MODEL ORDER SELECTION FOR NON-NEGATIVE MATRIX FACTORIZATION WITH APPLICATION TO SPEECH ENHANCEMENT

Nasser Mohammadiha and Arne Leijon

KTH Royal Institute of Technology, Sound and Image Processing Lab., Stockholm, Sweden
phone: + (46) 8 790 9267, nmoh@kth.se

ABSTRACT

This report deals with the application of non-negative matrix factorization (NMF) in speech processing. A Bayesian NMF is used to find the optimal number of basis vectors for the speech signal. The result is validated by performing a speech enhancement task for a set of different number of basis vectors. The algorithm performance is measured with the Source to Distortion Ratio (*SDR*) that represents the overall quality of speech. The results show that for medium input *SNR*s, 60 basis vectors for each speaker are sufficient to model the speech spectrogram. NMF produced better *SDR* results than a recently developed version of Spectral Subtraction algorithm. The window length was found to have a great effect on the results, but zero padding did not influence the results.

1. INTRODUCTION

Speech enhancement has focused on the suppression of additive background noise in the past decades. Recently Non-negative Matrix Factorization (NMF) and its extensions have been used for audio signal modeling. Consequently, they are used for blind audio source separation, speech enhancement, and other audio applications.

In general, given any non-negative matrix X of dimension $W \times K$, NMF finds non-negative matrices T and V such that $X \cong TV$, where T and V are of dimensions $W \times I$ and $I \times K$ respectively. To do the factorization, we have to define a cost function and minimize it. There are many possibilities to define the cost function $D(X\|TV)$. Information divergence has been found to produce good results in audio signal analysis [1], and a Bayesian framework is developed for minimizing it [2].

For NMF based speech enhancement or audio source separation, X is the spectrogram of the signal which is derived using a given window length, while spectra are stored column-wise in X. NMF is applied to factorize the spectrogram into non-negative factors, and the separation or enhancement is performed. In this factorization, different columns of T are basis vectors (spectra) of speech, and V is a mixing matrix.

The performance of this approach depends on the window length; so far, long windows have been considered in the literature. NMF was proposed for speech denoising in [3]; the time and frequency dependencies in the speech signal were modeled and used during the optimization. The results of the proposed approach outperformed the Wiener filtering approach; however, a long window (64 ms) was used in this report. Different priors were assigned to T and V in [4] and [5], and the result of factorization was used for speech separation; a window length around 40 ms was used in these works.

In the application of speech enhancement and source separation techniques for hearing aids, the imposed delay by an algorithm plays an important role. A delay of 20 ms is just acceptable while a delay of 40 ms is disrupting for subjects [6]; hence, it is important to investigate the performance of NMF with short delays.

The other important factor in NMF based approaches is the number of basis vectors (I), and there has not been any study specifically for speech enhancement application to show how one can choose an optimal number for I. In [7] the effect of different parameters including window length and the number of basis vectors are studied using a validation set for source separation problem (cross-talk speakers). NMF is used for single-channel speech separation in [8], and it is shown that increasing I will improve the performance of separation.

In this report, we study different aspects of NMF based speech enhancement. Following the proposed Bayesian framework in [2], an optimal number of basis vectors is obtained for the speech signal, and the result is validated by a performance-based model selection. The Source to Distortion Ratio (*SDR*) [9] is used for algorithm evaluation. The effect of window length and zero padding are also studied using a validation set. A recently developed version of Spectral Subtraction is used for comparison with NMF.

2. NON-NEGATIVE MATRIX FACTORIZATION

2.1 Maximum Likelihood Solution For NMF

We have used information divergence cost function which is defined as equation (1).

$$D(X\|TV) = \sum_{v,\tau}([X]_{v,\tau}\log\frac{[X]_{v,\tau}}{[TV]_{v,\tau}} + [TV]_{v,\tau} - [X]_{v,\tau}) \quad (1)$$

where $[\cdot]_{v,\tau}$ is used to refer to an entry of a matrix; furthermore, $v = 1...W$, $i = 1...I$ and $\tau = 1...K$. Traditionally the following multiplicative update rules have been used iteratively to minimize equation (1) [10]:

$$\begin{aligned} [T]_{v,i} &\leftarrow [T]_{v,i}\frac{\sum_\tau [V]_{i,\tau}\left([X]_{v,\tau}/[TV]_{v,\tau}\right)}{\sum_p [V]_{i,p}}, \\ [V]_{i,\tau} &\leftarrow [V]_{i,\tau}\frac{\sum_v [T]_{v,i}\left([X]_{v,\tau}/[TV]_{v,\tau}\right)}{\sum_q [T]_{q,i}}. \end{aligned} \quad (2)$$

where iterations will be continued to achieve a defined convergency criterion.

