# OPTIMIZED ESTIMATION OF SPECTRAL PARAMETERS FOR THE CODING OF NOISY SPEECH

*Rainer Martin, Ingo Wittke, and Peter Jax*

Institute of Communication Systems and Data Processing
Aachen University of Technology, D-52056 Aachen, Germany
Phone: +49 241 806984, Fax: +49 241 8888 186, E-mail: martin@ind.rwth-aachen.de

## ABSTRACT

In this contribution we optimize a speech enhancement preprocessor such that a distortion measure in the Line Spectral Frequency (LSF) domain is minimized. We can thus improve the estimation of spectral parameters of a speech coder when the input signal to the coder is a noisy speech signal. The optimization aims at the maximum noise reduction of the enhancement preprocessor. The average maximum noise reduction characteristic is determined as a function of the speech signal SNR and is approximated by an exponential function. Since LSF parameters are widely used in speech coding the results are applicable to a wide range of speech coders and enhancement preprocessors. We report experimental results for an MMSE Log Spectral Amplitude estimator in conjunction with the new ETSI Adaptive Multi-Rate (AMR) speech coder. We found that the method is most effective for the low bit rate coding modes.

## 1. INTRODUCTION

For many years speech coding research has aimed at reducing the bitrate of coded speech signals while maintaining a high level of speech quality and intelligibility. While this endeavour has been very successful for clean speech (as manifested in new standards such as the ITU G.729 and the ETSI AMR codec) the coding of noisy speech becomes significantly more difficult as bit rates are decreasing.

Informal and formal listening tests show that improvements are obtained when the speech coder is combined with a speech enhancement preprocessor. However, to obtain optimal results, the preprocessor needs to be specifically adapted for a given speech coder at a given bitrate [1]. This optimization should take the parameters of a speech coder, especially the spectral parameters like LPC or LSF coefficients, explicitly into account [1, 2]. The spectral parameters are in fact very important for speech intelligibility [2] and are influenced by the noise attenuation characteristics of the enhancement preprocessor. Since a low noise attenuation does not remove a sufficient amount of noise from the disturbed signal and a high noise attenuation will also distort speech components it is not quite obvious how much noise reduction should be applied to a noisy speech signal in order to minimize the spectral distortions of a speech coder.

In this contribution we provide an answer to this question. We optimize a speech enhancement preprocessor such that the weighted mean square distance measure in the LSF domain [3]

$$\Delta_{LSF}^2 = \frac{1}{10M} \sum_{m=1}^{M} \sum_{i=1}^{10} c_i (\tilde{f}_{i,m} - f_{i,m})^2 \qquad (1)$$

is minimized where $f_{i,m}$ and $\tilde{f}_{i,m}$ denote the $i$th LSF parameter of the $m$th signal frame of the clean and the enhanced noisy speech signal, respectively. $M$ denotes the total number of speech signal frames, and $c_i$ is given in Table 1 below.

| LSF index | 1 ... 8 | 9 | 10 |
|---|---|---|---|
| $c_i$ | 1.0 | 0.64 | 0.16 |

Table 1: Weights for the LSF distortion measure [3].

Throughout this paper we will use the spectral parameters of the ETSI Adaptive Multi-Rate coder to optimize the **maximum** noise reduction of our speech enhancement preprocessor. The speech enhancement preprocessor is based on the MMSE Log Spectral Amplitude (MMSE LSA) estimator approach as described in [4, 5]. Optimizing the maximum noise reduction is important since insufficient attenuation of the signal in between the speech formant regions distorts the overall spectral shape such that an LPC analysis does not yield accurate results [1]. Allowing a higher maximum noise reduction in between the formant regions therefore improves the estimation of the speech envelope.

The remainder of this paper is organized as follows: After briefly reviewing the AMR speech coder in Sec. 2 we explain why the maximum noise reduction of the preprocessor is a key parameter for the optimization. In Sec. 3 we summarize the MMSE LSA speech enhancement technique and show how the maximum noise attenuation can be limited by limiting the *a priori* SNR. In Sec. 4 we explain our optimization procedure and conclude with informal listening test results.

