SUPERVISED SPEECH DEREVERBERATION IN NOISY ENVIRONMENTS USING EXEMPLAR-BASED SPARSE REPRESENTATIONS

Deepak Baby and Hugo Van hamme

Department ESAT, KU Leuven, Belgium.

{Deepak.Baby, Hugo.Vanhamme}@esat.kuleuven.be

ABSTRACT

Exemplar-based techniques, where the noisy speech is decomposed as a linear combination of the speech and noise exemplars stored in a dictionary, have been successfully used for speech enhancement in noisy environments. This paper extends this technique to achieve speech dereverberation in noisy environments by means of a non-negative approximation of the noisy reverberant speech in the frequency domain. A novel approach for estimating the room impulse response (RIR) together with the speech and noise estimates using a non-negative matrix deconvolution (NMD) -based technique is proposed. In addition, we extend an existing technique based on non-negative matrix factorisation (NMF) that performs speech dereverberation in noise-free environments to noisy scenarios. New estimators for jointly obtaining the RIR and exemplar weights for the NMD and NMF -based formulations are presented. The proposed techniques are evaluated on the noise-free and noisy reverberant speech in the CHiME-2 WSJ0 database and are shown to yield better speech enhancement in terms of signal-to-distortion ratio (SDR), perceptual evaluation of speech quality (PESQ) and cepstral distance (CD) measures.

Index Terms— speech dereverberation, non-negative matrix deconvolution, non-negative matrix factorisation

1. INTRODUCTION

Speech recordings obtained using a distant microphone in a noisy enclosed space often have reduced intelligibility due to additive noise and room reverberation. Therefore it is desirable to have a mechanism for noise suppression and dereverberation in many applications such as hearing aids and automatic speech recognition. Most of the traditional systems first make use of a source separation or denoising technique followed by a dereverberation step. In this paper, we concentrate on a system that can jointly obtain speech denoising and dereverberation on single channel data.

There exist a few unsupervised techniques that consider simultaneous denoising and dereverberation. For example, a two-stage method is proposed in [1] does channel identification followed by signal estimation, which requires prior knowledge about single-talk periods for channel identification. The TRINICON technique proposed in [2,3] also performs joint denoising and dereverberation using the higher order statistics of speech. Another work presented in [4] aims at achieving a similar task in a tandem manner.

In this paper, we propose a supervised speech enhancement technique operating on the magnitude spectrogram domain that can jointly obtain speech denoising and dereverberation using exemplars of speech and noise. The proposed model assumes that the magnitudes of the short-time Fourier transform (STFT) of the noisy reverberant speech at every frequency bin can be approximated as a sum of magnitude STFT of the additive noise and a convolution of the magnitude STFT of clean speech signal with that of the room impulse response (RIR) in that frequency bin. Such an approximation based on the non-negative transfer function has been successfully used in noise-free scenarios for speech dereverberation [5–10].

The main contribution of this paper is to propose a speech denoising model using non-negative matrix deconvolution (NMD) [11] -based technique to separate speech and noise that is optimised jointly with a non-negative RIR model in the magnitude STFT domain for dereverberation. We make use of speech and noise exemplars that are stored in speech and noise dictionaries to decompose the noisy speech, and use the speech estimate to estimate the RIR. In addition, we extend a technique proposed in [6] that uses non-negative matrix factorisation (NMF)-based approximation for noise-free reverberant speech to noisy cases as well. A similar technique is also explored in [12] which also uses NMF-based formulation for dereverberation where the estimate of the RIR is based on both speech and noise estimates. However, we argue that the RIR estimate estimated from both speech and noise is unreliable when we have multiple and/or moving noise sources, and we reformulate the problem such that the RIR is estimated only based on the speech estimate.

The proposed approaches are evaluated on the CHiME-2 database which contains the speech data added with room reverberation and multi-source noises. In addition, we also evaluate on the noise-free reverberant data to identify which formulation is better in such scenarios. The experimental results show that the proposed techniques yield better speech enhancement in terms of various measures over the traditional NMD and NMF-based techniques that do not have a reverberation model.

