AN IMPROVED SPECTRAL SUBTRACTION SPEECH ENHANCEMENT SYSTEM BY USING AN ADAPTIVE SPECTRAL ESTIMATOR

Saeed Ayat1 ayat@mehr.sharif.edu Mohamad T. Manzuri2 manzuri@sharif.edu Roohollah Dianat3 dianat@ce.sharif.ac.ir Jahanshah Kabudian4 kabudian@ce.aut.ac.ir

1,2,3: Computer Engineering Department, Sharif University of Technology
Tehran, IRAN
4: Computer Engineering Department, Amirkabir University of Technology
Tehran, IRAN

Abstract

Spectral subtraction is one of the most famous and common-used methods for speech enhancement. The main weakness of this method is the production of an annoying noise called musical noise.

In this paper, we have reduced the musical noise and improved the quality of enhanced speech by increasing the accuracy of the system spectral estimator. This method is useful for speech enhancement systems in which the speech signal is degraded by stationary or near-stationary noises.

Different experimental results in different SNRs confirmed the better performance of our system from both objective and subjective views, which means better quality and more SNR improvement.

Keywords: Speech enhancement; spectral subtraction; spectral estimation.

1. Introduction

One of the most important topics in signal processing is the reduction or removal of undesirable signals or noise from the original signal. Speech processing in the real world requires robust algorithms against different kinds of noises. A lot of methods, from the earliest ones that uses spectral subtraction (for example [1]) to the latest ones that uses wavelet (for example [2]), have been proposed to reduce or remove the noise from speech signal. Speech enhancement can be used to increase the quality of the speech or improve the performance of a recognition system.

One of the first methods introduced for speech enhancement is spectral subtraction. Till now, different versions of spectral subtraction have been proposed to increase the performance of this method, for example [1, 3, 4]. Despite of its high noise removal, it can cause an annoying noise called musical noise and hence it can reduce overall quality. In this paper, we introduce a new method to reduce or remove the musical noise.

Musical noise is produced because, we don’t have the needed spectra exactly, so we have to use their estimations.

In our method after separating speech and silence frames in the noisy signal with a basic analysis frame, we can increase the analysis frame length until it covers all the current silence frames. As in periodogram estimator technique the accuracy improves by increasing the number of frame samples [5], by using this adaptive analysis frame length we can have a better spectral estimation for noise and noisy signal and so the system can produce a better enhanced signal with less musical noise. Also, we have evaluated our system with a new test in which we have increased the oversubtraction factor in the extended spectral subtraction method until it began to produce the musical noise and repeat the test for different resolutions of the spectra. This test showed that with better resolution that the system could have with our method we could increase the oversubtraction factor and then we could get more SNR improvements with the same quality in the enhanced speech.

The paper is organized as follows: In section 2 we have a review on spectral subtraction method. Section 3 describes noise estimation in spectral subtraction method. In section 4 we proposed our method and in section 5 we present the simulation results.

2. Spectral Subtraction

As we know a clean speech signal consists of some sections that have speech and some others that have no speech and we call them silences. In a noisy speech signal these silence sections have only noise and other sections have noisy speech signals. If the noise is stationary we can estimate its spectrum in the silent sections.

In spectral subtraction method, after framing the noisy speech signal we use a silence detector or a voice activity detector for separating noisy speech frames and noise frames. After that with applying, FFT we have the spectrum of each frame. By calculating the average of the noise frames spectra we have estimation for noise spectra. Now with subtracting this estimation of noise spectrum from the spectrum of each noisy speech frames we can achieve enhanced speech signal.

There are many different versions for spectral subtraction. In a generalized spectral subtraction [6] we have:
\[
\hat{S}(w) = \max \left\{ \left( |S(w)|^\alpha - \beta |N(w)|^\alpha \right)^\frac{1}{\gamma} |N(w)| \right\}.
\]

Where \( |S(w)| \), \( |N(w)| \) and \( \hat{S}(w) \) are magnitude spectrum of noisy speech, estimation of noise and enhanced speech. \( \beta \) is the oversubtraction factor and \( \gamma \) is spectral floor. Both \( \beta \) and \( \gamma \) are adjusted to improve the quality of enhanced speech.

### 3. Noise Estimation

The main problem of spectral subtraction method is the production of musical noise. Musical noise is produced because we don’t have the exact spectrum of the noise signal.

By the assuming that the noise is stationary, a good estimation can be resulted by computing the average of the noise in silence frames spectra. We called such average \( \overline{W}(w) \).

In presence of nonstationary noises, an adaptation technique can be used. Given an initial value \( \overline{W}_0(w) \), if the current frame is silence, \( \overline{W}_m(w) \) is updated using this equation:

\[
\overline{W}_m(w) = (1 - f) \overline{W}_{m-1}(w) + f \overline{Y}_m(w)
\]

In this formula \( \overline{Y}_m(w) \) is the spectrum of current silence frame and \( f \) is a coefficient called forgetting factor. This factor is changed depending on the noise changing rate.

### 4. Proposed Method

In the previous section we introduced the typical method for noise spectrum estimation by computing the average of the spectrum of silence frames. This method suffers from the musical noise and as mentioned previously this is the major drawback of the spectral subtraction.

This noise is produced because we don’t have the needed spectra, exactly; so we have to use their estimated values.

In our method that estimates the spectrum better than the basic averaging method, after separating speech and silence frames in the noisy signal with a basic analysis frame, we can increase the analysis frame length until it covers all the current silence frames. As in periodogram estimator technique the accuracy improves by increasing the number of signal samples, by using this adaptive analysis frame length we can have a better spectral estimation for noise and noisy signal and so the system can produce a better enhanced signal with less musical noise.

