MODIFIED SPECTRAL SUBTRACTION USING DIFFUSIVE GAIN FACTORS

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

This paper presents an effective method for single channel noise reduction based on the spectral subtraction algorithm. The standard spectral subtraction algorithm suffers from the fact that either the noise reduction is not satisfactory due to intelligibility of the speech is not enhanced by the introduction of annoying 'musical noise'. In order to prevent the musical tones it is important 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.

1. INTRODUCTION

Speech enhancement by noise removal for many applications, like hands-free telephoning in cars and speech recognition devices has become more and more popular. The processing required to increase both the communication comfort and the recognition rate of voice controlled systems, must suppress background noise. For single channel methods, spectral subtraction [1] is a commonly applied method, which offers the simple and computationally efficient tool for the suppression of an additive noise in a speech signal. However, its plain application suffers from essentially drawbacks: its noise estimator during speech pauses is not sufficient for the tracking of nonstationary noise and a subtraction rule tends to introduce a distortion, often called "musical noise," that is sometimes more annoying than the original noise. Many modified forms of spectral subtraction have been suggested primarily with the goal of avoiding musical noise by "over-subtraction" of the noise spectrum [2, 3].

Complete removal of all the residual noise is impossible in principle because the speech signal is too tightly interlaced with the background noise. Thus, in this paper, in order to achieve a balanced tradeoff between noise reduction and speech distortion we propose a very simple but highly effective real-time approach. Instead of the complete removal of the background noise a low level of naturally sounding background noise remains in the enhanced speech signal. This method is based on a concept we call "spectral minimum detection and Diffusive Gain Factors (DGF-Filtering)".

2. ALGORITHM DESCRIPTION

A simplified block diagram of our approach is shown in Fig. 1. A short-time spectral power function leads to a simple direct way of subtracting noise from noisy speech. Thus, the short-time Fourier analysis is applied to the input signal $x(t)$ by computing the DFT $X(f,T)$ of overlapping windowed frames, respectively, at time T and frequency f. The power spectral density $A(f,T)$ of the input signal $x(t)$ is $A(f,T)=|X(f,T)|^2$. The spectral weighting rule is performed by multiplying the magnitude spectrum $|X(f,T)|$ 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 and diffusive gain factor computation layer.
The filtered spectral values

$$O(f,T)=|X(f,T)|\cdot F(f,T) \qquad (1)$$

are transformed back into the time domain using Inverse Short-time Fourier Transformation in order to calculate the output signal $o(t)$.

2.1. Noise Estimation

The proposed first layer called "minimum detection layer" estimates the present noise level by a nonlinear estimator. Speech pause detection of the standard spectral subtraction algorithm is not needed, anymore.

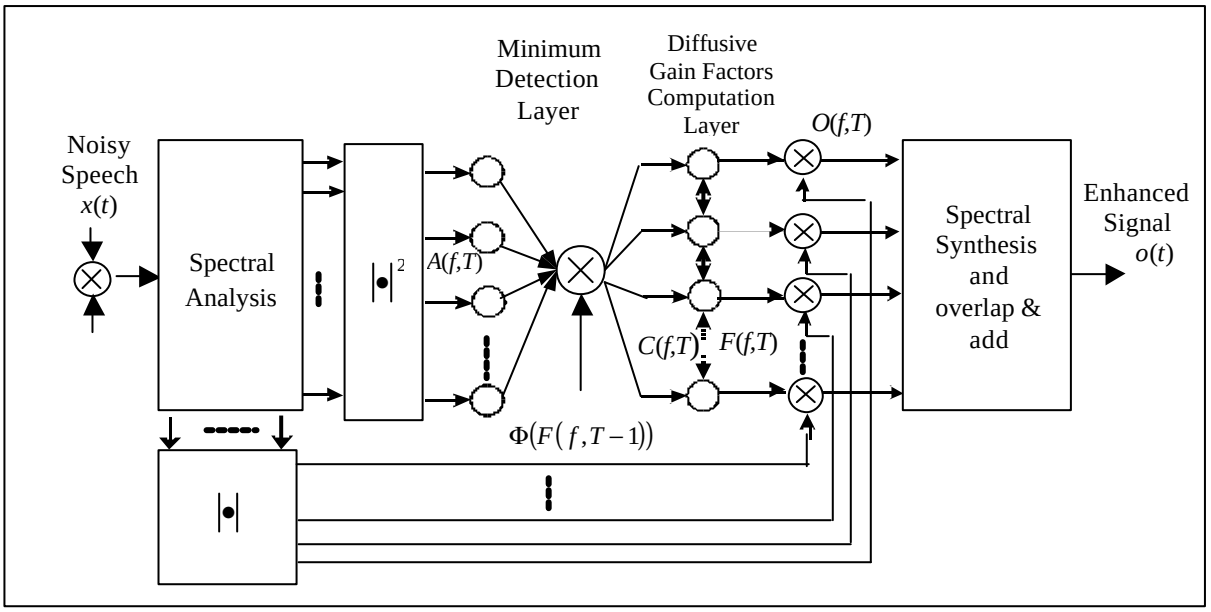

Fig. 1: Block scheme of the proposed noise suppression with modified spectral subtraction.
($S(f,T)$: magnitude spectrum of noisy speech, $A(f,T)$: power spectral density of noisy speech, $\Phi(F(f,T-1))$: nonlinear minimum control factor, $C(f,T)$: preliminary gain factor, $F(f,T)$: diffusive gain factor, $O(f,T)$: power spectral density of filtered speech)

