A BLIND SUBBAND-BASED DEREVERBERATION ALGORITHM

Oldooz Hazrati, Kostas Kokkinakis and Philipos C. Loizou

Center for Robust Speech Systems,
Department of Electrical Engineering, University of Texas at Dallas,
800 West Campbell Road, Richardson, TX 75080–3021, USA
hazrati@student.utdallas.edu, kokkinak@utdallas.edu, loizou@utdallas.edu

ABSTRACT

In this work, we develop a two-stage blind dereverberation scheme for reverberant speech enhancement. The proposed algorithm operates by first splitting the reverberant inputs into different subbands. In the first stage, the inverse filters are estimated using the blind multiple input-output inverse-filtering theorem (MINT), while in the second stage our approach suppresses late reverberant energy by subtracting the power spectrum of the late impulse components from the power spectrum of the inverse filtered speech signal. Experimental results with reverberant speech indicate that the performance of the proposed algorithm is substantially better when compared to existing dereverberation algorithms.

Index Terms— Blind dereverberation, psychoacoustic model, spectral subtraction, perceptual evaluation of speech quality (PESQ).

1. INTRODUCTION

Reverberation or the collection of reflected sounds from surfaces in an acoustic enclosure leads to temporal and spectral smearing, which distorts both the envelope and fine structure of speech [1]. Tackling speech degradation due to reverberation has recently become an area of intense research activity and has given rise to several speech dereverberation algorithms (e.g., see [2]–[5]). The most challenging issue when dealing with speech dereverberation is estimating the unknown room impulse responses (RIRs) between the speaker and the microphone(s). This essentially makes dereverberation a “blind” problem. The goal of blind dereverberation is to retrieve the (unknown) input speech signal and ideally the relevant features of the unknown RIRs, while only assuming knowledge of the microphone outputs, as well as some relevant information on the statistical properties of the original source [6].

The MINT method which is based on the multiple input-output inverse-filtering theorem (e.g., see [3]), remains one of the most efficient dereverberation algorithms to-date. The MINT has been found to perform near perfect towards finding accurate estimates of inverse impulse responses even in non-minimum phase reverberant environments
1. However, the main drawback of the MINT is that it requires explicit knowledge of the original impulse responses, making its use impractical in real-world reverberant settings, where information regarding the original RIRs may not be readily available. In addition, the MINT is only useful when background noise is weak or well-controlled because it is very sensitive to even small errors in the estimated channel impulse. To address these concerns, Furuya and Kataoka [4] have recently proposed a more robust speech dereverberation method using a blind deconvolution MINT-based inverse-filtering approach for early reverberation cancellation and spectral subtraction for late reverberation compensation. To overcome inaccuracies arising from the inverse filtering-based blind deconvolution stage, as well as tail fluctuations of real-world impulse responses, the proposed speech dereverberation technique utilizes a second spectral subtraction stage to minimize the influence of long-term reverberation [4]. Spectral subtraction methods can suppress late reverberant energy by estimating the power spectrum of clean speech after subtracting the power spectrum of reverberation from that of the reverberant speech (e.g., see [7, 8]).

In this paper, we propose a subband MINT-based blind deconvolution method for speech dereverberation. Motivated by a perceptual model that closely approximates auditory frequency selectivity, the reverberant speech is first passed through a gammatone filterbank [9]. The MINT-based blind deconvolution method proposed in [4] is then applied separately to each subband. The proposed subband inverse-filtering technique is based on the correlation of the input (reverberant) signals in different bands. This leads to a considerably more accurate inverse filter estimation, when compared against the original MINT-based blind deconvolution method, which uses the whole spectrum for correlation estimation. We therefore show that a subband-based speech dereverberation approach can compensate for the inverse filtering errors arising from traditional speech dereverberation algorithms. Furthermore, a spectral subtraction method relying on an asymmetric smoothing function is used for late reverberation reduction [8]. The performance of the proposed blind single-input two-output algorithm for reverberant speech enhancement is assessed through experiments in a two-microphone configuration using the perceptual evaluation of speech quality (PESQ) metric.

2. ALGORITHM FORMULATION

In this section, we analyze the blind deconvolution method based on the MINT inverse-filtering theorem for a single-input and two-microphone system configuration. Next, the proposed perceptually-motivated subband blind dereverberation method is presented. This is followed by a brief description of the spectral subtraction method used for late reverberation suppression.

