PRECISE NOISE ESTIMATION FOR HIGH NOISE REDUCTION

Michael Walker

sound-acoustics@onlinehome.de
Sound acoustics research, Schulstr.24/1, 73669 Lichtenwald, Germany

ABSTRACT

This contribution concerns single channel noise reduction principles (NR) in the frequency domain.
High noise reduction is needed for speech coders, packetized speech transmission (VOIP), speech recognition and many other applications. Discussed approaches [1] deal with a limited noise reduction degree (NRD) at poor environment conditions. In some applications NRD is reduced to provide full speech bandwidth or might be increased leading to quality loss in case of a poor signal to noise ratio (S/N).
This report describes a solution differentiating between various noise portions. Noise floor, short term and narrowband distortions are combined to a new noise portion estimator, taking also non stationary distortions as e. g. musical noise phenomena into account.
The modeling of noise portions by its properties offers new possibilities for improved speech processing. Combined in a hybrid solution, using the advantages of the frequency and time domain, an increase of the NRD from 15 dB to more than 40dB even under poor S/N (≥ 0dB) conditions becomes feasible.

1. INTRODUCTION

Many speech enhancement systems try to use the advantage of selective signal processing in the frequency domain [1]. For analysis and synthesis the block wise computed Fast Fourier Transformation (FFT) is preferred by economical reasons in spite of some difficulties [2].
The quality of NR systems depends strongly on the reliability, robustness and accuracy of the estimated noise to be eliminated. Exact noise estimation is still an unsolved problem as it is not feasible to determine all noise portions fast and precise enough with present solutions.
The main problem is caused by the fixed frequency bins of a frequency analysis which are not synchronized with the unknown noise. Thus the observed energy of the frequency bins changes from block to block with an unpredictable magnitude and periodicity acting as non stationary noise portions (see Fig. 1).

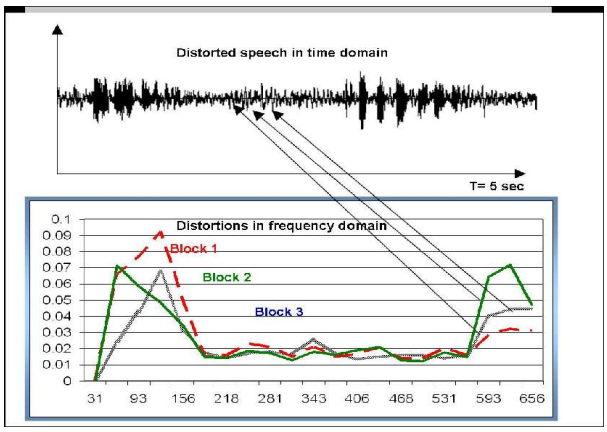

Fig. 1. *Noise in time and frequency domain*

Many of known noise estimation methods neglect the non stationary noise portions, leading to a loss in the feasible speech enhancement performance. The problem can be defused with a reliable detection and elimination of remaining noise portions.

2. OVERVIEW

Fig.2 shows the system arrangement of a typical noise reduction system in the frequency domain.

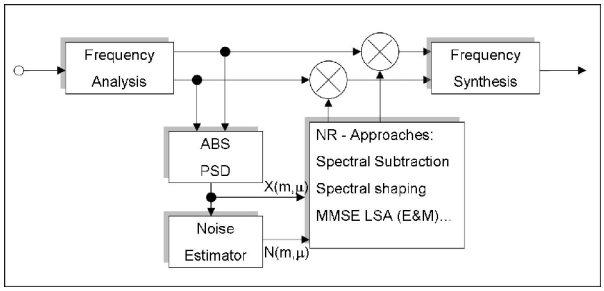

Fig. 2 *Noise reduction in frequency domain*

The estimated noise influences the NRD and quality of the noise suppression independent of the used NR approach.

Noise estimation might be carried out by following means:

2.1. Smoothing and averaging

A simple noise estimator might consist only of a low pass filter with a high response time. It is known, that this method provide only a row estimation value with poor quality results, as it follows changing environment noise not fast enough leading to poor noise reduction behavior. Speech degradation is introduced if the low pass filter is not controlled by a reliable voice activity detector (VAD) [3].