At first, the input power spectrum of each single mode is computed by using recursively smoothed periodograms:

$$\tilde{N}(f,T)=\alpha\cdot\tilde{N}(f,T-1)+(1-\alpha)\cdot A(f,T) \quad (2)$$

where α $(0<\alpha<1)$ is a smoothing constant and $\tilde{N}(f,T)$ is the estimated noise power. Power spectral minimum values $M(f,T)$ in the minimum detection layer can be obtained by detecting minimum values of the estimated noise $\tilde{N}(f,T)$ within windows of l frames. 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. The detected minimum values are computed by nonlinear minimum estimation function $\Phi(F(f,T-1))$ (see Fig.2) using the diffusive gain factor $F(f,T-1)$ (see Eq. (8)):

$$\text{if } |M(f,T)|\cdot\Phi(F(f,T-1))>|A(f,T)|$$
$$M(f,T)=\frac{|A(f,T)|}{\Phi(F(f,T-1))} \quad (3)$$

This noise spectrum estimation is capable of distinguishing non-speech segments in the noisy speech signal.
For all modes this noise estimation is independently performed by the nodes of the minimum detection layer, one mode by one node.

2.2. Gain Computation

Spectral subtraction supplies an intuitive estimate for $|X(f,T)|$ using Eq. (3) as

$$|O(f,T)|=|S(f,T)|-|K(f,T)\cdot M(f,T)|$$
$$\cong|S(f,T)|\cdot G(f,T) \quad (4)$$

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

$$G(f,T)=1-\sqrt{K(f,T)\frac{M(f,T)}{A(f,T)}}, \quad (5)$$

where $K(f,T)$ denotes an overestimation factor using spectral floor constant R $(0.90 < R < 1)$.

$$K(f,T) = R \cdot \Phi(F(f,T-1)) \quad (6)$$

This overestimation factor must take into account the balanced tradeoff between reducing musical tones and reverberation artifacts. Although these gain factors $G(f,T)$ lead to an effective removal of noise, there remains an unnaturally sounding residual noise. A more effective method to suppress the musical tones is smoothing the filter. At this point the performance is remarkably improved by a new processing step we call "spectral diffusion of gain factors" according to the Eqs. (7) and (8) . First, recursive smoothing over time

$$C(f,T) = b \cdot G(f,T-1) + (1-b) \cdot G(f,T) \quad (7)$$

is performed, where $C(f,T)$ denotes the recursively smoothed gain factors and the parameter b is a smoothing constant. The time smoothing is effective in reducing musical noise. Nevertheless, this smoothing should not be too intensive. Otherwise, it can lead to reverberation artifacts of the enhanced speech signal. As a second step, the diffusive gain factor interaction of neighboring modes is applied to the gain factors $C(f,T)$

$$F(f,T) = C(f,T) + D \cdot \frac{\partial^2 C(f,T)}{\partial f^2} \quad (8)$$

where $F(f,T)$ denotes the diffusive gain factors and performs between zero and one. D is the diffusion constant. This processing step leads to a very natural sound of the output signal $O(f,T)$ and helps to avoid the "musical tones". Smoothing over both time and frequency can be done to obtain more accurate SNR measurements and thus less distortion.
For distinct subtraction rule, it is necessary to determine a minimum control factor. The characteristic function to determine the parameters $\Phi(F(f,T-1))$ as a function of $F(f,T-1)$ is given in Fig. 2.

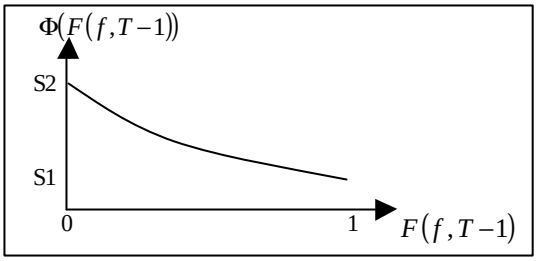

Fig. 2: Minimum Control factor $\Phi(F(f,T-1))$.

3. RESULTS

In order to visualize the functioning of the proposed noise suppression algorithm, a typical spectral power estimation and the associated gain factors are presented in Fig. 3.

