In Proc. of ICASSP2017
New Orleans, USA, March 2017

# SPEECH DEREVERBERATION USING NMF WITH REGULARIZED ROOM IMPULSE RESPONSE

*Nikhil Mohanan* *Rajbabu Velmurugan* *Preeti Rao*

Dept. of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India 400076
Email: {nikhilm, rajbabu, prao}@ee.iitb.ac.in

## ABSTRACT

In this paper, various regularizations on the room impulse response (RIR) are proposed to obtain better single-channel speech dereverberation in the non-negative matrix factorization (NMF) framework. The regularizations on the RIR are motivated by the spectral domain representation of the RIR. To obtain better estimates of the RIR and clean speech, we propose three modifications (i) to obtain a sparse RIR (ii) a frequency envelop constrained RIR and (iii) to include the early part of the RIR. The performance of the proposed regularizers are evaluated by considering speech enhancement measures. While it is observed that the regularizers lead to an improved estimate of the RIR, they do not necessarily lead to speech enhancement in all the cases. For the experiments conducted using the RIRs from the REVERB 2014 challenge and sentences from the TIMIT database, the regularizer that includes the early part of the RIR shows reasonable improvement in speech enhancement measures.

***Index Terms***— speech dereverberation, NMF, RIR, regularization

## 1. INTRODUCTION

Distant speech enhancement and recognition has been gaining importance over the past decade due to the prevalence of audio capturing instruments [1]. Speech processing (for recognition or enhancement) in such environments differs from traditional speech processing as it has to compensate for reverberation effects in the captured data. While speech dereverberation has been an active research area for a long time [2], it has gained interest recently [3] for the reason mentioned above. The effect of reverberation depends on both the speech signal and the room impulse response (RIR) under consideration. The characteristics of this RIR have a significant effect on the reverberant signal, and hence a good understanding of this is relevant for dereverberation. Speech dereverberation can be done using single- or multi-channel data depending on the application of interest. In this research we address single-channel dereverberation in the distant speech scenario.

Dereverberation methods can be classified as those which (i) cancel reverberation, e.g., blind deconvolution based methods (ii) suppress reverberation, e.g., spectral subtraction, linear prediction (LP) based methods, and statistical methods for spectral enhancement [2]. Here we consider reverberation suppression using non-negative matrix factorization (NMF). This uses the magnitude spectrum of the reverberant signal, and with minimal prior knowledge of the RIR, to obtain the dereverberated speech signal. The earliest work on NMF based dereverberation [4] used a convolutive NMF (referred to as N-CTF) and provided a statistical motivation for the use of NMF. More recently there have been several improvements over N-CTF in both single-channel [5, 6, 7, 8, 9], and multi-channel [10], [11], [12] scenarios. In [5, 7] the initial N-CTF model for speech dereverberation is improved by incorporating various NMF models for the speech signal within the original model leading to several N-CTF+NMF approaches. In [6], supervised N-CTF+NMF approaches have been proposed to handle reverberation in noisy environments. While most of these methods use the short-time Fourier transform (STFT) representation of the signal when performing NMF, the method proposed in [9] uses a Gammatone filtered spectrum and has shown improvements in automatic speech recognition (ASR) word error rates (WER) over the earlier NMF based methods [4]. These methods also have proposed incorporating a sparsity constraint on the speech signal as a regularizer to improve speech estimates. While the approaches in [4], [6],[5] have shown improvement in speech enhancement measures, other approaches [9], [8] have shown improvement in ASR tasks.

Single-channel NMF dereverberation problem is treated as a deconvolution in the sub-band domain. Obtaining the RIR from the observed reverberant signal is a unconstrained problem with possibly many solutions. Hence, it is required to impose appropriate regularizers or constraints to obtain the solution. Earlier NMF approaches obtained reasonable estimates for speech by imposing a sparsity constraint or by using appropriate models for speech. While the estimation accuracy of the speech signal is considered in all these approaches, they do not provide an evaluation of the accuracy of the RIR estimates. Though [6] claims to estimate the RIR, it does not provide any analysis or results on the RIR estimation.

