ROBUST SPEECH DEREVERBERATION BASED ON NON-NEGATIVITY AND SPARSE NATURE OF SPEECH SPECTROGRAMS

Hirokazu Kameoka, Tomohiro Nakatani, Takuya Yoshioka,

NTT Communication Science Laboratories, NTT Corporation,
3-1 Morinosato Wakamiya, Atsugi, Kanagawa 243-0198, Japan

ABSTRACT

This paper presents a blind dereverberation method designed to recover the subband envelope of an original speech signal from its reverberant version. The problem is formulated as a blind deconvolution problem with non-negative constraints, regularized by the sparse nature of speech spectrograms. We derive an iterative algorithm for its optimization, which can be seen as a special case of the non-negative matrix factor deconvolution. We confirmed through experiments that the algorithm is fast and robust to speaker movement.

Index Terms— Speech dereverberation, temporal envelope filtering, non-negative blind deconvolution, sparseness

1. INTRODUCTION

Reverberation in a room severely degrades the characteristics and quality of speech captured by distant microphones, thus posing a severe problem for many speech applications. The aim of speech dereverberation techniques is to recover a clean speech signal from its reverberant version. When only a reverberant speech signal is accessible, dereverberation systems must work 'blind'.

A number of blind dereverberation systems have already been developed. While there are several viable solutions in situations where multiple channels are available, single channel systems still pose a formidable challenge. For single channel systems two main approaches have been adopted. One approach involves applying inverse filtering techniques. This approach requires assumptions about clean speech (e.g. harmonicity, sparseness, non-stationarity, non-Gaussianity) and/or the use of a speech model (e.g. dual excitation model, autoregressive model, codebook) to assess 'speech naturalness'. Systems are then able to estimate the optimal inverse filter such that the speech naturalness of the inverse-filtered signal is maximized [1]–[9]. While inverse filtering methods are usually performed in the time domain, the short-time Fourier transform (STFT)-domain approach is more computationally efficient [8, 9]. Within this framework, since room impulse responses can vary rapidly according to such factors as the speaker's position, the problem of how robustly and adaptively a system can estimate an inverse filter from a short-term observation has attracted particular attention.

With a different approach, an attempt is made to recover the subband envelope (power envelope of a subband signal) of the original speech by applying an inverse filter of the modulation transfer function (MTF) [10]–[17]. As it is based on a model that implicitly assumes the additivity of power spectra, which holds only approximately, the performance may be limited to some extent. However, we hypothesize that it is advantageous in one sense. That is, the phase characteristic of a room impulse response is a characteristic that is especially sensitive to speaker movements. Therefore, the use of a convolution model of the subband envelopes, in which only the time-invariance of the amplitude characteristic is assumed, may allow the system to be robust against the changes in room characteristics.

On the basis of this hypothesis, we choose the latter approach in this paper. While most of the temporal envelope filtering techniques require strong assumptions about MTF (some require MTF to be measured prior to the dereverberation stage, while others employ a theoretically derived MTF model with only a few parameters such as the reverberation time RT_{60}), the motivation for this work has been to develop a blind temporal envelope filtering method that only assumes the non-negativity of the subband envelopes of speech and a room impulse response, and the sparseness of speech. The problem is thus formulated as a blind deconvolution problem with non-negativity constraints, regularized by a sparsity cost. For its optimization we derive an iterative algorithm that ensures a monotonic decrease in the objective function and the non-negativity of the parameters.

2. PRINCIPLE

2.1. Subband envelope model of reverberant speech

Let $s_k[t]$, $h_k[t]$ be the subband signals of speech and a room impulse response at the kth subband where t is the time index. The reverberant subband signal is then approximated by

$$x_k[t] = \sum_{\tau} s_k[\tau] h_k[t-\tau], \tag{1}$$

particularly where each subband signal is obtained by STFT[8]. Then, the subband envelope can be written as

$$\begin{aligned} \left|x_k[t]\right|^2 &= \sum_{\tau}\sum_{\tau'} s_k^*[t-\tau] h_k^*[\tau] s_k[t-\tau'] h_k[\tau'] \qquad (2)\\ &= \sum_{\tau,\tau'} s_k^*[t-\tau] s_k[t-\tau'] \left|h_k[\tau]\right| \left|h_k[\tau']\right| e^{-j\phi_k[\tau]} e^{j\phi_k[\tau']}, \end{aligned}$$