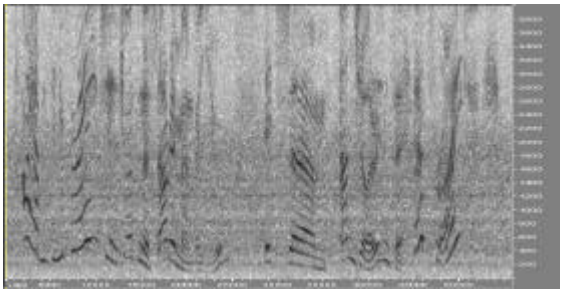

Fig. 3. (a) Spectrogram of noisy speech recorded in a car at a speed of 120 km/h with an SNR of about 5 dB.

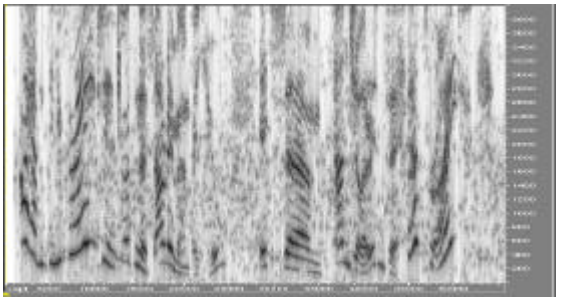

Fig. 3. (b) Spectrogram of enhanced speech with typical spectral subtraction based on wiener filter rule. Because this gain factors vary temporally very strong they have many single peaks randomly distributed over the entire time-frequency plane, which lead to the generation of the musical tones.

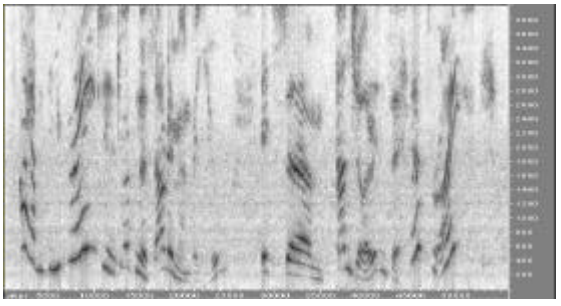

Fig. 3. (c) Spectrogram of enhanced speech based on diffusive gain factors In the output speech signals, no or only very weak musical tones are perceptible.

Fig. 3 Spectrograms of noisy and enhanced speech

To measure the performance of the proposed algorithm, the segmental signal-to-noise ratio (SNR) is computed for the filtered speech signals. The segmental SNR (*seg.SNR*) is defined as

$$seg.SNR = 10\log_{10}\frac{\sum_{t=0}^{N-1}(s(t))^2}{\sum_{t=0}^{N-1}[s(t)-o(t)]^2} \qquad (9)$$

where N is the length of the original signal $s(t)$ and $o(t)$ is the filtered output. For this, car noises were artificially added to different portions of the database at SNRs ranging from clean speech over 20 dB to 0 dB in steps of 5 dB. Ist improvement is computed as

$$improveSNR = seg.SNR_{out} - seg.SNR_{in} \cdot \qquad (10)$$

For this, car noises were artificially added to different portions of the database at SNRs ranging from clean over 20 dB to 0 dB in steps of 5 dB. The results of the SNR improvement (*improveSNR*) are presented in Fig. 4.

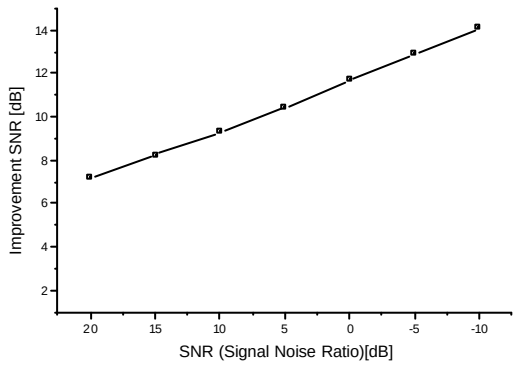

Fig. 4 Segmental SNR of a speech signal corrupted by car noise.

The results of the SNR improvement for different noise types are given in Table 1. The rows indicate the SNR results before (suffix "in") and after (suffix "out") noise reduction and the columns indicate three noise patterns; F16 noise, factory noise and pink noise.

SNR [dB]	**Noise Types (dB)**		
	F16	**Factory**	**Pink**
SNR in	-6	-9.96	-10.33
SNR out	2.47	1.66	1.64
SNR Gain	8.47	11.62	11.97

Table 1. The SNR improvement

To judge the performance, in a third experiment, we compare the recognition accuracy using the mel-scale frequency cepstral coefficient with and without our noise suppression algorithm in speaker independent isolated digit and continuous word recognizer automatic speech recognition. The proposed algorithm was used to clean up speech before it was passed to a speech recognition system, which was trained on clean speech. Test speech sentences were corrupted by additive car noise (SNR=6dB). The proposed DGF-filtering front-end was compared to a spectral subtraction (SS) front-end (see Table 2).

Speech Materials	**Correct Words (%)**	
	isolated digit	**continuous word**
Clean Speech	97.9	88
Noisy Speech	87.6	76.3
SS	92.7	79.7
DGF	94.6	84.1

Table 2. The results of speaker independent isolated digit and continuous word recognition

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

[3] 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