Source Separation Based on Non-Negative Matrix Factorization of the Synchrosqueezing Transform

Neha Singh
Jean Kuntzmann Laboratory
University Grenoble Alpes
Grenoble, France
nehairo.iitr@gmail.com

Sylvain Meignen
Jean Kuntzmann Laboratory
University Grenoble Alpes
Grenoble, France
sylvain.meignen@univ-grenoble-alpes.fr

Thomas Oberlin
ISAE-SUPAERO
University Grenoble Alpes
Toulouse, France
Thomas.OBERLIN@isae-supaero.fr

***Abstract*—In this paper, we consider the problem of single channel blind source separation, for which a very common and effective solution consists of applying non-negative matrix factorization to the spectrogram of the mixture. We here propose to replace the spectrogram by the modulus of the synchrosqueezing transform (SST), which achieves a sharper time-frequency representation. Then we introduce two methods for reconstructing the sources, one based on the direct reconstruction from the synchrosqueezed representation, and the other on a two-step procedure based on both the short-time Fourier transform (STFT) and SST, the latter technique being introduced to deal with large signals. Our experiments suggest that non-negative matrix factorization applied to SST enables a better source separation than when applied to the modulus of STFT, and that the proposed two-step procedure using SST and STFT also performs better than the classical technique based on STFT only.**

***Index Terms*—Synchrosqueezing transform, Non negative matrix factorization, Short-time Fourier transform, time-frequency reassignment.**

I. Introduction

Single channel blind source separation is the task of separating a set of sources from a mixed signal without (or very little) information on both the sources and the mixing process. There exists many different techniques to perform this task among which *fast fixed-point independent component analysis algorithms* (FastICA) [1], *principal component analysis* (PCA) [2] and *non-negative matrix factorization* (NMF) [3] are the most popular. Traditionally, when the separation of the sources is carried out with *non-negative matrix factorization* (NMF), the algorithm operates on the *time-frequency representation* (TFR) corresponding to the spectrogram [4], [5]. Such a technique is used in many different domains of applications as for instance in the analysis of electrocardiograms (ECGs) [6], phonocardiograms [7] , audio signals [8] or to separate the sources in a mixture in sound processing [9] and speech enhancement [10].

However, when the sources (or modes) to be separated are close in the *time-frequency* (TF) plane, the spectrogram contains interference the NMF cannot get rid of. To deal with this issue, TF reassignment techniques are therefore often used, in particular in music and speech signals to extract some useful information such as the correct onset, musical transients, the pitch of musical components [11], [12], or the active components in unseen noisy speech [13]. Nevertheless, the reassigned spectrogram does not enable direct source separation because phase information is missing. On the contrary, the *synchrosqueezing transform* (SST), an alternative reassignment technique, does not only perfectly localize the sources in the TF plane but allows for their reconstruction [14], [15].

Our goal in this paper is first to show the potential interest of applying NMF to the modulus of SST rather than to the spectrogram. Indeed, by reassigning first the TFR given by STFT, one obtains a sparser TFR that should enable a better mode separation by means of NMF. We thus define masks based on NMF applied either to the spectrogram or to SST, which we subsequently use for source separation. Nevertheless, such a procedure when applied to SST is not adapted to the separation of the modes of long signals since mode reconstruction is not possible with SST when the hop-size is larger than one. Therefore, to deal with large signals while exploiting the good behavior of NMF with SST, we propose a new two-step procedure based on both SST and STFT for mode separation.

After having recalled, in Section II, the basics on TFR and NMF, we first detail the algorithm for source separation based on NMF applied to the moduli of STFT or SST, in Section III, and then introduce the above mentioned two-step procedure in Section IV. We finally illustrate, in Section V, the benefits of the new proposed approaches.

