

A TIME-FREQUENCY METHOD FOR INCREASING THE SIGNAL-TO-NOISE RATIO IN SYSTEM IDENTIFICATION WITH EXPONENTIAL SWEEPS

Piotr Majdak[†], Peter Balazs[†], Wolfgang Kreuzer[†], Monika Dörfler[‡]

[†]Acoustics Research Institute, Austrian Academy of Sciences, Austria

[‡]Numerical Harmonic Analysis Group, Faculty of Mathematics, University of Vienna, Austria

ABSTRACT

Exponential sweeps are widely used to measure impulse responses of electro-acoustic systems. Measurements are often contaminated by environmental noise and nonlinear distortions. We propose a method to increase the signal-to-noise ratio (SNR) by denoising the recorded signal in the time-frequency plane. In contrast to state-of-the-art denoising methods, no assumption about the spectral characteristics of the noise is required. Numerical simulations demonstrate improvements in the SNR under low-SNR conditions even for measurements contaminated by colored noise.

Index Terms — Signal enhancement, time-frequency analysis, audio denoising, exponential sweeps, Gabor transform

1. INTRODUCTION

Exponential sweeps (ESs) are used in the field of audio engineering to measure impulse responses (IRs) of acoustic and electro-acoustic systems like speakers, rooms, in-situ absorption coefficients, or head-related transfer functions [1; 2]. Such measurements are usually contaminated by the environmental noise and are usually performed with a weakly non-linear equipment. With respect to these issues, the ES method provides a large signal-to-noise ratio (SNR) and is immune to non-linearities of the equipment [1].

At least two issues affect the SNR in the recording and thus the SNR of the measurement. First, most acoustic systems like rooms absorb more energy in the higher frequency bands compared to the lower ones and thus, the decay in the measured responses varies with frequency. Second, even though environmental noise is often modeled as an independent and identically distributed (i.i.d) process, most environmental noise sources have a non-flat spectral characteristics (colored noise). In this study, we propose a method that improves the SNR when systems with frequency-dependent response decay are measured under colored-noise conditions.

In acoustics, most denoising methods have been developed for speech or music. Many of these methods rely on the Wiener solution [3] which represents a mean-square-error (MSE) optimum for stationary signals assuming a contamination with an i.i.d process. For colored noise, spectral

subtraction is used where the spectral noise signature is subtracted from the recorded signal [4]. Those methods modify the signal in each time window independently which leads to artifacts like speech distortions or musical noise [5]. In [6], it was shown that the artifacts can be reduced by combining the information from both time and frequency. Recently proposed methods in the joined time-frequency (TF) domain further reduce musical noise [7-9]. However, these methods are speech related and thus do not consider any *a priori* knowledge about the TF representation of the ESs.

A sweep-based method for improving the SNR has been proposed in [10]. Even though this method shows promising results, it is limited to very short IRs, relies on the frequency-independent variance of the noise, and does not incorporate the properties of system-identification with ESs.

In contrast to [10], in our method, we use the *a priori* knowledge about the TF characteristic of ESs and the fact that the system response is decaying with time. Further, by using frame theory [11], we approach perfect reconstruction of clean signals and avoid artifacts like musical noise. The method does not rely on any assumption of the noise but stationarity and is able to handle any arbitrary broadband delay in the recorded signal.

In our method, we represent the recorded response to the ES in the TF plane and classify parts of that plane as either environmental noise or deterministic signal with the goal to obtain a connected region defined as the *signal region*. In contrast to most speech-denoising methods, our method uses hard thresholding: the parts considered as signal are not modified and the parts considered as noise are removed. The classification in either signal or noise is done for each frequency band independently and thus does not rely on any assumption of the spectral characteristic of the noise. By applying a Gabor multiplier [12; 13] corresponding to the signal region, we obtain a denoised version of the recorded signal which is used to estimate the IR of the measured system. Ideally, this method provides both accurate identification of the measured system and noise reduction. The method is evaluated by comparing the SNR in the noisy and denoised IRs.

