Efficient Speech Enhancement by Diffusive Gain Factors (DGF)

Hyoung-Gook Kim [1, 2]*, Klaus Obermayer*[2]*, Mathias Bode*[1]*, Dietmar Ruwisch* [1]

[1]Cortologic AG, Berlin, Germany

[2]Department of Computer Science, Technical University of Berlin, Germany

{kim, bode, ruwisch}@ cortologic.com, oby@ cs.tu-berlin.de

Abstract

In this paper we propose a very simple but highly effective algorithm for single channel noise reduction of speech signals. One of the main objectives is to find a balanced tradeoff between noise reduction and speech distortion in the processed signal. This is accomplished by a system based on spectral minimum detection and diffusive gain factors. Our approach to speech enhancement is capable of distinguishing between speech and noise interference in the microphone signal, even when they are located in the same frequency band.

1. Introduction

Noise removal in noisy speech is a very important research field with applications including suppression of environmental noise in machinery halls, voice communication systems and automatic speech recognition systems. One of the main objectives is to maximally reduce noise while minimizing speech distortion. To attain such an objective, many approaches have been reported in the literature for speech noise reduction such as many modified methods based on spectral amplitude estimation [1],[2] and Wiener filtering [3]. Even though various noise reduction methods remove the noise, they tend to introduce several realization problems in real-time processing. Furthermore, they tend to introduce a perceptually annoying residual noise usually called "musical tones". Complete removal of all the residual noise is impossible in principle because the speech signal is too tightly interlaced with the background noise.

We propose a very simple but highly effective real-time approach in order to achieve a balanced tradeoff between noise reduction and speech distortion. Instead of the complete removal of the background noise a low level of naturally sounding background noise remains in the enhanced speech signal during the proposed noise reduction. This method is based on a concept we call “spectral minimum detection and Diffusive Gain Factors (DGF-Filtering)”.

The paper is organized as follows. Section 2 describes the proposed noise suppression system. In Section 3 experimental results are presented. Finally, Section 4 concludes the paper.

2. Algorithm Description

According to Fig.1, we modify the spectral subtraction methods and [4] with practical solutions to the noise reduction problems. The input to the noise suppressor consists of the noisy speech samples s(t). As in almost all methods operating in the frequency domain the first signal processing step is the calculation of the short-time spectral magnitude $S(f,T)$ and the short-time power spectral density (PSD) $A(f,T)$ in a time frame T and frequency f of the noisy speech signal s(t). This is done using a 256 point FFT with Hann window and a frame increment of 128 samples. The actual spectral weighting is performed by multiplying the magnitude spectrum $S(f,T)$ with diffusive gain factors $F(f,T)$. The diffusive gain factors $F(f,T)$ are calculated in a two-layer structure (Fig.1):

- minimum detection layer to obtain noise power estimate using minimum values of a power spectral density of the noisy speech signal and
- diffusive gain factor computation layer which consists of two steps: preliminary gain factors and its diffusion.

Each node of a layer is responsible for a single mode of the power spectrum. The first layer called “minimum detection layer” estimates the present noise level. At first, the power spectral density of noise of each single mode is computed by using recursively smoothed periodograms:

$$N(f,T)=\alpha N(f,T\text{-}1)+(1\text{-}\alpha)A(f,T) \qquad (1)$$

where α ($0<\alpha<1$) is a smoothing constant and $N(f,T)$ is the estimation of the power spectral density (PSD) of the noise. The advantage of power spectrum estimation Eq.(1) is its computational simplicity and the fact that no measurement delay is introduced. A noise power spectrum estimate in the minimum detection layer can be obtained by detecting minimum values of the estimated noise $N(f,T)$ within windows of l frames.

