On Optimal Estimation of Compressed Speech for Hearing Aids
Dirk Mauler, Anil M. Nagathil, and Rainer Martin

Institute of Communication Acoustics (IKA), Ruhr-Universität Bochum, 44780 Bochum, Germany

dirk.mauler@rub.de, anil.nagathil@rub.de, rainer.martin@rub.de

Abstract
When noise reduction (NR) and dynamic compression (CP) systems are concatenated in a hearing aid or in a cochlear implant we observe undesired interaction effects like the degradation of the global SNR. A reason for this might be that the optimization of the NR algorithm is performed with respect to the uncompressed clean speech only. In this contribution we propose an alternative approach which integrates the CP task in the derivation of the NR algorithm. By this we get novel MMSE and MAP optimal estimators for the compressed clean speech. An analysis of the behavior of the proposed solutions reveals that the differences to a serial concatenation of NR and CP are in general small. In case of the widely used MMSE log spectral amplitude (LSA) estimator [1] we show that the combined optimization is identical to a serial concatenation.

Index Terms: noise reduction, dynamic compression, speech enhancement

1. Introduction
Dynamic compression (CP) and noise reduction (NR) algorithms are fundamental features of modern hearing instruments. The NR serves to improve the listening comfort by an attenuation of disturbing sounds as for example babble of competing speakers in a cafeteria, engine noise, etc. The rationale behind dynamic CP is manifold [2], necessitating different settings for CP parameters, like CP thresholds, CP ratio, or time constants, depending on the field of application. In this paper we focus on the fast dynamic CP. The purpose of a fast dynamic CP is to map the large range of acoustic stimuli to the reduced dynamic range of audibility of the hearing impaired. Fast compressors in state-of-the-art hearing aids are usually implemented as multiband compressors that allow for a different CP parameter setting per frequency band.

Typically, the CP system is applied in the spectral domain after NR [3]. This situation is illustrated in Figure 1a) where the frequency domain NR and CP system is shown, including a linear amplification by the factor \( a \). Without loss of generality we assume that a sliding window discrete Fourier transform (DFT) is used for spectral analysis. One of the standard NR schemes, e.g. [4], [1], [5], can be used to estimate for each discrete frequency band index, \( k \), a clean speech spectral magnitude, \( \hat{A}_k \), given the disturbed spectral coefficient, \( Y_k \). However, the serial concatenation with the CP produces undesired interaction effects. For instance, the global signal to noise ratio (SNR) degrades due to the fact that the linearly amplified speech is frequently attenuated by the CP, while in speech absence, the soft residual noise at the output of the NR component will usually become less compressed [6].

Let us denote the complex short time Fourier transform coefficients of clean and noisy speech as \( S_k = A_k \exp(j\omega_k) \) and \( Y_k = W_k \exp(j\theta_k) \), respectively, where \( k \) is the frequency bin index. Then, in a system designed according to Figure 1 a) or b) a non-linear system components are concatenated, that have been designed or optimized separately, assuming a stand-alone usage. Thus, an optimal system behavior of the concatenated systems in Figures 1 a) or b) can not be expected.

In this contribution, novel solutions are developed and discussed that optimize for both, NR and CP, in a common approach. In Figure 1 c) this concept is depicted as a single system block that performs both, NR and CP, in an optimal way. Furthermore, for the log spectral amplitude (LSA) NR estimator [1] we examine the properties of the arrangement given in Figure 1 b).

Section 2 starts with a definition of the dynamic CP that will be used subsequently. In Section 3, two minimum mean square error (MMSE) and a maximum \( a \) posteriori (MAP) optimal solution for the compressed cases are derived. For the example of the MAP estimator we also depict the input-output characteristics.

