INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Real-time Speech Enhancement with GCC-NMF

Sean UN Wood, Jean Rouat

NECOTIS, GEGI, Université de Sherbrooke, Canada
sean.wood@usherbrooke.ca, jean.rouat@usherbrooke.ca

Abstract

We develop an online variant of the GCC-NMF blind speech enhancement algorithm and study its performance on two-channel mixtures of speech and real-world noise from the SiSEC separation challenge. While GCC-NMF performs enhancement independently for each time frame, the NMF dictionary, its activation coefficients, and the target TDOA are derived using the entire mixture signal, thus precluding its use online. Prelearning the NMF dictionary using the CHiME dataset and inferring its activation coefficients online yields similar overall PEASS scores to the mixture-learned method, thus generalizing to new speakers, acoustic environments, and noise conditions. Surprisingly, if we forgo coefficient inference altogether, this approach outperforms both the mixture-learned method and most algorithms from the SiSEC challenge to date. Furthermore, the trade-off between interference suppression and target fidelity may be controlled online by adjusting the target TDOA window width. Finally, integrating online target localization with max-pooled GCC-PHAT yields only somewhat decreased performance compared to offline localization. We test a real-time implementation of the online GCC-NMF blind speech enhancement system on a variety of hardware platforms, with performance made to degrade smoothly with decreasing computational power using smaller pre-learned dictionaries.

Index Terms: real-time, speech enhancement, GCC, NMF, GCC-NMF, GCC-PHAT, CASA

1. Introduction

Real-world applications of speech processing including assistive listening devices and digital personal assistants rely on online speech separation and enhancement algorithms. However, a significant amount of research has focused on the offline setting, where many algorithms are unsuitable for real-time use due to batch processing or computational requirements. We recently presented the offline GCC-NMF speech enhancement algorithm, combining non-negative matrix factorization (NMF) with the generalized cross-correlation (GCC) localization method [1]. While GCC-NMF performs enhancement independently for each time frame, the NMF dictionary, its activation coefficients, and the target time delay of arrival (TDOA) are derived using the entire mixture signal, thus precluding its use online. In this work, we develop an online variant of GCC-NMF, and present a real-time implementation thereof.

We begin with a review of the foundations of GCC-NMF in Section 2, followed by a review of offline GCC-NMF and the development of the online variant in Section 3. We proceed with experimental analyses in Section 4, first showing that online GCC-NMF generalizes to new speakers and noise conditions from very little data. We then show that by forgoing NMF coefficient inference completely, thus performing enhancement using only a pre-learned dictionary and input phase differences, this approach outperforms the offline method. We also present various means to control the trade-off between interference suppression and target fidelity on a frame-by-frame basis, all but one having no effect on computational requirements. We finish with a description of the real-time implementation in Section 5, with performance made to decrease smoothly with decreasing computational power, followed by a conclusion in Section 6.

2. GCC and NMF

2.1. GCC

GCC is a robust approach to sound source localization in the presence of noise, interference, and reverberation [2, 3]. The GCC function extends the frequency domain cross-correlation definition with an arbitrary frequency-weighting function ψ_{ft}, providing control over the relative importance of the signal's constituent frequencies:

$$\mathrm{G}_{\tau t} = \sum_f \psi_{ft} \mathrm{V}_{lft} \mathrm{V}^*_{rft} e^{j2\pi f \tau} \tag{1}$$

where V_{lft} and V_{rft} are the left and right complex spectrograms computed with the short-time Fourier transform (STFT), * is complex conjugation, and f, t, and τ index frequency, time, and TDOA respectively. Many of the most robust localization methods are based on the GCC phase transform (GCC-PHAT) [4], in which frequencies are weighted equally, defining $\psi^{\mathrm{PHAT}}_{ft}$ as the inverse product of the magnitude spectrograms:

$$\mathrm{G}^{\mathrm{PHAT}}_{\tau t} = \sum_f \frac{\mathrm{V}_{lft} \mathrm{V}^*_{rft}}{|\mathrm{V}_{lft}|\,|\mathrm{V}_{rft}|} e^{j2\pi f \tau} \tag{2}$$

The resulting GCC-PHAT *angular spectrogram* can then be pooled over time, with the TDOA of the highest peaks corresponding to the source locations; see Figure 1a) for an example.

