SPEECH DEREVERBERATION USING NMF WITH REGULARIZED ROOM IMPULSE RESPONSE

Author(s) Name(s)

Author Affiliation(s)

ABSTRACT

In this paper, different regularizations for the room impulse response (RIR) are imposed to improve speech dereverberation using the non-negative matrix factorization (NMF) framework. The regularization on RIR is motivated by the spectral domain representation of reverberant impulse responses. RIRs in magnitude spectrogram are sparser during the late reverberation period when compared to the early reverberation period. Also, the effectiveness of imposing a structure across frequency for RIR spectrogram and retaining the early part RIR in dereverberation is studied. The NMF multiplicative updates are modified to incorporate this regularization. The performance of the algorithms is studied for both speech enhancement and automatic speech recognition task. Finally, the performance of a two-stage dereverberation algorithm using beamforming and proposed methods are analyzed.

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]. Such instruments could be mobile phones, voice recorders, microphones in a conference room or an automobile. Distant speech refers to scenarios where the audio source and capturing device are separated at least by a feet. 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. Speech dereverberation can be done using single- or multi-channel data depending on the application of interest. The main objective in this research is to address single-channel dereverberation in the distant speech scenario. The performance is evaluated for both speech enhancement and automatic speech recognition task.

The effect of reverberation on speech depends not only on the speech signal, but also on the room or environment under consideration. Reverberation alters speech in a way much different from other types of environmental noise. The environment or the room can be modelled as an impulse response or specifically as room impulse response (RIR). The characteristics of this RIR have a significant effect on the reverberant signal, and hence a good understanding of this is relevant for dereverberation. This is more important when addressing this problem as a blind deconvolution problem.

Dereverberation methods can be classified as those that (i) cancel reverberation or (ii) suppress reverberation. The reverberation cancellation methods include blind deconvolution based methods. The reveberation suppression methods include spectral substraction, linear prediction (LP) based methods, and statistical methods for spectral enhancement. Here we consider a non-negative matrix factorization (NMF) based approach that can be classified as reverberation suppression approach. This uses the magnitude spectrum of the reverberant signal, and with minimal prior knowledge of the RIR, to obtain the dereveberated speech signal. The earliest work to introduce the notion of NMF for dereverberation [4] provides a statistical motivation for use of NMF to solve the dereverberation problem. More recently there have been several improvements over the basic NMF based approache for dereverberation in both single-channel [5, 6, 7, 8, 9], and multi-channel [10], [11], [12] scenarios. In [5, 7] the initial NMD model for speech dereverberation is improved by integrating various NMF models for the speech signal within the original model leading to several NMD-NMF approaches. The work in [6] along with a NMD-NMF approach uses a NMF model for external additive noise and hence does both dereverberation and denoising. Most of these methods use the short-time Fourier transform (STFT) spectrum representation of the signal when performing NMF. The method proposed in [9] uses a gammatone filtered spectrum in the NMF framework and has shown improvements in word error rates (WER) over the earlier NMF based methods. These methods also have proposed incorporating a sparsity constraint on the speech signal as a regularizer to improve speech enhancement. All these approaches have demonstrated improvement in speech enhancement measures and or WER improvement. NMF based methods for dereverberation provide an estimate of the clean speech signal and the corresponding room impulse response (RIR). While the estimation accuracy of the

speech signal is considered in all existing approaches, they do not provide an evaluation of the accuracy of the RIR estimate. The 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 and possibly many solutions. Hence, its required to impose appropriate constraints in obtaining the solution. Earlier NMF approaches obtained reasonable estimates for speech by imposing a sparsity constraint or by using appropriate models for speech. The objective in this work is to consider the RIR estimates and relate them to the expected estimates, within the reverberation model constraints. Such an analysis is used further to provide regularizations or constraints on the RIR, leading to better estimates of both the clean speech signal and RIR of the system. These constraints are motivated from both time-domain and frequency-domain models of the revereberant RIR [2]. The proposed regularization on RIR will be evaluated using speech enhancement measures such as PESQ, SRMR, and CD. The speech enhancement task is evaluated using improvement in WER.

