ROTATIONAL RESET STRATEGY FOR ONLINE SEMI-SUPERVISED NMF-BASED SPEECH ENHANCEMENT FOR LONG RECORDINGS

*Jun Zhou**

Southwest University
Dept. of Computer Science
Beibei, Chongqing 400715, China
zhouj@swu.edu.cn

Shuo Chen, Zhiyao Duan

University of Rochester
Dept. of Electrical and Computer Engineering
Rochester, NY 14627, USA
schen76@ur.rochester.edu; zhiyao.duan@rochester.edu

ABSTRACT

Non-negative matrix factorization (NMF) has been successfully applied to speech enhancement in non-stationary noisy environments. Recently proposed online semi-supervised NMF algorithms are of particular interest as they carry the two nice properties (online and semi-supervised) of classical speech enhancement approaches. These algorithms, however, have only been evaluated using noisy mixtures shorter than 30 seconds. In this paper we find that these algorithms work well when it is run for less than 1 minute, but degradation of the enhanced speech signal starts to appear after 2 minutes. We analyze that the reason is due to the inappropriate dictionary update rule, which gradually loses its ability in updating the speech dictionary. We then propose a simple rotational reset strategy to solve the problem: Instead of continuously updating the entire speech dictionary, we periodically and rotationally select elements and reset their values to random numbers. Experiments show that this strategy successfully solves the degradation problem and the improved algorithm outperforms classical speech enhancement algorithms significantly even when they are run for 10 minutes.

Index Terms— Speech enhancement, non-stationary noise, non-negative matrix factorization, source separation

1. INTRODUCTION

Speech enhancement is widely used in telecommunications, hearing aids, and robust speech recognition. It aims to improve the quality and intelligibility of noisy speech by reducing noise [1]. Classical speech enhancement algorithms can be categorized into four kinds: spectral subtraction [2], Wiener filtering [3], statistical-model-based [4], and subspace algorithms [5]. These algorithms share two nice properties in real-world applications: First, they are *semi-supervised*, i.e., a statistical model for the noise is calculated from noise-only excerpts but not for speech. Second, they are *online* algorithms hence useful in real-time applications, i.e., the enhancement of the current time frame does not depend on future frames. However, these algorithms cannot work well with non-stationary noise such as computer-keyboard-typing noise and babble noise, due to the fundamental assumptions of the noise models [6].

Non-negative Matrix Factorization (NMF) [7] and its mathematical equivalence, Probabilistic Latent Component Analysis (PLCA) [8], have shown promising results in separating non-stationary sound sources, and have been applied in speech enhancement in non-stationary noisy environments [9]. Among the many

algorithms, online semi-supervised algorithms proposed in recent years [6, 10–12] are of particular interest, as they hold the two nice properties (online and semi-supervised) of classical speech enhancement methods. These algorithms pre-learns the noise (speech) dictionary from noise-only (speech-only) training excerpts, and then updates the speech (noise) dictionary during separation.

These online semi-supervised NMF-based algorithms have shown promising results in various experiments in non-stationary noisy environments. However, the noisy speech utterances in the experiments are all shorter than 30 seconds. In fact, to our best knowledge, most of the existing NMF-based (not only online semi-supervised) speech enhancement methods [9, 11, 13–15] only use files shorter than 30 seconds for evaluation. While the length of test files may not matter for supervised or offline NMF methods, we argue that it does matter for online semi-supervised approaches. For these approaches, the dictionary of one source needs to be updated from the past but no theoretical results exist to guarantee the appropriateness of the updates over a long period, especially when the underlying source whose dictionary needs to be updated evolves rapidly over time.

In this paper, we make the first investigation of the effect of the test file length to the performance of online semi-supervised NMF-based speech enhancement algorithms. We base our analysis on a representative algorithm [10], which has been shown to outperform classical algorithms in non-stationary noisy files of about 15 seconds long. In this algorithm, the noise dictionary is pre-learned and the speech dictionary is updated during separation. We find that severe distortion on the enhanced speech signals starts to appear when the algorithm is ran for more than 2 minutes. We analyze this problem and find that over time the speech dictionary becomes sparser and sparser hence explains less and less energy of the mixture spectrogram. This suggests that the speech dictionary multiplicative update rule is inappropriate. Other online semi-supervised NMF-based speech enhancement algorithms [6, 11, 12] use a similar multiplicative update rule and similar system designs (e.g., sliding window/buffer and warm initialization). Therefore, we believe that the degradation problem is universal in existing online semi-supervised NMF-based methods.

