A Noise Suppression System for the AMR Speech Codec

P. Jax, R. Martin, P. Vary, M. Adrat, I. Varga*, W. Frank*, M. Ihle*

Institute of Communication Systems and Data Processing,
RWTH Aachen, Templergraben 55, D-52056 Aachen
E-Mail: jax@ind.rwth-aachen.de

* Siemens AG, ICM CD
Grillparzer Straße 10-18, D-81675 München
E-Mail: imre.varga@mch.siemens.de

Abstract

In this paper we describe a noise reduction preprocessing algorithm for the adaptive multirate (AMR) speech codec of the GSM system. The algorithm is based on spectral weighting and explicitly takes into account the properties of the human auditory system. The weighting rule results in the smallest possible speech distortion under the constraint that the background noise should exhibit no audible distortions.
The algorithm was implemented in 16 Bit fixed-point arithmetic and submitted to the ETSI AMR noise reduction standardization contest. Compared to other algorithms, our noise reduction method gave very good results in CCR tests and good results in ACR tests.

1 Introduction

Today, mobile phones are used in various acoustic scenarios including environments with strong acoustic background noise of different kinds, e.g. car, street or babble noise, interfering talkers, music etc.
Thus, in 1998 ETSI decided to develop a noise suppression (NS) algorithm as an optional feature of the AMR codec [1]. The noise suppression function is a preprocessing module in front of the speech encoder of the mobile terminal. It is used to improve the signal to noise ratio (SNR) prior to speech coding and in this way improves speech quality and ease of conversation.
To guarantee minimum performance levels, ETSI subgroup SMG11 developed a set of design constraints concerning the subjective quality of the speech enhancement preprocessor and complexity. The performance of submitted algorithms was evaluated by a number of formal listening tests according to these constraints.
This paper describes an algorithm developed under the ETSI design constraints as well as test results obtained by this solution.

2 Algorithm

The proposed noise suppressor acts as a preprocessing front-end to the AMR encoder. The basic concept of our algorithm is to allow deviations from a constant frequency independent noise attenuation only when these deviations are masked by speech. Thus, in a psychoacoustical sense, a uniform and “musical noise” free noise reduction is achieved.
A block diagram of the algorithm is shown in **Fig. 1**. In the following subsections each of the blocks is described briefly.

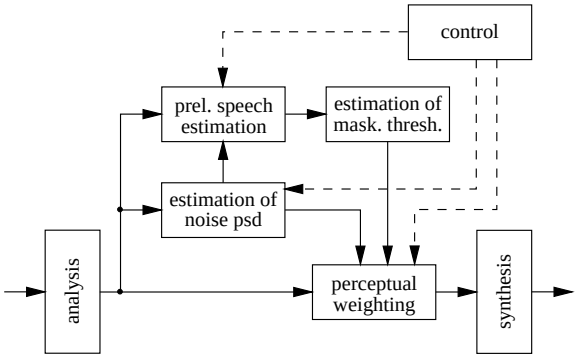

Fig. 1: Block diagram and main signal flow of the noise reduction algorithm

2.1 Analysis and Synthesis

Since the processing is performed on a frame-by-frame basis in the frequency domain, the noise reduction system employs a FFT based analysis-synthesis filterbank.
The noisy input signal is sampled with a sampling frequency of 8 kHz. The input signal is first seg-

mented into frames using the same frame rate of one frame per 160 samples (20 ms) as used by the AMR codec. Each frame consists of the 200 most recent input samples such that adjacent frames overlap by 40 samples (5 ms). Due to the overlapping of adjacent frames of 40 samples, the noise reduction causes an additional algorithmic delay of 5 ms compared to the stand-alone operation of the AMR codec.

A pre-emphasis filter amplifying high frequency components is applied to the signal segments. This helps to reduce quantization noise in the fixed-point implementation.

Before transforming into the frequency domain, each signal segment is multiplied by a flat-top Hann window. The rising and falling parts of this window function consist of 40 samples each. In-between, the window function is constantly one. Each signal segment is padded with zeros to a length of 256 samples. The signal blocks are then transformed into the frequency domain by means of the Fast Fourier Transform (FFT) algorithm.

The Fourier coefficients of the output signal are calculated by multiplying the Fourier coefficients of the input signal with a real-valued and positive weighting vector which is derived according to the weighting rule described in subsection 2.4. Hence, the phase of the Fourier coefficients is not modified.

After this modification of the amplitudes of the Fourier coefficients, the speech frame is transformed back into the time domain by an Inverse Fast Fourier Transform (IFFT). Afterwards, the overlapping speech segments are reassembled by the overlap-and-add method. Then a de-emphasis filter which is the inverse of the pre-emphasis filter is applied to the signal.

Finally, signal segments of the 160 latest fully reconstructed samples are transferred to the AMR codec.

2.2 Noise Estimation

The power spectral density (psd) of the background noise is estimated using the Minimum Statistics (MINSTAT) approach [2]. This method utilises the fact that a (short term) stationary background noise forms a “spectral floor” in the smoothed modified periodogram of the noisy signal.

