NMF-BASED SOURCE SEPARATION UTILIZING PRIOR KNOWLEDGE ON ENCODING VECTOR

Kisoo Kwon[⋆]*, Jong Won Shin*[†] *and Nam Soo Kim*[⋆]

[⋆]Dept. of Electrical and Computer Engineering and the INMC, Seoul National University, Seoul, Korea
[†]Dept. of Information and Communications, Gwangju Institute of Science and Technology, Gwangju, Korea
E-mail: kskwon@hi.snu.ac.kr, jwshin@gist.ac.kr, nkim@snu.ac.kr

ABSTRACT

Non-negative matrix factorization (NMF) is an unsupervised technique to represents a nonnegative data matrix with a product of nonnegative basis and encoding matrices. The encoding matrix for the training phase contains information on the pattern of how each basis vector is utilized. The histogram for each row of this matrix corresponding to a specific basis turned out to be sparse, while the level of sparsity varied significantly in each basis. In this paper, the distribution of each component of an encoding vector is modeled as an independent exponential or gamma distribution, and a new objective function with the log-likelihood of the current encoding vector is proposed. Experimental results on audio source separation demonstrate that the utilization of the prior knowledge on the encoding matrix based on sparse statistical models can enhance the source separation performance.

Index Terms— Nonnegative matrix factorization, encoding vectors, statistical model, source separation.

1. INTRODUCTION

Many approaches have been proposed for single channel source separation including independent component analysis, sparse decomposition, principal component analysis, and singular value decomposition [1]- [5]. Among them, the methods based on nonnegative matrix factorization (NMF) have shown impressive results [6]- [17]. NMF basically approximates a nonnegative data matrix $\mathbf{V}$ with a product of nonnegative basis and encoding matrices $\mathbf{W}$ and $\mathbf{H}$, i.e., $\mathbf{V} \approx \mathbf{WH}$ [18]. Since both $\mathbf{W}$ and $\mathbf{H}$ are nonnegative, NMF often leads to a part-based representation of the data, which may be desirable in many areas including image or visual signal processing, text information processing, audio signal processing, and music information retrieval [6], [19].

Most of the NMF-based source separation approaches compute the basis matrix $\mathbf{W}$ from a set of given training data and then it is used for specific source separation [6]- [14]. The encoding matrix for the training data, $\mathbf{H}^{train}$, is usually removed although it has some useful information on how often each basis was utilized. In [8], a multivariate Gaussian distribution is applied to model the distribution of the logarithm of the encoding vector, and the log-likelihood of the current estimate for the encoding vector is incorporated in the objective function. However, our analysis on $\mathbf{H}^{train}$ revealed that each row of this matrix was also highly sparse, which implies the lognormal distribution may not be the best form of prior knowledge. Another noteworthy observation was that the level of sparsity widely varies in each basis. It may suggest that each component in the encoding vector should contribute to the sparsity-related penalty differently.

In this paper, we propose the penalty terms based on the prior knowledge on $\mathbf{H}$ in the separation phase for NMF-based source separation. The new statistical models for the encoding vector are proposed based on the analysis of the distribution of the encoding matrix elements in the training data. The distribution of each component of an encoding vector is modeled as an independent exponential or gamma distribution of which the parameters are estimated from $\mathbf{H}^{train}$. The log-likelihood of $\mathbf{H}$ derived from these models is adopted in the objective function in the separation stage. The additional log-likelihood term derived from the exponential distribution turns out to be a weighted L_1 norm penalty. Experimental results on audio source separation show that the proposed method can enhance the separation performance in terms of the perceptual evaluation of speech quality (PESQ) [20] and the signal-to-distortion ratio (SDR) [21].

