Generalized Wiener filtering with fractional power spectrograms

Antoine Liutkus, Roland Badeau

GENERALIZED WIENER FILTERING WITH FRACTIONAL POWER SPECTROGRAMS

Antoine Liutkus[1] *Roland Badeau*[2]

[1]Inria, Speech processing team, Villers-lès-Nancy, France
[2]Institut Mines-Télécom, Télécom ParisTech, CNRS LTCI, France

ABSTRACT

In the recent years, many studies have focused on the single-sensor separation of independent waveforms using so-called soft-masking strategies, where the short term Fourier transform of the mixture is multiplied element-wise by a ratio of spectrogram models. When the signals are wide-sense stationary, this strategy is theoretically justified as an optimal Wiener filtering: the power spectrograms of the sources are supposed to add up to yield the power spectrogram of the mixture. However, experience shows that using fractional spectrograms instead, such as the amplitude, yields good performance in practice, because they experimentally better fit the additivity assumption. To the best of our knowledge, no probabilistic interpretation of this filtering procedure was available to date. In this paper, we show that assuming the additivity of fractional spectrograms for the purpose of building soft-masks can be understood as separating locally stationary α-stable harmonizable processes, α-harmonizable in short, thus justifying the procedure theoretically.

***Index Terms*—audio source separation, probability theory, harmonizable processes, α-stable random variables, soft-masks**

I. INTRODUCTION

In the past ten years, much research has focused on the *demixing* of musical signals. The objective of such research is to process a musical track so as to recover the original individual sounds that were used for its making. For instance, such a process would permit to automatically recover the voice signal from a song and thus automatically generate a karaoke version as well as solo vocals that could be used for resampling. In the scientific community, each constitutive component —or *stem*— from the mixture is called a *source*, and the problem of demixing is commonly called *audio source separation* [6], [32], [25], [19]. In the literature, both single-channel and multichannel audio source separation have been considered, depending on the number of channels of the mixture signal. For the sake of simplicity, we will only consider the single channel case in this study and leave the multichannel case for future developments.

For achieving single channel audio source separation, an efficient approach is focused on a *filtering* paradigm: each source estimate is obtained by applying a time-varying filter to the mixture. In practice, a time-frequency (TF) representation of the mixture is computed, such as its short-term Fourier transform (STFT), and each source is recovered by multiplying each element in this representation by a gain between 1 and 0, according to whether this point is identified as rather belonging to this source or not, respectively [4], [34], [8], [3]. For one given source, those gains form a *time-frequency mask*, and several ways of designing such masks have been considered in the past.

In the audio source separation literature, an important path of research is to consider the devising of TF masks as a *classification* problem. In that setting, the entries of the mask are either 0 or 1: it is typically assumed that only one source is active for any TF bin, so that the problem becomes to determine the source to which each entry of the mixture STFT is associated to. The separation algorithm hence inputs the mixture and performs a multi-class classification task, where each class corresponds to one source. Among those techniques, we can mention the celebrated DUET [34] and ADRESS [2] algorithms, that classify TF bins according to panning positions in the stereo plane. In the single-sensor case, other works attempt to separate sources with binary masks by using harmonicity assumptions: a melody line is first extracted, and then a binary comb-filter is generated to extract the corresponding source [27]. Other recent research considers deep neural network structures to generate the binary mask used to separate target sources [33].

Even if reducing the separation problem to a classification task is convenient, it comes with the drawback of bringing a characteristic and annoying *musical noise*, due to abrupt phase and amplitude transitions in the estimates. To address this issue, many researchers have focused on a *soft masking* strategy, where the TF mask is no longer binary, but rather lies in the continuous $[0\,1]$ interval. It has long been acknowledged that such strategies have the noticeable advantage of strongly reducing musical noise. Many different approaches were undertaken in the past for the purpose of building a soft TF mask. Among them, we can mention some studies where this mask is based on a divergence measure between the mixture and some model for the source: the further the observation is from the model, the smaller the weight, as in [21], [10], [26]. This approach has the advantage of requiring a model only for the target source to separate, but has the inconvenient to be unpractical if more than one source is to be extracted from the mixture.

