COLLABORATIVE SPEECH DEREVERBERATION: REGULARIZED TENSOR FACTORIZATION FOR CROWDSOURCED MULTI-CHANNEL RECORDINGS

Sanna Wager, Minje Kim

Indiana University
School of Informatics, Computing, and Engineering
Bloomington, IN 47408

scwager@indiana.edu, minje@indiana.edu

ABSTRACT

We propose a regularized Nonnegative Tensor Factorization (NTF) model for multi-channel speech dereverberation that incorporates prior knowledge about clean speech. The approach models the problem as recovering a signal convolved with different room impulse responses, so that the dereverberation problem can benefit from microphone arrays. Furthermore, as the factorization learns individual reverberation filters as well as the channel specific delays, we can employ an ad-hoc array with heterogeneous sensors (e.g. multi-channel recording by a crowd) even if they are not synchronized. For a more stable dereverberation performance, we introduce two regularization schemes. First, instead of a direct estimation of the commonly shared speech reconstruction across the NTF problem, we further factorize it using Nonnegative Matrix Factorization (NMF) with fixed and pre-trained clean speech basis vectors, so that the optimization process can focus on estimating only their activations. Furthermore, we constrain the NMF activation matrix to take on characteristics of dry signals using sparsity and total variation constraints. Empirical dereverberation results on different reverberation setups show that the proposed regularization improves both the recovered sound quality and the speech intelligibility.

***Index Terms*—** Dereverberation, nonnegative matrix factorization, nonnegative tensor factorization, collaborative audio enhancement

1. INTRODUCTION

Collaborative audio enhancement methods permit the recovery of a high-quality signal from multiple low-quality recordings, such that a clean, full-band signal is created from bandlimited, clipped, noisy, and reverberant signals, which are common artifacts in crowdsourced recordings [1, 2, 3]. This signal recovery is made more challenging by the presence of reverberation, because longer Room-Impulse Responses (RIRs) span multiple Short-Time Fourier Transform (STFT) windows. Reverberation is also of concern in applications such as automatic speech recognition (ASR). Recent studies on single-channel dereverberation [4, 5, 6, 7] have modeled a reverberant signal as a frequency-domain convolution between a clean signal and RIRs, both of which are unknown. Recovering the dry signal involves estimating a large number of parameters. This parameter estimation can be done using Nonnegative Matrix Factorization (NMF), but is hindered because such models are underdetermined [8], also by the NMF assumption of independence between frames.

We develop a Nonnegative Tensor Factorization (NTF)-based multi-channel dereverberation formulation based on [9], modeling the problem as recovering a signal convolved with different room-impulse responses. Ermiş *et al.* [10] describe the more general approach of finding the unknown common element among multiple signals as link prediction. The authors demonstrate that a tensor factorization model effectively harnesses information in large-scale data, and that its loss function optimizations often outperform commonly used loss functions for matrix factorizations.

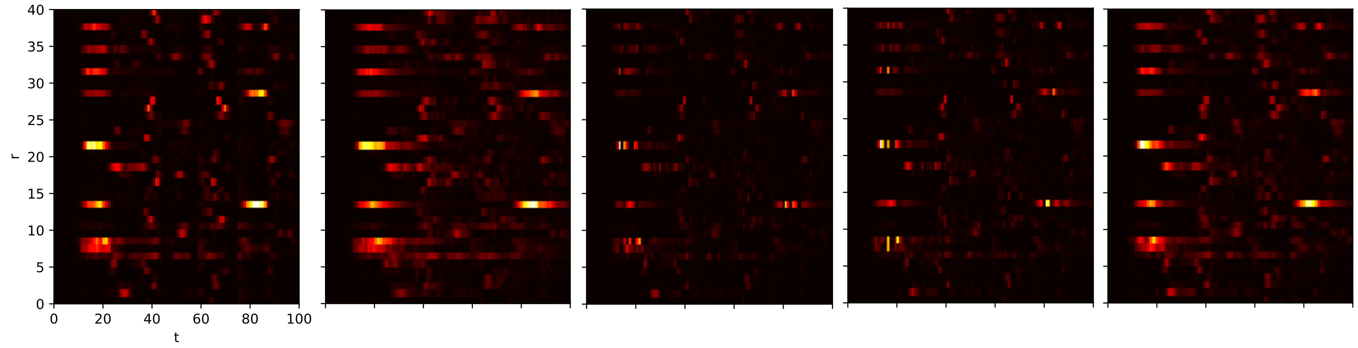
Examples of link prediction are common in the literature. Kim *et al.* [11] and Le Magoarou *et al.* [12] develop nonnegative matrix partial co-factorization. Kim *et al.* [11] separate a mixed signal into activations of priorly trained drum and non-percussive sounds, while Le Magoarou *et al.* [12] formulate a single-channel source separation problem into a joint factorization of the mixture and an example clean speech spectrogram. The model allows signals to share the speech-related commonalities while allowing divergence in other dimensions such as timbre, timing, and prosody. We follow these authors in adopting the use of clean speech basis vectors for dereverberation. Seichepine *et al.* [13] develop soft nonnegative co-factorization, which applies factorization to different signals while enforcing similarity between one of the factors across the two decompositions, thus taking advantage of the latent commonalities between the signals.