2. NMF-BASED SPEECH ENHANCEMENT

NMF is mainly applied on the speech magnitude or power spectrogram to obtain a set of basis vectors to represent a

This research was supported in part by the National Research Foundation of Koera (NRF) grant funded by the Korea government (MEST) (NRF-2015R1A2A1A15054343), and by the MSIP(Ministry of Science, ICT and Future Planning), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2015-H8501-15-1016) supervised by the IITP(Institute for Information & communications Technology Promotion).

speech signal. In our notations, an input nonnegative matrix $\mathbf{V} \in \mathbb{R}_+^{M \times N}$ is approximated by the product of a basis matrix $\mathbf{W} \in \mathbb{R}_+^{M \times r}$ and an encoding matrix $\mathbf{H} \in \mathbb{R}_+^{r \times N}$ ($\mathbf{V} \approx \mathbf{WH}$) where M and N denote the number of frequency bins and short-time frames, respectively, and r is the number of basis vectors. We apply NMF on the magnitude spectrogram of speech and noise. Hence, speech basis matrix $\mathbf{W}_s \in \mathbb{R}_+^{M \times r_s}$ is trained using a clean speech DB, and noise basis matrix $\mathbf{W}_n \in \mathbb{R}_+^{M \times r_n}$ is obtained from a noise DB, where r_s and r_n indicate the number of speech and noise basis vectors, respectively. In order to process an input with noisy magnitude spectrum, a joint basis matrix $\mathbf{W} = [\mathbf{W}_s \ \mathbf{W}_n] \in \mathbb{R}_+^{M \times (r_s + r_n)}$ is first constructed. In this paper, the Kullback-Leibler divergence (KLD) and multiplicative update rules (MuR) are used as a distance measure and an optimization method, respectively. The update rules for the encoding and basis matrices are as following [18]:

$$H_i \leftarrow H_i \otimes \frac{W_i^T \frac{V_i}{W_i H_i}}{W_i^T \mathbf{1}}, \quad W_i \leftarrow W_i \otimes \frac{\frac{V_i}{W_i H_i} H_i^T}{H_i^T \mathbf{1}} \qquad (1)$$

where $\bigotimes$ and $\frac{a}{b}$ denote the element-wise multiplication and division of matrices, and subscript i denotes either speech or noise, and $\mathbf{V}_i \in \mathbb{R}_+^{r \times N_i}$ is the magnitude spectrogram of the training signal where N_i is the total number of short-time frames in the training signal for source i, and $\mathbf{1}$ is a matrix of suitable size with all elements equal to one. H_i and W_i are obtained by iteratively applying the update rules (1) for a fixed number of iterations.

In the speech enhancement stage, speech and noise magnitude spectra are first approximated using NMF, and then a spectral gain is determined to obtain the enhanced speech signal. This is further explained in the following. Let $Y(t) \in \mathbb{C}^{M \times 1}$, $S(t) \in \mathbb{C}^{M \times 1}$, and $N(t) \in \mathbb{C}^{M \times 1}$ denote the short-time Fourier transform (STFT) coefficients of the noisy, clean speech, and noise signals, respectively, for the *t*-th frame. We assume an additive noise model, which is expressed as $Y(t) = S(t) + N(t)$. The input vector to the NMF analysis is the magnitude spectrum of the noisy signal in the present time frame, i.e., $V(t) = |Y(t)| \in \mathbb{R}_+^{M \times 1}$ where $|\cdot|$ denotes element-wise absolute value operator. The noisy magnitude spectrum $V(t)$ is approximated as $V(t) \approx \mathbf{W}H(t)$, where the basis matrix $\mathbf{W}$ is obtained during the training phase (as explained above), and $H(t) = [H_s(t)^T \ H_n(t)^T]^T \in \mathbb{R}_+^{(r_s)+r_n \times 1}$ denotes the encoding vector of the noisy signal in the *t*-th frame. With fixed $\mathbf{W}$, $H(t)$ is computed by iterating the left part of (1), in which $H_s(t)$ and $H_n(t)$ are initialized to nonnegative random numbers. After convergence or a fixed number of iterations of the algorithm, the speech and noise magnitude spectra are approximated as:

$$|\hat{S}(t)| = \mathbf{W}_s H_s(t), \qquad |\hat{N}(t)| = \mathbf{W}_n H_n(t). \qquad (2)$$

We adopt the gain function similar to Wiener filter to enhance the speech signal, where we use speech and noise approximations from (2). The gain function is given as

$$G(t) = \frac{|\hat{S}(t)|^2}{|\hat{S}(t)|^2 + |\hat{N}(t)|^2} \qquad (3)$$

The STFT coefficients of the speech signal at the *t*-th frame are now obtained according to $\hat{S}^{final}(t) = G(t) \bigotimes Y(t)$.

