SPEECH ENHANCEMENT IN NON-STATIONARY NOISE ENVIRONMENTS

Hyoung-Gook Kim1,2, Dietmar Ruwich1

1Cortologic AG, Berlin, Germany
2Technical University of Berlin, Germany

{kim, ruwich}@cortologic.com

ABSTRACT

This paper presents a speech enhancement using a noise estimation based on the ratio of the noisy speech and its minimum (NSMR) for non-stationary noise environments. The noise estimator is a very simple but highly effective real time approach for single channel noise reduction. The enhanced speech is free of musical tones and reverberation artifacts and sounds very natural compared to methods using other short-time spectrum attenuation techniques. The performance is measured by the segmental signal-to-noise ratio and MOS tests. To judge the performance the recognition accuracy of an Automatic Speech Recognition (ASR) system using Mel-scale Frequency Cepstral Coefficients (MFCC) features is measured with and without noise reduction. In another experiment we apply the NSMR noise reduction method to speech reconstruction at the back-end of a distributed speech recognition (DSR) system under various noise conditions.

1. INTRODUCTION

For single channel methods, spectral subtraction [1] is a commonly applied method due to its robustness and relatively simple implementation. It is known that the noise estimator is a very important component for the overall quality of the speech enhancement system in the presence of non-stationary noise. Many modified forms of spectral subtraction have been suggested in the literature to achieve good speech quality and avoid musical tones. Of these, spectral subtraction based on minimum statistics by Martin [2] is an attractive method to track non-stationary noise signals without speech activity detection and provides high noise reduction and low speech distortion. However, the smoothing with a fixed smoothing parameter will lead to inaccurate noise estimates as the sliding window for the minimum search might slip into broad peaks and the noise estimate is biased toward low energy phonemes. To overcome these limitations, the smoothing parameter and the bias compensation factor are made time and frequency dependent, and estimated for each spectral component and each time frame [3]. The only disadvantage is the memory consumption for the parameters.

The noise reduction algorithm presented in this paper focuses on system simplicity and low computational costs due to a small number of adaptive parameters. This is accomplished by a noise estimate based on the ratio of the noisy speech and its minimum (NSMR) and a gain function smoothed over time and frequency, which reduces fluctuations of the estimated spectral gain factors.

2. DESCRIPTION OF THE ALGORITHM

Assuming an additive uncorrelated noise model, the proposed noise reduction is based on short-time spectrum attenuation techniques and consists of two independent main parts, i.e. estimation of the noise spectrum and filtering of the noisy speech to obtain clean speech. A simplified block diagram of the approach is shown in figure 1.

Figure 1: Block scheme of the background noise reduction based on NSMR.

Let \( x(t) \) denote the input signal, which is assumed to be the sum of a clean speech signal \( s(t) \) and disturbing noise \( n(t) \). The short-time magnitude spectrum \( X(k, l) \) of \( x(t) \) in a time frame \( l \) and frequency bin \( k \) is estimated by computing a 256-point FFT with windowing and a frame increment of 128 samples. The noise spectrum is estimated by a noise spectrum tracker using NSMR. The spectral weighting is performed by multiplying the magnitude spectrum of the noisy speech signal with spectrally smoothed gain factors \( G_{NSMR}(k, l) \) calculated from the background noise estimation. The filtered spectral values are transformed back into the time domain using inverse short-time discrete Fourier transformation in order to calculate the enhanced speech \( \hat{s}(t) \).

2.1. Noise estimation using NSMR

If noise is not correlated with the clean speech, the noise component is additive in the magnitude spectrum of the signal.
Since only the noisy speech signal is available in single channel systems, the estimation of the disturbing noise has to be extracted from the noisy speech signal. To achieve an accurate noise estimation for various non-stationary noise types, the following three steps are performed:

**Step 1** Average of the short-time magnitude

\[
E(k,l) = \frac{1}{L} \sum_{i=1}^{L} |X(k,l)|
\]

(1)

\(E(k,l)\) is obtained as an average of the short-time spectral magnitude of the noisy speech signal over \(L\) frames.

**Step 2** Spectral minimum tracking

As a second step, we use the idea of minimum statistics by Martin. Minimum values \(\hat{M}(k,l)\) of the smoothed estimation \(E(k,l)\) are searched within windows of \(N\) frames whether speech is present or not. The minimum value for the current frame is found by a comparison with the stored minimum value. This step is capable of tracking non-stationary noise signals.

\[
\hat{M}(k,l) = \min \{ \hat{M}(k,l-1), E(k,l) \}
\]

(2)

**Step 3** Weighted noise estimation based on NSMR

The minimum tracking of step 2 provides a coarse estimate of the noise magnitude. If the noise estimate is too low, residual noise will be perceived. If the estimate is too high, low energy speech will be suppressed and intelligibility will be lost. To achieve good tracking capability, the noise estimation has to track the change of the noise characteristics in both speech and non-speech periods. To this end, we use the ratio \(T(k,l)\) between spectral amplitude of the noisy speech signal and its derived minimum; it is given by

\[
T(k,l) = \frac{|X(k,l)|}{\hat{M}(k,l)}.
\]