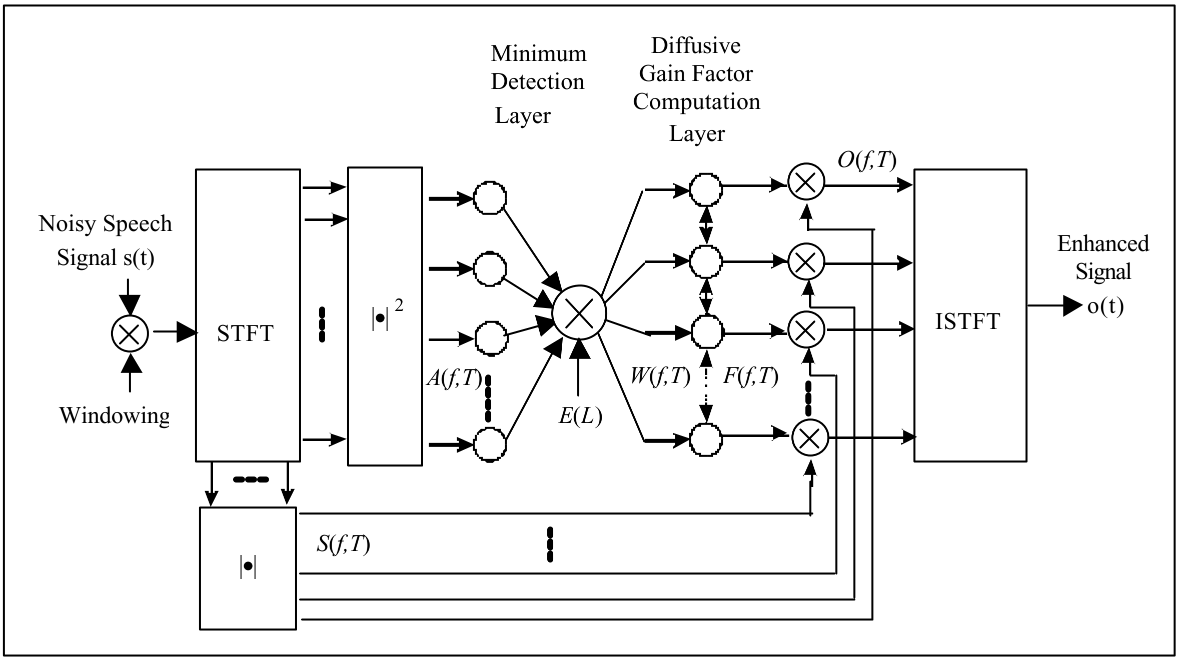

Figure 1: Block scheme of the proposed noise suppression with modified spectral subtraction. (STFT: short-time Fourier transform, ISTFT: inverse short-term Fourier transform, S(f,T): magnitude spectrum of noisy speech, A(f,T): power spectral density of noisy speech, E(L): nonlinear minimum control factor, W(f,T): actual power spectrum amplitude estimation of noisy speech, F(f,T): diffusive gain factor, O(f,T): power spectral density of filtered speech, L: normalization of the summation of the diffusive gain factor).

In noise-free speech all modes are zero from time to time. If there is a permanent offset in each mode it is supposed to be noise. Thus, a slowly varying component in each frequency channel is estimated as the minimum $M(f,T)$ by the minimum detection layer. The actual power spectrum amplitude estimation of noisy speech $W(f,T)$ is computed by nonlinear minimum estimation function $E(L)$ using the normalization of the summation of the diffusive gain factor L (see Fig.3) in order to prevent the suppression of the low energy phonemes:

If $M(f,T)E(L) > A(f,T)$

$$W(f,T) = \frac{A(f,T)}{E(L)}$$

else

$$W(f,T) = M(f,T). \qquad (2)$$

For all modes these minima are independently detected by the nodes of the minimum detection layer, one mode by one node. The Fig.2 shows our effective estimated noise spectrum of the noisy speech signal without speech pause detection. Noticeable peaks and valleys of Fig.2 exhibit a short time power spectral density of a mode of a noisy speech signal. To obtain an estimate of noise power in each mode the valleys of the noise power estimate can be used while the peaks correspond to speech activity.

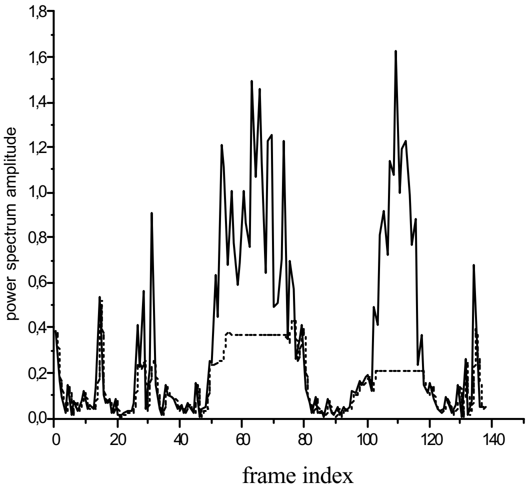

Figure 2. Example of spectral noise estimation for a noisy speech signal in mode f=2 (dotted line: the proposed estimated noise spectrum and solid line: the spectrum of the noisy speech signal

Using the above background noise estimation preliminary gain factors $C(f,T)$ can be found by "diffusive gain factor computation layer":

$$C(f,T) = 1 - \sqrt{K(L)\frac{W(f,T)}{A(f,T)}} \qquad (3)$$

where $K(L)$ denotes an overestimation factor. This overestimation factor takes into account the balanced tradeoff between reducing musical tones and reverberation artifacts. The characteristic of the overestimation factor is shown in Fig. 3.

