Incremental Approach to NMF Basis Estimation for Audio Source Separation

Kisoo Kwon*, Jong Won Shin†, Inkyu Choi*, Hyung Yong Kim*, Nam Soo Kim*
* Dept. of Electrical and Computer Engineering and the INMC,
Seoul National University, Seoul, Korea
E-mail: {kskwon, hykim, nkim}@hi.ac.kr Tel/Fax: +82-2-884-1824
† School of Electrical Engineering and Computer Science,
Gwangju Institute of Science and Technology, Gwangju, Korea.
E-mail: jwshin@gist.ac.kr Tel/Fax: +82-62-715-2235

***Abstract*—Nonnegative matrix factorization (NMF) is a matrix factorization technique that might find meaningful latent nonnegative components. Since, however, the objective function is non-convex, the source separation performance can degrade when the iterative update of the basis matrix is stuck to a poor local minimum. Most of the research updates basis iteratively to minimize certain objective function with random initialization, although a few approaches have been proposed for the systematic initialization of the basis matrix such as the singular value decomposition. In this paper, we propose a novel basis estimation method inspired by the similarity of the bases training with the vector quantization, which is similar to Linde-Buzo-Gray algorithm. Experiments of the audio source separation showed that the proposed method outperformed the NMF using random initialization by about 1.64 dB and 1.43 dB in signal-to-distortion ratio when its target sources were speech and violin, respectively.**

I. Introduction

Over the recent years, nonnegative matrix factorization (NMF) has been widely applied to many applications, and it has shown impressive performances particularly in image and audio signal processing [1]-[16]. NMF is a sort of latent factor analysis technique for which unsupervised learning algorithms are used to discover part-based representations underlying the given nonnegative data. NMF has shown certain benefits compared with other factorization schemes such as the independent component analysis and principal component analysis [1], [2], and it has been generally known that the part-based representation is suitable for audio signal analysis [3]. Several distance metrics are employed such as the Euclidean distance (EuD), Kullback-Leibler divergence (KLD), alpha-divergence, beta-divergence and Itakura Saito-divergence, and various optimization methods have been developed for each distance measure [2], [4], [5]. Since the publication of [1], numerous attempts have been made to improve NMF under some specific conditions, which include sparse NMF [6], Itakura-Saito NMF [2], convolutive NMF [7], discriminative NMF [8], and so on.

In the NMF-based source separation, the basis matrix $\mathbf{W}$ plays an important role. That is, the source separation performance depends on how $\mathbf{W}$ is trained. Therefore, because of the non-convexity of the specified objective function for NMF analysis, the form of $\mathbf{W}$ is different from the initialization and the optimized solution can be stuck on a local minima. It implies that the overall performance significantly may depend on the initial parameter values. For this reason, some previous works apply the centroids of k-means clustering, singular value decomposition (SVD)-based method for the basis and encoding matrices during the training phase [17]-[22]. Though some of these methods show a lower reconstruction error and a faster convergence speed than those of the random initialization, they do not consider the part-based feature of NMF.

The conventional vector quantization task can be interpreted as a special case of the matrix factorization where each basis vector corresponds to a codeword and only a single basis is activated at each time [23]. This analogy implies that the data clustering techniques may be used for the initialization or estimation of the NMF bases. Some of the previous works initialize the NMF basis by the centroids of the training data clusters [17]-[21]. Unfortunately, conventional codebook training approaches such as the k-means algorithm can only guarantee suboptimal solutions similar to the case of NMF bases estimation and the final centroids are sensitive to the initialization of the code vectors, and it is close to exemplar-based approach if the number of cluster is somewhat large. In [22], SVD-based approach is proposed and it shows a low reconstruction error, but it cannot support the over-completed bases.