While regularization of RIRs in single-channel scenario has not been considered before, multi-channel RIR regularization has been proposed in [11] leading to speech enhancement. The objective in this work is to obtain better single-channel RIR estimates using appropriate constraints on the RIR. These constraints are motivated both from time- and frequency-domain models of the RIR [2]. The improved estimates are expected to provide better estimates of both the clean speech signal and RIR. The proposed regularizations on the RIR are evaluated using speech enhancement measures such as perceptual evaluation of speech quality (PESQ), speech-reverberation modulation ratio (SRMR), and cepstral distance (CD) [3].

## 2. NMF BASED SPEECH DEREVERBERATION

For NMF to be used in speech dereverberation the spectral representation of the observed or reverberated signal needs to be understood. In time-domain, the observed signal $x[n]$ in a microphone as result of a clean speech signal $s[n]$ in a reverberant room with room im-

Authors acknowledge the support by Council of Scientific and Industrial Research (CSIR), India and Tata Consultancy Services (TCS), India

pulse response (RIR) $h[n]$, can be represented as

$$x[n] = s[n] * h[n] = \sum_{m=0}^{M-1} h[m]s[n-m], \quad (1)$$

where $*$ represents convolution in time-domain, and $M$ is the length of the RIR. In NMF, we are interested in a STFT representation of (1). It has been shown that the magnitude STFT of $x[n]$ [5],

$$X[k,n] \approx \sum_{m=0}^{L_h-1} H[k,m]S[k,n-m], k \in \{0,1,...,K-1\}, \quad (2)$$

where $H[k,n]$, $S[k,n]$ are the magnitude STFTs of $h[n]$, $s[n]$, respectively, $L_h$ is the length of the RIR in the STFT domain, and $K$ is the number of frequency bands in the STFT domain. To be consistent with [5] we will refer to (2) as the non-negative convolutive transfer function (N-CTF) model for reverberation. In matrix form, the reverberant spectrogram (2) can be represented as

$$\boldsymbol{X} \approx \sum_m \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}, \quad (3)$$

where the clean speech spectrogram of length $T$ frames is

$$\boldsymbol{S} = \left[\boldsymbol{s}_0, \boldsymbol{s}_1, \ldots, \boldsymbol{s}_{T-1}\right] \quad (4)$$

with $\boldsymbol{s}_i \in \mathbb{R}^{K\times 1}$, $^{m\rightarrow}$ indicates a column-shift by $m-1$ positions to the right, and the RIR spectrogram for a fixed frame $m$

$$\boldsymbol{H}_m = diag\left(H[0,m], \ldots, H[K-1,m]\right) \quad (5)$$

a diagonal matrix. As observed in [4], this representation for the reverberant signal comprises of components of clean speech spectrum being blurred by the temporal evolution of the $\boldsymbol{H}$ components.

Given an observed reverberant magnitude spectrogram $\boldsymbol{Y}$, we are interested in obtaining $\boldsymbol{H}$ and $\boldsymbol{S}$, which minimize the error between $\boldsymbol{Y}$ and $\boldsymbol{X}$. In [5], the Kullback-Leibler (KL) divergence between them is minimized under non-negativity constraints on $\boldsymbol{S}$ and $\boldsymbol{H}$, and an additional constraint on $H[k,m]$ to avoid indeterminacy in the estimates. The general form of the cost-function is

$$\begin{aligned} J(\boldsymbol{S},\boldsymbol{H}) &= D_{KL}(\boldsymbol{Y},\boldsymbol{X}) \\ &= D_{KL}(\boldsymbol{Y}, \sum_m \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}) \\ \text{s.t.} \sum_m H[k,m] = 1,\ k \in \{0,\ldots,K-1\}, \boldsymbol{H}_m \geq 0, \boldsymbol{S} \geq 0, \end{aligned} \quad (6)$$

where $\boldsymbol{X}$ is from (3) and $D_{KL}(\boldsymbol{Y},\boldsymbol{X})$ is defined as,

$$D_{KL}(\boldsymbol{Y},\boldsymbol{X}) = \sum_{k,m}(Y[k,m]ln(\frac{X[k,m]}{Y[k,m]}) - X[k,m] + Y[k,m]),$$

and can be solved using multiplicative update rules to obtain the estimates $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{H}}$. The cost function in (6) can be modified to include sparsity constraint on $\boldsymbol{S}$ leading to better estimates. The N-CTF model was further improved by incorporating a NMF model for the speech spectrum ($\boldsymbol{S} = \boldsymbol{W}\boldsymbol{A}$). The corresponding updated cost-function is