2.2. Soft decision methods

Soft decision methods try to overcome the problem of a hard VAD control with statistical approaches taking the probability of noise and speech presence [4] into account. These means provide faster adaptation behavior with less speech degradation.

2.3. Minimum statistics

The minima of the smoothed spectral coefficients [1] correspond to a low noise floor with advantageous fast adaptation behavior. Furthermore it provides quite natural speech quality, as it prevents over estimation. Non stationary noise portions might be handled as speech portions, reducing the maximum feasible NRD.

2.4. Combinations

The use of minimum statistics on a summarized spectrum [5] improves the NRD limitation by a loss in the available signal to noise margin. If the bandwidth of the observation window exceeds the range of distortions, speech coefficients might be affected impairing the speech quality by a loss in the frequency response.

2.5. Résumé

The strong variance of the noise in the frequency domain leads to noise portions, which are not really detected and removed. The robustness of NR systems might be achieved by a loss in the speech quality or by reduced NRD under poor S/N conditions with present solutions.

3. PRECISE NOISE ESTIMATION FOR HIGH NOISE REDUCTION

Noise is consisting of several noise portions in the frequency domain and can be expressed as:

$$N(m,\mu) = nmin(m,\mu) + nvar(m,\mu) + nnb(m,\mu) + nbb(m,\mu) \quad (1)$$

The noise portions are computed at each time block m and for each frequency bin μ separately. These dependencies will be dropped for simplification of following equations if not required.

3.1. Noise portion *nmin*

The noise portion *nmin* corresponds to the stationary noise floor.

$$nmin = \begin{cases} if\,(vad = 0) \\ \alpha 1 \cdot X + \beta 1 \cdot nmin \\ else \\ \min(X, nmin) \end{cases} \quad (2)$$

A voice activity detector controls the kind of update with the input signal *X* (Fig. 2). During speech pauses (*vad=0*) the update is carried out by a recursive smoothing filter. The filter coefficients (*a1=1-b1*) determine the response time and are frequency and time dependent. During speech activity (*vad=1*) the minimum is used as shown in (2).

3.2. Noise portion *nvar*

The margin between this noise floor and the required estimated noise for the computation of clean speech should be as less as possible, but is needed to achieve noise reduction at all. In some approaches a fixed threshold is determined for this purpose. The variance portion *nvar* (3) of the noise is used to determine an adaptive margin fulfilling this requirement with better performance.

$$nvar = \begin{cases} if\,(vad = 0) \\ \alpha 2 \cdot |X - n_{var}| + \beta 2 \cdot nvar \end{cases} \quad (3)$$

The weighted sum *nmv* of both noise portions (4) corresponds to the stationary noise without any other noise portion as e. g. narrowband (NBD) and short term distortions (STD).

$$nmv = w1 \cdot nmin + w2 \cdot nvar \quad (4)$$

nmv is needed for the detection of the following non stationary noise portions.

3.3. Noise portion *nnb*

The property of a narrow band distortion (NBD) is the small bandwidth. The number of the current excited frequency bins *enb(m)* within a time interval determines the current bandwidth of the present signal. Speech signal is consisting of high number excited frequencies (formants). A small number of excitations are expected to be narrowband distortions.

$$enb(m) = \sum_{i=0}^{M} \begin{cases} if\left(X(i) > \gamma \cdot nmv(i)\right) \\ enb + 1 \end{cases} \quad (5)$$

With *i(k)=i(k-1)*MSC* and the initialization *enb=0*, the count of excited frequency bins, distributed according to the MEL scale [6], is carried out acc to (5). This MEL – distribution is needed to avoid the count of adjacent excited frequency bins. With γ the lower threshold of the observation window is determined.

