JOINT ACOUSTIC AND SPECTRAL MODELING FOR SPEECH DEREVERBERATION USING NON-NEGATIVE REPRESENTATIONS

Nasser Mohammadiha[†] *Paris Smaragdis*[‡] *Simon Doclo*[†]

[†] Dept. of Medical Physics and Acoustics and Cluster of Excellence Hearing4all
University of Oldenburg, Germany
[‡] University of Illinois at Urbana-Champaign, USA
Adobe Systems Inc.

ABSTRACT

This paper proposes a single-channel speech dereverberation method enhancing the spectrum of the reverberant speech signal. The proposed method uses a non-negative approximation of the convolutive transfer function (N-CTF) to simultaneously estimate the magnitude spectrograms of the speech signal and the room impulse response (RIR). To utilize the speech spectral structure, we propose to model the speech spectrum using non-negative matrix factorization, which is directly used in the N-CTF model resulting in a new cost function. We derive new estimators for the parameters by minimizing the obtained cost function. Additionally, to investigate the effect of the speech temporal dynamics for dereverberation, we use a frame stacking method and derive optimal estimators. Experiments are performed for two measured RIRs and the performance of the proposed method is compared to the performance of a state-of-the-art dereverberation method enhancing the speech spectrum. Experimental results show that the proposed method improved instrumental speech quality measures, where using speech temporal dynamics was found to be beneficial in severe reverberation conditions.

Index Terms— Non-negative convolutive transfer function, non-negative matrix factorization, dictionary-based processing

1. INTRODUCTION

The quality and intelligibility of speech signals recorded using a distant microphone in an enclosed space may highly degrade due to reverberation, i.e., the reflections from the surrounding objects. Therefore, in many applications, such as hearing aids and automatic speech recognition, it is important to recover the non-reverberant clean speech signal [1].

Several single-channel dereverberation approaches have been proposed in the literature aiming to blindly estimate the speech signal from a reverberant recording. Most single-channel approaches are based on either inverse filtering [2–4] or speech spectral enhancement [5, 6].

In this paper, we propose a single-channel dereverberation method operating in the magnitude spectrogram domain. We assume that the magnitudes of the short-time Fourier transform (STFT) coefficients of the reverberant signal in each frequency bin are obtained by convolving the STFT magnitudes of the clean speech signal and RIR in that frequency bin. Such non-negative convolutive transfer function (N-CTF) model, which only holds approximately, can be advantageous as it does not model the RIR phase variations which are difficult to be robustly modeled [7]. Methods based on similar approximations have been recently proposed [7–11], where the speech spectrogram is additionally assumed to be sparse. These methods utilize the N-CTF model to form an optimization problem to blindly estimate the speech spectrogram. Therefore, these estimation methods are purely based on the acoustic N-CTF model, while ignoring the spectral structure of the speech signal.

The main contribution of this paper is to propose a blind single-channel speech dereverberation method by jointly modeling the room acoustic using the N-CTF model and the speech spectrogram using non-negative matrix factorization (NMF). We propose a new model by combining the N-CTF and NMF models and construct a cost function. We derive new estimators for the model parameters by minimizing the obtained cost function. Additionally, we present a method based on the frame stacking concept [12] to utilized the speech temporal dynamics. Experimental results show that by additionally using the NMF-based spectral model the Perceptual Evaluation of Speech Quality (PESQ) [13] scores improve substantially and become superior to that of a state-of-the-art dereverberation method based on spectral enhancement [6]. An additional improvement is obtained by learning the NMF model offline from clean speech training data, while using the speech temporal dynamics was found to be beneficial in relatively severe reverberation conditions and when a low-rank NMF model was used.