$$\begin{aligned} J(\boldsymbol{W},\boldsymbol{A},\boldsymbol{H}) &= D_{KL}\left(\boldsymbol{Y}, \sum_m \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}\right) \\ &= D_{KL}\left(\boldsymbol{Y}, \sum_m \boldsymbol{H}_m \overset{m\rightarrow}{(\boldsymbol{W}\boldsymbol{A})}\right) + \| \boldsymbol{A} \|_1 \\ \text{s.t.} \sum_m H[k,m] = 1, \boldsymbol{H}_m \geq 0, \boldsymbol{W} \geq 0, \boldsymbol{A} \geq 0. \end{aligned} \quad (7)$$

where $\| \cdot \|_1$ denotes $l1$-norm and promotes sparsity. They also suggested another weighted method that combined the N-CTF model for dereverberation along with the NMF speech model. However, their experiments and results suggest that the integrated model in (7) performs better and hence not discussed here. They proposed three possible NMF models for the speech signal which are either unsupervised or semi-supervised. In the unsupervised method of speech modeling the basis vectors $\boldsymbol{W}$ were learnt online from the reverberant signal and referred to as N-CTF+NMF. The other two methods were semi-supervised approaches where the basis matrix $\boldsymbol{W}$ was learnt offline from training data and are not discussed here. Their results indicate improved speech enhancement measures and do not provide any speech recognition results. In [6], NMF based approaches in [4], [5] have been extended to do both dereverberation and denoising in a supervised setting. Based on their speech enhancement results, they conclude that the N-CTF+NMF in a supervised context provides a better estimate of the RIR if only the speech estimates are used than using both speech and noise estimates.

All the NMF based methods discussed have demonstrated improvements in either speech enhancement or speech recognition measures (WER) indicating successful dereverberation. However, these existing approaches have not considered the estimates obtained for the RIR, i.e., $\boldsymbol{H}$. In the next section, we motivate this problem and present our proposed modification to handle this.

## 3. PROPOSED REGULARIZATION FOR RIR ($\boldsymbol{H}$)

As seen in (2), the underlying assumption in NMF based approaches to solve the dereverberation problem is that the reverberant signal spectrum for a specific frequency band can be considered as a convolution of the clean speech spectrogram and $\boldsymbol{H}$ for that band.

It should be noted that the magnitude spectrogram $\boldsymbol{H}$ in (2) does not correspond to the magnitude STFT of $h[n]$, but is an approximation assuming cross-band effects can be neglected in the reverberant signal spectrum [5], [13]. However, it is still valid to use (2) and operate in the sub-band domain to perform dereverberation [11]. We will not be able to verify the accuracy of estimated $\boldsymbol{H}$ by comparing it to the STFT of the true or actual RIRs $\boldsymbol{H}_{\text{true}}$ . However, we can compare the estimated $\boldsymbol{H}$ to the STFT of an approximate RIR $\tilde{\boldsymbol{H}}$ obtained from $\boldsymbol{H}_{\text{true}}$. Following the approach suggested in [13], we compared $\tilde{\boldsymbol{H}}$ and $\boldsymbol{H}$ for different frequency bands and found them to be similar. Hence, we can compare the estimated $\hat{\boldsymbol{H}}$ with $\boldsymbol{H}$, to evaluate the accuracy of RIR estimation from the reverberated speech signal.

We compare the narrow band estimate of the RIR ($\hat{\boldsymbol{H}}_k$) obtained using N-CTF to $\boldsymbol{H}_k$ for a RIR of $T_{60} \approx 700$ ms in Fig. 1. It can be seen that the estimated RIR matches the expected RIR more closely during the later part of the RIR. In the early part of the RIRs, the estimates are more erroneous, when compared to the later parts. This is one of the main motivations for the proposed approach, where we intend to use appropriate regularizer on $\boldsymbol{H}$ in the NMF framework so that $\hat{\boldsymbol{H}}$ is improved. Such an improved estimate for RIR, will also lead to a better estimate of the speech signal leading to improved speech enhancement and automatic speech recognition measures.

We propose three possible regularizations to $\boldsymbol{H}$ for obtaining better estimates of $\boldsymbol{H}$. We discuss the three choices and the corresponding cost function.