The most popular approach to soft-masking for source separation today is based on estimating a nonnegative time-frequency energy distribution for each source, which is most commonly called a "spectrogram" in a loose acceptation. Then, the soft mask is computed for each source as the ratio of its estimated spectrogram over the sum of them all. This strategy guarantees that the sum of all soft masks equals 1 for each TF bin, so that the sum of all estimated sources is identical to the mixture, which is a desirable property. For the purpose of estimating those spectrograms, it is typically assumed that they simply add up to yield the observable spectrogram of the mixture, notwithstanding destructive interferences. Given some assumptions on how those spectrograms should look like, such as a specific parametric form [25] or local regularities [20], [22], estimation is performed as a latent variable decomposition of the spectrogram of the mixture.

It has long been acknowledged [4], [3], [5], [18] that when the spectrogram is understood as an estimate of the time-varying Power Spectral Density (PSD) of the source, this weighting strategy is theoretically justified as an optimal Wiener filtering performed independently in each frame. This filter provides the Minimun Mean Squared Error (MMSE) linear estimator of the sources given the mixture. Furthermore, theory does suggest that the PSDs of uncorrelated wide-sense stationary (WSS) processes do add up to yield the PSD of their sum [18]. For all this framework to hold,

This work was partly supported under the research programme EDiSon3D (ANR-13-CORD-0008-01) funded by ANR, the French State agency for research.

the spectrograms to be used must hence be estimates of PSDs, i.e. *squared* modulus of STFTs. We should emphasize here that this acceptation is actually the original and only rigorous one.

However, much research undertaken in the recent years has commonly understood the term "spectrogram" with a different meaning. Instead of seeing it as an estimate of the PSD, many researchers have used the word "spectrogram" to denote the modulus of the STFT raised to some arbitrary exponent $\alpha \in]0\,2]$ (see [29], [11], [14], [30]). Choosing $\alpha = 1$ is common. In the sequel, the term α-spectrogram will be used for clarity to denote this wider acceptation of the word. Just like in the WSS case with $\alpha = 2$, it is then typically assumed that the α-spectrograms of the sources add up to form the α-spectrogram of the mixture, and soft masks are derived in the same way as for the Wiener filter. Experience shows that such a procedure *does* often lead to improved performance. However, no theoretical foundation was available to explain and support this approach: to the best of our knowledge, both additivity of the α-spectrograms and soft-masking filtering are only justified theoretically for $\alpha = 2$.

In this paper, we show that using general α-spectrograms for sources modeling and separation is the optimal procedure if the sources are not understood as WSS processes, but rather as *locally stationary stable harmonizable* processes [28], α-harmonizable processes in short. Note that for $\alpha = 2$, such processes coincide with Gaussian processes [18]. They fall under the umbrella of α-stable distributions [24], [28]. Several studies demonstrated that those distributions are often better models for audio signals than the Gaussian distribution, due to their ability to handle very large deviations from the mean, which is important for such impulsive phenomena as music or sound signals in general that exhibit a large dynamic range [16], [12]. Whereas some papers focused on the separation of independent and identically distributed (i.i.d.) α-stable random variables [16], no study so far considered the separation of locally stationary and harmonizable stable processes. As we show, they provide the exact probabilistic framework needed to assume additivity of α-spectrograms as well as a justification for the design of the corresponding soft-masks.

This paper is structured as follows. In section II, we study the empirical validity of the additivity assumption for α-spectrograms. In section III, we quickly introduce α-harmonizable processes and show how they can be separated using soft masking strategies. In section IV, we compare the music separation performance of this stable harmonizable model as a function of the exponent α. Finally, we draw some tracks for future research as a conclusion.

II. ADDITIVITY OF α-SPECTROGRAMS

II-A. Notations and background

Let $\tilde{x}(t)$ be the audio signal to be separated, which is assumed regularly sampled. In typical audio applications, it is the waveform of the single channel song to be unmixed and for this reason, $\tilde{x}$ is called the *mixture* in the following. The mixture is assumed to be the simple sum of J underlying signals $\tilde{s}_j(t)$ called sources, that correspond to the individual waveforms of the different instruments playing in the mixture, such as voice, bass, guitar, percussions, etc.