2. SYSTEM IDENTIFICATION WITH EXPONENTIAL SWEEPS

Let $x[l]$ be an exponential-frequency sweep from f_b to f_e with the length of L_x samples:

$$x[l] = \sin \left[\frac{f_b}{f_s} \frac{L_x}{c} \cdot (e^{l \cdot c/L_x} - 1) \right], \quad l \in [0, L_x] \quad (1)$$

where l is the sample index, c/L_x the slew rate with $c = \ln(f_e/f_b)$, and f_s the sampling rate, see Fig. 1(b).

When using the sweep as the excitation signal for an unknown LTI system with the IR h , the recorded response signal \tilde{y} (length $L_y = L_{lat} + L_x + L_h + L_r$) consists of a broadband latency L_{lat} , the sweep (length L_x), the response of the system (length L_h), and a recording trail (length L_r), see Fig. 1(a). Denoting noise by $\epsilon[l]$, we obtain:

$$\tilde{y}[l] = (h * x)[l - L_{lat}] + \epsilon[l]. \quad (2)$$

The recorded response may also contain nonlinear harmonics, as the measurement is considered to be performed using a weakly non-linear equipment.

The measured IR h_f is defined by:

$$h_f[l] = (\tilde{y} * x^{-1})[l] \quad (3)$$

where $x^{-1}[l]$ is the inverse sweep¹ such that $\delta[l] = (x * x^{-1})[l]$. Then, h_f has the length of $L_y + L_x - 1$ and thus, h_f is usually cropped to \tilde{h} using a simple rectangular window with the length of $L_h[1]$.

3. TIME-FREQUENCY REPRESENTATION

We represent $x[l]$ in the TF domain by using discrete Gabor transform (DGT) [14]:

$$X_{m,n} = \sum_{l=0}^{L_x-1} x[l] \cdot e^{-2\pi j m l / M} \cdot g^*[l - an] = \langle x, g_{m,n} \rangle \quad (4)$$

where M is the number of frequency bins indexed by $m \in [0, M-1]$, a is the hop size of the time-shifts indexed by $n \in [0, L_x/a - 1]$, and $g[l]$ represents the analysis window. We zero-pad x such that a is a divisor of L_x . We choose Gabor atoms $g_{m,n} = e^{2\pi j m l / M} \cdot g[l - an]$ which form a frame implying that there exists a synthesis window $\tilde{g}_{m,n}$ [15] such that:

$$x[l] = \sum_{m,n} \langle x, g_{m,n} \rangle \tilde{g}_{m,n}[l] \quad \text{for all } x \in \mathbb{R}^{L_x}. \quad (5)$$

Further, we define the Gabor multiplier $M_{\mu,g,\tilde{g}}$ [12] as DGT, followed by multiplication with a mask μ , and inverse DGT:

$$M_{\mu,g,\tilde{g}} x[l] = \sum_{m,n} \mu_{m,n} \langle x, g_{m,n} \rangle \tilde{g}_{m,n}[l]. \quad (6)$$

4. PROPOSED METHOD

The SNR in the recorded response signal $\tilde{y}[l]$ is improved by applying $M_{\mu,g,\tilde{g}}$:

$$\hat{y}[l] = M_{\mu,g,\tilde{g}} \tilde{y}[l]. \quad (7)$$

Here μ is the mask containing 1's (defining the signal region) and 0's elsewhere. In each frequency bin, the start of the signal region in μ is determined according to our knowledge of the excitation signal (1) and corrected by the estimated broadband latency of the system L_{lat} . Since the system response is decaying, the end of the signal region in μ can be estimated by classifying parts of $\tilde{Y}_{m,n} = \langle \tilde{y}, g_{m,n} \rangle$ in either noise or the deterministic signal.