### 3.1. Sparsity of RIR

The time-domain model by Pollack [2] is a reasonable characterization based on the $T_{60}$ of a room. In time-domain the RIR is expo-

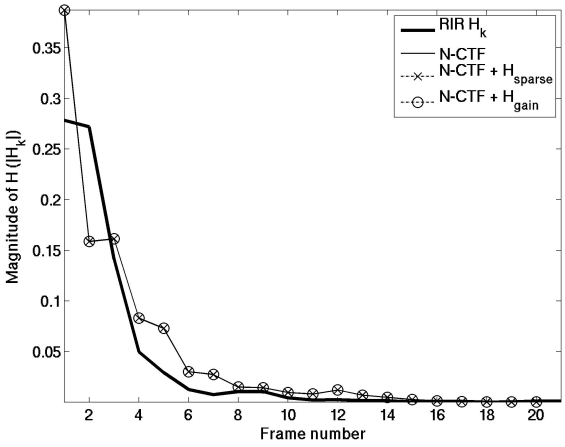


**Fig. 1**. Comparison of normalized estimates of the RIR ($\hat{\boldsymbol{H}}_k$) with the actual RIR ($\boldsymbol{H}_k$) for a specific RIR with $T_{60} = 700$ ms and frequency band ($k = 10$). It can be seen that while their late parts are closer, early parts are different. Further, the proposed regularizers $\boldsymbol{H}_{gain}$, $\boldsymbol{H}_{sparse}$ do not lead to a better RIR estimate when used with just the N-CTF approach.

nentially decaying with a decay factor dependent on the $T_{60}$. Correspondingly the magnitude STFT domain representation of the RIR has larger magnitude values during the early part of the reverberation and dies down to smaller values during the late part, indicating that most entries in $\boldsymbol{H}$ have values close to zero. Hence, the simplest constraint is to assume that $\boldsymbol{H}$ has a sparse structure. We incorporate this into the basic N-CTF framework and have the modified cost-function,

$$J(\boldsymbol{S},\boldsymbol{H}) = D_{KL}\left(\boldsymbol{Y}, \sum_m \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}\right) + \lambda \parallel \boldsymbol{H} \parallel_1$$
$$\text{s.t.} \sum_m H[k,m] = 1, k = 0,\ldots,K-1; \boldsymbol{H}_m \geq 0, \boldsymbol{S} \geq 0. \quad (8)$$

and $\lambda$ decides the weight given to sparsity of $\boldsymbol{H}$. The value of $\lambda$ was chosen empirically based on enhancements obtained during the experiments. This method will be referred to as N-CTF+$\boldsymbol{H}_{sparse}$. The multiplicative updates for $\boldsymbol{H}$ and $\boldsymbol{S}$ can be obtained in a way similar to that in [5].

### 3.2. Sub-band gains constrained RIR

Depending on the $T_{60}$ of the room and the distance between the source and microphone the sub-band gain ($g[k]$) for the RIRs can be modelled as a function of the frequency $k$ for a fixed frame. Such models can be obtained by fitting polynomial functions on existing magnitude STFTs of the RIRs. If the RIRs for a room are available, this can be included in the N-CTF framework as opposed to constraining the sub-band sums of $\boldsymbol{H}$ to be unity. The corresponding cost function is,

$$J(\boldsymbol{S},\boldsymbol{H}) = D_{KL}\left(\boldsymbol{Y}, \sum_m \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}\right)$$
$$\text{s.t.} \sum_m H[k,m] = g[k], k = 0,\ldots,K-1; \boldsymbol{H}_m \geq 0, \boldsymbol{S} \geq 0. \quad (9)$$

This method will be referred to as N-CTF+$\boldsymbol{H}_{gain}$.

### 3.3. Inclusion of early part of the RIR

The RIR can be broadly divided into two regions - early reverberation and reverberation tail (late reverberation) [2]. The early part accounts for the reflections up to 50 ms after the direct path and the rest forms the reverberation tail. The early part modifies the spectrum within a phone region, whereas the late reverberation results in changing the spectral characteristics of the present phone by the preceding phone. It is shown in the literature that late reverberation causes degradation of speech and need to be removed [2]. However, retaining the early part improves speech intelligibility and ASR performance [14]. Once the clean speech spectrum is estimated ($\hat{\boldsymbol{S}}$) using the N-CTF framework, we propose to use the early part of the estimated RIR ($\hat{\boldsymbol{H}}_{early}$) as shown below to obtain an improved dereverberated spectrum ($\boldsymbol{S}_{new}$).

$$\boldsymbol{S}_{new}(n,k) = \hat{\boldsymbol{S}}(n,k) * \hat{\boldsymbol{H}}_{early}(n,k), k \in \{0,1,...,K-1\} \quad (10)$$

This method will be referred to as N-CTF+$\boldsymbol{H}_{early}$.