2. NON-NEGATIVE CONVOLUTIVE TRANSFER FUNCTION (N-CTF)

Let $s(n)$ and $h(n)$ denote the discrete-time clean speech signal and M-tap RIR, where n denotes the sample index. The reverberant speech signal $y(n)$ is obtained by convolving $s(n)$ and $h(n)$ as:

$$y\left(n\right) = \sum_{m=0}^{M-1} h\left(m\right) s\left(n-m\right). \tag{1}$$

In the STFT domain, (1) can be approximated as [14]:

$$y_c\left(k,t\right) \approx \sum_{\tau=0}^{L_h-1} h_c\left(k,\tau\right) s_c\left(k,t-\tau\right), \tag{2}$$

where $y_c(k,\tau)$, $s_c(k,\tau)$, and $h_c(k,\tau)$ denote the complex-valued STFT coefficients of reverberant speech signal, clean speech signal, and RIR, respectively, k and t denote the frequency and frame indices, respectively, and L_h is the RIR length in the STFT domain [15]. Based on (2), it has been proposed in [7] to approximate

This research was supported by the Cluster of Excellence 1077 "Hearing4all", funded by the German Research Foundation (DFG).

the spectral power $|y_c(k,t)|^2$ as

$$|y_c(k,t)|^2 \approx \sum_{\tau=0}^{L_h-1} |h_c(k,\tau)|^2 |s_c(k,t-\tau)|^2, \quad (3)$$

where $|\cdot|$ denotes the absolute value operator. This approximation can be justified in the sense of mathematical expectation assuming that the RIR phase component is an independent uniformly-distributed random variable [7]. In this paper, we will use the STFT magnitude coefficients instead of the magnitude-squared coefficients since experimental results showed that this resulted in better-quality dereverberated speech signals. A similar observation has been also made in [9]. Thus, we have

$$y(k,t) \approx \sum_{\tau=0}^{L_h-1} h(k,\tau) s(k,t-\tau), \quad (4)$$

where $y(k,t) = |y_c(k,t)|$, and $s(k,t)$ and $h(k,t)$ are defined similarly. We refer to (4) as the non-negative convolutive transfer function (N-CTF) model.

3. PROPOSED METHOD

Based on the N-CTF model (4), we can estimate the magnitude spectrogram of the clean speech signal $s(k,t)$ by minimizing the distance between the left and right hand sides of (4). We use the Kullback-Leibler (KL) divergence between y and $\tilde{y}$ as a possible distance measure, which is defined as:

$$KL(y|\tilde{y}) = y \log \frac{y}{\tilde{y}} + \tilde{y} - y. \quad (5)$$

Accordingly, to estimate s and h, the following cost function should be minimized:

$$Q = \sum_{k,t} KL\left(y(k,t) \middle| \sum_{\tau} h(k,\tau) s(k,t-\tau)\right). \quad (6)$$

Because of the sparse nature of the speech spectrograms, it is also beneficial to add a regularization term to (6) to obtain a sparse estimate for s. We use the l_1-norm of the speech spectrogram as a measure of sparseness. The l_1-norm of the speech spectrogram has been also used in [7, 8] to obtain a regularized Euclidean distance. Therefore, the following regularized cost function should be minimized in order to estimate s and h:

$$Q = \sum_{k,t} KL\left(y(k,t) \middle| \sum_{\tau} h(k,\tau) s(k,t-\tau)\right) + \lambda \sum_{k,t} s(k,t), \quad (7)$$

where λ is a weighting parameter to encourage sparse estimates for s. Note that the cost function Q does not include any criterion related to the structure of the speech spectra (except its sparsity), e.g., individual frequency bins are treated independently. In order to incorporate some knowledge about the structure of the speech spectra, e.g., its low-rank nature and dependency across frequencies, we propose to use a speech spectral model, resulting in a new cost function. Motivated by the successful modeling of speech spectra using non-negative matrix factorization (NMF) in different applications, e.g., [16–19], we propose to use an NMF-based spectral model:

$$s(k,t) \approx \sum_{r=1}^{R} w(k,r) x(r,t), \quad (8)$$

where R is the number of columns in the dictionary $\mathbf{W} = [w(k,r)]$. In matrix notations, (8) can be written as $\mathbf{S} \approx \mathbf{W}\mathbf{X}$, where $\mathbf{S} = [s(k,t)]$ and $\mathbf{X} = [x(r,t)]$ denote the speech magnitude spectrogram, and the activation matrix, respectively. If R is chosen to be smaller than the dimensions of $\mathbf{S}$, (8) imposes a low-rank structure on the speech spectrogram.