In Section 3.1, we will show that individual NMF dictionary atoms can be used as GCC frequency-weighting functions, such that their TDOAs may be estimated at each point in time.

2.2. NMF

NMF is known to learn parts-based representations of non-negative input data in a purely unsupervised fashion [5]. In the context of speech separation and enhancement, input typically consists of a magnitude spectrogram $|\mathrm{V}_{ft}|$, with f and t indexing frequency and time as above. NMF decomposes the spectrogram into two non-negative matrices: a dictionary W_{fd} whose columns comprise atomic spectra indexed by d and set of corresponding activation coefficients H_{dt} such that $|\mathrm{V}| \approx \mathrm{WH}$; see Figure 1b) for example dictionary atoms. Each column of the input spectrogram $|\mathrm{V}|$, i.e. each frame t, is thus approximated as a linear combination of the dictionary atoms with the coefficients from the corresponding column of H. For the stereo spectrograms we study here, we may set $\mathrm{V}_{ft} = [\mathrm{V}_{lft}|\mathrm{V}_{rft}]$, with the corresponding stereo coefficients $\mathrm{H}_{dt} = [\mathrm{H}_{ldt}|\mathrm{H}_{rdt}]$, where the matrices are concatenated in time.

http://dx.doi.org/10.21437/Interspeech.2017-1458

In traditional NMF, dictionary learning and coefficient inference are performed together by initializing the dictionary and coefficient matrices randomly, and updating them iteratively according to the following rules,

$$\mathrm{H} \leftarrow \mathrm{H} \odot \frac{\mathrm{W}^\top \left(|\mathrm{V}| \odot \Lambda^{\beta-2} \right)}{\mathrm{W}^\top \Lambda^{\beta-1}} \tag{3}$$

$$\mathrm{W} \leftarrow \mathrm{W} \odot \frac{\left(\Lambda^{\beta-2} \odot |\mathrm{V}| \right) \mathrm{H}^\top}{\Lambda^{\beta-1} \mathrm{H}^\top} \tag{4}$$

where $\Lambda = \mathrm{WH}$ is the reconstructed input, β parameterizes the reconstruction cost function $d_\beta\left(|\mathrm{V}|, \Lambda\right)$[1], and the matrix exponentials, divisions, and $\odot$ product are computed element-wise. Dictionary atoms are typically normalized after each update, and their coefficients scaled accordingly. Since all input examples are required prior to optimization, this is an offline approach. As described in Section 3.2, we will instead pre-learn a dictionary offline, and infer the coefficients for each input frame online by initializing the coefficient vector randomly and iteratively performing (3) while keeping the dictionary fixed.

3. Online GCC-NMF

3.1. Offline GCC-NMF

As NMF dictionary atoms are non-negative functions of frequency, they may be used to construct a set of atom-specific GCC frequency weighting functions,

$$\psi^{\mathrm{NMF}}_{dft} = \frac{1}{|\mathrm{V}_{lft}|\,|\mathrm{V}_{rft}|} \frac{\mathrm{W}_{fd}}{\sum_f \mathrm{W}_{fd}} \tag{5}$$

such that for a given atom d, frequencies are weighted according to their relative magnitude in the atom. The resulting GCC-NMF atom-specific angular spectrograms are then defined as follows, with examples shown in Figure 1c),

$$\mathrm{G}^{\mathrm{NMF}}_{d\tau t} = \sum_f \psi^{\mathrm{NMF}}_{dft} \mathrm{V}_{lft} \mathrm{V}^*_{rft} e^{j2\pi f \tau} \tag{6}$$

