000$ samples. G_{dist} is minimized in most cases for $F = 15$. For short median filter lengths, the gain distortion is lower on fading intervals than on the whole signal for SNR of 20dB and 10dB. This reveals the distortion induced in the fading gain by over-smoothing.

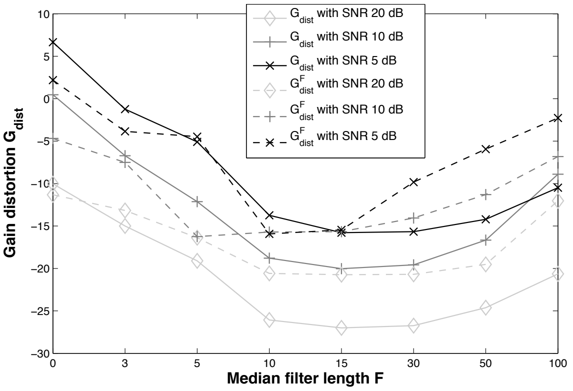


Figure 5: Evolution of G_{dist} (solid) and G^F_{dist} (dashed) with the median filter length ($N = 2000$), for different SNR values.

4.3. Unknown input estimation

In this next experiment, the scope of evaluation is restricted to the fading intervals. The previous experiment has provided some clues to calibrate F and N. If this noise is replaced by a speech track, the minimal gain distortion differs only by a few dB, and is still observed in most cases for $N = 2000$, as shown in table 2(a), where each column sums up the optimal configuration for a given Mix to Additional track Radio (i.e. MAR = $\|\mathbf{M}\|^2$ / $\|\mathbf{U}\|^2$). The gain estimation is evaluated for different values of MAR ranging from 20 dB to -5 dB. For low MAR values (i.e. a stronger added signal) the gain distortion is much higher, and reaches -7dB in the best case for a -5dB MAR. The best configuration is clearly a median filter length of 30 samples and a window of $N = 2000$.

MAR (dB)	20	10	5	0	-5
F	30	50	30	30	30
N	1000	500	2000	2000	2000
$\mathbf{G^F_{dist}}$ (dB)	-22.3	-16.0	-13.8	-10.0	-7.0
(a) Optimal N, F and gain distortion G^F_{dist}					
F	5	5	10	5	5
N	500	2000	500	500	500
$\mathbf{SIR^F}$ (dB)	48.9	51.8	51.4	50.1	56.0
(b) Optimal N, F and Signal to Interference Ratio SIR$^{\mathrm{F}}$					
F	10	10	15	10	10
N	1000	1000	1000	1000	2000
$\mathbf{SAR^F}$ (dB)	18.2	22.9	24.3	25.1	26.8
(c) Optimal N, F and Signal to Artefacts Ratio SAR$^{\mathrm{F}}$					

Table 2: Optimal values on fading intervals for different MAR.

The Signal to Interference Ratio, presented above, evaluates the separation of the unknown track $\mathbf{U}$ by quantifying the proportion of the mixed signal $\mathbf{M}$ present in the estimation $\hat{\mathbf{U}}$. N varies from 500 to 4000, and the median filter length F between 5 and 50. The SIR$^{\mathrm{F}}$ criterion on the sole fading intervals helps judging the separation capability for the different configurations. Table 2(b) shows, for each MAR value, the optimal N and F values, along with the maximum SIR$^{\mathrm{F}}$. The latter criterion increases when the MAR gets higher, which shows that the G_{dist} criterion is not relevant in evaluating source separation since it has an opposite behaviour. The SIR$^{\mathrm{F}}$ is maximized to 56dB with -5dB MAR.

Nevertheless, the artefacts are a much important part of the in noise induced in the source separation, than the interference. Table 2(c) shows the optimal Signal to Artifacts Ratio measured in the same experiment. The latter increases as well when the unknown track energy increases, and reaches 26.8 dB for a -5dB MAR, with $F = 10$ and $N = 2000$. Since the SAR$^{\mathrm{F}}$ is 30 dB lower than the SIR$^{\mathrm{F}}$, the latter is considered negligible, and the global distortion measure SDR$^{\mathrm{F}}$ is considered equal to SAR$^{\mathrm{F}}$. The optimal N and F are thus very close to the values estimated in Section 4.2 above.

A last study is done on the step length R. For each MAR value, the SAR$^{\mathrm{F}}$ score is measured for $R \in \left[\frac{N}{8}\frac{N}{4}\frac{N}{2}\right]$. A systematic improvement is observed with $R = \frac{N}{8}$ and $F' = 4F$. Table 3 shows the gain measured on the SAR$^{\mathrm{F}}$ evaluation criterion, when compared to $R = N$ and $F' = F$.

MAR (dB)	20	10	5	0	-5
$\mathbf{SAR^F}$ (dB)	24.3	26.5	27.4	28.4	28.9
$\mathbf{\Delta SAR^F}$ (dB)	6.1	3.6	3.1	3.3	2.1

Table 3: Gain on the Signal to Interference Ratio on fadings SAR$^{\mathrm{F}}$ with a window hop of $R = \frac{N}{8}$ and $F' = 4F$ (where F is the optimal value with $R = N$) for different MAR values.

5. CONCLUSION

We have presented here an efficient and very simple algorithm for the estimation of a mix with the prior knowledge of the input and output signals. The optimal gain distortion is -20dB on the whole signal and -16dB on the gain fading transitions. The extraction of an added unknown track has shown very reliable since the global signal to distortion measured on the estimation reaches 28.9dB, this distortion is mostly due to artefacts induced by the algorithm.

The major weakness of our algorithm, though, lies in the need of a prior knowledge of the filters applied on each track by the mixing console. An interesting perspective would be the dynamic estimation of the filters response coupled with the mix estimation.

6. REFERENCES

[1] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. ISMIR '02*, October 13-17 2002.

[2] A. Li-Chun Wang, "An industrial-strength audio search algorithm," in *Proc. ISMIR '03*, 2003.

[3] M. Ramona and G. Peeters, "Audio identification based on spectral modeling of bark-bands energy and synchronisation through onset detection," in *Proc. ICASSP*, May 2011.

[4] S. Foster, "Impulse response measurement using golay codes," in *Proc. ICASSP '86*, April 1986, vol. 11, pp. 929–932.

[5] P. F. Welleman, "Robust nonlinear data smoothers: Definitions and recommendations," in *Proc. National Academy of Sciences '77*, February 1977, vol. 74, pp. 434–436.

[6] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC music database: Music genre database and musical instrument sound database," in *Proc. ISMIR*, October 2003, pp. 229–230.

[7] R. Gribonval, L. Benaroya, E. Vincent, and C. Févotte, "Proposals for performance measurement in source separation," in *Proc. ICA*, April 1-4 2003, pp. 763–768.