2. NMF BASED SPEECH DEREVERBERATION

The reverberant speech model considered when using NMF for dereverberation and a brief description of existing NMF approaches to solve this problem are described in this section. The goal of NMF algorithms is to factorize a non-negative matrix $\boldsymbol{V} \in \mathbb{R}^{F\times T}$, consisting of elements $v_{i,j} > 0, \forall i,j$, into a product of two non-negative matrices $\boldsymbol{W} \in \mathbb{R}^{F\times R}$ and $\boldsymbol{U} \in \mathbb{R}^{R\times T}$, such that

$$\boldsymbol{V} = \boldsymbol{W}\boldsymbol{U}. \tag{1}$$

Here, the columns $w_i \in \mathbb{R}^{F\times 1}, i \in \{1,\ldots,R\}$ of $\boldsymbol{W}$ can be considered as a set of basis vectors and the rows $u_i \in \mathbb{R}^{1\times T}, i \in \{1,\ldots,R\}$ of $\boldsymbol{U}$ are the weights or activations to obtain the columns $v_i, i \in \{1,\ldots,T\}$ of $\boldsymbol{V}$. Efficient algorithms to solve (1) using multiplicative-updates were initially proposed in [13]. Since then NMF algorithms have been used in several applications. In speech or audio processing, NMF can be applied if the magnitude spectrum of the signal is considered as $\boldsymbol{V}$. Several variants of NMF have been proposed for source separation, denoising, and enhancement of speech and music signals. Here, we will restrict our discussion to the use of NMF for speech enhancement or 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 impulse response (RIR) $h[n]$, can be represented as

$$x[n] = s[n] * h[n] = \sum_{m=0}^{M-1} h[m]s[n-m], \tag{2}$$

where $*$ represents convolution in time-domain, and M is the length of the RIR. In NMF we require a frequency-domain representation of the signal. Hence, we are interested either in the short-time Fourier transform (STFT) representation or any equivalent transform where a representation as in (1) is possible. Here, we will consider the magnitude of the STFT spectrum of the signal and follow a representation motivated in [4] and later followed in [9], [5]. One main aspect of this representation is that it handles the reverberant RIRs which are typically longer (in 100s of ms) compared to the STFT analysis window duration of 20-40 ms considered in speech analysis. In this STFT model, the STFT of the reverberated signal is considered as a convolution of the $s[n]$ and $h[n]$ in each of the sub-bands [14], [4] i.e.,

$$x[k,n] \approx \sum_{m=0}^{L_h-1} h[k,m]s[k,n-m],\ k \in \{0,1,...,K-1\} \tag{3}$$

where $x[k,m]$, $h[k,m]$, $s[k,m]$ are the STFT representation of reverberated signal, RIR, and clean speech signal, respectively, L_h is the length of RIR in the STFT domain, and K is the number of frequency bands in the STFT representation. With this underlying model, it has been observed in [9], [5] that a similar representation can be obtained when using the magnitude of the STFT coefficients. i.e.,

$$|x[k,n]| \approx \sum_{m=0}^{L_h-1} |h[k,m]||s[k,n-m]|, k \in \{0,1,...,K-1\}. \tag{4}$$

For ease of notation, we drop the $|.|$ and use the following equation to denote the model that uses the magnitude of STFT spectrum,

$$X[k,n] \approx \sum_{m=0}^{L_h-1} H[k,m]S[k,n-m], k \in \{0,1,...,K-1\}. \tag{5}$$

To be consistent with [5] we will refer to (5) as the non-negative convolutive transfer function (N-CTF) model for reverberation.

The earliest attempt to solve (5) using the non-negativity of the magnitude squared of the spectrum was in [4]. Given the observed sub-band signal $Y[k,m]$ (denotes $|y(k,m)|^2$),

$$Y[k,n] = X[k,n] + \epsilon[k,n]] \tag{6}$$

where ϵ is considered as the reconstruction error. They obtained a solution for $S[k,n]$ and $H[k,n]$ to minimize the reconstruction error assuming it to be Gaussian white noise. In addition to non-negativity constraints on $S[k,n]$ and $H[k,n]$, they also assumed $\sum_n H[k,n] = 1, \forall k \in \{0,1,...,K\}$ to avoid indeterminacy in the estimates obtained. They proposed multiplicative update rules to obtain S and H, and also related it to the non-negative matrix factor deconvolution (NMFD) problem [15]. The NMFD framework is a variant of the original NMF formulation in (1), to use a convolutive model as

in

$$\boldsymbol{V} \approx \sum_{j=1}^{J} \boldsymbol{W}_j \overset{j\rightarrow}{\boldsymbol{U}}, \tag{7}$$

where $\boldsymbol{W}$ and $\boldsymbol{U}$ are the basis and activation matrices. The $j\rightarrow$ indicates a column-shift by $j-1$ positions to the right. Using this NMFD representation, the reverberation model in (5) can be represented as