## 2. SPECTRAL PARAMETERS OF THE AMR CODER

The AMR concept allows a flexible distribution of the gross bit rate to source and channel coding and thus achieves high speech quality for good channel conditions and a high

0-7803-6293-4/00/$10.00 ©2000 IEEE.

degree of robustness for heavily disturbed channels. The ETSI AMR speech coder implements eight different net bit rates between 4.75 kbps and 12.2 kbps. In the 12.2 kbps mode two sets of LSF coefficients are computed for each speech frame of 20 ms. All other modes extract only one set of LSF coefficients per speech frame. Regardless of the mode each set contains 10 LSF coefficients. When the input speech is disturbed by noise the LSF coefficients as computed in the encoder deviate from the coefficients of the clean speech signal. Especially for the lower bit rate modes, these deviations lead to significant quality and intelligibility impairments in the decoded speech signal.

As an example, Fig. 1-A shows the magnitude squared DFT coefficients (dotted) and the LPC spectrum for a given frame of noisy speech (recorded in a Nissan Sunny at 140 km/h; dashed) as well as the LPC spectrum of the corresponding clean speech signal (solid). Obviously, as the SNR of the noisy speech sample is about 3 dB the additive noise results in a significant distortion of the LPC spectrum.

Fig. 1-B shows the three spectra for the same frame of speech after enhancement. In this case the maximum noise attenuation had been limited to about 10 dB. Although the enhancement improves the LPC spectrum, there is still a significant deviation from the clean speech LPC spectrum. Finally, Fig. 1-C shows the magnitude squared DFT coefficients and the LPC spectrum for the enhanced signal where the maximum noise attenuation was limited to about 20 dB. Note, that the clean speech and the enhanced speech LPC spectra are now much closer and that also the first two formants are more pronounced.

We conclude that a high maximum noise reduction is beneficial for the enhancement of low SNR speech when the enhanced speech is the input to a speech coder. However, for high SNR speech a high maximum noise reduction might lead to undesirable speech distortions and, during speech pause, it results in spectral distortions of the background noise ('musical noise'). It is therefore of interest to optimize the maximum noise reduction as a function of the speech signal SNR such that the LPC or LSF spectra are optimally reproduced.

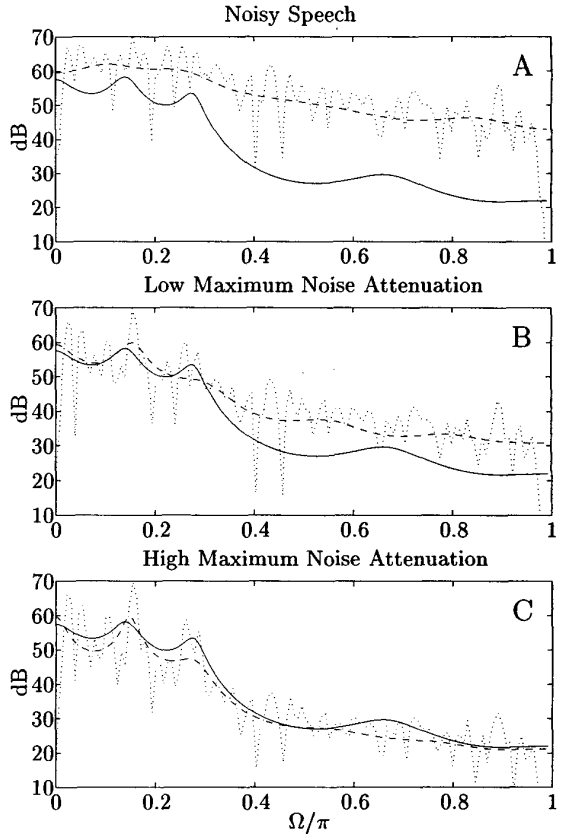


Figure 1: A: Noisy speech, B: enhanced speech with low maximum attenuation, C: enhanced speech with high maximum attenuation. Dotted: magnitude squared DFT spectrum, solid: LPC spectrum of clean speech, dashed: LPC spectrum of enhanced speech.

## 3. MMSE LSA SPEECH ENHANCEMENT

Our speech enhancement algorithm consists of three major components: a spectral analysis/synthesis system (realized by means of a windowed FFT/ IFFT and overlap/add), a noise estimation algorithm (using the 'Minimum Statistics' approach [6]), and a spectral gain computation. While all of these components have significant impact on the overall quality of the enhanced signal we focus here on the spectral gain computation. Fig. 2 shows a block diagram of the speech enhancement algorithm. The spectral gain is computed on the basis of the Fourier magnitudes and modifies only the Fourier magnitudes of an input frame. Since the noise cannot completely removed without distorting the speech signal the gain function has to strike a balance between the amount of noise reduction and the amount of speech signal distortion. We might add that in frequency bins which contain mostly noise any implementation of an optimal gain function suffers from estimation errors in the power spectral density of the noise. It is therefore very sensible to introduce additional measures such as an adaptive limiting mechanism for the gain function.

