# $L_P$-NORM NON-NEGATIVE MATRIX FACTORIZATION AND ITS APPLICATION TO SINGING VOICE ENHANCEMENT

*Tomohiko Nakamura*[†] *and Hirokazu Kameoka*[†,‡]

[†]Graduate School of Information Science and Technology, The University of Tokyo.
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan.
[‡]NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation.
3-1, Morinosato Wakamiya, Atsugi, Kanagawa, 243-0198, Japan
Tomohiko_Nakamura@ipc.i.u-tokyo.ac.jp,kameoka.hirokazu@lab.ntt.co.jp

**ABSTRACT**

Measures of sparsity are useful in many aspects of audio signal processing including speech enhancement, audio coding and singing voice enhancement, and the well-known method for these applications is non-negative matrix factorization (NMF), which decomposes a non-negative data matrix into two non-negative matrices. Although previous studies on NMF have focused on the sparsity of the two matrices, the sparsity of reconstruction errors between a data matrix and the two matrices is also important, since designing the sparsity is equivalent to assuming the nature of the errors. We propose a new NMF technique, which we called $L_p$-norm NMF, that minimizes the $L_p$ norm of the reconstruction errors, and derive a computationally efficient algorithm for $L_p$-norm NMF according to an auxiliary function principle. This algorithm can be generalized for the factorization of a real-valued matrix into the product of two real-valued matrices. We apply the algorithm to singing voice enhancement and show that adequately selecting $p$ improves the enhancement.

***Index Terms***— Non-negative matrix factorization, $L_p$ norm, auxiliary function principle

## 1. INTRODUCTION

Non-negative matrix factorization (NMF) [1] is a powerful technique that approximates a data matrix $Y$ by using the product of two non-negative matrices $W$ and $H$. NMF has been actively studied in many scientific and engineering fields in recent years (see [2]). In particular, in the field of music signal processing, successful results were obtained by regarding a magnitude spectrogram as a non-negative matrix [3].

NMF is formulated as the problem of minimizing a measure between a data matrix and a model. One standard measure is the Frobenius norm, and it corresponds to assuming that the reconstruction errors are additive Gaussian noise. While NMF with the Frobenius norm works well for small errors, NMF performance deteriorates for large errors such as spike noise, even if there are a few entries with errors. We thus need to explore appropriate measures for large errors.

Measures of sparsity (e.g. $L_1$ norm) have been widely used in many audio signal processing techniques such as NMF and robust principle component analysis [4, 5]. Many previous studies of NMF have used sparseness measures [6, 7], regularizers [8] and priors [9] on $W$ and $H$, which induce sparse solutions. However, it is difficult to use the measures between a data matrix and a model directly since the measures are often non-linear and not differentiable, and it is intractable to solve the NMF problem.

In this paper, we propose a new NMF technique, which we called $L_p$-norm NMF, that minimizes the $L_p$ norm of reconstruction errors between a data matrix and a model. We formulate $L_p$-norm NMF with $0 < p \leq 2$ and derive a convergence-guaranteed algorithm that consists of multiplicative update equations for $W$ and $H$ based on the auxiliary function principle [10, 11]. The constant $p$ controls the sparsity of the reconstruction errors. To examine the effect of varying $p$, we apply the proposed algorithm to singing voice enhancement for monaural audio signals.

We henceforth denote sets of real values and non-negative real values as $\mathbb{R}$ and $\mathbb{R}_{\geq 0}$, respectively.

## 2. $L_P$-NORM NON-NEGATIVE MATRIX FACTORIZATION

### 2.1. Problem setting

Let us define frequency, time, and basis indexes, respectively, as $\omega \in [0, \Omega - 1]$, $t \in [0, T - 1]$ and $k \in [0, K - 1]$ such that $K < \Omega$ and $K < T$. Given a non-negative data matrix $Y := \{y_{\omega,t}\}_{0\leq\omega\leq\Omega-1,0\leq t\leq T-1}$, $L_p$-norm NMF is the problem of minimizing the $L_p$ norm

$$\mathcal{L}(W, H) := \|Y - WH\|_p^p = \sum_{\omega,t} \left| y_{w,t} - \sum_k w_{\omega,k} h_{k,t} \right|^p \tag{1}$$

subject to

$$\forall k,\ \sum_{\omega} W_{\omega,k} = 1. \tag{2}$$

Here $W = \{w_{\omega,k}\}_{\omega,k}$ is a $\Omega \times K$ non-negative matrix and $H = \{h_{k,t}\}_{k,t}$ is a $K \times T$ non-negative matrix. Eq. (2) is introduced to avoid an indeterminacy in the scaling. When $Y$ is a magnitude spectrogram, the columns of $W$ represent spectral templates and the rows of $H$ represent the temporal activities of the spectral templates.