In typical source separation procedures, the mixture is processed so as to compute its STFT denoted $x(f,n)$, where f is a frequency index and n is a frame index. x is thus a $N_f \times N_n$ matrix, where N_f is the total number of frequency bands[1] and N_n the total number of time frames. (f,n) is called a TF bin. For music source separation, experience shows that having frames approximately 80ms long with 80% overlap yields good results. Since the STFT is a linear transform, the simple mixing model we choose leads to:

$$\forall (f,n), x(f,n) = \sum_{j=1}^{J} s_j(f,n),$$

[1]Since $\tilde{x}$ is a real signal in audio, its spectrum is Hermitian. We assume that the redundant information in the Fourier transform of each frame has been discarded.

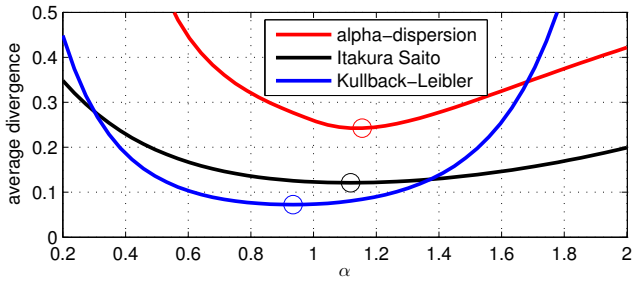

Fig. 1. Average L_α, Itakura-Saito and Kullback-Leibler divergences between the sum of the α-spectrograms of the sources and the α-spectrogram of the mixture, as a function of α. Minimal values are marked with a circle.

where s_j is the STFT of source j. For convenience, the modulus of the STFT is denoted p in the following[2]:

$$p(f,n) \triangleq |x(f,n)|.$$

Throughout this paper, the α-spectrogram p^α is defined as $p^\alpha(f,n) \triangleq p(f,n)^\alpha$. Similarly, p_j^α corresponds to the α-spectrogram of source j. As we see, the 2-spectrogram is the *power* spectrogram, which is the estimate of the PSD.

Most audio source separation methods can be understood as assuming that we basically have:

$$\forall (f,n), p^\alpha(f,n) \approx \sum_{j=1}^{J} p_j^\alpha(f,n). \quad (1)$$

As seen above, this assumption is justified theoretically when $\alpha = 2$ if we assume that the sources are locally stationary Gaussian processes [18]. For any other $\alpha \in]0\,2]$, no such probabilistic framework is available even though (1) is often assumed [29], [11], [14], [30].

II-B. Experimental study

The objective of this section is to study the validity of the additivity assumption (1) for α-spectrograms, as a function of α. To this purpose, we consider the 8 complete songs of different musical genres found in the QUASI database[3], for which the constitutive sources are available. For a set of 50 α values ranging from 0.2 to 2, we computed the *α-dispersion* between the mixture α-spectrogram and the sum of the α-spectrograms of the sources:

$$L_\alpha(f,n) = \left| p^\alpha(f,n) - \sum_{j=1}^{J} p_j^\alpha(f,n) \right|^{1/\alpha}, \quad (2)$$

as well as the popular Itakura-Saito (IS) and Kullback-Leibler (KL) divergences, commonly used in audio source separation [9], [7]. Then, the average of each divergence over all songs and all TF bins was computed, as a function of α. The results are displayed in Fig. 1.

II-C. Discussion

As can be noticed in Fig. 1, the additivity assumption (1) is not equally valid for all α. On the contrary, we clearly see that a value $\alpha \approx 1$ is much more empirically appropriate than the value $\alpha = 2$, for all divergences considered.

[2]$\triangleq$ denotes a definition.

[3]www.tsi.telecom-paristech.fr/aao/en/2012/03/12/quasi/

This result has already been noticed, e.g. in [13], [17], and demonstrates that assuming additivity of the power spectrograms, even if justified theoretically under Gaussian assumptions, is not mostly appropriate. On the contrary, assuming additivity of the moduli p_j of the STFT of the sources for audio processing, as in most Probabilistic Latent Component Analysis studies (PLCA, see [30] and references therein), is indeed a good idea[4].