2. Dynamic compression
In this section we introduce a CP system for spectral magnitudes. The output \( X_{out, dB} \) of the CP is a piecewise linear function of the input \( X_{in, dB} \) in the logarithmic domain (Figure 2). In the linear domain the compressed output amplitude, \( \hat{x}_{out} \),

\[ X_{out, dB} = \begin{cases} a \exp(j\omega_k) & \text{if } |S_k| > |Y_k| \\ 0 & \text{otherwise} \end{cases} \]

\[ \hat{x}_{out} = \begin{cases} a \exp(j\omega_k) & \text{if } |S_k| > |Y_k| \\ 0 & \text{otherwise} \end{cases} \]

\[ X_{out, dB} = \begin{cases} a \exp(j\omega_k) & \text{if } |S_k| > |Y_k| \\ 0 & \text{otherwise} \end{cases} \]
given an input amplitude, $x_{in}$, is defined by

$$\hat{x}_{out} = \begin{cases} \frac{1}{c_{th}} x_{in}, & X_{in,db} > C_{th,db} \\ x_{in}, & X_{in,db} \leq C_{th,db} \end{cases}$$  \hspace{1cm} (1)$$

where $C_{th,db} = 10 \log_{10}(c_{th})$ denotes the CP threshold in dB, $c_{th}$ the CP threshold in the linear domain, and $R$ is the CP ratio defined as the inverse of the slope of the compression curve $R = \frac{\Delta X_{out,db}}{\Delta X_{in,db}}$. Note that for input levels less than the CP threshold, $X_{in,db} \leq C_{th,db}$, the CP system is transparent (no compression). In this case only NR is performed so that NR algorithms known from the literature can be used. In the following subsections we focus on the compression case, $X_{in,db} > C_{th,db}$, and derive MMSE optimal or MAP optimal solutions. In a complete system, the speech estimate is then assembled depending on the level of the input signal. This means that in case of an inaccurate compression the traditional optimal NR solution is used while for an active compression the solution from the following derivation will be applied.

To be compatible with a DFT-based NR with relatively high spectral resolution we consider a multiband CP scheme that allows for a DFT-binswise CP of the spectral magnitudes over time. If for the CP a lower spectral resolution is desired, several DFT channels can be grouped together to control the CP. In all equations the letter $k$ denotes the frequency bin index.

### 3. Optimal estimators for the compressed clean speech magnitude

In the following we derive MMSE and MAP optimal estimators for the compressed clean speech spectral magnitude, assuming a single system that integrates NR and CP (Figure 1 c), and in Section 3.1. also for a system that consists of the serial concatenation of CP and NR as shown in Figure 1 b). The optimal short time phase for all estimators under investigation is known to be the phase of the disturbed input signal [4], [5]. Thus, in the following we concentrate on the estimation of the magnitudes only.

#### 3.1. MMSE optimal LSA estimator

The MMSE optimal estimator, $\hat{A}_k$, that minimizes the mean distance to a logarithmically weighted clean speech magnitude, log $A_k$, given the noisy speech spectral coefficient, $Y_k$, is

$$E \left\{ \log (A_k - \log \hat{A}_k) \right\} \rightarrow \min,$$  \hspace{1cm} (2)

may be written as [1]

$$\hat{A}_k = \exp (E \{ \ln A_k | Y_k \}).$$  \hspace{1cm} (3)

For Gaussian distributed and uncorrelated speech and noise spectral components a closed form solution results [1, (20)].

If a dynamic CP shall be integrated in (2) the reference in the minimization is the clean and compressed speech spectral amplitude. Thus, using (1), we derive the estimator that minimizes the mean distance to the compressed clean speech spectral log-amplitude

$$E \left\{ \left( \log \left( \frac{1}{c_{th}} \hat{A}_k \right) - \log \hat{A}_k \right)^2 \right\} \rightarrow \min.$$  \hspace{1cm} (4)

Setting the first derivative with respect to $\hat{A}_k$ to zero and solving for $\hat{A}_k$, the optimal estimator for the compressed clean speech magnitude is obtained:

$$\hat{A}_k = c_{th} \left( \exp \left\{ \ln A_k \right\} \right)\frac{1}{\hat{A}_k}.$$  \hspace{1cm} (5)

We recognize that this result is composed of the solution for the uncompressed case, (3), modified by a CP approach as defined in (1). In other words, the estimator for compressed speech that is optimal in the sense of (4) (cf. Figure 1 c)) results in the same solution that is obtained by a serial concatenation of a NR first, followed by a dynamic CP (Figure 1 a).