To avoid strong influence on small speech parts, mainly at the beginning of an utterance under poor S/N conditions, *enb(m)* is averaged acc. to (6) with L-th order.

$$enb_{av}(m) = \frac{1}{L} \cdot \sum_{l=0}^{L} enb(m-l) \quad (6)$$

In (7) the noise portion of NBD distortions is determined. The threshold ϑ corresponds to the least expected number of excitations.

$$nnb(m) = \begin{cases} if\left(enb_{av}(m) < \vartheta\right) \\ X(m) - nmv \end{cases} \quad (7)$$

3.4. Noise portion *nbb*

The *nbb* portion concerns STD distortions introduced by short broad band excitations. For the detection of these disturbances the intermediate result

$$ni = nmin + nvar + nnb \quad (8)$$

of the previous noise portions is needed (8). The condition for the excitation within a time interval of a FFT block is indicated by *ebb(m)* (9).

$$ebb(m) = \begin{cases} 1\ if\ (X(m,\mu) > ni(m,\mu) \\ 0\ else \end{cases} \quad (9)$$

The *nbb* noise portion depends on the a priori and a posteriori excitations. Speech presence can not be assumed, if only one short term excitation in between occurs.

$$nbb(m) = \begin{cases} if\left(\overline{ebb(m+1)} \wedge \overline{ebb(m-1)}\right) \\ X(m) \\ else\ \ 0 \end{cases} \quad (10)$$

The time interval in (10) depends on the block length of the FFT removing block distortions caused by the asynchronous operation concerning the noise frequencies. The length of this time interval can be adapted to another noise portion as e. g. crackle noise and other.

3.5. Arrangement in time and frequency domain

Fig. 3 shows the block diagram of the new noise estimation principle. The accurateness depends still on the reliability of the VAD.

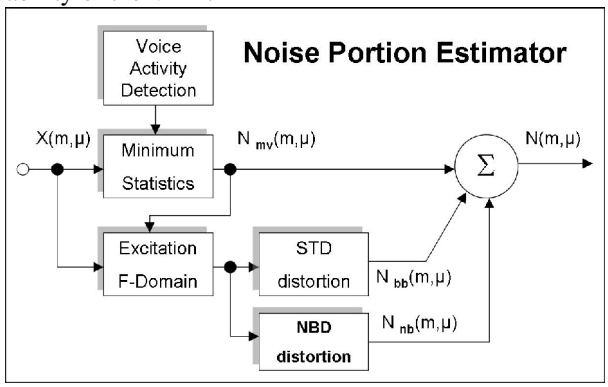

Fig. 3: *Principle o the proposed noise estimator*

The advantage of the computation in the frequency domain is the selective differentiation of course. The draw back in this domain is the high variance of the frequency bins. The advantage in the time domain is the high number of available samples, making a system change earlier and with more reliability detectable. With the system arrangement in Fig. 4 the advantages of both domains can be used.

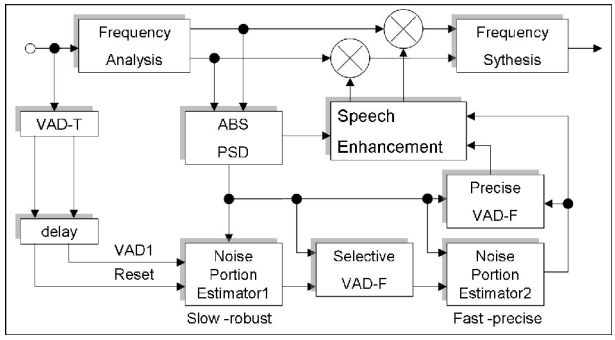

Fig. 4: *Hybrid solution in time and frequency domain*

A first noise portion estimator is controlled by a VAD which is processed in the time domain. This noise estimator is used for a second selective VAD in the frequency domain. A second noise portion estimator, controlled by a selective VAD allows fast and precise noise estimation in the frequency domain.

The system can be extended by a third VAD indicating the speech presence if only one frequency bin exceeds the estimated noise. The third VAD might be used to

control data transfer for packetized speech transmission or to improve the reliability of speech recognizers.