The constant $p$ controls the sparsity of the reconstruction error. A smaller $p$ induces the sparsity, and most of the entries of the estimated data matrix $WH$ are the same as those of $Y$, while the other entries differ greatly from those of $Y$. On the other hand, a larger $p$ does not induce the sparsity, and all the entries of the estimated data matrix may be non-zero. We thus assume that the $L_p$ norm satisfies $0 < p < 2$, which promotes sparsity. Note that when $p = 2$, the objective function equals the NMF with Frobenius norm.

This work was partly supported by JSPS Grant-in-Aid for Young Scientists B Grant Number 26730100.

### 2.2. Efficient algorithm based on auxiliary function principle

Since $\mathcal{L}(W,H)$ involves a summation over $k$ in the $L_p$ norm, it is intractable to solve the current minimization problem analytically. However, we can develop a computationally efficient algorithm to find a locally optimal solution based on the auxiliary function principle [10, 11].

When applying an auxiliary function principle to the minimization problem, the first step is to define an upper bound function for the objective function. Introducing auxiliary variables $\xi = \{\xi_{\omega,t} \in \mathbb{R}_{\geq 0}\}_{\omega,t}$, we derive

$$\left|y_{w,t} - \sum_k w_{\omega,k}h_{k,t}\right|^p \leq p\xi_{\omega,t}^{p-2}\left|y_{w,t} - \sum_k w_{\omega,k}h_{k,t}\right|^2 + (2-p)\xi_{\omega,t}^p. \quad (3)$$

The proof of the inequality is described in Lemma 2 of [12]. The equality of Eq. (3) holds if and only if

$$\xi_{\omega,t} = \left|y_{\omega,t} - \sum_k w_{\omega,t}h_{k,t}\right|. \quad (4)$$

The term in the square function of Eq. (3) includes a summation over $k$, and we further derive the upper bound of the right-hand side of Eq. (3). Since the square function is a convex function, we can invoke Jensen's inequality:

$$\Big(\sum_k w_{\omega,k}h_{k,t}\Big)^2 \leq \sum_k \frac{1}{\lambda_{\omega,t,k}}(w_{\omega,k}h_{k,t})^2 \quad (5)$$

where $\lambda = \{\lambda_{\omega,t,k} \in \mathbb{R}_{\geq 0}\}_{\omega,t,k}$ are auxiliary variables that sum to unity: $\sum_k \lambda_{\omega,t,k} = 1$. The equality of Eq. (5) holds if and only if

$$\lambda_{\omega,t,k} = \frac{w_{\omega,k}h_{k,t}}{\sum_{k'} w_{\omega,k'}h_{k',t}}. \quad (6)$$

The upper bound of $\mathcal{L}(W,H)$ can thus be described as

$$\begin{aligned}&\mathcal{L}^+(W,H,\lambda,\xi)\\ &= \sum_{\omega,t} p\xi_{\omega,t}^{p-2}\Big\{y_{w,t}^2 - y_{\omega,t}\sum_k w_{\omega,k}h_{k,t} + \sum_k \frac{1}{\lambda_{\omega,t,k}}(w_{\omega,k}h_{k,t})^2\Big\}\\ &\quad + \sum_{\omega,t}(2-p)\xi_{\omega,t}^p. \end{aligned} \quad (7)$$

We can derive update equations for the parameters, using the above upper bound. By setting the partial derivative of $\mathcal{L}^+(W,H,\xi,\lambda)$ with respect to $W$ and $H$ at zero and using Eqs. (4) and (6), we obtain

$$W \leftarrow W \odot [\{(Y \oslash C)H^\top\} \oslash \{(WH \oslash C)H^\top\}] \quad (8)$$

$$H \leftarrow H \odot [\{W^\top(Y \oslash C)\} \oslash \{W^\top(WH \oslash C)\}] \quad (9)$$

where $\odot, \oslash$ denote element-wise product and division, and $C$ is an $\Omega \times T$ matrix, whose $(\omega,t)$-th element is

$$c_{\omega,t} = \left|y_{\omega,t} - \sum_k w_{\omega,k}h_{k,t}\right|^{2-p}. \quad (10)$$

It is worth noting that each update equation consists of multiplication by a non-negative factor. Hence, it is guaranteed that the entries of $W$ and $H$ are always non-negative when their initial values are set at non-negative values.