4.1. Start of the signal region in the mask

An idealized binary sparse representation $X_{m,n}^{sp}$ of $X_{m,n}$, see Fig. 1(c), is generated as:

$$X_{m,n}^{sp} = \begin{cases} 1 & \forall (\gamma^n(n), n) \quad \text{for } n \in [0, \lfloor L_x/a \rfloor] \\ 1 & \forall (m, \gamma^m(m)) \quad \text{for } m \in [\lfloor f_b M / f_s \rfloor, \lfloor f_e M / f_s \rfloor] \\ 0 & \text{else} \end{cases} \quad (8)$$

where $\gamma^n(n) = \left\lfloor \frac{f_b M}{f_s} \cdot e^{c a n / L_x} \right\rfloor$ and $\gamma^m(m) = \frac{L_x f_s}{c a} \ln \left(\frac{m f_s}{M f_b} \right)$.

Then, we estimate the TF extension n_{lat} of the broadband latency L_{lat} . Fixing a time-shift n_0 , a restricted 2-D correlation between $X_{m,n}^{sp}$ and sub-blocks of $\tilde{Y}_{m,n}$ is calculated:

$$r[n_0] = \sum_{m=f_0}^{f_1} \sum_{n=0}^{L_x/a} X_{m,n}^{sp} \cdot |\tilde{Y}_{m,n+n_0}|^2 \quad (9)$$

where $f_0 = \lfloor f_{lat} M / f_s \rfloor$ and $f_1 = \lfloor f_e M / f_s \rfloor$. The TF representation of the sweep is well localized in time for higher frequencies and thus the correlation is calculated only for $f_{lat} \gg f_b$. Then $n_{lat} = \arg \max_{n_0} r[n_0]$, where $a n_{lat}$ equals L_{lat} up to the resolution of a .

Finally, for each m_0 in the interval $[\lfloor f_b M / f_s \rfloor, \lfloor f_e M / f_s \rfloor]$, the start of the signal region is:

$$n_{b,m_0} = \left\lfloor \frac{L_x f_s}{c a} \ln \frac{f_{m_0}}{f_b} \right\rfloor + n_{lat} \quad (10)$$

with $f_{m_0} = m_0 f_s / M \in [f_b, f_e]$.

4.2. End of the signal region in the mask

For each m_0 in the interval $[\lfloor f_b M / f_s \rfloor, \lfloor f_e M / f_s \rfloor]$, we consider $S_{m_0}[n] = |\tilde{Y}_{m_0,n}|$. The mean μ_{m_0} and standard deviation σ_{m_0} of the noise are estimated from the trailing part of $\log S_{m_0}[n]$, $n \in [L_y - (1 - \alpha_r) L_r, L_y - \alpha_r L_r]$. Introducing $\alpha_r \in [0, 1]$ allows to exclude the start and the end of the trailing part in the estimation, which excludes potential artifacts appearing at the boundaries like fading of the recorded signal. Then, $S_{m_0}[n]$ is smoothed by applying a zero-phase moving-average filter with width W proportional to $L_r \cdot f_s / a$:

$$\bar{S}_{m_0}[n] = (S_{m_0} * h_{MA})[n] \quad (11)$$

where $h_{MA}[n] = \frac{1}{2W}$ for $n \in [-W, W]$ and 0 elsewhere. We

¹ Note that the analytical derivation of the inverse sweep is trivial [1].