Utilizing the prior knowledge about the source, especially dry speech, is also a popular idea in speech dereverberation. For example, Kumar *et al.* [6] and Liang *et al.* [4] both model single-channel dereverberation with a sparsity constraint. Liang *et al.* [4] model reverberation as a convolutive process applied to clean speech, pre-training it on clean speech using Bayesian methods. They use statistical inference to recover the process and estimate the clean speech in the single-channel case.

In this paper we show that incorporation of prior knowledge about dry speech improves the standard NTF model, where the dry speech source is shared as a target to recover. Although each channel-specific dereveberation can benefit from this shared component, the direct estimation of the large speech spectrogram is still challenging. To this end, we first propose to further factorize the shared speech spectrogram by using a fixed basis vector array prelearned from dry speech. In this way, the estimation procedure can focus on the activation matrix with a smaller number of parameters. We also regularize the NTF problem by enforcing the activation to be sparse, but at the same time does not vary too frequently.

2. MULTI-CHANNEL REVERBERATION MODEL

Our approach is based on the NTF model for the multi-channel scenario developed in [9]. This model approximates the complex-



(a) (b) (c) (d) (e)

Fig. 1. Appearance of activation matrix $\mathbf{A}$ learned using (a) NMF directly on clean speech (b) NMF directly on reverberant speech (c) NTF baseline with no constraints (d) NTF with sparsity constraint only (e) NTF with both sparsity and total variation constraints (proposed).

domain convolution theorem in the magnitude domain, and applies it to the case where the RIR lasts multiple DFT frames. For the subband-wise convolution, we introduce a notation $\circledast$ as follows:

$$\mathbf{X}^{(i)} = \mathbf{H}^{(i)} \circledast \mathbf{S} \iff \mathbf{X}^{(i)}_{f,t} = \sum_{p=0}^{L-1} \mathbf{H}^{(i)}_{f,p} \mathbf{S}_{f,t-p}, \tag{1}$$

where $\mathbf{S}, \mathbf{X}^{(i)} \in \mathbb{R}_+^{F \times T}$ are the magnitude STFTs of the clean signal and i-th channel of the input signal, respectively. $\mathbf{H}^{(i)} \in \mathbb{R}_+^{F \times L}$ is magnitude STFT of the i-th channel RIR. Therefore, the multi-channel data in $\mathbf{X}$ and $\mathbf{H}$ is represented as three-dimensional tensors whose third axis— the channels—is indexed by the superscript (i).

Reverberant signals are typically misaligned because the source-to-sensor distances vary between sensors [14]. Aligning the signals directly using approaches such as cross-correlation is difficult because the RIRs make the channel waveforms very different from each other. The NTF model automatically aligns the signals in time by adjusting $\mathbf{H}$ for each channel. These models do not require training data, so offer the advantage of small overhead.

2.1. The baseline dereverberation model using NTF

We are based on the NTF-based multi-channel speech dereverberation mode [9]. This approach seeks to minimize the objective function (2) to learn estimates for the filter tensor $\mathbf{H}$ and clean signal $\mathbf{S}$:

$$\mathcal{J}_0 = \sum_i \left\| \mathbf{X}^{(i)} - \mathbf{Z}^{(i)} \right\|_F^2, \quad \mathbf{Z}^{(i)} = \mathbf{H}^{(i)} \circledast \mathbf{S} \tag{2}$$

The objective function represents Euclidean distance between the reverberant input signals $\mathbf{X}$ and their approximations $\mathbf{Z}$ using the estimates of the clean signal and filters. Like regular NMF [15], it uses iterative update rules with a nonnegativity constraint.