## 3. $L_p$-NORM MATRIX FACTORIZATION

### 3.1. Problem setting

Although the algorithm in the previous section is for non-negative matrices, it can be generalized for real-valued matrices, according to the auxiliary function principle. We call the problem $L_p$-norm matrix factorization ($L_p$-norm MF).

$L_p$-norm MF is the problem of finding real-valued matrices $W, H$ for a given real-valued data matrix $Y$ such that

$$\min_{W\in\mathbb{R}^{\Omega\times K}, H\in\mathbb{R}^{K\times T}} \mathcal{J}(W,H) = \|Y - WH\|_p^p,$$
$$\text{subject to } \forall k,\ \sum_\omega W_{\omega,k} = 1$$

where $0 < p < 2$ and $\mathcal{J}(W,H)$ is the objective function.

### 3.2. Iterative algorithm based on auxiliary function principle

As in Sec. 2, $\mathcal{J}(W,H)$ involves summation over $k$ in the $L_p$ norm, and we consider the upper bound of $\mathcal{J}(W,H)$ to apply the auxiliary function principle to the intractable minimization problem. The inequality used in Eq. (3) is applicable to $\mathcal{J}(W,H)$, and the upper bound of $\mathcal{J}(W,H)$ is the same as the right-hand side of Eq. (3). To derive the upper bound of the square function of Eq. (3), we can use the generalization of Jensen's inequality for $y_{\omega,t}, w_{\omega,k}, h_{k,t} \in \mathbb{R}$, which was employed in [12]:

$$\left|y_{\omega,t} - \sum_k w_{\omega,k}h_{k,t}\right|^2 \leq \sum_k \frac{|\alpha_{\omega,t,k} - w_{\omega,k}h_{k,t}|^2}{\beta_{\omega,t,k}} \quad (11)$$

where $\alpha = \{\alpha_{\omega,t,k} \in \mathbb{R}\}_{\omega,t,k}, \beta = \{\beta_{\omega,t,k} \in [0,1]\}_{\omega,t,k}$ are auxiliary variables subject to $\sum_k \alpha_{\omega,t,k} = y_{\omega,t}, \sum_k \beta_{\omega,t,k} = 1$. The equality of Eq. (11) holds if and only if

$$\alpha_{\omega,t,k} = w_{\omega,k}h_{k,t} - \beta_{\omega,t,k}\Big(\sum_{k'} w_{\omega,k'}h_{k',t} - y_{\omega,t}\Big). \quad (12)$$

The upper bound of $\mathcal{J}^+(W,H)$ can thus be described as

$$\mathcal{J}^+(W,H,\xi,\alpha,\beta) = \sum_{\omega,t} p\xi_{\omega,t}^{p-2}\sum_k \frac{|\alpha_{\omega,t,k} - w_{\omega,k}h_{k,t}|^2}{\beta_{\omega,t,k}} + (2-p)\xi_{\omega,t}^p. \quad (13)$$

By differentiating $\mathcal{J}^+(W,H,\xi,\alpha,\beta)$ partially with respect to $W$ and $H$ and setting them at zero, we can obtain the update equations:

$$w_{\omega,k} \leftarrow \frac{\sum_t c_{\omega,t}^{-1}h_{k,t}\Big(\beta_{\omega,t,k}^{-1}w_{\omega,k}h_{k,t} + y_{\omega,t} - \sum_{k'} w_{\omega,k'}h_{k',t}\Big)}{\sum_t c_{\omega,t}^{-1}\beta_{\omega,t,k}^{-1}h_{k,t}^2} \quad (14)$$

$$h_{k,t} \leftarrow \frac{\sum_\omega c_{\omega,t}^{-1}w_{\omega,k}\Big(\beta_{\omega,t,k}^{-1}w_{\omega,k}h_{k,t} + y_{\omega,t} - \sum_{k'} w_{\omega,k'}h_{k',t}\Big)}{\sum_\omega c_{\omega,t}^{-1}\beta_{\omega,t,k}^{-1}w_{\omega,k}^2} \quad (15)$$

where $c_{\omega,t}$ is defined as Eq. (10).

