REAL-TIME NOISE CANCELLING BASED ON SPECTRAL MINIMUM DETECTION AND 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

ABSRACT: In this paper we propose an efficient algorithm for a one channel noise reduction in audio 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 language and noise interference in the microphone signal, even when they are located in the same frequency band.

1. INTRODUCTION

The speech enhancement of noisy speech is a very important research field with applications including suppression of environmental noise in machinery halls, mobile voice communication systems and noise suppressors for automatic speech recognition systems with the need of higher quality and intelligibility of voice. One of the main objectives is to maximally reduce noise while minimizing speech distortion. To attain such an objective, there are many different ways to perform the noise reduction in both the time and the frequency domain. Among them are methods based on spectral amplitude estimation (Boll ,1979) (Sovka et al., 1996) (Martin, 1994) (Gustafsson et al., 1998) and adaptive Wiener filtering (Hermansky et al., 1994) (Anderson et al.,1998) (Widrow et al., 1975). Even though various noise reduction methods remove the noise, they tend to introduce several realization problems in real-time processing.

Spectral amplitude estimation techniques estimate spectral information about the background noise during non-speech periods and remove this portion from the spectrum of the signal during speech periods. The noise spectrum can either be assumed as known or can be found by averaging many samples of the signal spectrum during speech pauses.

The adaptive Wiener filtering using least mean squares (LMS) or recursive least squares (RLS) based on training data is most efficient on disturbances similar to those present in the training data. However, during the operation on data with unknown noise, the noise level can be underestimated and the suppression can be slightly milder. The disadvantage of these methods is the need to record all the noise sources, which most of the time is not feasible or even impossible. Furthermore, they tend to introduce a perceptually annoying residual noise 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 in the noisy speech signal.

In this paper, we propose a very simple but highly effective psychoacoustically motivated real-time approach without assuming the training data-derived filter as in Wiener filtering and the known nonstationary noise 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 our proposed noise reduction processing. This method is based on a concept we call "spectral minimum detection and diffusive gain factors".

2. ALGORITHM DESCRIPTION

A simplified block diagram of our approach is shown in Fig. 1. 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 power spectrum A(f,T) in a time frame T of the noisy speech signal s(t). The short time power spectrum A(f,T) is estimated using a 256 point FFT with hanning window and a frame step of 128 samples. By estimating the background noise of the short time power spectrum the system calculates diffusive gain values F(f,T) in real time. The diffusive gain factors F(f,T) are calculated in a two-layer structure (Fig.1): Each node of a layer is responsible for a single mode of the power spectrum. The first layer called "minimum detection layer" collects the present noise level and provides preliminary gain factors C(f,T). The second layer performs diffusion of the gain factors C(f,T) in the "diffusion layer" to obtain the final factors F(f,T). In the frequency domain, a filtering operation is performed by multiplying the noisy speech power spectrum A(f,T) by the diffusive gain factors F(f,T) to yield O(f,T). Finally the filtered signal spectrum O(f,T)=A(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).

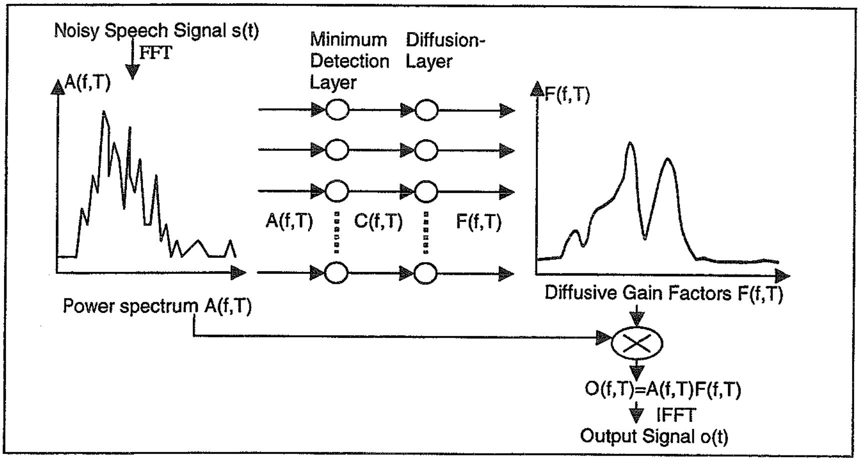


Fig. 1: Calculation of the diffusive gain factors F(f,T).

The idea of estimating the noise level within each mode is very easy: The corrupting background noise is usually assumed to be stationary and the spectral characteristics of the noise changes markedly slower than that of the speech. The suppression of slowly varying components in the noisy speech makes a good engineering sense. This fact is the basis of the minimum detection layer of our approach. A noise power estimate in the minimum detection layer can be obtained by detecting minimum values of a short time signal power spectrum of the noisy speech signal. 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, the present noise level in every single mode is assumed to be the minimum of the short time power estimate within a time interval of given length. For all modes these minima are independently detected by the nodes of the minimum detection layer, one mode by one node (Fig. 2). At first, the short-time noise spectral power of each single mode is computed by using recursively smoothed periodograms with the smoothing constant α for slowly changing signals, respectively:

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

where N(f,T) is the estimated power spectral density of the noise at the time T and frequency f and the smoothing constant is set to values between α = 0.3 ... 0.7.

