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, allowing the dereverberation problem to benefit from microphone arrays. The factorization learns both individual reverberation filters and channel-specific delays, which makes it possible to employ an ad-hoc microphone array with heterogeneous sensors (such as multi-channel recordings by a crowd) even if they are not synchronized. We integrate two prior-knowledge regularization schemes to increase the stability of dereverberation performance. First, a Nonnegative Matrix Factorization (NMF) inner routine is introduced to inform the original NTF problem of the pre-trained clean speech basis vectors, so that the optimization process can focus on estimating their activations rather than the whole clean speech spectra. Second, the NMF activation matrix is further regularized to take on characteristics of dry signals using sparsity and smoothness constraints. Empirical dereverberation results on different simulated reverberation setups show that the prior-knowledge regularization schemes improve both recovered sound quality and speech intelligibility compared to a baseline NTF approach.

Index Terms— multi-channel dereverberation, nonnegative matrix factorization, nonnegative tensor factorization, collaborative audio enhancement, speech enhancement

1. INTRODUCTION

Collaborative audio enhancement methods recover a clean signal from multiple low-quality recordings, each of which retains elements of the common desired signal while also being corrupted in some way. Removal of common distortions in crowdsourced recordings such as clipping, additive noise, and bandlimiting has been addressed in [1, 2, 3]. We add to the collaborative-audio-enhancement toolbox by proposing a model that reduces reverberation. Reverberation decreases a signal's perceptual audio quality, and is of concern in applications such as automatic speech recognition.

We formulate dereverberation as link prediction, modeling a reverberant signal as a frequency-domain convolution of a clean signal with different unknown Room-Impulse Responses (RIRs) as shown in (1), where we introduce $\circledast$ for simplicity of notation of subband-wise convolution:

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

In this equation, $\mathbf{S}, \mathbf{X}^{(i)} \in \mathbb{R}_+^{F \times T}$ denote 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 the 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).

This model approximates the complex-domain convolution theorem in the magnitude domain and applies to the case where the RIR lasts multiple Discrete Fourier Transform (DFT) frames. Our framing of the search for an unknown common element across multiple signals as link prediction is similar to that of Ermiş *et al.* [4], who demonstrate the ability of a tensor factorization model to effectively harness information in large-scale data.

The NTF-based formulation applied to the model in (1) has been shown to be effective in dereverberation. Mirsamadi *et al.* [5] demonstrate improvements in automatic speech recognition results over the single-channel baseline [6]. Our model builds on this tensor factorization approach, noting in passing the probabilistic multi-channel models proposed by [7, 8, 9].

Although each channel-specific signal recovery can benefit from the shared component, direct estimation of a large speech spectrogram is challenging. In reverberant recordings, longer RIRs span multiple DFT frames, while tensor factorization models assume independence between frames. The convolution involved in reverberation results in a large number of unknown parameters. This issue is exacerbated by the underdetermined nature [10] of a tensor factorization model. Many ways of incorporating prior knowledge about the source have been used to restrict the solution space and improve the results of speech dereverberation. A common approach is to use the ℓ_1- or ℓ_2-norm sparsity constraint to enforce the characteristics of dry speech [11, 12]: such models treat every frame independently. Nakatani *et al.* [13] propose a time-dependent model that captures the relationship between neighboring frames by parameterizing the variance between nearby frames in an expectation-maximization framework.

Mohammida *et al.* [14] improve sparse convolutive single-channel dereverberation by further factorizing the shared speech spectrogram by using Nonnegative Matrix Factorization (NMF). They either train basis vectors online from the source estimate, or pre-train them offline on speaker-independent clean-speech data. The use of clean-speech basis vectors pre-trained offline is a common technique in matrix factorization [15]. Mohammida *et al.*'s integrated method yields a considerably higher speech quality than both the baseline convolutive approach and a state-of-the-art spectral enhancement method [16, 17]. We adapt this method of nesting NMF inside of a matrix factorization to the NTF model for analyzing

multi-channel signals.

