ASSESSMENT OF MUSICAL NOISE USING LOCALIZATION OF ISOLATED PEAKS IN TIME-FREQUENCY DOMAIN

Ronan Hamon[*]*, Valentin Emiya*

Aix Marseille Univ, CNRS, LIF
Marseille, France

Lucas Rencker[†]*, Wenwu Wang, Mark Plumbley*[‡]

Centre for Vision, Speech and Signal Processing
University of Surrey, Guildford, UK

ABSTRACT

Musical noise is a recurrent issue that appears in spectral techniques for denoising or blind source separation. Due to localised errors of estimation, isolated peaks may appear in the processed spectrograms, resulting in annoying tonal sounds after synthesis known as "musical noise". In this paper, we propose a method to assess the amount of musical noise in an audio signal, by characterising the impact of these artificial isolated peaks on the processed sound. It turns out that because of the constraints between STFT coefficients, the isolated peaks are described as time-frequency "spots" in the spectrogram of the processed audio signal. The quantification of these "spots", achieved through the adaptation of a method for localisation of significant STFT regions, allows for an evaluation of the amount of musical noise. We believe that this will pave the way to an objective measure and a better understanding of this phenomenon.

***Index Terms*—** Musical noise, spectrogram, time-frequency analysis, Delaunay triangulation, denoising

1. INTRODUCTION

Noise reduction techniques have been extensively studied over decades, with the objective of attenuating the background environmental noise while preserving the underlying signal. Many techniques rely on the attenuation of short-time spectral coefficients of spectrogram, such as the squared magnitudes of Short-Time Fourier Transform (STFT), to remove the noise from the signal. These techniques have nevertheless one major disadvantage: Due to the tonal components that emerge in the denoised audio signal, they generate a residual noise, usually called "musical noise" [1], that is unpleasing for the listener [2]. There does not appear to be a formal definition of musical noise, although it is usually associated with the presence of isolated power spectral components in spectrograms [3]. These peaks often result from the poor estimation of time-frequency masks, whether it is in a denoising context or for source separation. As described in [1], peaks and valleys naturally arise in STFT areas containing white noise at random frequencies. After subtraction of the estimation of noise, the peaks remain and the narrowest ones are perceived as time-varying tones known as musical noise.

In the literature, a wide collection of methods have been proposed to reduce musical noise, by improving the estimation of time-frequency masks [4, 5, 6, 7, 8], post-processing techniques [2, 9, 10, 11], or in the context of sparse representations [12, 13]. The evaluation of the effectiveness is in most cases exclusively based on listening tests assessing the overall quality of denoised signals. The influence of musical noise is nevertheless not precisely measured, as other effects of different natures, such as distortion or interferences, may also appear. Some researches have then introduced objective measures of musical noise, by capturing the presence of isolated spectrogram peaks. In [3, 14], the hypothesis is that the isolated peaks impact the probability density function of the power spectral coefficients. They introduce a measure to evaluate the amount of musical noise in a speech signal after denoising by spectral subtraction [3] and Wiener filtering [14], by comparing the processed signals with the change of kurtosis. In [15], the proposed objective measure involves several features, such as the distinctiveness and the non-harmonicity of isolated peaks as well as the similarity between consecutive frames. The complexity of each step makes this measure difficult to interpret and to implement for audio analysis.

In this contribution, we outline a related approach to assess the musical noise in already-processed audio signals. The novelty lies in the detection of the isolated peaks after synthesis of the processed signals. The musical noise is then measured regardless of the spectral technique that generated it, unlike [3, 14]. A better understanding of how isolated power spectral components are incorporated in the synthesised signal is given: indeed, due to the constraints that bind the coefficients of the time-frequency plane, these peaks are not concentrated on one unique time-frequency coefficient, but rather spread in time and in frequency after synthesis to form time-frequency "spots". A characterisation of the shape of these "spots" is given, and a method to localise them in the spectrogram is introduced, based on the work introduced in [16]. Estimating these regions gives then access to the estimation of the underlying musical noise.

This paper is organised as follows. In Section 2, a characterisation of musical noise is introduced, by first recalling some STFT properties and then deriving expected shape for the isolated peaks. In Section 3, after briefly discussing the method introduced in [16] to detect relevant areas in the spectrogram, the main elements of its implementation for the detection of musical noise in audio spectrograms are given. An evaluation of the relevance of this approach is finally given in Section 4.