In this paper, we propose a simple way to solve this problem by periodically reset elements in the speech dictionary to random values. These elements are selected in a rotational fashion. By doing so we "reboot" the update process of the speech dictionary. We compare the improved algorithm with the original one [10] and four classical speech enhancement algorithms, on long noisy speech files that contain multiple non-overlapping speakers. Results show that the improved algorithm successfully solves the degradation prob-

---

lem, and it outperforms the comparison methods significantly in various SNR conditions.

2. EXISTING ONLINE SEMI-SUPERVISED PLCA AND ITS DEGRADATION PROBLEM

PLCA is a mathematical equivalence of NMF. The basic idea of PLCA-based separation is to approximate each magnitude spectrum of the mixture signal $P_t(f)$ with $Q_t(f)$, a linear combination of spectral basis vectors from sources' dictionaries:

$$P_t(f) \approx Q_t(f) = \sum_{z \in \mathcal{S} \cup \mathcal{N}} P(f|z)P_t(z), \quad (1)$$

where $P(f|z)$ for $z \in \mathcal{S}$ represents the speech dictionary, and for $z \in \mathcal{N}$ represents the noise dictionary. $P_t(z)$ are the combination coefficients (or *activation weights*). The enhanced speech magnitude spectrum can then be obtained by $\sum_{z \in \mathcal{S}} P(f|z)P_t(z)$, and its time domain signal can be reconstructed by taking an inverse Fourier transform using the mixture signal's phase.

[10] is a representative online semi-supervised PLCA algorithm applied to speech enhancement. It assumes that the training data for noise but not for speech is available beforehand to train a noise dictionary. During separation, the noise dictionary is fixed while the speech dictionary and the activation weights of both dictionaries are estimated. Note that it is because of the varying activation weights that the fixed noise dictionary can model non-stationary noise. To make the separation online without having the estimated speech dictionary overfit the current mixture frame, the algorithm collects a moving buffer of past mixture frames that are likely to contain speech signals (detected by a Voice Activity Detection (VAD) module), and approximates the current mixture frame as well as the weighted buffer frames:

$$\underset{\substack{P(f|z) \text{ for } z \in \mathcal{S} \\ P_t(z) \text{ for } z \in \mathcal{S} \cup \mathcal{N}}}{\arg\min} d(P_t(f)||Q_t(f)) + \frac{\alpha}{L} \sum_{s \in \mathcal{B}} d(P_s(f)||Q_s(f)), \quad (2)$$

where $d(\cdot||\cdot)$ measures the mismatch between the mixture signal and its approximation. $\mathcal{B}$ represents the set of the L buffer frames; α is the tradeoff between the approximation of the current frame t and that of buffer frames.

To reduce the computational complexity, the algorithm updates the speech dictionary from its past values whenever it receives a new frame. The algorithm also fixes the activation weights of buffer frames as what have been estimated when enhancing those frames. These operations constitute a so called "warm initialization" strategy. The benefit is that the algorithm is much faster as it inherits information from the past. It is noted that the moving buffer (window) and warm initialization strategies are commonly used in other online semi-supervised NMF algorithms [6, 11, 12] as well.