2. NON-NEGATIVE REPRESENTATION OF REVERBERANT SPEECH

This section details the non-negative formulation of reverberation in the magnitude short-time Fourier transform (STFT) domain. Let $y[n]$ and $h[n]$ denote the clean speech signal and room impulse response (RIR) of length L_t, respectively. The resulting reverberant signal is obtained by convolving the speech signal with the RIR, i.e., $z[n] = y[n] * h[n] = \sum_m h[m]y[n-m]$. In the STFT domain, for the f-th frequency bin at frame t, this can be approximated as [5,6]:

$$\mathcal{Z}(f,t) \approx \sum_{p=1}^{L} \mathcal{H}(f,p)\mathcal{Y}(f,t-p+1) \quad (1)$$

This work has been funded with support from the European Commission under Contract FP7-PEOPLE-2011-290000 (INSPIRE).

where $\mathcal{Z}$, $\mathcal{H}$ and $\mathcal{Y}$ denote the complex-valued STFT of $z[n]$, $h[n]$ and $y[n]$, respectively. L denote the length of the RIR in the STFT space. Let the STFT be obtained for $2K$ frequency bins and $\mathcal{Z}$ contains F frames.

For the non-negative formulation the magnitude STFT of the reverberant signal is considered, which is approximated as $\mathbf{Z}(f,t) \approx \sum_{p=1}^{L} \mathbf{H}(f,p)\mathbf{Y}(f,t-p+1)$, where $\mathbf{Z} = |\mathcal{Z}|$, $\mathbf{H} = |\mathcal{H}|$ and $\mathbf{Y} = |\mathcal{Y}|$. Such an approximation has been successfully used for speech dereverberation in [6,9].

The approximation can be expressed as a matrix operation as

$$\mathbf{Z} \approx \sum_{p=1}^{L} [\mathbf{H}]_p \circledcirc \overset{(p-1)\rightarrow}{\mathbf{Y}} \tag{2}$$

where $[\mathbf{H}]_p$ is the p-th column of the matrix $\mathbf{H}$ and $\overset{p\rightarrow}{\mathbf{Y}}$ denotes the right-shifting operation by adding p columns of zeros to the left and removing the last p columns of $\mathbf{Y}$. The operation $\mathbf{h} \circledcirc \mathbf{Y}$ stands for the element-wise multiplication of a vector $\mathbf{h}$ with the all the columns of $\mathbf{Y}$.

3. METHODOLOGY

This paper aims at speech dereverberation in noisy environments where the reverberant speech is corrupted with additive noise $w[n]$. In this work, we assume an additive model for the noisy reverberant speech as:

$$\mathbf{Z} \approx \tilde{\mathbf{Z}} = \sum_{p=1}^{L} [\mathbf{H}]_p \circledcirc \overset{(p-1)\rightarrow}{\mathbf{Y}} \; + \; \mathbf{W} \tag{3}$$

where, $\mathbf{W}$ is the magnitude STFT of $w[n]$. We also assume that the reverberation is to be modelled only with the speech signal. Such an assumption is also beneficial to obtain a better and reliable estimate of $\mathbf{H}$ since we assume a fixed RIR between the speaker and the microphone, whereas such assumptions are often invalid for a real background noise source.

The goal of this work is thus to estimate $\mathbf{H}$, $\mathbf{Y}$ and $\mathbf{W}$ from the magnitude STFT of the noisy reverberant speech signal $\mathbf{Z}$. We use an exemplar-based technique to decompose $\mathbf{Z}$ as the sum of reverberant speech and noise estimates. Only the positive half of the magnitude STFT is used resulting in a $\mathbf{Z}$ of size $K \times F$. Exemplar-based techniques make use of speech and noise dictionaries $\mathbf{S}$ and $\mathbf{N}$ containing J_s clean speech and J_n noise exemplars randomly sampled from the training data, respectively. To model the temporal continuity of speech, exemplars that span T frames are considered. Thus the speech and noise dictionaries are of size $K \cdot T \times J_s$ and $K \cdot T \times J_n$, respectively.

Notice that the magnitude STFT exemplars are also non-negative. In this work, we make use of two popular exemplar-based decomposition schemes: the non-negative matrix deconvolution (NMD) and non-negative matrix factorisation (NMF), which are detailed below.