2. TIME-FREQUENCY CHARACTERISATION OF MUSICAL NOISE

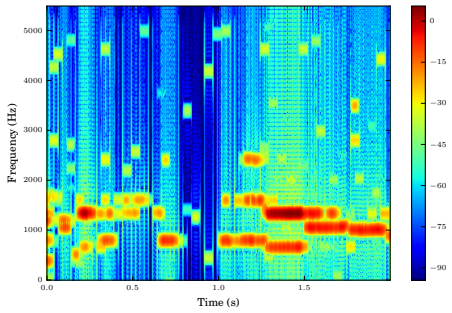
2.1. Consistent representation of isolated peaks

Spectral representations such as spectrograms are efficient and simple tools for analysing and processing audio signals [18]. The STFT

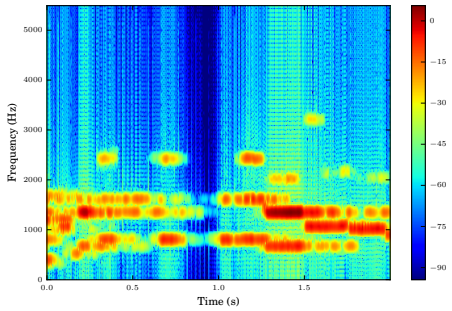
[*]This work was supported by ANR JCJC program MAD (ANR- 14-CE27-0002)

[†]The research leading to these results has received funding from the European Union's H2020 Framework Programme (H2020-MSCA-ITN-2014) under grant agreement no 642685 MacSeNet.

[‡]This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) of the UK grant number EP/L027119/2.



(a) Spectral subtraction [17]



(b) Time-frequency block thresholding [7]

Fig. 1: Log-spectrograms of the output of two denoising techniques applied on a music excerpt [7].

of a real and discrete signal x is defined as:

$$\mathbf{F}(n,m) = \text{STFT}(x) = \sum_{k=0}^{N-1} x(k+Rm)w(k)e^{-i2\pi\frac{kn}{N}} \quad (1)$$

where n is the frequency index, m is the frame index, w is the analysis window of length N and R is the hop size. The corresponding spectrogram simply follows as

$$\mathbf{S}(n,m) = |\mathbf{F}(n,m)|^2. \quad (2)$$

Conversely, the inversion of the STFT is given by

$$\tilde{x}(l) = \sum_m s(l-mR) \sum_n \mathbf{F}(n,m)e^{i2\pi n\frac{l-mR}{N}}, \quad (3)$$

where s is the synthesis window. The STFT representation $\mathbf{F}$ has internal constraints that bind the time-frequency coefficients, that are expressed by the following reproducing identity [19]

$$\mathbf{F} = \text{STFT}[\text{STFT}^{-1}(\mathbf{F})]. \quad (4)$$

After filtering, the resulting spectrogram is generally not consistent, that is to say, the reproducing identity is not verified anymore. When isolated peaks appear in the processed spectrograms, the synthesis into a new audio signal $\tilde{x}$ causes the spread of this highly-localised energy both in time and frequency, leading to visible "spots" in the spectrogram of $\tilde{x}$.

An analytical description of this spot is derived, by considering a STFT representation $\mathbf{F}_0$ with a unique isolated peak at the coordinates (α, β):

$$\mathbf{F}_0(n,m) = \begin{cases} 1+0i & \text{if } n=\alpha \text{ and } m=\beta \\ 0+0i & \text{otherwise.} \end{cases} \quad (5)$$

The synthesis of this spectrogram gives the signal $\tilde{\mathbf{x}}[l]$ using (3)

$$\tilde{\mathbf{x}}[l] = s[l-\beta R]e^{i2\pi\alpha\frac{l-\beta R}{N}}. \quad (6)$$

The time-frequency representation $\hat{\mathbf{F}}_0$ of $\tilde{\mathbf{x}}$ is obtained by using (1)

$$\hat{\mathbf{F}}_0(n,m) = \Big| \sum_{k=0}^{N-1} e^{-i2\pi\frac{k(n-\alpha)-R(m-\beta)}{N}} w(k)s[k-R(\beta-m)] \Big|^2. \quad (7)$$

The study of (7) shows that the energy of the spectrogram of a signal, obtained after synthesis of a spectrogram with a unique isolated peak, is localised in a specific region of the time-frequency domain. This region has boundaries which depend on the usual parameters of the STFT, such as the synthesis and the analysis windows and the hop size. We will consider in the following that these parameters are known, i.e., the shape of the "spots" is known.