This algorithm has been shown to outperform four kinds of classical speech enhancement algorithms in non-stationary noisy environments in [6], on noisy speech files about 15 seconds long of the same speaker in each file. As discussed in the introduction, we think that the length of test files may affect the performance significantly. Therefore, we create a number of long noisy speech files, each of which contains multiple speakers, to test the algorithm. Interestingly, enhancement performance degrades significantly over time. Figure 1 shows an example of the degradation phenomenon. Figure 1(a) shows the average Signal-to-Distortion Ratio (SDR) calculated by the the BSS_EVAL 3.0 toolbox [16] over 10 pieces
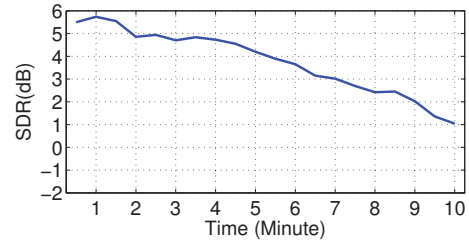
(a) Enhancement performance over time.
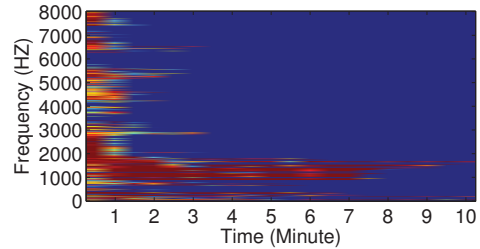
(b) Evolution of one basis vector in the speech dictionary over time. Red/blue shows high/low energy, respectively.

Figure 1: Illustration of the speech degradation problem of the original algorithm in [10].

of noisy speech files, each of which is 10 minutes long, created by mixing a clean speech file with a motorcycle noise file at the Signal-to-Noise Ratio (SNR) of 0 dB. Each speech file was created by concatenating speech sentences from 10 different speakers in a sequence with alternating genders, where each speaker takes about 1 minute. We can see that the SDR value starts to degrade significantly at around 2 minutes from the beginning, and never rebounds. We listened to the enhanced speech signals carefully and found that they sounded thinner (less full) over time. By the end of the file, the speech sounded very thin, although not much noise interference could be heard either.

3. PROBLEM ANALYSIS AND PROPOSED SOLUTION

This leads us to reason that the speech dictionary may gradually become sparser over time, so it cannot extract enough energy that should belong to speech from the mixture. To verify this thought, we visualize the evolution of the speech dictionary over time. There are in total 7 basis vectors and they all behave similarly. In Figure 1 we show the evolution of one vector. We can see that the basis vector indeed gradually becomes sparser. At the beginning of the file, many elements of the basis vector take large values. By the end of the file, however, most elements are close to zero. No wonder why the enhanced speech was thin at the end, as it was reconstructed using basis spectra that contained only a few sinusoids!

This indicates that there is some problem in the speech dictionary update process. In [10], the commonly used multiplicative update rule [17] is adopted to update the speech dictionary and activation weights. In each iteration, the speech dictionary $P(f|z)$ and the activation weights $P_t(z)$ are updated from their previous values by multiplying some factor:

$$P(f|z) \leftarrow \frac{1}{C_1} \sum_{s \in \mathcal{B} \cup \{t\}} V_{fs} P_s(z) \cdot P(f|z), \text{ for } z \in \mathcal{S}, \quad (3)$$

$$P_t(z) \leftarrow \frac{1}{C_2} \sum_f V_{ft} P(f|z) \cdot P_t(z), \text{ for } z \in \mathcal{S} \cup \mathcal{N}. \quad (4)$$

One problem of multiplicative update rule is that zero (or close-to-zero) elements will not get updated (or will be updated slowly). The warm initialization adopted in [10] initializes the speech dictionary in a new time frame as that has been updated in the previous frame. This speeds up the algorithm convergence when the speech characteristics do not change much. However, when the characteristics do change much (e.g., change of speaker or drastic pitch shift of the same speaker), the dictionary cannot be updated appropriately. For example, suppose the speaker changes from a female to a male, then the dictionary basis vector corresponding to a vowel of the male cannot be effectively updated from the basis vector of the female speaker, because the male vector should show high energy at his fundamental frequency, but the female vector is likely to show low energy at this frequency. Instead, the vector is likely to remain a low value at this frequency in the future. Therefore, the basis vector will become sparser and sparser over time. In other words, the speech dictionary will gradually lose its ability to adapt to new speech signals.

Having identified and analyzed the degradation problem, here we propose a simple solution for it. Instead of always initializing the speech dictionary with previously updated values, we reset the dictionary to random values once after a while. This will bring back the speech dictionary's potential to be adapted to new speech signals and prevent degradation in the enhanced speech. The problem of this solution, however, is that the random dictionary resulted from each reset will take much more iterations of updates before it could well explain the speech signal. This will cause significant fluctuations of speech dictionary quality and computation complexity of dictionary updates.