We estimate the TDOA of each atom d at each time t as the τ for which GCC-NMF reaches its maximum value: $\mathrm{argmax}_\tau \mathrm{G}^{\mathrm{NMF}}_{d\tau t}$. Atoms are then associated with the target if their estimated TDOA lies within a window of size ϵ around the target TDOA τ^*_t, otherwise they are associated with the interference. This defines a binary coefficient mask,

$$\mathrm{M}_{dt} = \begin{cases} 1 & \text{if } |\tau^*_t - \mathrm{argmax}_\tau \mathrm{G}^{\mathrm{NMF}}_{d\tau t}| < \epsilon/2 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

Multiplying M_{dt} with the coefficients H_{dt} and reconstructing as usual then yields the estimate target magnitude spectrogram,

$$|\hat{\mathrm{X}}_{ft}| = \sum_d \mathrm{W}_{fd} \mathrm{H}_{dt} \mathrm{M}_{dt} \tag{8}$$

As is typical in NMF-based separation, the target estimate signal is then reconstructed by applying a time-varying Wiener-like filter to the input signal. The filter is constructed in the frequency domain as the ratio between the target and mixture estimate spectrograms, and is multiplied with the complex input spectrogram V_{cft}, yielding the complex target spectrogram,

[1]The beta divergence $d_\beta\left(|\mathrm{V}|, \Lambda\right)$ is equivalent to the Euclidian distance for $\beta = 2$, the generalized KL divergence for $\beta = 1$, and the IS divergence for $\beta = 0$ [6].

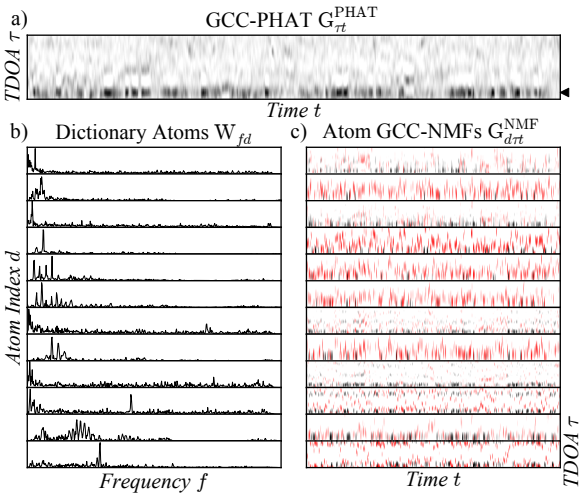

Figure 1: *Elements of the GCC-NMF speech enhancement algorithm for a 10 second mixture of speech and noise. a) The GCC-PHAT angular spectrogram, with resulting target TDOA estimate indicated with a triangle marker. b) Subset of the NMF dictionary atoms* W_{fd}*, with corresponding GCC-NMF angular spectrograms* $\mathrm{G}^{\mathrm{NMF}}_{d\tau}$ *shown in c). When an atom is associated with the target (see Section 3.1), its angular spectra is colored in black, otherwise it is colored in red. Angular spectrograms are rectified here for clarity with* $\max(0, x)$.

$$\hat{\mathrm{X}}_{cft} = \frac{|\hat{\mathrm{X}}_{ft}|}{\Lambda_{ft}} \mathrm{V}_{cft} \tag{9}$$

where c is the channel index. The complex target spectrogram is then transformed to the time domain with the inverse STFT.

3.2. Online GCC-NMF

Since the coefficient mask M_{dt} is generated independently for each frame, GCC-NMF has potential be performed online. However, dictionary learning, coefficient inference, and target localization are performed using the entire mixture signal, thus precluding online use. We proceed to address each of these elements now, as we develop the online variant of GCC-NMF.