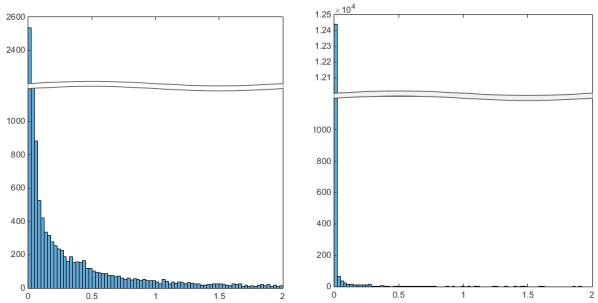


Fig. 1. The histograms of some rows of $\mathbf{H}_S^{train}$ corresponding to the most frequently and rarely used basis vectors.

3. NMF-BASED SOURCE SEPARATION UTILIZING PRIOR KNOWLEDGE ON ENCODING VECTOR

Though most of the previous works only use the trained basis matrix during the source separation phase, the encoding matrix obtained for the training data is regarded to posses important information as to how frequently each basis is utilized to reconstruct the clean source signals. In the training procedure, $\mathbf{W}_S \in \mathbb{R}_+^{M \times r_s}$ and $\mathbf{H}_S^{train} \in \mathbb{R}_+^{r_s \times N_s}$ are obtained through the NMF analysis of the clean target signal data $\mathbf{V}_S^{train} \in \mathbb{R}_+^{M \times N_s}$ while $\mathbf{W}_N \in \mathbb{R}_+^{M \times r_n}$ and $\mathbf{H}_N^{train} \in \mathbb{R}_+^{r_n \times N_n}$ are computed from the noise signal data $\mathbf{V}_S^{train} \in \mathbb{R}_+^{M \times N_n}$. In [8], the distribution of the logarithm of the elements of an encoding vector is modeled as r-dimensional multivariate Gaussian distribution of which the parameters are estimated from $\mathbf{H}_S^{train}$ and $\mathbf{H}_N^{train}$ assuming that H_S and H_N are independent. Based on this statistical model, the log-likelihood of the current estimate of H for the test data V is subtracted from the objective function used in the source separation phase. Although the utilization of the prior knowledge on H brought about performance improvement, the method in [8] has a high complexity and the chosen statistical model does not fit to the actual distribution of the data as the $\mathbf{H}_S^{train}$ or $\mathbf{H}_N^{train}$. In this paper in order to alleviate this, we propose new statistical models for the elements of H. To analyze the actual histograms of the row elements of $\mathbf{H}_S^{train}$, we performed the standard NMF analysis on a database of clean speech magnitude spectra, i.e., $\mathbf{V}_S^{train} = [|Y(1)||Y(2)| \cdots |Y(N)|] \in \mathbb{R}_+^{M \times N}$ where $Y(t) \in \mathbb{C}^{M \times 1}$ is the $2(M-1)$-point STFT of speech at frame t and N is the total number of frames in the training data.

TIMIT database with a sampling rate of 16 kHz was used for the speech data, $M = 257$, and $N = 12,728$. Fig. 1 shows the histograms corresponding to the two rows of $\mathbf{H}_S^{train}$ having the one largest and smallest L_1 norms. There rows roughly correspond to the most frequently and rarely used basis vectors. The shape of the histograms shows sparse distributions which can be better approximated by a gamma or an exponential distribution rather than a lognormal distribution as used in [8]. Moreover, a wide range of sparsity level was observed along different bases. The analysis on the $\mathbf{H}_N^{train}$ for various noise signals also showed similar tendency.