---

---

---

---

---

---
2.1. Blind deconvolution based on MINT

In the conventional MINT algorithm [3], the original RIRs are used to compute the inverse filters. Since the RIRs are unknown, the MINT-based blind inverse-filtering method was proposed to estimate the inverse filters by the use of correlation matrix of received signals which contains useful information about the impulse responses [4]. The inverse filters in the MINT-based blind deconvolution method are estimated by solving the equation

\[ b = R h \]  

where \( b = [1, 0, 0, \ldots, 0]^T \) is the \( NL \times 1 \) target vector and \( h = [h^{(1)}_1, h^{(2)}_1]^T \) is the \( NL \times 1 \) vector of the estimated inverse filters, defined as

\[ h_j = [h_j(0), h_j(1), \ldots, h_j(L-1)]^T \]  

where \( j = 1, 2 \) is the microphone index, \( N \) denotes the number of microphones and \( L \) is the length of the estimated inverse filters. The correlation matrix of the signals observed at the system microphones is given by

\[ R = E \{ x(k) x^T(k) \} \]  

where \( E[\cdot] \) represents the expectation operator, \( k \) is the sample index and vector

\[ x(k) = [x^1(k), x^2(k)]^T \]  

of size \( NL \times 1 \) denotes the received signals, defined as

\[ x_j(k) = [x_j(k), x_j(k-1), \ldots, x_j(k-L+1)]^T \]  

Assuming that the source signal is stationary and statistically white, the relationship between \( R \) and the \( NL \times NL \) matrix of the original impulse responses \( G = [G_1, G_2] \) is given by

\[ R = G^T G \]  

where the \( j \)th sub-matrix of \( G \) is given by

\[ G_j = \begin{pmatrix} g_j(0) & 0 & \ldots & 0 \\ g_j(1) & g_j(0) & \ldots & \vdots \\ \vdots & g_j(1) & \ddots & 0 \\ 0 & g_j(K-1) & \ddots & g_j(0) \\ 0 & 0 & \ddots & g_j(K-1) \\ \end{pmatrix} \]  

with \( K \) denoting the length of the RIRs. Since in practice the original speech source is not statistically white, we need to remove its temporal correlation before any further processing. The reverberant signals are first whitened and the correlation matrix shown in (3) is computed for the whitened signals. The inverse filters are then estimated by using this correlation matrix and solving (1).

![Fig. 1. Magnitude responses of a 10-channel gammatone filterbank in the 50–4,000 Hz range.](image-url)

2.2. Perceptually-motivated subband blind dereverberation

Reverberation increases the prominence of low-frequency energy that masks the speech spectrum by filling the dips in the temporal envelope of speech [1]. Thus, different frequency bands are affected differently by reverberation. Similarly to multi-band noise reduction techniques, blind deconvolution may be more effective if performed in a multi-band manner. The gammatone auditory filterbank can provide a fairly accurate perceptual model of the basilar membrane movement in the ear and also a much better characterization of the human auditory system. Hence, to obtain a better performance out of the MINT-based blind deconvolution method, we resort to a gammatone filterbank as a preprocessing stage before inverse filtering.

By utilizing a subband filtering approach, we can also estimate the inverse impulse responses more accurately and consequently achieve a more reliable estimation of the original clean speech signal. As depicted in Fig. 1, each filter in the filterbank is essentially a bandpass filter whose impulse response is the product of a gamma function and a tone and represents the frequency response associated with a particular point on the basilar membrane of the cochlea as a function of time. A gammatone filter is defined in the time-domain by its impulse response, written as [9]

\[ z(k) = \delta k^{r-1} \exp(-2\pi kb(\nu)) \cos(2\pi F_c k + \theta) \]  

where \( \tau \) denotes the filter order, \( \delta \) defines the amplitude of the response, \( b(\nu) \) is the bandwidth defined in Hz, \( F_c \) is the center frequency and \( \theta \) denotes phase. The subband center frequencies of each filter are set according to measurements of the equivalent rectangular bandwidth (ERB) of the human auditory filter and are quasi-logarithmically spaced proportional to their bandwidths from 50–4,000 Hz. In turn, each ERB band is calculated using

\[ \text{ERB} = 24.7 + 0.108 F_c \]  

where \( F_c \) is the center frequency in Hz. After the signals are split into different subbands, the blind deconvolution MINT-based inverse filtering approach is performed on each subband by estimating the correlation of the whitened versions of the microphone signals. The inverse filters are estimated separately in each subband by solving

\[ h_m = R^{-1}_m b_m \]
where $R_m$ indicates the correlation matrix in the $m$th subband and $m = 0, 1, \ldots, M - 1$ denotes the subband index. After solving (10) for each subband, the signals in each subband are inverse filtered and added together, yielding the clean signal

$$\tilde{x}_m = \sum_{j=1}^{2} h_{m,j} \ast x_{m,j}$$

(11)

where $\ast$ defines convolution in the time-domain. In a similar manner to [10], the resulting output of each subband is time-reversed, passed through the gammatone filter, time-reversed again, and then summed across all subbands to obtain the final (enhanced) speech signal.