However, this empirical fact does raise an important question. When estimates of the α-spectrograms p_j^α of the sources have been obtained by any appropriate method, the estimation of the STFT of source j is then typically achieved through:

$$\hat{s}_j(f,n) = \frac{p_j^\alpha(f,n)}{\sum_{j'} p_{j'}^\alpha(f,n)} x(f,n), \tag{3}$$

which we call an α-Wiener filter in the following. Is this procedure any good and does it come with any flavor of optimality? If so, in which sense? The current lack of a probabilistic model justifying (1) for $\alpha \neq 2$ also prevented answering these questions so far. As we now show, assuming that each source is a locally stationary and α-stable harmonizable process naturally leads to (1) and, for $0 < \alpha \leq 2$, establishes (3) as the conditional expectation of $s_j(f,n)$ given $x(f,n)$, thus providing a theoretical understanding for the validity of the procedure.

III. α-HARMONIZABLE PROCESSES

We define an α-harmonizable process as a process that can locally be approximated as a stationary harmonizable α-stable process. In practice, the audio is split into overlapping frames, which are then assumed independent and each one of them is assumed stationary α-stable harmonizable.

In this section, we briefly present stationary α-stable harmonizable processes, which have been the topic of much research since the 70s and are particular cases of α-stable processes [23], [28], [24], [31], [12]. Due to space constraints, only some important facts which are of interest in our study are recalled here and the interested reader is referred to the very thorough overview of α-stable processes given in [28] and references therein for a more comprehensive treatment.

III-A. Symmetric α-stable distributions and processes

Let v be a random vector of dimension $T \times 1$. We say that v is strictly stable if for any positive numbers A and B, there is a positive number C such that

$$Av^{(1)} + Bv^{(2)} \overset{d}{=} Cv, \tag{4}$$

where $v^{(1)}$ and $v^{(2)}$ are independent copies of v and $\overset{d}{=}$ denotes equality in distribution. It can be shown [28, p. 58] that for any random vector v satisfying (4), there is one constant $\alpha \in]0\,2]$ called the *characteristic exponent* such that C in (4) is given by:

$$C = (A^\alpha + B^\alpha)^{1/\alpha}.$$

We then say that v is α-stable. If v and $-v$ furthermore have the same distribution, v is called symmetric α-stable, abbreviated as SαS. An important result is that the simple property (4) of an α-stable random vector permits to derive its characteristic function. No expression for the α-stable probability density functions is available in general, but only for $\alpha = 2$ and $\alpha = 1$, that respectively coincide with the Gaussian and Cauchy distributions.

α-stable distributions have an important number of desirable properties. One of the most famous is their ability to model data with very large deviations, making them a practical model for impulsive data in the field of robust signal processing [24]. In practice, the closest α is to 0, the heavier are the tails of an α-stable distribution. In a source separation context, the *stability* property (4) is fundamental. It basically means that provided the sources are modeled as α-stable, so will be their mixture.

We say that a collection $\{\tilde{z}(t)\}_t$ of random variables is an α-stable random *process* if the vector $\tilde{z}_T \triangleq [\tilde{z}(t_1), \ldots, \tilde{z}(t_T)]^\top$ (where $^\top$ denotes transposition) is α-stable for any choice and any number of sample positions $t_1, \ldots, t_T$.

III-B. Isotropic complex SαS random variables

Because it will be useful in the sequel, we mention here that a complex random variable (r.v.) $z = v_1 + iv_2$ is called SαS if the random vector $\left[v_1^\top\, v_2^\top\right]^\top$ is SαS. A particular case of interest in our context is the special case where a complex SαS r.v. z is isotropic, or circular, abbreviated $S\alpha S_c$, meaning that:

$$\forall \theta \in [0\,2\pi[\,, \exp(i\theta)\, z \overset{d}{=} z.$$

It can be shown that in the Gaussian case $\alpha = 2$ this is equivalent to v_1 and v_2 being independent and identically distributed (i.i.d.) Gaussian r.v., whereas for the case $\alpha < 2$, isotropy leads to the particular characteristic function [28, p. 85]:

$$\begin{aligned} z = v_1 + iv_2 \sim S\alpha S_c \\ \Leftrightarrow \mathbb{E}\left[\exp\left(i\left(\theta_1 v_1 + \theta_2 v_2\right)\right)\right] = \exp\left(-\sigma^\alpha |\boldsymbol{\theta}|^\alpha\right), \end{aligned} \tag{5}$$

where $|\boldsymbol{\theta}|$ is the Euclidean norm of the vector $[\theta_1\, \theta_2]$, and $\sigma > 0$ is a scale parameter[5]. The real and imaginary parts of an isotropic complex SαS r.v. are *not* independent in general. As can be seen, the isotropic complex SαS distribution is only parameterized by the scale parameter σ. For convenience, we denote it $S\alpha S_c(\sigma^\alpha)$. We trivially have:

$$\begin{aligned} z_1 \sim S\alpha S_c(\sigma_1^\alpha) \text{ and } z_2 \sim S\alpha S_c(\sigma_2^\alpha),\ z_1 \text{ and } z_2 \text{ independent} \\ \Rightarrow z_1 + z_2 \sim S\alpha S_c(\sigma_1^\alpha + \sigma_2^\alpha). \end{aligned} \tag{6}$$

III-C. Stationary harmonizable α-stable processes

An harmonizable process $\tilde{z}(t)$ is defined as the inverse Fourier transform of a complex random measure $z(\omega)$ with independent increments:

$$\tilde{z}(t) = \int_{-\infty}^{\infty} \exp(i\omega t)\, z(\omega)\, d\omega. \tag{7}$$

In expression (7), the r.v. $z(\omega)$ may be understood as the spectrum of $\tilde{z}$, taken at angular frequency ω. Stating that z has independent increments basically means that all frequencies of the spectrum of $\tilde{z}$ are asymptotically independent, if the frame is long enough. It is a classical result that when $z(\omega)$ is an isotropic complex Gaussian random measure, $\tilde{z}(t)$ is furthermore stationary. Since audio signals can be considered stationary for the whole duration of each frame, assuming $z(\omega)$ to be an isotropic complex Gaussian is a popular assumption in the audio processing literature (see e.g. [18]).

However, assuming an isotropic complex Gaussian spectral measure is not the only way of guaranteeing that an harmonizable process $\tilde{z}$ is stationary. In particular, a very important result in our context [28, p. 292] is that taking z as an isotropic complex SαS random measure is equivalent to having $\tilde{z}$ being both a stationary and an SαS random process, which is the natural extension of the Gaussian case to $\alpha < 2$. We then model $z(\omega) \sim S\alpha S_c(\sigma_z^\alpha(\omega))$, where σ_z^α is called the fractional power spectral density of $\tilde{z}$ [31], abbreviated α-PSD in the following.

[4]Remarkably, Fig. 1 also suggests to use KL rather than IS for $\alpha = 1$, and IS rather than KL for $\alpha = 2$, as done in the literature.

[5]Since we only consider isotropic complex SαS r.v., we do not linger here on the topic of the so-called "spectral measure" of $\left[v_1^\top\, v_2^\top\right]^\top$, which is important for general SαS multivariate distributions [28, p. 65].

The main interest of the α-harmonizable model is to account for signals that both include large deviations and are stationary. It is thus interesting for audio signals, because they are stationary on short time-frames and often feature large dynamic ranges.

III-D. Separation

Let the J source waveforms $\tilde{s}_1, \ldots, \tilde{s}_J$ defined in section II be modeled as independent α-harmonizable processes. Due to the stability property (4), their mixture $\tilde{x}$ is also α-harmonizable and using (6), we have:

$$x\left(f,n\right) \sim S\alpha S_c\left(\sum_{j=1}^{J} \sigma_j^\alpha\left(f,n\right)\right),$$

where σ_j^α is the α-PSD of source j. Since the α-spectrogram p_j^α defined in section II-A is an estimate of the α-PSD [6], we see that the α-harmonizable model indeed leads to the additivity assumption (1) over the α-spectrograms of the sources.

