Available at https://jtsiskom.undip.ac.id (7 December 2020)
DOI:10.14710/jtsiskom.2020.13858

Jurnal Teknologi dan Sistem Komputer, 9(1), 2021, 41-48

Sampled and discretized of short-time Fourier transform and non-negative matrix factorization: the single-channel source separation case

Jans Hendry*), Isnan Nur Rifai, Yoga Mileniandi

Department of Electrical Engineering and Informatics, Vocational College, Universitas Gadjah Mada
Yacaranda st., Sekip Unit IV, Yogyakarta, Indonesia 55281

How to cite: J. Hendry, I. N. Rifai, and Y. Mileniandi, "Sampled and discretized of short-time Fourier transform and non-negative matrix factorization: the single-channel source separation case," *Jurnal Teknologi dan Sistem Komputer*, vol. 9, no. 1, pp. 41-48, 2021. doi: 10.14710/jtsiskom.2020.13858, [online].

Abstract – The Short-time Fourier transform (STFT) is a popular time-frequency representation in many source separation problems. In this work, the sampled and discretized version of Discrete Gabor Transform (DGT) is proposed to replace STFT within the single-channel source separation problem of the Non-negative Matrix Factorization (NMF) framework. The result shows that NMF-DGT is better than NMF-STFT according to Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR), and Signal-to-Distortion Ratio (SDR). In the supervised scheme, NMF-DGT has a SIR of 18.60 dB compared to 16.24 dB in NMF-STFT, SAR of 13.77 dB to 13.69 dB, and SDR of 12.45 dB to 11.16 dB. In the unsupervised scheme, NMF-DGT has a SIR of 0.40 dB compared to 0.27 dB by NMF-STFT, SAR of -10.21 dB to -10.36 dB, and SDR of -15.01 dB to -15.23 dB.

Keywords *– DGT; STFT; NMF; single-channel source separation; time-frequency representation*

I. INTRODUCTION

The source separation problem is an effort to extract interests from mixtures of time-varying data sequences. In the signal processing field, this kind of data is called signals. The final goal of extraction can be any type, like signals filtering and signals separation. The researchers have found and proposed many solutions to this problem from various approaches. Moreover, they are exploited to solve multiple issues in different fields like medical issues [1]-[3], robotics [4], [5], fault monitoring [6], imaging [7]-[9], and even uncountably in the field of speech processing.

The non-Negative Matrix Factorization (NMF) is a frequently used method in source separation like music and speech for single-channel source separation. The single-channel source separation problem is to extract or separate mixed sources from just one mixture. The NMF can approximate the power spectrum of a single-channel mixture as the product of two non-negative matrices (every matrices element is prohibited from containing negative values). The signal-source mixture is then decomposed into its constituents by using the Wiener filter and reusing its original phase [10], [11]. The NMF utilizes complex Time-Frequency representation in the process, usually Short-time Fourier Transform (STFT). The STFT is a Fourier transform in instantaneous time that uses a window to smooth the transition flow of frames. The representation of STFT is a matrix filled with complex elements. These elements can give enormous information like power spectrum (magnitudes), instantaneous phases, and instantaneous frequencies. The STFT is utilized in many new source separation methods and can be divided into three categories: underdetermined cases [12]-[16], convolutive cases [17], [18], and overdetermined cases [19], [20].

The STFT is a fully redundant time-frequency representation as it has a window that translates frame by frame. In order to avoid this redundancy, the under-sampling can be applied. As a result, the sampled and discretized version of STFT is constructed. Furthermore, this version is called Discrete Gabor transform (DGT). Similar to STFT, the DGT is also invertible. The invertibility is essential to reconstruct the separated sources or signals into their time representation.

The DGT is another representation of time-frequency used in this work to replace widely used representation, STFT. According to our knowledge, the DGT and its reciprocal is rarely used in a single-channel source separation problem, especially in NMF. Hence, this research aims to examine the performance of DGT being implemented in the single-channel source separation. It also benefits from its capability to avoid redundancies in representing a signal. Furthermore, speech signals are used herein to evaluate the system as they have broad frequency bins compare to merely monochrome signal (single frequency signal).

II. RESEARCH METHODS

It is indispensable to use Time-Frequency (T-F) analysis to localize information, especially in the short duration of oscillation signals. In NMF-based single-channel source separation, T-F is utilized to calculate the basis vector and activation matrix. A mixture is created from two clean speeches. T-F matrix is calculated from a mixture using DGT instead of STFT.

*) Corresponding author (Jans Hendry)
Email: jans.hendry@ugm.ac.id

Submitted: 5 August 2020; Revised: 27 October 2020; Accepted: 27 November 2020; Published: 31 January 2021

The basis vector and dictionary are calculated using NMF. A mask is created from a basis vector and a dictionary. Wiener filter and the mask are used to get the power of each component of the mixture. Their phases are restored and the DGT is inverted to get the time-domain components.