4. RESULTS

The described principle needs about 3MIPS computational power at a sample rate of 8 kHz. Several tests for the evaluation of the quality improvement have been carried out:

4.1. Speech recognition and packetized Speech transmission

a) The speech recognition rate was tested with a database consisting of 2830 Words under highly mismatched conditions. The word recognition rate could be improved from 62% with a conventional noise estimator to 86,11% with the proposed estimator.

b) The data transmission to a speech recognizer has been controlled by the third VAD of the described principle i. o. to reduce the data transfer and so the observation time of the recognizer. With a selection of 5 most critical speech files (S/N=0dB) the recognition rate could be improved from 35% to 100%! Subjective tests with cancelled speech pauses proved the necessity for high noise reduction to avoid the perception of cut out noise parts.

4.2. Objective measurement results

Objective measurements [7] have been carried out with a mixed database consisting of car and street noise with S/N from <-2 to >22 dB.

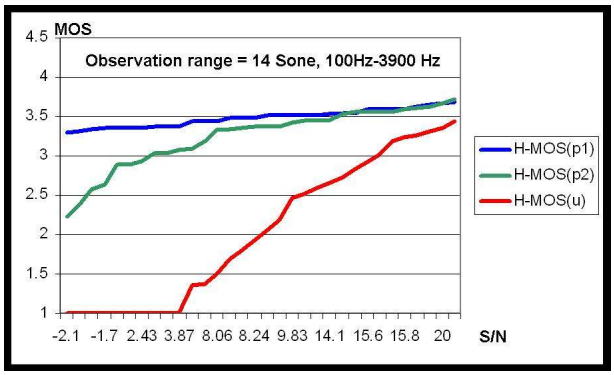

Fig. 5: *Objective results*

The quality results in Fig. 5 show the MOS values of the new principle H-MOS (p1) compared with the original distorted data H-MOS (u) and the processed Data with a conventional noise estimator H-MOS (p2).
Subjective tests correspond to the results in Fig. 5.

4.3. Real time tests

The new principle was implemented in a Windows application program using the Media Control Interface for sound blaster cards. Thus different hardware platforms (pc, soundcard, microphone…) with various sampling rates could be tested in real time.
Real time tests proved high long term stability and reliable fast adaptation behavior, if e.g. the microphone was moved from the table to a running fan of a pc, or the vacuum cleaner was switched on during operation.
Higher sampling rate sounds more natural. A quite remarkable quality increase is perceptible from 8 to 11 kHz according to subjective quality assessment.

5. CONCLUSION AND OUTLOOK

With the noise portion estimator a strong and reliable noise reduction became feasible in spite of a low computational complexity. The principle might be extended if new realizations about another distortion models exist. New application- dependent noise portions as e. g. rest echoes to be cancelled or codec distortions to be suppressed are imaginable.

6. REFERENCES

[1] Rainer Martin: “Statistical methods for enhancement of noisy speech” *International Workshop on Acoustic Echo and noise Control (IWAENC),* September 2003

[2] Michael Walker "Speech improvement by noise reduction based on a Continuous Fourier Transformation” *International Workshop on Acoustic Echo and noise Control (IWAENC),* September 2001

[3] Virgine Gilg, Christophe Beaugeant, Martin Schönle, Bern Andrassy „Methodology for the design of a robust voice activity detector for speech enhancement” *International Workshop on Acoustic Echo and noise Control (IWAENC)* ,September 2003

[4] Jongseo Sohn and Wonyong Sung “A voice activity detector employing soft decision based noise spectrum adaptation” *IEEE 6/98, pp 365-368*

[5] Markus Schwab, Hyoung-Gook Kim, Wiriady and Peter Noll “Robust Noise Estimation Applied to Different Speech Estimators” *Department of Communication Systems Technical University of Berlin, Germany.*

[6] E. Zwicker„Psychoakustik“ *Springerverlag 1982, ISBN 3-540-11401-7*

[7] Michael Walker "Hearing Adequate Signal Quality Estimator” *DPA patent pending, internal documents by Sound acoustics.*