2.2. Illustration

To highlight the connection between isolated peaks and musical noise, an illustration is given here on signals with different levels of musical noise. In [7], two denoising techniques are applied on a music except, and the authors assess the amount of musical noise generated by these methods through extensive listening tests. To assess the amount of musical noise in these signals, it is necessary to consider time-frequency "spots" as explained previously.

The spectrogram of the signal obtained after spectral subtraction [17], with a subjective score that indicates that it contains a high amount of musical noise, is displayed in Figure 1a. It shows more spots in the spectrogram, compared to the spectrogram of the signal obtained after time-frequency block thresholding [7], displayed in Figure 1b, which has higher quality score. These observations suggest that the number of "spots" could be a good indicator of the perceived amount of musical noise in an audio signal.

Based in this hypothesis, which is consistent with the observations given in previous works [1, 10, 8, 3, 15], we develop in the next section a methodology to extract from a spectrogram the regions that correspond to the shape of the expected "spots", in order to quantify them.

3. DOMAIN LOCALISATION IN AUDIO SPECTROGRAMS

3.1. Identification using the zeros of the STFT

Recently, a new method has been proposed in [16] to capture the high-energy regions in a spectrogram. Considering signals composed of modulated components, this method allows for the segmentation of their spectrograms and the unmixing and the extraction of the different modes. The identification of these high-energy areas, called domains, is based on the zero coefficients, which have a distribution that is shown to be a signature of the underlying signal. In concrete terms, the zeros are shown to be homogeneously distributed in the low-energy areas, while the high-energy areas do not exhibit zeros, except at their borders. From this observation, the rationale is that when the distance between two zeros is low, the corresponding area is identified as being of low-energy, and conversely. The Delaunay triangulation [20] is then applied to connect all zeros: triangles with edges with a length, i.e., the distance between the corresponding zeros, longer than the expected length of edges of non-significant areas, are identified and merged to construct the significant time-frequency domains. Identification of time-frequency

Algorithm 1 Localisation of Time-Frequency domains

Input: A spectrogram of an audio signal
Output: A list of simplices in the time-frequency plane
Step 1: Detect small local minima of the spectrogram
Step 2: Perform Delaunay triangulation [20] over local minima
Step 3: Select triangles with at least one edge longer than a threshold
Step 4: Group adjacent triangles in domains

components using this approach has been proven to be of interest in tasks such as mode extraction [21] or filtering [16].

3.2. Implementation to the detection of musical noise

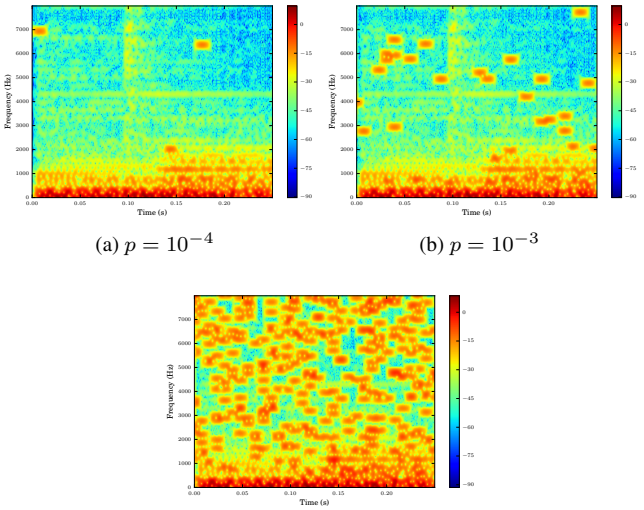
Unlike signals considered in [16, 21], audio signals are too complex to be easily described as the combination of simple modulated components, and then do not exhibit isolated areas of interest. One notable exception is precisely the spots that appear in audio signals with a lot of musical noise, as described in Section 2. Considering that these spots are nearly isolated in the spectrogram, and that their shape is known, it becomes possible to use the methodology described in [16] to locate time-frequency spots corresponding to musical noise in an audio spectrogram. These choices are discussed in the following and are summarised in Algorithm 1.

Step 1 Due to the complexity of audio spectrograms, there is in practice no zero values in spectrograms but rather local minima, as highlighted by the implementation proposed in [16]. When dealing with signals with a few modulated components, it is observed that the minima are always located in the low-energy regions and there is then no need to control their value. In the case of audio spectrograms, it may occur that local minima appear in high-energy areas, as visible for instance in the log-spectrograms in Figure 1. To prevent this problem, a maximal value τ_0 is set in such a way zeros only appear in low-energy regions.