In this paper, we propose a novel approach to estimate the basis for the NMF analysis. The proposed method is based on Linde-Buzo-Gray (LBG) [24] algorithm well-known in the area of data compression and clustering. One of the prominent features of this algorithms is that it increases the number of bases in each step of the procedure. In order to evaluate the performance of the proposed technique, we carry out an experiment on target source separation. In the experimental result, we can see that the proposed method outperformed the other basis estimation methods.

II. NMF-based audio source separation

In this paper, we particularly focus on audio signal separation where a given data set $\mathbf{V}\in \mathbb{R}^{m\times n}$ represents a data set of n magnitude spectra. It is noted that the variables in boldface capital letters denote matrices, while those which are not in the boldface represent vectors. NMF approximates the data set $\mathbf{V}$ as 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, n, and r denote the numbers of frequency bins, short-time frames, and basis vectors, respectively. For the objective function, Euclidean distance (EuD) is employed as the distance measure in this work. The optimization of the objective function is based on the projected gradient descent (PGD) method where the learning rate is decided according to [4]. The update rules of the encoding and basis matrices during the training phase are given as [4]

$$\mathbf{H} \leftarrow \mathbf{H} - \alpha_H(\mathbf{W}^T\mathbf{WH} - \mathbf{WW}^T\mathbf{V}), \tag{1}$$

$$\mathbf{W} \leftarrow \mathbf{W} - \alpha_W(\mathbf{WHH}^T - \mathbf{VH}^T). \tag{2}$$

Each learning rate, α_H and α_W, becomes bigger until the condition which guarantees the sufficient decrease of the objective function is satisfied [4]. $\mathbf{H}$ and $\mathbf{W}$ are obtained by iterative application of the update rules (1) and (2) for a fixed number of iterations, and $\mathbf{W}$ and $\mathbf{H}$ are usually initialized with nonnegative random values [1]-[16]. During the training phase, the target and interfering basis matrices, $\mathbf{W}_S$ and $\mathbf{W}_N$, are trained by each clean data set, separately.

In the separation phase, a noisy magnitude spectrum $|Y(t)|$ is approximated as $|Y(t)| \approx \mathbf{W}H(t)$ at each frame *t* with the fixed basis matrix $\mathbf{W} = [\mathbf{W}_S\ \mathbf{W}_N]$ obtained during the training phase where $H(t) = [H_S(t)^T\ H_N(t)^T]^T\in \mathbb{R}^{(r_s+r_n)\times 1}$ denotes the encoding vector of the mixed signal in the *t*-th frame, and $Y(t)$ and $|\cdot|$ denote the short-time Fourier transform (STFT) coefficients of the noisy input and the element-wise magnitude, respectively. Keeping $\mathbf{W}$ fixed, $H(t)$ is computed by iterating (1) for a fixed number of times in which $H_S(t)$ and $H_N(t)$ are initialized to nonnegative random numbers. After a fixed number of iterations, the magnitude spectra of the target and interfering signals are estimated as follows:

$$|\hat{S}(t)| = \mathbf{W}_S H_S(t), \qquad |\hat{N}(t)| = \mathbf{W}_N H_N(t). \tag{3}$$

Instead of directly using the estimated magnitude spectra in (3), a spectral gain function similar to the Wiener filter is adopted in [3] and [14].

III. Incremental approach to the NMF basis estimation

Because the general objective function of NMF is biconvex in $\mathbf{W}$ and $\mathbf{H}$, different algorithms and their initializations lead to different solutions. It means that NMF algorithm does not satisfy the uniqueness [25]. To get more accurate solutions with complex models, carefully designed initializations or regularizations may be needed. Since, however, it is hard to exploit a general prior knowledge of the parts of a source data, nonnegative random values have been widely applied for NMF initialization. Although this method has shown a somewhat proper performance experimentally [1], [20], due to the non-covexity of the objective function and iterative nature of the algorithm, it cannot be considered to provide an optimal initial point for successful NMF analysis.