(3)

The weighted noise estimation \(\lambda(k,l)\) is accomplished by

\[
\lambda(k,l) = \begin{cases} 
\frac{\hat{M}(k,l)}{E(k,l)} & \text{if } T(k,l) > \psi \\
E(k,l) & \text{else} 
\end{cases}
\]

(4)

where \(\lambda(k,l)\) is the weighted noise estimation and \(\psi\) is typically in the range \(2 < \psi < 7\). \(T(k,l) > \psi\) indicates hypothetical speech activity and the noise estimation updates the minimum values \(\hat{M}(k,l)\) in order to prevent suppression of low energy phonemes. In figure 2 the results of three steps of the noise estimation are shown. Figure 2 (a) depicts the spectral magnitude of the clean speech. In figure 2 (b) f16 cockpit noise was artificially added to the clean speech at an SNR of 7 dB. Its spectral magnitude at frequency bin \(f = 9\) is illustrated. The valleys of figure 2 (c) present a minimum \(\hat{M}(k,l)\) as bold line while peaks correspond to the spectral amplitude of the noisy speech signal (thin line). Figure 2 (d) shows that NSMR performs the minimum tracking following the peaks of the speech signal. During speech pauses the noise magnitude is fairly accurately estimated.

### 2.2. Spectrally smoothed gain computation

Using the noise estimation \(\lambda(k,l)\), preliminary gain factors \(G(k,l)\) can be computed as:

\[
G(k,l) = 1 - \frac{\lambda(k,l)}{|X(k,l)|}.
\]

(5)

For each spectral magnitude \(X(k,l)\) the separate noise estimation \(\lambda(k,l)\) is performed and suffers from variations over time and frequency. To reduce fluctuations of the preliminary spectral gains over time, recursive smoothing of the gain factors over time is performed as

\[
G_{IS}(k,l) = \alpha G_{IS}(k,l-1) + (1 - \alpha) G(k,l),
\]

(6)

where \(G_{IS}(k,l)\) denotes the time-smoothed gain factors and the parameter \(\alpha (0 < \alpha < 1)\) is a time smoothing constant. With increasing \(\alpha\), the smoothing of the gain factors is increased. This time smoothing is very effective in reducing musical noise. Nevertheless, smoothing should not be too intensive. Otherwise, it may lead to reverberation artifacts in the enhanced speech signal.
The last step of the smoothed gain function computation is the interaction of the time-smoothing gain factors $G_{TS}(k,l)$ of neighboring frequency bins

$$G_{TS}(k,l) = \beta G_{TS}(k-1,l) + (1-2\beta)G_{TS}(k,l) + \beta G_{TS}(k+1,l)$$

(7)

where $G_{TS}(k,l)$ denotes the smoothed gain factors over frequency and $\beta (0 < \beta \leq 0.5)$ is the frequency smoothing constant. This approach decreases fluctuations of the estimated spectral gains, and therefore the speech distortion is low and speech sounds natural.

3. EXPERIMENTAL RESULTS

The performance of the proposed algorithm is measured using the following criteria: segmental SNR improvement in speech segments, recognition accuracy improvement, subjective study of speech spectrograms, and listening test.

3.1. Segmental SNR improvement

To measure the performance of the proposed algorithm in comparison to other single channel noise reduction methods, the segmental signal-to-noise ratio ($\text{seg.SNR}$) is computed for the enhanced speech signals. For this, three types of background noise - white noise, car noise and factory noise - were artificially added to different portions of the data at SNR of 5 dB and -5 dB. While the white noise is completely stationary, the car noise has some fluctuations and factory noise is highly non-stationary. The average improvement of the segmental SNR during speech frames,

$$\text{improve.SNR} = \text{seg.SNR}_{\text{out}} - \text{seg.SNR}_{\text{in}}$$

(8)

for NSMR and two other noise reduction methods is shown in table 1.

<table>
<thead>
<tr>
<th>methods</th>
<th>Input SNR [dB]</th>
<th>white</th>
<th>car</th>
<th>factory</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>5</td>
<td>-5</td>
<td>5</td>
<td>-5</td>
</tr>
<tr>
<td>PSS</td>
<td>4.3</td>
<td>7.3</td>
<td>5.3</td>
<td>8.1</td>
</tr>
<tr>
<td>MS</td>
<td>7.8</td>
<td>12.8</td>
<td>8.4</td>
<td>14</td>
</tr>
<tr>
<td>NSMR</td>
<td>9.9</td>
<td>13.6</td>
<td>10.8</td>
<td>15</td>
</tr>
</tbody>
</table>

Table 1: Comparison of segmental SNR improvement of different single channel noise estimation methods (PSS: Power Spectral Subtraction [1], MS: spectral subtraction based on minimum statistics [2] and NSMR)

Minimum statistics of noise estimation follows Martin [2]. The following parameters are used: smoothing constant is 0.95, window length 100, sub-window length 25 and bias compensation factor 1.5. Table 1 shows that NSMR algorithm gives best results for input SNR 5 dB and -5 dB compared to the results of PSS and MS. For input SNR -5 dB of factory noise the performance of NMSR is only slightly better than MS.