First, the smoothed modified periodogram of the input signal is calculated. Then, for each frame and each frequency bin i the power spectral density $R_n(i)$ of the noise component is estimated by determining the minima of the periodogram of the input signal over a sliding window of a fixed number of previous frames and applying an over-estimation factor.

This noise estimation algorithm needs no voice activity detector. Hence, it allows a continuous adaptation of the estimated noise psd also during periods of speech activity. As a result, fast tracking of non-stationary background noise is achieved.

2.3 Preliminary Clean Speech Signal Estimation

The aim of this part of the algorithm is to derive a first estimate of the clean speech signal. This estimate is used as the input to the algorithm which estimates the masking threshold that is needed for the final weighting rule.

The core of this preliminary speech estimation procedure is the well-known weighting rule proposed by Ephraim and Malah which aims at minimizing the mean-squared error of the log-spectral amplitudes (MMSE LSA) of the Fourier coefficients of the speech estimate [3, 4]. Furthermore, the MMSE LSA gain vector is weighted by a soft-decision vector which takes the probability of speech absence (or presence) into account [5, 6]. For this purpose the weighting rule uses three input quantities, namely the a posteriori and the a priori Signal-to-Noise Ratios (SNRs) as well as speech absence probabilities.

The a posteriori SNR is defined as the ratio between the current periodogram of the noisy input signal and the psd of the noise. Since both quantities can be easily estimated, the calculation of the a posteriori SNR is straight forward.

The a priori SNR is defined as the ratio between the psd of the clean speech and the psd of the background noise. Since the clean speech psd is not explicitly available, the estimate of the a priori SNR is based on the a posteriori SNR and the output signal of the noise reduction algorithm for the previous frame (*decision-directed approach* [3]).

Due to the fact that speech is non-stationary and may not be present in every frequency bin, especially during voiced speech, the speech absence probabilities are tracked individually for each frequency bin and continuously over time [6]. This tracking procedure is based primarily on exploiting the a posteriori SNR.

The preliminary speech estimate is finally calculated by multiplying the Fourier coefficients of the input speech signal by the weighting vector derived according to the MMSE LSA weighting rule.

2.4 Weighting Rule based on Psychoacoustic Criteria

This final weighting rule is based on masking properties of the human auditory system [7]. The masking threshold $R_t(i)$ is estimated using the preliminary clean speech estimate as the masker. This estimation is performed by means of a simple auditory model and involves several steps. First, the result of an initial

critical band analysis is convolved by a spreading function. Then a threshold offset is applied and normalizations are performed.
The desired amount of noise reduction in the psychoacoustical sense is defined by a scalar noise attenuation factor ζ . Accordingly, the weighting factors $H(i)$ for the individual frequency bins i are chosen in such a way that all components of the residual noise which exceed the desired amount are just "hidden" below the estimated masking threshold:

$$H(i) = \sqrt{\frac{R_t(i)}{R_n(i)}} + \zeta$$

The value of $H(i)$ is then limited to values smaller than one.
This weighting method results in the smallest possible speech distortion for the desired amount of noise reduction.

2.5 Control of the Algorithm

In order to obtain optimal results for various kinds of acoustic situations, the averaged a posteriori SNR of the noisy input signal is continuously scanned. The noise reduction algorithm is adjusted according to this parameter [8].

3 Implementation

The noise suppression algorithm was implemented in 16 Bit fixed-point arithmetic using ANSI C and ETSI basic operations [9]. These basic operations include a mechanism to measure the maximal computational complexity of the algorithm which is expressed in *weighted million operations per second* (WMOPS). Additionally the usage of the different kinds of memories had to be evaluated. The measured complexities are summarized in **Table 1**.

	Complexity	design constraint
WMOPS	3.386	5
Dynamic RAM (words)	2234	3039
Static RAM (words)	718	1500
Data ROM (words)	863	1000
Program ROM (ETSI basic operations)	772	2000

Table 1: Summary of the computational complexity of the noise suppression algorithm and the ETSI design constraints.

4 Evaluation

During the AMR noise suppression selection phase the proposals were tested in a variety of test conditions. These tests took place in different independent test laboratories and were performed in several languages.
The ETSI testing rules define a number of experiments as well as minimum performance levels for the evaluation of the different test conditions:

- Quality during the initial convergence time (informal test with expert listeners)
- Degradation in clean speech (pair comparison test)
- Artifacts and clipping effects in background noise conditions (ACR test)
- Performances in background noise conditions (CCR test)
- Performance in background noise: Influence of propagation errors (CCR test)
- Performances in background noise: Influence of VAD/DTX (CCR test)
- Influence of the input signal + noise level and performances with special noises (ACR test)

Due to the large number of test conditions, only a small subset of the test results will be described in the following subsections.