3.2.1. Dictionary Pre-learning

A typical approach for supervised speech enhancement with NMF is to *pre-learn* a pair of dictionaries on isolated speech and noise signals, and subsequently infer their coefficients for the mixture signal while keeping the dictionaries fixed [7, 8, 9]. We take inspiration from this approach and pre-learn a single NMF dictionary from a dataset containing both isolated speech and noise signals. Contrary to the supervised approach, this approach remains purely unsupervised as a single dictionary is learned for both speech and noise. Individual atoms are then associated with either the target or interference at each point in time according to (7). In Section 4.1, we will see that this dictionary pre-learning approach generalizes to different speakers, acoustic environments, noise conditions, and recording setups.

3.2.2. Coefficient Inference

The activation coefficients of the pre-learned dictionary can be inferred for the input mixture on a frame-by-frame basis by ini-

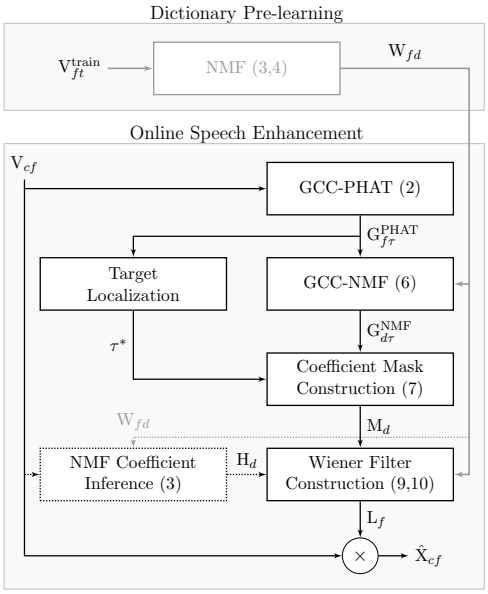

Figure 2: *Block diagram of online GCC-NMF consisting of offline dictionary pre-learning and online speech enhancement. Online, offline, and optional components are drawn with black, gray, and dotted lines respectively, with the blocks' equations listed in parentheses.*

tializing the coefficient vector randomly, and updating it iteratively according to (3). However, we will see in Section 4.2 that better overall performance can in fact be achieved by forgoing coefficient inference completely. In this case, replacing the coefficients with the all-ones vector, the Wiener-like filtering process defined in (9) reduces to,

$$\hat{X}_{cft} = \frac{\sum_d W_{fd} M_{dt}}{\sum_d W_{fd}} V_{cft} \tag{10}$$

3.2.3. Online Localization

With offline GCC-NMF, target localization was performed using a max-pooled GCC-PHAT technique [4] where the target TDOA is that at which the global maximum occurred in the GCC-PHAT angular spectrogram (2), i.e. $\text{argmax}_\tau G_{\tau t}^{PHAT}$. We adapt this approach to the online setting by considering the only the current and previous angular spectrogram frames. While this approach works well for the static speaker case we consider here, a more complex localization and tracking approach will be incorporated in future work to handle moving speakers.

4. Experiments

We proceed to evaluate online GCC-NMF on the SiSEC 2016 speech in noise *dev* dataset, consisting of two-channel mixtures of speech and background noise [10]. Dictionary pre-learning is performed on a subset of the CHiME 2016 development set [11], taking an equal number of randomly selected frames from the isolated speech and background noise signals. The sample rate for both SiSEC and CHiME is 16 kHz, and we use an STFT with 1024-sample windows (64 ms), 16-sample hop size / frame advance (4 ms), and a Hann window function. Default GCC-NMF parameters are dictionary size = 1024, number of updates = 100, $\beta = 1$, number of TDOA samples = 128, and target TDOA window size = 5% (6 samples). Enhancement performance is measured with the PEASS open source toolkit quantifying overall quality, target fidelity, interference suppression, and lack of artifacts, where higher scores are better. PEASS is a perceptually-motivated method that better correlates with human assessments than the traditional SNR-based measures [12].

We first study the effects on enhancement performance of the pre-learned dictionary size and the amount of data used for pre-learning, followed by the number of training and inference iterations, and the target TDOA window width ϵ. These evaluations are performed with offline target TDOA estimation. We then compare performance using online and offline localization, and compare results with other speech enhancement algorithms from the SiSEC challenge, in addition to an oracle baseline.