II. Background

A. Time-Frequency Representations

For a given signal f of length L and a window $w \in [0 : N-1]$, STFT is defined as

$$\mathbf{S}_f^w[n,k] = \sum_{l\in\mathbb{Z}} f[l]w[l-nH]e^{-i2\pi\frac{k(l-nH)}{N}} \qquad (1)$$

where $k \in [0 : N-1]$ is the frequency index, $H \leq N$ the hop-size, N the frequency resolution, and $f[l] = f(\frac{l}{L})$. Assuming

This work was supported in part by the University Grenoble Alpes under IRS Grant "AMUSETE" and the ANR ASCETE project with grant number ANR-19-CE48-0001-01.

the length of w is smaller than N, the signal is traditionally reconstructed through overlap-add (OLA) [16]:

$$f[l] = \frac{\sum\limits_{n\in\mathbb{Z}} w[l-nH]\left(\frac{1}{N}\sum\limits_{k=0}^{N-1}\mathbf{S}_f^w[n,k]e^{i2\pi\frac{k(l-nH)}{N}}\right)}{\sum\limits_{n\in\mathbb{Z}} w[l-nH]^2}. \quad (2)$$

To obtain a sharper *time-frequency representation* (TFR), one can alternatively consider the *synchrosqueezing transform* (SST) [15], [17], which consists of reassigning $\mathbf{S}_f^w[n,k]$ to $[n, \lfloor \hat{\mathbf{m}}_f[n,k]\frac{N}{L}\rceil]$, where $\lfloor X\rceil$ denotes the nearest integer to X and in which:

$$\hat{\mathbf{m}}_f[n,k] = \frac{mL}{N} - \Im\left(\frac{\mathbf{S}_f^{w'}[n,k]}{\mathbf{S}_f^w[n,k]}\right) \quad (3)$$

where w' represents the derivative of window w and $\Im(Z)$ the imaginary part of Z. Note that $\lfloor \hat{\mathbf{m}}_f[n,k]\rceil$ can be viewed as the projection on the frequency grid of the *instantaneous frequency* (IF) evaluated at time $\frac{n}{L}$ and frequency $\frac{kL}{N}$. SST is then defined as follows [15], [18]:

$$\hat{\mathbf{S}}_f^w[n,k] = \sum_{k'\in\mathbb{Z}} \mathbf{S}_f^w[n,k']\delta_{k',\lfloor \hat{\mathbf{m}}_f[n,k]\frac{N}{L}\rceil}, \quad (4)$$

where $\delta_{i,j}$ is the Kronecker symbol. Signal reconstruction is then carried out as:

$$f[n] = \frac{1}{w[0]N}\sum_{k=0}^{N-1}\hat{\mathbf{S}}_f^w[n,k]. \quad (5)$$

B. Non Negative Matrix Factorization

NMF decomposes a given non negative data matrix $\mathbf{X} \in \mathbb{R}^{N\times L}$ into two non negative matrices, the *dictionary matrix* $\mathbf{W} \in \mathbb{R}^{N\times R}$ and the *activation matrix* $\mathbf{H} \in \mathbb{R}^{R\times L}$ such that $\mathbf{X} \approx \mathbf{WH}$. In that context, R stands for the number of components in the dictionary. The decomposition is based on minimizing the reconstruction error of $\mathbf{X}$ through $\mathbf{WH}$, which can be formulated as [19]

$$\min_{\mathbf{W},\mathbf{H}} \quad D(\mathbf{X}|\mathbf{WH}) \quad \text{subject to} \quad \mathbf{W}\geq 0, \mathbf{H}\geq 0. \quad (6)$$

The most popular cost functions D are the Euclidean distance, Kullback-Leibler (KL) divergence and Itakura-Saito (IS) distance, which are special cases of β-divergence defined by

$$D_\beta(\mathbf{A}|\mathbf{B}) = \begin{cases} \frac{\mathbf{A}^{\cdot\beta} - \mathbf{B}^{\cdot\beta} - \beta\mathbf{B}^{\cdot(\beta-1)}\odot(\mathbf{A}-\mathbf{B})}{\beta(\beta-1)}, \beta\in\mathbb{R}\setminus\{0,1\} \\ \mathbf{A}\odot\log(\mathbf{A}\oslash\mathbf{B}) + (\mathbf{A}-\mathbf{B}), \beta = 1 \\ \mathbf{A}\oslash\mathbf{B} - \log(\mathbf{A}\oslash\mathbf{B}) - 1, \beta = 0, \end{cases} \quad (7)$$