NMF- and NTF-based models permit a large number of equivalent factorizations, which leads to two concerns. First, the model given by (2) risks at worst that $\mathbf{H}$ represents a filter where all but the first column is zero-valued, with $\mathbf{S}$ an averaged reverberant signal. Second, direct estimation of $\mathbf{S}$ means that the number of parameters to be learned can become very large.

3. PROPOSED MODEL

We address both of the concerns about the NTF-based model by incorporating prior knowledge about clean speech. In our proposed model, instead of directly estimating the clean signal $\hat{\mathbf{S}}$, we assume a further NMF approximation, setting $\mathbf{S} \approx \mathbf{WA}$, where $\mathbf{W} \in \mathbb{R}_+^{F \times R}$ and $\mathbf{A} \in \mathbb{R}_+^{R \times T}$. A lower rank approximation, i.e. $R < F, T$, is common. $\mathbf{W}$ is a trained dictionary of clean speech components. Given that $\mathbf{W}$ is fixed, we then only need to estimate the activation matrix $\mathbf{A}$, which reduces the number of parameters that need to be learned from $F \times T$ to $R \times T$. Furthermore, we enforce the structure of $\mathbf{A}$ to correspond to that of a clean signal, using constraints described below: sparsity and total variation. This ensures that the model represents the reverberation in the estimated reverberation filters $\mathbf{H}$, not in $\mathbf{S}$. Finally, given that King et al. have demonstrated empirically that the generalized KL-divergence produces better results in NMF applications to audio source separation than the Frobenius norm [16], we use this metric in our objective function.

3.1. Objective function

The KL-divergence with additional constraints on A forms our new objective function $\mathcal{J}_p$:

$$\mathcal{J}_p = \sum_i \left\| \mathbf{X}^{(\mathbf{i})} \log \frac{\mathbf{X}^{(\mathbf{i})}}{\mathbf{Z}^{(\mathbf{i})}} - \mathbf{X}^{(\mathbf{i})} + \mathbf{Z}^{(\mathbf{i})} \right\|_1 + \gamma \Psi(\mathbf{A}) + \zeta \Phi(\mathbf{A}), \tag{3}$$

where we define

$$\mathbf{Z}^{(i)} = \mathbf{H}^{(i)} \circledast [\mathbf{WA}], \tag{4}$$

$$\Psi(\mathbf{A}) = \sum_{i,j} \log(\mathbf{A}_{i,j} + \epsilon), \tag{5}$$

$$\Phi(\mathbf{A}) = \sum_{i,j} |\mathbf{A}_{i,j+1} - \mathbf{A}_{i,j}|^2 = \|\mathbf{AL}\|_F^2, \ \mathbf{L}_{i,j} = \begin{cases} +1 & \text{if } i = j+1 \\ -1 & \text{if } i = j \\ \ \ 0 & \text{otherwise} \end{cases} \tag{6}$$

respectively. Note that (4) is similar to the reconstruction in (2), except that $\mathbf{S}$ is replaced by its NMF approximation $\mathbf{S} \approx \mathbf{WA}$. The values of γ, ϵ and ζ control the contribution of the constraints to the objective function. Given that every basis component in $\mathbf{W}$ is activated separately, the constraints used in the objective function apply to the rows of $\mathbf{A}$.

- ***Sparsity***: The sparsity cost Ψ in (5) described in [17, 18, 19] is the sum of the logarithms of the row-wise products of $\mathbf{A}$. This constraint is equivalent to L_1 norm regularization. A small value ϵ is added to the argument to avoid zero-values. Minimizing this cost encourages most of the energy to be distributed in fewer bins.

- ***Total Variation***: The sparsity constraint alone can result in "gaps" in the sound, given that it does not take the smoothness between adjacent bins into consideration. The total variation constraint Φ in (6), a computer vision technique used for denoising [20, 21], is also adapted here to address this issue by minimizing variation between row-wise adjacent bins. It consists of minimizing the norm of the product of $\mathbf{A}$ first-order derivative operator $\mathbf{L}$.