Fig. 1. Pseudo code for the proposed incremental approach to the NMF basis estimation

```
Input: Matrix $\mathbf{V} = (V_1, V_2, \cdots, V_n)\in \mathbb{R}^{m\times n}$,
       integer $k$
Output: Matrix $\mathbf{W}\in \mathbb{R}^{m\times 2^k}$
  1. $W^0 = \frac{centroid(\mathbf{V})}{\|centroid(\mathbf{V})\|_1 \mathbf{1}}$
  2. $H^0 = \mathbf{V}^T W^0$
  for $i = 0 : k-1$
    3. $\mathbf{W}^{i+} = \mathbf{W}^i + \epsilon$, $\mathbf{W}^{i-} = \mathbf{W}^i - \epsilon$
    5. $\mathbf{W}^{temp} = [\mathbf{W}^{i+}, \mathbf{W}^{i-}]\in \mathbb{R}^{m\times 2^{i+1}}$
    6. $\mathbf{H}^{temp} = [(\mathbf{H}^i/2)^T, (\mathbf{H}^i/2)^T]^T\in \mathbb{R}^{2^{i+1}\times n}$
    7. Do NMF process by $\mathbf{W}^{temp}$ and $\mathbf{H}^{temp}$
       $\longrightarrow \mathbf{W}^{i+1}$, $\mathbf{H}^{i+1}$
    8. $\mathbf{W}^{i+1} = \frac{\mathbf{W}^{i+1}}{\|\mathbf{W}^{i+1}\|_1 \mathbf{1}}$
  end
  9. $\mathbf{W} = \mathbf{W}^k$
```

In this section, we propose novel approach to estimate the basis for NMF analysis which is based on the clustering approach. The proposed method is motivated by accepting a general premise that the best basis is the centroid of the whole training DB when the number of bases is set to one. The core idea of the proposed approach is to estimate the bases incrementally. The incremental approach where a new centroid is searched at each step can be a good strategy for basis estimation.

LBG algorithm is the most cited and widely used algorithm on designing the vector quantization codebook [24], and it is similar to the k-means algorithm in data clustering. At each iteration of LBG, each vector is split into two new vectors. LBG algorithm can be employed in the making of the basis. In this case, its number of bases doubles in every procedure.

The pseudo code for the proposed incremental approach to the NMF basis estimation is given in Fig. 1. The input of the algorithm is the training data matrix $\mathbf{V}\in \mathbb{R}^{m\times n}$ and the integer k. The output is the matrix composed of 2^k basis vectors. $\mathbf{W}^i$ and $\mathbf{H}^i$ in Fig. 1 denote the basis matrix with 2^i bases and the corresponding encoding matrix, respectively. In order to decide the single basis case, $W^0\in \mathbb{R}^{m\times 2^0}$ is obtained by the centroid of $\mathbf{V}$ with unit-norm normalization. $\|\cdot\|_1$, $\mathbf{1}$, and *centroid*$(\mathbf{V})$ represent unit-norm, a vector of a proper size with all elements equal to one, and the centroid of the matrix $\mathbf{V}$, respectively. For the encoding of W^0, H^0 which minimizes the Euclidean distance is given in a closed-form as $H^0 = \mathbf{V}^T W^0$. For the bases $\mathbf{W}^1\in \mathbb{R}^{m\times 2^1}$, $[W^{0+}, W^{0-}]$ is applied for the initialization of *NMF process* where $W^{0+}\in \mathbb{R}^{m\times 2^0}$ and $W^{0-}\in \mathbb{R}^{m\times 2^0}$ are obtained by addition and subtraction of a very small value ϵ to W^0, respectively. *NMF process* indicates the alternative update phase, (1) and (2), for a fixed number of iterations. The initial value of $\mathbf{H}^1$ is given as $[(\mathbf{H}^0/2)^T, (\mathbf{H}^0/2)^T]^T\in \mathbb{R}^{2^1\times n}$. The procedures from 3 to 8 are repeated until we get 2^k bases.