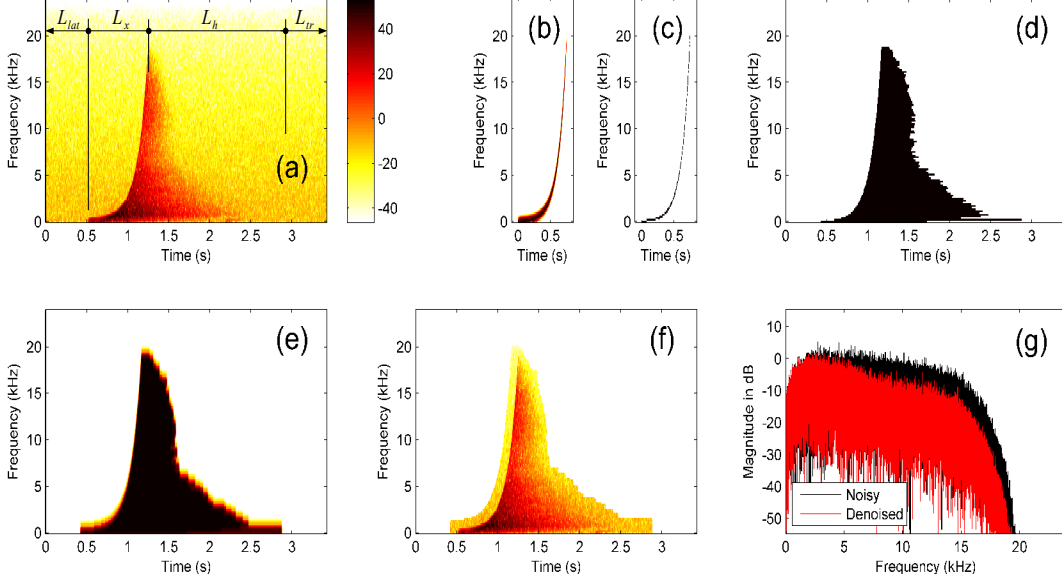


Fig. 1: TF representations of the recorded signal (a) and the exponential sweep (b). Sparse binary TF representation of the sweep (c). Sparse (d) and broadened mask (e) containing the signal region. TF representation of the output of the Gabor multiplier (f). Spectral representation of the differences in the noisy (g, black) and the denoised (g, colored) IRs relative to the clean IR. Note the smaller errors in the denoised condition especially for the higher frequencies.

compare $\log \bar{S}_{m_0}[n]$ with the threshold $(\mu_{m_0} + \varepsilon_{m_0})$ where $\varepsilon_{m_0} = N_\sigma \cdot \sigma_{m_0}$ and N_σ is deduced from the desired classification sensitivity. If $\log \bar{S}_{m_0}[n]$ decays below the threshold, it is assumed that the system response is damped below noise level. To estimate the end of the signal region, the time shift \tilde{n}_{m_0} and the log-amplitude \tilde{S}_{m_0} of the last element above $(\mu_{m_0} + \varepsilon_{m_0})$ are defined as:

$$\begin{aligned} \tilde{n}_{m_0} &= 1 + \max_n [(\log \bar{S}_{m_0}[n]) > (\mu_{m_0} + \varepsilon_{m_0})] \\ \tilde{S}_{m_0} &= \log \bar{S}_{m_0}[\tilde{n}_{m_0}] \end{aligned} \quad (12)$$

The position \hat{n}_{m_0} and log-amplitude \hat{S}_{m_0} of the peak in $\log S_{m_0}[n]$ are determined by:

$$\hat{n}_{m_0} = \arg \max_n \log S_{m_0}[n], \quad \hat{S}_{m_0} = \log S_{m_0}[\hat{n}_{m_0}]. \quad (13)$$

The end of the signal region is defined as the time shift n_{e, m_0} where the linear extrapolation between \hat{n}_{m_0} and \tilde{n}_{m_0} crosses μ_{m_0} :

$$n_{e, m_0} = \frac{\mu_{m_0} - \hat{S}_{m_0}}{\hat{S}_{m_0} - \tilde{S}_{m_0}} (\hat{n}_{m_0} - \tilde{n}_{m_0}) + \hat{n}_{m_0}. \quad (14)$$

4.3. Mask generation

For all m_0 for which $n_{e, m_0} \geq n_{b, m_0}$, the mask is given as:

$$\mu_{m_0, n}^{sp} = \begin{cases} 1 & \forall n \in [n_{b, m_0}, n_{e, m_0}] \\ 0 & \text{otherwise} \end{cases} \quad (15)$$

where 1's represent the signal region, see Fig. 1(d). If $n_{e, m_0} < n_{b, m_0}$, the system response is below noise level and thus $\mu_{m_0, n}^{sp} = 0 \forall n$.