Fig. 1 compares the estimate of $\mathbf{A}$ learned (a) directly from a clean signal and (b) from a reverberant signal. The reverberant case is much less sparse than the clean case, as the reverberant activations decrease slowly instead of quickly vanishing. In (c) we see that the recovered $\mathbf{A}$ suffers from discontinuity in activation patterns, which can be partly addressed by introducing sparsity as in (d). But, eventually with the additional total variation constraint we can recover $\mathbf{A}$ that is most similar to (a).

3.2. Multiplicative update rules

As in many other NMF-related algorithms, we first calculate the gradients for the parameters $\mathbf{H}$ and $\mathbf{A}$, then choose the step sizes so that the gradient descent updates turn into multiplicative update rules:

$$\tilde{\mathbf{H}}^{(i)} \leftarrow \tilde{\mathbf{H}}^{(i)} \odot \frac{\frac{\overleftrightarrow{\mathbf{X}^{(i)}}}{\overleftrightarrow{\mathbf{Z}^{(i)}}} \circledast [\mathbf{W}\mathbf{A}]}{\mathbf{1} \circledast [\mathbf{W}\mathbf{A}]}, \;\; \mathbf{H}^{(i)} \leftarrow \tilde{\mathbf{H}}^{(i)}_{:,T-1:T-1+p}, \;\; \forall i, \quad (7)$$

$$\mathbf{A} \leftarrow \tilde{\mathbf{A}} \odot \frac{\mathbf{W}^\top \sum_i \overleftrightarrow{\mathbf{H}^{(i)}} \circledast \frac{\mathbf{X}^{(i)}}{\mathbf{Z}^{(i)}} + \zeta \min\left(\tilde{\mathbf{A}}(\mathbf{L}\mathbf{L}^\top), \mathbf{0}\right)}{\mathbf{W}^\top \sum_i \overleftrightarrow{\mathbf{H}^{(i)}} \circledast \mathbf{1} + \frac{\gamma}{\tilde{\mathbf{A}}+\epsilon} + \zeta \max\left(\tilde{\mathbf{A}}(\mathbf{L}\mathbf{L}^\top), \mathbf{0}\right)},$$

$$\mathbf{A} \leftarrow \tilde{\mathbf{A}}_{:,L-1:L-1+T}, \quad (8)$$

where $\odot$ denotes the Hadamard product. Another important new notation $\overleftrightarrow{\mathbf{X}^{(i)}}$ indicates the left-right flipping operation, i.e. $\overleftrightarrow{\mathbf{X}^{(i)}} = [\mathbf{X}^{(i)}_{:,T-1}, \mathbf{X}^{(i)}_{:,T-2}, \cdots, \mathbf{X}^{(i)}_{:,0}]$, which is a procedure to make sure the use of $\circledast$ in the update rules is for deconvolution as opposed to that in (1). We can also see that the gradient of total variation, $\mathbf{A}(\mathbf{L}\mathbf{L}^\top)$ is separated into its negative components in the numerator and its positive components in the denominator by using the element-wise min and max functions. Note that the division is also element-wise. $\mathbf{1}$ and $\mathbf{0}$ are the matrix of 1's and 0's whose sizes are $F \times T$ and $R \times T$. After every update, we have an estimation of the filter $\tilde{\mathbf{H}}^{(i)}$ and the NMF activation $\tilde{\mathbf{A}}$ that are with zero padding in the beginning due to the delays, i.e. $\tilde{\mathbf{H}}^{(i)} = \left[\mathbf{0}, \mathbf{H}^{(i)}\right]$ and $\mathbf{0}$ here has dimensions $F \times (T-1)$ and $\tilde{\mathbf{A}} = [\mathbf{0}, \mathbf{A}]$ and $\mathbf{0}$ here has dimensions $R \times (L-1)$. Hence we discard them after every update by a shifting operation.

An additional constraint addresses the scaling indeterminacy of the model, ultimately causing the signals that differ more from the rest to be weighted less, following [9]. We normalize the filter estimate $\mathbf{H}$ at every iteration: $\sum_f \mathbf{H}^{(i)}_{f,p} = 1$.

4. EXPERIMENT

4.1. Room, source and sensor configurations

We generated RIRs using the `roomsimove` toolbox [22]. This program implements the image method [23], which simulates the RIR of a rectangular room. We simulated three rooms, with T60 reverberation times approximately 0.6, 1.2, and 1.6 seconds, values which are challenging for applications such as automatic speech recognition. In each room, we fixed the sensor at 75, 75, and 5 per cent of the room dimensions. In order to simulate the scenario where the sensors are not equidistant from the source, we generated 40 random four-channel sensor configurations for each room, making sure that each sensor was at least one meter away from the other sensors and from the source.