Step 3 After Delaunay triangulation in Step 2, all local minima are connected into triangles. When the triangles are located in low-energy areas, the length of their edges is lower than an empirically chosen value [16]. Conversely, triangles in high-energy areas have longer edges, identified as outliers. In our context, it has been observed that areas of interest are mostly defined by vertical edges, i.e., edges which have a dominant frequency contribution. Outlier edges are then identified by their lengths along the frequency axis. The threshold is chosen by using the expected size of "spots" obtained using (7).

Step 4 The retained triangles are grouped into disjoint areas, called domains. First, the edges which are not shared by two retained triangles constitute the contours of the domains, and vertices are browsed by following the edges until a domain is closed. Domains are then selected according to the expected duration and the frequency band size of "spots" corresponding to musical noise given by (7).

This methodology gives access to the regions of the spectrogram that are considered as responsible for the musical noise. By measuring different statistics, such as the energy of these peaks or their distribution in time and in frequency, they may lead to define an objective measure of the tonal sounds audible in the audio signal. In the following, this measure is defined by simply taking the number of detected "spots".



(a) $p = 10^{-4}$ (b) $p = 10^{-3}$

(c) $p = 10^{-2}$

Fig. 2: Log-spectrogram of a music signal after adding artificial peaks and synthesis, for different value of p.

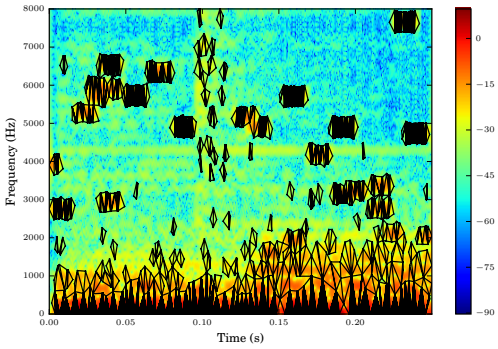
4. PRELIMINARY EXPERIMENTS

We introduce in this section two methods to generate isolated peaks in the spectrogram of an audio signal. We discuss the number of detected spots, with respect to the expected number of generated isolated peaks, and the resulting musical noise assessed through informal listening. All the experiments are performed on the first two seconds of the audio signal *dev1_bearlin_roads_snip_85_99_mix* from the SISEC 11 dataset [22]. The STFT is calculated in the detection method using a hamming window of size 128 with a hop size of 1. The synthesis in both generation methods is performed using a sine window of size 256 with a hop size of 128.

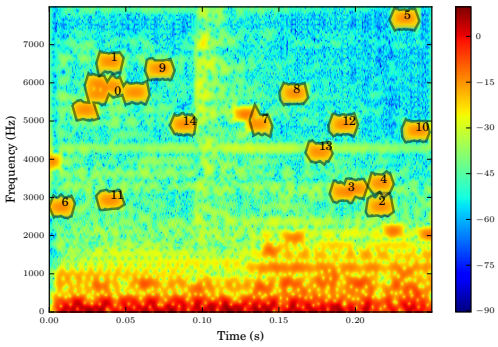
4.1. Addition of artificial isolated peaks

Based on the hypothesis that musical noise is generated by isolated spectral components, random isolated peaks are artificially added in the spectrogram, as introduced in [23] to generate "artificial noise". From a spectrogram $\mathbf{S}$, the resulting spectrogram is expressed by $\tilde{\mathbf{S}} = \mathbf{S} + \mu\mathbf{M}$ where $\mathbf{M}$ is a Bernoulli matrix with the same shape as $\mathbf{S}$ of parameter p, and μ is a scalar value such that the generated peaks are audible in the synthesised signals as a tonal sound. In this generator, the value of p controls the number of isolated peaks, and then the amount of musical noise. Figure 2 shows the log-spectrograms obtained after adding artificial isolated peaks, for $p \in \left\{10^{-4}, 10^{-3}, 10^{-2}\right\}$. When the value of p is small, the number of spots is directly linked with the number of isolated peaks. These peaks are clearly audible in the resulting audio signals as tonal sounds, revealing the presence of musical noise. Above a certain value of p, the peaks are not isolated anymore and tend to merge together, leading to an audio signal sounding like white noise.

Figure 3 shows the log-spectrogram of a musical excerpt with musical noise and the outputs of Algorithm 1. In Figure 3a, the retained triangulation is displayed, highlighting the fact that edges are mostly vertical in high-energy areas. Domains obtained after Step 4, displayed in Figure 3b, are only present in time-frequency spots after selection of the domain with the expected shape.



