Speech Enhancement Using Masking Properties in Adverse Environments

Atanu Saha\(^1\), and Tetsuya Shimamura\(^1\)

Graduate School of Science and Engineering, Saitama University, Saitama, Japan
\{saha\(^1\),shima\(^1\)}@eie.ics.saitama-u.ac.jp

Abstract

In this paper, we propose a speech enhancement method by exploiting masking properties of human auditory system. The masking properties are exploited to calculate a masking threshold. The spectral components which lie above the threshold are audible to human listeners. These audible spectral components in the proposed method are suppressed as a predefined attenuation factor of the original noise. The evaluation is conducted in the experiments. The experimental results show that the proposed method provides significant performance compared to the conventional approaches.

Index Terms: Speech enhancement, masking properties, audible noise, adverse environment

1. Introduction

Since the performance of speech processing systems deteriorates under the presence of additive background noise, it is essential for these systems to rely on effective speech enhancement techniques. The motivation behind the enhancement techniques is to enhance one or more perceptual aspects of speech, such as speech quality and intelligibility while extracting the desired signal from its corrupted observations. Though research on speech enhancement has been going on so far for a long time, the optimization between the perceptual aspects remains unsolved. Several algorithms have been proposed to address this problem with varying degrees of success.

The MMSE estimator [1], which minimizes the mean-square error of the short-time spectral magnitudes, has received a lot of attention among speech enhancement algorithms. An improvement to the estimator [1] was reported in another approach [2], called log-MMSE. The log-MMSE estimator, which minimizes the mean-square error of the log-magnitude spectra, is more suitable in reducing musical noise. A further modification has been conducted by incorporating signal presence uncertainty in [3]. For multiplicative modification, Cohen [4] also proposed an optimally modified log-MMSE estimator, which takes a lower bound threshold into account. However, most of the algorithms mentioned above are not perceptually motivated and do not take intelligibility into account.

In an effort to make perceptually motivated, the phenomenon of auditory masking has been exploited in many speech enhancement algorithms. The rationale behind the use of the masking properties to speech enhancement, in general, is to calculate a masking threshold, below which all spectral components are inaudible and so perceptually are not important components to suppress. The goal of masking-based enhancement algorithms, then, is to minimize only the audible portion of the noise components for making perceptual to human listeners. In the perceptual sense, Virag [5] proposed an enhancement technique that incorporates an auditory model into a generalized spectral subtraction process. In the enhancement process, the subtractive parameters were adapted based on a masking threshold. However, the optimal estimation of speech signal cannot be obtained due to imperfections in the mathematical model of generalized spectral subtraction. In contrast, masking-based $\beta$-order MMSE was proposed in [6]. The value of $\beta$ in [6] was made to be a function of frame SNR (signal-to-noise ratio) and masking threshold. The function for $\beta$-value was set by an empirical formula with a few parameters, which are extremely difficult to choose for all speech and noise conditions. In the audible noise suppression approach proposed by Tsoukalas et al. [7], the frequency masking properties were used to extract some information on audible noise before performing a nonlinear spectrum filtering. Based on audible noise information, the spectrum of noisy speech was filtered by a parametric Wiener-type function. An evaluation of the speech processed by the approach [7] was reported in [8]. Small improvements were observed in some but not all of the conditions tested. This is mainly due to distortions of the signal, resulting from the exaggerated suppression of acoustic noise, particularly at very low SNRs. On the other hand, in order to preserve the background noise characteristics, a constrained minimization approach was proposed in [9] through a predefined fixed attenuation factor of original noise. Although this gives a promising result, improvements are expected by using the predefined noise level to be a function of time and frequency, since the noise characteristics are not stationary at all frequencies and at all times.

The objective of this paper is to develop a speech enhancement technique so as to modify the predefined attenuation factor to be a function of audible noise. The audible noise is estimated by exploiting masking properties of human auditory system. As opposed to the approach in [7], the audible noise in the proposed technique is suppressed in a direct manner. This makes the processed speech produced by the proposed technique to be more pleasant without introducing significant speech distortion.