3.1. Compositional model using NMD

Here, we approximate the frame level speech and noise spectra using the NMD-based model [11],

$$\mathbf{Y} \approx \tilde{\mathbf{Y}} = \sum_{t=1}^{T} \mathbf{S_t} \overset{(t-1)\rightarrow}{\mathbf{X_s}} \quad \text{and} \quad \mathbf{W} \approx \sum_{t=1}^{T} \mathbf{N_t} \overset{(t-1)\rightarrow}{\mathbf{X_n}}. \tag{4}$$

The matrix $\mathbf{S_t}$ denotes the t-th block matrix obtained by partitioning $\mathbf{S}$ into T block rows each of size $K \times J_s$ [11]. $\mathbf{N_t}$ is also defined in the same manner from $\mathbf{N}$. The approximation is obtained such that mixing weights or activations $\mathbf{X_s}$ and $\mathbf{X_n}$ are also non-negative. This paper proposes a compositional model for noisy reverberant speech as:

$$\tilde{\mathbf{Z}} = \sum_{p=1}^{L}\sum_{t=1}^{T} [\mathbf{H}]_p \circledcirc \mathbf{S_t} \overset{\tau\rightarrow}{\mathbf{X_s}} \; + \; \sum_{t=1}^{T} \mathbf{N_t} \overset{(t-1)\rightarrow}{\mathbf{X_n}} \tag{5}$$

using (3) and (4), where $\tau = p + t - 2$. The problem thus boils down to estimating $\mathbf{H}$ and the activations, which are estimated so as to minimize the Kullback-Leibler divergence between $\mathbf{Z}$ and $\tilde{\mathbf{Z}}$ which is defined as:

$$D_{KLD}\left(z \| \tilde{z}\right) = z \, \log \frac{z}{\tilde{z}} + \tilde{z} - z. \tag{6}$$

In addition, we also add sparsity constraints on the activations to obtain a reliable approximation of speech and noise spectra with randomly sampled exemplars. The resulting cost function is,

$$\mathcal{C} = D_{KLD}\left(\mathbf{Z} \| \tilde{\mathbf{Z}}\right) + \lambda_s \cdot \mathbf{X_s} + \lambda_n \cdot \mathbf{X_n}. \tag{7}$$

λ_s and λ_n penalise dense speech and noise activations, respectively. To obtain $\mathbf{H}$ and the activations that minimize (7), we make use of an iterative gradient-descend technique using multiplicative updates given by

$$\mathbf{H} \leftarrow \mathbf{H} \odot \frac{\bigtriangledown^{-}_{\mathbf{H}} \mathcal{C}}{\bigtriangledown^{+}_{\mathbf{H}} \mathcal{C}} \tag{8}$$

where, $\bigtriangledown^{-}_{\mathbf{H}} \mathcal{C}$ and $\bigtriangledown^{+}_{\mathbf{H}} \mathcal{C}$ are the positive and the negative parts of the derivative $\partial\mathcal{C}/\partial\mathbf{H}$. To obtain the required derivatives, we apply the chain rule,

$$\frac{\partial \mathcal{C}}{\partial \mathbf{H}} = \underbrace{- \frac{\mathbf{Z}}{\tilde{\mathbf{Z}}} \frac{\partial \tilde{\mathbf{Z}}}{\partial \mathbf{H}}}_{\bigtriangledown^{-}_{\mathbf{H}}} \; + \; \underbrace{\frac{\partial \tilde{\mathbf{Z}}}{\partial \mathbf{H}}}_{\bigtriangledown^{+}_{\mathbf{H}}} \tag{9}$$

The ratio $\mathbf{Z}/\tilde{\mathbf{Z}}$ is done element-wise and let it be denoted as $\mathbf{R}$. The updates for the activations are also obtained in the same manner. The multiplicative updates for all the unknowns can be obtained using (9), (7) and (5) as given below ($\intercal$ denotes the matrix transpose).

$$[\mathbf{H}]_p \leftarrow [\mathbf{H}]_p \odot \frac{\sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1} \odot \; [\mathbf{R}]_l}{\sum_{l=1}^{F} [\tilde{\mathbf{Y}}]_{l-p+1}}$$

$$\mathbf{X_s} \leftarrow \mathbf{X_s} \odot \frac{\sum_{t=1}^{T}\sum_{p=1}^{L} \mathbf{S_t}^{\intercal} \left([\mathbf{H}]_p \circledcirc \overset{\leftarrow\tau}{\mathbf{R}} \right)}{\sum_{t=1}^{T}\sum_{p=1}^{L} \mathbf{S_t}^{\intercal} \left([\mathbf{H}]_p \circledcirc \overset{\leftarrow\tau}{\mathbf{1}} \right) + \lambda_s}$$