This work was supported by the EU Initial Training Network AUDIS (grant 2008-214699).

Following the statistical framework from [2], we can assume Poisson distribution for the data as it is shown below:

$$[X]_{v,\tau} = \sum_i [S]_{v,i,\tau} \qquad [S]_{v,i,\tau} \sim PO([S]_{v,i,\tau}; [T]_{v,i}[V]_{i,\tau})$$

where latent sources $[S]_{v,i,\tau}$ are assumed to have Poisson probability mass function which is defined as

$$PO(s;\lambda) = \frac{1}{\Gamma(s+1)}\lambda^s e^{-\lambda}$$

and $\Gamma(s+1) = s!$. Since the sum of some Poisson random variables will have a Poisson distribution, the probability mass function of $[X]_{v,\tau}$ is given by: $p([X]_{v,\tau}) = PO([X]_{v,\tau}; \sum_i [T]_{v,i}[V]_{i,\tau})$; thus, X is assumed to be discrete.

The Maximum Likelihood estimates of T and V were derived using the Expectation Maximization algorithm and the result is identical to equation (2).

2.2 Bayesian NMF

In the Bayesian framework, a conjugate prior for T and V is the Gamma distribution:

$$[T]_{v,i} \sim G([T]_{v,i}; a^t, b^t) \quad [V]_{i,\tau} \sim G([V]_{i,\tau}; a^v, b^v) \qquad (3)$$

where a^t, b^t, a^v, and b^v are prior hyperparameters. The Gamma probability density function is defined as

$$G(t;a,b) = \frac{1}{b^a \Gamma(a)} t^{a-1} e^{-\frac{t}{b}}$$

By using these priors, we assume that each component of the basis matrix (T) and mixing matrix (V) are drawn independently from the distributions given in (3). In the Bayesian approach, in addition to the posterior distributions of the parameters ($p(T,V,S|X)$), an important quantity is the *model evidence* which is obtained by integrating out all the parameters. Let $\Theta = \{a^t, b^t, a^v, b^v\}$ denote all the prior hyperparameters, and I be the number of basis vectors in the NMF model.

Suppose we wish to compare a set of NMF models with different number of basis vectors. The *model evidence* is obtained as

$$p(X;\Theta,I) = \int \sum_S p(X|S,T,V;\Theta,I)p(S,T,V;\Theta,I)dTdV \qquad (4)$$

In a simple approximation, the *model evidence* can be viewed as subtraction of two terms; the first term, with plus sign, is the likelihood of data under the model assumption which increases with increasing the model complexity, the number of basis vectors here. The other term is the complexity penalty which again increases with increasing the model complexity[11]. In general, *model evidence* can be used as a model selection criterion which gives a compromise between data fitting and model complexity. The *model evidence* can be also used for optimizing the prior hyperparameters:

$$\Theta^* = \arg\max_{\Theta} p(X;\Theta,I)$$

Even with the conjugate priors, the required integration to calculate the *model evidence* is not tractable; hence, using the Variational Inference (VI) a lower bound can be derived and maximized in an iterative approach. It turns that in the VI framework, the posterior distributions for T and V are gamma distributions while the posterior distribution for S is a multinomial distribution. The (posterior) hyperparameters for these distributions are updated iteratively until a defined convergency criterion is reached. The prior hyperparameters (Θ) are also optimized in each iteration by maximizing this lower bound . The derivations and update rules can be found in [2].