A. Discrete Gabor transform

Assume *x(t)* denotes a real continuous signal and transformed with STFT, as expressed in Eq. 1 [21]. It yields T-F representation in radian frequency vs time, $X(\omega,t) \in \mathbb{C}$. The window $\overline{g()}$ is a complex conjugate of translated-fixed window function with $g(\tau) \neq 0$. Eq. 1 can also be defined in linear frequency, *f*, as shown in Eq. 2 by merely replacing ω with $2\pi f$.

$$X(\omega,t)=\int_{-\infty}^{\infty} x(\tau)e^{-i\omega\tau}\overline{g(\tau-t)}\,d\tau \tag{1}$$

$$X(f,t)=\int_{-\infty}^{\infty} x(\tau)e^{-i2\pi ft}\overline{g(\tau-t)}\,d\tau \tag{2}$$

To analyze using computing machines upon oscillating signals like sounds and speeches, they must be discretized (sampled and quantized). The T-F analysis method should also be in discrete form. Hence, in discrete form, STFT can be defined as is given by Eq. 3 [21]. *K* denotes wide of the signal, and *n* is time-discrete, $K, n \in \mathbb{Z}$. The final product of STFT will be completely redundant as a result of window translation. One way to avoid redundancy is by reducing the number of points involved in the calculation process. It is naturally the Gabor transform principle. Gabor transformation for a continuous signal is expressed in Eq. 4.

$$X[m,n]=\sum_{-\infty}^{\infty} x(k)e^{-i2\pi mk/K}\overline{g(k-n)} \tag{3}$$

$$X(m,n)=\int_{-\infty}^{\infty} x(\tau)e^{-i2\pi mbt}\overline{g(\tau-an)}\,d\tau \tag{4}$$

In the discrete form, this formula turns to be Eq. 5. The Discrete Gabor coefficients is a complex matrix, $\boldsymbol{X}[m, n] \in \mathbb{C}^{MxN}$, while window coefficients and signal sequence can be (not necessarily in) complex, *x(k)*, $g(k) \in \mathbb{C}^{L}$ [21].

$$X[m,n]=\sum_{l=0}^{L-1} x(k)e^{-i2\pi bmk/K}\overline{g(k-an)} \tag{5}$$

T-F shift parameters are denoted with $a, b > 0$, which is a hop factor in time and frequency, respectively. While *g(k)* is a fixed window function, $m = 0, \ldots, M-1$ and $n = 0, \ldots, N-1$, $m, n \in \mathbb{Z}$ of which $M = L/b$ and $N = L/a$ represent the number of channels and number of time shifts, respectively. The length of the signal *L* should be divisible to *a* and *b*, and zero paddings are commonly used to fulfill this condition. Eq. 5 also describes how DGT is a sort of *sampled and discretized* version of STFT. As such, the signal should be finite with periodic boundaries [21]-[23].

B. Inversion of discrete Gabor transform

Similar to STFT, DGT is also invertible. The reciprocal guarantees that the analyzed signal can be synthesized to a time-domain signal with a significantly small error. The inverse of DGT (IDGT) is expressed in Eq. 6 where $\Omega(k)$ denotes the dual window of $g(k)$ [21].

$$x[k]=\sum_{n=0}^{N-1}\sum_{m=0}^{M-1} X[m,n]\Omega[k-an]e^{i2\pi bmk/K} \tag{6}$$

The STFT can be considered superior when dealing with inversion. One can use any window in the reconstruction phase. Meanwhile, the synthesis window in IDGT is restricted to appropriateness with respect to the analysis window.

C. Single-channel source separation problem

Source separation has been one of the popular topics in signal processing. It is intended to separate a mixture of signals or sources. Assume there are P sources and Q sensors in hand. The separation problem can then be classified into three groups, which are underdetermined (P>Q), determined (P=Q), and overdetermined (P<Q). The mixtures can be considered as mathematical entanglement of each source. The way they are mathematically entangled is classified into instantaneous, anechoic, and convolutive [24].

In this work, the determined problem with a linearly mixed-instantaneous model has been used, of which two speech signals were mixed to form a single mixture. Using this signal as the data test is because of its multiple spectrum frequency and closely spaced compared to musical instruments notes. In other words, musical instruments' notes can be considered sharper than what speech has. Further, the framework (conventional NMF) used in the test was chosen as simple as possible. Hence, it can help to firmly conclude which is the most superior to the other (DGT or STFT) in a basic implementation.

In a nutshell, a mixture of signals used in this work can be described by Eq. 7 where $P = Q$, $\boldsymbol{A}_{PxQ}$ is a mixing matrix $\in \mathbb{R}^{PxQ}$, *x(t)* is the mixtures of sources $\in \mathbb{R}^{PxL}$, *s(t)* is the sources $\in \mathbb{R}^{QxL}$, and *L* is the signal length [12]. The mixing matrix was generated randomly for normal distribution. In the end, this matrix would not be discovered as it was not the aim of the NMF algorithm. Eq. 7 can be rewritten in matrix form as shown by Eq. 8 with 2x2 mixing matrix. Solely one arbitrary mixture is used for the separation.

$$x(t)=A_{P\times Q}s(t) \tag{7}$$

$$X(f,t)=\int_{-\infty}^{\infty} x(\tau)e^{-i2\pi ft}\overline{g(\tau-t)}\,d\tau \tag{8}$$