As we know if the frame length is \( L \) the frequency resolution in Fourier spectral analysis is \( \frac{F_s}{L} \). For example if \( F_s=11025 \text{Hz} \) and \( L=256 \) then \( \frac{F_s}{L} \) is 43Hz and this resolution may not be enough for speech signal.

In our method we first apply a SAD algorithm with \( L=256 \) and \( L/2=128 \) points overlap to detect the silence frames. Now we can increase the analysis frame length until it covers all the current silence frames. By this method we have larger window length and hence better frequency resolution. If we have several silence areas with the new frame length, the average of them is the overall noise spectrum.

By applying such method we have better noise spectrum estimation with less musical noises.

In section 5 we give experimental results that confirm this improvement clearly.

### 5. Experimental Results and Analysis

In this section we explain our simulation. The speech signal that used for these tests was pronounced with a male and recorded by sampling frequency \( F_s=11025 \text{Hz} \) in 16 bits, and degraded by additive Gaussian white noise, so we can have the noisy signal in required SNRs. For evaluating our method we use both objective and subjective tests.

In objective test, we calculate SNR improvement as below. If \( s(n) \) is the clean speech, \( y(n) \) the noisy, \( \hat{s}(n) \) the enhanced signal and \( w(n) \) the noise then we have:

\[
y(n) = s(n) + w(n)
\]

and the SNR improvement is computed as follows:

\[
\text{SNR}_{\text{imp}} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}
\]

In which \( \text{SNR}_{\text{in}} \) and \( \text{SNR}_{\text{out}} \) are the SNRs for noisy and enhanced:

\[
\text{SNR}_{\text{in}} = 10 \log_{10} \frac{\sum s^2(n)}{\sum (y(n) - s(n))^2}
\]

\[
\text{SNR}_{\text{out}} = 10 \log_{10} \frac{\sum s^2(n)}{\sum (\hat{s}(n) - s(n))^2}
\]

#### 5.1. First Experiment: Comparison of SNR Improvement at the Start of Musical Noise

In this experiment a listener listen to the enhanced signal and increase \( \beta \) until the musical noise appears in the enhanced signal. At this point, \( \beta \) and SNR improvement is recorded. This is done for SNRs equal to -5, 0, +5 and different frame lengths with 256, 512, 1024, 2048 and 4096 samples. \( \alpha \) is fixed to 1.0 and \( \gamma \) to 0.0. Note that the frame length is 256 in silence detection step.
Tables 1 to 3 show the results for $\beta$ and SNR improvement at the appearance of musical noise in the enhanced signal for tested SNRs.

**Table 1: $\beta$ and SNR improvement at the start of musical noise (SNR=-5db)**

<table>
<thead>
<tr>
<th>$L$</th>
<th>256</th>
<th>512</th>
<th>1024</th>
<th>2048</th>
<th>4096</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR$_{imp}$</td>
<td>0.8</td>
<td>1.2</td>
<td>2.1</td>
<td>2.4</td>
<td>4.1</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
<td>0.6</td>
<td>1.2</td>
</tr>
</tbody>
</table>

**Table 2: $\beta$ and SNR improvement at the start of musical noise (SNR=0db)**

<table>
<thead>
<tr>
<th>$L$</th>
<th>256</th>
<th>512</th>
<th>1024</th>
<th>2048</th>
<th>4096</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR$_{imp}$</td>
<td>0.8</td>
<td>1.6</td>
<td>2.0</td>
<td>3.0</td>
<td>4.2</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.2</td>
<td>0.3</td>
<td>0.5</td>
<td>0.8</td>
<td>1.6</td>
</tr>
</tbody>
</table>

**Table 3: $\beta$ and SNR improvement at the start of musical noise (SNR=5db)**

<table>
<thead>
<tr>
<th>$L$</th>
<th>256</th>
<th>512</th>
<th>1024</th>
<th>2048</th>
<th>4096</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNR$_{imp}$</td>
<td>0.4</td>
<td>1.1</td>
<td>1.8</td>
<td>2.5</td>
<td>3.3</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.1</td>
<td>0.3</td>
<td>0.5</td>
<td>0.8</td>
<td>1.5</td>
</tr>
</tbody>
</table>

As we can see the SNR improvement is better for longer frame lengths. This show that the musical noise arises from inaccurate noise estimation and reduces as the frame length increases, so we can choose a greater $\beta$ without production of musical noise and by increasing it we can have less noise in the enhanced signal and then achieve more SNR improvement.

Figures 1 to 3 show the same results in a clearer manner.

**5.2. Second Experiment: Musical Noise Comparison with the Same SNR Improvement**

In this experiment we want to compare subjective quality of enhanced signal with various frame lengths. For this sake the $\beta$ parameter is accurately set such that the SNR improvements are equal to each others for different frame lengths. This test is also performed for different input SNRs. After listening to these signals and subjective comparison of enhanced signals in various frame lengths with equal SNR$_{imp}$ we saw that despite of equal SNR$_{imp}$ the subjective quality of enhanced signal in greater frame length is better and its musical noise is lower.
6. Conclusions

Spectral subtraction is one of the most famous and commonly used methods for speech enhancement. The main weakness of this method is the production of an annoying noise called musical noise. This noise is produced because we don’t have the needed spectrums for noise and noisy signal exactly, so we have to use their estimations.

In this paper we proposed an improved spectral subtraction method by increasing the accuracy of spectral estimator. This adaptive estimator can give better spectral estimation by increasing the analysis frame length that achieves in silence regions. In this method for separating silence frames we use a basic analyzing frame and for estimation the spectrum we use an adaptive frame length that can increase until it covers all current silence region. By this method we could have a better spectral estimation for noise and noisy signal and so the system can produce a better enhanced signal with less musical noise.

Acknowledgement

This research was supported in part by Iran Telecommunication Research Center (ITRC).

References