The outline of this paper is as follows. Section 2 presents some preliminary notations. Section 3 describes the masking-based proposed method. We then evaluate the experimental results in Section 4. Finally, Section 5 concludes the paper.

2. Preliminary Notations

In this section, we introduce the notations that we will use throughout this paper.

Let us assume that noisy speech $x(n)$ at sampling time index $n$ consists of clean speech $s(n)$ and additive noise $d(n)$. The noisy signal in the time domain, $x(n) = s(n) + d(n)$, is transformed into the frequency domain by the application of a window function and analyzed using the discrete Fourier transform (DFT). In the frequency domain, we have

$$X(\lambda, k) = S(\lambda, k) + D(\lambda, k)$$  \hspace{1cm} (1)

where $X(\lambda, k)$, $S(\lambda, k)$ and $D(\lambda, k)$ are the DFT coefficients...
at frequency bin $k$ and frame index $\lambda$ of the noisy speech, clean speech and noise, respectively. The power spectra of $X(\lambda,k)$, $S(\lambda,k)$ and $D(\lambda,k)$ are obtained by squaring their respective magnitude spectra and are denoted as $P_x(\lambda,k)$, $P_s(\lambda,k)$ and $P_d(\lambda,k)$, respectively. Based on the noisy speech signal, an estimate of the clean speech spectrum is computed as

$$\hat{S}(\lambda,k) = G(\lambda,k)X(\lambda,k)$$

(2)

where $G(\lambda,k)$ is the gain function. The proposed method is devoted to deriving the gain function $G(\lambda,k)$.

3. Masking-Based Proposed Speech Enhancement Method

In this section, we first derive the gain function $G(\lambda,k)$ and estimate its corresponding parameters, and then we describe the suppression characteristics for the derivation of $G(\lambda,k)$ under the influence of the parameters.

3.1. Gain Function Derivation

The gain function in the proposed algorithm is determined based on a simplified constrained minimization approach. The constrained minimization approach minimizes an error criterion that considers the speech and noise distortions as constraints. The error spectrum between the estimated spectrum and the desired spectrum can be expressed as

$$\varepsilon(\lambda,k) = \hat{S}(\lambda,k) - S_d(\lambda,k)$$

(3)

where $S_d(\lambda,k)$ is the desired signal spectrum defined as

$$S_d(\lambda,k) = S(\lambda,k) + A_d(\lambda,k)D(\lambda,k).$$

(4)

In (4), $A_d(\lambda,k)$ is the predefined attenuation factor which controls a noise level according to its degree of freedom. In our implementation, we set the attenuation factor as audible spectral components. Substituting (2) and (4) into (3), we obtain the error spectrum as

$$\varepsilon(\lambda,k) = G(\lambda,k)X(\lambda,k) - S(\lambda,k) - A_d(\lambda,k)D(\lambda,k)$$

$$= G(\lambda,k)(S(\lambda,k) + D(\lambda,k)) - S(\lambda,k) - A_d(\lambda,k)D(\lambda,k)$$

$$= (G(\lambda,k) - 1)S(\lambda,k) + (G(\lambda,k) - A_d(\lambda,k))D(\lambda,k)$$

$$= \varepsilon_s(\lambda,k) + \varepsilon_d(\lambda,k)$$

(5)

where $\varepsilon_s(\lambda,k)$ and $\varepsilon_d(\lambda,k)$ represent the spectra of speech and noise distortions, respectively. The power spectral density of the error can be expressed as the sum of two components, i.e.,

$$\varepsilon^2(\lambda,k) = \varepsilon^2_s(\lambda,k) + \varepsilon^2_d(\lambda,k)$$

(6)

where

$$\varepsilon^2_s(\lambda,k) = (G(\lambda,k) - 1)^2 P_s(\lambda,k)$$

(7)

$$\varepsilon^2_d(\lambda,k) = (G(\lambda,k) - A_d(\lambda,k))^2 P_d(\lambda,k).$$

(8)