where $\odot$ (resp. $\oslash$) stands for the Hadamard product (resp. division), and '.' means entry-wise power. Note that (7) corresponds to the matrix form of the divergence, and to obtain the actual divergence, one has to sum the coefficient of the matrix $D_\beta(\mathbf{A}|\mathbf{B})$. Note that $\beta = 2$ corresponds to Euclidean distance, $\beta = 1$ to KL divergence and $\beta = 0$ to IS divergence. To solve the problem defined in (6), a majorization-minimization algorithm provides multiplicative updates [19] that are widely used [20] and are given by

$$\begin{aligned} &\mathbf{H} \leftarrow \\ &\mathbf{H}\odot\left[\left(\mathbf{W}^T\odot\mathbf{X}\odot(\mathbf{WH})^{\cdot(\beta-2)}\right)\oslash\left(\mathbf{W}^T\odot(\mathbf{WH})^{\cdot(\beta-1)}\right)\right] \\ &\mathbf{W} \leftarrow \\ &\mathbf{W}\odot\left[\left((\mathbf{WH})^{\cdot(\beta-2)}\odot\mathbf{X}\odot\mathbf{H}^T\right)\oslash\left((\mathbf{WH})^{\cdot(\beta-1)}\odot\mathbf{H}^T\right)\right] \end{aligned} \quad (8)$$

III. Source Separation from SST

We first propose here to apply NMF to SST modulus rather than STFT modulus. This helps us define TF masks which we use for source separation exploiting the fact that SST is invertible. We detail here the case when the hop-size $H = 1$ for which source separation is straightforward. Indeed, in that case the SST is invertible from equation (5). One thus only needs to compute *soft masks* for each individual source, which can be seen as a Wiener filtering. First, individual source STFT or SST moduli are estimated through NMF and then used to make the corresponding soft masks, which are multiplied point-wise with signal STFT or SST to obtain the TFR of individual sources. This can be formally described by:

$$\mathbf{S}_k = \mathbf{M}_k\odot\mathbf{S}, \quad (9)$$

where $\mathbf{S}_k$ represents the estimated STFT or SST of the k^{th} source, the soft mask $\mathbf{M}_k$ being defined by:

$$\mathbf{M}_k = \frac{\mathbf{X}_k}{\sum\limits_{r=1}^{R}\mathbf{X}_r}, \quad (10)$$

where $\mathbf{X}_k$ corresponds to the k^{th} source in NMF decomposition, namely $\mathbf{W}_{:,k}\mathbf{H}_{k,:}$. Having defined $\mathbf{S}_k$, one reconstructs the modes using either STFT or SST by replacing $\mathbf{S}_f^w$ by $\mathbf{S}_k$ in (2) of in (4) respectively. In what follows, we call STFT-NMF and SST-NMF the reconstruction processes based on NMF applied to STFT and SST, respectively.

IV. Source Separation from STFT and SST

Though the previous approach is interesting, it is somewhat limited because reconstruction with SST is not tractable when the hop-size is larger than 1, and is therefore not adapted to the processing of large signals. For that purpose, we now introduce a novel source separation procedure that exploits the nice properties of SST while circumventing the limitation regarding the hop-size. When $H > 1$, source recovery is not straightforward, since SST is no longer invertible. Yet, this case is of particular interest in audio processing where one often has to deal with long signals.

To circumvent this limitation, we here introduce a two-step approach: we first apply NMF to the SST and only keep the corresponding activation matrix $\mathbf{H}$, and then recompute a dictionary matrix $\mathbf{W}$ from the spectrogram. With these new matrices, we are able to build soft masks following (9), and