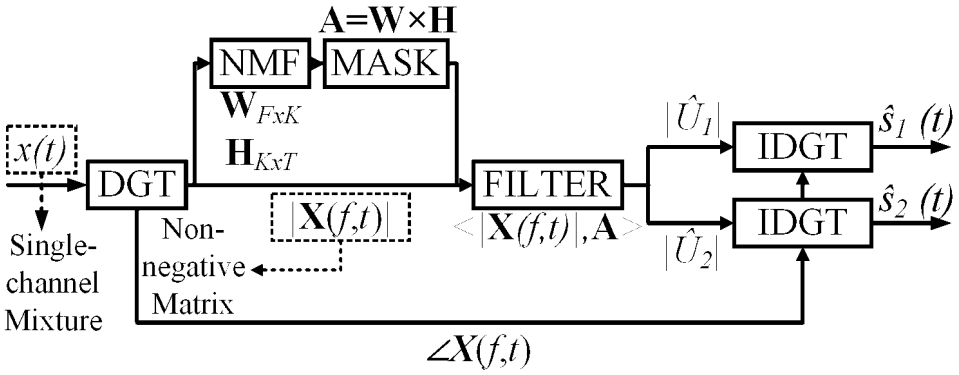


Figure 1. Single-channel separation process diagram for 2 x 2 mixing matrix

D. Non-negative matrix factorization

In the NMF framework, a set of mixtures can be expressed in terms of their decomposition as shown in Eq. 9. **X** is the mixtures, **W** is the basis vector or basis spectral or dictionary, and **H** is the activation matrix or weight matrix. All elements of these matrices are real and positive. The NMF does not require the components to be statistically independent, as ICA does. This method recovers the sources without pre-knowledge about estimating the first and second moments (mean and variance, respectively) of the sources like in ICA. However, this method cannot deliver a unique solution like most ICA-based solutions [25]-[28].

$$X \stackrel{def}{=} WH \; ; X \in R_{+}^{FxT}, W \in R_{+}^{FxK}, H \in R_{+}^{KxT} \qquad (9)$$

Figure 1 shows the block diagram of how the single-channel source separation is conducted. The NMF depends only on the power spectrum of instantaneous T-F coefficients. This method is sensitive to the initial value, **W**. Based on how the initial value is set up, NMF can be divided into two classes: supervised and unsupervised. The difference between them is how the initial value of the basis spectral is calculated. The supervised NMF requires an initial basis spectral built from real speech signal while the initial dictionaries for unsupervised NMF are generated randomly. The supervised NMF often outperforms unsupervised because of this initial value. The NMF is also said to be non-convex and capable of finding local optimum only.

Based on how the initial value is set up, NMF can be divided into two classes, namely supervised and unsupervised. The difference between them is how the initial value of basis spectral is calculated. The supervised NMF requires an initial basis spectral built from real speech signal while the initial dictionaries for unsupervised NMF are generated randomly. The supervised NMF often outperforms unsupervised because of this initial value. The NMF is also said to be non-convex and capable of finding local optimum only.

There are various types of NMF especially related to how the divergence between **WH** and **X** is calculated, $D(\mathbf{V}||\mathbf{WH})$, like Euclidean and Kullback-Leibler (KL). In this work, we simplify the NMF to use ordinary iterations merely. Algorithm 1 elaborates every single step of the diagram block above. Algorithm 2 explains how the basis spectral and vector weight are updated.

Algorithm 1. Single-channel source separation

Input: *x(t)* ; *single-channel mixture*
K ; *K-selected basis*
Num_Sources
Output: $\hat{s}_1(t) \dots \hat{s}_P(t)$; *separated sources*

X(t,f) ← DGT(*x(t)*) ; *T-F representation*
$\ddot{\mathbf{X}}$ ← |**X**(t,f)| ; *non-negative coeff.*
[**W, H**] ← NMF(**X**(t,f)) ; *final value of* **W**, **H**
θ ← *angle*(**X**(t,f)) ; *mixture's phase*

iterate Num_Sources:
Mask ← **W**(K) * **H**(K)
S ← $\ddot{\mathbf{X}}$.* Mask ; *filter*
$\hat{U}_1 \dots \hat{U}_P$ ← S .* $e^{j\theta}$
$\hat{s}_1(t) \dots \hat{s}_P(t)$ ← IDGT($\hat{U}_P$)

Algorithm 2. NMF update

Input: W, **H** ; *Initial values*
Output: W, **H** ; *Final values*

iterate MAX_ITER:

$$\mathbf{H} \leftarrow \mathbf{H} .* \frac{W^{'} \frac{\ddot{X}}{W H + eps(1)}}{W^{'} 1_{F \times T}}$$

$$\mathbf{W} \leftarrow \mathbf{W} .* \frac{\frac{\ddot{X}}{W H + eps(1)} H^{'}}{1_{F \times T} H^{'}}$$

The benefit of NMF is it is guaranteed to be always in convergence. But the number of selected basis spectral, *K*, has to be defined prior to separation, typically $K < F < T$. The selection of *K* indicates the use of a low-rank matrix in NMF. This method also requires the knowledge of the number of sources, which is often unknown. Also, it is hard to get real-world signals to lack thereof of noise and reverberances in real practices. Hence, additional complementary algorithms are

inevitable when dealing with this kind of signal. In our work, the signals used for evaluation were well provided, of which the noise and reverberances have been filtered out.

E. Materials

The machine and software specifications used to simulate the whole process here are described in Table 1. The speech signals were taken from TIMIT Corpus [29]. The STFT and NMF program was built on our own. Meanwhile, the DGT program was a combination of our wrapper and the one accessed from the source [21].