### 2.3. Spectral subtraction

In order to compensate for the inaccuracies arising from the blind deconvolution stage and to further eliminate the tail fluctuations of the impulse responses, a spectral subtraction method is used in the third stage of our dereverberation approach. Here, to suppress the late reverberant energy, the power spectrum of the late impulse components is subtracted from the power spectrum of the inverse filtered speech signal. The signal power spectrum of the clean speech signal can be estimated as in [8]

$$|S_u(t, \ell)|^2 = |S_{\tilde{x}}(t, \ell)|^2 \cdot$$

$$\max \left( \frac{|S_{\tilde{x}}(t, \ell) - \gamma w(t-\rho) \ast |S_{\tilde{x}}(t, \ell)|^2}{|S_{\tilde{x}}(t, \ell)|^2}, \varepsilon \right)$$

(12)

where $t$ and $\ell$ are the frame and frequency bin indices, respectively, and $|S_{\tilde{x}}(t, \ell)|^2$ and $|S_u(t, \ell)|^2$ are the power spectra of the subband inverse filtered signal and the estimated clean signal after the spectral subtraction stage. In order to approximate the coefficients of the late impulse response components, we apply an asymmetrical smoothing function resembling the Rayleigh distribution. This smoothing function has the form

$$w(t) = \begin{cases} \frac{t + a}{a^2} \exp \left( -\frac{(t + a)^2}{2a^2} \right), & \text{if } t > -a \\ 0, \text{ otherwise} \end{cases}$$

(13)

where $\gamma$ is the scaling factor, $\rho$ is the relative delay of the late impulse response components, $a$ denotes the control parameter and $\varepsilon$ is the maximum attenuation floor.

### 3. EXPERIMENTAL RESULTS

#### 3.1. Speech material

The performance of the proposed blind speech dereverberation algorithm is evaluated on a test set of 10 speech signals comprised of 5 randomly selected sentences each. The duration of each speech set (A–J) is approximately 15 s. All signals are recorded at a sampling rate of 8 kHz. To create the speech test set we used the IEEE database, which consists of phonetically balanced sentences, with each sentence being composed of approximately 7 to 12 words [11]. All signals have the same onset and are normalized to their maximum amplitude. The reverberant speech signals are generated by convolving the clean speech signal with synthetically generated RIRs using the image method for a $6.60 \times 4.60 \times 3.10$ m room (e.g., see [12]). The distance between the single source signal and the two microphones is 4.80 m and the two microphones are positioned around 10 cm apart.

#### 3.2. Algorithm parameters

In the whitening stage, the inverse of the autoregressive vocal tract filter is estimated by linear prediction coding (LPC) analysis of the reverberant speech signal. Prior studies have suggested that such a whitening strategy can lead to a better speech correlation cancelation when compared to just computing the whitening filter based on the inverse of the long-term average of the speech spectrum (e.g., see [13]). The whitening is performed on non-overlapping frames of 25 ms duration corresponding to 200 samples at 8 kHz and the whitening filter is obtained using 10th-order LPC analysis.

In addition, a gammatone filterbank of 128 4th-order filters is used to divide the frequency range of 50–4000 Hz to subbands similar to the human auditory system, which uses filters with quasi-logarithmically spaced center frequencies. The correlation matrix is estimated for approximately 16,000 sample blocks and is averaged with a moving average factor of 0.9 for each set. The reverberation time of the room is $T_{60} = 500$ ms in the 20–4000 Hz frequency band, which is a typical value for a severely reverberant environment. The length of the RIRs and the estimated inverse filters are approximately 1,500$^2$ and 700 taps, respectively. The parameters $a$, $\gamma$, $\rho$ and $\varepsilon$ used in the spectral subtraction stage are set to 5, 0.3, 7 and 0.03, respectively.