Based on these empirical analyses, we propose in this work to prior distribution of each component of the encoding vector as an independent gamma or exponential distribution of which the parameters are estimated from the corresponding row of $\mathbf{H}_S^{train}$ or $\mathbf{H}_N^{train}$. The probability density function (pdf) of the gamma distribution is given by

$$P(x) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}, \tag{4}$$

where k, θ, and $\Gamma(\cdot)$ indicate a shape parameter, a scale parameter, and the gamma function, respectively. Because the correlation coefficients among different components of the encoding vector were found not so significant, we assumed that each component of the encoding vector is statistically independent to avoid heavy computation. The log-likelihood of the current estimate for H based on the assumed model is subtracted from the original objective function to form the modified objective function for the separation phase as given by

$$f(H) = D(V \mid \mathbf{W}H) - \gamma_g \sum_{i=1}^{r} [(k_i - 1) log H_i - \frac{H_i}{\theta_i}] \tag{5}$$

in which the constant term irrelevant of H is ignored. The MuR with KLD is now modified to

$$H_i \leftarrow H_i \frac{\sum_{k=1}^{M} \frac{\mathbf{W}_{k,i} V_k}{\sum_{f=1}^{r} \mathbf{W}_{k,f} H_f}}{\sum_{k=1}^{M} \mathbf{W}_{k,i} + \gamma_g (\frac{1-k_i}{H_i} + \frac{1}{\theta_i})}. \tag{6}$$

It is noted that $k_i \leq 1$ to match the shape of the distribution shown in Fig. 1.

Alternatively, we can employ an exponential distribution instead of the gamma distribution of which the pdf is given by

$$f(x;\eta) = \eta e^{-\eta x} \qquad x \geq 0, \tag{7}$$

where η is the rate parameter, which is the reciprocal of the mean. It is noted that it is a special case of the gamma distribution with $k = 1$. As in the case of the gamma distribution, the log-likelihood of H is combined with the KLD between V and $\mathbf{W}H$ so that the final objective function becomes

$$f(H) = D(V \mid \mathbf{W}H) + \gamma_e \sum_{i=1}^{r} (\eta_i H_i) \tag{8}$$

where η_i is the rate parameter for the distribution of H_i, and γ_e is the parameter controlling the trade-off between the reconstruction error and the log-likelihood. The MuR with KLD is now given by

$$H_i \leftarrow H_i \frac{\sum_{k=1}^{M} \frac{\mathbf{W}_{k,i} V_k}{\sum_{f=1}^{r} \mathbf{W}_{k,f} H_f}}{\sum_{k=1}^{M} \mathbf{W}_{k,i} + \gamma_e \eta_i}. \tag{9}$$

It is noteworthy that the penalty term in (8) becomes essentially the same as that of sparse NMF in [19] except that the weighting by η_i, which is the reciprocal of the L_1 norm or mean of the corresponding row of $\mathbf{H}_S^{train}$ or $\mathbf{H}_N^{train}$.

4. EXPERIMENTS

To evaluate the performance of the proposed algorithm, it is applied to audio source separation in which the target signal is speech. The proposed constraints in the separation stage based on the prior distribution of H modeled by the gamma and exponential distributions were compared with that based on the lognormal distribution and L_1 norm constraint in terms of the PESQ and and SDR.

Speech samples were chosen from TIMIT database while the noise signals used for the experiments were *F-16*, *factory1*, *babble*, and *machinegun* noises from the NOISEX-92 DB. Each signal was sampled at 16 kHz, and the Hamming window and a 512-point discrete Fourier transform with 75% overlap were applied to form a spectrogram. Training DB for speech contains of 102-second long speech data spoken by 40 different speakers, while the noise data for training were 117-second long in total for each type of noise which has the same level with speech data. To test the proposed and conventional methods, 32 sentences spoken by 32 different speakers which weren't included in the training DB were mixed with the aforementioned four types of noise data which were not used in the training at 0 dB SNR to construct the test DB. The MuR was applied with the distance measure of KLD in the NMF analysis, and the numbers of iterations for the training and test phases were 100 and 30, respectively, and the each number of bases was 128 ($r_s = r_n = 128$).