3.2. Recognition accuracy in an automatic speech recognition (ASR) system

To judge the performance in a second experiment, we compare the recognition accuracy using a mel-scale frequency cepstral coefficient ASR with and without the noise reduction algorithm in speaker independent isolated digit recognition. For the training of the phoneme models, the TIDIGITS database is used. For the additive noise in these experiments, we used factory noise. The test set consists of 200 sentences from 50 speakers that are prepared by adding factory noise to a clean subset of the TIDIGITS database, i.e. factory noise has been artificially added to different portions of the database at the SNR ratios ranging from clean speech over 20 dB to 0 dB in steps of -5 dB. The proposed algorithm was combined with feature extraction and trained on clean speech. The feature vector consists of 26 parameters: 13 mel-scale frequency cepstral coefficients along with their delta coefficients. The mel-scale cepstrum coefficients are derived from 26 equidistant channels in the mel frequency domain. The input to the MLP for the non-linear transformation consists of 9 frames (current frames, 4 frame in the past and 4 frames in the future) as commonly used in a hybrid HMM/MLP ASR system (234 inputs, 420 hidden units and 61 outputs). The proposed NSMR-filtering front-end was compared to a power spectral subtraction (PSS) and DGF [4] front-end. Figure 3 describes the results of the recognition accuracy.

![Figure 3: The results of speaker independent isolated digit recognition: NSMR-MFCC (○), DGF-MFCC (▲), PSS-MFCC (■), and MFCC (×).](image)

When the test speech is corrupted by additive factory noise, the recognition rate of mel frequency cepstral coefficients (MFCC) is seriously decreased. The combination of MFCC with DGF yields a higher recognition rate than the MFCC method alone or MFCC with PSS. A still higher recognition rate is achieved by combining MFCC with NSMR.

3.3. Speech spectrograms and listening test

In order to visualize the effect of the noise reduction algorithm based on NSMR, the spectrograms of the clean speech, noisy
speech, and the enhanced speech are shown in figure 4. Dark gray areas correspond to the speech components while background noise is light gray. In order to evaluate the quality of the noise reduction algorithm based on NSMR, a subjective Mean-Opinion-Score (MOS) test was performed. Recently, speech reconstruction using the MFCC parameters for the back-end of the distributed speech recognition has been developed by other authors [5, 6]. Our reconstruction system tested here is combined with the proposed noise reduction method to improve performance under noisy conditions. For the reconstruction using a sinusoidal synthesis is performed. The determination of the pitch period is based on a normalized autocorrelation in time domain and spectral prominent peak search corresponding to the pitch harmonics in order to compensate the pitch errors. The speech is encoded with a low data rate (1 kbit/s) according to the bit allocation (7 bits for the pitch period, 9 bits for the 13 MFCCs, total 16 bits per 16 ms). In table 2, MOS test results of the speech reconstruction using the MFCC parameters combined with NSMR are compared to those of other noise reduction methods. The MOS test in this experiment involved 24 listeners, and a total of 10 sentences added to car noise, street noise and F-16 were included. Additionally, the spectrogram of the reconstructed speech is presented in figure 4 (d).

In table 2, MOS test results of the speech reconstruction using the MFCC parameters combined with NSMR are compared to those of other noise reduction methods. The MOS test in this experiment involved 24 listeners, and a total of 10 sentences added to car noise, street noise and F-16 were included. Additionally, the spectrogram of the reconstructed speech is presented in figure 4 (d).

Table 2: The results of MOS test: RC-N is the reconstructed speech using the MFCC parameters in concatenation with NSMR.

<table>
<thead>
<tr>
<th>methods</th>
<th>MOS car [dB]</th>
<th>MOS street [dB]</th>
<th>MOS F-16 [dB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>PSS</td>
<td>2.3</td>
<td>2.5</td>
<td>2.2</td>
</tr>
<tr>
<td>DGF</td>
<td>2.8</td>
<td>3.2</td>
<td>2.7</td>
</tr>
<tr>
<td>NSMR</td>
<td>3.1</td>
<td>3.5</td>
<td>2.9</td>
</tr>
<tr>
<td>RC-N</td>
<td>2.3</td>
<td>2.5</td>
<td>2.2</td>
</tr>
</tbody>
</table>

Figure 4: Spectrograms of clean speech, noisy speech, enhanced speech and reconstructed speech

(d) Spectrogram of reconstructed speech from MFCC parameters (1 kbit/s) combined with NSMR.

4. CONCLUSIONS

A speech enhancement system based on NSMR (spectral amplitude of the noisy speech to its minimum ratio) for non-stationary noise environments. Experiments confirmed that this method, while having only low computational requirements significantly improves the quality of speech and ASR performance in the presence of non-stationary noise.

5. REFERENCES