In (7) and (8), $\varepsilon^2_s(\lambda,k)$ and $\varepsilon^2_d(\lambda,k)$ correspond to the power spectra of speech and noise distortions, respectively. To make audible, both the distortions should be masked. A complete masking of all distortions, however, is not always possible [9]. A possible way for deriving the gain function consists of trying to mask the noise power distortions. This is typically done by keeping the noise power distortions exactly at the masking threshold level, thereby minimizing the speech distortion, that is,

$$\varepsilon^2_d(\lambda,k) = (G(\lambda,k) - A_d(\lambda,k))^2 P_d(\lambda,k) = T(\lambda,k)$$

(9)

where $T(\lambda,k)$ is the masking threshold. Solving (9) with the constraint $A_d(\lambda,k) \leq G(\lambda,k) \leq 1$, the gain function $G(\lambda,k)$ is obtained as

$$G(\lambda,k) = \min \left( \frac{T(\lambda,k)}{P_d(\lambda,k)} + A_d(\lambda,k), 1 \right)$$

(10)

where $P_d(\lambda,k)$ is the estimated noise power spectrum obtained during non-speech activity. In (10), $A_d(\lambda,k)$ represents the estimated audible noise and $T(\lambda,k)$ corresponds to an auditory masking threshold.

3.2. Masking Threshold Determination

The masking threshold is computed by exploiting the masking properties of human auditory system. The estimation of the masking threshold includes rough estimation of the speech, critical band analysis, spreading function, signal tonality, absolute threshold, normalization and comparison to the absolute threshold [10]. The rough estimation of the speech $\hat{S}(\lambda,k)$ in our implementation is conducted using simple spectral subtraction.

3.3. Audible Noise Estimation

The audible noise spectrum is estimated as the approach proposed in [7]. Generally, the spectral components which lie above the masking threshold are perceived as noise, and refereed to audible noise spectrum. The audible noise spectrum, $A_d(\lambda,k)$, can be expressed as the difference between the audible spectrum of the noisy speech $A_x(\lambda,k)$ and the audible spectrum of the clean speech $A_s(\lambda,k)$, that is,

$$A_d(\lambda,k) = A_x(\lambda,k) - A_s(\lambda,k).$$

(11)

The audible spectral components for $A_x(\lambda,k)$ and $A_s(\lambda,k)$ in (11) are obtained by taking the maximum between the power spectrum of speech and the corresponding masking threshold per frequency component. The approach proposed in [7] extracted some information on $A_d(\lambda,k)$ and modified the noisy spectrum in a way that made the perceived noise inaudible by a parametric Wiener-type function. In contrast to the approach in [7], the proposed method estimates the audible noise using the power spectrum of the roughly estimated speech $P_d(\lambda,k)$. A more analytical expression for estimation of the audible noise can be found in (12), at the top of the next page.

3.4. Suppression Characteristics

The underlying principle for the derivation of gain function $G(\lambda,k)$ obtained from (10) can be interpreted with Figure 1. Figure 1 shows the suppression characteristics with the respective influence of audible noise and masking threshold. The observed characteristics can be summarized as follows.

1. Low attenuation is obtained for relatively high value of masking threshold.
2. Strong suppression is acquired for high audible noise.
3. Suppression remains almost constant for high audible noise.
approaches. In ANS, an iterative procedure is used to get a good estimate of the masking threshold values. This is typically done for two iterations, since it provides a good trade-off between noise reduction and speech distortion.

As mentioned earlier, speech enhancement algorithms typically degrade the signal components while suppressing the background noise, particularly at very low SNRs. This, in turn, degrades the intelligibility of the processed speech. This is mainly due to exaggerated suppression which causes the speech and noise distortions simultaneously. To judge in terms of individual contribution of noise and speech distortions to overall quality, the enhancement algorithms are evaluated by composite measures. The composite measures, which are developed in [11], combine the existing objective measures linearly by regression analysis. Three kinds of composite measures are used.

- SIG: A composite measure for signal distortion formed by linearly combining the log-likelihood ratio (LLR), perceptual evaluation of speech quality (PESQ), and weighted spectral slope distance (WSS).
- BAK: A composite measure for background noise distortion formed by linearly combining the segmental SNR (segSNR), PESQ, and WSS measures.
- OVL: A composite measure for overall quality formed by linearly combining the PESQ, LLR, and WSS measures.