IV. Experiment

To evaluate the performance of the proposed algorithm, audio source separation was performed in a variety of noisy conditions, and the target sources were speech and violin signals. The whole data for the training and test was not overlapped. A 512-point discrete Fourier transform with 75% overlap was used to form the spectrogram with a sampling rate of 16 kHz ($m = 257$).

The performance of the proposed methods were evaluated in terms of PESQ [26] and SDR [27]. To demonstrate the performance improvement achieved by the proposed methods, four source separation systems for which only the basis matrices were trained in different ways were compared:

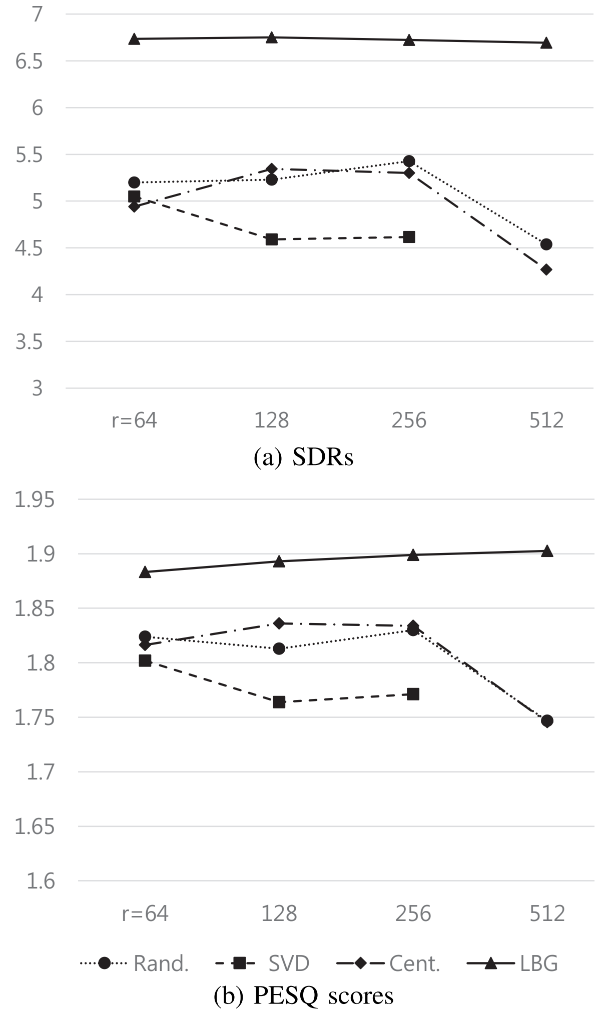
- *Rand.*: the initialization of nonnegative random values [1]
- *SVD*: the initialization in [22] which utilizes the result of singular value decomposition
- *Cent.*: the initialization in [17] which utilizes the centroids from the k-means clustering (The number of the cluster is the same to the number of the basis vectors.)
- *LBG*: the proposed method using the LBG-based basis estimation for NMF ($\epsilon = 1^{-14}$)

The whole bases of above systems were trained by PGD [4], on the other hand, the source separation was performed by multiplicative update rule with KLD [1]. This is because such a system experimentally shows a high performance and a low computational time of the separation when random nonnegative values are used for NMF initialization. Each number of iteration during the training phase was decided based on the separation performance [28]. (*Rand.*=*SVD*=*Cent.*=100, *LBG*=10.)

A. The target source: speech signal

Speech and noise samples were selected from TIMIT [29] and NOISEX-92 [30] DBs, respectively. The basis matrix for each noise types was obtained from about 120-second long noise signal, and the speech DB for the training was 130-second long spoken by 56 different speakers. The speech test data set consisted of 32 sentences from 32 different speakers. We tested 4 different types of noises including *F-16*, *factory1*, *babble*, and *white* noises. The numbers of the bases were 64, 128, 256, and 512, and the number of iteration for the separation phase was 30.