In this paper, we propose a *rotational reset* strategy: we periodically select and reset a subset of speech dictionary elements to random values, where the subsets are selected in a fixed rotational fashion. Let T be the *reset period*, M be the number of elements selected for reset in each period (*reset element amount*). Then the average *reset rate* is M/T. Compared to resetting the entire speech dictionary once for a while, this rotational reset strategy "smoothes out" the dictionary update process. While newly reset elements are recovering their potentials to adapt to new speech signals, old elements keep the continuity of the dictionary to prevent sudden changes in the enhanced signals.

Figure 2(a) shows the speech enhancement result and dictionary basis vector evolution over time using the proposed rotational reset strategy, on the same 10 noisy speech files as in Figure 1(a). We can see that the SDR of the proposed method stays around 5 dB and does not decrease over time. The basis vector in Figure 2(b) does not become sparse over time either. In fact, the values of each frequency bin can change from high to low and also low to high, to be adapted to different speech signals at different time frames.

4. EXPERIMENTS

We test the proposed strategy using 500 noisy speech files, each of which is about 10 minutes long. These files are obtained by adding clean speech files with noise-only files at different SNRs. We select 5 male and 5 female speakers from the PTDB-TUG speech corpus [18] and concatenate their randomly selected utterances to generate 10 different clean speech files. During the concatenation, male and female speakers are alternated to maximize the change of speech
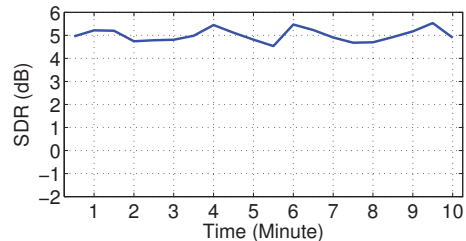
(a) Enhancement performance over time.
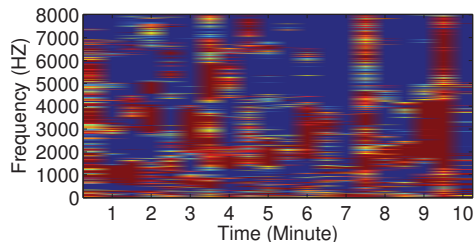
(b) Evolution of one basis vector in the speech dictionary over time. Red/blue shows high/low energy, respectively.

Figure 2: The proposed rotational reset strategy solves the speech degradation problem.

signals over time. Noise-only files are generated using the non-stationary noise dataset created in [10]. There are in total 10 kinds of noise: birds, casino, cicadas, computer keyboard, eating chips, frogs, jungle, machine guns, motorcycles, and ocean. Each noise file is at least one minute long. The first twenty seconds are used to train the noise dictionary beforehand. The rest is duplicated and concatenated to generate a long noise-only file to match up with each clean speech file. Clean speech files and their corresponding noise-only files are finally mixed with 5 different SNRs: -10, -5, 0, 5, 10 dB. The sampling rate of all the files is 16 kHz.

We first compare the proposed algorithm with four classical speech enhancement algorithms: spectral subtraction (*MB*) [2], Wiener filtering (*Wiener-as*) [3], statistical-model-based (*log-MMSE*) [4] and subspace algorithm (*KLT*) [5]. We use Loizou's implementations of these algorithms, as provided in [1]. Noise models of these algorithms are also calculated from the twenty seconds noise training excerpts and kept fixed. It is noted that noise tracking methods have been proposed in recent years to adapt noise models for non-stationary noise for the classical algorithms [19, 20]. In this paper, however, we only compare to the widely used basic algorithms. We also compare the improved algorithm with the original algorithm in [10]. We use two kinds of evaluation metrics. The first is PESQ [21] , which is a widely used objective speech quality measure. It ranges from 0.5 to 4.5, with a larger value for better quality. The second is Signal-to-Distortion Ratio (SDR), calculated using the BSS-EVAL 3.0 [16] toolbox. SDR is widely used in evaluating source separation algorithms, and it accounts for both interference removal and artifact introduction in the separated sources.