Then, the minimum of the input power spectrum is detected within a windows of l frames. These minima M(f,T) at present time T are transformed into a gain factor function C(f,T) as described in Fig. 2 using a reduction control parameter K. With this reduction control parameter K between 0.90 ... 1.0 the filter level can be adjusted:

1) K=0 leads to C(f,T)=1, that means no filtering at all.
2) If K=1 then the full estimated noise level is removed.

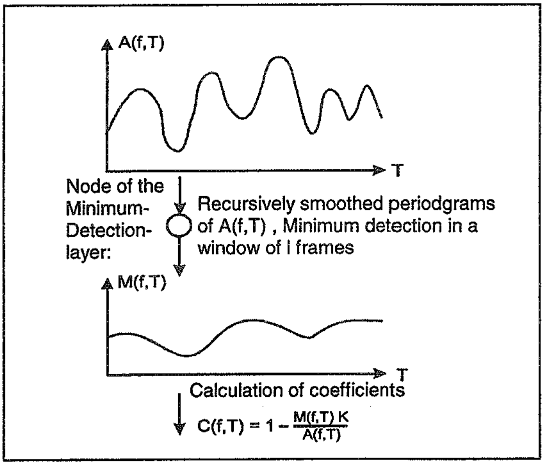

A(f,T)
T
Node of the Minimum-Detection-layer:
Recursively smoothed periodgrams of A(f,T) , Minimum detection in a window of l frames
M(f,T)
T
Calculation of coefficients
$C(f,T) = 1 - \frac{M(f,T)\,K}{A(f,T)}$

Fig. 2: The minimum detection layer estimates the present noise level by using recursively smoothed periodograms of A(f,T) for each mode and detecting the minimum M(f,T) of each mode within the last l frames. M(f,T) is transformed into a gain factor function C(f,T).

While for high values of the reduction control parameter K the algorithm essentially eliminates residual spectral peaks it affects speech quality such that some of the low energy phonemes are suppressed. To limit this undesirable effect the minima M(f,t) are adjusted according to the mode- signal to noise ratio (SNR) levels. The reduction control parameter can be smaller in case of a better signal to noise ratio.

The obtained gain factor C(f,T) offers a performance superior to conventional spectral subtraction with a speech activity detector and Wiener optimum filters based on minimizing the mean squared error based on training data. Although these gain factors C(f,T) leads to an effective removal of noise, there still remains a small low-level unnaturally sounding residual noise in parts of nonstationary car noise. Thus, the new processing step called the "diffusion of gain factors" is performed in the diffusion layer (Fig. 3). The diffusive gain factor interaction of neighboring modes quoted in Fig. 3 leads to a smoothing of the filter coefficients C(f,T). This processing step leads to a very natural sound of the output signal o(t) and helps to avoid the “musical tones” being a common problem of similar noise suppressing algorithms.

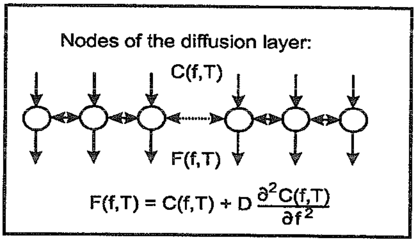
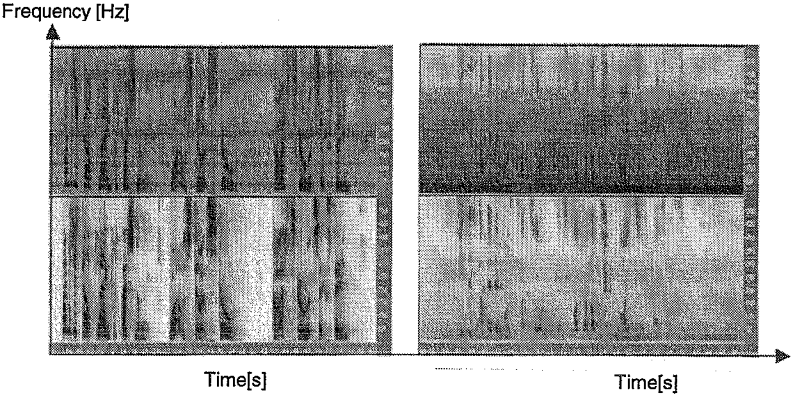

Nodes of the diffusion layer:
C(f,T)
F(f,T)
F(f,T) = C(f,T) + D ∂²C(f,T)/∂f²

Fig. 3: In the diffusion layer there is an interaction of neighboring modes leading to a smoother shape of the diffusive gain function F(f,T) (D can be interpreted as a diffusion constant). This " diffusion of the gain factors" avoids the "musical tones" that are a common problem of noise reduction algorithms.