Table 1 lists the mean scores for SIG, BAK, and OVL scales for speech processed by 4 different methods at different types of noise and levels. The mean scores of the noisy speech are also listed for reference.

As is evident from Table 1, the proposed method performs better than the other algorithms in terms of overall quality. Relatively, lower speech distortion (i.e., higher SIG scores) was observed with our proposed method in all cases. Lower noise distortion (i.e., higher BAK scores) was also observed with our proposed method in most conditions.

The comparison between the enhanced speech and the noisy speech provides information as to which algorithms perform well in terms of intelligibility. Such comparison is important because the processed speech can even be more disturbing to the human listeners than the noisy speech. As can be seen from Table 1, the proposed method performs the best in all conditions when compared to the noisy speech. An informal listening test also confirms that the proposed approach performs significantly better than the conventional approaches.

4. Experimental Results

In this section, we evaluate the performance of the masking-based proposed method. The NOIZEUS speech corpus [11] is used for evaluation in the experiments. The corpus comes with non-stationary noises at different SNRs. Two kinds of noises taken from the corpus are used in the experiments. These are exhibition noise and train noise. A total of four utterances from the NOIZEUS corpus are used in the experiments. The corpus comes with two kinds of noises at different SNRs. Two kinds of noises are from male speakers and half are from female speakers. The test utterances are sampled at 8 kHz. A 20-msec analysis Hamming window is used with 50% overlap between frames.

The performance of the proposed method is investigated by comparing with the following methods:

1. OML: Optimally Modified log-MMSE [4]
2. ANS: Audible Noise Suppression [7]
3. JND: Just Notable Distortion [9].

The typical parameter selection is the same as that in those approaches. In ANS, an iterative procedure is used to get a good estimate of the masking threshold values. This is typically done for two iterations, since it provides a good trade-off between noise reduction and speech distortion.

As mentioned earlier, speech enhancement algorithms typically degrade the signal components while suppressing the background noise, particularly at very low SNRs. This, in turn, degrades the intelligibility of the processed speech. This is mainly due to exaggerated suppression which causes the speech and noise distortions simultaneously. To judge in terms of individual contribution of noise and speech distortions to overall quality, the enhancement algorithms are evaluated by composite measures. The composite measures, which are developed in [11], combine the existing objective measures linearly by regression analysis. Three kinds of composite measures are used.

- SIG: A composite measure for signal distortion formed by linearly combining the log-likelihood ratio (LLR), perceptual evaluation of speech quality (PESQ), and weighted spectral slope distance (WSS).
- BAK: A composite measure for background noise distortion formed by linearly combining the segmental SNR (segSNR), PESQ, and WSS measures.
- OVL: A composite measure for overall quality formed by linearly combining the PESQ, LLR, and WSS measures.

Table 1 lists the mean scores for SIG, BAK, and OVL scales for speech processed by 4 different methods at different types of noise and levels. The mean scores of the noisy speech are also listed for reference.

As is evident from Table 1, the proposed method performs better than the other algorithms in terms of overall quality. Relatively, lower speech distortion (i.e., higher SIG scores) was observed with our proposed method in all cases. Lower noise distortion (i.e., higher BAK scores) was also observed with our proposed method in most conditions.

The comparison between the enhanced speech and the noisy speech provides information as to which algorithms perform well in terms of intelligibility. Such comparison is important because the processed speech can even be more disturbing to the human listeners than the noisy speech. As can be seen from Table 1, the proposed method performs the best in all conditions when compared to the noisy speech. An informal listening test also confirms that the proposed approach performs significantly better than the conventional approaches.