F. Performance evaluation

The single-channel source separation performance utilizing DGT was evaluated with three popular ratios of interest, namely Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR), and the total error by Signal-to-Distortion Ratio (SDR) [30], [31] (Eq. 10 – Eq. 12). In this work, the performance of DGT was compared to STFT in the same framework of NMF.

$$SIR = 10\log_{10}\frac{\left\|s_{target}(t)\right\|^2}{\left\|e_{interf}(t)\right\|^2} \quad (10)$$

$$SAR = 10\log_{10}\frac{\left\|s_{target}(t)+e_{interf}(t)\right\|^2}{\left\|e_{artif}(t)\right\|^2} \quad (11)$$

$$SDR = 10\log_{10}\frac{\left\|s_{target}(t)\right\|^2}{\left\|e_{interf}(t)+e_{artif}(t)\right\|^2} \quad (12)$$

The first three parameters can be calculated by Eq. 13 – Eq. 15. $s_{target}(t)$ denotes the target signal, e_{interf} is the error due to interferences, e_{artif} is the error due to artifacts, and $\|.\|^2$ is the squared 2-norm. $\{s_i(t)\}$ denotes the original anechoic signals, $s_j(t)$ is the anechoic target signal, and $\hat{s}_j(t)$ is the estimated target signal. While the $P(.)$ denotes projection operator. Higher values of all three ratios show better performance.

$$s_{target}(t) = P(\hat{s}_j(t), s_j(t)) \quad (13)$$

$$e_{interf}(t) = P\left(\hat{s}_j(t), \left\{s_i(t)\right\}\right) - P(\hat{s}_j(t), s_j(t)) \quad (14)$$

$$e_{artif}(t) = \hat{s}_j(t) - P\left(\hat{s}_j(t), \left\{s_i(t)\right\}\right) \quad (15)$$

III. Results and Discussion

The matrix used to mix speeches was randomly generated. The speeches were resampled to 8 kHz with a duration of around 2 seconds. The speeches are anechoic and noise-free. If there is no initial value provided for basis spectral (dictionary), then a random number generator is used for **W** and **H**. The sample of the mixture is shown in Figure 2.

In this work, generating a random number is an iterative process. They were iterated about 50,000 times and averaged. The **W** and **H** update inside NMF's loop

Table 1. Machine specifications

Materials	Specification
Matlab®	R2018b
CPU	Intel® I5 2.5 GHz
RAM	12 GB

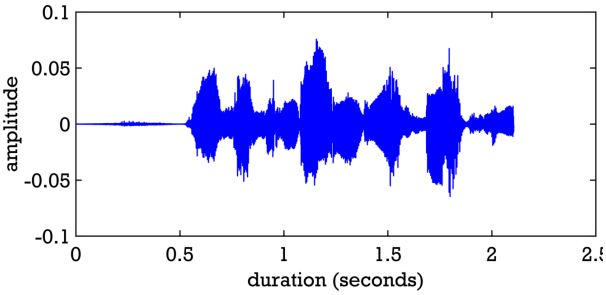


Figure 2. The mixture

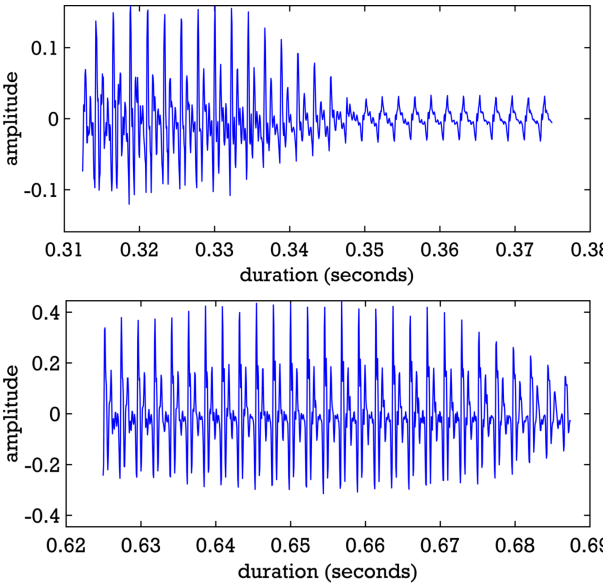


Figure 3. First and second original speeches

was iterated 200 times. While the separation process (the outer loop) was also iterated 1,000 times. It means there is 1,000 x (50,000 + 200) iterations in total. The purpose of this huge iteration is to support reliable decisions. The number of frequency bins, window size, window type, and overlap factor are 1024, 512, Hanning, and 50 %, respectively.

A. Supervised NMF

In this scheme, two speeches are used to build the initial dictionary, **W**, as shown in Figure 3. Only 1000 samples are shown to have better visualization. The goal is to estimate **H** and calculate **W** update. The amount of basis vector, *K*, used was 25 for every source.