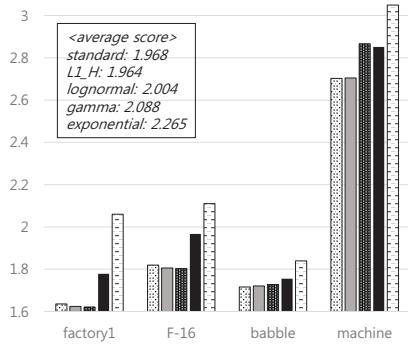
In this experiment, the training procedure is implemented as in II.A without any constraint, and various penalty terms were utilized to compute H in the source separation stage which is then used to enhance the signal as in II.B. The penalty terms used in the experiments were:

•*standard*: no constraint as in (1)

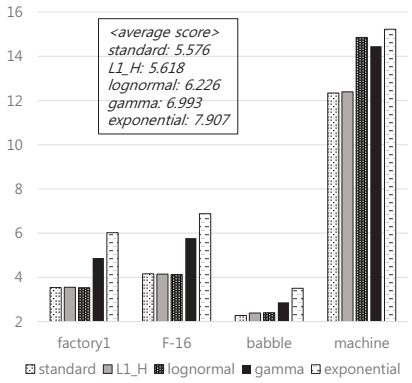
•*L1*: L_1 norm of H in (9) with $\eta_i = 1$

•*lognormal*: the negative log-likelihood of $logH$ assuming that H follows lognormal distribution as in [8] where $logA$ denotes element-wise logarithm.

•*gamma*: the negative log-likelihood of H in which the PDF of H is modeled as an independent gamma distribution, which is shown in (5).



(a) The PESQ scores



(b) SDRs

Fig. 2. Source separation performance of NMF methods with various prior models of H ($r_s = r_n = 128$).

•*exponential*: the negative log-likelihood of H where the distribution for H is assumed to be an independent exponential distribution, which is given in (8).

L1 is included since the penalty term based on the exponential distribution results in the weighted L_1 norm of H. The parameters for the lognormal and exponential distributions were determined by the 1st and 2nd order moments of $\mathbf{H}_S^{train}$ and $\mathbf{H}_N^{train}$, while those for the gamma distribution were achieved by the maximum likelihood estimation through the MATLAB function "fitdist". The parameters for each penalty term were chosen to maximize the source separation performance, which fall in the ranges $\lambda_{L1} \in [0.001, 1]$, $\gamma_{lognormal} \in [0.001, 0.3]$, $\gamma_g \in [0.01, 0.04]$ and $\gamma_e \in [0.005, 0.02]$. Fig. 2 shows the performance of the source separation when the input SNR was 0 dB. For both of the cases, the penalty terms based on sparse distributions outperformed other penalty terms. One interesting observation is that the system based on the exponential distribution performed better than that based on the gamma distribution although the exponential pdf is a special case of the gamma pdf. One possible interpretation is that the objective function in (5) is not convex if $k_i < 1$, in contrast to the exponential modeling which leads to a convex objective function. It can be seen from (6) that once H_i has a very small value, it becomes smaller very quickly and cannot reach the global optimum. Also, it is noted that the *lognormal* required heavy computation when r_s and r_n are large due to the multivariate modeling, while the proposed approaches were processed much faster. One potential issue for the prior model-based approaches is that the performance of the systems is questionable if the power level of the test data is significantly different from that of the training data based on which the model parameters are estimated. The effect of the power level mismatch may not be crucial since the KLD term would regulate the difference between $\mathbf{V}$ and $\mathbf{W}H$ and the effect of the power mismatch have impact on all the element of H. To verify this, another set of experiments have been carried out for the test data which are with 10 dB higher or 10 dB lower power level. Table 1 summarizes the source separation performance with these data and the original data with the matched level, when the same parameters to Fig. 2 were used. The proposed method based on the exponential model resulted to be very robust to the level mismatch. As for the system based on the gamma distribution, the performance increased as the input level got higher, possibly because it was less likely to fell into local optima near the small values of H_i if the input level was higher.