Because of its close relation to the Itakura-Saito measure we use the Minimum Mean Square Error Log Spectral Amplitude estimator (MMSE LSA) [4] to compute the gain function. The MMSE LSA estimator minimizes $E\{(\log \widehat{A}_k - \log A_k)^2\}$ where $A_k = |S_k|$ denotes the spectral speech amplitude in the $k$th DFT bin and $\widehat{A}_k$ its optimal estimate. The solution to the minimization problem is given by the expected value for the clean speech amplitude $A_k$ given the noisy DFT coefficient $Y_k = S_k + N_k$ and is obtained by applying a real gain function $G(\xi_k, \gamma_k)$ [4] to the noisy spectral coefficient $Y_k$:

$$\widehat{A}_k = \exp(E\{\ln(A_k)|Y_k\}) = Y_k G(\xi_k, \gamma_k) \quad (2)$$

$$= \frac{\xi_k}{1+\xi_k} Y_k \exp(\frac{1}{2}\int_{v_k}^{\infty} \frac{\exp\{-t\}}{t} dt)\,. \quad (3)$$

$\xi_k$ and $\gamma_k$ denote the *a priori* and *a posteriori* SNR values for bin $k$ [7], respectively, and $v_k = \frac{\xi_k}{1+\xi_k}\gamma_k$. According to

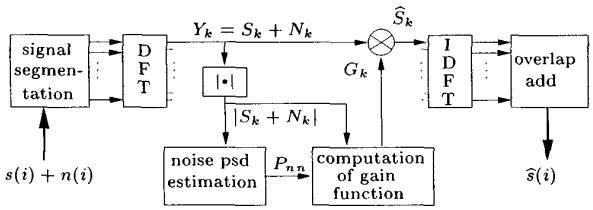


Figure 2: Block diagram of our single microphone speech enhancement algorithm.

[5] the *a priori* SNR should be conditioned on the presence of speech, i.e. $\xi_k = \eta_k/(1-q_k)$, where $\eta_k$ is the unconditional *a priori* SNR (obtained via the 'decision-directed' estimation approach [7]) and $q_k$ is the probability of speech absence. Further improvements to the quality of the enhanced signal are obtained by using the multiplicatively modified MMSE-LSA estimator [5] which accounts for the probability of speech absence. With

$$\mu_k = \frac{1-q_k}{q_k} \tag{4}$$

the gain modifier $G_M(\xi_k, \gamma_k, q_k)$ is given by

$$G_M(\xi_k, \gamma_k, q_k) = \frac{\mu_k}{\mu_k + (1+\xi_k)\exp(-v_k)} \tag{5}$$

and the total gain $\widetilde{G}(\xi_k, \gamma_k, q_k)$ by

$$\widetilde{G}(\xi_k, \gamma_k, q_k) = G(\xi_k, \gamma_k) G_M(\xi_k, \gamma_k, q_k)\,. \tag{6}$$

The probability of speech absence $q_k$ is updated for each new frame of speech by using the hard-decision approach described in [8, 5].

Similar to the MMSE STSA estimator [7] the MMSE LSA estimator can be approximated by a power subtraction rule for frequency bins which contain only noise. In fact for $v_k \ll 1$ the MMSE LSA gain can be approximated by

$$G(\xi_k, \gamma_k) \approx \sqrt{\frac{\xi_k}{1+\xi_k}}\,\frac{\exp(-0.5\mathfrak{C})}{\sqrt{\gamma_k}} \tag{7}$$

where $\mathfrak{C}$ is Euler's constant. Similarly, the gain modifier approaches

