QUALITY OF NETWORK BASED ACOUSTIC NOISE REDUCTION

Matthias Pawig and Peter Vary

Institute of Communication Systems and Data Processing, RWTH Aachen University
{*pawig, vary*}*@ind.rwth-aachen.de*

Abstract: In this contribution, we analyze the behavior of a single channel noise reduction system when it is moved from the mobile telephone to a network based unit, i.e., when it suppresses the noise after transmission with a speech codec. State-of-the-art algorithms have been designed primarily for small terminal equipment, taking the constraints of low complexity and low memory consumption into account. In order to ease these constraints and to reduce the signal processing complexity at the mobile terminal, the acoustic noise reduction may be moved to a network based unit. This contribution shows by simulations for the exemplary AMR 12.2 kBit codec that the difference between single channel background noise reduction before and after a speech codec is far less severe than usually assumed, and the objective measures of both systems are very close for a variety of scenarios.

1 Introduction

The reduction of background noise noise has been widely studied in the past and remains an active field of research. Generally, single microphone systems depend on an estimation of the background noise power spectrum using, e.g., the Minimum Statistics approach [12, 17]. Once the power spectral density (PSD) of the noise has been estimated, a spectral weighting rule is employed to calculate the spectral weighting gains and to suppress the noise in the frequency domain.

State-of-the-art algorithms have been designed primarily for small terminal equipment, taking the constraints of low complexity and low memory consumption into account. In order to ease these constraints and to reduce the signal processing complexity at the mobile terminal, the acoustic noise reduction can be moved to a network based unit (NBU), e.g. at the entry point into the provider network like the transcoding unit in GSM (TRAU) or at a management unit for audio conferencing.

For noise reduction in a NBU, the additional effects of the speech codec have to be considered. State-of-the-art codecs like the Adaptive Multi-Rate codec (AMR) are parametric codecs, which have been optimized for speech signals. Generally, it is assumed that noise reduction before encoding improves the overall transmission quality significantly as the parametric codecs have been optimized for the properties of clean speech signals.

In this paper, extensive simulations will be described which quantify the differences between noise suppression at the mobile terminal and noise suppression in a network based unit. The first simulated system is a single channel noise reduction system followed by encoding and decoding with the AMR 12.2 kBit codec. In the second system, encoding and decoding is performed *first* and followed by the same single channel noise reduction system as before. Both systems are then rated by the Perceptual Evaluation of Speech Quality (PESQ) and the segmental noise attenuation and the segmental speech attenuation as well as by informal listening tests.

The paper is structured as follows: In Section 2, the compared systems are explained in detail, including a description of the noise reduction algorithm and an overview of the used speech codecs. In Section 3, the simulation results are be presented, followed by some conclusions.

2 System Description

2.1 Single-Channel Noise Reduction

The single channel noise reduction system employed in this paper is based on spectral decomposition of the noisy input signal using statistical noise suppression techniques. An overview is shown in Figure 1.

First, the noisy microphone signal $y(k)$ consisting of the clean speech $s(k)$ and the noise $n(k)$ is segmented into overlapping frames with frame index λ. These frames are transformed into the frequency domain by short-time fourier transformation (STFT). The noise suppression is applied to the signal by multiplying frequency gains $G(\lambda,\mu)$ which aim to minimize noise components while preserving the speech signal depending on a mathematical cost function. The frequency bins are denoted by μ. After applying the gains, the output signal $S'(\lambda,\mu)$ is transformed back into the time-domain resulting in the output signal $s'(k)$.

In order to calculate the gains $G(\lambda,\mu)$, the noise power spectral density (PSD) as well as the signal to noise ratio (SNR) are estimated using the noisy input spectrum $Y(\lambda,\mu)$. The estimation of the noise PSD $\sigma_N^2(\lambda,\mu)$ is a crucial component of speech enhancement systems. Various single-channel noise PSD estimation algorithms can be found in the literature. A comparison of some state-of-the-art estimators is presented in [15]. In this paper, the *Minimum Statistics* algorithm is employed which is able to update the estimated noise PSD even during speech activity [12]. This approach estimates the noise by tracking the minimum of the noisy PSD As this minimum is always smaller or equal to the mean noise power, a bias correction is necessary. For further details we refer to the literature.