Applying a Gabor multiplier (6) with μ^{sp} to the signal would attenuate important components of the recorded signal near the border of the mask because the support of the reproducing kernel $\langle \tilde{g}, g_{m, n} \rangle$ is broader than a single Gabor coefficient. Thus, μ^{sp} is broadened to incorporate the most

important parts of the reproducing kernel. Note that a natural way would be to perform a synthesis and re-analysis of the mask, however, the mask μ^{sp} is not in the range of DGT and the synthesis and re-analysis would produce artifacts due to the phase effects in the DGT. Hence, we broaden the mask with the magnitude of the reproducing kernel by using 2-D convolution:

$$\mu^b = \mu^{sp} * |\langle \tilde{g}, g_{m, n} \rangle|. \quad (16)$$

The broadened mask is smooth, see Fig. 1(e), and to maintain accurate reconstruction in the signal region, it is converted to binary with a tolerance α_b close to zero:

$$\mu_{m, n} = \begin{cases} 1 & \forall \mu_{m, n}^b > \alpha_b \\ 0 & \text{otherwise} \end{cases}. \quad (17)$$

4.4. Application of the mask

The Gabor multiplier (6) with the output mask μ is applied to the recorded signal \tilde{y} . The resulting masked signal \hat{y} is shown in Fig. 1(f). To obtain the denoised IR \hat{h} , deconvolution (3) is applied to \hat{y} . Note that using a Gabor multiplier with mask μ is similar to cropping h_j , but in a frequency-dependent manner.

5. APPLICATION TO MEASUREMENT OF ROOM IMPULSE RESPONSES

Our method was evaluated by simulating a measurement of the IR of a theater². The measurement was simulated by convolving the IR with an exponential sweep, soft-clipping the response to simulate non-linearities and superimposing additional noise.

Different noise levels were used to simulate different SNRs of the measurements, see Tab. I. Two types of noise

² Municipal Buero Vallejo theater, Alcorc3n, Spain, reverberation time of 1.3 s, available at <http://piotr.majdak.com/download/soundlib/raw/ir1.wav>

were tested: 1) spectrally-flat Gaussian white noise; 2) noise with a spectral slope of 6 dB/oct (brown noise).

The parameters of the simulations were $f_b=50$ Hz, $f_e=20$ kHz, $L_x=0.75$ s, $f_s=48$ kHz, and $L_w=0.3$ s. The first and last 50 ms of the sweep were faded and the amount of non-linear distortions was equivalent to -40 dB.

IRs h , \tilde{h} , and \hat{h} were calculated for the clean, noisy, and denoised cases of system identification by applying (3) to $y=(h*x)$, \tilde{y} , and \hat{y} , respectively. The denoising method was set to $M=256$, $a=M/4$, $N_\sigma=3$, $W=24$, $f_{lat}=4$ kHz, $\alpha_r=0.1$, and $\alpha_b=10^{-5}$ and $g[l]$ was a Gaussian window with the TF-support ratio of aM/L_x .

To quantify the effect of our method, we define:

$$SNR(x, \hat{x})=20 \log_{10} \left(\frac{\|x\|_2}{\|x-\hat{x}\|_2} \right) \quad (18)$$

and calculate $SNR(h, \tilde{h})$ and $SNR(h, \hat{h})$ for the different conditions. The improvement from denoising is given by the SNR gain, $SNR(h, \hat{h})-SNR(h, \tilde{h})$.

The simulation results are provided in Tab. I. Also, we provide results from the block-thresholding method [8], which, as shown in [8], outperforms spectral subtraction [4] and minimum mean-square error log-spectral amplitude estimator [6] and thus seems to be the state-of-the art basis for a comparison.

6. CONCLUSIONS

The proposed method improves the SNR in the impulse response measured with exponential sweeps. It has the following properties:

- The system response is preserved until it decays below the frequency-dependently estimated noise;
- The signal parts considered as noise are removed;
- In the low-SNR conditions, the SNR improves compared to direct measurement and/or block-thresholding;
- In the high-SNR conditions, the method does not fail, i.e., it does not introduce artifacts.

Assuming stationary noise, decaying system response, and an exponential sweep as the excitation signal, our method shows promising results in denoising measurements of electro-acoustic systems.