$$\mathbf{X_n} \leftarrow \mathbf{X_n} \odot \frac{\sum_{t=1}^{T} \mathbf{N_t}^{\intercal} \overset{\leftarrow\tau}{\mathbf{R}}}{\sum_{t=1}^{T} \mathbf{N_t}^{\intercal} \overset{\leftarrow\tau}{\mathbf{1}} + \lambda_n}$$

where, $\tau = p + t - 2$, $\mathbf{1}$ is a matrix of ones of the same size as $\mathbf{Z}$ and $\odot$ denotes element-wise multiplication. The operation $\overset{\leftarrow\tau}{\mathbf{R}}$ shifts the matrix to the left by removing the first τ columns and adding τ zero columns to the right. The optimal parameters are obtained after updating the RIR and activations in an alternating fashion for several iterations. After every iteration, we apply a regularisation over $\mathbf{H}$ by element-wise dividing all its columns by the first column and clamp every column such that $\mathbf{H}(f,p) < \mathbf{H}(f,p-1)$. The rows of $\mathbf{H}$ are also normalised to sum to one to obtain a bounded estimate.

The optimal frame-level estimates for clean speech $\tilde{\mathbf{Y}}$ and noisy reverberant speech $\tilde{\mathbf{Z}}$ are then found using (4) and (5). From these

estimates, we construct a time-varying filter $\mathbf{G}$ to obtain the enhanced complex-valued STFT, $\mathbf{G} \odot \mathcal{Z}$, where $\mathbf{G}$ is the element-wise ratio $\mathbf{G} = \hat{\mathbf{Y}} \oslash \tilde{\mathbf{Z}}$. The enhanced speech is obtained using the overlap method from the enhanced complex-valued STFT.

3.2. Compositional model using NMF

This is based on the work presented in [6] where a similar model is used for dereverberation of noise-free reverberant speech. We extend this technique such that it also models noise and estimates $\mathbf{H}$ from the clean speech estimate.

The NMF formulation operates on stacked vectors formed by T consecutive frames of data [6]. For this, we use a sliding window of length T frames over the frame axis of $\mathbf{Z}$ and the features belonging to each window are stacked and stored as columns in the data matrix $\mathbf{Z_{st}}$. For $\mathbf{Z}$ with F frames this will result in $N_w = F - T + 1$ windows and $\mathbf{Z_{st}}$ will be of size $K \cdot T \times N_w$. The RIR matrix $\mathbf{H}$ is also stacked T times to obtain the stacked RIR $\mathbf{H_{st}}$ of size $K \cdot T \times L$.

In this setting, the noisy reverberant speech is represented as,

$$\mathbf{Z_{st}} \approx \tilde{\mathbf{Z}}_{\mathbf{st}} = \sum_{p=1}^{L} [\mathbf{H_{st}}]_p \circledcirc \mathbf{S}\mathbf{X_s} \; + \; \mathbf{N}\mathbf{X_n} \tag{10}$$

The optimal set of parameters are obtained such that they minimise the cost function (7) by replacing the frame-level features with the stacked features. The multiplicative updates are obtained in similar manner as explained in Section 3.1 and from [6] as:

$$\mathbf{H}(k,p) \leftarrow \mathbf{H}(k,p) \odot \frac{\sum_{t=1}^{T}\sum_{l=1}^{N_w} \tilde{\mathbf{Y}}_{\mathbf{st}}(r, l-p+1)\mathbf{R_{st}}(r,l)}{\sum_{t=1}^{T}\sum_{l=1}^{N_w} \tilde{\mathbf{Y}}_{\mathbf{st}}(r, l-p+1)}$$

$$\mathbf{X_s} \leftarrow \mathbf{X_s} \odot \frac{\sum_{p=1}^{L} \mathbf{S}^{\intercal}\left([\mathbf{H_{st}}]_p \circledcirc \overset{\leftarrow(p-1)}{\mathbf{R_{st}}}\right)}{\sum_{p=1}^{L} \mathbf{S}^{\intercal}\left([\mathbf{H_{st}}]_p \circledcirc \overset{\leftarrow(p-1)}{\mathbf{1}}\right) + \lambda_s}$$

$$\mathbf{X_n} \leftarrow \mathbf{X_n} \odot \frac{\mathbf{N_t}^{\intercal}\mathbf{R_{st}}}{\mathbf{N_t}^{\intercal}\mathbf{1} + \lambda_n}$$

where, $\tilde{\mathbf{Y}}_{\mathbf{st}} = \mathbf{S}\mathbf{X_{st}}$ obtained from the current estimate of $\mathbf{X_{st}}$, $r = k + (t-1)K$ and $\mathbf{R_{st}}$ is the element-wise ratio between $\mathbf{Z_{st}}$ and $\tilde{\mathbf{Z}}_{\mathbf{st}}$. The optimal values are obtained by iteratively applying the above updates until convergence. Notice that the update is only applied on the frame-level $\mathbf{H}$ at every iteration followed by stacking it to obtain $\mathbf{H_{st}}$. We also apply the same kind of regularisation on $\mathbf{H}$ as in Section 3.1 during every iteration.