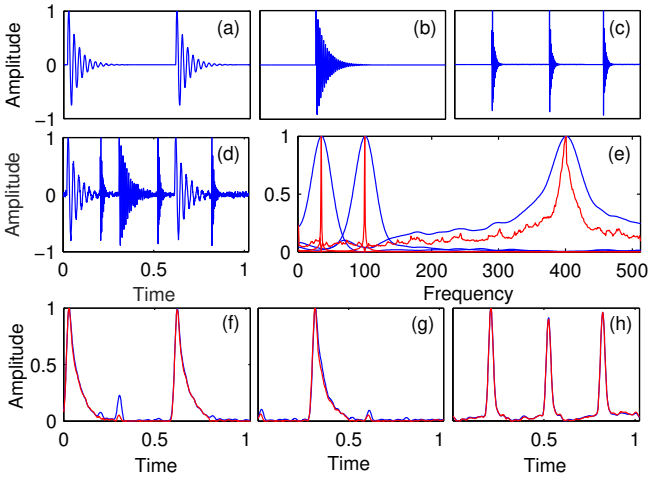

Fig. 1. (a)-(c): Time representation of three different modes; (d): Signal made of the sum of these modes (input SNR 20 dB); (e): First three columns of matrix $\mathbf{W}$ corresponding to SST-NMF and STFT-NMF (the signal is assumed to contain three modes) (f)-(h): rows of the activation matrices corresponding to these methods

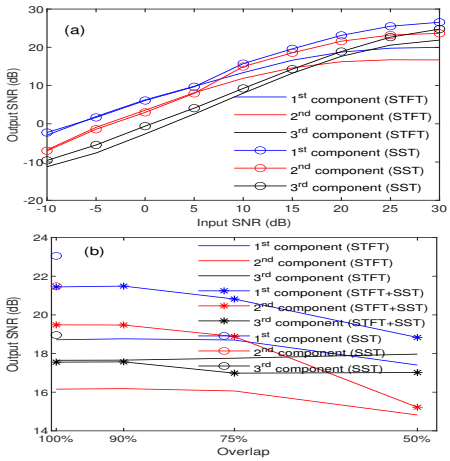

Fig. 2. (a): Output SNR corresponding to the reconstruction of the individual sources using either SST-NMF or STFT-NMF (denoted by STFT and SST in the graph) for the signal of Fig. 1 (d) and for different input SNRs;(b): Output SNR corresponding to the reconstruction of individual sources using either STFT-NMF or STFT+SST-NMF (denoted by STFT and STFT+SST on that figure), for the signal of Fig. 1 (d), for input SNR 20 dB and when the percentage of overlap varies, the results are averaged over 20 realizations

then proceed with source separation. The just mentioned two-step procedure formally reads:

$$
\begin{aligned}
(\mathbf{W}, \mathbf{H}) &= \mathrm{NMF}(|\hat{\mathbf{S}}_f^w|) \\
\mathbf{W}' &= \mathrm{NMF}'(|\mathbf{S}_f^w|, \mathbf{H}) \\
\mathbf{X}_k &= \mathbf{W}'_{:,k}\mathbf{H}_{k,:},
\end{aligned}
$$

where NMF' denotes the NMF where $\mathbf{H}$ is fixed.

The rationale behind this procedure is the following: as SST is sparser than STFT, it should enable better source separation, thus providing more accurate activation and dictionary matrices. But these cannot be directly used to recover the sources since SST is no longer invertible in that case. To circumvent this limitation, we recompute a dictionary from NMF applied to STFT modulus, and in which the activation matrix corresponds to that of NMF based on SST. Such an activation matrix is supposed to be consistent with STFT since SST reassigns harmonic signals to the local maxima of the spectrogram along the frequency axis. The last step of the algorithm is a simple constrained problem, which should converge faster than NMF and is convex if the divergence is convex. In what follows, the technique is denoted by STFT+SST-NMF.

V. Numerical Results

In this section, we investigate the behavior of the source separation procedures introduced above when these are applied to either synthetic or real signals.