where $h_k[\tau] = |h_k[\tau]| e^{j\phi_k[\tau]}$. As hypohesized in 1, the phase $\phi_k[\tau]$ can vary sensitively with respect to the reverberant conditions. We shall thus find it convenient to treat the phase as a

random variable and integrate it out of the model. Hence, by assuming that $\phi_k[t]$ is an independent random variable uniformly distributed on the interval $D = [-\pi, \pi)$, the expectation of $|x_k[t]|^2$ leads to

$$\mathbb{E}\big[|x_k[t]|^2\big] = \sum_{\tau} \big|s_k[t-\tau]\big|^2 \big|h_k[\tau]\big|^2, \tag{3}$$

which suggests that the subband envelope of reverberant speech can be represented, in an expectation sense, as the convolution of subband envelopes of clean speech and the room impulse response.

2.2. Problem setting

For simplicity of notation, let $S_k[t] \equiv |s_k[t]|^2$, $H_k[t] \equiv |h_k[t]|^2$. According to the discussion in 2.1, we model the kth subband envelope of reverberant speech as

$$X_k[t] \equiv \sum_{\tau} S_k[\tau] H_k[t-\tau], \tag{4}$$

and assume $\sum_t H_k[t] = 1$ in order to avoid an indeterminacy in the scaling. Given an observed subband envelope, $Y_k[t]$, the goal is to find an approximation such that $Y_k[t] \simeq X_k[t]$ in which $S_k[t]$ is sparse, based on the sparse nature of the speech spectrogram. It should be noted that the 'sparseness of speech' is the only assumption we make about clean speech. We now assume the following generative model

$$Y_k[t] = X_k[t] + \epsilon_k[t]. \tag{5}$$

The reconstruction error $\epsilon_k[t]$ is assumed to include any errors resulting from the approximations in 2.1. Assuming $\epsilon_k[t]$ is Gaussian white noise that follows $\mathcal{N}(0, \sigma^2)$, the likelihood of $S \equiv \{S_k[1], \cdots, S_k[T]\}_{k=1}^K$ and $H \equiv \{H_k[1], \cdots, H_k[T]\}_{k=1}^K$, given $Y \equiv \{Y_k[1], \cdots, Y_k[T]\}_{k=1}^K$, is written as

$$P(Y|S,H) = \prod_{k,t} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(Y_k[t] - X_k[t])^2}{2\sigma^2}\right). \tag{6}$$

We use a generalized Gaussian prior for $P(S)$

$$P(S) = \prod_{k,t} \frac{1}{2\Gamma\big(1+\frac{1}{p}\big)b} \exp\left(-\frac{|S_k[t]|^p}{b^p}\right), \tag{7}$$

and assume for convenience that $P(H_k[1], \cdots, H_k[T])$ is independent over k and follows a uniform distribution (or more accurately, a Dirichlet distribution whose parameters are all set at 1, as H is constrained to $\sum_t H_k[t] = 1$). When $0 < p < 2$, $P(S)$ becomes super-Gaussian and promotes sparsity if the norm of S is bounded. The likely values of S and H can thus be inferred from the posterior density

$$P(S,H|Y) \propto P(Y|S,H)P(S). \tag{8}$$

Regarding p, σ^2 and b as constant parameters, we arrive at the following objective function:

$$f(S,H) \equiv \sum_{k,t} \big(Y_k[t] - X_k[t]\big)^2 + 2\lambda \sum_{k,t} |S_k[t]|^p. \tag{9}$$

λ, determined by p, σ^2 and b, weighs the importance of the sparsity cost relative to the accurate reconstruction. We notice that, as S and H are non-negative in nature, we must seek to minimize $f(S,H)$ subject to $S_k[t] \geq 0$ and $H_k[t] \geq 0$. We are therefore led to solve the following optimization problem:

$$\begin{aligned} &\text{minimize} \quad f(S,H) \text{ with respect to } S \text{ and } H \\ &\text{subject to} \quad \sum_t H_k[t] = 1,\ H_k[t] \geq 0,\ S_k[t] \geq 0. \end{aligned} \tag{10}$$

2.3. Multiplicative update algorithm

Guided by the idea of NMF[18], we derive an efficient iterative algorithm that ensures a monotonic decrease (convergence to a stationary point) in the objective function and, simultaneously, the non-negativity of the parameters.