### 3.3. Relation to $L_p$-norm NMF

The parameters $\beta$ can be chosen arbitrarily subject to $\beta_{\omega,t,k} \in \mathbb{R}_{\geq 0}$ and $\sum_k \beta_{\omega,k,t} = 1$. Choosing $\beta$ as in [12], we obtain the multiplicative update equations as in Eqs. (8) and (9). Therefore, we can say that the algorithm of $L_p$ NMF is a special case of that of $L_p$ MF.

## 4. APPLICATION TO SINGING VOICE ENHANCEMENT

### 4.1. Singing voice enhancement

In this section, we apply $L_p$-norm NMF to singing voice enhancement whose aim is to extract a singing voice from a monaural mixed audio signal consisting of vocal and accompaniment parts. Singing voice enhancement is often used in music information retrieval (MIR) applications such as automatic lyrics recognition [13, 14], automatic singer identification [15], and automatic karaoke generators [16].

When input audio signals are multichannel, we can use spatial cues, based on the fact that the vocal parts of music audio signals are frequently mixed in the center of the sound field. However, for monaural inputs, we need other cues instead of the spatial cues.

### 4.2. Enhancement algorithm

We utilize the fact that there is a difference in spectrogram between accompaniments and singing voices. Accompaniment sounds are generated by musical instruments, which reproduce approximately the same sounds every time they are played. We can see the spectrograms of accompaniment signals as a low-rank matrix. In constrast, singers fluctuate their singing voices to obtain musical effects such as vibrato. The spectrograms of singing voices are relatively high rank and sparse, and corresponds to reconstructions errors of $L_p$-norm NMF.

As mentioned in Sec. 2.1, the model spectrogram $WH$ of $L_p$-norm NMF is a non-negative matrix with the rank of $K$, and small $p$ induces sparsity of the reconstruction errors. Setting $K$ and $p$ at sufficiently small values, we expect the low-rank matrix $WH$ to contain accompaniment signals and the reconstruction errors $Y-WH$ to contain singing voice signals.

Let the estimated magnitude spectrogram of a singing voice be $\hat{S} = \{\hat{s}_{\omega,t}\}_{0\le\omega\le\Omega-1,0\le t\le T-1}$. First, an input signal containing a singing voice and accompaniment is converted into a complex spectrogram with the short-time Fourier transform (STFT). We regard the magnitude spectrogram of the input signal as a non-negative matrix, and apply $L_p$-norm NMF to the magnitude spectrogram. The obtained model spectrogram $WH$ should correspond to the accompaniment, and the reconstruction errors between the observed magnitude spectrogram and the model spectrogram corresponds to the singing voice. The time-frequency components of the model spectrogram may be larger than those of the observed spectrogram, and we derive $\hat{S}$ as

$$\hat{s}_{\omega,t} = \begin{cases} y_{\omega,t} - \sum_k w_{\omega,k}h_{k,t} & (y_{\omega,t} \ge \sum_k w_{\omega,k}h_{k,t}) \\ 0 & (y_{\omega,t} < \sum_k w_{\omega,k}h_{k,t}) \end{cases}. \quad (16)$$

The estimated magnitude spectrogram with the phase of the original complex spectrogram is converted into an audio signal by the inverse STFT.

## 5. EXPERIMENTAL EVALUATION

### 5.1. Experimental conditions

To evaluate the performance of the proposed method, we conducted two experiments on singing voice enhancement: an evaluation of the effect of $p$, which controls sparsity, and frame length $F$, and a comparison of our results with the state-of-the-art [5, 17, 18].

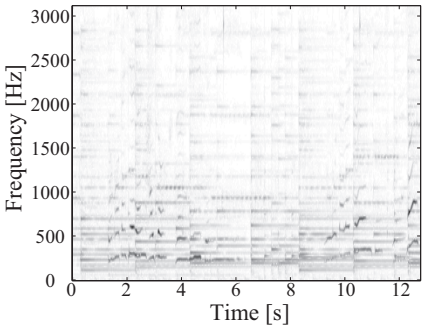
The criteria for evaluating the singing voice enhancement were the normalized signal-to-distortion ratio (NSDR) and the global NSDR (GNSDR), given as

$$\mathrm{NSDR}(\{\hat{f}_t\}_t;\{f_t\}_t,\{x_t\}_t) = \mathrm{SDR}(\{\hat{f}_t\}_t,\{f_t\}_t) - \mathrm{SDR}(\{x_t\}_t,\{f_t\}_t), \quad (17)$$