3. RESULTS

The algorithm was tested with different speech signals disturbed by car noise. The noisy spectrograms shown in the upper images of Fig.4 recorded in a busy street with a signal to noise ratio (SNR) of about 5 dB (a) and in a car at a speed of 130 km/h with an SNR of about –10 dB (b). The spectrograms of the enhanced speech signals obtained by the proposed algorithm are depicted in the lower parts (a) and (b). Dark gray areas correspond to the speech components. Intervals between speech utterances which are dominated by the noisy background appear as medium gray regions in the upper diagrams and in a very light gray in the lower line, respectively. This picture clearly indicates that only speech portions pass the system whereas the noise is suppressed. A low-level naturally sounding background noise derived from the original noise of the processed signal is preserved and gives the far end user a feeling of the atmosphere at the near end.



(a) car noise in a busy street (b) car noise at a speed of 130 km/h

Fig. 4: Spectrograms of the noisy signal (upper) and the enhanced signal (below)

In addition to the acoustic impression of the noise suppression system there is a lucid description of the filter performance, namely giving its modulation characteristic. In our real-time noise suppressor the parameters are tuned in a way that the maximum transmission for each mode is found at modulation frequencies of approximately 1 Hz (Fig. 5), corresponding to the typical modulation of human speech. Much slower and faster modulating signal components are recognized as not belonging to the speech signal and thus are considerably damped.

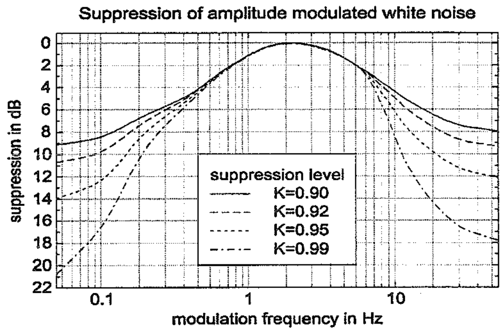


Fig. 5: Modulation characteristic of the Real-Time Noise Suppressor with parameters l=348, D=0.25.

Methods	Assumption	Training for the Adaption	Musical Tones	Calculation time	Listening test (MOS)
Power Spectral Subtraction (PSS)	Voice Activity Detector	No training	very noticeable	very short	unpleasant (2.48 ± 0.12)
Wiener Filtering & PSS	All the Noise Source	necessary	noticeable	short	speech distortion (3.97 ± 0.45)
Kalman Filtering	Noise Variance	necessary	weak musical tones	large	light speech distortion (3.65 ± 0.35)
Dual Extended Kalmam Filtering	Noise Variance	necessary	no musical tones	very large	unnatural smoothing (4.42 ± 0.39)
Real-Time Noise Cancelling	No Assumption	No training (self adjustment)	no musical tones or natural low-level noise	short	minor speech distortion (4.66 ± 0.18)

Table. 1: The proposed algorithm was compared with other methods according to an informal listening test (MOS) with various speech material. The values shown in the table are MOS points with standard deviation.

4. CONCLUSION

In this paper, a real-time approach for a one 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 tne sense that it produces a naturally sounding speech signal and suppresses the musical tones. Additionally, our approach is simple and the computational power needed to excute the algorithm is small. Five main features of this proposed noise suppression algorithm are shown in Table. 1 in comparison with other methods.

REFERENCES

Anderson, D. V., Clements, M. A. (1988) *Noise Suppression in Speech Using Multi-resolution Sinusoidal Modeling,* presented at the Fall 1998 Meeting of the Acoustical Society of America, Norfolk, VA.

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

Gustafsson, S., Jax, P. & Vary, P.(1998) *A Novel Psychoacoustically Motivated Audio Enhancement Algorithm Preserving Background Noise Characteristics,* Proceedings ICASSP'98, Seattle, USA.

Hermansky, Hynek., Wan, E. A., Avendano, Carlos (1994) *Noise Suppression in Cellular Communications,* in Proceedings IEEE IVTTA'94, pp. 85-88, Kyoto, Japan.

Martin, R. (1994) *Spectral Subtraction Base on Minmum Statistics,* Proc. Seventh European Siganl Processing Conference, pp. 1182-1185.

Sovka, P., Pollak, P., Kybic, J. (1996) *Extended Spectral Subtraction,* Signal Processing VIII Theories and Applications, volume 2, pp.963-966, EUSIPCO-96.

Widrow, B., Grover, J. R., McCool, J. M., Kaunitz, J., Williams, C. S., Goodlin, R.C., (1975) *Adaptive Noise Cancelling: Principles and Applications,* Proceedings of the IEEE, 63:1692-1716.