4.1. Dictionary pre-learning

PEASS scores for varying train set and dictionary sizes are shown in Figure 3. For a given dictionary size, we note that performance converges quickly with increasing train set size, such that performance is near maximal for most measures with only 2^{10} (1024) frames, with interference suppression reaching its maximum at larger training sets in some cases. Contrary to many supervised approaches, therefore, unsupervised dictionary pre-learning only requires a small amount of training data. We also note that overall, target, and artifact performance increase smoothly with increasing dictionary size, as was the case with offline GCC-NMF, albeit with diminishing returns, with interference suppression showing a slight decrease for larger dictionaries. Finally, since the overall scores are similar to those presented previously for offline GCC-NMF [1], this dictionary pre-learning technique generalizes to new speakers, noise and acoustic conditions, and recording setups.

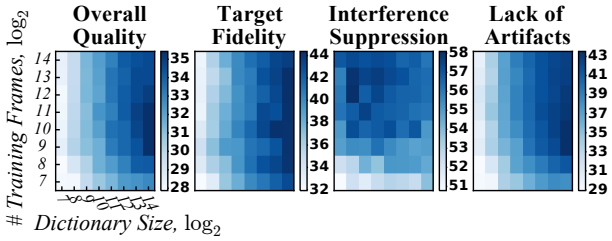

Figure 3: *PEASS scores for varying number of dictionary training frames (vertical axes) and dictionary sizes (horizontal axes), with both varying from 2^7 (128) to 2^{14} (16 384) exponentially. Colorbars indicate the range for each of score type.*

4.2. Number of training and inference updates

The effect of the number of dictionary pre-learning updates on enhancement performance is presented in Figure 4a). As was the case for offline GCC-NMF, increasing the number of training iterations results in increased interference suppression. Overall, target, and artifact scores, however, increase until approximately 100 iterations, decreasing thereafter. The choice of the number of training iterations therefore offers offline control of the trade-off between target fidelity and interference suppression. One could learn a set of dictionaries spanning a range of training iterations, and subsequently control the trade-off online by selecting the desired dictionary on a frame-by-frame basis.

The number of online inference iterations is presented in Figure 4b), showing similar effects to the number of training iterations for large values. For small number of iterations, how-

ever, we note an opposite effect for overall, target, and artifact scores, as they continue to increase with decreasing number of iterations. Surprisingly, then, the best overall performance is in fact achieved when no inference is performed, i.e. 0 coefficient updates. As mentioned in Section 3.2.2, we can thus forego the coefficient inference stage completely, and perform the Weiner-like filtering using only the pre-learned dictionary and input phase differences as in (10).

Finally, we note that both the number of training and inference iterations offer control over the target fidelity vs. interference suppression trade-off. While the dictionary pre-learning is performed offline, and thus has no computational effect online, increasing the number of inference iterations comes with a computational cost at runtime.

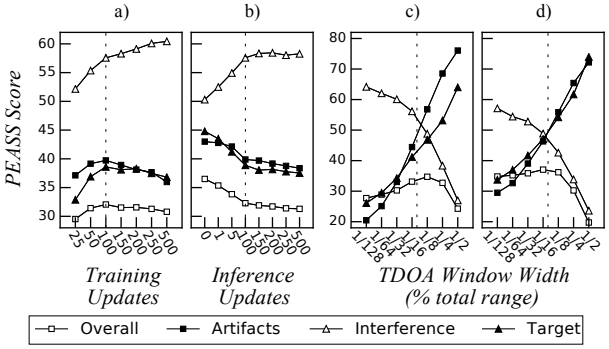

Figure 4: *Effect on average PEASS scores of a) the number of NMF pre-learning updates; b) the number of NMF coefficient inference updates at test time; the target TDOA window width for c) 100 coefficient inference updates, and d) 0 updates.*