Figure 4 shows the approach of the original T-F representation (matrix) of a mixture. The T-F matrix is complex values calculated using DGT. Another term for this matrix is the spectrogram. The spectrogram

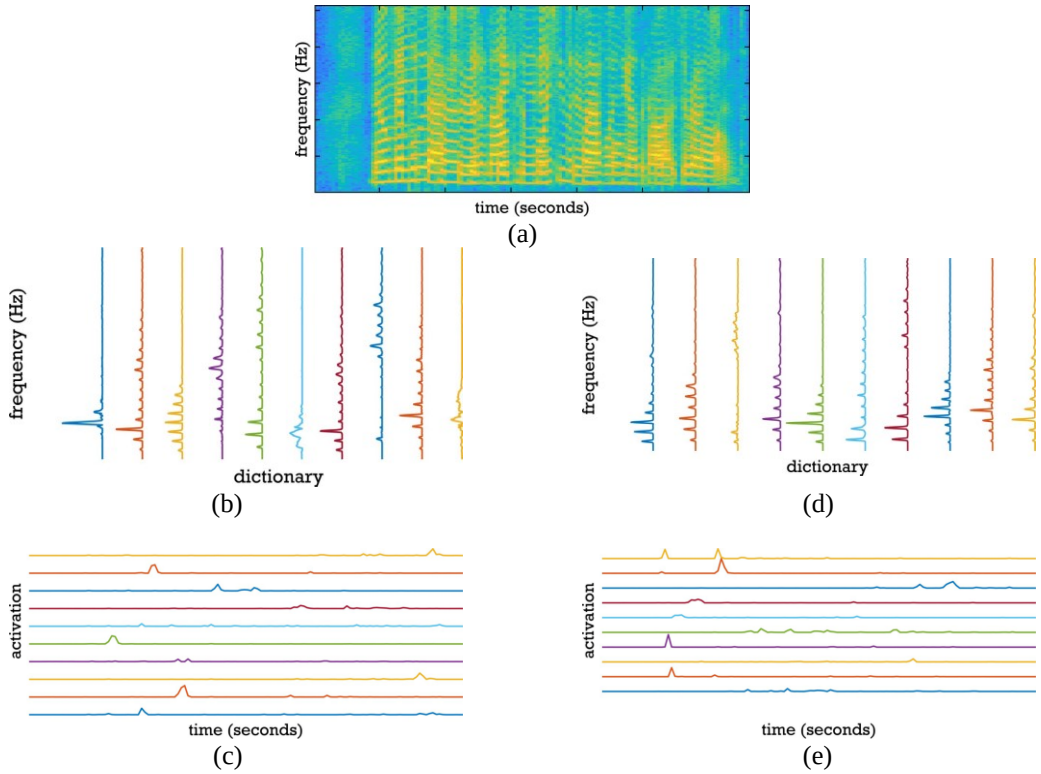


Figure 4. The dictionaries and activations of supervised NMF: (a) mixture's spectrogram, (b)-(c) first source's dictionary and activation, (d)-(e) second source's dictionary and activation

frequently calculated using STFT, but in this work, it is shown how DGT can be used to yield a spectrogram. This spectrogram tells frequencies distribution in every time frame. Yellowish color depicts the presence of particular frequencies.

As DGT employs Fourier transform in its process as well as STFT, the dictionary matrices are filled with estimated non-negative (power) spectrum, as shown by Figure 4(b) and Figure 4(d). While the activation matrices, which are also non-negative elements, as shown by Figure 4(c) and Figure 4(e), contain weight values that control the contribution of every element of the dictionary in a mixture. That is why this concept can be grafted in this single-channel source separation problem flawlessly.

In Figure 5, the reconstruction of successfully separated sources is shown—the reconstruction of first and second sources close to the original. According to Figure 5(b), the reconstructed source even follows the flow of the original sequence. The problem with reconstruction mostly happens in the detail or high frequencies part. They are changing at a high pace from a certain frequency to another. However, this reconstruction still can estimate where the flow goes on. The other difference between reconstructed sources and original is on amplitudes. This difference only tells about the strength of the portion of sound (in the speech signal).

To help the reconstruction before they are inverted to the time domain, a Wiener filter is applied. The filter coefficients are selected from **W** and **H** according to *K*. This filter is multiplied with the power spectrum (non-negative matrix) of the mixture. The masks for the first and second sources are shown in Figure 6.

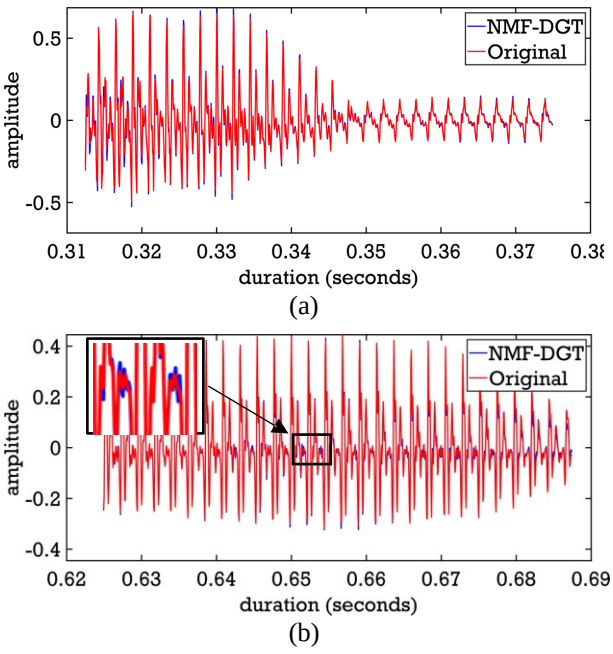


Figure 5. The reconstructed sources of supervised NMF: (a) first signal and (b) second signal

In Figure 7, the comparisons between DGT and STFT are shown. The overall performances are measured using three popular ratios. The benefit of using these ratios is their ability to neglect phase changes on reconstructed sources [21]-[23]. In the case

of speech, most of the time, the phase changes to the known value π, which will not change how it sounds. Nevertheless, this is risky when different evaluations are used to measure the error between reconstructed and original.