In addition to the estimation of the noise PSD, most statistical noise suppression techniques require estimates of the *a posteriori SNR* $\gamma(\lambda,\mu)$ and *a priori SNR* $\xi(\lambda,\mu)$. The a posteriori

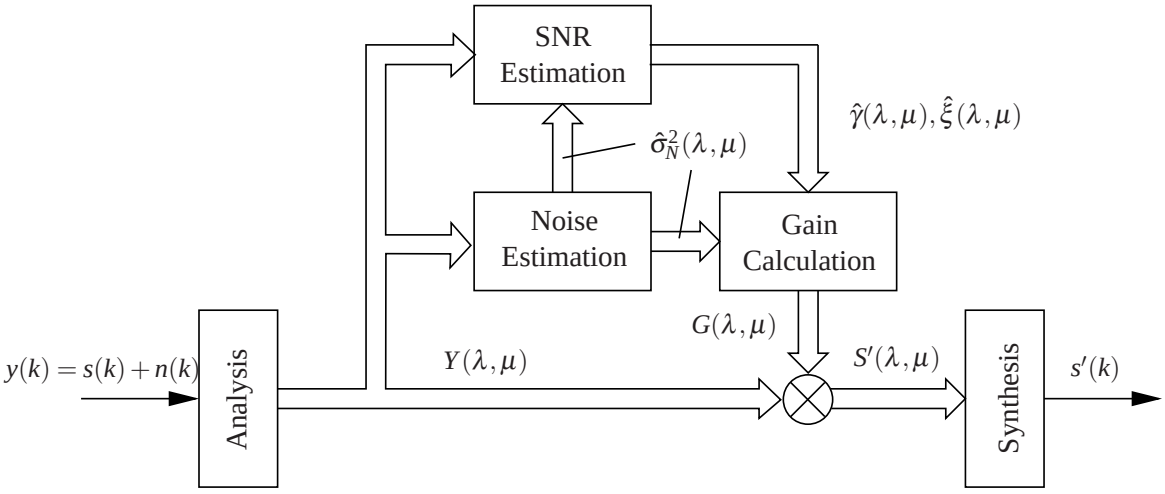


Figure 1 - System block diagram of a conventional noise suppression system working in the frequency domain.

SNR is defined as the ratio between the noisy periodogram and the noise PSD according to:

$$\gamma(\lambda,\mu) = \frac{|Y(\lambda,\mu)|^2}{\sigma_N^2(\lambda,\mu)}. \tag{1}$$

The a posteriori SNR can easily be calculated with the estimated noise PSD $\sigma_N^2(\lambda,\mu)$. In order to estimate the a priori SNR $\xi(\lambda,\mu)$, the widely accepted decision-directed approach [2] is used. This approach linearly combines estimates from previous frames with an instantaneous SNR realization relying on the a posteriori SNR according to:

$$\hat{\xi}(\lambda,\mu) = \alpha_{\mathrm{DD}} \frac{|\hat{S}(\lambda-1,\mu)|^2}{\hat{\sigma}_N^2(\lambda-1,\mu)} + (1-\alpha_{\mathrm{DD}}) \cdot \max\big(\hat{\gamma}(\lambda,\mu)-1, 0\big), \tag{2}$$

where $\max(\cdot,\cdot)$ returns the maximum of its two arguments. The smoothing factor α_{DD} adjusts the trade off between noise reduction and speech distortions. Here it is chosen as $\alpha_{\mathrm{DD}} = 0.98$.

The estimated noise PSD and a priori SNR as well as a posteriori SNRs are employed to calculate the weighting gains $G(\lambda,\mu)$. Two different weighting rules have been considered:

1. The *Wiener filter* is derived from the optimal filter theory [18, 11]. It is a linear estimator that minimizes the mean square error between the clean speech DFT coefficients $S(\lambda,\mu)$ and the enhanced DFT coefficients $S'(\lambda,\mu)$. The derivation [18] results in a weighting gain $G_W(\lambda,\mu)$ which is only dependent on the a priori SNR $\xi(\lambda,\mu)$:

$$G_W(\lambda,\mu) = \frac{\xi(\lambda,\mu)}{\xi(\lambda,\mu)+1}. \tag{3}$$

2. The *minimum mean square error log spectral amplitude* (MMSE-LSA) estimator determines only the magnitudes of the short-time Fourier coefficients of the clean speech signal [3] since information of the phases has minimal influence on the performance of a noise reduction system. In order to put more emphasis on small speech spectral amplitudes which are very important for speech intelligibility, this weighting rule minimizes the mean square error of the logarithmically weighted amplitudes as follows:

$$\mathrm{E}\{(\ln(A(\lambda,\mu)) - \ln(\hat{A}(\lambda,\mu)))^2\} \quad \rightarrow \quad \min. \tag{4}$$

   The derivation results in

$$G_{LSA}(\lambda,\mu) = \frac{\xi(\lambda,\mu)}{1+\xi(\lambda,\mu)} \exp\left(\frac{1}{2}\int_{\nu(\lambda,\mu)}^{\infty} \frac{\exp(-t)}{t} dt\right) \tag{5}$$

   where $\nu(\lambda,\mu) = \frac{\xi(\lambda,\mu)}{1+\xi(\lambda,\mu)}\gamma(\lambda,\mu)$.

2.2 GSM-EFR/ AMR Codec

The influence of a state-of-the-art transmission of the speech signal to the network based unit is taken into account by using the most commonly used codec for mobile communication, the *GSM Enhanced Full Rate* (GSM-EFR) codec [10, 4]. The 12.2 kBit mode of the *Adaptive Multi-Rate* (AMR) codec [5, 1] is used, which is identical to GSM-EFR. In the following a short description with respect to the elements which could deteriorate noise reduction performance will be given.

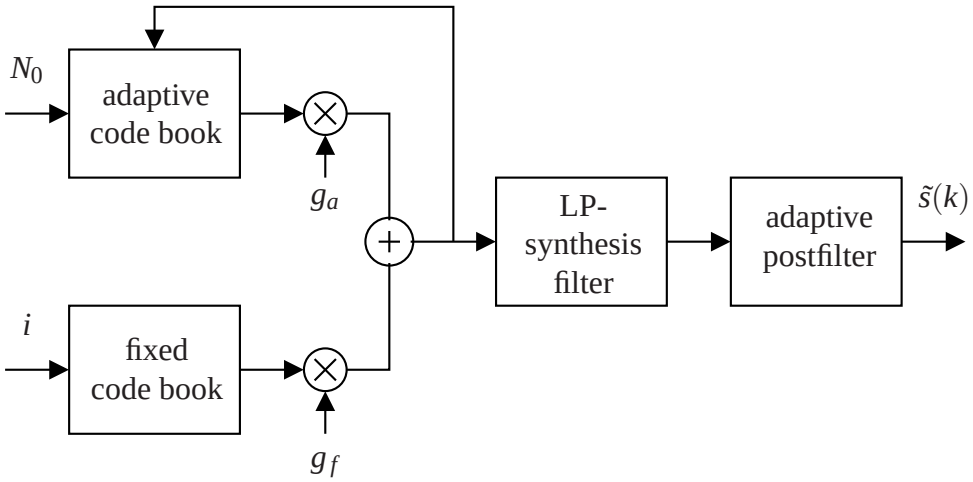


Figure 2 - Simplified block diagram of the AMR 12.2 kBit decoder

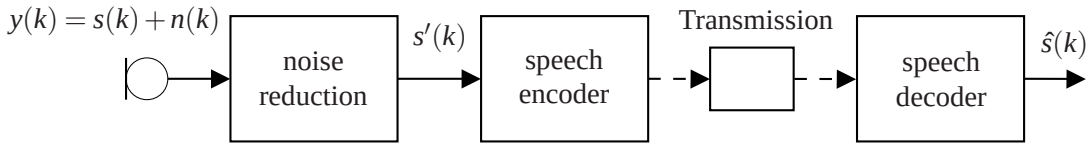
The codec used is a parametric codec, which means that the input signal is analyzed and several parameters are transmitted which can be used at the decoder to synthesize an output signal which sounds very close to the input signal. The AMR codec uses analysis-by-synthesis; the input vector is systematically encoded with different parameters. The parameter set which is transmitted as the coded bitstream is the set which minimizes the perceptively weighted error between input and output signal. Therefore, only some aspects of the decoder will be explained here, for a more detailed explanation we refer to the literature.