Fig. 2 shows the PESQ scores and SDRs when the input signal-to-noise ratio (SNR) was 0 dB. For all cases of r, the proposed algorithm outperformed other methods in terms of both the PESQ score and the SDR. In particular, *LBG* produced the best separation performance, and the optimal r for the performance was different from the case of *Rand.*. In the over-complete case, $r = 512 > m$, the performance of *Rand.* was the lowest, but the proposed methods maintained the performance or outperformed the case of $r \leqq 256$. One of the previous method, *SVD*, showed the minimum reconstruction error during our training phase, but it cannot support the over-complete case and its separation performance was lower than *Rand*. This result denotes that the performance of source separation is not proportional to the reconstruction error during the training phase, and the proposed method which utilizes the incremental approach may extract the proper character of the source to the bases.



(a) SDRs

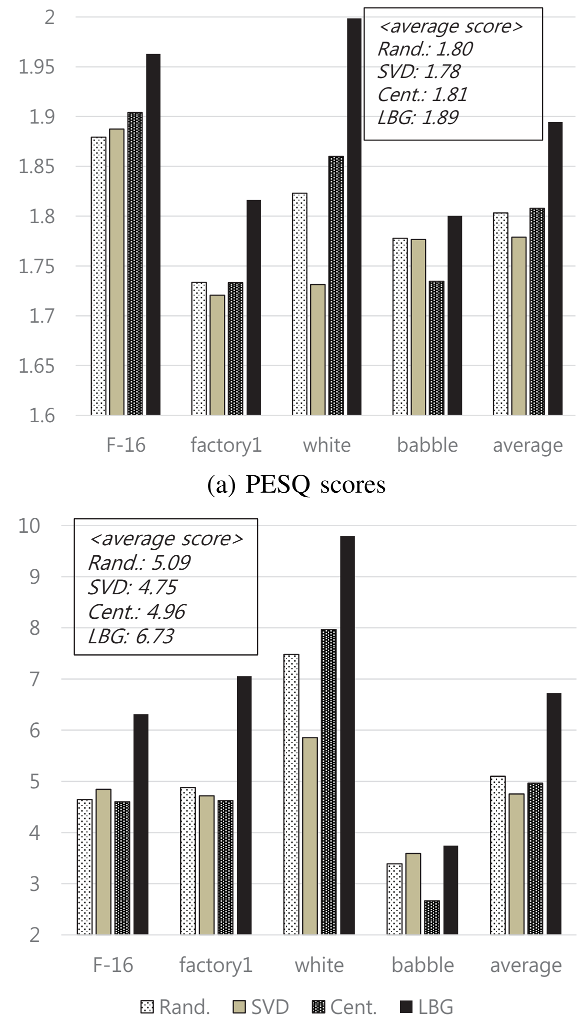
(b) PESQ scores

Fig. 2. The source separation performances based on the numbers of basis (target source = speech, input SNR = 0 dB)

Fig. 3 denotes the source separation performance as the interfering sources, *F-16*, *factory1*, *babble*, and *white* signals. In the same manner as Fig. 2, the proposed method *LBG* outperformed other methods in the face of all interfering sources. The previous works, *SVD* and *Cent.*, show a similar performance as *Rand.*, but the performance degraded when *white* and *babble* signals are mixed, respectively. *LBG* made a performance improvement at every interfering sources, and it showed a prominent improvement in term of SDR.

B. The target source: violin signal

For violin and piano data, we used four songs to the training and separation phases, and Table. I shows the information of the data. The interfering sources were 4 different types of sources including *piano*, *factory1*, *babble*, and *machinegun* sources, and the test data set of violin consisted of 10 clips



(a) PESQ scores

(b) SDRs

Fig. 3. The source separation performances based on the interfering sources (target source = speech, input SNR = 0 dB)

TABLE I
THE INFORMATION OF THE DATA FOR THE BASES ESTIMATION AND SOURCE SEPARATION (RESAMPLED TO 16KHZ/S)

The phase	Title	Artist	Instrument
Training	Partita No.1 (BWV 1002) - Double	Ida Haendel	Violin
	Blind Film	Yiruma	Piano
Separation	Sonata No.2 (BWV 1003) - Allegro	Ida Haendel	Violin
	Waltz In C Minor (Only For Piano)	Yiruma	Piano

which are 5 seconds long each.