To jointly model the room acoustics and the speech spectra, we propose to directly replace $s(k,t)$ in (4) with its NMF approximation in (8), leading to:

$$Q = \sum_{k,t} KL\left(y(k,t) \middle| \sum_{\tau} h(k,\tau) \sum_{r} w(k,r) x(r,t-\tau)\right) + \lambda \sum_{r,t} x(r,t). \quad (9)$$

Since s does not directly appear in (9), as can be seen, the sparsity regularization is now imposed on the activations x. This will implicitly help to obtain sparse estimates for s, considering the relation between s and x in (8). The model in (9) includes the N-CTF model (4) as a special case when the dictionary $\mathbf{W}$ is a $K \times K$-dimensional identity matrix, where K denotes the number of frequency bins. A similar idea has been used in [10], which is also a special case of (9) when $\mathbf{W}$ is a fixed matrix.

To minimize (9), we use an auxiliary function method, similar to [20], which leads to iterative multiplicative update rules for h, w, and x. Hence, h, w, and x are updated iteratively and when updating one of the variables the other two variables are held fixed using their estimates from the previous iteration.

Let us first consider the optimization with respect to (w.r.t.) h, and let $Q(h)$ denote all terms depending on h in (9). Also, let estimates of h, w, and x at the i-th iteration be denoted by h^i, w^i, and x^i, respectively. The following lemma can be stated [20].

Lemma 1. *Let $G(h,h^i)$ be an auxiliary function for $Q(h)$ such that $G(h,h) = Q(h)$ and $G(h,h^i) \geq Q(h)$ for a given h^i and all h. Let h^{i+1} be the new estimate obtained by minimizing $G(h,h^i)$ w.r.t. h. $Q(h)$ is non-increasing under this update, i.e., $Q(h^{i+1}) \leq Q(h^i)$, where equality holds only when h^i is a local minimum of $Q(h)$.*

Theorem 1. *The function $Q(h)$ is non-increasing under the following update rule:*

$$h^{i+1}(k,\tau) = h^i(k,\tau) \frac{\sum_t y(k,t)\tilde{s}(k,t-\tau)/\tilde{y}(k,t)}{\sum_t \tilde{s}(k,t-\tau)}, \quad (10)$$

where $\tilde{s}(k,t) = \sum_r w^i(k,r)x^i(r,t)$, and $\tilde{y}(k,t) = \sum_\tau h^i(k,\tau) \times \tilde{s}(k,t-\tau)$.

Proof. Since $-\log \sum_\tau h(k,\tau)\tilde{s}(k,t-\tau)$ is convex, using Jensen's inequality [21] with $h(k,\tau)$ as the variable we have:

$$-\log \sum_{\tau} h(k,\tau)\tilde{s}(k,t-\tau) \leq -\sum_{\tau} \frac{h^i(k,\tau)\tilde{s}(k,t-\tau)}{\tilde{y}(k,t)} \log \frac{\tilde{y}(k,t) h(k,\tau)\tilde{s}(k,t-\tau)}{h^i(k,\tau)\tilde{s}(k,t-\tau)}. \quad (11)$$