Now, given $x\left(f,n\right)$ and assuming the α-PSD σ_j^α of the sources are known, is there a way to estimate $s_j\left(f,n\right)$ in order to proceed to source separation? Interestingly, the answer is yes. If $0 < \alpha \leq 2$, and considering that (i) $x\left(f,n\right)$ is the sum of J independent $S\alpha S_c$ r.v. $s_j\left(f,n\right)$ and that (ii) $x\left(f,n\right)$ and $s_j\left(f,n\right)$ are jointly SαS, we have[7]:

$$\mathbb{E}\left[s_j\left(f,n\right) \mid x\left(f,n\right), \left\{\sigma_j^\alpha\right\}_j\right] = \frac{\sigma_j^\alpha\left(f,n\right)}{\sum_{j'} \sigma_{j'}^\alpha\left(f,n\right)} x\left(f,n\right). \quad (8)$$

Equation (3) can thus be interpreted as a practical estimate $\hat{s}_j(f,n)$ of $s_j(f,n)$ given $x(f,n)$, where the α-PSD σ_j^α in equation (8) has been replaced by its estimate p_j^α. We can conclude that for $0 < \alpha \leq 2$, the α-Wiener filter (3) corresponds to estimating the separated sources as their conditional expectation given the mixture x under an α-harmonizable model.

IV. EVALUATION

IV-A. Data and metrics

For evaluating the performance of the proposed α-Wiener filter for source separation, we processed the 8 songs of the QUASI database in the following way:

First, the α-spectrograms p_j^α of the true sources were computed. Then, separation was performed through (3) to obtain the best possible estimates $\hat{s}_j$ under an α-harmonizable model. After this, the resulting waveforms were obtained through an inverse STFT.

For evaluation, all separated sources were split into 30s excerpts, yielding a total of 182 separated source excerpts. The Perceptual Similarity Measure (PSM, from PEMO-Q [15]) was finally used to compare the estimated sources with the true ones, on all the excerpts and for 19 values of α between 0.2 and 2. The PSM lies between 0 (mediocre) to 1 (identical) and is frequently used in assessing audio quality. Results are displayed in Fig. 2.

IV-B. Discussion

As can be noticed in Fig. 2, the α-Wiener filter yields approximately the same performance for $\alpha \in [1, 2]$. This justifies both common practice in the source separation community and the α-harmonizable model that establishes it on solid theoretical grounds for $0 < \alpha \leq 2$. That said, two further remarks may be done here.

[6] Actually, p_j^α as defined in section II should be multiplied by a constant depending only on α, in order to get an asymptotically unbiased estimate of σ_j^α. Even so, it is important to note that this constant would vanish in equation (3). In [31], an asymptotically unbiased and consistent estimator of σ_j^α is proposed, which additionally involves a stage of spectral smoothing.

[7] The proof of this result is available in [1]. It is the natural extension of [28, th. 4.1.2 p. 175] to the isotropic complex $S\alpha S_c$ case, and to the whole range $\alpha \in]0, 2]$.

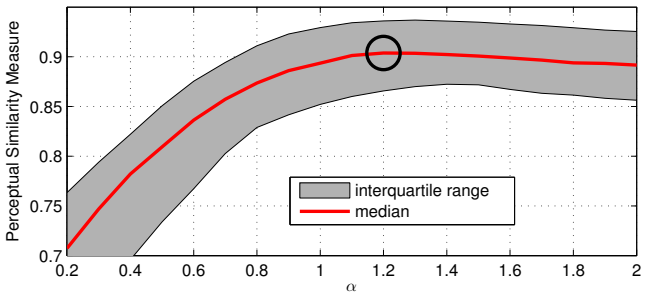

Fig. 2. Distribution of the Perceptual Similarity Measure between the true sources and those obtained by the α-Wiener filter (3), as a function of α. $\alpha = 2$ corresponds to classical Wiener filtering. The best performance is marked with a circle.

First, we see that choosing an α-harmonizable model with $\alpha < 2$ *does* improve the separation performance. In particular, the classical 2-Wiener filter is outperformed in our experiments by an α-Wiener filter with $\alpha \approx 1.2$, even if the improvement is only of a few percents.