### 3.4. Analysis of estimated RIRs

In Figures 1 and 2 we compare the RIR estimates ($\hat{\boldsymbol{H}}$) obtained for various regularization on the RIR for a particular narrow band. From Fig. 1 it can be seen that the estimates $\hat{\boldsymbol{H}}$ obtained with regularization when using the basic N-CTF model do not show any improvement, i.e., N-CTF+$\boldsymbol{H}_{gain}$ and N-CTF+$\boldsymbol{H}_{sparse}$ do not improve the RIR estimate as compared to N-CTF. However, when regularization is used in addition to the speech model, the estimated $\hat{\boldsymbol{H}}$ is closer to $\boldsymbol{H}$ as seen in Fig. 2, i.e., $\boldsymbol{H}$ estimation improved in regularized N-CTF+NMF+$\boldsymbol{H}_{sparse}$ method as compared to the reference method N-CTF+NMF. Though not included here, we also considered mean squared error between $\hat{\boldsymbol{H}}$ and $\boldsymbol{H}$ across all bands. This did not show any significant improvement. We attribute this to the fact that the measure is an average across multiple frequency bands where some bands show improvement, while others do not. Together it can be inferred that regularization of the RIR when used with speech models leads to better RIR estimation in many frequency bands.

## 4. RESULTS

The performance of the algorithms discussed in Sec. 3 were evaluated for speech enhancement.

### 4.1. Database and measures

The speech enhancement performance was compared using the TIMIT database [15]. A sub-set of 16 different sentences spoken by 16 distinct speakers was used. Measured RIRs available from the REVERB challenge [3] were used for the evaluation. The RIRs correspond to an 8 channel circular array of diameter 10 cm. Each TIMIT sentence was convolved with each of these RIRs to obtain 8 independent reverberated recordings. Hence for each reverberant condition, we have used $8 \times 16$ sentences to conduct the experiments. The improvement in speech enhancement task was compared using improvement in three objective measures - PESQ, CD and SRMR [3], [16]. Dereverberated speech has larger PESQ, SRMR scores and lower CD when compared to the reverberated speech. The effectiveness of the dereverberation algorithms is obtained using the relative change in these measures $\Delta PESQ$, $\Delta CD$ and $\Delta SRMR$ when comparing the dereverberated to the reverberated speech.

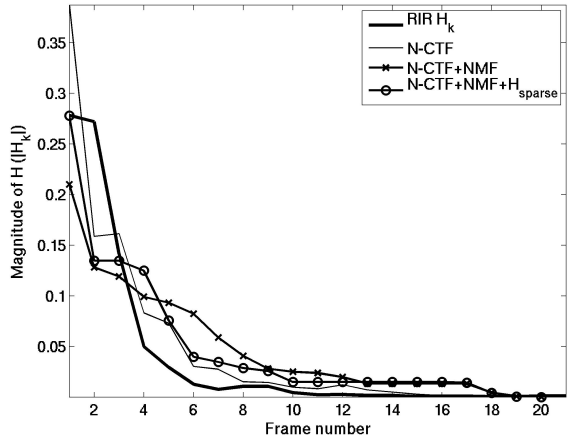


**Fig. 2**. Comparison of normalized estimates of the RIR ($\hat{\boldsymbol{H}}_k$) with the actual RIR ($\boldsymbol{H}_k$) for a specific RIR with $T_{60} = 700$ ms and frequency band ($k = 10$). It can be seen that the proposed regularizer $\boldsymbol{H}_{sparse}$ leads to a better RIR estimate when used with N-CTF+NMF approach.