Using the above inequality in the definition of $Q(h)$ (which is omitted here due to space limit), we obtain the following auxiliary func-

tion for $Q(h)$:

$$Q(h) \leq G(h,h^i) = -\sum_{k,t,\tau}\Bigg(y\left(k,t\right)\frac{h^i\left(k,\tau\right)\tilde{s}\left(k,t-\tau\right)}{\tilde{y}\left(k,t\right)}\times \log h\left(k,\tau\right)\Bigg) + \sum_{k,t,\tau}\left(h\left(k,\tau\right)\tilde{s}\left(k,t-\tau\right) - C\left(k,t,\tau\right)\right), \quad (12)$$

where $C(k,t,\tau) = y\left(k,t\right)\frac{h^i(k,\tau)\tilde{s}(k,t-\tau)}{\tilde{y}(k,t)}\log\frac{\tilde{y}(k,t)}{h^i(k,\tau)}$ is a constant. Differentiating $G(h,h^i)$ w.r.t. $h(k,\tau)$ and setting it to zero yields (10). Recalling Lemma 1, the proof is complete. □

Update rules for w and x can be derived similarly:

$$w^{i+1}(k,r) = w^i(k,r)\frac{\sum_{t,\tau} y\left(k,t\right)h^{i+1}\left(k,\tau\right)x^i\left(r,t-\tau\right)/\tilde{y}\left(k,t\right)}{\sum_{t,\tau} h^{i+1}\left(k,\tau\right)x^i\left(r,t-\tau\right)}, \quad (13)$$

$$x^{i+1}(r,t) = x^i\left(r,t\right)\frac{\sum_{k,\tau} y\left(k,t+\tau\right)h^{i+1}(k,\tau)w^{i+1}(k,r)/\tilde{y}(k,t+\tau)}{\sum_{k,\tau} h^{i+1}\left(k,\tau\right)w^{i+1}\left(k,r\right)+\lambda}, \quad (14)$$

where $\tilde{y}(k,t)$, defined after (10), is computed using the latest estimates of the parameters. These update rules can be efficiently implemented using the fast Fourier transform (FFT) [7]. To remove the scale ambiguity[1], after each iteration, each column of $\mathbf{W}$ is normalized to sum to one, and the columns of $\mathbf{H}$ are element-wise divided by its first column, and $h(k,\tau)$ is clamped to satisfy $h(k,\tau) < h(k,\tau-1)$ for all τ.

Let $\hat{\mathbf{W}} = [\hat{w}(k,r)]$, $\hat{\mathbf{X}} = [\hat{x}(k,r)]$, and $\hat{\mathbf{H}} = [\hat{h}(k,r)]$ denote the obtained estimates after convergence of the iterative algorithm. One possible estimate for the speech magnitude spectrogram $\mathbf{S}$ is given by $\hat{\mathbf{S}} = \hat{\mathbf{W}}\hat{\mathbf{X}}$. Alternatively, we suggest to estimate the speech spectrogram using a time-varying gain function as

$$\hat{s}(k,t) = G(k,t)y(k,t), \quad (15)$$

where the gain function $G(k,t)$ is given by

$$G\left(k,t\right) = \frac{\sum_r \hat{w}\left(k,r\right)\hat{x}\left(r,t\right)}{\sum_{r,\tau}\hat{h}\left(k,\tau\right)\hat{w}\left(k,r\right)\hat{x}\left(r,t-\tau\right)}. \quad (16)$$

This was found to be particularly advantageous when the dictionary $\mathbf{W}$ was learned offline from speech training data and was held fixed for dereverberation.

Since temporal correlations are an important aspect of speech signals, we describe an extension of the above algorithm where we stack the consecutive frames to form super-vectors to model the temporal correlations [12]. Let $\mathbf{y}(t)$ denote the $t-$th column of $\mathbf{Y} = [y(k,t)]$, and let K denote the number of frequency bins or dimension of $\mathbf{y}(t)$. We define the KT_{st}-dimensional vector $\mathbf{y}_{st}(t)$ as $\mathbf{y}_{st}^T(t) = [\mathbf{y}^T(t)\ldots\mathbf{y}^T(t+T_{st}-1)]$. $\mathbf{s}_{st}(t)$ is defined similarly. Also, let $\mathbf{h}_{st}(t)$ be the KT_{st}-dimensional vector defined as $\mathbf{h}_{st}^T(t) = [\mathbf{h}^T(t)\ldots\mathbf{h}^T(t)]$. The new cost function is obtained by replacing y and h in (9) by their stacked counterparts y_{st} and h_{st}, where $\mathbf{W}$ is a $KT_{st}\times R$ matrix. The update rules for w and x remain identical to (13) and (14). The update rule for h can be derived similarly to Theorem 1 and is given by