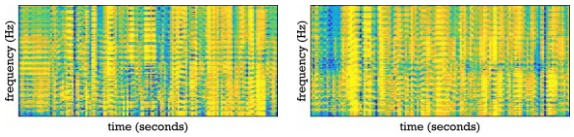
The SIR, SAR, and SDR of NMF-DGT outperform the NMF-STFT with scores of 18.60 dB compared to 16.24 dB, 13.77 dB to 13.69 dB, and 12.45 dB to 11.16 dB, respectively. The SIR value indicates that reconstructed signal using DGT has lower interferences and induces slightly low artifacts compared to STFT. When combined, the total error of DGT is much lower than STFT, as indicated by SDR around 1.29 dB. It concludes that in terms of supervised NMF, the DGT has better performance than STFT, consistent with [25]-[28].

B. Unsupervised NMF

The difference between supervised and unsupervised NMF is in the way the initial value of **W** is provided. Unlike the previously explained supervised scheme, the initial value of **W** is generated randomly. However, the amount of *K* is predetermined. In this case, the value of *K* is set to 40 for every source.

The pairs of dictionary and activation matrices are shown in Figure 8. They reasonably much vary compared to the supervised scheme. The overlay visual between original and reconstructed sources are shown in Figure 9. Similar to the supervised scheme, the reconstructed sources have the same phase as the original. However, according to Figure 9(a), the reconstructed sources seem to lose their details. The loss is because of the lack of information about high frequencies (with low amplitude) that occupied a certain time domain that generated the initial dictionary. It happens to both reconstructed sources. The masks that the dictionary and activation matrices created are shown in Figure 10.

The evaluation compares to NMF-STFT is shown in Figure 11. Again, the NMF-DGT outperforms NMF-STFT even though the difference is not significant. The scores are very low for both T-F representations. Even more, the SAR and SDR lie below zero. Negative values for both evaluation methods mean some artifacts and interferences cause a total error bigger. The only reason is that the initial dictionary containing basis spectral (frequency bins) cannot estimate the basis spectral correctly as more interferences are preserved and artifacts induced. However, the probability that a randomly generated initial dictionary would estimate correct frequencies is extremely low. Overall, the supervised NMF gives better performances compared to unsupervised NMF [25]-[28]. The comparison values of NMF-DGT and NMF-STFT are 0.40 dB vs 0.27 dB, -10.21 dB vs -10.36, and -15.01 dB vs -15.23 dB, respectively. In this round, NMF-DGT is slightly better than NMF-STFT.



(a) (b)

Figure 6. Mask filter of supervised NMF: (a) first signal and (b) second signal

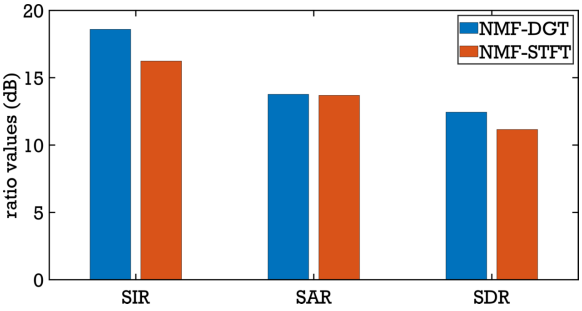
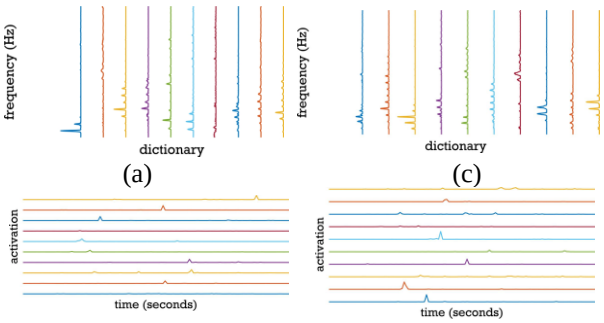


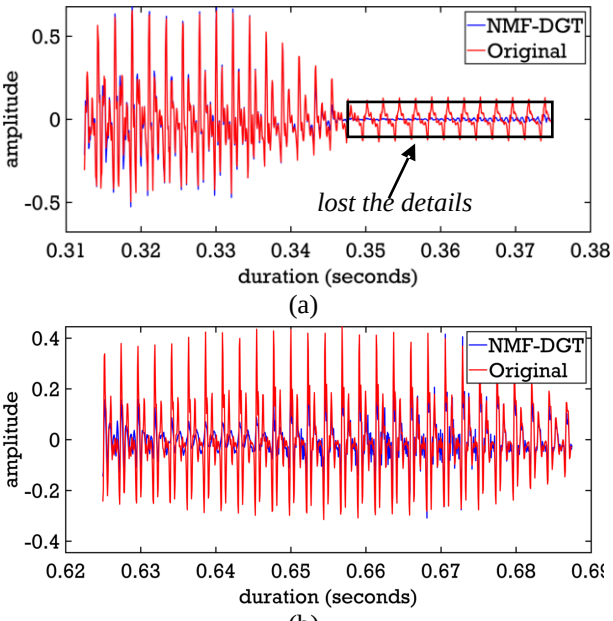
Figure 7. The evaluation performance of supervised NMF