**Table 1**. SRMR improvements obtained using the proposed methods are compared with existing NMF based approaches for four different RIRs. The RIRs labeled as $RIR_1$, $RIR_2$, $RIR_3$ and $RIR_4$ have ($T_{60}$, source - microphone separation) of $(700ms, 2m)$, $(700ms, 0.5m)$, $(600ms, 2m)$, and $(600ms, 0.5m)$, respectively.

| Methods | $RIR_1$ | $RIR_2$ | $RIR_3$ | $RIR_4$ |
|---|---|---|---|---|
| N-CTF [5] | 1.2000 | 1.3177 | 0.9544 | 1.2804 |
| N-CTF + $\boldsymbol{H}_{sparse}$ | 1.2000 | 1.3177 | 0.9544 | 1.2804 |
| N-CTF + $\boldsymbol{H}_{gain}$ | 1.0080 | 1.1144 | 0.5893 | 1.1456 |
| N-CTF + NMF [5] | 1.7100 | 1.9496 | 0.9798 | 1.3655 |
| N-CTF + NMF + $\boldsymbol{H}_{sparse}$ | 1.8830 | 1.9576 | 1.0389 | 1.3993 |

### 4.2. Experiment setup

The magnitude spectrogram of the reverberant signal was obtained using a 64 ms window with a hop-size of 16 ms. The square root of Hanning window was used in analysis and synthesis. The magnitude spectrogram of the RIR ($\boldsymbol{H}$) was represented using $L_h = 20$ frames, with the first 2 frames (length of $\boldsymbol{H}_{early}$) representing the combination of direct path and early reverberation. Each narrowband $\boldsymbol{H}_k$ was initialized as a linearly decreasing function, and $\boldsymbol{S}$ was initialized using the spectrogram of reverberated speech. For algorithms with speech model, initial values for the basis and the activations were obtained by performing NMF decomposition on the spectrogram of reverberated speech. Since the algorithms converge fast, 20 iterations are performed for each algorithm to obtain the estimates of $\hat{\boldsymbol{S}}$ and $\hat{\boldsymbol{H}}$.

### 4.3. Speech enhancement using the proposed methods

The performance of the proposed algorithms was compared to two dereverberation algorithms, N-CTF and N-CTF with speech model (N-CTF+NMF) [5], for 4 different reverberation conditions. Initially the comparison was based on SRMR improvements and is shown in Table 1. As observed in Fig. 1 and Sec. 3.4 the RIR regularization on N-CTF does not improve the RIR estimates, and hence the SRMR does not improve in these cases (Rows 1, 2, 3 in Table 1). However, similar regularization on N-CTF+NMF resulted in better RIR estimates and leads to SRMR improvements (Rows 4, 5 in Table 1).

**Table 2**. The improvement in objective measures obtained using the proposed RIR regularization are compared with existing NMF based approaches for a RIR with $T_{60} = 700$ ms. The reverberant signal has an average PESQ=1.39, average SRMR=2.13, and average CD=3.85. The best three values are shown in **bold**.

| Methods | $\Delta PESQ$ | $\Delta CD$ | $\Delta SRMR$ |
|---|---|---|---|
| N-CTF [5] | 0.286 | 0.671 | 1.200 |
| N-CTF + $\boldsymbol{H}_{sparse}$ | 0.286 | 0.671 | 1.200 |
| N-CTF + $\boldsymbol{H}_{gain}$ | 0.278 | 0.659 | 1.008 |
| N-CTF + $\boldsymbol{H}_{early}$ | 0.356 | 0.722 | **1.611** |
| N-CTF + $\boldsymbol{H}_{early}$ + $\boldsymbol{H}_{gain}$ | 0.364 | 0.718 | 1.557 |
| N-CTF + NMF [5] | **0.570** | **0.909** | 1.236 |
| N-CTF + NMF + $\boldsymbol{H}_{sparse}$ | **0.527** | **0.730** | **1.710** |
| N-CTF + NMF + $\boldsymbol{H}_{early}$ | **0.525** | **0.914** | **1.883** |

Among the RIRs considered, since the SRMR improvement was significant for $RIR_1$, other objective measures for $RIR_1$ were also considered and this is shown in Table 2. It can be observed from Table 2 that the baseline dereverberation algorithms N-CTF and N-CTF+NMF are able to enhance the reverberated speech. Inducing sparsity (N-CTF+$\boldsymbol{H}_{sparse}$) or frequency envelope on $\boldsymbol{H}$ (N-CTF+$\boldsymbol{H}_{gain}$) does not improve performance. The reason for the sparse $\boldsymbol{H}_{sparse}$ showing no improvement could be that narrow band $\boldsymbol{H}$ roughly has the exponentially decaying structure which we hope to achieve. The multiplicative update for the cost function in (9) is obtained such that it minimizes the error in each narrowband independently. Hence, effectively the cost-function remains the same even after having a frequency envelope on $\boldsymbol{H}$. All the objective measures show improvement for N-CTF+$\boldsymbol{H}_{early}$. Both, N-CTF+NMF+$\boldsymbol{H}_{sparse}$ and N-CTF+NMF+$\boldsymbol{H}_{early}$ show significant improvements in SRMR, and other measures change marginally. One possible explanation can be that the regularization reduces reverberation, but also adds distortion to the dereverberated speech.