A. Application to Synthetic Drum Sound Signals

We first consider a simple synthetic signal made to closely mimic drum sound signals: its components have exponential decay, correspond to different frequencies and have different time durations. We display the modes making up such a signal in Fig. 1 (a), (b) and (c), associated with respective frequencies 35 Hz, 100 Hz and 400 Hz. Finally, summing these components and adding some white Gaussian noise with input SNR equal to 20 dB results in the signal displayed in Fig. 1 (d). Then, we perform STFT-NMF and SST-NMF on that signal. The three columns of matrix $\mathbf{W}$ obtained with both techniques are shown in Fig. 1 (e), highlighting a much more peaky representation with SST-NMF than with STFT-NMF. In Fig. 1 (f)-(h) we finally display the rows of the activation matrix in both cases, and notice that mode-mixing is also visible on that related to STFT-NMF.

Then to investigate the influence of the noise on source separation, we compute the output SNRs corresponding to the reconstruction of each source making up the signal of Fig. 1 (d) and for different input SNRs. The results depicted in Fig. 2 (a), plead in favor of using SST-NMF rather than STFT-NMF when the SNR is medium to high, while the two techniques behave similarly when the noise level increases. It is worth also noting that the improvement brought by SST-NMF is less obvious for the third component of the signal, suggesting that the sparsity of matrix $\mathbf{W}$ is important to improve source separation.

We also investigate the performance of the two-step procedure called STFT+SST-NMF introduced in section IV for $H > 1$ by comparing it with STFT-NMF, when the hop-size varies and for the signal of Fig. 1 (d). As the hop-size is relative to the window length, we prefer to display, in Fig. 2 (b), the reconstruction results with respect to the percentage of overlap. To plot that figure, we consider an input SNR of 20 dB, so that an easy comparison can be made with the study of the case $H = 1$ displayed in Fig. 2 (a). In this regard, along with the results corresponding to STFT-NMF and STFT+SST-NMF we plot the results obtained with SST-NMF in the case $H = 1$ (i.e. 100 % overlap). Looking at the results, we notice

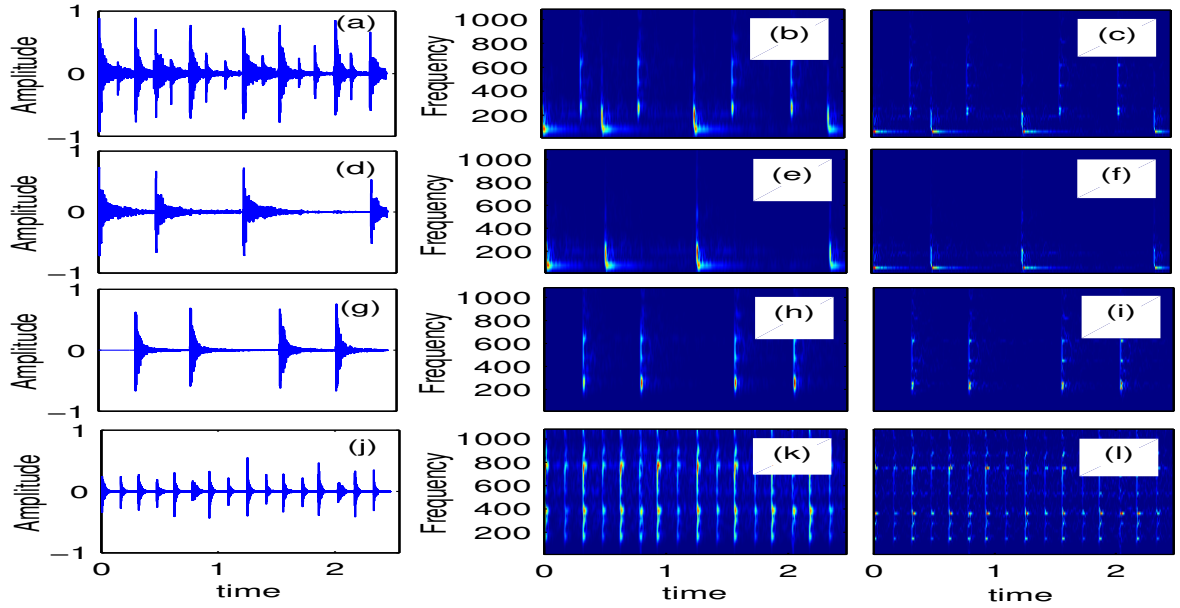

Fig. 3. first column: Mixture signal of KD, SD and HH components (a) with KD, SD and HH components displayed in (d),(g) and (h); second column : STFT of the signals of the first column; third column: SST of the signals of the first column