$$G_M(\xi_k, \gamma_k, q_k) \approx \frac{\mu_k}{1+\mu_k} \tag{8}$$

for low *a priori* signal-to-noise ratios. Due to the adaptation strategy for $q_k$ [8, 5] $G_M(\xi_k, \gamma_k, q_k)$ is close to zero for frequency bins which never contain speech (e.g. because of a bandpass filter) and close to 0.5 for bins which contain mostly noise but might also contain some speech at a different time. Therefore, the maximum noise attenuation for bins within the band of speech frequencies which contain mostly noise can be limited by limiting $\eta_k$. However, the approximations of eqs. 7 and 8 do not hold if $\eta_k$ is limited to values larger than 0.1 since then $v_k$ becomes too large for noise only bins. In this case the interaction between the estimated probability of speech absence $q_k$ and the *a priori* SNR $\eta_k$ results in a total gain for noise only bins which is heavily influenced by the multiplicative modifier $G_M(\xi_k, \gamma_k, q_k)$. It is therefore reasonable to also limit $G_M(\xi_k, \gamma_k, q_k)$ in a similar way as $\eta_k$. However, to avoid an unnecessary increase in the complexity of the optimization task we limited $G_M(\xi_k, \gamma_k, q_k)$ to a fixed value $G_{Mmin} = 0.2$ which gives close to optimal results for a wide range of *a priori* SNR values.

With the help of the optimization procedure as outlined in the next Section the lower limit $\eta_{min}$ of the unconditional *a priori* SNR $\eta_k$ was determined as a function of the overall speech signal SNR which was estimated and updated by a first order recursive system for each speech frame. We found that the lower limit for the unconditional *a priori* SNR $\eta_k$, i.e. the maximum noise reduction, can be approximated by (see Section 4):

$$10\log_{10}(\eta_{min}) = \sqrt{SNR} - 16.5 \tag{9}$$

where SNR denotes the estimated average signal-to-noise ratio (on a linear scale). This SNR value can be computed as the ratio of the recursively smoothed average speech power and the recursively smoothed average noise power. The adaptive limit $\eta_{min}$ is only applied to signal frames which contain speech. Noise only frames are limited to a constant limit $\eta_{min} = 0.12$ to avoid musical fluctuations during speech pause. The resulting time varying $\eta_{min}$ was recursively smoothed with a smoothing parameter of $\alpha_\eta = 0.8$.

## 4. OPTIMIZATION PROCEDURE AND RESULTS

To obtain the optimal limit for the maximum noise attenuation $\eta_{min}$ in equ. 9 as a function of the input speech SNR we added computer generated white Gaussian noise to 60 s of male and 60 s of female speech at various average signal-to-noise ratios ($10\log(SNR)$ = 0, 6, 12, 18, 24 dB). These files were processed with the enhancement preprocessor as described in Section 3 for several fixed values of $\eta_{min}$ ($10\log(\eta_{min})$ = -3, -8, -11, -17, -20, -23, -26, -29, -32, -35, -38, -40 dB). The enhanced speech files were then fed into the AMR coder (4.75 and 12.2 kbps modes) and the LSF parameters were recorded for each speech frame. For each SNR and $\eta_{min}$ value the distortion measure $\Delta^2_{LSF}$ was computed and an optimal $\eta_{min}$ value was determined after the measurement points were interpolated by cubic splines. Fig. 3 plots $\Delta^2_{LSF}$ vs. $\eta_{min}$ for SNR $\approx$ 0 dB and female speech. We found that for this condition as well as for all other SNR values the $\Delta^2_{LSF}$ vs. $10\log(\eta_{min})$ plot showed a pronounced minimum. After computing the optimum for all other SNR values the resulting minima were plotted in Fig. 4 for male and female speech as a function of $10\log(SNR)$. Fig. 4 also includes the approximation given in equ. 9 which was then used to implement the maximum attenuation characteristic of the enhancement preprocessor.

For the sake of standardized experimental conditions the optimization was performed with white Gaussian noise. However, similar improvements were obtained with other noise types. Since frequency bins which contain mostly speech are not much affected by the maximum noise attenuation the algorithm performed almost equally well with

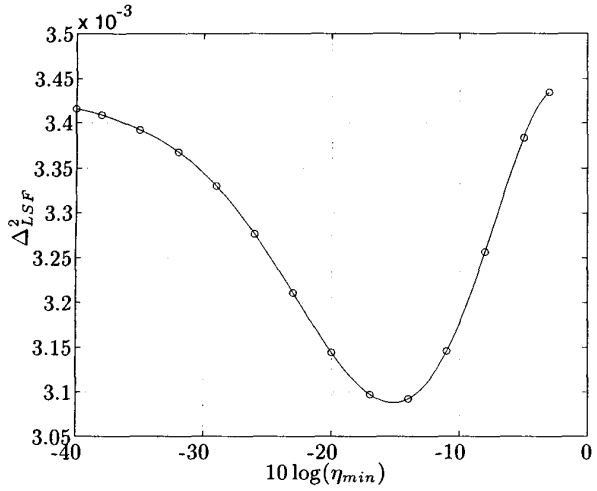


Figure 3: LSF distortion measure $\Delta^2_{LSF}$ vs. $10\log(\eta_{min})$ for SNR $\approx$ 0 dB and female speech.