First we derive the update formula for S. Let S' and H' be the parameters at the previous iteration such that $S'_k[t] \geq 0$ and $H'_k[t] \geq 0$. We then have an inequality

$$\begin{aligned} f(S,H') \leq \sum_{k,t,\tau} \frac{S'_k[\tau]H'_k[t-\tau]}{X'_k[t]} \left(Y_{k,t} - \frac{S_k[\tau]}{S'_k[\tau]} X'_k[t]\right)^2 \\ + \sum_{k,t} \Big(pS'_k[t]^{p-2} S_k[t]^2 + 2|S'_k[t]|^p - p|S'_k[t]|^p\Big), \end{aligned} \tag{11}$$

where the equality holds when $S_k[t] = S'_k[t]$. Its proof is omitted owing space limitations. We write the right-hand side of the inequality as $\bar{f}(S)$. It can be proved that the minimization of $\bar{f}(S)$ w.r.t. $S_k[t]$ leads to a certain decrease of $f(S,H')$. Thus, differentiating $\bar{f}(S)$ partially w.r.t. $S_k[t]$ and setting it at 0, we obtain the update formula, which is often referred to as the 'multiplicative update rule', for $S_k[t]$

$$S_k[\tau] = S'_k[\tau] \frac{\sum_t H'_k[t-\tau] Y_k[t]}{\sum_t H'_k[t-\tau] X'_k[t] + \lambda p |S'{}_k[\tau]|^{p-1}}, \tag{12}$$

where

$$X'_k[t] = \sum_{\tau} S'_k[\tau] H'_k[t-\tau]. \tag{13}$$

We note that Eq. (12) comprises the product and sum of non-negative entities, and so the non-negativity of the parameter update is thus guaranteed. Next, we derive the update formula for H. Similarly, we have an inequality

$$\begin{aligned} f(S',H) \leq \sum_{k,t,\tau} \frac{S'_k[t-\tau]H'_k[\tau]}{X'_k[t]} \left(Y_k[t] - \frac{H_k[\tau]}{H'_k[\tau]} X'_k[t]\right)^2 \\ + 2\lambda \sum_{k,t} |S'_k[t]|^p. \end{aligned} \tag{14}$$

In the same way, differentiating the right-hand side of this inequality partially w.r.t. $H_k[t]$ and setting it at 0, we obtain

the update formula for $H_k[t]$

$$H_k[\tau] = H'_k[\tau] \frac{\sum_t S'_k[t-\tau] Y_k[t]}{\sum_t S'_k[t-\tau] X'_k[t]}. \quad (15)$$

Note, however, that as we did not take into account the constraint $\sum_t H_k[t] = 1$ in the above, a convenient update procedure for $H_k[t]$ shall consist of computing Eq. (15) and then projecting it onto the constraint space, that is, $H_k[t] \leftarrow H_k[t] / \sum_{t'} H_k[t']$.

2.4. Modulation frequency domain processing

As can be seen from an inspection of Eqs. (12), (15) there is a need to compute the convolution of S' and H', and the cross-correlations of H' and Y, H' and X', S' and Y, and S' and X'. Fortunately, they can be computed very efficiently using the Fast Fourier Transform (FFT) as described below. Let $\mathscr{F}$ be the Fourier transform operator such that $\mathcal{S}'_k[u] = \mathscr{F}\{S'_k[t]\}_u$, $\mathcal{H}'_k[u] = \mathscr{F}\{H'_k[t]\}_u$ and $\mathcal{Y}_k[u] = \mathscr{F}\{Y_k[t]\}_u$, where u corresponds to the modulation frequency index and $\mathcal{H}'_k[u]$ is therefore nothing else than the modulation transfer function (MTF). Then, the update formulae for H and S can be rewritten as