A simplified block diagram of the decoder structure is shown in Figure 2. The basic principle employed is *Code Excited Linear Prediction* (CELP), which means that the linear prediction residual is quantized by a code book which represents a vector quantizer. The transmitted parameters are the coefficients of the linear prediction (LP) filter, the pitch lag N_0, the fixed code book index i and the weighting gains g_a for the adaptive code book and g_f for the fixed code book. All of these blocks are optimized for clean speech signals, which means that the performance will be reduced for noisy input signals.

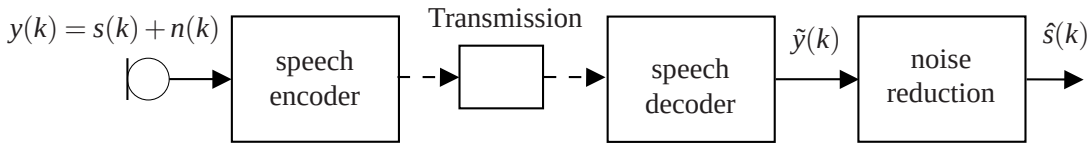
In particular, the adaptive code book represents a form of long term prediction, which is suitable for the pitch structure of voiced sounds, but less for noise signals. The fixed code book is used to quantize the linear prediction residual. Since it has a limited number of entries, a similar but not identical representation of the residual is received at the decoder. Noisy input signals should not influence the fixed code book too much, since the residual is spectrally flat for perfect linear prediction. The linear prediction filter is used to recover the spectral envelope of the speech signal. In case of noisy input signals, a part of the coefficient set will represent the envelope of undesired noise components instead of speech components, reducing the coding quality for the speech signal. The adaptive post-filter can reduce the perceived quantization noise of the speech codec while enhancing the structure of the speech components. It should not be influenced too much by a noisy input signal and will reduce some of the influence of the noise.

2.3 System Models

The block diagram of the two systems which are compared with each other are illustrated in Figure 3. System a) is a model for most noise reduction systems employed in today's mobile terminals. The noisy microphone signal $y(k)$ is fed into a single-channel noise reduction system resulting in a noise reduced signal $s'(k)$. The influence of the codec on the signal is then modelled by processing $s'(k)$ with a speech encoder followed by a speech decoder. Finally, the



(a) State-of-the-art system



(b) Proposed system with noise reduction after transmission

Figure 3 - Block diagram of both systems

resulting output signal of the overall system is denoted as $\hat{s}(k)$.

System b) is the model for a network based noise reduction system. In this case, the input signal $s(k)$ is transmitted without processing over the transmission model of speech encoder and decoder, resulting in the signal $\tilde{y}(k)$. The model for processing in the network is the same noise reduction algorithm as in system a) employed for the signal $\tilde{y}(k)$. The output of this stage is then the overall system output signal $\hat{s}(k)$.