For the configuration of a CP first, followed by the NR (Figure 1 b)) we replace $Y_k$ in (4) by the compressed noisy spectral coefficient, $\hat{Y}_k$. The minimization then leads to

$$\hat{A}_k = c_{th} \left( \exp \left\{ \ln \hat{A}_k \right\} \right)\frac{1}{\hat{A}_k}.$$  \hspace{1cm} (6)

We solve (6) as in [1] utilizing the moment generating function of $\ln A_k$ given $\hat{Y}_k$. Furthermore, we have to transform the density of the noisy spectral coefficients given the clean speech amplitude and phase

$$p(Y_k | A_k, \alpha_k) = \frac{1}{\pi \sigma_{N,k}^2} \exp \left( \frac{-|Y_k - A_k e^{i\alpha_k}|^2}{\sigma_{N,k}^2} \right),$$  \hspace{1cm} (7)

to the compressed domain

$$p(\hat{Y}_k | A_k, \alpha_k) = \frac{c_{th}^{R-1}}{\pi \sigma_{N,k}^2} \frac{R}{\sigma_{N,k}^2} \exp \left( \frac{-|\hat{Y}_k - A_k e^{i\alpha_k}|^2}{\sigma_{N,k}^2} \right).$$  \hspace{1cm} (8)

Using this, it turns out that

$$\exp E \left\{ \ln A_k | \hat{Y}_k \right\} = \exp E \{ \ln A_k | Y_k \}$$  \hspace{1cm} (9)

and thus (6) equals (5). We conclude that due to the logarithmic weighting of the MMSE-LSA approach on the one hand and the exponential compression function on the other hand, the order in which NR and CP are applied has no impact on the optimality of the processing result. In both cases the result equals the optimal solution that is obtained from a NR approach that incorporates a dynamic CP.
3.2. MMSE optimal STSA estimator

In [4] the MMSE optimal estimator that minimizes the mean distance to a clean speech magnitude reference, \( A_k \),

\[
E\left\{ (A_k - \hat{A}_k)^2 | Y_k \right\} \rightarrow \min
\]

(10)

has been derived for a given disturbed speech spectral coefficient, \( Y_k \), and assumed Gaussian distributed uncorrelated speech and noise spectral components and may be written as

\[
\hat{A}_k = \Gamma(1.5) \frac{\sqrt{\gamma_k} \cdot M(-0.5; 1; -\nu_k)}{\gamma_k} |Y_k|, \quad \nu_k = \frac{\xi_k}{1 + \xi_k} \gamma_k, \quad \xi_k = \frac{\sigma_{S,k}^2}{\sigma_{N,k}^2}, \quad \gamma_k = \frac{W_k^2}{\sigma_{N,k}^2},
\]

where \( \xi_k \) and \( \gamma_k \) denote the \textit{a priori} and the \textit{a posteriori} SNR in frequency bin \( k \), respectively. \( \sigma_{S,k}^2 \) and \( \sigma_{N,k}^2 \) denote the speech and the noise power, respectively, and \( \Gamma(\cdot) \) and \( M(\cdot; c; x) \) denote the gamma and the confluent hypergeometric function, respectively.

To derive the estimator that minimizes the mean distance to the \textit{compressed} clean speech spectral magnitude we minimize

\[
E\left\{ (c_{th}^{1/2} A_k - \hat{A}_k)^2 | Y_k \right\} \rightarrow \min
\]

(12)

and after calculations analogous to those in [4] we obtain the result

\[
\tilde{A}_k = c_{th}^{1/2} \left( \frac{\sqrt{\gamma_k}}{\gamma_k} \right) \tilde{\gamma} \Gamma \left( \frac{1}{2 R} + 1 \right) M \left( \frac{-1}{2 R}; 1; -\nu_k \right) |Y_k|^{1/2}.
\]

(13)

Note, that for \( R = 1 \) (i.e. no compression) (13) equals (11). The solution (13) differs from the solution of a serial concatenation of NR and CP which would be obtained if the compression (1) is applied to the estimator (11). However, the differences are too small for being perceivable. A numerical evaluation of the ratio of the speech magnitudes estimated with the solution of the integrated approach (13) to the speech magnitudes estimated with (11) and subsequent CP reveals that the ratio as a function of the disturbed input speech magnitude, \( |Y_k| \), is approximated to

\[
\frac{A_k}{\tilde{A}_k} \sim 0.96
\]

with (11) and subsequent CP revealing that the ratio as a function of the disturbed input speech magnitude, \( |Y_k| \), is close to 1, and does never fall below 0.96. The minimal ratio is observed for SNR \( \rightarrow -\infty \) dB and a CP ratio \( R \approx 2.1 \).