The gain function $\mathbf{G}'$ to enhance the noisy STFT using this setting requires converting the stacked parameters into the frame-level estimates. Notice that the estimate of a frame appears over different overlapping windows and we sum those to obtain the frame-level estimates. Scaling with the number of overlapping windows is omitted as it appears both on the numerator and the denominator of the gain function. This procedure is in fact exactly the same as the operations defined in Equations (4) and (5) to obtain $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Z}}$. The gain function and the enhanced STFT are obtained as in Section 3.1.

4. EVALUATION SETUP

To evaluate and compare the settings described in this paper, development set of the CHiME-2 WSJ0 corpus is used. It contains 410 utterances taken from the WSJ0 development set corpus that are artificially reverberated and added with realistic background noise [13]. The sampling frequency is 16 kHz. The database contains binaural noisy reverberant speech at SNRs ranging from -6 dB to 9 dB in steps of 3 dB. In addition, the performance on the noise-free reverberant speech is also evaluated. The stereo data is averaged across the channels to obtain the single channel data.

For the NMD and NMF -based approaches, the STFT frame length and frame shift are set to 25 ms and 10 ms, respectively. A temporal context of $T = 10$ frames is used in all cases. To obtain the exemplars for creating the dictionaries, we randomly choose training data spanning 10 frames (115 ms) and its magnitude STFT is taken. Only the positive half of the magnitude STFT is considered and are reshaped to a vector to obtain an exemplar of length 2 560.

The clean training corpus of WSJ0 corpus is used to create the speech dictionary which contain $J_s = 5000$ randomly chosen exemplars. The noise dictionary used in this work consists of two parts: a fixed and a *sniffed* part. The fixed part of the dictionary is constructed using 2 500 randomly chosen noise exemplars taken from the background noise recordings provided with the CHiME-2 dataset. The sniffed noise dictionary is created on the fly from the embedded noisy utterances present in the database, that contains 5 seconds of noise context immediately before and after the utterance. This provides knowledge about the noise from the immediate context which can be beneficial for a better noise modelling in such difficult tasks. The sniffed noise exemplars are created from these 10 seconds of data which yields almost 1000 sniffed exemplars. This part is updated for every test utterance. The noise dictionary thus contains 3 500 noise exemplars.

A sparsity penalty of $\lambda_s = 1.6$ and $\lambda_n = 0.8$ is used in all formulations as used in our previous works [14,15]. The multiplicative updates are applied for 100 iterations with randomly initialised set of $\mathbf{H}$ and activations. We first evaluate the NMF and NMD settings without any reverberation model (denoted as *NMF* and *NMD*, respectively) and then various experiments are conducted with incorporating the proposed reverberant speech model (denoted as *NMF+R* and *NMD+R*, respectively) for various choices of RIR lengths L.

To evaluate and compare the speech enhancement quality, we use the signal-to-distortion ratio (SDR), PESQ [16], cepstral distance (CD) and segmental SNR (segSNR) measurements. The SDR is obtained using the BSS evaluation toolkit [17] and CD is obtained using the tool provided with the REVERB challenge [18]. We also make use of improvements in these measures (ΔSDR, ΔPESQ, ΔCD and ΔsegSNR) for comparing the results. The Δs are obtained by subtracting the metric obtained on the noisy data from that of the enhanced data for PESQ, SDR and segSNR measures (because higher measures mean better performance). On the other hand, since a lower CD implies a better performance, ΔCD is obtained by subtracting the CD obtained for enhanced speech from that of the noisy speech. In short, for all the Δ measurements, higher values mean a better performance.

The MATLAB codes for implementing the NMF and NMD -based updates for jointly estimating the RIR and activations are available in our webpage[1]. Some examples of the noisy and enhanced speech using these techniques are also provided.