5. Conclusions

In this paper, we have proposed a speech enhancement technique by exploiting perceptual properties of human auditory system. Despite the simplicity, the proposed method has shown to have significant performance compared to the other existing algorithms, particularly in adverse conditions. This is due to the time- and frequency-dependent attenuation factor, which is considered as a function of audible spectral components. The
<table>
<thead>
<tr>
<th>Noise Types</th>
<th>Measures</th>
<th>Method</th>
<th>SNR = 0 dB</th>
<th>SNR = 5 dB</th>
<th>SNR = 10 dB</th>
<th>SNR = 15 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exhibition</td>
<td>SIG</td>
<td>Noisy</td>
<td>1.96</td>
<td>2.56</td>
<td>3.13</td>
<td>3.75</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OML</td>
<td>1.46</td>
<td>2.17</td>
<td>2.95</td>
<td>3.53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANS</td>
<td>1.14</td>
<td>1.96</td>
<td>2.74</td>
<td>3.19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JND</td>
<td>1.94</td>
<td>2.63</td>
<td>3.20</td>
<td>3.79</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proposed</td>
<td><strong>2.04</strong></td>
<td><strong>2.65</strong></td>
<td><strong>3.23</strong></td>
<td><strong>3.81</strong></td>
</tr>
<tr>
<td></td>
<td>BAK</td>
<td>Noisy</td>
<td>1.71</td>
<td>2.10</td>
<td>2.54</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OML</td>
<td>1.64</td>
<td>2.15</td>
<td>2.58</td>
<td>3.18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANS</td>
<td>1.59</td>
<td>2.17</td>
<td>2.59</td>
<td><strong>3.19</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>JND</td>
<td>1.78</td>
<td>2.13</td>
<td>2.56</td>
<td>3.04</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proposed</td>
<td><strong>1.82</strong></td>
<td><strong>2.20</strong></td>
<td><strong>2.62</strong></td>
<td><strong>3.04</strong></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>Noisy</td>
<td>1.68</td>
<td>2.13</td>
<td>2.58</td>
<td>3.12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OML</td>
<td>1.17</td>
<td>1.89</td>
<td>2.60</td>
<td>3.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANS</td>
<td>1.05</td>
<td>1.76</td>
<td>2.47</td>
<td>2.91</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JND</td>
<td>1.66</td>
<td>2.20</td>
<td>2.65</td>
<td>3.17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proposed</td>
<td><strong>1.73</strong></td>
<td><strong>2.21</strong></td>
<td><strong>2.67</strong></td>
<td><strong>3.19</strong></td>
</tr>
<tr>
<td>Train</td>
<td>SIG</td>
<td>Noisy</td>
<td>1.99</td>
<td>2.52</td>
<td>3.09</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OML</td>
<td>1.50</td>
<td>2.42</td>
<td>2.90</td>
<td>3.51</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANS</td>
<td>1.08</td>
<td>1.81</td>
<td>2.36</td>
<td>3.14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JND</td>
<td>1.80</td>
<td>2.56</td>
<td>3.23</td>
<td>3.71</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proposed</td>
<td><strong>2.02</strong></td>
<td><strong>2.59</strong></td>
<td><strong>3.25</strong></td>
<td><strong>3.72</strong></td>
</tr>
<tr>
<td></td>
<td>BAK</td>
<td>Noisy</td>
<td>1.72</td>
<td>2.13</td>
<td>2.56</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OML</td>
<td>1.73</td>
<td>2.18</td>
<td>2.67</td>
<td><strong>3.20</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANS</td>
<td>1.39</td>
<td>2.08</td>
<td>2.65</td>
<td>3.15</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JND</td>
<td>1.77</td>
<td>2.21</td>
<td>2.64</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proposed</td>
<td><strong>1.82</strong></td>
<td><strong>2.27</strong></td>
<td><strong>2.69</strong></td>
<td><strong>3.02</strong></td>
</tr>
<tr>
<td></td>
<td>OVL</td>
<td>Noisy</td>
<td>1.70</td>
<td>2.12</td>
<td>2.58</td>
<td>3.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>OML</td>
<td>1.32</td>
<td>2.07</td>
<td>2.60</td>
<td><strong>3.11</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ANS</td>
<td>1.02</td>
<td>1.60</td>
<td>2.19</td>
<td>2.86</td>
</tr>
<tr>
<td></td>
<td></td>
<td>JND</td>
<td>1.60</td>
<td>2.17</td>
<td>2.70</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Proposed</td>
<td><strong>1.73</strong></td>
<td><strong>2.19</strong></td>
<td><strong>2.72</strong></td>
<td><strong>3.11</strong></td>
</tr>
</tbody>
</table>

6. References