$$\begin{aligned} H_k[t] &= H'_k[t] \frac{\mathscr{F}^{-1}\{\mathscr{F}\{S'_k[t]\}^*_u \mathscr{F}\{Y_k[t]\}_u\}_t}{\mathscr{F}^{-1}\{\mathscr{F}\{S'_k[t]\}^*_u \mathscr{F}\{X'_k[t]\}_u\}_t} \\ &= H'_k[t] \frac{\mathscr{F}^{-1}\{\mathscr{F}\{S'_k[t]\}^*_u \mathscr{F}\{Y_k[t]\}_u\}_t}{\mathscr{F}^{-1}\{\mathscr{F}\{S'_k[t]\}^*_u \mathscr{F}\{S'_k[t]\}_u \mathscr{F}\{H'_k[t]\}_u\}_t} \\ &= H'_k[t] \frac{\mathscr{F}^{-1}\{\mathcal{S}'^*_k[u] \mathcal{Y}_k[u]\}_t}{\mathscr{F}^{-1}\{|\mathcal{S}'_k[u]|^2 \mathcal{H}'_k[u]\}_t}, \quad (16) \end{aligned}$$

$$S_k[t] = S'_k[t] \frac{\mathscr{F}^{-1}\{\mathcal{H}'^*_k[u] \mathcal{Y}_k[u]\}_t}{\mathscr{F}^{-1}\{\mathcal{S}'_k[u] |\mathcal{H}'_k[u]|^2\}_t + \lambda p |S'_k[t]|^{p-1}}, \quad (17)$$

from which we see that the computation of the multiplicative factors can be performed in the modulation frequency domain, which is fast and also easy to implement.

2.5. Interpretation as 'Diagonal' Sparse NMFD

The algorithm presented here turns out to be a special case of the non-negative matrix factor deconvolution (NMFD)[19] when $\lambda = 0$. While NMF uses a model of the form $\boldsymbol{Y} \simeq \boldsymbol{W}\boldsymbol{U}$, NMFD employs an extended model

$$\boldsymbol{Y} \simeq \sum_{j=1}^{J} \boldsymbol{W}_j \overset{j\rightarrow}{\boldsymbol{U}}, \quad (18)$$

where $\boldsymbol{Y}$ is the matrix to be decomposed, and $\boldsymbol{W}_t$ and $\boldsymbol{H}$ are the bases and weight matrices. The $j\rightarrow$ operator shifts the columns of its argument by $j-1$ positions to the right, i.e.,

$$\boldsymbol{A} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}, \quad \overset{1\rightarrow}{\boldsymbol{A}} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix} \quad (19)$$

$$\overset{2\rightarrow}{\boldsymbol{A}} = \begin{pmatrix} 0 & 1 & 2 & 3 \\ 0 & 5 & 6 & 7 \end{pmatrix}, \quad \overset{3\rightarrow}{\boldsymbol{A}} = \begin{pmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 5 & 6 \end{pmatrix}. \quad (20)$$

We will write Eq. (4) using the notation defined above. Let $\boldsymbol{H}_t$ be the diagonal matrix whose entries of the main diagonal are $H_1[t], H_2[t], \cdots, H_K[t]$ such that

$$\boldsymbol{H}_t \equiv \begin{pmatrix} H_1[t] & & 0 \\ & \ddots & \\ 0 & & H_K[t] \end{pmatrix}, \quad (21)$$

and $\boldsymbol{S}$ be the 'spectrogram' matrix of clean speech such that

$$\boldsymbol{S} \equiv \begin{pmatrix} S_1[1] & \cdots & S_1[T] \\ \vdots & & \vdots \\ S_K[1] & \cdots & S_K[T] \end{pmatrix}. \quad (22)$$

Eq. (4) can then be written in a matrix notation

$$\boldsymbol{X} = \sum_{t=1}^{T} \boldsymbol{H}_t \overset{t\rightarrow}{\boldsymbol{S}}, \quad (23)$$

which has the same form as the NMFD model. This method can thus be explained as a particular case of NMFD where the bases matrices $\boldsymbol{H}_t$ are constrained to diagonal matrices.