3. NMF BASED SPEECH ENHANCEMENT

The speech signal is first windowed into frames and their spectra are obtained. Stacking these spectra into a matrix yields a spectrogram. We have used the magnitude spectrogram in our simulations. Although NMF based approach can be unsupervised (mainly for music separation) or semisupervised, here we consider a supervised algorithm only. NMF based speech enhancement is a supervised approach and needs training. During the training, NMF is applied to the clean speech and noise signals, and the basis matrices for speech and noise are found:

$$\begin{aligned}(T_S^*, V_S^*) &= \arg\min_{T_S,V_S} D(X_S \| T_S V_S) \\ (T_N^*, V_N^*) &= \arg\min_{T_N,V_N} D(X_N \| T_N V_N)\end{aligned} \qquad (5)$$

In equation (5), X_S and X_N are speech and noise spectrograms respectively. We have used one speech model for each speaker. T_S, V_S are the basis and mixing matrices for speech, and T_N, V_N are the basis and mixing matrices for the noise signal.

Now, we put T_S^* and T_N^* into a bigger basis matrix: $T^* = [T_S^* \quad T_N^*]$. Having a noisy signal, we calculate its spectrogram, X_M. Keeping the basis matrix fixed, NMF is applied and the mixing matrix is obtained as:

$$V_M^* = \arg\min_{V_M} D(X_M \| T^* V_M) \qquad (6)$$

V_M^* is partitioned as $V_M^* = [V_{SM}^T \quad V_{NM}^T]^T$ where V_{SM} describes the contribution of speech, and V_{NM} describes the contribution of noise in the observed noisy signal. Now the following mask is calculated:

$$M = \frac{T_S^* V_{SM}}{T^* V_M^*} \qquad (7)$$

In (7) division is performed element-wise. Usually the obtained mask is applied to the noisy signal as $X_{\tilde{S}} = M .* X_M$ where $X_{\tilde{S}}$ is the magnitude spectrogram of the enhanced speech, and $.*$ is element-wise multiplication. We now give a new interpretation to what is being done; the algorithm consists of two steps: 1-Calculate the instantaneous estimate of the magnitude spectrogram of the noise in the current frame using (8).

$$X_{\tilde{N}} = (1-M) .* X_M = \frac{T_N^* V_{NM}}{T^* V_M^*} .* X_M \qquad (8)$$

where 1 ia a matrix with all ones. Like (7), division is element-wise in (8) again. 2-Obtain the instantaneous estimate of the magnitude spectrogram of the enhanced speech

by $X_{\tilde{S}} = \max(0, X_M - X_{\tilde{N}})$. In fact, this is just an instantaneous magnitude subtraction algorithm with NMF as a method for estimating the magnitude spectrogram of noise. $X_{\tilde{S}}$ can be converted to a time domain signal with the spectrogram inversion and using the phase information of the noisy signal.

Applying this mask will introduce time aliasing distortion because of the inherent properties of DFT; to prevent that, the obtained mask must be zero padded first, and then the zero-padded mask is applied to the noisy signal. To do so, M is converted to its time domain form by inverse FFT ; then the zero padding is performed in the time domain; by applying FFT to the zero-padded filter the frequency response of the zero-padded filter is obtained, and is used for the enhancement purpose. After this zero padding, the frequency response of the filter will be smoother.

Implementation of equations (5) and (6) are done in the Maximum Likelihood manner using the multiplicative rules from equation (2) due to its simplicity and high speed. However, we use the Bayesian NMF to find the optimal number of basis vectors. The magnitude spectrogram of the training data of each speaker is used to obtain the *model evidence* using equation (4) for that particular speaker. As mentioned in section 2.2, an iterative approach based on the Variational Inference is used to calculate a lower bound for the integration in equation (4); the algorithm was found to be very sensitive to the initial values. Since the speech signal is sparse, we assign some sparse and broad prior distributions to T and V according to equation (3). For this purpose, a^t, b^t are chosen such that the mean of the prior distribution for T is close to zero, and its variance is very high. On the other hand, a^v, b^v are chosen such that the prior distribution of V has a high mean accounting for the data scale and high variance to represent uncertainty. To have good initializations for posterior means, the multiplicative NMF from equation (2) is applied first to find an approximately locally optimum solution for the KL divergence, and the result is used as the initial values for the posterior means. Using these initializations, the Variational Inference approach becomes more stable. Finally, the lower bounds for the logarithm of the *model evidences* from different speakers are summed, and the result is considered as the criterion for the model selection and used in the simulations.