Second, these scores correspond to the oracle performance of the method, i.e. when the true α-spectrograms of the sources are known. In real applications, they need to be estimated from the mixture and the additivity assumption (1) is critical for this purpose. Since we saw in section II that (1) is much better verified when $\alpha \approx 1$ than in the Gaussian case, we see that the α-harmonizable model may be advantageous in practice, because it is the only one we know of that justifies both this popular assumption and the resulting filtering procedure (3).

V. CONCLUSION

In a single channel audio source separation context, it is often convenient to assume some linear relationship between the spectrogram of the mixture and the spectrograms of the sources. Identifying the spectrograms of the sources is indeed important to devise soft TF masks used for separation.

When we model the sources as independent and locally wide-sense stationary processes, we have recalled that this assumption is valid for power spectrograms. In that case, a natural TF mask is the classical Wiener filter.

However and as we empirically showed here, assuming the power spectrograms of the sources to add up to form the power spectrogram of the mixture is generally a rough assumption for real audio signals. After introducing the α-spectrogram as the magnitude of the STFT raised to the power $\alpha \in]0\,2]$, we demonstrated that the additivity assumption rather holds for α-spectrograms for some $\alpha < 2$. This fact has already been pointed out by some studies in the dedicated literature.

In this paper, we have modeled the sources as locally stationary α-stable harmonizable processes, abbreviated α-harmonizable, and showed that this naturally leads to the additivity of their α-spectrograms. Furthermore, that probabilistic framework does yield a natural way of separating such signals through a soft TF mask which is analogous to the Wiener filter.

This study could be extended in two main and important directions. First, the case of multichannel mixtures is important for audio processing, because audio signals often come in several channels, as in stereophonic music. Second, this paper was only concerned with the oracle performance of the separation of stationary α-harmonizable processes, i.e. assuming that the true α-spectrograms were known. An interesting question concerns the implications of this model with respect to the blind estimation of the α-spectrograms of the sources when only the mixture is available.

VI. REFERENCES

[1] R. Badeau and A. Liutkus. Proof of Wiener-like linear regression of isotropic complex symmetric alpha-stable random variables. Technical report, September 2014.

[2] D. Barry, B. Lawlor, and E. Coyle. Real-time sound source separation using azimuth discrimination and resynthesis. In *117th Audio Engineering Society (AES) Convention*, San Francisco, CA, USA, October 2004.

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

[4] L. Benaroya, L. McDonagh, F. Bimbot, and R. Gribonval. Non negative sparse representation for Wiener based source separation with a single sensor. In *IEEE International Conference Acoustics Speech Signal Processing (ICASSP)*, pages 613–616, Hong-Kong, April 2003.

[5] A.T. Cemgil, P. Peeling, O. Dikmen, and S. Godsill. Prior structures for time-frequency energy distributions. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 151–154, New Paltz, NY, USA, October 2007.

[6] P. Comon and C. Jutten, editors. *Handbook of Blind Source Separation: Independent Component Analysis and Blind Deconvolution*. Academic Press, 2010.

[7] C. Févotte, N. Bertin, and J.-L. Durrieu. Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. *Neural Computation*, 21(3):793–830, March 2009.

[8] C. Févotte and J.-F. Cardoso. Maximum likelihood approach for blind audio source separation using time-frequency Gaussian models. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 78–81, New Paltz, NY, USA, Oct. 2005.

[9] D. FitzGerald, M. Cranitch, and E. Coyle. On the use of the beta divergence for musical source separation. In *Irish Signals and Systems Conference (ISSC)*, Galway, Ireland, June 2008.

[10] D. Fitzgerald and R. Jaiswal. On the use of masking filters in sound source separation. In *International Conference on Digital Audio Effects, (DAFx-12)*, York, UK, September 2012.

[11] J. Ganseman, G. J. Mysore, J.S. Abel, and P. Scheunders. Source separation by score synthesis. In *International Computer Music Conference (ICMC)*, New York, NY, USA, June 2010.

[12] P. Georgiou, P. Tsakalides, and C. Kyriakakis. Alpha-stable modeling of noise and robust time-delay estimation in the presence of impulsive noise. *IEEE Transactions on Multimedia*, 1(3):291–301, September 1999.