(a) Triangulation obtained after Step 3



(b) Selected domains after Step 4

Fig. 3: Log-spectrograms of a musical excerpt with musical noise and the output of Algorithm 1 after Steps 3 and 4.

Figure 4 shows the average number of domains in the spectrogram and the corresponding standard deviation for p varying from 10^{-4} to 10^{-1}, and averaged over 10 repetitions. The number of detected "spots" increases when p is comprised between 10^{-4} and 10^{-2}. When p becomes higher, the number of domains decreases as the spots tend to merge together to form larger regions, no more associated with musical noise.

4.2. Orthogonal Matching Pursuit denoising

Sparse representations have been efficiently used for audio denoising [24, 25, 12, 13]. Orthogonal Matching Pursuit (OMP) [26, 27] is a well-studied greedy algorithm that can iteratively estimate a sparse approximation of a signal from noisy observations. The algorithm iteratively reduces the residual error, until an approximation error ϵ is reached. ϵ is usually chosen close to the noise level ($\epsilon = k\sigma^2$), hoping that no noise component will be captured during the process. This parameter acts like a threshold between good signal reconstruction and noise residual: low threshold leads to poor signal reconstruction, while a high threshold leads to isolated noise components being captured, hence producing musical noise.

Figure 5 shows the number of domains in the spectrogram of the resulting signal after OMP of a signal with additive white noise (SNR $=$ 10dB), for different value of the approximation error $\epsilon = k\sigma^2$, and averaged over 10 repetitions of input noisy signal. As for the previous experiment, these results confirm that the method introduced in Section 3 localises spots corresponding to isolated peaks.

In these preliminary experiments, two methods have been introduced to generate isolated peaks, that have been related to the amount of musical noise. The results show that the proposed approach is able to capture these spots, and return a measure that corresponds to the observations.

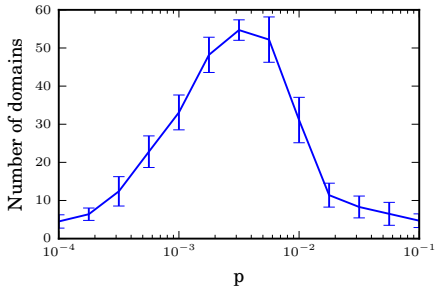


Fig. 4: Average number of domains (or spots) and standard deviation according to the value of p, for 10 draws of random isolated peaks.

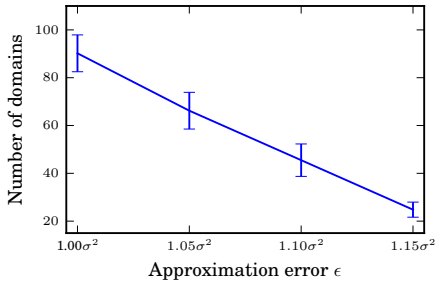


Fig. 5: Average number of domains (or spots) and standard deviation according to the value of OMP parameter $\epsilon = k\sigma^2$, for 10 draws of additive white noise in the input signals (SNR $=$ 10dB).

5. CONCLUSION

The presence of musical noise is a significant factor to consider when assessing the effectiveness of spectral techniques such as denoising. The overall quality of processed audio signals is highly affected by generated spectral artefacts causing annoying tonal sounds. In this paper, a novel approach to evaluate musical noise in an audio signal, regardless of the spectral technique, is introduced, based on the localisation of time-frequency spots, causing musical noise. Preliminary results have highlighted the ability of the proposed method to detect these time-frequency spots, and therefore to assess the perceived amount of musical noise, even if they need to be confirmed with formal listening tests. Further experiments, including a wider range of spectral techniques known to generate annoying musical noise and a comparison with the state-of-the-art objective measures of musical noise, would be also interesting.

More generally speaking, this study proposes new insights in the understanding of musical noise, which is little known and mainly measured in related works through subjective quality assessment. The extraction of the regions of the spectrogram may be used to describe more finely the sound texture of musical noise, through for instance the energy of corresponding spots, or their distribution over the time-frequency plane. It then allows for a better understanding of musical noise, which is an essential step to the development of new strategies for the reduction of the annoying effect it causes.

6. REFERENCES

[1] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *Proc. ICASSP*. IEEE, 1979, vol. 4, pp. 208–211.

[2] Z. Goh, K.-C. Tan, and B. T. G. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 287–292, May 1998.