4. EXPERIMENTAL SETUP

We used the *Grid Corpus* and *NOISEX-92* databases in our simulations. All the training and test sentences are downsampled to 16 kilohertz. 4 male and 4 female speakers are chosen, and the final results are obtained by averaging over all the speakers. For training, a 120-second speech signal is obtained by concatenating 100 sentences for each speaker, and for noise training a 60-second noise signal is used. For the test purposes, a 40-second speech signal is obtained by concatenating 30 sentences and used for each speaker; moreover, the test sentences are chosen carefully to avoid having a sentence in both the training and test sets. In this report, we present the results for Babble noise since it is a difficult noise to deal with, but similar results were obtained for factory and car noises as well. To study the effect of input Signal to Noise Ratio (*SNR*) the simulations are carried out for 0, 5, and 10 dB *SNR*s. For the noise signal 100 basis vectors (I) are derived but to evaluate the effect of the number of basis vectors for speech, it is varied between 15 to 250.

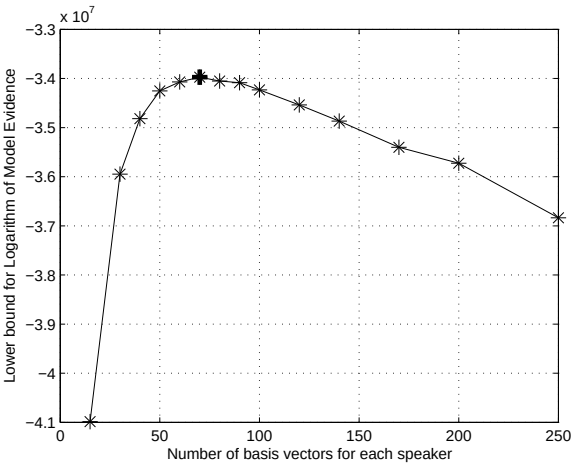

Figure 1: *Bayesian Model Selection for speakers. A window with 800 samples is used for the spectrogram derivation. The lower bounds for the logarithm of the model evidences from different speakers are summed, and the result is shown in the figure.*

To evaluate the performance of algorithms the Source to Distortion Ratio (*SDR*) is used:

$$SDR = 10\log \frac{\|s_{target}\|^2}{\|e_{interf} + e_{artif}\|^2}$$

where s_{target},e_{interf} and e_{artif} are target speech signal, interference, and artifact error terms defined in [9], and $\|\cdot\|^2$ denotes the energy. *SDR* has been shown to represent the overall quality of speech when reducing noise and absence of "burbling" artifacts are equally important and it has a high correlation with the Mean Opinion Score [12]. We used an overlap-add procedure with *hann* window and 50% overlap between adjacent frames. Let us denote the window length by w. In the following, NMF and NMFzp refer to NMF algorithm used without zero padding and with zero padding respectively. FFT length is chosen as w and $2w$ samples for NMF and NMFzp in turn. The total delay imposed by the NMF and NMFzp is around $1.5w$ and $2w$ samples correspondingly. To study the effect of the window length, we use a long window with $w = 800$ samples, 75 ms delay for NMF and 100 ms delay for NMFzp, and a short window with $w = 160$ samples, 15 ms delay for NMF and 20 ms delay for NMFzp which is just acceptable for hearing aid users.

5. RESULTS AND DISCUSSIONS

5.1 Model Selection

Figure 1 presents the result of the Bayesian model selection. This result is obtained by applying Bayesian NMF to the training clean speech with a narrow band spectrogram (window length = 800 samples). We see that the lower bound for the logarithm of the *model evidence* is maximized for $I = 70$, but the lower bound is rather flat from $I = 50$ to $I = 90$. If we use a wide band spectrogram instead, the lower bound will be maximized for some smaller number of basis vectors.

To check the result of the Bayesian model selection, a

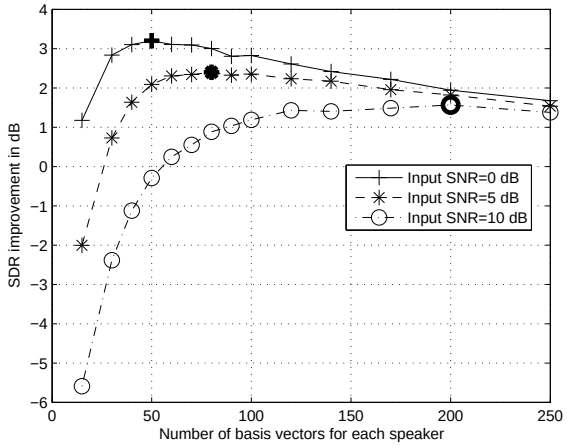

Figure 2: *Performance-based Model Selection for speakers using NMF based speech enhancement. "SDR improvement" is used to evaluate the performance of a given model. A window with 800 samples is used for the overlap-add procedure.*