(a) (c)

(b) (d)

Figure 8. The dictionaries and activations of unsupervised NMF: (a)-(b) first source's dictionary and activation, (c)-(d) second source's dictionary and activation



(a)

(b)

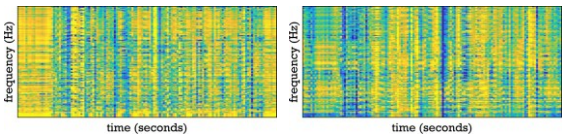
Figure 9. The reconstructed sources of unsupervised NMF: (a) first signal and (b) second signal

IV. Conclusion

The performance of DGT being utilized to replace the STFT for a single-channel source separation problem is better in both schemes: a supervised NMF and unsupervised NMF. Even though NMF-DGT also exhibits low performance in unsupervised NMF, but it is still better than NMF-STFT. It shows that DGT can replace STFT in single-channel source separation, especially in the NMF framework. Possibly, this T-F representation can be beneficial in a wide area of the signal processing field. In the future, the DGT will be utilized to design a novel method to estimate the mixing matrix based on active single-point estimation.

Acknowledgment

This work is supported by a research grant under a competitive scheme from Vocational College of Universitas Gadjah Mada with contract number 83/UN1.SV/KPT/2020.



(a) (b)

Figure 10. Mask filter of unsupervised NMF: (a) first source and (b) second source

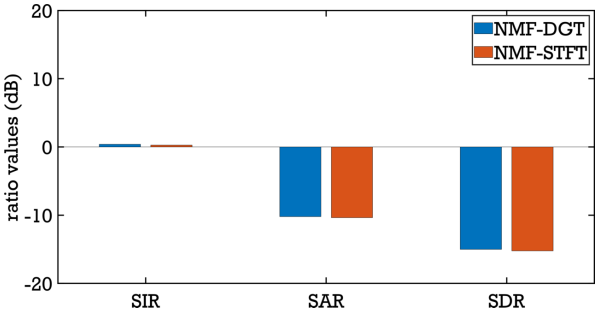


Figure 11. The evaluation performance unsupervised NMF

References

[1] M. F. Issa and Z. Juhasz, "Improved EOG artifact removal using wavelet enhanced independent component analysis," *Brain Sciences*, vol. 9, no. 12, 355, 2019. doi: 10.3390/brainsci9120355

[2] A. Ghazdali, A. Hakim, A. Laghrib, N. Mamouni, and S. Raghay, "A new method for the extraction of fetal ecg from the dependent abdominal signals using blind source separation and adaptive noise cancellation techniques," *Theoritical Biology and Medical Modelling*, vol. 12, no. 25, pp. 1-20, 2015. doi: 10.1186/s12976-015-0021-2

[3] H. Qi, Z. Guo, X. Chen, Z. Shen, Z. J. Wang, "Video-based human heart rate measurement using joint blind source separation," *Biomedical Signal Processing and Control*, vol. 31, pp. 309-320, 2017. doi: 10.1016/j.bspc.2016.08.020

[4] M. Maazaoui, K. Abed-Meraim, and Y. Grenier, "Blind source separation for robot audition using fixed HRTF beamforming," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, 58, 2012. doi: 10.1186/1687-6180-2012-58

[5] H. Lee, "Simultaneous blind separation and recognition of speech mixtures using two microphones to control a robot cleaner," *International Journal of Advanced Robotic Systems*, vol. 10, no. 2, pp. 1-10, 2017. doi: 10.5772/55408

[6] K. Zhang, G. Tian, and T. Lan, "Blind source separation based on JADE algorithm and application," in *3rd International Conference on Mechatronics, Robotics and Automation*, Shenzhen, China, Apr. 2015, pp. 252-255. doi: 10.2991/icmra-15.2015.50

[7] C-Y. Yu, Y. Li, B. Fei, and W-L. Li, "Blind source separation based x-ray image denoising from an image sequence," *Review of Scientific Instruments*, vol. 86, no. 9, 2015, doi: 10.1063/1.4928815

[8] M. M. Hossain, B. E. Levy, D. Thapa, A. L. Oldenburg, and C. M. Gallippi, "Blind source separation-based motion detector for imaging super-paramagnetic iron oxide (SPIO) particles in magnetomotive ultrasound imaging," *IEEE Transactions on Medical Imaging*, vol. 37, no. 10, pp. 2356-2366, 2018. doi: 10.1109/TMI.2018.2848204

[9] R. R. Wildeboer et al., "Blind source separation for clutter and noise suppression in ultrasound imaging: review for different applications," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 67, no. 8, pp. 1497-1512, 2020. doi: 10.1109/TUFFC.2020.2975483

[10] K. Yoshii, R. Tomioka, D. Mochihashi, M. Goto, "Beyond NMF: time-domain audio source separation without phase reconstruction," in *14th International Society for Music Information Retrieval Conference*, Curitiba, Brazil, Nov. 2013.