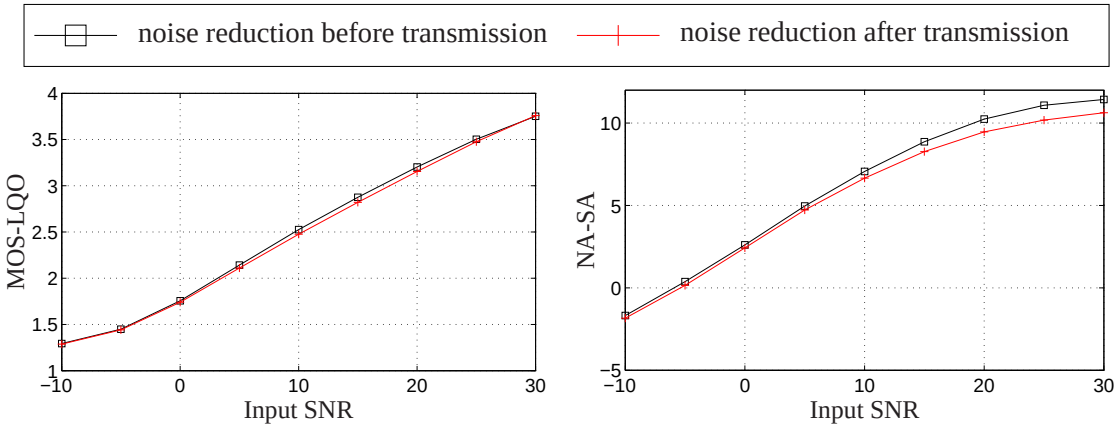
3 Simulation Results

In order to quantify the effects of moving the noise reduction algorithm to the network, simulations were carried out for noise reduction systems using either of the two weighting rules described in 2.1. The speech input signals were taken from the NTT database [13], using a total of around 1 hour of speech from different speakers both male and female. The sampling rate was 8000 Hz. These inputs signals were then disturbed by two different noise signals 'babble' and 'factory1' taken from the Noisex database [16] and fed through the Systems a) and b) described in Section 2.3.

The input and output signals were then analyzed by two different objective measures. The first measure used is the *Perceptual Evaluation of Speech Quality* (PESQ) [14, 8] which perceptually rates the difference between reference signal and noisy signal and can be mapped [9] to a *Mean Opinion Score* (MOS-LQO).

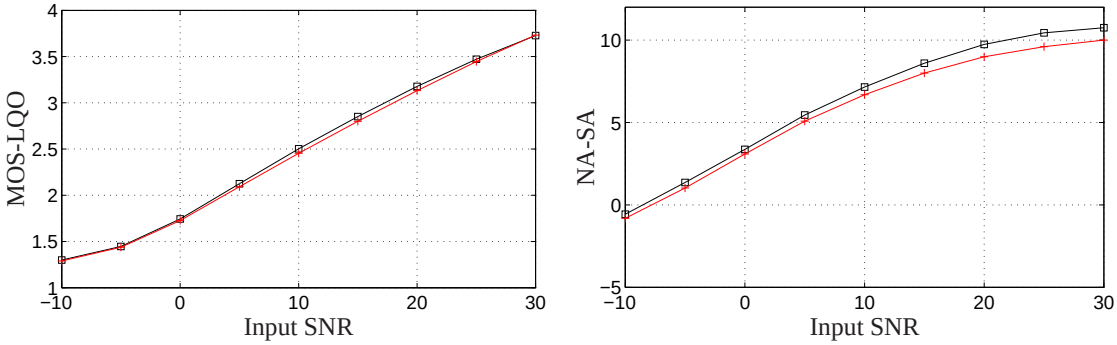
The second objective measure is the difference between *Noise Attenuation* and segmental *Speech Attenuation* (NA-SA). The goal of noise reduction is to maximize the noise attenuation while keeping the speech attenuation as small as possible, so this difference gives an indication of the performance of the analyzed algorithm. Noise attenuation was determined by comparing the noise signal $n(k)$ with the signal $\hat{n}(k)$ produced by filtering and transmission of the noise signal alone. Both signal were segmented into L segments of length M with 20 ms duration and used to calculate the noise attenuation as

$$NA = 10\log_{10}\left(\frac{1}{L}\sum_{\lambda=1}^{L}\frac{\frac{1}{M}\sum_{m=0}^{M-1}n(\lambda M+m)^2}{\frac{1}{M}\sum_{m=0}^{M-1}\hat{n}(\lambda M+m)^2}\right). \tag{6}$$



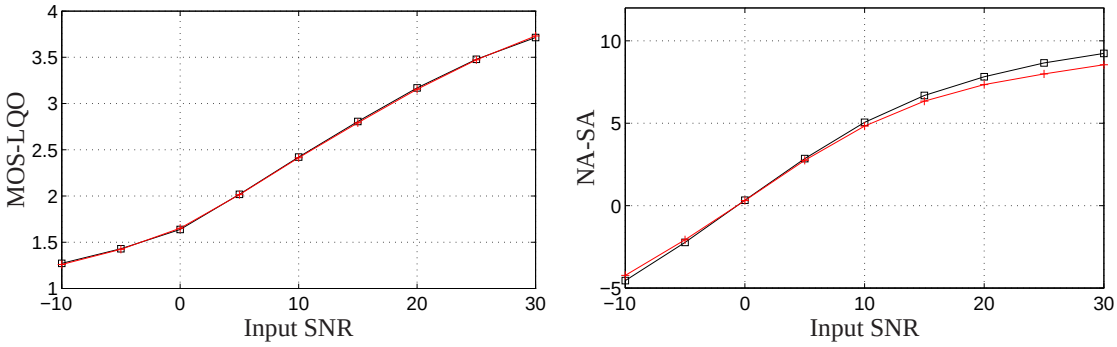
(a) Speech quality according to PESQ (b) Noise attenuation minus speech attenuation

Figure 4 - Simulation results for factory noise, Wiener filter.