$$\boldsymbol{X} \approx \sum_{m} \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}, \tag{8}$$

where

$$\boldsymbol{S} = \left[\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_T\right] \tag{9}$$

with $\boldsymbol{s}_i \in \mathbb{R}^{K\times 1}, i \in \{1,\ldots,T\}$ and

$$\boldsymbol{H}_m = \begin{pmatrix} H[1,m] & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & H[K,m] \end{pmatrix}. \tag{10}$$

a diagonal matrix. As observed in [4], this representation for the reverberant signal comprises of components of clean speech spectrum are being blurred by the temporal evolution of the $\boldsymbol{H}$ components. With this the $\boldsymbol{H}$ corresponds to a basis matrix constrained to be diagonal and the clean speech spectrum corresponds to the activation matrix. The solution proposed in [16] corresponds to a NMF setting where the error to be minimized is the Kullback - Leibler (KL) divergence between $\boldsymbol{Y}$ and $\boldsymbol{X}$. The general form of the cost-function to be minimized can be represented as

$$J(\boldsymbol{S},\boldsymbol{H}) = D_{KL}(\boldsymbol{Y},\boldsymbol{X}), \tag{11}$$

$$= D_{KL}(\boldsymbol{Y}, \sum_{m} \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}) \tag{12}$$

$$\text{s.t.} \sum_{m} H[k,m] = 1,\ k \in \{1,\ldots,K-1\}, \boldsymbol{H}_m \geq 0, \boldsymbol{S} \geq 0. \tag{13}$$

where $\boldsymbol{X}$ is from (7) and $D_{KL}(Y \parallel X)$ is defined as follows.

$$D_{KL}(Y \parallel X) = \sum_{k,m} (Y[k,m] ln(\frac{X[k,m]}{Y[k,m]}) - X[k,m] + Y[k,m]) \tag{14}$$

The cost functionin (13) is modified to include sparsity constraint on $\boldsymbol{S}$, which leads to better estimates. Improving on [4], the work in [9] observed that using a Gammatone filtered magnitude spectrum provides improved estimates compared to that of using just the Fourier transform. This work also demonstrated the effectiveness of using the magnitude of the spectrum as opposed to that of magnitude squared spectrum suggested in [4]. The effectiveness of the NMF algorithm on speech recognition was also demonstrated here by considering WER measures. In the more recent work [5], the NMFD reverberation model was further improved by incorporating a NMF model for the speech spectrum. The NMF model for the speech signal has been used in several applications. In matrix notation, the NMF factorization of the clean speech signal spectrogram can be denoted as

$$\boldsymbol{S} \approx \boldsymbol{W}\boldsymbol{X} \tag{15}$$

where $\boldsymbol{W} \in \mathbb{R}^{F\times R}$ represents a matrix of basis vectors for clean speech signal and $\boldsymbol{X} \in \mathbb{R}^{R\times T}$ is a matrix of activations. Given the speech spectrogram, the NMF multiplicative updates can be used to obtain the basis vectors $\boldsymbol{W}$ and activations $\boldsymbol{X}$. In [5] such a representation is integrated into the NMFD model for revereberation, and the corresponding updated cost-function is

$$J(\boldsymbol{W},\boldsymbol{X},\boldsymbol{H}) = D_{KL}\left(\boldsymbol{Y},\boldsymbol{X}\right), \tag{16}$$

$$= D_{KL}\left(\boldsymbol{Y}, \sum_{m} \boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}\right) \tag{17}$$

$$= D_{KL}\left(\boldsymbol{Y}, \sum_{m} \boldsymbol{H}_m \overset{m\rightarrow}{(\boldsymbol{W}\boldsymbol{X})}\right) + \parallel \boldsymbol{X} \parallel_1 \tag{18}$$

$$\text{s.t.} \sum_{m} H[k,m] = 1, \boldsymbol{H}_m \geq 0, \boldsymbol{W} \geq 0, \boldsymbol{X} \geq 0. \tag{19}$$

where $\parallel \cdot \parallel_1$ denotes $l1$-norm and promotes sparsity. They also suggested another weighted method that combined the NMFD model for dereverberation along with the NMF speech model. However, their experiments and results suggest that the integrated model in (19) does performs better and hence not discussed here. The results in [5] indicate improved speech enhancement measures and do not provide any speech recognition results. They proposed three possible NMF models for the speech signal which are either unsupervised or semisupervised. 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. Here they considered two approaches (i) a low-rank NMF model where $\boldsymbol{W}$ was obtained from clean speech training data denoted as N-CTF+NMF+LR (ii) an overcomplete NMF model where the $\boldsymbol{W}$ was obtained using a random walk model to select basis vectors from those obtained using training data denoted as N-CTF+NMF+OC.