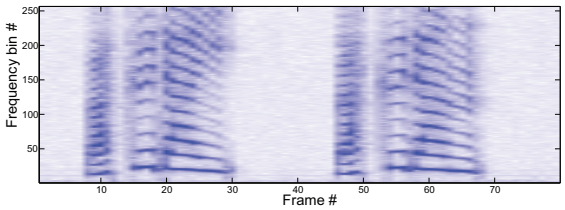
By analogy with NMFD, the temporal envelope of the spectrogram blurring effect caused by reverberation corresponds to the temporal evolution of the basis component and interestingly, the clean speech spectrogram corresponds to the activation matrix. The reverberant spectrogram is thus considered to comprise blurring envelopes 'activated' by the components of a clean speech spectrogram.

3. EXPERIMENTS

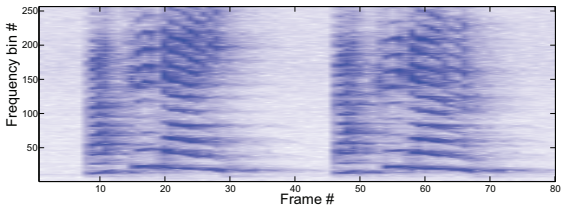
In this section we report some results for speech data recorded under several different reverberant conditions. All the speech data were monaural and sampled at 16kHz. The STFT was computed using a Hanning window that was 64ms long with a 32ms overlap. p and λ were set at $p = 1.2$ and $\lambda = E^{2-p}$ where $E = \sum_{k,t} Y_k[t] \times 10^{-8}$. The algorithm was run for 20 iterations. $S_k[t]$ was initially set equal to $Y_k[t]$. $H_k[t]$ were initially set at a decaying exponential envelope. The final reconstruction of the speech waveform was performed using the original phase function of the observation.

For the first experiment we tested our method on a synthesized reverberant speech signal. The test data were created by convolving a clean speech signal from a female speaker, excerpted from the ATR speech database, with a room impulse response, measured in a room with an RT_{60} of 0.5s. Fig. 1 (a), (b) show the spectrograms of the clean speech signal and the test data, respectively. The spectrograms of the dereverberated signal obtained with the conventional and present methods can be seen in Fig. 1 (c), (d). Here and subsequently, when we refer to the conventional method, we are referring to the single channel version of [8]. We used signal-to-noise ratio (SNR) as a quantitative measure of the dereverberation performance. As a result, the present method improved the SNR from 2.30 to 2.94dB, while the conventional method was only able to improve it to 2.53dB.

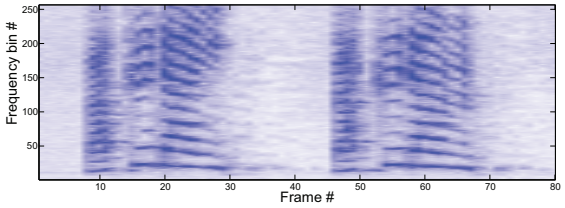
For the second exeperiment we used a speech signal recorded in a reverberant room. The clean speech signal was



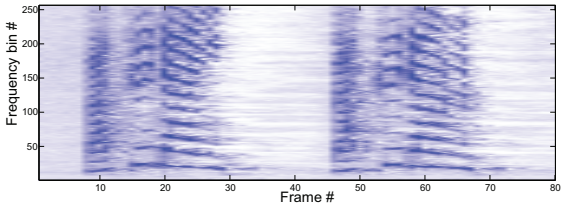
(a) Clean speech (female speaker)



(b) Synthesized reverberant speech (RT_{60} =0.5s)

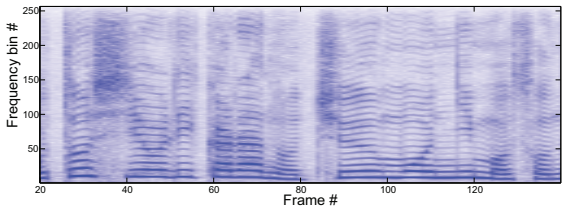


(c) Dereverberated signal obtained with the 1ch version of [8]