[11] M. N. Schmidt, "Single-channel source separation using non-negative matrix factorization," *thesis*, Technical University of Denmark, Denmark, 2009.

[12] Y. Li, Y. Wang, and Q. Dong, "A novel mixing matrix estimation algorithm in instantaneous underdetermined blind source separation," *Signal, Image and Video Processing*, vol. 14, pp. 1001-1008, 2020. doi: 10.1007/s11760-019-01632-z

[13] Y. Zhang, Z. Zhang, H. Tao, and Y. Lin, "A single source point detection algorithm for underdetermined blind source separation problem," in *ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2019*, S. Liu and G. Yang (Eds.): ADHIP 2018, LNICST 279, 2019, pp. 68–76. doi: 10.1007/978-3-030-19086-6_8

[14] W. Guan, L. Dong, Y. Cai, J. Yan, and Y. Han, "Sparse component analysis with optimized clustering for underdetermined blind modal

identification," *Measurement Science and Technology*, vol. 30, no. 12, 2019. doi: 10.1088/1361-6501/ab3054

[15] H. Zhang, G. Hua, L. Yu, Y. Cai, and G. Bi, "Underdetermined blind separation of overlapped speech mixtures in time-frequency domain with estimated number of sources," *Speech Communication*, vol. 89, pp. 1-16, 2017. doi: 10.1016/j.specom.2017.02.003

[16] T. Peng, Y. Chen, and Z. Liu, "A time–frequency domain blind source separation method for underdetermined instantaneous mixtures," *Circuits System Signal Processing*, vol. 34, pp. 3883-3895, 2015. doi: 10.1007/s00034-015-0035-3

[17] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Convolutive blind source separation for more than two sources in the frequency domain," in *International Conference on Acoustics, Speech, and Signal Processing*, Montreal, Canada, May 2004, pp. iii-885. doi: 10.1109/ICASSP.2004.1326687

[18] M. Jafari, E. Vincent, S. Abdallah, M. Plumbley, and M. Davies, "Blind source separation of convolutive audio using an adaptive stereo basis," *UK ICA Research Network Workshop*, Southampton, United Kingdom, Sep. 2006. Available: https://hal.inria.fr/inria-00544290. [Accessed: August 2, 2020].

[19] T. Asamizu, S. Saito, K. Oishi, and T. Furukawa, "Overdetermined blind source separation using approximate joint diagonalization," in *60th International Midwest Symposium on Circuits and Systems*, Boston, MA, USA, Oct. 2017, pp. 168-171. doi: 10.1109/MWSCAS.2017.8052887

[20] L. Wang, J. D. Reiss, and A. Cavallaro, "Over-determined source separation and localization using distributed microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 9, pp. 170-177, 2016. doi: 10.1109/TASLP.2016.2573048

[21] K. Yatabe, Y. Masuyama, T. Kusano, and Y. Oikawa, "Representation of complex spectrogram via phase conversion," in *Acoustical Science and Technology*, vol. 40, no. 3, pp. 170-177, 2019. doi: 10.1250/ast.40.170

[22] H. G. Feichtinger and T. Strohmer, Eds., *Gabor Analysis and Algorithms: Theory and Applications*. Birkhäuser, Boston, 1998. doi: 10.1007/978-1-4612-2016-9

[23] Z. Průša, "STFT and DGT phase conventions and phase derivatives interpretation," *The Large Time-Frequency Analysis Toolbox (LTFAT Notes)*, 2016. Available: https://www.ltfat.github.io/notes/ltfatnote042.pdf . [Accessed: July 29, 2020].

[24] S. A. Rafiei, H. Sheikhzadeh, and M. Sabbaqi, "A new reduced-interference source separation method based on a complementary combination of masking algorithm and mixing matrix estimation," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 44, pp. 1529-1547, 2020, doi: 10.1007/s40998-020-00326-4

[25] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *13th International Conference on Neural Information Information Systems*, Cambridge, USA, Jan. 2000, pp. 556-562.

[26] M. W. Berry, M. Brown, A. N. Langville, V. P. Pauca, and R. J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 155-173, 2007. doi: 10.1016/j.csda.2006.11.006

[27] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411-430, 2000. doi: 10.1016/S0893-6080(00)00026-5

[28] A. Hyvärinen, "Fast and robust fixed-point algorithms for independent component analysis," *IEEE Trans Neural Network*, vol. 10, no. 3, pp. 626-634, 1999. doi: 10.1109/72.761722

[29] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus LDC93S1," Philadelphia: Linguistic Data Consortium, 1993. Available: https://catalog.ldc.upenn.edu/LDC93S1. [Accessed: August 2, 2020].

[30] E. Vincent et al., "The signal separation evaluation campaign (2007-2010): Achievements and remaining challenges," *Signal Processing*, vol. 92, no. 8, pp. 1928-1936, 2012. doi: 10.1016/j.sigpro.2011.10.007

[31] M. I. Mandel, S. Bressler, B. Shinn-Cunningham, and D. P. W. Ellis, "Evaluating source separation algorithms with Reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1872-1883, 2010. doi: 10.1109/TASL.2010.2052252