In another recent work [6], the NMF based approaches in [5], [4] have been extended to do both dereverberation and denoising in a supervised setting. They introduce a NMF model to handle other background noises, which does not include reverberation, and NMF model for the speech signal. They learn basis matrices for both speech signal and noise signal from training data. They demonstrate improvement in speech enhancement measures using NMF models for both the speech signal and noise signal. Based on the speech enhancement results, they also conclude that the N-CTF+NMF

in a supervised context provides a better estimate of 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

As discussed in equation (3), the basic underlying assumption in the NMF based approaches to solve the dereverberation problem is that the reverberant signal spectrum in a single frequency band can be considered as a convolution of the clean speech spectrogram and $\boldsymbol{H}$ for that specific band.

It should be noted that the magnitude spectrogram $\boldsymbol{H}$ in (3) does not correspond to the STFT of the RIR $h[n]$, but it is an approximation assuming cross-band effects can be neglected in the reverberant signal spectrum [5], [14]. However, it is still valid to treat (3) and operate in the subband domain to perform dereverberation. 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 estimates $\boldsymbol{H}$ to the STFT of an approximate RIR $\tilde{\boldsymbol{H}}$ obtained from the $\boldsymbol{H}_{\text{true}}$. To illustrate this, we have considered a reverberant RIR of duration 1 s for a room with $T_{60} = 700$ ms. Using a sampling frequency of 16 kHz, a synthesis window of length $N = 1024$, type square root of Hamming, overlap of 75%, and corresponding analysis window, we obtained the true STFT $\boldsymbol{H}_{\text{true}}$. Then using the approach suggested in [5] with similar STFT settings, we obtained $\tilde{\boldsymbol{H}}$. In Fig. 1, we compare these STFTs for a specific frequency band k. It can be seen that they are reasonably similar, though do not match exactly. Hence, we can compare the estimated $\boldsymbol{H}$ with $\tilde{\boldsymbol{H}}$, to evaluate the accuracy of RIR estimation from the revereberated speech signal.

We compare the estimated $\boldsymbol{H}$ obtained using the reference approaches (N-CTF and N-CTF+NMF) to the $\tilde{\boldsymbol{H}}$ for a RIR of T_{60} approximately 700 ms. These narrow band estimate of RIR obtained by reference methods with the actual value is plotted in Fig. 2. 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 approach so that the estimated $\boldsymbol{H}$ 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 the $\boldsymbol{H}$ for