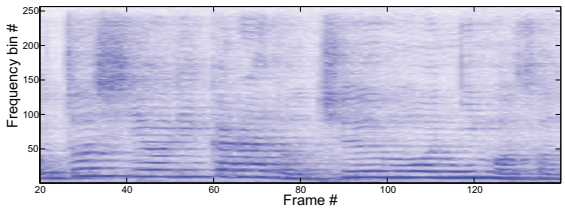


(d) Dereverberated signal obtained with the present method

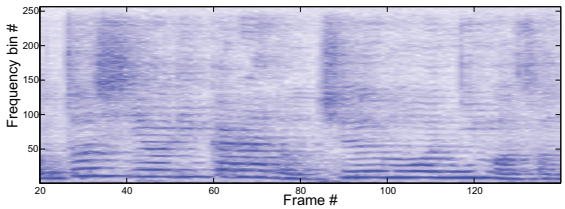
Fig. 1. Test on synthesized reverberant speech data



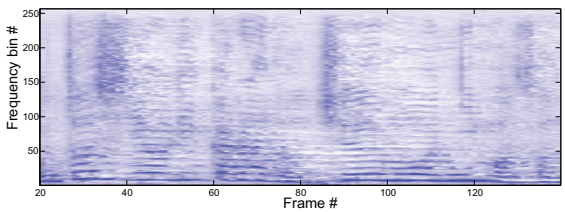
(a) Clean speech data (female speaker)



(b) Speech data recorded by a distant microphone in a reverberant room



(c) Dereverberated signal obtained with the 1ch version of [8]



(d) Dereverberated signal obtained with the present method

Fig. 2. Test on synthesized reverberant speech data

played by a loudspeaker and recorded by a distant microphone. The loudspeaker was carried by a participant, who was asked to move around the room. Fig. 2 (a), (b) show the spectrograms of the clean speech signal and the test data, respectively. The spectrograms of the dereverberated signal obtained with the conventional and present methods can be seen in Fig. 1 (c), (d). The present method improved the SNR from 1.29 to 1.68dB, while the conventional method was only able to improve it to 1.52dB. The workstation used to perform the experiments had a Core 2 Duo processor with a 2.66GHz clock speed and a 1.99GB memory. The algorithm was implemented in Matlab on a Windows platform. The algorithm usually converged within fewer than 20 iteration cycles at a real time factor of around 1/3.

4. CONCLUDING REMARKS

In this paper we developed a new dereverberation method designed to recover the subband envelope of an original speech signal from its reverberant version. The discussions in this paper mainly focused on the validity of the subband envelope model of reverberant speech, problem setting based on the non-negativity and the sparse nature of speech spectrograms leading to a regularized optimization problem with non-negativity constraints, and an iterative algorithm for its optimization, which can be seen as a special case of NMFD. The algorithm presented in this paper is fast, easy to implement, and robust to speaker movement.

5. REFERENCES

[1] B. Yegnanarayana, Proc. ICASSP'98, 1, 405–408, 1998.
[2] M.S. Brandstein, Proc. ICASSP'98, 6, 3613–3616, 1998.
[3] B. W. Gillespie et al., ICASSP'01, 6, 3701–3704, 2001.
[4] T. Nakatani et al., IEEE Trans. ASLP, 15(1), 80–95, 2007.
[5] K. Kinoshita et al., Proc. ICASSP'06, 1, 817–820, 2006.
[6] T. Yoshioka et al., Proc. IWAENC'06, 2006.
[7] T. Nakatani et al., Proc. ICASSP'07, 193–197, 2007.
[8] T. Nakatani et al., Proc. ICASSP'08, 85–88, 2008.
[9] T. Yoshioka et al., Proc. ICASSP'08, 4585–4588, 2008.
[10] T. Langhans et al., ICASSP'82, 156–159, 1982.
[11] J. Mourjopoulos et al., ICASSP'83, 1144–1147, 1983.
[12] H. G. Hirsch, EURASIP, 1177–1180, 1988.
[13] H. Hermansky et al., ICASSP'95, 405–408, 1995.
[14] C. Avendano et al., Proc. ICSLP'96, 2, 889–892, 1996.
[15] S. Hirobayashi et al., IEICE Trans. A, J81A(10), 1323–1330, 1998.
[16] M. Unoki et al., ICASSP'03, 1, 840–843, 2003.
[17] E. A. P. Habets, Proc. ProRISC'04, 250–254, 2004.
[18] D. D. Lee et al., Proc. NIPS'00, 556–562, 2000.
[19] P. Smaragdis et al., Proc. ICA'04, 494–499, 2004.