The number of bases r for each source was set to 128, which provided a good trade-off between the reconstruction error and the computational complexity. The experimental results when the input SNR is 0 are illustrated in Fig. 4. The proposed algorithms outperformed other methods at all interfering types. In particular, LBG shows a high performance improvement when *piano* or *factory1* source is mixed. The performance improvements of *LBG* were 1.43 and 2.05 in term of the SDR over *Rand.* and *SVD*, respectively.

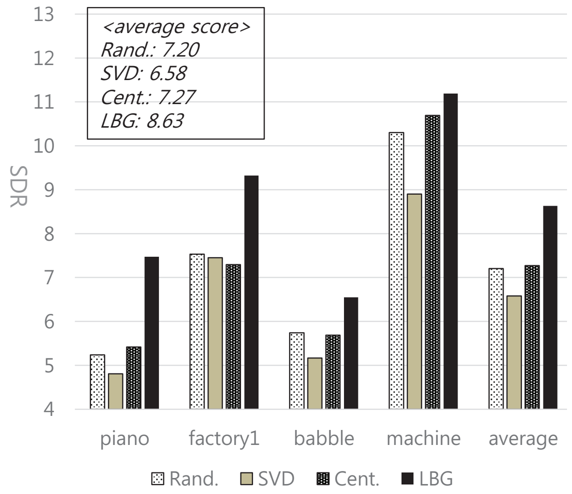


Fig. 4. The source separation performances as the interfering sources (target source = violin, input SNR = 0 dB, r=128)

V. CONCLUSION

This paper proposed the basis estimation based on the incremental approach. Since the objective function of NMF is non-convex, the optimized solution can be stuck to a local minima which implies that the overall performance significantly depends on the initial parameter values, and this influences not only the reconstruction error during the training phase, but also the performance of the source separation. The novel proposed algorithm applies LBG algorithm which is the vector quantization codebook. Since the vector quantization task can be interpreted as a special case of the matrix factorization, the vector quantization algorithm like LBG can be utilized to NMF algorithm. The whole experimental results of this paper may imply that the incremental approach for the training of bases is effective in making a good basis matrix for the audio source separation.

ACKNOWLEDGMENT

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 Basic Science Research Program through the National Research Foundation of Korea (NRF) (NRF-2016R1C1B1015291).

REFERENCES

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

[2] 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.

[3] K. W. Wilson, B. Raj, and P. Smaragdis, “Regularized non-negative matrix factorization with temporal dependencies for speech denoising,” *INTERSPEECH*, pp. 411-414, 2008.

[4] C. J. Lin, “Projected gradient methods for nonnegative matrix factorization,” *Neural computation*, vol. 19, no. 10, pp. 2756-2779, 2007.

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

[6] P. O. Hoyer, “Non-negative matrix factorization with sparseness constraints,” *The Journal of Machine Learning Research*, vol. 5, pp. 1457-1469, 2004.

[7] P. D. O’grady and B. A. Pearlmutter, “Convolutive non-negative matrix factorisation with a sparseness constraint,” *Machine Learning for Signal Processing 2006. Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on. IEEE*, pp. 427-432, 2006.

[8] F. Weninger, J. L. Roux, J.R. Hershey, and S. Watanabe, “Discriminative NMF and its application to single-channel source separation,” *Proc. of ISCA Interspeech*, pp.865-869, 2014.

[9] K. Kwon, J. W. Shin, and N. S. Kim, “Target source separation based on discriminative nonnegative matrix factorization incorporating cross-reconstruction error,” *IEICE Letter on Information and Systems*, Vol. E98-D, No. 11, pp.2017-2020, Nov. 2015.