$$h^{i+1}\left(k,\tau\right) = h^i\left(k,\tau\right)\frac{\sum_{l=1}^{T_{st}}\sum_t y_{st}\left(f,t\right)\tilde{s}_{st}\left(f,t-\tau\right)/\tilde{y}_{st}\left(f,t\right)}{\sum_{l=1}^{T_{st}}\sum_t \tilde{s}_{st}\left(f,t-\tau\right)}, \quad (17)$$

where $f = k + K(l-1)$, $\tilde{s}_{st}(f,t) = \sum_r w^i\left(f,r\right)x^i\left(r,t\right)$, and $\tilde{y}_{st}\left(f,t\right) = \sum_\tau h_{st}^i\left(f,\tau\right)\tilde{s}_{st}\left(f,t-\tau\right)$, where KT_{st}-dimensional vector $\mathbf{h}_{st}^i(t)$ is defined as $\mathbf{h}_{st}^{i,T}(t) = [\mathbf{h}^{i,T}(t)\ldots\mathbf{h}^{i,T}(t)]$. After convergence of the iterative algorithm, the speech magnitude spectrogram is estimated as $\hat{s}(k,t) = G(k,t)y(k,t)$, where $G(k,t)$ is obtained by averaging over the overlapping segments:

$$G\left(k,t\right) = \frac{\sum_{l=1}^{T_{st}}\sum_r \hat{w}\left(f,r\right)\hat{x}\left(r,t\right)}{\sum_{l=1}^{T_{st}}\sum_{r,\tau}\hat{h}_{st}\left(f,\tau\right)\hat{w}\left(f,r\right)\hat{x}\left(r,t-\tau\right)}, \quad (18)$$

where $\hat{}$ is used to denote the obtained estimates after convergence. After estimating the speech spectrogram, the time-domain clean speech signal $s(n)$ is estimated by applying inverse STFT on the estimated spectrogram $\hat{s}(k,t)$, where the reverberant phase spectra and the overlap-add procedure are used.

4. EXPERIMENTAL RESULTS

We applied our proposed method to dereverberate speech signals obtained by convolving clean speech signals with two measured RIRs with reverberation times $T_{60} \approx 430$ and $T_{60} \approx 680$ ms, and direct-to-reverberation ratio (DRR) around 5 dB and 0 dB, respectively. The proposed methods were applied on 16 different speech sentences (uttered by different speakers) from the TIMIT database [22] to make the results independent of the speech material. The sampling frequency was 16 kHz and the STFT frame length and overlap length were set to 64 ms and 32 ms, respectively, where a square-root Hann window was used for both STFT analysis and synthesis.

The dereverberation performance is measured using PESQ [13] with clean speech signal as the reference. For the proposed method, we investigate two possible ways to learn the dictionary $\mathbf{W}$. First, a dictionary $\mathbf{W}$ with $R = 100$ dictionary elements was learned online from the reverberant signal *(N-CTF+NMF)*. Alternatively, $\mathbf{W}$ was learned offline from training data (consisting of 250 sentences uttered by 27 speakers, disjoint from the test data set) and it was held fixed. We report results for two cases: 1) a low-rank NMF with $R = 100$ was learned (*N-CTF+NMF-Dic100*), 2) an overcomplete speaker-independent dictionary with $R = 4000$ (*N-CTF+NMF-Dic4000*) was constructed by sampling from the magnitude spectrogram of the speech training data using a uniform random walk method [23]. Moreover, the performance of the N-CTF based dereverberation method (with $\mathbf{W}$ being fixed to an identity matrix) is evaluated in the experiments. Additionally, the proposed method is compared to a speech spectral enhancement (*SE*) method where the late reverberant spectral variance was estimated using [6] (with T_{60} and DRR computed from the RIR), and speech log-spectral amplitude estimator was used to enhance the reverberant speech [24].