5. RESULTS AND DISCUSSION

5.1. Evaluation on noise-free data

This section details the experiments conducted on noise-free reverberant data. The experiments are conducted for various RIR lengths

[1] http://www.esat.kuleuven.be/psi/spraak/downloads/

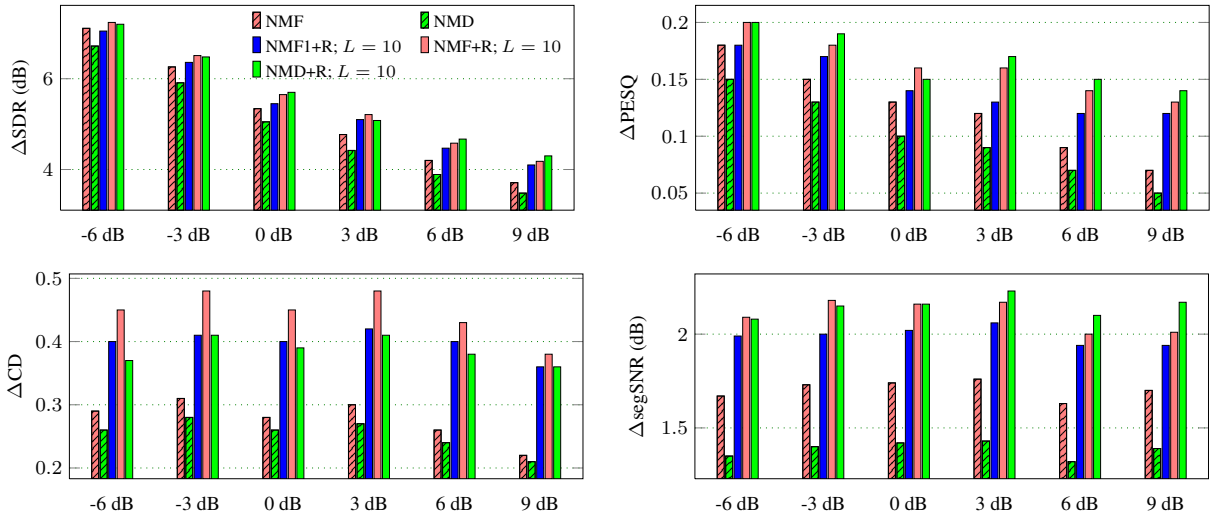

Fig. 1. *Improvements in SDR, PESQ, CD and segSNR obtained for the various evaluated settings on the CHiME-2 WSJ0 database as a function of the input SNRs. All figures use the same legends.*

Setting	SDR (dB)	PESQ	CD
No Enhancement	6.00	2.48	4.03
NMF	6.09	2.49	4.16
+RIR; $L = 5$	6.00	2.51	3.92
+RIR; $L = 7$	5.99	2.51	3.92
+RIR; $L = 10$	5.99	2.52	3.92
NMD	5.93	2.46	4.04
+RIR; $L = 5$	6.75	2.55	3.86
+RIR; $L = 7$	6.91	2.55	3.83
+RIR; $L = 10$	**7.00**	**2.56**	**3.82**

Table 1. *Speech enhancement results obtained for the various evaluated settings with varying RIR length L on noise-free reverberant speech. Best scores are highlighted in bold font.*

and the corresponding SDR, PESQ and CD obtained are summarised in Table 1. It is also verified that the multiplicative updates always result in a decreasing cost.

It can be seen that introducing the proposed approaches result in improvements for all the measures. The NMF-based approach does not introduce noticeable improvements when evaluated on the CHiME-2 noise-free reverberant data and a increasing the value of L beyond 5 does not introduce much improvements. On the other hand, the NMD-based approach provides significant improvements for all the evaluated measures and increasing the RIR length yields further improvements.

5.2. Evaluation on noisy reverberant data

The proposed approaches are evaluated on the noisy reverberant data with a RIR length of $L = 10$. Figure 1 summarises the improvements obtained for various measures. It can be seen that the proposed approaches always result in a performance improvement when compared to the settings where no reverberation model is used. In addition, we also include a baseline setting where the RIR is estimated from both the speech and noise estimates as used in [12], which is denoted as *NMF1+R*.