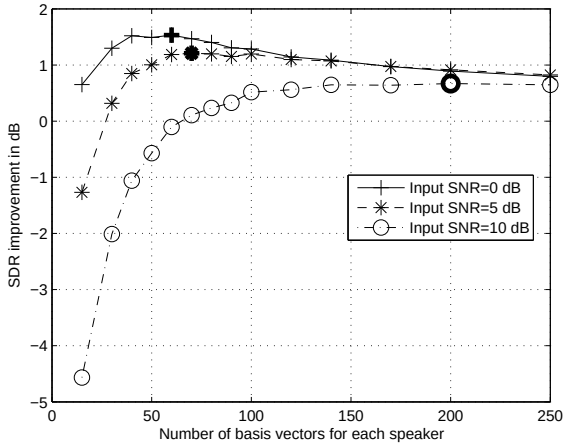

Figure 3: *Performance-based Model Selection for speakers using NMF based speech enhancement. "SDR improvement" is used to evaluate the performance of a given model. A window with 160 samples is used for the overlap-add procedure.*

performance-based model selection is carried out. To do so, the speech enhancement task is performed for a set of different parameters, and the *SDR* is measured. Figures 2 and 3 show '*SDR* improvement' versus 'number of basis vectors' for three different input *SNR*s and two window lengths.

For a long window and input *SNR*s equal to 0 and 5 dB, the result of the performance-based model selection coincides with the result of the Bayesian model selection well (Figures 1 and 2). In this case the best choice for the number of basis vectors is around 60 to 70. For the shorter window the maximum performance is obtained for I in the range $50-70$. We should remember that with a short window with $w = 160$ samples, the spectrogram has 80 frequency bins. In general $(80I + IK)$ parameters must be estimated from $80K$ data points. When $I > 80$ the problem becomes underdetermined. This causes the lower bound to fluctuate for $I > 80$ in the Bayesian model selection, which is not shown here. Nevertheless, considering the performance-based model selection we see that even for $I > 80$, the performance does not reduce rapidly. We can conclude that the number of basis vectors does not depend heavily on the window length, at least considering the performance-based model selection, and a choice around 60 basis vectors for each speaker can produce nearly optimal results for input *SNR*s around $0-5$ dB.

With increasing the input *SNR*, more basis vectors are required to get the best performance. Since the quality of the noisy signal improves with increasing the input *SNR*, adding some basis vectors will capture some high quality data into the enhanced signal and the overall performance will increase. From Figures 2 and 3 we see that for 10 dB input *SNR*, the *SDR* improvement is maximized for $I = 200$ basis vectors; however, using only 120 basis vectors is sufficient to get a nearly maximum performance.

5.2 Performance Comparison

In this section, we study the effect of zero padding, window length, and input *SNR* in the enhancement performance. To compare the performance of NMF with other approaches, a perceptually optimized Spectral Subtraction (PSS) algorithm is used [13]. The result is shown in Figure 4. Based on the result of Model Selection in section 5.1, which was obtained using the training data, 60 basis vectors are used for each speaker for 0 and 5 dB *SNR*s, and 120 basis vectors are used for 10 dB *SNR* case.

We see that zero padding does not improve the performance; though, it causes some extra delay. It seems that omitting zero padding will introduce a distortion which is small compared to the other distortions. We also see that the window length has a great effect on the performance such that the *SDR* improvement is almost twice for the longer window. With increasing *SNR*, up to a limit, the performance of the PSS algorithm improves while the performance of the NMF approach degrades (the amount of enhancement is reduced). PSS is based on Spectral Subtraction and with a high quality signal, it can achieve better performance, but NMF based algorithm is good at noise reduction, and when the input signal has a high quality, its performance is limited.