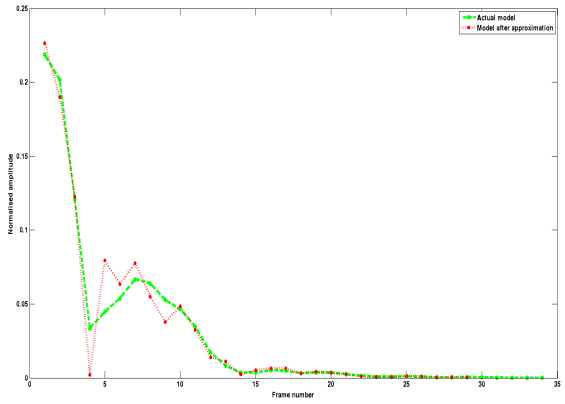

Fig. 1. Comparison of estimate of H with the actual H for k fixed at 50

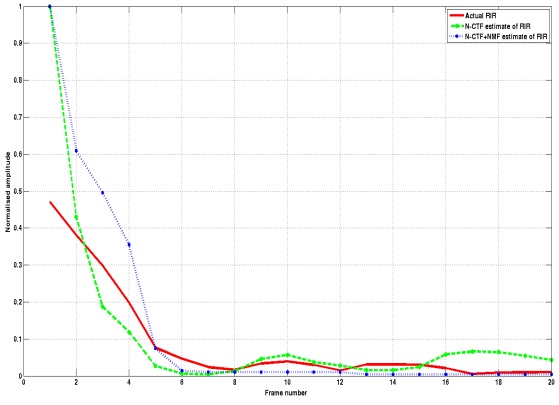

Fig. 2. Comparison of estimate of H with the actual H for k fixed at 50

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

3.1. Sparsity of RIR

Characterising RIRs of a reverberant room is challenging, as the RIRs depends on the room characteristics and distance between the source and the microphone acquiring the signal. However, the time-domain model by Pollack is a reasonable characterisation based on the T_{60} of a room model. Similarly, the frequency-domain characterisation by Schroder is also a good model that has been used extensively [2]. It does not provide a characterisation of the frequency-response but provides a characterisation of the possible distribution of power with respect to frequency. In time-domain the RIR is exponentially decaying with the decay factor dependent on the T_{60}. Correspondingly the frequency domain representation of the RIR also has larger magnitude values during the early part of the reverberation and the magnitude spectrum dies down to smaller values during the late part. Most of the entries in H has a value very close to zero. Hence the simplest constraint is to assume that the $\boldsymbol{H}$ has a sparse structure. This is different from assuming sparsity on the speech signal. Assuming sparsity on $\boldsymbol{H}$ is similar to assuming a sparse basis vector matrix in the standard NMF domain. We incorporate this into the basic NMD framework and have the modified cost-function,

$$J(\boldsymbol{W},\boldsymbol{X},\boldsymbol{H}) = D_{KL}\left(\boldsymbol{Y},\boldsymbol{X}\right), \tag{20}$$

$$= D_{LS}\left(\boldsymbol{Y},\sum_{m}\boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}\right) + \lambda \parallel \boldsymbol{H} \parallel_1 \tag{21}$$

$$\text{s.t.}\boldsymbol{H}_m \geq 0, \boldsymbol{S} \geq 0. \tag{22}$$

λ decides the weight given to sparsity of H. In the work, $\lambda = 1$. This method is referred to as N-CTF+Sparse H.

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 in the RIRs can be modelled as a function of the frequency $Gain[k]$. Such models can be obtained by fitting polynomial functions on existing recorded RIRs and their STFTs. This can be included in the NMD framework as opposed to constraining the sub-band sums of $\boldsymbol{H}$ to be unity. The corresponding cost function

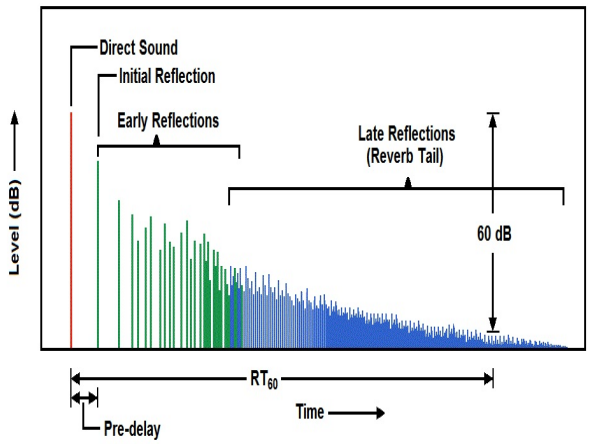

Fig. 3. Time domain representation of a typical RIR [18]

is,

$$J(\boldsymbol{S},\boldsymbol{H}) = D_{KL}(\boldsymbol{Y},\boldsymbol{X}), \tag{23}$$

$$= D_{KL}\left(\boldsymbol{Y},\sum_{m}\boldsymbol{H}_m \overset{m\rightarrow}{\boldsymbol{S}}\right) \tag{24}$$

$$\text{s.t.}\sum_{m} H[k,m] = Gain[k], k \in \{0,1,...,K-1\} \tag{25}$$

$$\boldsymbol{H}_m \geq 0, \boldsymbol{S} \geq 0 \tag{26}$$

where the $Gain[k]$ is to be obtained from existing RIRs or RIR models. This model is referred to as N-CTF+Gain H.

3.3. Inclusion of early part of RIR

RIR can be broadly divided into two regions - early reverberation and reverberation tale (late reverberation). The early part accounts for the reflections which come up to 50ms after the direct path. The reflections which come after 50 ms forms the reverberation tale. A sample RIR is shown in the fig. 3. 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. However, retaining the early part improves speech intelligibility and ASR performance [2][17]. Once the clean speech spectrum (S) is estimated, the early part of RIR (H_{ear}) is attached to it using the following equation to obtain an improved dereverberated spectrum (S_{new}).

$$S_{new}(n,k) = S(n,k) * H_{early}(n,k), k \in \{0,1,...,K\} \tag{27}$$

This method is referred to as N-CTF+H_{ear}. The algorithm using both N-CTF+H_{ear} and N-CTF+Gain H is refereed to as N-CTF+H_{ear}+Gain H.

4. RESULTS

The performance of the algorithms discussed in section 3 are evaluated for both speech enhancement and automatic speech recognition tasks (ASR).