Table 1. Source separation performance when the power level of the test DB differs from that of the training DB.

distribution	Power level of the test data	PESQ	SDR
standard	original	1.9681	5.5763
lognormal	original	2.0043	6.2264
	+10dB	1.9904	6.1562
	-10dB	2.0052	6.0510
gamma	original	2.0878	6.9930
	+10dB	2.1717	7.4682
	-10dB	1.9320	4.0748
exponential	original	2.2648	7.9071
	+10dB	2.2415	7.9153
	-10dB	2.2188	7.8422

5. CONCLUSION

In this paper, to utilize the statistical information on the encoding vector obtained during the training, we propose an additional penalty term in the test phase which is the negative log-likelihood of the encoding vector based on a sparse distribution such as an exponential or a gamma distribution. Experimental results show that the empirical distribution of each encoding vector component was actually sparse, and the proposed methods can enhance the source separation performance in terms of PESQ and SDR when applied the audio source separation task in which the target signal is speech.

6. REFERENCES

[1] A. Belouchrani, K. A. Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique based on second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, pp. 434-44, Feb. 1997.

[2] M. Zibulevsky and B. Pearlmutter, "Blind source separation by sparse decomposition in a signal dictionary," *Neural Comput.*, vol. 13, no. 4, 2001.

[3] M. E. Davies and C. J. James, "Source separation using single channel ICA," *Signal Process.*, vol. 87, no. 8, pp. 1819-1832, 2007.

[4] A. Hyvarinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Net*, vol. 13, no. 4-5, pp. 411-430, Jun. 2000.

[5] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation*. New York: Wiley, 2009.

[6] P. Smaragdis, C. Fevotte, G. Mysore, N. Mohammadiha, and M. Hoffman, "Static and dynamic source separation using nonnegative factorizations: A unified view," *IEEE Signal Process. Mag.*, vol. 31, no. 3, pp. 66-75, 2014.

[7] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis," *Neural computation*, vol. 21, no. 3, pp. 793-830, 2009.

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

[9] K. Kwon, J. W. Shin, and N. S. Kim, "Target Source Separation Based on Discriminative Nonnegative Matrix Factorization Incorporating Cross-Reconstruction Error," *IEICE Trans. on Information and Systems*, vol. E98-D. no. 11, pp.-, Nov. 2015.

[10] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 3, pp. 1066-1074, 2007.

[11] K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based speech enhancement using bases update," *IEEE Signal Process. Lett.*, vol. 22, no. 4, pp. 450-454, Apr. 2015.

[12] T. G. Kang, K. Kwon, J. W. Shin, and N. S. Kim, "NMF-based target source separation using deep neural network," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 229-233, Feb. 2015.

[13] N. Mohammadiha, T. Gerkmann, and A. Leijon, "A new linear MMSE filter for single channel speech enhancement based on nonnegative matrix factorization," *In Applications of Signal Processing to Audio and Acoustics (WASPAA), 2011 IEEE Workshop on*, pp. 45-48, 2011.

[14] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. on Audio, Speech, and Language process.*, vol. 15, no. 1, Jan. 2007.

[15] G. J. Mysore, P. Smaragdis, and B. Raj, "Non-negative hidden markov modeling of audio with application to source separation," in *Int. Conf. Latent Variable Analysis and Signal Separation*, pp. 140-148, 2010.

[16] C. Joder, F. Weninger, D. Virette, and B. Schuller, "A Comparative Study on Sparsity Penalties for NMF-based Speech Separation: Beyond LP-Norms," *Acoustics, Speech and Signal Processing (ICASSP)*, pp. 858862, 2013.

[17] G. Zhou, A. Cichocki, Q. Zhao, and S. Xie, "Nonnegative matrix and tensor factorizations," *IEEE Signal Process. Mag.*, vol.31, no.3, pp. 54-65, May 2014.

[18] D. D. Lee and H. S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[19] A. Pascual-Montano, J. Carazo, K. Kochi, D. Lehmann, and R. D. Pascual-Marqui, "Nonsmooth nonnegative matrix factorization (NSNMF)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 3, pp. 403-415, Mar. 2006.

[20] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," Tech. Rep. ITU-T P.862, 2001.

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