We also see that the NMF approach with a long window produces much better results than the PSS. For example for $SNR = 0$ dB, the difference for the *SDR* improvement is around 1.7 dB. Using a shorter window will deteriorate the performance of the NMF; nevertheless, even with a window having only 160 samples, the NMF produces a higher *SDR* improvement than the PSS. Although the *SDR* is found to represent the overall quality of speech and is a compromise between reducing the noise and introducing artifacts, the final judgement about the quality of the enhanced signal must be done by actual listeners, and the results of this study might need some perceptual evaluations. Some audio examples are available at "http://www.ee.kth.se/∼nmoh/ModelOrderSelection".

6. CONCLUSIONS

We applied a Bayesian NMF to the spectrogram of the speech signal. We used the *model evidence* to find the optimal

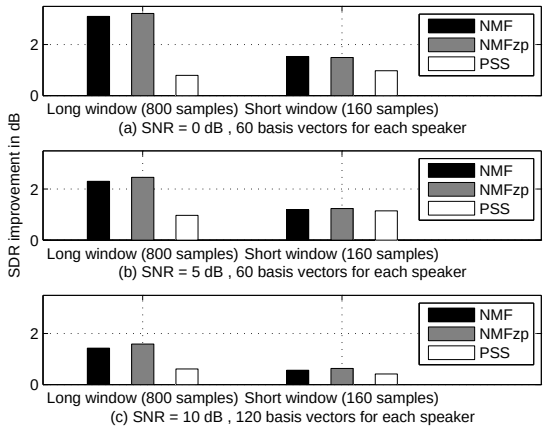

Figure 4: *Comparing the performance of different algorithms. 'NMF' is the NMF approach with multiplicative update rules. 'NMFzp' is the NMF approach with zero padding. 'PSS' is the perceptually optimized Spectral Subtraction*

number of basis vectors and the result was validated with a performance-based model selection. The result showed that for an input *SNR* in the order of $0-5$ dB, 60 basis vectors for each speaker are sufficient to get a nearly optimal enhancement while for higher *SNR*s more basis vectors are required. The simulations showed that by using a short window in the overlap-add procedure the performance of the NMF degrades, but still it can achieve better results than the PSS. The extra zero padding was shown to have no influence in the results.

REFERENCES

[1] Tuomas Virtanen, "Monaural sound source separation by non-negative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, 2007.

[2] Ali Taylan Cemgil, "Bayesian inference for nonnegative matrix factorisation models," *Technical Report CUED/F-INFENG/TR.609,Cambridge University Engineering Department*, 2008.

[3] Kevin Wilson , Bhiksha Raj, and Paris Smaragdis, "Regularized non-negative matrix factorization with temporal dependencies for speech denoising," in *Int. Conf. on Spoken Language Processing*, 2008, pp. 411–414.

[4] Tuomas Virtanen, and Ali Taylan Cemgil, "Mixtures of gamma priors for non-negative matrix factorization based speech separation," in *Int. Conf. ICA and BSS*, 2009, pp. 646–653.

[5] Steven J. Rennie , John R. Hershey, and Peder A. Olsen, "Efficient model-based speech enhancement and denoising using non-negative subspace analysis," in *Int. Conf. on Communications, Circuits and Systems*,2008, pp. 1833–1836.

[6] M.A. Stone, and B.C.J. Moore, "Tolerable hearing-aid delays. I. Estimation of limits imposed by the auditory path alone using simulated hearing losses," *Ear Hearing*, vol. 20, no. 3,pp. 182–192, 1999.

[7] Paris Smaragdis, "Convolutive Speech Bases and Their Application to Supervised Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1–12, 2007.

[8] Mikkel N. Schmidt, and Rasmus K. Olsson, "Single-channel speech separation using sparse non-negative matrix factorization," in *Int. Conf. on Spoken Language Processing*, 2006.

[9] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[10] Daniel D. Lee and H. Sebastian Seung, "Algorithms for Non-negative Matrix Factorization," in *NIPS*,2000.

[11] Christopher M. Bishop, *Pattern Recognition and machine learning*. Springer, 2006.

[12] Mingu Lee, and Inseok Heo, and Nakjin Choi, and Koeng-Mo Sung, "On evaluation of blind audio source separation," in *Int. Conf. of Audio Engineering Society* , 2008.

[13] H. Luts, and K. Eneman, and J. Wouters, et al, "Multicenter evaluation of signal enhancement algorithms for hearing aids," *Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1491–1505, 2010.