The NMF+R setting is found to outperform the NMF1+R setting in all cases, validating the claim that the RIR estimate will be less reliable when it is estimated using both the speech and noise estimates. Notice that, in the absence of a reverberation model, the NMF-based technique yields a better denoising performance when comapared to the NMD-based technique. This suggests that the NMF-based model results in a better speech and noise estimate. However, the NMD+R approach is still able to yield more or less comparable performance with NMF+R once the reverberation model is introduced, suggesting that the NMD+R formulation is a better model for estimating the RIR.

The NMD+R approach outperformed the NMF+R technique for positive SNRs in terms of PESQ, SDR and segSNR measurements. This implies that adding the reverberation model to the traditional NMD-based formulation equips the setting to estimate better approximations of the underlying speech and noise.

6. CONCLUSIONS AND FUTURE WORK

This paper proposed a supervised speech derevereration technique based on exemplar-based sparse representations for jointly estimating the RIR together with the speech and noise estimates. A novel formulation based on the NMD-based decomposition of noisy speech is proposed along with an extension to an existing NMF-based model. We also provide the update equations for estimating the various parameters for both the formulations such that they minimize the Kullback-Leibler divergence between the observed noisy reverberant data and its approximation. Evaluations on the development set of CHiME-2 WSJ0 data show that the proposed techniques yield better speech enhancement quality.

Introducing better regularisations on the RIR and incorporating it as part of the cost function is a future work. Investigating the utility of such a setting as a front-end for automatic speech recognition is also a remaining work.

7. REFERENCES

[1] Y.A. Huang, J. Benesty, and Jingdong Chen, "A blind channel identification-based two-stage approach to separation and dereverberation of speech signals in a reverberant environment," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 5, pp. 882–895, Sept 2005.

[2] H. Buchner, R. Aichner, and W. Kellermann, "A generalization of blind source separation algorithms for convolutive mixtures based on second-order statistics," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 1, pp. 120–134, Jan 2005.

[3] H. Buchner, R. Aichner, and W. Kellermann, "Trinicon-based blind system identification with application to multiple-source localization and separation," in *Blind Speech Separation*, S. Makino, H. Sawada, and T. Lee, Eds., Signals and Communication Technology, pp. 101–147. Springer Netherlands, 2007.

[4] T. Yoshioka, T. Nakatani, M. Miyoshi, and H.G. Okuno, "Blind separation and dereverberation of speech mixtures by joint optimization," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 69–84, Jan 2011.

[5] R. Talmon, I. Cohen, and S. Gannot, "Relative Transfer Function Identification Using Convolutive Transfer Function Approximation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 4, pp. 546–555, May 2009.

[6] N. Mohammadiha, P. Smaragdis, and S. Doclo, "Joint acoustic and spectral modeling for speech dereverberation using nonnegative representations," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4410–4414.

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

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

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

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

[11] P. Smaragdis, "Non-negative matrix factor deconvolution; extraction of multiple sound sources from monophonic inputs," in *Independent Component Analysis and Blind Signal Separation*, C. Puntonet and A. Prieto, Eds., vol. 3195 of *Lecture Notes in Computer Science*, pp. 494–499. Springer Berlin Heidelberg, 2004.

[12] H. Kallasjoki, J. F. Gemmeke, K. J. Pallomaki, and A. V. Beeston, "Recognition of reverberant speech by missing data imputation and NMF feature enhancement," in *Proc. REVERB Workshop*, Florence, Italy, May 2014.

[13] E. Vincent, J. Barker, S. Watanabe, J. Le Roux, F. Nesta, and M. Matassoni, "The second 'CHiME' speech separation and recognition challenge: Datasets, tasks and baselines," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 126–130.

[14] D. Baby, T. Virtanen, J. F. Gemmeke, and H. Van hamme, "Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 23, no. 11, pp. 1788 –1799, Nov. 2015.

[15] D. Baby, J. F. Gemmeke, T. Virtanen, and H. Van hamme, "Exemplar-based Speech Enhancement for Deep Neural Network based Automatic Speech Recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 4485–4489.

[16] P. C. Loizou, *Speech Enhancement: Theory and Practice (Signal Processing and Communications)*, CRC Press, 1 edition, 2007.

[17] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 4, pp. 1462–1469, 2006.

[18] K. Kinoshita, M. Delcroix, T. Yoshioka, T. Nakatani, A. Sehr, W. Kellermann, and R. Maas, "The reverb challenge: Acommon evaluation framework for dereverberation and recognition of reverberant speech," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, Oct 2013, pp. 1–4.