However, our method seems to be far from the optimal solution. For example, the noise estimator does not use any statistical model – using an appropriate statistical model for the amplitude of the overcomplete TF representation of the noise may help. Also, the separation acuity is low in the low-frequency region and may be improved using the non-stationary Gabor transform [16] in terms of a constant-Q transform.

7. ACKNOWLEDGMENTS

This work was partly supported by the WWTF project MulAc (MA07-025) and by the Austrian Science Fund (FWF) project LOCATIF (T384-N13).

Signal		Our Method		Block Thresholding	
Noise level	SNR	White	Brown	White	Brown
-60	74.46	0.52	0.01	-22.61	-22.52
-40	54.47	1.44	-0.14	-2.98	-3.45
-20	34.39	5.93	3.00	1.93	1.11
0	14.44	8.66	5.23	0.02	0.02

Table I: Simulation results (see text) in dB. Noise level: level of the simulated environmental noise. SNR: SNR resulting from the unprocessed system identification. Our Method: Gain in SNR using the proposed method. Block Thresholding: Gain in SNR using [8]. White: Simulations with spectrally-flat Gaussian noise. Brown: Simulations with noise with a spectral slope of 6 dB/oct.

8. REFERENCES

- [1] S. Müller, and P. Massarani, "Transfer-function measurement with sweeps," *J Audio Eng Soc*, vol. 49, pp. 443-471, 2001.
- [2] P. Majdak, P. Balazs, and B. Laback, "Multiple exponential sweep method for fast measurement of head-related transfer functions," *J Audio Eng Soc*, vol. 55, pp. 623-637, July/August 2007.
- [3] N. Wiener, *Extrapolation, interpolation, and smoothing of stationary time series* (John Wiley & Sons, 1949)
- [4] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans Acoust Speech Sig Proc*, vol. ASSP-27, pp. 113-120, 1979.
- [5] P. Vary, "Noise suppression by spectral magnitude estimation - mechanism and theoretical limits," *Sig Proc*, vol. 8, pp. 387-400, 1985.
- [6] Y. Ephraim, and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans Acoust Speech Sig Proc*, vol. 32, pp. 1109-1121, Dec. 1984.
- [7] F. Millioz, J. Huillery, and N. Martin, "Short time fourier transform probability distribution for time-frequency segmentation," in *Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2006, pp. III-III.
- [8] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Trans on Sig Proc*, vol. 56, pp. 1830-1839, 2008.
- [9] G. Matz, and F. Hlawatsch, "Minimax robust nonstationary signal estimation based on a p-point uncertainty model," *J Franklin Inst*, vol. 337, pp. 403-419, 2000.
- [10] X. Xia, "System identification using chirp signals and time-variant filters in the joint time-frequency domain," *IEEE Trans Sig Proc*, vol. 45, pp. 2072-2084, Aug. 1997.
- [11] H. Bölskei, F. Hlawatsch, and H. G. Feichtinger, "Frame-theoretic analysis of oversampled filter banks," *IEEE Trans Sig Proc*, vol. 46, pp. 3256-3268, 1998.
- [12] H. G. Feichtinger and T. Strohmer, *Advances in Gabor analysis* (Birkhäuser, Basel, 2003)
- [13] P. Balazs, "Basic definition and properties of Bessel multipliers," *J Math Anal Appl*, vol. 325, pp. 571-585, January 2007.
- [14] H. G. Feichtinger and T. Strohmer, *Gabor analysis and algorithms* (Birkhäuser, Boston, 1998)
- [15] P. Balazs, H. G. Feichtinger, M. Hampejs, and G. Kracher, "Double preconditioning for Gabor frames," *IEEE Trans Sig Proc*, vol. 54, pp. 4597-4610, 2006.
- [16] F. Jaillet, P. Balazs, M. Dörfler, and N. Engelputzeder, "Non-stationary Gabor Frames," presented at the *8th International Conference on Sampling Theory and Applications (SAMPTA)*, Marseille 2009.