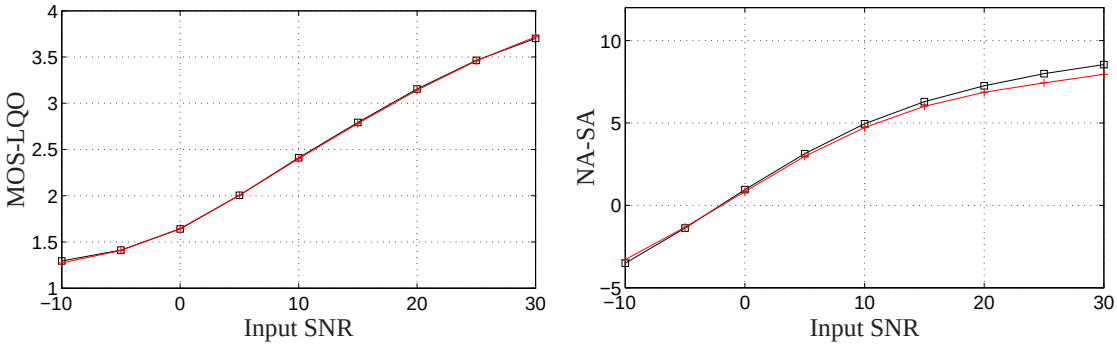
(a) Speech quality according to PESQ (b) Noise attenuation minus speech attenuation

Figure 5 - Simulation results for factory noise, MMSE-LSA filter.



(a) Speech quality according to PESQ (b) Noise attenuation minus speech attenuation

Figure 6 - Simulation results for babble noise, Wiener filter.



(a) Speech quality according to PESQ (b) Noise attenuation minus speech attenuation

Figure 7 - Simulation results for babble noise, MMSE-LSA filter.

Speech attenuation was determined in the same way for the speech input $s(k)$ and signal $\hat{s}'(k)$ produced by filtering and transmitting the speech signal $s(k)$ alone. However, only the L' segments with speech activity are used for the calculation

$$SA = 10\log_{10}\left(\frac{1}{L'}\sum_{\lambda=1}^{L'}\frac{\frac{1}{M}\sum_{m=0}^{M-1}s(\lambda M+m)^2}{\frac{1}{M}\sum_{m=0}^{M-1}\hat{s}'(\lambda M+m)^2}\right). \tag{7}$$

In order to produce the signals $\hat{n}(k)$ and $\hat{s}'(k)$, the weighting gains of the noise reduction algorithm were used to filter the clean speech signal $s(k)$ and the noise signal $n(k)$ separately. The effect of the transmission and thus the codec were approximated by using the black box approach of [7], calculating a linear filter approximation of the codec and filtering speech and noise alone.

The simulation results are shown in Figures 4 to 7. It can be observed that all simulations show the same tendency. For the simulations with factory noise, a very small difference can be seen between both systems, rating the systems with noise reduction before transmission slightly higher. The same tendency is true for the NA-SA measure, where the state-of-the-art systems also rates slightly higher. However, the differences in both results is extremely small and in a range where it is barely audible. In case of babble noise, the difference in the PESQ measurements disappears completely. For the NA-SA rating, the gap between both systems closes even more.

4 Conclusions

In this paper, the effect of moving the noise reduction algorithm from a mobile terminal to a network based unit was quantified. Simulations for different noise scenarios and noise reduction filters were carried out for the example of a transmission with the widely used GSM-EFR/ AMR 12.2 kBit codec.