For the proposed algorithm, we segment each noisy speech file into frames of 64 ms with 48 ms overlap. We set the rotational reset period T to 60 seconds, and the reset element amount M to 4, as this parameter combination achieves good performance on the motorcycle noise with 0 dB SNR. All the other parameters (e.g., speech and noise dictionary sizes, buffer size, buffer tradeoff factor, number of

Table 1: Effect analysis of the rotational reset period (rows) and reset element amount (columns) on speech enhancement performance, using 10 noisy speech files with the motorcycle noise in 0 dB SNR.

SDR (mean±std)	1	2	3	4	5	6	7
5s	4.67±.07	4.54±.04	4.47±.03	4.39±.02	4.35±.04	4.30±.05	4.26±.04
15s	4.92±.07	4.91±.06	4.82±.05	4.79±.04	4.76±.05	4.71±.05	4.66±.05
30s	4.84±.16	4.95±.09	4.98±.06	4.97±.05	4.91±.05	4.92±.06	4.88±.07
60s	4.46±.54	4.94±.11	4.95±.05	5.03±.08	5.01±.04	5.00±.10	4.99±.07
120s	4.34±.78	4.72±.28	4.73±.37	4.75±.20	4.85±.17	4.91±.12	4.91±.14
240s	4.00±.57	4.14±.99	4.22±.78	4.35±.86	4.24±.87	4.50±.80	4.74±.13

iterations in each frame) are set to the same as those used in [10]. In particular, the speech dictionary size is 7, providing a compact dictionary with good speech reconstruction. The number of iterations in each frame is 20, which is enough for the convergence of the multiplicative update rule.
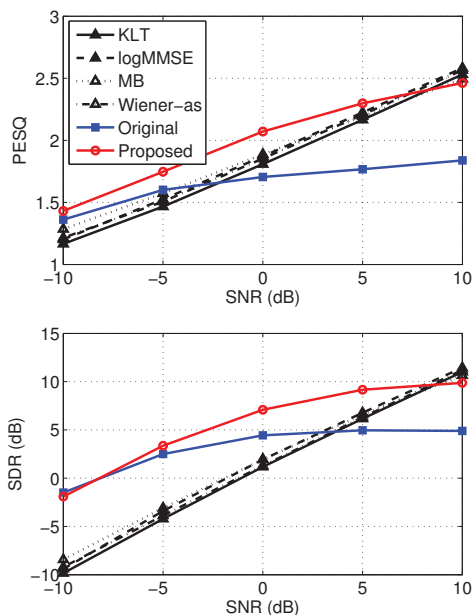

Figure 3: Overall comparison of the proposed algorithm with four classical speech enhancement methods and the original algorithm at different SNR conditions.

Figure 3 shows the comparison results. Each data point shows the average over 100 noisy speech files (10 files for each of the 10 kinds of noise). It can be seen that for both PESQ and SDR, the proposed algorithm improves from the original algorithm significantly for SNRs larger than -5 dB, and the improvement becomes more apparent as the SNR increases. This is reasonable as the strong speech signals in the mixture may "trap" the speech dictionary elements more easily after convergence in each frame in the original algorithm, causing more severe degradation. The improved algorithm achieves significantly better results than all the four classical algorithms for all SNRs less than 10 dB. As the original algorithm achieves worse results than classical algorithms for SNR larger than 5 dB due to degradation, the proposed strategy has successfully solved the problem.

In the second experiment, we conduct parameter sensitivity analysis on the two rotational reset parameters T and M. For T, we take values of 5, 15, 30, 60, 120, and 240 seconds, and for M, we take values of 1, 2, 3, 4, 5, 6, and 7. We run the algorithm with all these parameter combinations. As the reset rate equals to M/T, multiple combinations may share the same reset rate. One interesting question for this experiment is whether the reset rate is the key parameter, i.e., whether combinations corresponding to the same reset rate achieves similar results. We take 10 noisy speech files corresponding to the motorcycle noise to do the analysis.