[13] R. Hennequin. *Décomposition de spectrogrammes musicaux informée par des modèles de synthèse spectrale*. PhD thesis, Telecom ParisTech, Paris, France, December 2011.

[14] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson. Singing-voice separation from monaural recordings using robust principal component analysis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 57–60, Kyoto, Japan, March 2012.

[15] R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of auditory perception. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(6):1902 –1911, November 2006.

[16] P. Kidmose. *Blind separation of heavy tail signals*. PhD thesis, Technical University of Denmark, Lyngby, Denmark, 2001.

[17] B. King, C. Févotte, and P. Smaragdis. Optimal cost function and magnitude power for nmf-based speech separation and music interpolation. In *Machine Learning for Signal Processing (MLSP), 2012 IEEE International Workshop on*, pages 1–6. IEEE, 2012.

[18] A. Liutkus, R. Badeau, and G. Richard. Gaussian processes for underdetermined source separation. *IEEE Transactions on Signal Processing*, 59(7):3155 –3167, July 2011.

[19] A. Liutkus, J-L. Durrieu, L. Daudet, and G. Richard. An overview of informed audio source separation. In *Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, pages 1–4, Paris, France, July 2013.

[20] A. Liutkus, D. Fitzgerald, Z. Rafii, B. Pardo, and L. Daudet. Kernel additive models for source separation. *IEEE Transactions on Signal Processing*, 62(16):4298–4310, Aug 2014.

[21] A. Liutkus, Z. Rafii, R. Badeau, B. Pardo, and G. Richard. Adaptive filtering for music/voice separation exploiting the repeating musical structure. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 53–56, Kyoto, Japan, March 2012.

[22] A. Liutkus, Z. Rafii, B. Pardo, D. Fitzgerald, and L. Daudet. Kernel Spectrogram models for source separation. In *Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Nancy, France, May 2014.

[23] G. Miller. Properties of certain symmetric stable distributions. *Journal of Multivariate Analysis*, 8(3):346 – 360, 1978.

[24] C. Nikias and M. Shao. *Signal processing with alpha-stable distributions and applications*. Wiley-Interscience, 1995.

[25] A. Ozerov, E. Vincent, and F. Bimbot. A general flexible framework for the handling of prior information in audio source separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(4):1118–1133, May 2012.

[26] Z. Rafii and B. Pardo. Repeating pattern extraction technique (REPET): A simple method for music/voice separation. *IEEE Transactions on Audio, Speech & Language Processing*, 21(1):71–82, January 2013.

[27] C. Raphael. A classifier-based approach to score-guided source separation of musical audio. *Computer Music Journal*, 32(1):51–59, March 2008.

[28] G. Samoradnitsky and M. Taqqu. *Stable non-Gaussian random processes: stochastic models with infinite variance*, volume 1. CRC Press, 1994.

[29] P. Smaragdis. Separation by humming : User-guided sound extraction from monophonic mixtures. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, New Paltz, NY, USA, October 2009.

[30] P. Smaragdis, C. Févotte, G.J. Mysore, N. Mohammadiha, and M. Hoffman. Static and dynamic source separation using nonnegative factorizations: A unified view. *IEEE Signal Processing Magazine*, 31(3):66–75, May 2014.

[31] G.A. Tsihrintzis, P. Tsakalides, and C.L. Nikias. Spectral methods for stationary harmonizable alpha-stable processes. In *European signal processing conference (EUSIPCO)*, pages 1833–1836, Rhodes, Greece, September 1998.

[32] E. Vincent, S. Araki, F. Theis, G. Nolte, P. Bofill, H. Sawada, A. Ozerov, B. Gowreesunker, D. Lutter, and N. Duong. The signal separation evaluation campaign (2007–2010): Achievements and remaining challenges. *Signal Processing*, 92(8):1928–1936, August 2012.

[33] Y. Wang and D. Wang. Towards scaling up classification-based speech separation. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(7):1381–1390, July 2013.

[34] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Transactions on Signal Processing*, 52(7):1830–1847, July 2004.