## 5. CONCLUSIONS

The estimated RIRs ($\hat{\boldsymbol{H}}$) obtained using the proposed regularizers on $\boldsymbol{H}$ are better approximations to the $\boldsymbol{H}$ when compared to those obtained without any regularization on $\boldsymbol{H}$. But, these improvements were substantial only in the late reverberant part. Speech enhancement results were also obtained using the proposed constraints on the RIR. The addition of sparsity and frequency envelope of the RIR does not change the speech enhancement performance of the algorithm, when compared to the baseline N-CTF model. However, inclusion of the early part of the estimated RIR ($\hat{\boldsymbol{H}}_{early}$) with the estimated speech lead to speech enhancement. The objective measures for an existing N-CTF method that uses a speech model (N-CTF+NMF) also showed improvement when the proposed $\boldsymbol{H}_{early}$ was included. Though there were improvements in speech enhancement using the proposed regularizers, our initial experiments with speech recognition have not shown such clear improvements in the ASR WER. Future work will consider this and other modifications to the N-CTF model for representing reverberant speech.

## 6. REFERENCES


[1] Kenichi Kumatani, John McDonough, and Bhiksha Raj, "Microphone array processing for distant speech recognition," *IEEE Signal Process. Mag.*, pp. 127–140, Nov. 2012.

[2] Patrick A Naylor and Nikolay D Gaubitch, *Speech Dereverberation*, Springer, New York, 2010.

[3] "REVERB 2014," `http://reverb2014.dereverberation.com/workshop/proceedings.html`, Online accessed: 2016-03-23.

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

[5] Nasser Mohammadiha and Simon Doclo, "Speech dereverberation using non-negative convolutive transfer function and spectro-temporal modeling," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 24, no. 2, pp. 276–289, 2016.

[6] Deepak Baby and Hugo Van hamme, "Supervised speech dereverberation in noisy environments using exemplar-based sparse representations," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 156–160.

[7] Nasser Mohammadiha, Peter Smaragdis, and Simon Doclo, "Joint acoustic and spectral modeling for speech dereverberation using non-negative representations," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015.

[8] Heikki Kallasjoki, Jort F. Gemmeke, Kalle J. Palomaki, Amy V. Beeston, and Guy J. Brown, "Recognition of reverberant speech by missing data imputation and NMF feature enhancement," in *Proc. REVERB Workshop*, May 2014.

[9] Kshitiz Kumar, Rita Singh, Bhiksha Raj, and Richard Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011.

[10] Meng Yu, *Multi-channel speech enhancement by regularized optimization*, Ph.D. thesis, University of California, Irvine, 2012.

[11] Meng Yu and Frank K. Soong, "Speech dereverberation by constrained and regularized multi-channel spectral decomposition: evaluated on REVERB challenge," in *Proc. REVERB Workshop*, May 2014.

[12] Seyedmahdad Mirsamadi and John H L Hansen, "Multichannel speech dereverberation based on convolutive nonnegative tensor factorization for ASR applications," in *Proc. of Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[13] Yekutiel Avargel and Israel Cohen, "System identification in the short-time fourier transform domain with crossband filtering," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 15, no. 4, pp. 1305–1319, 2007.

[14] Marc Delcroix, Takuya Yoshioka, Atsunori Ogawa, Yotaro Kubo, Masakiyo Fujimoto, Nobutaka Ito, Keisuke Kinoshita, Miquel Espi, Takaaki Hori, Tomohiro Nakatani, et al., "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the reverb challenge," in *REVERB Workshop*, 2014.

[15] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, David S Pallett, Nancy L Dahlgren, and Victor Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic data consortium, Philadelphia*, vol. 33, 1993.

[16] Tiago H Falk, Chenxi Zheng, and Wai-Yip Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 18, no. 7, pp. 1766–1774, 2010.