It was shown that the performance of a network based noise reduction system is reduced by the transmission slightly for this scenario, however the effect is barely measurable. Since a network based unit would be less limited by the constraints of a mobile terminal in terms of complexity (typically 5 MOPS according to ETSI [6]) and memory, it should be possible to improve the overall system performance even more by using a more sophisticated and thus more complex algorithm in the network, while still saving signal processing complexity at the mobile terminal. The reduced complexity of the mobile terminal results in a reduction of the required energy of the signal processor and thus in a longer battery life time. It should be noted that this result also shows that noise reduction at the receiver can be beneficial if the transmitting device has insufficient noise reduction performance.

References

[1] EKUDDEN, E. ; HAGEN, R. ; JOHANSSON, I. ; SVEDBERG, J. : The adaptive multi-rate speech coder. In: *Proc. IEEE Workshop Speech Coding*, 1999, pp. 117–119

[2] EPHRAIM, Y. ; MALAH, D. : Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. In: *IEEE Transactions on Speech and Audio Processing* 32 (1984), No. 6, pp. 1109–1121

[3] EPHRAIM, Y. ; MALAH, D. : Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator. In: *IEEE Transactions on Speech and Audio Processing* 33 (1985), Apr., No. 2, pp. 443–445

[4] ETSI: *ETSI EN 300 726: Digital cellular telecommunications system (Phase 2+) (GSM); Enhanced Full Rate (EFR) speech transcoding (GSM 06.60 version 8.0.1 Release 1999)*. 2000

[5] ETSI: *ETSI EN 301 704: Digital cellular telecommunications system (Phase 2+) (GSM); Adaptive Multi-Rate (AMR) speech transcoding (GSM 06.90 version 7.2.1 Release 1998)*. 2000

[6] ETSI: *ETSI TR 126 978: Digital cellular telecommunications system (Phase 2+); Universal Mobile Telecommunications System (UMTS); Results of the AMR noise suppression selection phase (3GPP TR 26.978 version 6.0.0 Release 6)*. 2004

[7] FINGSCHEIDT, T. ; SUHADI, S. ; STEINERT, K. : Towards objective quality assessment of speech enhancement systems in a black box approach. In: *ICASSP*, 2008, pp. 273–276

[8] ITU-T: *ITU-T Rec. P.862: Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*. 2001

[9] ITU-T: *ITU-T Rec. P.862.1: Mapping function for transforming P.862 raw result scores to MOS-LQO*. 2003

[10] JARVINEN, K. ; VAINIO, J. ; KAPANEN, P. ; HOONKANEN, T. ; HAAVISTO, P. ; SALAMI, R. ; LAFLAMME, C. ; ADOUL, J.-P. : GSM enhanced full rate speech codec. Munich, Germany, 1997, pp. 771–774

[11] LIM, J. S. ; OPPENHEIM, A. V.: Enhancement and Bandwidth Compression of Noisy Speech. 67 (1979), Dec., No. 12, pp. 1586–1604

[12] MARTIN, R. : Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics. In: *IEEE Transactions on Speech and Audio Processing* 9 (2001), No. 5, pp. 501–512

[13] NTT-CORPORATION: *Multi-Lingual Speech Database for Telephonometry*. Tokyo, Japan, 1994

[14] RIX, A. W. ; BEERENDS, J. G. ; HOLLIER, M. P. ; HEKSTRA, A. P.: Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In: *ICASSP*. Washington, DC, USA, 2001. – ISBN 0–7803–7041–4, pp. 749–752

[15] TAGHIA, J. ; TAGHIA, J. ; MOHAMMADIHA, N. ; SANG, J. ; BOUSE, V. ; MARTIN, R. : An Evaluation of Noise Power Spectral Density Estimation Algorithms in Adverse Acoustic Environments. Prague, Czech Republic, May 2011

[16] VARGA, A. P. ; STEENEKEN, H. J. M. ; TOMLINSON, M. ; JONES, D. : The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition / Speech Research Unit, Defense Research Agency. Malvern, UK, 1992. – Forschungsbericht

[17] VARY, P. ; MARTIN, R. : *Digital Speech Transmission*. John Wiley and Sons, LTD, 2006

[18] VASEGHI, S. V.: *Advanced Signal Processing and Digital Noise Reduction*. Chichester, UK : John Wiley & Sons, 1996