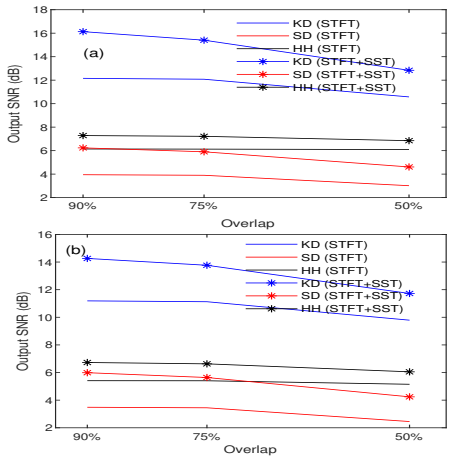

Fig. 4. Output SNR corresponding to the reconstruction of KD, SD and HH components averaged over 60 drum loops with varying percentage overlaps for the analysis window; (a) without noise; (b) for input SNR 20 dB; and when the percentage of overlap varies, the results are averaged over 20 realizations

that while STFT+SST-NMF behaves slightly worse than SST-NMF when $H = 1$, it remains much better than STFT-NMF for a wide range of overlap values. Again we notice, that in accordance with the results of Fig. 2 (a), STFT+SST-NMF does not perform any better than STFT-NMF on the third signal.

B. Performance Evaluation on Drum Source Separation (DSS)

We now evaluate the performance of STFT-NMF and STFT+SST-NMF for DSS. The signals are sampled at 44.1 Kz and the STFT is performed with the same Gaussian window with length 1024 with width parameter $\sigma = 0.1s$. The running examples considered for this task are synthetic drum sounds of a Roland TR-808 drum machine, of which an example is shown in Fig. 3 (a). Such signals are composed of 3 different types of components called kick drum (KD), snare drum (SD) and hi-hat (HH) with very different time and frequency behaviors (Fig.3, second and third columns).

The experiments are conducted on publicly available "Wavedrum02" subset of "IDMT-SMT-DRUMS" dataset [21]. It consists of 60 drum loops of each KD, SD and HH (oracle) drum components in uncompressed 16-bit mono PCM WAV format with 44.1 KHz sampling rate, and the corresponding 60 mixture signals. The advantage of using this data-set is that it mimics real world break-beats and also that the ground truth is available for the individual components.

As previously, we apply STFT-NMF and STFT+SST-NMF techniques to the 60 "Wavedrum02" drum loops of "IDMT-SMT-DRUMS" dataset. For the first experiment we investigate the decomposition of the noiseless signals and compute the output SNRs for the three components, i.e. KD, SD and HH for all the 60 drum loops, and finally compute the averaged output SNR for all the three components. Observing the third column of Fig. 3, SST leading to a highly concentrated TFR for KD and SD components, it is quite expected to have better source separation using STFT+SST-NMF as compared to STFT-NMF for these components. This is confirmed in Fig. 4 (a), in which we also notice that this remains true when the percentage of overlap varies. For the third component, namely HH, the benefit of using STFT+SST-NMF rather than STFT-NMF is less obvious. For the second experiment, we carry out the comparison between STFT-NMF and STFT still on the signal of Fig. 3 (a), but when some noise is added (20 dB input SNR). The results depicted in Fig. 4 (b) show that

the benefit of using the proposed STFT+SST-NMF instead of STFT-NMF is even greater in that case than in the noiseless case.

VI. CONCLUSION

In this paper, NMF based on STFT and SST were first compared. We noticed that the latter provides more concentrated dictionaries and observed that the overlapping between components is also less important with SST-based NMF than with its counterpart based on STFT. Performing source separation using soft masks built from SST-based NMF then proved to be more relevant than the same approach based on STFT. To study long signals for which SST-based NMF is not relevant, since SST is not invertible, we proposed a novel approach we stamped STFT+SST-NMF, which proved to outperform NMF based on STFT for the separation of the sources in real and synthetic drum signals. In a near future, the behavior of different variants of NMF such as non-negative matrix factor deconvolution (NMFD) should be investigated and that of NMF on SST variants should also be clarified.