The multi-channel speech dereverberation approach [5] 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 [18], it uses multiplicative update rules to ensure nonnegativity of parameters.

The model shown in (2) offers two advantages: scalability and robustness to issues such as misalignment. It is scalable to unknown geometric configurations of the source and sensors thanks to its unsupervised nature. Also, it is robust to issues such as misalignment common to crowdsourced recordings due to the fact that source-to-sensor distances vary between sensors and are unknown. Other approaches to alignment of the signals— such as cross-correlation— are not always applicable because the RIRs make the channel waveforms differ significantly from each other. The NTF model automatically aligns the signals in time by adjusting the RIR filter $\mathbf{H}$ for each channel. It also automatically gives lower weights to bad quality recordings and can address issues such as bandlimiting by zeroing out corresponding frequencies in the filter.

Two concerns arise from the large number of equivalent factorization processes permitted by this model. The model given by (2) risks a trivial solution where $\mathbf{H}$ represents a filter where all but the first column is zero-valued, with $\mathbf{S}$ an averaged reverberant signal. Also, direct estimation of $\mathbf{S}$ means that the number of parameters to be learned can become very large. Our objective is to address these concerns.

2. MODEL DEVELOPMENT

In this section, we develop a model that integrates NTF and NMF to dereverberate sound in the multi-channel case. To this model we incorporate prior knowledge about clean speech in order to reduce the number of equivalent factorizations. We cut down on the number of parameters in the activation matrix by factorizing the shared speech spectrogram using a fixed-basis vector array pre-learned from dry speech. We then regularize the NTF problem by enforcing the structure of the activation matrix to correspond to that of a clean signal. This ensures that the model represents the reverberation in the estimated reverberation filters $\mathbf{H}$, not in $\mathbf{S}$. Although we impose sparsity on the NMF activations, the additional total variation constraint on them enhances the smoothness of the activations over time, a commonly used technique in computer vision [19, 20].

Specifically, instead of directly estimating the clean signal $\mathbf{S}$, we set $\mathbf{S} \approx \mathbf{WA}$, where $\mathbf{W} \in \mathbb{R}_+^{F \times R}$ and $\mathbf{A} \in \mathbb{R}_+^{R \times T}$. The lower rank approximation $R < F, T$ is common. $\mathbf{W}$ is a previously trained NMF basis vectors to represent clean speech sound. 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 $R \times T$.

Finally, we use the generalized KL-divergence in our objective function. This metric has been demonstrated by King et al. [21] to produce better results in NMF applications to audio source separation than the Frobenius norm.

2.1. Objective function

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

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

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 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 [22] 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 activation matrix, 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 [19, 20], 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 NTF baseline with the proposed inner NMF routine recovers $\mathbf{A}$ with a significant amount of discontinuity, 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).

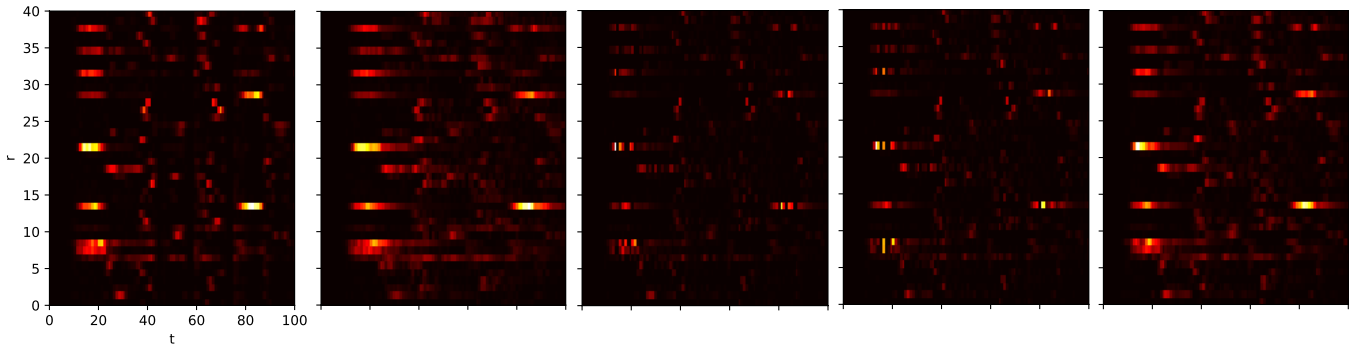
2.2. Multiplicative update rules

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:

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

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

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



(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).