Although these gain factors $C(f,T)$ lead to an effective removal of noise, there still remains a small low-level unnaturally sounding residual noise in parts of nonstationary car noise. At this point the performance is remarkably improved by a new processing step we call "spectral diffusion of gain factors". Thus, the diffusive gain factor interaction of neighboring modes leads to a smoothing of the gain factors $C(f,T)$:

$$F(f,T) = C(f,T) + D\,\frac{\partial^2 C(f,T)}{\partial f^2} \qquad (4)$$

where $F(f,T)$ denotes the diffusive gain factor and D is the diffusion constant. This processing step leads to a very natural sound of the output signal o(t) and helps to avoid the "musical tones".

For distinct subtraction rule, it is necessary to determine a minimum control factor and an overestimation factor. The normalization of the summation of the diffusive gain factor L is performed to limit the maximum range to values between zero and one:

$$L = \frac{\sum_{f=0}^{Z-1} F(f,T)}{Z} \qquad (5)$$

The characteristic functions to determine the parameters $E(L)$ and $K(L)$ as a function of L are given in Fig. 3. In order to preserve the low energy phonemes the power spectral amplitude estimation of noisy speech $M(f,T)$ in Eq.(2) are adjusted according to the minimum control factor $E(L)$. The overestimation factor $K(L)$ (with regard to a fixed noise floor) raises the estimated noise level in order to prevent an annoying residual noise due to fluctuations of the noise PSD and reverberation artifacts of the enhanced speech signal that result from suppression of speech component. Some of the natural sound of the speech signal can be preserved by a fixed noise floor between the overestimation factor $K(L)$ and minimum control factor $E(L)$, which additionally masks remaining musical tones with a residual noise.

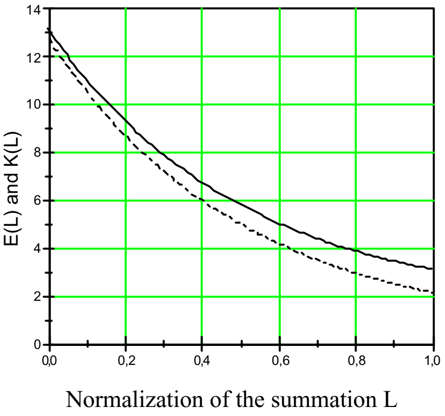

Figure 3. Overestimation factor K(L) (dotted line) and minimum control factor E(L) (solid line). The space between overestimation factor and minimum control factor denotes the fixed noise floor.

In the frequency domain, a filtering operation is performed by multiplying the noisy speech magnitude spectrum $S(f,T)$ by the diffusive gain factors $F(f,T)$ to yield $O(f,T)$. Finally, the filtered signal spectrum $O(f,T)=S(f,T)F(f,T)$ is transferred to the time domain by an inverse Fourier transform with original phases in order to calculate the output signal o(t).

3. Results

In order to assess performance of the DGF noise suppression algorithm in comparison with spectral subtraction algorithm experiments were executed using the following criteria:

- listening tests for subjective speech quality, intelligibility, and distortion
- recognition accuracy improvement
- global segmental SNR improvement

First, the algorithm was tested with different speech signals disturbed by car noise. The spectrogram of noisy speech is shown in Fig.4.(a) recorded in a car at a speed of 130 km/h with an SNR of about 0 dB. Dark gray areas correspond to the speech components. Intervals between speech utterances which are dominated by the noisy background appear as medium gray regions and in a very light gray in the Fig.4.(b) and (c) diagrams, respectively. The spectrogram of the enhanced speech signals obtained by the spectral weighting coefficients (gain factors) based on wiener filter rule [1] is depicted in Fig.4.(b). As this gain factors vary temporally very strong they have many single peaks randomly distributed over the entire time-frequency plane, which lead to the generation of the musical tones. Fig.4.(c) shows the spectrogram of the enhanced speech signal obtained by the diffusive gain factor algorithm. The speech is free of musical tones and sounds more natural compared to speech enhanced using typical spectral subtraction methods.