REFERENCES

[1] Erkki Oja and Zhijian Yuan, "The fastica algorithm revisited: Convergence analysis," *IEEE transactions on Neural Networks*, vol. 17, no. 6, pp. 1370–1381, 2006.

[2] Juha Karhunen, Petteri Pajunen, and Erkki Oja, "The nonlinear pca criterion in blind source separation: Relations with other approaches," *Neurocomputing*, vol. 22, no. 1-3, pp. 5–20, 1998.

[3] Felix Weninger, Jonathan Le Roux, John R Hershey, and Shinji Watanabe, "Discriminative nmf and its application to single-channel source separation," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[4] C. Dittmar and M. Mller, "Reverse engineering the amen break score-informed separation and restoration applied to drum recordings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 1535–1547, 2016.

[5] Sean U.N. Wood and Jean Rouat, "Blind speech separation with GCC-NMF," in *Interspeech 2016*, 2016, pp. 3329–3333.

[6] Pengju He and Xiaomeng Chen, "A method for extracting fetal ecg based on emd-nmf single channel blind source separation algorithm," *Technology and Health Care*, vol. 24, no. s1, pp. S17–S26, 2016.

[7] D. Pham, S. Meignen, N. Dia, J. Fontecave-Jallon, and B. Rivet, "Phonocardiogram signal denoising based on nonnegative matrix factorization and adaptive contour representation computation," *IEEE Signal Processing Letters*, vol. 25, no. 10, pp. 1475–1479, 2018.

[8] A. Ozerov and C. Fevotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 3, pp. 550–563, 2010.

[9] Christian Dittmar and Daniel Gärtner, "Real-time transcription and separation of drum recordings based on NMF decomposition.," in *DAFx*, 2014, pp. 187–194.

[10] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4029–4032.

[11] Stephen W Hainsworth and Patrick J Wolfe, "Time-frequency reassignment for music analysis," in *ICMC*. Citeseer, 2001.

[12] S. W. Hainsworth, M. D. Macleod, and P. J. Wolfe, "Analysis of reassigned spectrograms for musical transcription," in *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics*, 2001, pp. 23–26.

[13] Maarten Van Segbroeck and Hugo Van_hamme, "Applying non-negative matrix factorization on time-frequency reassignment spectra for missing data mask estimation," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[14] Gaurav Thakur, Eugene Brevdo, Neven S. FučKar, and Hau-Tieng Wu, "The synchrosqueezing algorithm for time-varying spectral analysis: Robustness properties and new paleoclimate applications," *Signal Processing*, vol. 93, no. 5, pp. 1079–1094, May 2013.

[15] T. Oberlin, S. Meignen, and V. Perrier, "The Fourier-based synchrosqueezing transform," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 315–319.

[16] Sylvain Meignen and Duong-Hung Pham, "Retrieval of the modes of multicomponent signals from downsampled short-time fourier transform," *IEEE Transactions on Signal Processing*, vol. 66, no. 23, pp. 6204–6215, 2018.

[17] Ingrid Daubechies, Jianfeng Lu, and Hau-Tieng Wu, "Synchrosqueezed wavelet transforms: an empirical mode decomposition-like tool," *Applied and Computational Harmonic Analysis*, vol. 30, no. 2, pp. 243–261, 2011.

[18] Gaurav Thakur and Hau-Tieng Wu, "Synchrosqueezing-based recovery of instantaneous frequency from nonuniform samples.," *SIAM J. Math. Analysis*, vol. 43, no. 5, pp. 2078–2095, 2011.

[19] Cédric Févotte and Jérôme Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[20] V. Leplat, N. Gillis, and A. M. S. Ang, "Blind audio source separation with minimum-volume beta-divergence NMF," *IEEE Transactions on Signal Processing*, vol. 68, pp. 3400–3410, 2020.

[21] *https://www.idmt.fraunhofer.de/en/business*$_u nits/m2d/smt/drums.html$.