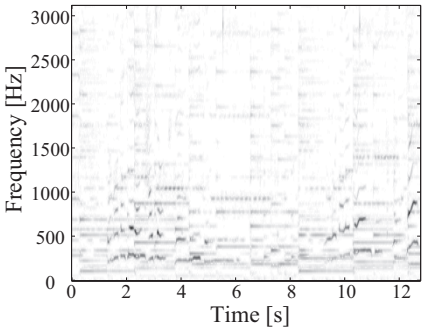
$$\mathrm{GNSDR} = \frac{\sum_i w_i \mathrm{NSDR}(\{\hat{f}_{i,t}\}_t;\{f_{i,t}\}_t,\{x_{i,t}\}_t)}{\sum_i w_i}, \quad (18)$$

$$\mathrm{SDR}(\{\hat{f}_t\}_t,\{f_t\}_t) = 10\log_{10}\frac{\sum_t \hat{f}_t f_t}{\left(\sum_t \hat{f}_t\right)\left(\sum_t f_t\right) - \sum_t \hat{f}_t f_t}, \quad (19)$$

where $\hat{f}_{i,t}$, $f_{i,t}$ and $x_{i,t}$ denote the estimated signal, the target signal and the input signal of the $i$-th piece. NSDR represents the improvement in SDR, and GNSDR denotes the weighted averages of the NSDR of all the music pieces by the length of the $i$-th piece, $\{w_i\}_i$. These criteria have also been employed in many previous studies [5, 17–21]. To calculate the SDR, we used the BSS Eval Toolbox [22, 23].




(a) Spectrogram of input audio signal.




(b) Spectrogram of audio signal after enhancement.

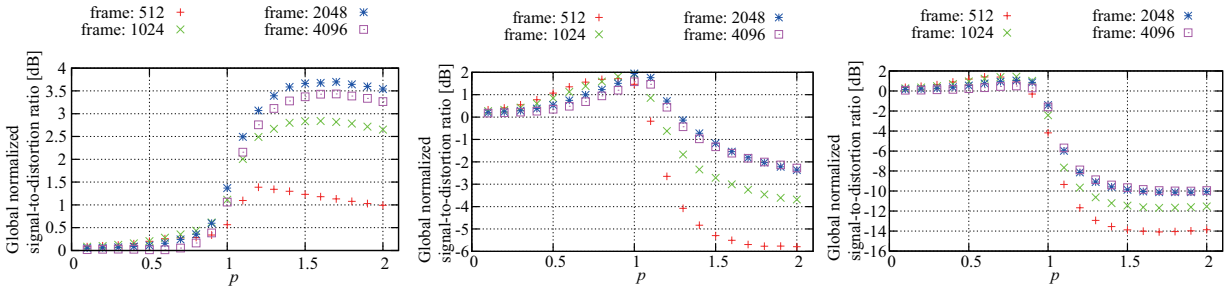
**Fig. 1**. Spectrogram examples of (a) a mixed audio signal and (b) the singing-voice-enhanced audio signal obtained with the proposed method.

As an evaluation dataset, we used the MIR-1K dataset [24], following the evaluation framework in [5, 17, 18]. The dataset consists of 1000 Chinese song clips performed by amateur singers. The durations of the clips range from 4 to 13 s, and the audio signals are monaural with a sampling rate of 16 kHz. The accompaniment and vocal parts were recorded separately, and we could mix them with any signal-to-noise ratio (SNR), where the SNR corresponds to the voice to accompaniment ratio. The accompaniment and vocal parts for each clip were mixed at −5 dB (accompaniment is louder), 0 dB (same level) and 5 dB (vocal is louder) SNRs.

### 5.2. Effect of sparsity and frame lengths

We first compared the proposed method in $p$ and $F$. We used $p = 0.1, 0.2, \cdots, 2.0$ and $F = 512, 1024, 2048, 4096$ sample points. For STFT, the window function was the sine window, and the frame shifts were half the length of the frames. The number of bases was set at $K = 10$. The entries of $W$ and $H$ were initialized randomly. There were 200 iterations, which is supposed to be sufficient empirically.