4.2. Input data and parameter settings

The RIRs were convolved with a clean signal from the TIMIT dataset [24] to generate the reverberant input signals. The prior clean speech components $\mathbf{W}$ were learned in advance from 200 utterances from the TIMIT dataset, with the same speaker gender and accent region as the input, but not including the input. The number of basis vectors was set to $R = 40$. For each of the 120 room-sensor configurations, we applied the three different dereverberation algorithms: (1) the NTF baseline using KL-divergence as our objective function, which is a slight modification of [9] (2) the NTF model using the speech prior for a further factorization (3) the NTF model using the speech prior as well as the sparsity and total variation regularization. The regularization parameters for the third case, γ, ϵ and θ, were fixed to a single value for all three T60 settings to simulate the scenario where the T60 time is unknown. Values that were empirically shown to produce the best results according to SNR and STOI were $\gamma = 10^{-6}$, $\epsilon = 10^{-5}$ and $\zeta = 1$.[1] The large value of ζ indicates that the total variation constraint needed a high weight to avoid gaps between frames.

$\mathbf{A}$ was initialized using small nonnegative random values. $\mathbf{H}^{(i)}$ was initialized as $\mathbf{H}^{(i)}_{f,p} \leftarrow 1 - p/2L + \alpha$, $(f = 0, ..., L-1)$, where α is a small nonnegative random value, based on the assumption that the earlier arrivals are louder.

4.3. Data pre- and post-processing

The input signals were standardized, and then multiplied by a normalizing constant 0.01. This pre-processing makes the algorithm perform more consistently given other parameters. However, it does not keep the model from assigning different weights to the channels in $\mathbf{H}$ depending on their similarity to the other channels, thus reducing the error introduced by varying qualities of input signals [9].

The phase of the clean signal spectrogram was estimated using the phase of an input signal chosen at random. Although this provided an estimate that was fairly rough because of the reverberant nature of the input phase and the likely misalignment of the reconstruction with the input, in practice it provided better results than the Griffin-Lim algorithm [25]. The difficulty of estimating the phase highlights the usefulness of deriving a complex-valued NMF model for the problem [26].