The sparsity parameter λ was set to $\frac{0.1}{KT}\sum_{k,t} y(k,t)$ where T is the total number of frames. Additionally, to encourage sparser solutions, the estimates of x and s, after each iteration, were raised to a power ϕ_x as proposed in [25], where we experimentally set $\phi_x = 1.02$ when $R = 100$, and $\phi_x = 1.05$ when $R = 4000$. The iterative update rules were executed for 50 iterations for all methods, and the RIR length L_h was set to 10, independent of T_{60}. Each row of $\mathbf{H}$ was initialized identically using a linearly-decaying envelope, while $\mathbf{W}$ and $\mathbf{X}$ were initialized by iterating the standard NMF update rules [20] (with random initializations) on the spectrogram of the reverberant signal for 10 times.

Fig. 1 shows an example of the spectrograms of the reverberant, dereverberated, and clean speech signals. As can be observed, the reverberation effects have been reduced when an NMF-based spec-

[1]Note that if $\hat{\mathbf{H}}$, $\hat{\mathbf{W}}$, and $\hat{\mathbf{X}}$ are a solution to (9), the same optimal value for Q can be obtained using $\alpha\hat{\mathbf{H}}$, $\hat{\mathbf{W}}/\alpha$, and $\hat{\mathbf{X}}$ where α is a random non-negative number.

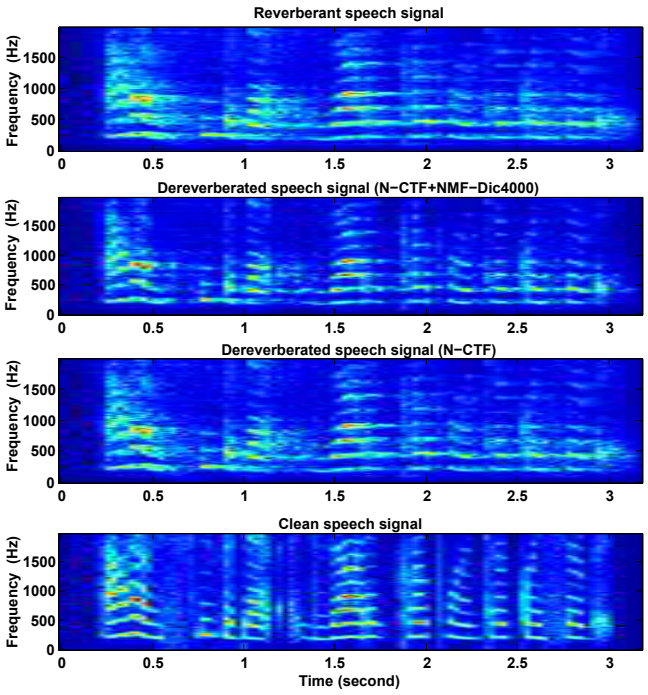

Fig. 1: Spectrograms of reverberant, dereverberated (using the *N-CTF+NMF-Dic4000*, second panel, and *N-CTF*, third panel, methods), and clean speech signals for a RIR with $T_{60} = 680$ ms.

tral model is additionally used in the N-CTF based dereverberation.

To quantitatively study the dereverberation performance, the PESQ improvements, averaged over all speech sentences, for the two RIRs are shown in Fig. 2. As can be seen, the dereverberation method using only the N-CTF model and the speech enhancement (*SE*) method lead to a comparable PESQ improvements. By introducing a low-rank structure on the speech signal (*N-CTF+NMF*), the performance has substantially improved for both RIRs. The results show that the performance of the *N-CTF+NMF-Dic100* method with $R = 100$ offline-learned dictionary elements is worse than the online counterpart *N-CTF+NMF*. However, by using a richer dictionary (*N-CTF+NMF-Dic4000*) the performance is further improved, where the proposed *N-CTF+NMF-Dic4000* method outperforms the *SE* method by more than 0.2 MOS points.