Fig. 1 shows one of the enhanced results obtained with the proposed method. The figures show the spectrograms of the input signal and the enhanced result. We can see that most of the accompanying sounds (vertically and horizontally smooth components) are suppressed, and the singing voice component of the spectrogram is clearer than that of the input spectrogram.




(a) GNSDR at −5 dB SNR. (b) GNSDR at 0 dB SNR. (c) GNSDR at 5 dB SNR.

**Fig. 2**. Global normalized signal-to-distortion ratio (GNSDR) of the proposed method with respect to $p$ of the $L_p$ norm and frame length $F$. Red, blue, green and purple points correspond to $F = 512, 1024, 2048, 4096$ sample points. The results are for (a) −5 dB, (b) 0 dB, and (c) 5 dB SNRs.

**Table 1**. Global normalized signal-to-distortion ratio (GNSDR) comparison with previous studies. $F$ denotes the length of a frame in sample point. Hsu, Rafii and Huang represent the corresponding methods [5, 17, 18].

| Input SNR [dB] | Proposed method $F = 1024$ | Proposed method $F = 2048$ | Hsu [17] | Rafii [18] | Huang [5] |
|---|---|---|---|---|---|
| −5 | 2.84 ($p = 1.6$) | 3.70 ($p = 1.7$) | −0.51 | 0.52 | 1.51 |
| 0 | 1.93 ($p = 1.0$) | 1.95 ($p = 1.0$) | 0.91 | 1.11 | 2.37 |
| 5 | 1.43 ($p = 0.8$) | 1.04 ($p = 0.8$) | 0.17 | 1.10 | 2.57 |

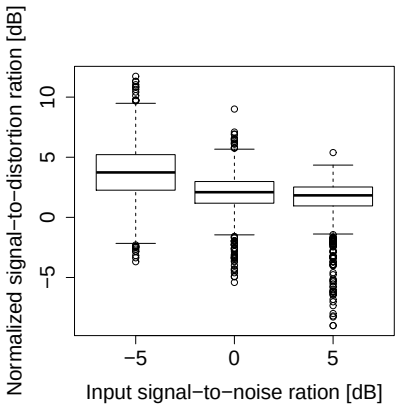



**Fig. 3**. Box plot of the normalized signal-to-distortion ratio (NSDR) of the proposed method for the MIR-1K dataset. The results for SNRs of −5, 0 and 5 dB are for $(p, F) = (1.7, 2048), (1.0, 2048)$ and $(0.8, 1024)$, respectively.

As illustrated in Fig. 2 for SNRs of $-5, 0, 5$ dB, the results show that the GNSDRs depended strongly on $p$ for all frame lengths. The highest GNSDRs for all SNRs were 3.7 at $(p, F) = (1.7, 2048)$ for −5 dB SNR, 1.95 at $(p, F) = (1.0, 2048)$ for 0 dB SNR, and 1.43 at $(p, F) = (0.8, 1024)$ for 5 dB SNR. We can see that $p$ at which GNSDR was the highest for each input SNR decreased as the input SNR became higher. With a high input SNR, the non-zero time-frequency components of the singing voice spectrogram are large, and increasing the sparsity is preferred. On the other hand, with a low input SNR, the non-zero time-frequency components are small, and decreasing the sparsity is preferred. The obtained results are consistent with this idea.

Fig. 3 shows the distributions of NSDRs for each SNR. The results were for $(p, F) = (1.7, 2048), (1.0, 2048), (0.8, 1024)$ for SNRs of $-5, 0, 5$ dB. Since most of the NSDRs exceeded 0 dB, and we can confirm that the proposed method worked well for most of the input signals.

### 5.3. Comparison with previous studies

Finally, we compared the proposed method with the state-of-the-art [5, 17, 18]. The results are summarized in Tab. 1. The proposed method with $F = 1024$ outperformed two previous methods for all input SNRs. While the GNSDRs of the proposed method were lower than those of [5] at SNRs of 0 and 5 dB, the GNSDR with the proposed method was 0.4 to 2.4 dB larger than those of three previous methods at a SNR of −5 dB. This result indicates that the proposed method works well particularly in a low SNR environment.

