Speech Enhancement Based on a Combination of Spectral Subtraction and a Minimum Mean-Square Error Short-Time Log-Spectral Amplitude Estimator in Wavelet Domain

Farid Ykhlef, A. Guessoum and D. Berkani
LATSI laboratory, Department of Electronics, Faculty of Engineering Sciences, University of Blida, Algeria.
f_ykhlef@yahoo.fr

Abstract- The corruption of speech due to presence of additive background noise causes severe difficulties in various communication environments. This paper presents a novel noise reduction technique based upon a combination of cascaded spectral subtraction and minimum mean-square error short-time log-spectral amplitude estimator methods in the wavelet domain. The scheme’s performance is illustrated by experiments in noisy car environment, in comparison with spectral subtraction and a minimum mean-square error short-time log-spectral amplitude estimator.

Keywords: spectral subtraction, minimum mean-square error short-time log-spectral amplitude estimator, wavelet domain.

I. INTRODUCTION

Noise reduction is a subject of research in many different fields [1]. Depending on the environment, the application, the source signals, the noise, and so on, the solutions look very different. Here we consider noise reduction for speech signals, and concentrate on common acoustic environments such as an office room or inside a car. The goal of noise reduction is to reduce the noise level without distorting speech, thus to reduce the stress on the listener and ideally to increase intelligibility [2]. There are many different ways to perform the noise reduction. Principally, the solutions can be split into two classes: single and multi-microphone systems. Whereas multi-microphone systems exploit the spatial properties of speech and noise, a single-microphone system usually relies on temporal characteristics. A fundamental requirement for most single-microphone systems is that speech and noise are additive and result form uncorrelated statistical processes, and that the spectral characteristics of the noise change markedly slower than those of the speech.

The noise reduction method discussed in this paper is a single channel method based on converting successive short segments of speech into the frequency domain. In the frequency domain, the noise is removed by adjusting the discrete frequency “bins” on a frame by frame basis, usually by reducing the amplitude based on an estimate of the noise. The various methods (differentiated by the suppression rule, noise estimate and other details) are collectively known as, Short-Time Spectral Amplitude (STSA), Spectral Weighting, or Spectral Subtraction methods [3] [4]. Other approaches have been reported in the literature for speech noise reduction, such as the signal subspace approach in [5] and the human auditory system model-based approaches in [6] and [7]. This paper addresses the problem of noise reduction of additive background noise in speech based on the combination in cascade of spectral subtraction and a Minimum Mean-Square Error (MMSE) log-STSA estimator in the wavelet domain. The rest of this paper is organized as follows; Section II and Section III give a review of the noise reduction strategies based on spectral subtraction and MMSE log-STSA. Section IV discusses and develops the new proposed scheme. Results are presented in Section V. Section VI gives conclusion.

II. NOISE REDUCTION BY SPECTRAL WEIGHTING

Spectral weighting means that different spectral regions of the mixed signal of speech and noise are attenuated with different factors. The aim of this process is a speech signal which contains less noise than the original one. Besides requiring a minimal distortion of the original speech, it is also important that the residual noise, i.e. the noise remaining in the processed signal, does not sound unnatural. The spectral weighting is usually performed in a transformed domain (the frequency domain). A common transform is the Fourier transform which provides an equidistant frequency solution. Let s(t) and n(t) denote speech and uncorrelated additive noise signals, and let x(t) represent the noisy observed signal. We can write:

\[ x(t) = s(t) + n(t) \]  

(1)

In the short-term Fourier domain we have:

\[ X(m, f) = S(m, f) + N(m, f) \]  

(2)

where \( m \) is the current frame and \( f \) is the frequency index. The actual spectral weighting is now performed by multiplying the spectrum \( X(m, f) \) with a real weighting function \( G(m, f) \geq 0 \). We call \( G(m, f) \) a weighting function or weighting rule. The result \( S(m, f) \) is then,

\[ S(m, f) = G(m, f)X(m, f) \]  