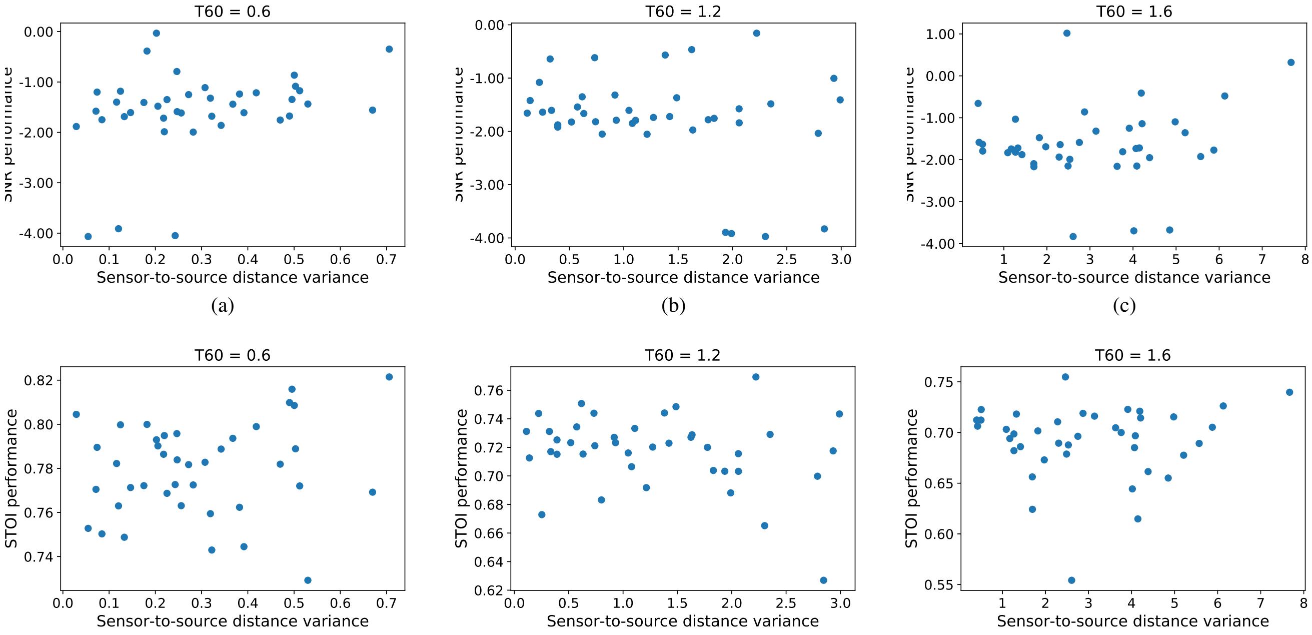
5. EVALUATION

Table 1 displays average STOI and SNR for reverberant input signals in the three T60 scenarios, and for clean speech reconstructions using (1) the NTF baseline using KL-divergence as our objective function, (2) the NTF model using the speech prior (3) the NTF model using the speech prior with regularization. Given that the delay between reverberant signals and the clean signal is unknown, we aligned the reverberant signals using cross correlation for the evaluation. Results show an increase in quality from deploying the baseline model, then further significant improvements first when adding the

[1] These values depend on the magnitude of the input signal. In our experiment, the data was normalized as described in 4.3.

Table 1. Average STOI and SNR results over 40 random sensor configurations per T60 value.

T60		0.6 sec.		1.2 sec.		1.6 sec.	
Reverberant	SNR	-1.90		-2.19		-2.12	
input (average)	STOI	0.66		0.55		0.53	
NTF baseline	SNR	-2.03	(-0.13)	-1.93	$(+0.26)$	-1.82	$(+0.30)$
(with KL-div)	STOI	0.61	(-0.05)	0.61	$(+0.06)$	0.61	$(+0.08)$
NTF with	SNR	-1.79	$(+0.11)$	-1.85	$(+0.34)$	-1.65	$(+0.47)$
speech prior	STOI	0.75	$(+0.09)$	0.69	$(+0.14)$	0.67	$(+0.14)$
NTF w/ speech prior	SNR	-1.57	$(+0.33)$	-1.74	$(+0.45)$	-1.64	$(+0.48)$
and regularization	STOI	0.77	$(+0.11)$	0.72	$(+0.17)$	0.69	$(+0.16)$



(a) (b) (c)

(d) (e) (f)

Fig. 2. Performance based on the variance of the sensor-to-source distance.

prior, and then when adding the regularization. The baseline only improves the quality of the signal when the reverberation time is long enough. The proposed model, however, increases the quality even in the case of T60 at 0.6 seconds. These measured improvements were obtained despite artifacts introduced in the form of missing phase information: a model that properly estimates phase would most likely further improve the quality of the output.[2].

Fig. 2 displays how the variance of distances from the source to the sensors affects performance. We can see that the proposed method performs consistently across different scenarios.

6. CONCLUSION

We develop a model based on Mirsamadi *et al.*'s NTF multi-channel dereverberation [9], where channel-specific dereveberation tasks are tied up by having the speech source as the common anchor across channels. In this way the filter matrices estimate the sensor-wise RIR while the model estimate the dry source spectrogram directly. To improve this model that can converge at an unattractive local minima, such as the source learning reverberation instead of the filter, we propose a regularized NTF model by introducing our prior knowledge about the source. The proposed speech basis vectors further help decompose the source estimate so that the model can focus on estimating the dry activation. We also regularize the NMF activations using sparsity and total variation, because sparsity reduces reverberation by penalizing spread of energy over time, while total variation helps avoid gaps across frames.

This algorithm learns a magnitude-domain estimate of the clean signal and of the room impulse response filters. The next steps include deriving the complex tensor factorization version of this work to avoid the artifacts induced by the lack of phase information and involving other types of prior knowledge of the clean speech such as probabilistic models [4] extended to the multi-channel case. Also, NMF is limited by its assumption of independence between frames. Recurrent neural networks or hidden Markov models may be useful in addressing this limitation [27, 28].

[2] Input and reconstructed signals can be heard at `http://homes.sice.indiana.edu/scwager/collaborative_dereverberation.html`

7. REFERENCES

[1] M. Kim and P. Smaragdis, “Collaborative audio enhancement using probabilistic latent component sharing,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 896–900.

[2] ——, “Efficient neighborhood-based topic modeling for collaborative audio enhancement on massive crowdsourced recordings,” in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 41–45.