#### 3.3. Performance evaluation

The overall quality of the enhanced signals is assessed with the perceptual evaluation of speech quality (PESQ) score [14]. The PESQ employs a sensory model to compare the original (unprocessed) with the enhanced (processed) signal, which is the output of the dereverberation algorithm, by relying on a perceptual model of the human auditory system. The PESQ score has been shown to exhibit a high correlation coefficient (Pearson’s correlation) of $r = 0.91$ with subjective listening quality tests [15]. The PESQ measures the subjective assessment quality of the dereverberated speech rated as a value between 1 and 5 according to the five grade mean opinion score (MOS) scale. Here, we use a modified PESQ measure referred to as $m$PESQ, with parameters optimized towards assessing speech signal distortion, calculated as a linear combination of the average disturbance value $D_{\text{ind}}$ and the average asymmetrical disturbance values $A_{\text{ind}}$ [14, 15]

$$m\text{PESQ} = a_0 + a_1 D_{\text{ind}} + a_2 A_{\text{ind}}$$

(14)

such that

$$a_0 = 4.959, \ a_1 = -0.191 \text{ and } a_2 = -0.006$$

(15)

By definition, a high value of $m$PESQ indicates low speech signal distortion, whereas a low value suggests high distortion with considerable degradation present. The $m$PESQ score is presumed to be inversely proportional to reverberation time and is expected to increase as reverberant energy decreases (e.g., see [5]).

---

2For computational simplicity, 4,000-tap RIRs corresponding to $T_{60} = 500$ ms are truncated to 1,500 taps.
### Table 1. Speech dereverberation performance of the SBD+SS method in terms of $m$PESQ values averaged across both microphones.

<table>
<thead>
<tr>
<th>TEST SET</th>
<th>INPUT</th>
<th>BD</th>
<th>SBD</th>
<th>SBD+SS</th>
<th>MINT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2.56</td>
<td>2.96</td>
<td>3.18</td>
<td>3.30</td>
<td>4.95</td>
</tr>
<tr>
<td>B</td>
<td>2.55</td>
<td>2.91</td>
<td>3.20</td>
<td>3.33</td>
<td>4.95</td>
</tr>
<tr>
<td>C</td>
<td>2.58</td>
<td>2.99</td>
<td>3.18</td>
<td>3.31</td>
<td>4.92</td>
</tr>
<tr>
<td>D</td>
<td>2.53</td>
<td>2.78</td>
<td>2.99</td>
<td>3.21</td>
<td>4.89</td>
</tr>
<tr>
<td>E</td>
<td>2.49</td>
<td>2.79</td>
<td>2.95</td>
<td>3.14</td>
<td>4.94</td>
</tr>
<tr>
<td>F</td>
<td>2.59</td>
<td>2.83</td>
<td>2.97</td>
<td>3.23</td>
<td>4.90</td>
</tr>
<tr>
<td>G</td>
<td>2.61</td>
<td>2.87</td>
<td>3.09</td>
<td>3.29</td>
<td>4.95</td>
</tr>
<tr>
<td>H</td>
<td>2.58</td>
<td>2.88</td>
<td>3.07</td>
<td>3.38</td>
<td>4.88</td>
</tr>
<tr>
<td>I</td>
<td>2.54</td>
<td>2.90</td>
<td>2.99</td>
<td>3.20</td>
<td>4.91</td>
</tr>
<tr>
<td>J</td>
<td>2.63</td>
<td>2.80</td>
<td>3.09</td>
<td>3.31</td>
<td>4.94</td>
</tr>
</tbody>
</table>

Mean $±$ STD: $2.57 ± 0.04$ $2.87 ± 0.07$ $3.07 ± 0.09$ $3.27 ± 0.07$ $4.92 ± 0.03$

### 3.4. Discussion

Table 1 shows the performance of the proposed method in terms of the $m$PESQ scores calculated from the test set created from the IEEE database. Results for the proposed subband-based blind dereverberation inverse-filtering technique without spectral subtraction (SBD) and with the spectral subtraction stage (SBD+SS) are presented separately to better demonstrate the impact of each stage in the entire dereverberation process. The $m$PESQ scores obtained from the blind deconvolution-based MINT inverse-filtering approach (BD) [4] and the conventional (non-blind) MINT method [3] are also reported to provide a benchmark for assessing the performance of the proposed method.

The average improvement in $m$PESQ is approximately 0.7 when comparing SBD+SS to the baseline input $m$PESQ, which indicates a relatively high improvement in speech quality. Although the conventional MINT performs substantially better, the proposed method still provides an adequate improvement by blindly estimating the RIRs with only about half of the length of the original RIRs. By comparing SBD to BD, it is evident that the proposed method improves the BD approach by making use of different subbands in a manner similar to human auditory frequency selectivity. This leads to a better channel correlation estimation and consequently more accurate inverse filters. In addition, a more precise estimation of the inverse filter coefficients yields fewer errors, thus enabling the spectral subtraction to suppress the late reverberation more effectively. The improvement obtained solely due to the spectral subtraction stage is on average 0.2 (SBD to SBD+SS).

### 5. REFERENCES