4.3. Target TDOA window size

We present the effect of the target TDOA window size, i.e. ϵ in (7), for both 100 inference iterations and 0 iterations in Figure 4c) and d). We first note that the 0 iterations case generally yields higher overall scores with higher target fidelity and decreased interference suppression. Second, we note in both cases a drastic effect on the target vs. interference trade-off, as widening the TDOA window results in reduced interference suppression and higher target fidelity. Since the target TDOA window width can be controlled online, this provides the most significant control of the target fidelity vs. interference suppression trade-off with respect to the parameters presented thus far, with no effect on computational requirements. The highest overall score is achieved for 1/8 (100 iterations) and 1/16 (0 iterations) of the total TDOA range.

4.4. Comparison between approaches

In Table 1, we compare online GCC-NMF with dictionary pre-learning and no coefficient inference for both offline and online max-pooled GCC-NMF localization methods. Offline GCC-NMF and other algorithms from the 2013, 2015, and 2016 SiSEC separation challenges are included for comparison [13, 14, 10]. We first note that the proposed online GCC-NMF approaches yield better overall and artifact scores than offline GCC-NMF, with reduced interference suppression and somewhat reduced target fidelity. The online localization method results in somewhat decreased performance when compared to offline localization, suggesting that more complex localization methods should be investigated. Finally, online GCC-NMF outperforms all but one of the previous methods, most of which rely on supervised learning or are unsuitable in online settings. Online GCC-NMF therefore holds significant potential for future research, especially given that it remains purely unsupervised, conceptually simple, easy to implement, and generalizes across speakers, noise conditions, and recording setups.

Table 1: *Mean PEASS scores for different speech enhancement algorithms taken over the SiSEC speech and noise mixtures dev dataset. The GCC-NMF methods include the previous offline mixture-learned approach[1], the dictionary pre-learning approach both with online localization[2] and offline localization[3]. Other approaches from the SiSEC challenges are presented for comparison, where * are computed using the subset of examples as reported in [10], and the ideal binary mask † (IBM) is an oracle baseline.*

	Overall	Target	Interference	Artifacts
Offline[1]	34.01±6.00	44.60±14.77	57.06±6.83	39.56±5.65
Pre-trained[2]	35.34±5.72	41.19±12.88	50.59±9.32	42.80±6.36
Pre-trained[3]	36.50±6.10	43.00±12.05	50.30±9.49	44.80±5.28
*Liu** [10]	14.93±4.76	43.53±4.45	17.13±1.33	69.13±6.98
*Duong** [15]	16.57±5.47	70.03±2.93	11.53±5.35	73.30±4.75
Rafii [16]	30.81±5.69	56.60±7.74	31.53±9.89	55.56±4.78
Magoarou [17, 18]	32.66±7.51	66.10±22.10	34.84±12.71	44.13±11.87
Wang [19, 20]	38.01±5.79	53.91±9.25	54.50±6.12	50.60±7.02
IBM† [14]	38.37±9.33	56.61±8.96	73.69±1.10	38.43±10.54

5. Real-time Implementation

A real-time GCC-NMF software implementation was written in Python, using the Theano optimizing compiler, with an interactive graphical interface using PyQt and pyqtgraph [21]. Parameters may be manipulated in real-time, such that their effects on subjective enhancement quality can be studied interactively. The software has been tested on a range of hardware platforms including a desktop PC with an NVIDIA K40 GPU, an NVIDIA TX1 embedded system on a chip (SoC), the low-cost Raspberry Pi 3, and a 2010 MacBook Pro. Performance can be made to degrade smoothly with decreasing computational power by using smaller pre-trained dictionaries, as shown in Figure 3. The source code for real-time GCC-NMF will be made available at `https://www.github.com/seanwood/gcc-nmf`.