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. As introduced in [5], we normalize the filter estimate $\mathbf{H}$ at every iteration: $\sum_f \mathbf{H}^{(i)}_{f,p} = 1$. This constraint causes the signals that differ more from the rest to be weighted less, thus improving the overall result.

3. EXPERIMENT

3.1. Room, source and sensor configurations

We generated RIRs using the `roomsimove` toolbox [23]. This program implements the image method [24], 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 source at 75 per cent of the width and 50 per cent of the height of the room. In order to simulate the crowdsourced scenario where the sensors are not equidistant from the source causing the signals to be misaligned, 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.

3.2. Input data and parameter settings

The RIRs were convolved with a clean signal from the TIMIT dataset [25] 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 [5] (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.

3.3. Data pre- and post-processing

The input signals were standardized, and then multiplied by a normalizing constant 0.01. The STFT frame size was set to 64 ms and the hop size to 32 ms. This pre-processing makes the algorithm perform more consistently given other parameters. Additionally, it does not hinder the model from assigning higher weights to the channels in $\mathbf{H}$ depending on their similarity to the other channels, thus reducing the error introduced by anomalous low quality input signals.

The phase of the clean signal spectrogram was estimated both using the phase of an input signal chosen at random and the Griffin-Lim algorithm [26]. Neither estimate is ideal and results were similar using either approach. This difficulty highlights the usefulness of deriving a complex-valued NMF model for the problem [27].

4. EVALUATION

Table 1 displays average Short-Term Objective Intelligibility (STOI) and Signal-to-Noise Ratio (SNR) for reverberant input signals in the three T60 scenarios, and for clean-speech reconstructions using (1)

[1] These values depend on the magnitude of the input signal. In our experiment, the data was normalized as described in 3.3.