Next, we present the dereverberation results using the proposed method where the frame stacking method with 6 stacked frames, i.e., $T_{st} = 6$, is used to utilize the speech temporal dynamics. The obtained PESQ improvements are shown in Fig. 3, where the extension "*(S)*" is used to identify the methods with stacked vectors. Results show that in the mild reverberation conditions, top panel of Fig. 3, using the temporal dynamics degrades (or does not improve) the performance of the dereverberation methods. For relatively severe reverberation conditions, however, bottom panel of Fig. 3, using the temporal dynamics improves the performance of the *N-CTF+NMF-Dic100* method while it does not improve the performance of the *N-CTF+NMF* and the *N-CTF+NMF-Dic4000* methods.

The results show that the best performance is obtained using the *N-CTF+NMF-Dic4000* method, where the *N-CTF+NMF* method also leads to a good performance. While the *N-CTF+NMF-Dic4000* method requires a large 512×4000-dimensional matrix to be stored in the memory, the *N-CTF+NMF* method is computationally slightly more complex because it additionally updates the dictionary $\mathbf{W}$.

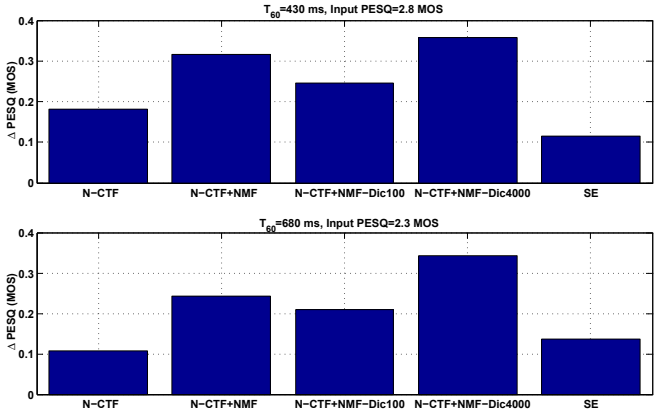

Fig. 2: PESQ improvements obtained using the proposed method with different parameters.

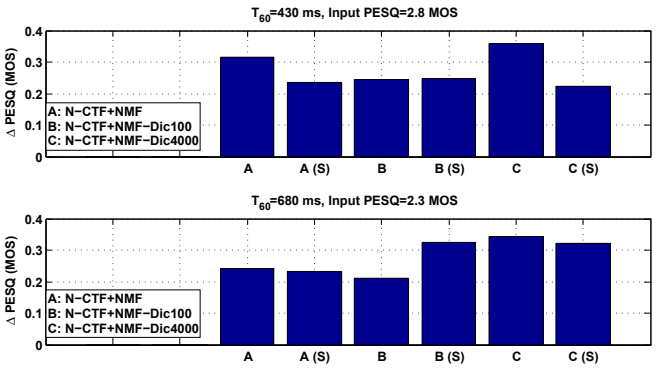

Fig. 3: PESQ improvements obtained using the proposed method with and without use of speech temporal dynamics, where the extension "*(S)*" is used to identify the methods with temporal dynamics.

Hence, depending on the available resources, one of the methods may be preferred.

5. CONCLUSION

In this paper, we developed a speech dereverberation method using the N-CTF model, where we proposed a method to additionally model the speech spectrum using NMF. The NMF model was directly used inside the N-CTF model resulting in a new cost function. The obtained cost function was then minimized to estimate the RIR magnitude spectrogram, NMF dictionary and NMF activation matrix. The speech magnitude spectrogram was then estimated using a time-varying gain function. To utilize the speech temporal dynamics for dereverberation, a frame stacking method was additionally used and corresponding optimal estimators were derived. Experimental results using two measured RIRs with $T_{60} \approx 680$ ms and $T_{60} \approx 430$ ms show that the dereverberation performance improves substantially when the NMF-based spectral method is utilized. Moreover, an additional improvement was observed when an overcomplete speaker-independent NMF dictionary was learned offline from speech training data, outperforming a state-of-the-art speech spectral enhancement method for dereverberation by 0.2 MOS points. The results show that, using the speech temporal dynamics can improve the dereverberation performance for severe reverberation conditions, while it degrades (or does not improve) the performance for mild reverberation conditions.