Table 1 shows the results. There are several interesting findings. First, cells with the same or similar reset rate do show similar mean SDR values. This indicates that the reset rate is indeed the key parameter of the rotational strategy. For example, cells (15s, 1), (30s, 2), (60s, 4), and (120s, 7) all have about 4.95 dB, while cells (60s, 1), (120s, 2), and (240s, 4) all shows about 4.5 dB. Second, speech enhancement performance generally increases when the reset rate decreases from the upper right corner (5s, 7) and reaches the highest values in the middle part e.g., (30s, 4), but then decreases again when the reset rate becomes too fast e.g., (240s, 1). This suggests that the dictionary elements should not be reset too frequently, as doing so may prevent useful information learned from the past being passed to future frames. However, the degradation phenomenon starts to happen if the dictionary elements are not reset frequently enough, which is also suggested by the larger variances in the lower left corner cells. Nevertheless, the performance is not very sensitive to the rotational reset parameters as many cells in the middle range give good results.

5. CONCLUSIONS

We conducted the first experiment of using long (about 10 minutes) noisy speech files containing multiple speakers to evaluate speech enhancement performance of online semi-supervised PLCA-based approaches, while existing papers all use files shorter than 30 seconds. We found that the enhanced speech signal started to degrade after the algorithm was ran for 2 minutes. We analyzed the problem and found that the reason was due to the inappropriate update of the speech dictionary. We then proposed a simple solution to periodically and rotationally reset speech dictionary elements. Experiments showed that this simple strategy indeed solved the problem. The improved algorithm outperformed the original algorithm and four classical speech enhancement algorithms significantly in non-stationary noisy environments in various SNR conditions. Furthermore, parameter analysis showed that the enhancement performance was not very sensitive to the strategy's parameters.

6. REFERENCES

[1] P. C. Loizou, *SPEECH ENHANCEMENT: THEORY and PRACTICE.* CRC press, 2013.

[2] S. Kamath and P. C. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002, pp. 4164–4164.

[3] P. Scalart, "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1996, pp. 629–632.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, no. 2, pp. 443–445, 1985.

[5] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[6] Z. Duan, G. J. Mysore, and P. Smaragdis, "Speech enhancement by online non-negative spectrogram decomposition in non-stationary noise environments." in *Proc. INTERSPEECH*, 2012.

[7] 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.

[8] P. Smaragdis, B. Raj, and M. Shashanka, "A probabilistic latent variable model for acoustic modeling," in *Proc. Advances in Models for Acoustic Processing (NIPS)*, vol. 148, 2006.

[9] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using nonnegative matrix factorization," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2140–2151, 2013.

[10] Z. Duan, G. J. Mysore, and P. Smaragdis, "Online PLCA for real-time semi-supervised source separation," in *Proc. Latent Variable Analysis and Signal Separation*, 2012, pp. 34–41.

[11] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-time speech separation by semi-supervised nonnegative matrix factorization," in *Proc. Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 322–329.

[12] L. S. Simon and E. Vincent, "A general framework for online audio source separation," in *Proc. Latent Variable Analysis and Signal Separation*. Springer, 2012, pp. 397–404.

[13] N. Guan, L. Lan, D. Tao, Z. Luo, and X. Yang, "Transductive nonnegative matrix factorization for semi-supervised high-performance speech separation," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 2534–2538.

[14] X. Jaureguiberry, E. Vincent, and G. Richard, "Multiple-order non-negative matrix factorization for speech enhancement," in *Proc. INTERSPEECH*, 2014, p. 4.

[15] F. G. Germain and G. J. Mysore, "Stopping criteria for non-negative matrix factorization based supervised and semi-supervised source separation," *IEEE Signal Processing Letters*, vol. 21, no. 10, pp. 1284–1288, 2014.

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

[17] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proc. Advances in Neural Information Processing Systems*, 2001, pp. 556–562.

[18] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario." in *Proc. INTERSPEECH*, 2011, pp. 1509–1512.

[19] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 4266–4269.

[20] R. C. Hendriks, J. Jensen, and R. Heusdens, "Noise tracking using DFT domain subspace decompositions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 541–553, 2008.

[21] L. Di Persia, D. Milone, H. L. Rufiner, and M. Yanagida, "Perceptual evaluation of blind source separation for robust speech recognition," *Signal Processing*, vol. 88, no. 10, pp. 2578–2583, 2008.