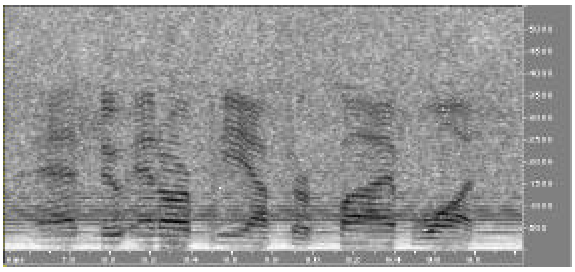

Figure 4. (a) Spectrogram of noisy speech

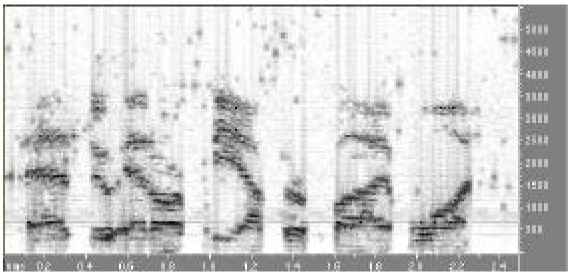

Figure 4. (b) Spectrogram of enhanced speech with typical gain factors based on wiener filter rule

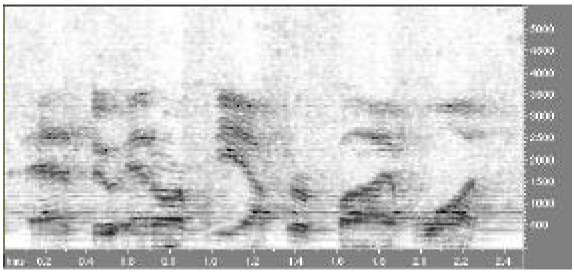

Fig.4. (c) Spectrogram of enhanced speech based on diffusive gain factors

Fig.4: Spectrograms of noisy and enhanced speech

To judge the performance, in a second experiment, we compare the recognition accuracy using the mel-scale frequency cepstral coefficient with and without our real-time noise suppression. The proposed algorithm was used to clean up speech before it is passed to a digit recognition system, which was trained on clean speech. Test speech sentences were corrupted by additive white noise (SNR=6dB). The proposed noise algorithm (DGF-filtering) front-end was compared to a spectral subtraction (SS) front-end (see Table 1).

Table 1: The isolated word recognition results and informal listening test (MOS: 1=Bad, 2=Poor, 3=Fair, 4=Good, 5=Excellent) with various speech materials. The values shown in the table are MOS points with standard deviation.

Speech Materials	Correct Words (%)	MOS
Clean Speech	95.9	4.3±0.17
Noisy speech	58.6	2.25±0.23
SS	75.7	2.48±0.12
DGF	81.7	3.7±0.16

As another signal quality measure of the proposed algorithm, the segmental signal-to-noise ratio (SNR) is computed for the filtered speech signals. The average SNR improvement of the proposed DGF Filtering is about 5.5 dB for noisy speech at 6 dB input SNR.

4. Conclusions

In this paper, a real-time approach for single channel noise reduction algorithm was presented based on spectral minimum detection and diffusive gain factors in order to maximize noise reduction while minimizing speech distortion. For stationary noise and non stationary noise it is effective in the sense that it produces a naturally sounding speech signal and suppresses the musical tones. The proposed noise suppression algorithm based on diffusive gain factors has four main features: 1) No a-priori information on the present noise and no training-phase is needed in which the pure noise and/or speech signal is present; i.e. the filter self-adjusts within a very short time (~1 sec.) during operation. 2) The signal delay is very short (~ 20 ms). 3) The suppression level can be easily adjusted and the character of the residual noise is unchanged, i.e. no or only very weak distortions appear. 4) The necessary computational power is rather small (~10 MIPS).

5. References

[1] Boll, S., "*Suppression of Acoustic Noise in Speech Using Spectral Subtraction*", IEEE Trans. on Speech and Audio Processing, vol.27, no.2, pp.113-120, 1979.

[2] Martin, R., "*Spectral Subtraction Base on Minmum Statistics*", Proc. Seventh European Siganl Processing Conference, pp. 1182-1185, 1994.

[3] Jiang, W., H. S. Malvar, "*Adaptive Noise Reduction of speech Signals*", in Technical Report MSR-TR-2000-86, July 2000.

[4] H.-G. Kim, K. Obermayer, M. Bode, and D. Ruwisch, "*Real-Time Noise Cancelling based on Spectral Minimum Detection and Diffusive Gain Factors*", In Proceedings of the 8th Australian International Conference on Speech Science & Technology, pages 256-261, 2000