## 6. CONCLUSION

We proposed a new NMF that minimizes the $L_p$ norm of the reconstruction errors between a data matrix and the model. A computationally efficient algorithm was derived according to the auxiliary function principle, and it has multiplicative update equations, which guarantee the non-negativity of $W$ and $H$. The algorithm of $L_p$-norm NMF can be generalized for that of $L_p$-norm MF for real-valued matrices. We applied $L_p$-norm NMF to singing voice enhancement and showed by experiments that adequately selecting $p$ improves the enhancement quality, and the proposed method outperformed three previous works under a low SNR situation.

There are several ways to extend $L_p$-norm NMF to other applications. One promising application is speech enhancement, since the spectrogram of background noise is sometimes approximated as low rank and the speech spectrogram is relatively sparse.

## 7. REFERENCES


[1] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[2] Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, 2013.

[3] P. Smaragdis and J. C. Brown, "Non-negative matrix factorization for polyphonic music transcription," in *Proc. IEEE Workshop Applications Signal Process. Audio Acoust.* IEEE, 2003, pp. 177–180.

[4] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[5] P.-S. Huang, S. D. Chen, P. Smaragdis, and M. Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2012, pp. 57–60.

[6] P. O. Hoyer, "Non-negative sparse coding," in *IEEE Workshop Neural Networks for Signal Process.*, 2002, pp. 557–565.

[7] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *J. Mach. Learn. Res.*, vol. 5, pp. 1457–1469, 2004.

[8] B. Shen, L. Si, R. Ji, and B.-D. Liu, "Robust nonnegative matrix factorization via $l_1$ norm regularization," *CoRR*, vol. abs/1204.2311, 2012.

[9] T. Virtanen, A. T. Cemgil, and S. Godsill, "Bayesian extensions to non-negative matrix factorisation for audio signal modelling," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Mar. 2008, pp. 1825–1828.

[10] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Number 30. 2000.

[11] H. Kameoka, M. Goto, and S. Sagayama, "Selective amplifier of periodic and non-periodic components in concurrent audio signals with spectral control envelops," in *Proc. SIG Tech. Reports on Music and Computer of IPSJ*, Aug. 2006, vol. 2006-MUS-66, pp. 77–84, in Japanese.

[12] H. Kameoka, N. Ono, K. Kashino, and S. Sagayama, "Complex NMF: A new sparse representation for acoustic signals," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2009, pp. 3437–3440.

[13] M. Suzuki, T. Hosoya, A. Ito, and S. Makino, "Music information retrieval from a singing voice using lyrics and melody information," *EURASIP J. Applied Signal Process.*, vol. 2007, no. 1, pp. 151–151, 2007.

[14] A. Mesaros and T. Virtanen, "Automatic recognition of lyrics in singing," *EURASIP J. Audio, Speech, and Music Process.*, vol. 2010, no. 4, pp. 1–7, 2010.

[15] H. Fujihara, T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Singer identification based on accompaniment sound reduction and reliable frame selection," in *Proc. Int. Symposium Music Info. Retrieval*, 2005, pp. 329–336.

[16] M. Ryynänen, M. Virtanen, J. Paulus, and A. Klapuri, "Accompaniment separation and karaoke application based on automatic melody transcription," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2008, pp. 1417–1420.

[17] C.-L. Hsu and J.-S. R. Jang, "On the improvement of singing voice separation for monaural recordings using the mir-1k dataset," *IEEE Trans. Acoust., Speech, and Language Process.*, vol. 18, no. 2, pp. 310–319, 2010.

[18] Z. Rafii and B. Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 221–224.

[19] A. Ozerov, P. Philippe, R. Gribonval, and F. Bimbot, "One microphone singing voice separation using source-adapted models," in *Proc. IEEE Workshop Applications Signal Process. Audio Acoust.*, 2005, pp. 90–93.

[20] H. Tachibana, N. Ono, and S. Sagayama, "Singing voice enhancement in monaural music signals based on two-stage harmonic/percussive sound separation on multiple resolution spectrograms," *IEEE/ACM Trans. Audio, Speech and Language Process.*, vol. 22, no. 1, pp. 228–237, 2014.

[21] Z. Rafii, Z. Duan, and B. Pardo, "Combining rhythm-based and pitch-based methods for background and melody separation," *IEEE/ACM Trans. Audio, Speech and Language Process.*, vol. 22, no. 12, pp. 1884–1893, 2014.

[22] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Acoust., Speech, and Language Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[23] "Bss eval," `http://bass-db.gforge.inria.fr/bss_eval/`.

[24] "Mir-1k dataset," `https://sites.google.com/site/unvoicedsoundseparation/mir-1k`.