[3] Y. Uemura, Y. Takahashi, H. Saruwatari, K. S., and K. Kondo, "Automatic optimization scheme of spectral subtraction based on musical noise assessment via higher-order statistics," in *Proc. IWAENC*, Seattle, WA, USA, Sept. 2008.

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

[5] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, 1994.

[6] C.-T. Lu, "Reduction of musical residual noise for speech enhancement using masking properties and optimal smoothing," *Pattern Recognit. Lett.*, vol. 28, no. 11, pp. 1300–1306, 2007.

[7] G. Yu, S. Mallat, and E. Bacry, "Audio denoising by time-frequency block thresholding," *IEEE Trans. Signal Process.*, vol. 56, no. 5, pp. 1830–1839, 2008.

[8] S. Ben Jebara, "A perceptual approach to reduce musical noise phenomenon with Wiener denoising technique," in *Proc. ICASSP*, May 2006, vol. 3.

[9] S. Araki, S. Makino, H. Sawada, and R. Mukai, "Reducing musical noise by a fine-shift overlap-add method applied to source separation using a time-frequency mask," in *Proc. ICASSP*, Mar. 2005, vol. 3, pp. 81–84.

[10] T. Esch and P. Vary, "Efficient musical noise suppression for speech enhancement system," in *Proc. ICASSP*. IEEE, 2009, pp. 4409–4412.

[11] Y. Uemura, Y. Takahashi, H. Saruwatari, K. Shikano, and K. Kondo, "Musical noise generation analysis for noise reduction methods based on spectral subtraction and MMSE STSA estimation," in *Proc. ICASSP*, Apr. 2009, pp. 4433–4436.

[12] C. Févotte, B. Torrésani, L. Daudet, and S. J. Godsill, "Sparse linear regression with structured priors and application to denoising of musical audio," *IEEE Trans. Audio Speech Lang. Process.*, vol. 16, no. 1, pp. 174–185, 2008.

[13] M. D. Plumbley, T. Blumensath, L. Daudet, R. Gribonval, and M. E. Davies, "Sparse representations in audio and music: from coding to source separation," *Proc. IEEE*, vol. 98, no. 6, pp. 995–1005, 2010.

[14] T. Inoue, H. Saruwatari, K. Shikano, and K. Kondo, "Theoretical analysis of musical noise in Wiener filter via higher-order statistics," in *Proc. APSIPA*, Biopolis, Singapore, Dec. 2010.

[15] N. Derakhshan, M. Rahmani, A. Akbari, and A. Ayatollahi, "An objective measure for the musical noise assessment in noise reduction systems," in *Proc. ICASSP*. IEEE, 2009, pp. 4429–4432.

[16] P. Flandrin, "Time–frequency filtering based on spectrogram zeros," *IEEE Signal Proc. Let.*, vol. 22, no. 11, pp. 2137–2141, 2015.

[17] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, no. 2, pp. 113–120, 1979.

[18] P. Flandrin, *Time-frequency/Time-scale Analysis*, vol. 10, Academic Press, 1998.

[19] N. Sturmel and L. Daudet, "Signal reconstruction from stft magnitude: a state of the art," in *Proc. DAFx-11*, 2011, pp. 375–386.

[20] A. Okabe, B. Boots, K. Sugihara, and S. N. Chiu, *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, vol. 501, John Wiley & Sons, 2009.

[21] S. Meignen, T. Oberlin, P. Depalle, P. Flandrin, and S. McLaughlin, "Adaptive multimode signal reconstruction from time–frequency representations," *Phil. Trans. R. Soc. A*, vol. 374, no. 2065, 2016.

[22] S. Araki, F. Nesta, E. Vincent, Z.ěk Koldovskỳ, G. Nolte, A. Ziehe, and A. Benichoux, "The 2011 signal separation evaluation campaign (SiSEC2011):-audio source separation," in *Proc. LVA/ICA*. Springer, 2012, pp. 414–422.

[23] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 7, pp. 2046–2057, Sept. 2011.

[24] M. G. Jafari and M. D. Plumbley, "Speech denoising based on a greedy adaptive dictionary algorithm," in *Proc. EUSIPCO*. IEEE, 2009, pp. 1423–1426.

[25] C. Kereliuk and P. Depalle, "Sparse atomic modeling of audio: A review," in *Proc. DAFx-11*, 2011.

[26] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition," in *Proc. Asilomar*, Nov. 1993, pp. 40–44vol.1.

[27] A. Adler, V. Emiya, M. G. Jafari, M. Elad, R. Gribonval, and M. D. Plumbley, "Audio inpainting," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 3, pp. 922–932, 2012.