6. REFERENCES

[1] P. A. Naylor and N. D. Gaubitch, Eds., *Speech Dereverberation*, 1st ed. New York, USA: Springer, 2010.

[2] B. W. Gillespie, H. S. Malvar, and D. A. F. Florêncio, "Speech dereverberation via maximum-kurtosis subband adaptive filtering," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 6, Salt Lake City, Utah, USA, 2001, pp. 3701–3704.

[3] T. Nakatani, K. Kinoshita, and M. Miyoshi, "Harmonicity-based blind dereverberation for single-channel speech signals," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 80–95, Jan. 2007.

[4] I. Kodrasi, T. Gerkmann, and S. Doclo, "Frequency-domain single-channel inverse filtering for speech dereverberation: Theory and practice," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Florence, Italy, May 2014, pp. 5177–5181.

[5] K. Lebart, J. M. Bouche, and P. N. Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acoustica*, vol. 87, no. 3, pp. 359–366, 2001.

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

[7] H. Kameoka, T. Nakatani, and T. Yoshioka, "Robust speech dereverberation based on non-negativity and sparse nature of speech spectrograms," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Taipei, Taiwan, Apr. 2009, pp. 45–48.

[8] R. Singh, B. Raj, and P. Smaragdis, "Latent-variable decomposition based dereverberation of monaural and multi-channel signals," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, Dallas, Texas, USA, Mar. 2010, pp. 1914–1917.

[9] K. Kumar, R. Singh, B. Raj, and R. Stern, "Gammatone sub-band magnitude-domain dereverberation for ASR," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, May 2011, pp. 4604–4607.

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

[11] M. Yu and F. K. Soong, "Constrained multichannel speech dereverberation," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, Portland, Oregon, USA, 2012.

[12] J. F. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 19, no. 7, pp. 2067–2080, Sep. 2011.

[13] I.-T. P.862, "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assesment of narrowband telephone networks and speech codecs," Tech. Rep., 2000.

[14] R. Talmon, I. Cohen, and S. Gannot, "Relative transfer function identification using convolutive transfer function approximation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 17, no. 4, pp. 546–555, May 2009.

[15] Y. Avargel and I. Cohen, "System identification in the short-time Fourier transform domain with crossband filtering," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 4, pp. 1305–1319, May 2007.

[16] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 15, no. 1, pp. 1–12, Jan. 2007.

[17] C. Févotte, N. Bertin, and J. L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence: with application to music analysis," *Neural Computation*, vol. 21, pp. 793–830, 2009.

[18] N. Mohammadiha, P. Smaragdis, and A. Leijon, "Supervised and unsupervised speech enhancement using NMF," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 21, no. 10, pp. 2140–2151, Oct. 2013.

[19] N. Mohammadiha, "Speech enhancement using non-negative matrix factorization and hidden Markov models," Ph.D. dissertation, KTH - Royal Institute of Technology, Stockholm, Sweden, 2013. [Online]. Available: http://theses.eurasip.org/theses/499/speech-enhancement-using-nonnegative-matrix/download/

[20] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Process. Systems (NIPS)*. MIT Press, 2000, pp. 556–562.

[21] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd ed. New York: Wiley, 1973.

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

[23] N. Mohammadiha and S. Doclo, "Single-channel dynamic exemplar-based speech enhancement," in *Proc. Int. Conf. Spoken Language Process. (Interspeech)*, Singapore, Sep. 2014, pp. 2690–2694.

[24] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[25] A. Cichocki, R. Zdunek, and S. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 5, Toulouse, France, May 2006.