[10] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, “Online learning for matrix factorization and sparse coding,” *The Journal of Machine Learning Research*, vol.11, pp. 19-60, 2010.

[11] N. Mohammadiha, P. Smaragdis, and A. Leijon, “Supervised and unsupervised speech enhancement using nonnegative matrix factorization,” *IEEE Trans. Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140-2151, 2013.

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

[13] N. Guan, D. Tao, Z. Luo, and B. Yuan, “Online nonnegative matrix factorization with robust stochastic approximation,” *IEEE Trans. Neural networks and Learning systems*, vol. 23, no. 7, pp. 1087-1099 Jul. 2012.

[14] K. Kwon, J. W. Shin, and N. S. Kim, “NMF-based speech enhancement using bases update,” *IEEE Signal Processing Letters*, vol. 22, no. 4, pp. 450-454, Apr. 2015.

[15] 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.

[16] M. N. Schmidt, J. Larsen, and F. T. Hsiao, “Wind noise reduction using non-negative sparse coding,” *Machine Learning for Signal Processing, 2007 IEEE Workshop on*. IEEE, pp.431-436, 2007.

[17] S. Wild, “Seeding non-negative matrix factorizations with spherical k-means clustering,” *Master’s thesis*, University of Colorado, 2003.

[18] S. Wild, J. Curry, and A. Dougherty, “Improving non-negative matrix factorizations through structured initialization,” *Pattern Recognition*, vol. 37, no. 11, pp. 2217-2232, 2004.

[19] Z. Zheng, J. Yang, and Y. Zhu, “Initialization enhancer for non-negative matrix factorization,” *Engineering Applications of Artificial Intelligence*, vol. 20, no. 1, pp. 101-110, 2007.

[20] R. Albright, J. Cox, D. Duling, A. N. Langville, and C. D. Meyer, “Algorithms, initializations, and convergence for the nonnegative matrix factorization,” *presented at the 12th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, Aug. 2006.

[21] Gong, Liyun, and Asoke K. Nandi. “An enhanced initialization method for non-negative matrix factorization.” Machine Learning for Signal Processing (MLSP), 2013 IEEE International Workshop on. IEEE, 2013.

[22] C. Boutsidis and E. Gallopoulos, “SVD based initialization: A head start for nonnegative matrix factorization,” *Pattern Recognition*, vol. 41, no. 4, pp. 1350-1362, 2008.

[23] A. Singh and G. Gordon, “A unified view of matrix factorization models,” *Mach. Learn. Knowledge Discovery Databases*, vol. 5212, pp.358-373, 2008.

[24] Y. Linde, A. Buzo, and R. Gray, “An algorithm for vector quantizer design,”*IEEE Transactions on Communications*, vol. com-28, no. 1, pp. 84-95, Jan. 1980.

[25] T. Virtanen, J. F. Gemmeke, B, Raj, and P. Smaragdis, “Compositional models for audio processing: Uncovering the structure of sound mixtures,” *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 125-144, 2015.

[26] “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.

[27] E. Vincent, R. Gribonval, and C. Fvotte, “Performance measurement in blind audio source separation,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp.1462-1469, 2006.

[28] P. Smaragdis, “Convolutive speech bases and their application to supervised speech separation,” *IEEE Transactions on Audio, Speech, and Language processing*, vol. 15, no. 1, pp. 1-12, Jan. 2007.

[29] L. Larnel, R. Kassel, and S. Seneff, “Speech database development: design and analysis of the acoustic-phonetic corpus,” *Proc. DARPA Speech Recognition Workshop*, pp. 26-32, Mar. 1987.

[30] A. Varga, H.J.M Steenneken, M. Tomlinson, and D. Jones, “The NOISEX-92 study on the effect of additive noise on automatic speech recognition,” 1992. Documentation included in the NOISEX-92 CD-ROMs.