(3)

and the cleaned output signal \( \hat{s}(t) \) of the system is obtained by transforming \( \hat{S}(m, f) \) back into the time domain. Because of the short-time stationary property of speech, the processing has to be done on a frame-by-frame basis. A basic system for this is shown in Fig. 1. In addition to the functions shown, other ones such as framing, windowing and overlap-and-add, are also necessary [8]. Because \( G(m, f) \) is a real function, only the magnitude of \( X(m, f) \) is changed. The phase is retained for the reconstruction. The weighting function \( G(m, f) \) is usually a function of the magnitude spectra \( |\hat{S}(m, f)| \) and \( |N(m, f)| \), or of the power spectral
densities $\tilde{S}(m,f)$ and $[N(m,f)]^2$. Thus, to calculate $G(m,f)$, some estimate of the noise which should be reduced is necessary. The spectrum of the noise during speech periods is not exactly known. The basic idea is to measure the noise spectrum only when there is no speech. However, it can be estimated, since the noise is assumed to be a short-time stationary process. The estimate of the noise is taken from the speech pauses which are identified using a voice activity detector (VAD).

$$G(m,f) = \frac{1}{X(m,f)} - \frac{1}{[\tilde{N}(m,f)]^2}$$  \hspace{1cm} (5)

To prevent $\tilde{S}(m,f)$ from being negative, $\frac{1}{\tilde{N}(m,f)}$ must not be greater than $X(m,f)$. Although the noise level is reduced by the spectral subtraction, a serious disadvantage is that there will remain an unnatural sounding residual noise. It can be easily explained by the statistical nature of the noise [12]. For example, consider some frequency of the instantaneous spectrum which does not contain any speech, $X(m,f) = N(m,f)$. On the one hand, the effect of too small a noise estimate, $\tilde{N}(m,f) < [N(m,f)]$, is a remaining excitation at this frequency.

On the other hand, if the noise is estimated to be higher than it actually is, the result will be zero due to the necessary bounding, $\tilde{S}(m,f) = 0$. The result is short sinusoids randomly distributed over time and frequency, which remain in the processed signal. Normally this kind of noise is called musical noise.

B. MMSE log-STSA

To be more consistent with human auditory perception, Ephraim and Malah [4] proposed a minimum mean-squared error (MMSE) amplitude estimator in the log-spectral domain instead of in the power spectral domain as used by spectral subtraction. The MMSE log-STSA weighting rule is defined for discrete frequencies $f$ and for a frame $m$. It minimizes the mean squared error of the logarithmic spectra of the original undisturbed speech signal and the processed output signal.

$$E[(\log_{10}|S(m,f)| - \log_{10}|\tilde{S}(m,f)|)^2]$$ \hspace{1cm} (7)

where $S(m,f)$ represents the modified Bessel functions of zero and first order.

In the above equation, $I_0$ and $I_1$ represent the modified Bessel functions of zero and first order. The SNRs are calculated by:

$$R_{\text{pre}} = \beta \frac{[\tilde{N}(m,f)]^2}{|N(m,f)|} + (1-\beta)P(R_{\text{pre}})$$ \hspace{1cm} (10)

$$R_{\text{post}} = \frac{|X(m,f)|^2}{|\tilde{N}(m,f)|^2} - 1$$ \hspace{1cm} (11)

with $(0 < \beta < 1)$ is the smoothing factor, which on the basis of simulations was set to about 0.98 and $P(y) = \frac{1}{2}(y + |y|)$. The a priori SNR ($R_{\text{pre}}$) is the dominant parameter in this weighting function. Strong attenuation is obtained only if $R_{\text{pre}}$ is low, and little attenuation is obtained if $R_{\text{pre}}$ is high. The a posteriori SNR ($R_{\text{post}}$) acts as a correction factor whose influence is limited to the case where $R_{\text{pre}}$ is low. However, this reduces the musical tones artefact: For low values of $R_{\text{pre}}$ concurrent with high values of the $R_{\text{post}}$, a larger attenuation is assigned. Thus values of the
A novel noise reduction structure is proposed here based upon a combination of cascaded spectral subtraction and MMSE log-STSA methods in wavelet domain. The proposed hybrid system needs the use of the Discrete Wavelet Transform. For that, we will provide a brief introduction of wavelet transforms.