4.1. Database

Two databases (TIMIT and TIDIGITS) have been used for the experiments. TIMIT contains two sets of data (train and test). Train set contains 10 sentences spoken by 377 different speakers and test set contains 10 sentences spoken by 168 distinct speakers. There are no common speakers in test and train sets. Out of the 10 sentences, 2 Sentences is common across all speakers. The performance in speech enhancement task is compared using a subset of TIMIT database. A set of 16 different sentences spoken by 16 distinct speakers is used. TIDIGITS contains 112 distinct adult speakers (55 male and 57 female) in train set and 113 different adult speakers (56 male and 57 female) in the test set. Each speaker has 77 utterances. The database contains 11 distinct words (zero to nine and oh). The ASR performance is evaluated using adult data set of TIDIGITS.

The word error rate (WER) of TIMIT database is relatively high ($\approx 28\%$) for clean speech. Since the best achievable WER is itself high, a reliable WER improvement of various dereverberation algorithms cannot be analyzed using the database. The relative improvement in objective measures for the reference method using TIMIT database is available in [16]. To verify the implementation of reference method, TIMIT database is used for speech enhancement task. TIDIGITS is a limited vocabulary database with very low WER ($\approx 1.5\%$). So TIDIGITS is used for comparing the performance of various algorithms for ASR task.

4.2. Evaluation measures

The improvement in speech enhancement task is compared using improvement in three objective measures - perceptual evaluation of speech quality (PESQ), cepstral distance (CD) and speech-to-reverberation modulation energy ratio (SRMR) [19] [3]. Dereverberated speech increases PESQ and SRMR score while CD tend to decrease as compared to reverberated speech. The effectiveness of the dereverberation algorithms is compared using the relative change in the objective measures of dereverberated speech ($\Delta PESQ$, ΔCD and $\Delta SRMR$) as compared with the reverberated speech.

Speech recognition performance is compared using the improvement in word error rate (WER) as compared to the reverberated speech. Speech recognition is performed using Kaldi toolkit trained on TIDIGITS. The speech recognition system has three sources of error - insertion, deletion, and substitution. For this task, the WER is defined as the ratio of the total number of misclassified phones to the total number of phones in the test set. Mathematically,

$$WER = \frac{INS + DEL + SUB}{Total\ number\ of\ phones\ in\ test\ set} * 100\% \tag{28}$$

where INS, DEL, SUB represents the total number of insertion, deletion and substitution errors while decoding in the test set respectively.

4.3. Reference methods

Two dereverberation algorithms NCTF and NCTF with speech model [16] are used as reference methods for comparison. In the reference paper, the objective measures PESQ, and CD were used for verifying the performance of the algorithm. The algorithm was tested on reverberated data generated using 16 different TIMIT sentences which are spoken by 16 distinct speakers. Reverberation is simulated using a measured RIR having a T_{60} of around 680 ms and DRR of 0 dB was used. The NCTF model showed an improvement of about 0.3 in PESQ score and 0.5 in CD as compared to reverberated speech. The NCTF with speech model showed an improvement of 0.4 in PESQ score and 0.4 CD as compared to reverberated speech. The implementation of this method is verified by trying to reproduce the results in [16]. Although the exact RIR and set of sentences used in the experiments are unknown, we chose an RIR which has a T_{60} and DRR of roughly 700 ms and 0 dB. Also, a subset of 16 sentences spoken by different speakers from TIMIT database was used for performance evaluation. A PESQ score improvement of 0.46 and CD improvement of 0.08 is obtained for N-CTF method. Also, for N-CTF with speech model, a PESQ and CD improvement of 0.65 and 0.1 were obtained. The improvement obtained in PESQ score is comparable with the score obtained in [16]. However, improvements in CD was not comparable with the reference paper. The difference in obtained objective measure scores with the reference paper could be because of the different RIR and set of sentences used in the experiments. The implemented of algorithms are tested for other RIRs. Table 1 shows the objective measure improvement for a different set of RIRs. Both PESQ and CD improves significantly improved for the reference methods.

4.4. Experiment setup

The magnitude spectrogram is obtained using a 64 ms window with a hop size of 16 ms. The square root of Hanning window is used in analysis and synthesis side. The magnitude spectrogram of RIR (H) is represented using 20 frames. Out of which the first 2 frames is related to the combination of direct path and early reverberation. H for each narrowband is initialized as a linearly decreasing function. S is initialized as the spectrogram of reverberated speech. For models with speech model, initial values for the basis and the activations are 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 S and H.