6. Conclusion

We presented an online variant of the GCC-NMF speech enhancement algorithm, and studied its performance on stereo mixtures of speech and real-world noise. We showed that pre-learning the NMF dictionary on a different dataset and inferring its activation coefficients frame-by-frame generalizes to new speakers, noise conditions, and recording setups from very little data. By foregoing the coefficient inference step completely, thus using only the pre-learned dictionary and input phase differences, this approach yields better overall performance than the offline method, and outperforms all but one of the previous algorithms submitted to the SiSEC speech enhancement challenge. The trade-off between interference suppression and target fidelity may be controlled online via several different parameters, with the target TDOA window width offering the most control, and having no effect on computational requirements. Finally, a real-time, open source Python implementation was developed, allowing a subjective analysis of the effects of various parameters to be studied interactively in real-time.

Acknowledgements: NSERC discovery grant, FQRNT (CHIST-ERA, IGLU)

7. References

[1] S. U. N. Wood, J. Rouat, S. Dupont, and G. Pironkov, "Blind speech separation and enhancement with GCC-NMF," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 745–755, 2017.

[2] C. H. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 4, pp. 320–327, 1976.

[3] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, Universitat Politècnica de Catalunya, 2006.

[4] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source TDOA estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, vol. 92, no. 8, pp. 1950–1960, 2012.

[5] 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.

[6] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the β-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[7] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 163–176, 2006.

[8] F. Weninger, J. Le Roux, J. R. Hershey, and S. Watanabe, "Discriminative NMF and its application to single-channel source separation." in *INTERSPEECH*, 2014, pp. 865–869.

[9] E. Vincent, N. Bertin, R. Gribonval, and F. Bimbot, "From blind to guided audio source separation: How models and side information can improve the separation of sound," *IEEE Signal Processing Magazine*, vol. 31, no. 3, pp. 107–115, 2014.

[10] A. Liutkus, F.-R. Stöter, Z. Rafii, D. Kitamura, B. Rivet, N. Ito, N. Ono, and J. Fontecave, "The 2016 signal separation evaluation campaign," in *International Conference on Latent Variable Analysis and Signal Separation*. Springer, 2017, pp. 323–332.

[11] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2016.

[12] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2046–2057, 2011.

[13] N. Ono, Z. Koldovsky, S. Miyabe, and N. Ito, "The 2013 signal separation evaluation campaign," *Proc. International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2013.

[14] N. Ono, Z. Rafii, D. Kitamura, N. Ito, and A. Liutkus, "The 2015 signal separation evaluation campaign," in *Latent Variable Analysis and Signal Separation*. Springer, 2015, pp. 387–395.

[15] H.-T. T. Duong, Q.-C. Nguyen, C.-P. Nguyen, T.-H. Tran, and N. Q. Duong, "Speech enhancement based on nonnegative matrix factorization with mixed group sparsity constraint," in *Proceedings of the Sixth International Symposium on Information and Communication Technology*. ACM, 2015, pp. 247–251.

[16] Z. Rafii and B. Pardo, "Online REPET-SIM for real-time speech enhancement," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 848–852.

[17] S. Arberet, A. Ozerov, N. Q. Duong, E. Vincent, R. Gribonval, F. Bimbot, and P. Vandergheynst, "Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation," in *Information Sciences Signal Processing and their Applications (ISSPA), 2010 10th International Conference on*. IEEE, 2010, pp. 1–4.

[18] L. Le Magoarou, A. Ozerov, and N. Q. Duong, "Text-informed audio source separation using nonnegative matrix partial cofactorization," in *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE, 2013, pp. 1–6.

[19] L. Wang, H. Ding, and F. Yin, "A region-growing permutation alignment approach in frequency-domain blind source separation of speech mixtures," *IEEE transactions on audio, speech, and language processing*, vol. 19, no. 3, pp. 549–557, 2011.

[20] (2017, June). [Online]. Available: http://www.onn.nii.ac.jp/sisec13/evaluation_result/BGN/Kayser.txt

[21] S. U. N. Wood and J. Rouat, "Real-time speech enhancement with GCC-NMF: Demonstration on the Raspberry Pi and NVIDIA Jetson," in *Interspeech 2017*, 2017.