[3] N. Stefanakis and A. Mouchtaris, “Maximum component elimination in mixing of user generated audio recordings,” in *IEEE International Workshop on Multimedia Signal Processing*, 2017.

[4] D. Liang, M. D. Hoffman, and G. J. Mysore, “Speech dereverberation using a learned speech model,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1871–1875.

[5] R. Maas, E. A. Habets, A. Sehr, and W. Kellermann, “On the application of reverberation suppression to robust speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 297–300.

[6] K. Kumar, R. Singh, B. Raj, and R. Stern, “Gammatone subband magnitude-domain dereverberation for ASR,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4604–4607.

[7] J. S. Erkelens and R. Heusdens, “Correlation-based and model-based blind single-channel late-reverberation suppression in noisy time-varying acoustical environments,” *IEEE transactions on audio, speech, and language processing*, vol. 18, no. 7, pp. 1746–1765, 2010.

[8] N. Gillis, “Sparse and unique nonnegative matrix factorization through data preprocessing,” *Journal of Machine Learning Research*, vol. 13, no. Nov, pp. 3349–3386, 2012.

[9] S. Mirsamadi and J. Hansen, “Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications,” in *Interspeech*, 2014, pp. 2828–2832.

[10] B. Ermiş, E. Acar, and A. T. Cemgil, “Link prediction in heterogeneous data via generalized coupled tensor factorization,” *Data Mining and Knowledge Discovery*, vol. 29, no. 1, pp. 203–236, 2015.

[11] M. Kim, J. Yoo, K. Kang, and S. Choi, “Nonnegative matrix partial co-factorization for spectral and temporal drum source separation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 6, pp. 1192–1204, 2011.

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

[13] N. Seichepine, S. Essid, C. Févotte, and O. Cappé, “Soft nonnegative matrix co-factorization,” *IEEE Transactions on Signal Processing*, vol. 62, no. 22, pp. 5940–5949, 2014.

[14] N. Ono, H. Kohno, N. Ito, and S. Sagayama, “Blind alignment of asynchronously recorded signals for distributed microphone array,” in *Applications of Signal Processing to Audio and Acoustics, 2009. WASPAA’09. IEEE Workshop on*. IEEE, 2009, pp. 161–164.

[15] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *Advances in neural information processing systems*, 2001, pp. 556–562.

[16] B. King, C. Févotte, and P. Smaragdis, “Optimal cost function and magnitude power for nmf-based speech separation and music interpolation,” in *IEEE Machine Learning for Signal Processing Conference*, 2012.

[17] A. Lefevre, F. Bach, and C. Févotte, “Itakura-Saito nonnegative matrix factorization with group sparsity,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 21–24.

[18] M. Kim and P. Smaragdis, “Mixtures of local dictionaries for unsupervised speech enhancement,” *IEEE Signal processing letters*, vol. 22, no. 3, pp. 293–297, 2015.

[19] D. L. Sun and G. J. Mysore, “Universal speech models for speaker independent single channel source separation,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 141–145.

[20] D. M. Sima, “Regularization techniques in model fitting and parameter estimation,” 2006.

[21] L. I. Rudin, S. Osher, and E. Fatemi, “Nonlinear total variation based noise removal algorithms,” *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[22] E. Vincent and B. R. Campbell, “Matlab roomsimove toolbox,” 2008. [Online]. Available: http://homepages.loria.fr/evincent/software/Roomsimove_1.4.zip

[23] J. B. Allen and D. A. Berkley, “Image method for efficiently simulating small-room acoustics,” *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[24] L. D. Consortium *et al.*, “Timit acoustic-phonetic continuous speech corpus,” *URL http://www. ldc. upenn. edu/Catalog/CatalogEntry. jsp*, 1993.

[25] D. Griffin and J. Lim, “Signal estimation from modified short-time Fourier transform,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[26] H. Sawada, H. Kameoka, S. Araki, and N. Ueda, “Formulations and algorithms for multichannel complex NMF,” in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 229–232.

[27] G. J. Mysore, P. Smaragdis, and B. Raj, “Non-negative Hidden Markov Modeling of audio with application to source separation.” in *LVA/ICA*. Springer, 2010, pp. 140–148.

[28] N. Boulanger-Lewandowski, G. J. Mysore, and M. Hoffman, “Exploiting long-term temporal dependencies in NMF using recurrent neural networks with application to source separation,” in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6969–6973.