Measured RIRs available from REVERB challenge [3] is used for the evaluation. The RIRs are measured using an 8 channel circular array of diameter 10 cm. The RIRs of three rooms (small, medium and large) with T_{60} of 250 ms, 500 ms and 700 ms are available. The recordings are done for a microphone array placed 0.5 m and 2 m away from the source. The particular RIRs used for the experiment were measured in a room having a T_{60} of 700 ms with the microphone array placed 2 m distant from the source. A similar trend in the dereverberation performance was observed for other RIRs. The dereverberation algorithms are performed by treating the 8 channel reverberated data as 8 independent recordings.

4.5. Speech enhancement for the proposed methods

The performance of proposed dereverberation algorithms is compared using different objective measures. Since the microphones are placed very close to each other, the relative improvement in objective measures obtained across 8 channels is very close to each other. The net enhancement in the performance of different algorithms is obtained as an average of improvement in the objective measures for 16 sentences across 8 channels. The average objective measure improvement for various algorithms is shown in the Table 1.

From Table 1, it is observed that the baseline dereverberation algorithms N-CTF and N-CTF+NMF are able to enhance the reverberated speech. Inducing sparsity (N-CTF+Sparse H) and frequency envelope on H (N-CTF+Gain H) does not improve the performance. The reason for the sparse H showing no improvement could be that narrow band H roughly has the exponentially decaying structure which we hope to achieve. The multiplicative update for the cost function in (26) is obtained such that it minimizes the error in each narrowband independently. So the in effect cost function remains the same even after putting frequency envelope in H. All the objective measures show improvement for N-CNMF+H_{early}. However, for N-CNMF+NMF+H_{early}, SRMR shows a significant improvement while other measures remain unchanged. The possible explanation can be that the algorithm reduces reverberation but also adds distortion to the dereverberated speech.

Table 1. *The improvement in objective measures obtained using the proposed regularization to the RIR are compared with existing NMF based approaches.*

Methods	$\Delta PESQ$	ΔCD	$\Delta SRMR$
N-CTF	0.286	0.671	1.200
N-CTF + $\boldsymbol{H_{early}}$	0.356	0.722	1.611
N-CTF + Gain $\boldsymbol{H}$	0.278	0.659	1.008
N-CTF + $\boldsymbol{H_{early}}$ + Gain $\boldsymbol{H}$	0.364	0.718	1.557
N-CTF + Sparse $\boldsymbol{H}$	0.286	0.671	1.200
N-CTF + NMF speech	0.570	0.909	1.236
N-CTF + NMF speech + $\boldsymbol{H_{early}}$	0.525	0.914	1.883

The N-CTF model shows significant improvement in WER. N-CTF+Sparse H and N-CTF+Gain H methods show no improved the performance. N-CTF+H_{early} shows marginal improvement. Methods including speech model (N-CTF+NMF and N-CTF+NMF+H_{early}) also show marginal improvement over N-CTF even though the objective measures shown significant improvement. The reason could be that dereverberation using speech model induced artifacts which do not affect the enhancement but affected the ASR performance. This can be observed in the estimate of H which shows distortions whereas for the N-CTF, the H estimate is smooth. N-CTF+NMF+H_{early} has no significant improvement in ASR results as compared to N-CTF+NMF.

Table 2. *The improvement obtained for ASR task using the proposed regularization to the RIR are compared with existing NMF based approaches.*

Methods	$WER(\%)$	$\Delta WER(\%)$
Clean speech	1.76	–
Reverberated speech	29.03	–
N-CTF	19.97	9.06
N-CTF + $\boldsymbol{H_{early}}$	19.26	9.77
N-CTF + Gain $\boldsymbol{H}$	20.06	8.97
N-CTF + $\boldsymbol{H_{early}}$ + Gain $\boldsymbol{H}$	19.61	9.42
N-CTF + Sparse $\boldsymbol{H}$	19.50	9.53
N-CTF + NMF speech	19.24	9.79
N-CTF + NMF speech + $\boldsymbol{H_{early}}$	19.21	9.82

4.6. ASR improvement for the proposed methods

The speech recognition system using Kaldi toolkit was trained using a Gaussian mixture model (GMM) - hidden Markov model (HMM) system. The system was trained using training data set of TIDIGITS. It contains 56 male and 58 female speakers speaking 77 utterances each. The GMM is an 8 Gaussian mixture. The HMM is a monophone with 1000 tied states. The ASR results obtained for an RIR with T_{60} of 700 ms and 2 m away from the source is shown in Table 2.