A. Discrete Wavelet Transform

Wavelet transform is a recent and promising set of tools and techniques for speech processing. Wavelets have generated a tremendous interest in both theoretical and applied areas, especially over the past few years. There exists an extensive literature addressing the wavelet transform. The discrete wavelet transform DWT can be simply thought of in terms of filter banks. A filter bank is defined as a set of filters which are applied to a signal together with changes in sampling rates. The simplest case is the two-channel filter bank which consists of a low-pass and a high-pass filter, represented by the coefficients h and g respectively. An efficient way to implement this scheme using filters was developed in 1988 by Mallat [13]. The low-pass coefficients cA can be thought of as representing a coarser approximation of the data and are known as the approximation coefficients. Correspondingly, the high-pass coefficients cD represent more detailed information in the data and are known as the detail coefficients. The other half of the story is how those components can be assembled back into the original signal with no loss of information. This process is called reconstruction, or synthesis. The mathematical manipulation that effects synthesis is called the inverse discrete wavelet transform (IDWT). Where wavelet analysis involves filtering and downsampling, the wavelet reconstruction process consists of upsampling and filtering. The algorithm of wavelet signal decomposition and reconstruction is illustrated in Fig. 2. Moreover, we must add that there are different types of wavelets such as: Haar, Daubechies, Coiflets, Symlet, Biorthogonal and etc.

In our case, we chose the Daubechies wavelet.

B. Hybrid System

The proposed speech enhancement scheme is illustrated by Fig. 3. The idea consists in enhancing the approximation coefficients (cA) and the detail coefficients (cD) resulting from DWT transformation of the noisy speech signal by Spectral subtraction and MMSE log-STSA respectively. The cleaned approximation and detail coefficients (cA, cD) are transformed in time domain using the IDWT transformation.

V. EXPERIMENTAL RESULTS

Experiments are carried out for the three different methods discussed in Section IV. We will use for these experiments frames of 25 ms with an overlap of 40% between two successive frames. To demonstrate the usefulness of the proposed scheme in the context of noise reduction application, we will compare the Spectral Subtraction and MMSE log-STSA with the proposed scheme. For the hybrid system, we will choose Daubechies wavelets of an order equal to 7. The test sentence was originally recorded under controlled conditions at a sampling frequency of 16 KHz, using 16-bits. Noise was taken from the interior of an automobile in rainy conditions. Basic objective measures, such as the Signal-to-Noise Ratio (SNR) and the Segmental SNR (SNRseg) used below, are compactly computable functions that can be uniformly applied to all forms of speech distortion in order to estimate subjective quality. Table I shows the results obtained using different input Signal to Noise Ratio SNRinput (vehicle interior noise). We can notice that the hybrid system exhibits better results than those obtained with the Spectral Subtraction and MMSE log-STSA in terms of SNRout and also of noise reduction.

For a second experiment, the noisy signal is a mix of speech and car noise. The noise is relatively stationary and has its energy concentrated in the lower frequencies. The input SNR is about 5 dB. Fig. 4 and Fig. 5 show the time evolutions and the spectrograms of a recorded noisy speech signal with the cleaned speech using the three different methods.

![Figure 2. Decomposition and reconstruction Algorithm](image1)

**Figure 2. Decomposition and reconstruction Algorithm**

h = low-pass decomposition filter; g = high-pass decomposition filter; ↓2 = down-sampling operation.

h' = low pass reconstruction filter; g' = high-pass reconstruction filter; ↑2 = up-sampling operation

![Figure 3. Hybrid system.](image2)
VI. CONCLUSION

In this paper a new scheme based upon a combination of cascaded spectral subtraction and MMSE log-STSA methods in wavelet domain was proposed for noise reduction fields. A comparative study between with other known methods was carried out to evaluate the performance of the proposed system. The experimental results have shown that our proposed hybrid system is capable of reducing noise and is an adequate procedure to improving the quality of the speech enhancement application.

REFERENCES


<table>
<thead>
<tr>
<th>( \text{SNR}_{\text{out}} ) (dB)</th>
<th>Spectral Subtraction</th>
<th>MMSE log-STSA</th>
<th>Hybrid System</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>11.22/1.60</td>
<td>8.54/0.40</td>
<td>11.20/1.56</td>
</tr>
<tr>
<td>5</td>
<td>10.30/0.75</td>
<td>7.57/0.44</td>
<td>10.20/0.61</td>
</tr>
<tr>
<td>0</td>
<td>8.23/-0.54</td>
<td>5.81/-1.61</td>
<td>7.08/-1.22</td>
</tr>
</tbody>
</table>

TABLE I

Figure 4. Temporal representations.

Figure 5. Spectrograms.