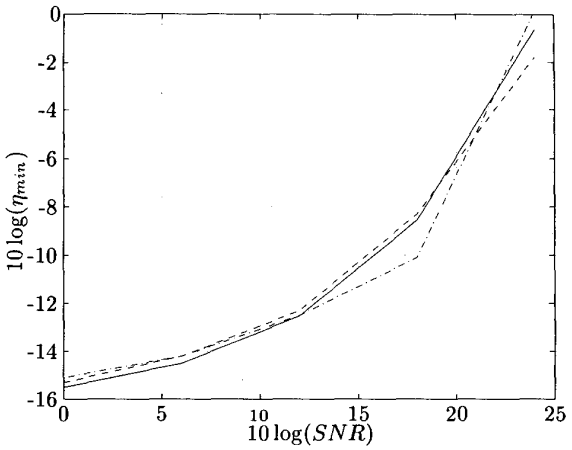


Figure 4: $10\log(\eta_{min})$ vs. speech $10\log(SNR)$ for male speech (dashed), female speech (dash-dotted), and the approximation of equ. 9 (solid).

other noises. Also, as Fig. 4 indicates, we could not find a significant difference between male and female speech.

Informal listening tests of speech processed with the joint enhancement and coding system were conducted with speech disturbed by stationary car noise at three different signal-to-noise ratios (6, 12, 18 dB) and two coding modes (4.75 and 12.2 kbps). These tests compared the new approach to a system which applies a fixed limit of $10log(\eta_{min}) = -9dB$ to all signal frames. The listening tests indicated that also in the new approach $10\log(\eta_{min})$ should be limited to values equal or above $-12$ dB to avoid distortions for low SNR speech. For the low coding rate the optimized limit gave audibly improved speech quality while at the higher rate and for high SNR speech the improvements were less pronounced.

## 5. CONCLUSIONS

In this paper we showed how the performance of a joint speech enhancement and coding system can be improved by optimizing the maximum noise reduction of the enhancement preprocessor. We showed that for the widely used MMSE LSA speech estimator the *a priori* SNR can be used to limit the maximum noise reduction and determined the optimal lower limit of the *a priori* SNR such that a distortion measure in the LSF domain is minimized. The optimized system gave improved results in conjunction with the ETSI AMR coder which were most notable for the 4.75 kbps mode and low SNR conditions. Our listening test suggest that for the higher bit rates it is crucial to also include other coder parameters such as the linear prediction residual in the optimization procedure [9].

## 6. REFERENCES


[1] R. Martin and R. Cox, "New Speech Enhancement Techniques for Low Bit Rate Speech Coding," in *Proc. IEEE Workshop on Speech Coding*, (Porvoo,Finland), pp. 165–167, 1999.

[2] G. Guilmin, R. L. Bouquin-Jeannes, and P. Gournay, "Study of the Influence of Noise Pre-Processing on the Performance of a Low Bit Rate Parametric Speech Coder," in *Proc. EUROSPEECH*, vol. 3, (Budapest, Hungary), pp. 2367–2370, September 1999.

[3] K. Paliwal and B. Atal, "Efficient Vector Quantization of LPC Parameters at 24 Bits/Frame," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, pp. 661–664, 1991.

[4] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 33, pp. 443–445, April 1985.

[5] D. Malah, R. Cox, and A. Accardi, "Tracking Speech-Presence Uncertainty to Improve Speech Enhancement in Non-Stationary Noise Environments," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, 1999.

[6] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. Euro. Signal Processing Conf. (EUSIPCO)*, pp. 1182–1185, 1994.

[7] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 32, pp. 1109–1121, December 1984.

[8] I. Soon, S. Koh, and C. Yeo, "Improved noise suppression filter using self-adaptive estimator of probability of speech absence," *Signal Processing, Elsevier*, vol. 75, pp. 151–159, 1999.

[9] A. Accardi and R. Cox, "A Modular Approach to Speech Enhancement with an Application to Speech Coding," in *Proc. IEEE Intl. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, 1999.