4.7. ASR improvement on beamformed output

This section discusses the ASR performance a 2 stage dereverberation system. In the first stage, beamforming is performed using BeamformIt toolkit [20] to obtain an enhanced single channel data obtained from 8 channel data. The second stage is a single channel dereverberation methods using the proposed methods. The second stage reduces the residual reverberation present after beamforming. The reverberation conditions and training set are same as in earlier experiments,

but testing is done in a reduced set of 6 speakers (3 males and 3 females).

Table 3 shows the ASR performance of dereverberation using beamforming and proposed methods. The reference method N-CTF shows improvement in WER as compared with the beamformed output. This shows that the algorithm is able to suppress the residual reverberation present in the channel after beamforming. Other proposed algorithms failed to improve the WER any further. Table 4 compares the performance of proposed methods for the beamformed output in the presence of reverberation and noise. The algorithms perform poorly. The reason could be that the algorithm is designed to model reverberation alone. So in the presence of residual noise after beamforming, the algorithm fails to produce a good estimate of clean speech.

Table 3. *The improvement in reverb speech ASR obtained using the beamforing followed by proposed methods.*

Methods	$WER(\%)$	$\Delta WER(\%)$
Clean speech	1.32	–
Reverberated speech	27.15	–
Beamforming	22.2	4.96
N-CTF	16.34	10.82
N-CTF + $\boldsymbol{H_{early}}$	16.47	10.69
N-CTF + NMF speech	16.6	10.56
N-CTF + NMF speech + $\boldsymbol{H_{early}}$	17.33	9.83

Table 4. *The improvement in reverb+noise(10dB) speech ASR obtained using the beamfoming followed by proposed methods.*

Methods	$WER(\%)$	$\Delta WER(\%)$
Clean speech	1.32	–
Reverberated speech	56.19	–
Beamforming	41.63	17.88
N-CTF	77.93	−18.42
N-CTF + $\boldsymbol{H_{early}}$	77.87	−18.36
N-CTF + NMF speech	73.06	−13.55
N-CTF + NMF speech + $\boldsymbol{H_{early}}$	53.69	5.82

5. CONCLUSIONS AND FUTURE WORK

The speech enhancement and ASR performance of speech dereverberation algorithms for various constraints on RIR are studied. The addition of sparsity and frequency envelope of RIR does not change the speech enhancement performance of the algorithm as compared with the baseline N-CTF model. But the inclusion of early part of RIR (H_{early}) with the estimate of clean speech improved the objective measures. The objective measures for reference method with speech model (N-CTF+NMF) also showed improvement when H_{early} was included in the model.

The dereverberation using C-NMF reduced significantly the ASR performance. The inclusion of sparsity and frequency envelope did not change the WER. $N-CTF+H_{early}$ method shows a marginal improvement in ASR performance. The inclusion of speech model ($N - CTF + NMF$) also did not show significant improvement in ASR performance. However, the algorithms were able to residual reverberation present in a beamformed output, but the algorithm fails to improve the performance in the presence of background noise.

The dereverberation algorithms are not able to completely remove the effects of reverberation. A better model for reverberation and speech could improve the performance of dereverberation algorithm. Replacing NMF model for the clean speech by a convolutive NMF model could be one such approach. The use of multichannel data helps in dereverberation [2]. The proposed methods can be modified to handle multichannel data. Also, other multichannel dereverberation algorithms like beamforming have to be studied. The effects of noise in distance speech recording are unavoidable. The performance of dereverberation algorithms is affected by the presence of noise. The proposed methods perform poorly in the presence of noise as discussed in section. So there is a need for modifying the proposed models to include the effects of noise. Other methods which perform dereverberation and denoising together should also need to be studied.

6. REFERENCES

[1] Kenichi Kumatani, John McDonough, and Bhiksha Raj, "Microphone array processing for distant speech recognition," *IEEE Signal Processing Magazine*, 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 *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 Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 276–289, 2016.

[6] Deepak Baby et al., "Supervised speech dereverberation in noisy environments using exemplar-based sparse representations," in *2016 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 *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 *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 *Fifteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014.

[13] Daniel D. Lee and H. Sebastian Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

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

[15] Paris Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation: Fifth International Conference*. Sept. 2004, vol. 3195, pp. 494+, Springer-Verlag GmbH.

[16] Nasser Mohammadiha and Simon 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.

[17] 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.

[18] "Reverberation," `http://http://lossenderosstudio.com/newsletter.php?issue=66`, Online accessed: 2016-08-16.

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

[20] Xavier Anguera, Chuck Wooters, and Javier Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.