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

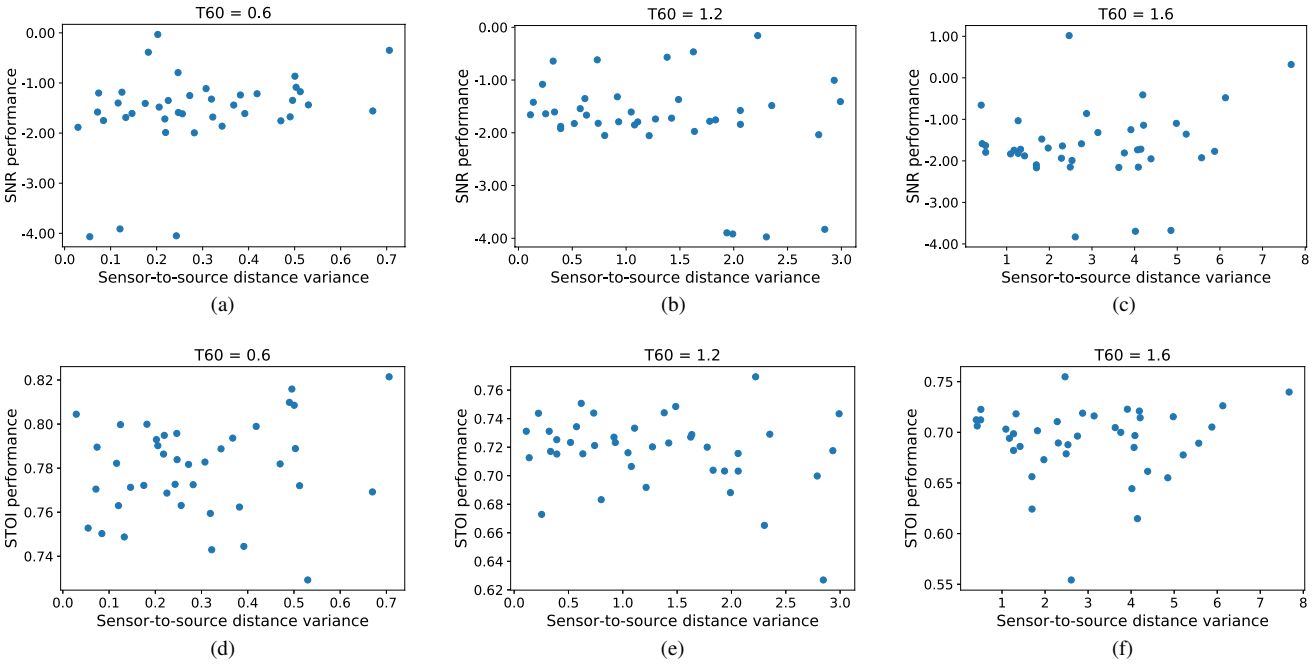


Fig. 2. Performance based on the variance of the sensor-to-source distance.

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 prior, and then when adding the regularization. The baseline multichannel model only improves the quality of the signal when T60 is long enough. The proposed model, however, increases the quality even in the shortest case of 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, lending itself to recordings with ad-hoc microphone arrays.

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

5. CONCLUSION

We develop an NTF model for multichannel speech dereverberation with integrated NMF of the dry source estimate. The speech source estimation common across channels anchors dereverberation tasks. The filter matrices provide estimates of the sensor-wise RIRs, and the inner factorization estimates the dry source spectrogram. To avoid the model's converging at an unattractive local minimum—such as the source learning reverberation instead of the filter—we introduce source prior knowledge through regularization, smoothing, and clean-speech basis vectors. The resulting decomposition of the source estimate focuses the model on estimating the activations. Regularization on the NMF activations imposes sparsity using the ℓ_1-norm to reduce reverberation by penalizing the spread of energy over time, and imposes smoothness using the total variation constraint to avoid discontinuities in energy across frames.

This algorithm learns a magnitude-domain estimate of the clean signal and of the RIR filters. Next steps include deriving the complex tensor factorization version of this work to avoid the artifacts induced by the lack of phase information. This work contributes to the set of algorithms available for collaborative audio enhancement.

6. REFERENCES

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

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

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

[6] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2009, pp. 45–48.

[7] R. Singh, B. Raj, and P. Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multichannel signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 1914–1917.

[8] I. Kodrasi, A. Jukić, and S. Doclo, "Robust sparsity-promoting acoustic multi-channel equalization for speech dereverberation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 166–170.

[9] M. Yu and F. K. Soong, "Constrained multichannel speech dereverberation," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

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

[11] A. Jukic, T. van Waterschoot, T. Gerkmann, and S. Doclo, "A general framework for incorporating time–frequency domain sparsity in multichannel speech dereverberation," *Journal of the Audio Engineering Society*, vol. 65, no. 1/2, pp. 17–30, 2017.

[12] F. Ibarrola, L. Di Persia, and R. Spies, "On the use of convolutive nonnegative matrix factorization with mixed penalization for blind speech dereverberation," 2017.

[13] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi, and B.-H. Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[14] N. Mohammadiha and S. Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 2, pp. 276–289, 2016.

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

[16] E. A. Habets, S. Gannot, and I. Cohen, "Late reverberant spectral variance estimation based on a statistical model," *IEEE Signal Processing Letters*, vol. 16, no. 9, pp. 770–773, 2009.

[17] 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.

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

[19] D. M. Sima, "Regularization techniques in model fitting and parameter estimation," Ph.D. dissertation, Katholieke Universiteit Leuven Leuven, Belgium, 2006.

[20] 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.

[21] 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.

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

[23] E. Vincent and B. R. Campbell, "Matlab roomsimove toolbox," 2008. [Online]. Available: http://homepages.loria.fr/evincent/software/Roomsimove/_1.4.zip

[24] 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.

[25] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus, 1993," *Linguistic Data Consortium, Philadelphia*, 1993.

[26] 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.

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