4.1 Artifacts and Clipping in Background Noise

The goal of this test is to assess the subjective quality of the background noise in the processed speech signal. The test was performed as an *Absolute Category Rating* (ACR) test, i.e. the listeners had to assess each of the presented speech samples using an absolute *Mean Opinion Score* (MOS). The candidates were instructed to make their judgement of the sample "considering unnatural sound during the complete sample".
Results of this experiment for the English language and for different noise situations are shown in **Fig. 2** for low SNR and in **Fig. 3** for high SNR. In the selected sub-experiment the AMR coder operates at its highest bit rate of 12.2 kBit/s.
From both Fig. 2 and Fig. 3 it can be seen, that the noise suppression preprocessing helps the AMR coder to reduce unnatural sounds and artifacts in the background noise. Such unnatural sounds typically occur as coding artifacts when coding speech with high level background noise.
The advantage of the noise suppression preprocessing is especially dominant for stationary noises (car and street noise). In babble noise, the stand-alone AMR

coder already performs very well and thus the effect of additional NS preprocessing is low. Furthermore, for the instationary babble noise the amount of noise reduction is lower than for the other noise types.

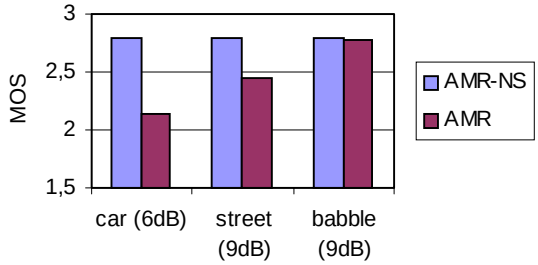

Fig. 2: Results of the test concerning "artifacts and clipping in background noise" for low SNR.

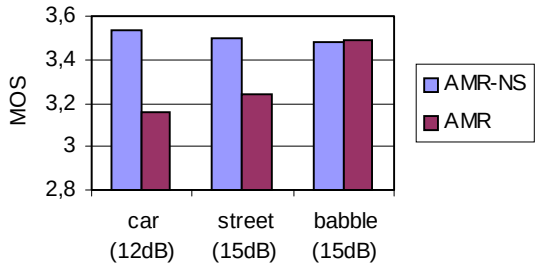

Fig. 3: Results of the test concerning "artifacts and clipping in background noise" for high SNR.

4.2 Performance in Background Noise

The performance of the NS preprocessor in background noise was evaluated formally by a *Comparison Category Rating* (CCR) test. In this test the listener has to assess the quality differences between two samples – a reference sample and the sample under test. A rating of zero indicates that there is no difference between the samples. The reference samples for the results presented in **Fig. 4** are the speech samples processed by the stand-alone AMR codec at a bitrate of 12.2 kBit/s. The low and high SNR of the input speech was 6 and 12 dB for the car noise and 9 and 15 dB for street and babble noise, respectively.

The results from Fig. 4 show a significant preference of the listeners for those samples, which were preprocessed by the noise suppression algorithm. The speech enhancement system yields best results for stationary noises such as car noise. For non-stationary and more speech-like background signals such as babble noise, the CMOS rating is smaller. In such cases the noise suppression algorithm does not succeed in reducing the noise as much as for stationary noises. However, also for such background noises the enhanced and coded signal is still significantly preferred over the signal processed by the stand-alone AMR codec.

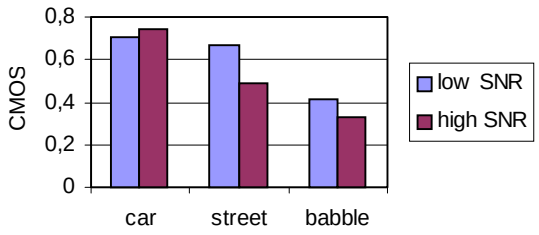

Fig. 4: Results of the CCR test comparing the transmission with noise suppression and AMR codec with the stand-alone AMR coder.

5 Conclusion

The proposed algorithm in conjunction with the AMR speech coder results in significant improvements for various background noise situations such as car noise, street noise and office babble. Furthermore, it has been shown, that a fixed-point implementation of an advanced speech enhancement algorithm is possible within the tight ETSI design constraints.

6 References

[1] ETSI, „Digital Cellular Telecommunications System (Phase 2+); Noise Suppression for the AMR Codec; Service Description; Stage 1". GSM 02.76

[2] R. Martin, „Spectral Subtraction Based on Minimum Statistics". Proc. EUSIPCO, 1994

[3] Y. Ephraim, D. Malah, „Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator". IEEE Trans. ASSP, Dec. 1984

[4] Y. Ephraim, D. Malah, „Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator". IEEE Trans. ASSP, Apr. 1985

[5] R.J. McAulay, M.L. Malpass, „Speech Enhancement Using a Soft-Decision Noise Suppression Filter". IEEE Trans. ASSP, 1980

[6] D. Malah, R.V. Cox, A.J. Accardi, „Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments". Proc. ICASSP, May 1999

[7] S. Gustafsson, P. Jax, P. Vary, „A Novel Psychoacoustically Motivated Audio Enhancement Algorithm Preserving Background Noise Characteristics". Proc. ICASSP, May 1998

[8] R. Martin, I. Wittke, P. Jax, „Optimized Estimation of Spectral Parameters for the Coding of Noisy Speech". Proc. ICASSP, Istanbul, June 2000

[9] ETSI, „AMR permanent document (AMR-9) : Complexity and Delay Assessment". Tdoc SMG11 AMR 44/98