3.3. MAP optimal estimator

In [5] a MAP estimator for the clean speech magnitudes, \( A_k \), given an observation of the disturbed spectral coefficient, \( Y_k \),

\[
\tilde{A}_k = \arg \max_{A_k} p(A_k, \alpha_k | Y_k)
\]

\[
= \arg \max_{A_k} \frac{p(Y_k | A_k, \alpha_k)}{p(A_k, \alpha_k)}.
\]

(14)

has been derived for a given MAP optimal speech phase \( \alpha_k = \theta_k \) to

\[
\tilde{A}_k = \left( u + \frac{v^2 + v}{2 \gamma_k} \right) |Y_k|, \quad u = \frac{1}{2} - \frac{\mu}{4 \sqrt{\gamma_k} \xi_k},
\]

(15)

where Gaussian distributed noise spectral magnitudes are assumed, (7) is used, and \( \mu \) and \( \nu \) are parameters of the parametric distribution for the speech spectral amplitudes

\[
p(A_k) = \frac{\mu^{\nu+1}}{\Gamma(\nu + 1)} \sigma_{S,k}^{\nu+1} \exp \left( -\frac{A_k}{\sigma_{S,k}} \right).
\]

(16)

The MAP estimator for the \textit{compressed} clean speech magnitude

\[
\hat{A}_k = \arg \max_{A_k} p(\hat{A}_k, \alpha_k | Y_k)
\]

(17)

requires the transformation of the distribution (16) according to the compression function (1). The transformed density reads

\[
p(\hat{A}_k) = \frac{\mu^{\nu+1}}{\Gamma(\nu + 1)} \frac{R c_{th}^{(\nu+1)(1-R)} A_k^{\nu+1} R^{-1}}{A_k^{\nu+1} R^{-1}} \exp\left( -\mu c_{th}^{1-R} A_k^{R} \right).
\]

(18)

where \( \hat{A}_k \) denotes the magnitude of the compressed speech spectral coefficient. Equation (7) does not need to be transformed as it provides the distribution of the \textit{uncompressed} noisy input spectral coefficient, \( Y_k \). However, it has to be evaluated given the compressed speech magnitude, i.e.

\[
p(Y_k | \hat{A}_k, \alpha_k) = p(Y_k | A_k, \alpha_k) \bigg|_{A_k = \hat{A}_k, \alpha_k = \alpha_k}.
\]

(19)

The phase \( \alpha_k \) of the clean speech spectral coefficient, \( S_k \), is preserved by the CP defined in (1).

Using (18) and (19) we solve (17) and obtain

\[
\tilde{A}_k = c_{th}^{1/2} \left( u + \sqrt{\frac{v^2 + v}{2 \gamma_k}} \right) \frac{1}{R} |Y_k|^{1/2},
\]

(20)

where \( u \) is defined as in (15). Again, for \( R = 1 \) (i.e. no compression) (20) equals the solution of the estimator without CP. (15). Note that due to the term \( \frac{1}{R} \) under the square root, (20) differs from the solution of a serial concatenation of NR and CP which would be obtained if the compression (1) is applied to the estimator (15).

Figure 3 shows the results for the estimated speech spectral magnitude, \( \tilde{A}_k \), and the estimated and compressed speech spectral magnitudes, \( \tilde{A}_k, \hat{A}_k \), as a function of the disturbed input spectral magnitude, \( |Y_k| \). The \textit{a priori} SNR is 0 dB while the power \( \sigma_{F,k}^2 = \sigma_{S,k}^2 + \sigma_{N,k}^2 = 2 \) is constant. The CP threshold is set to \( c_{th} = 1 \) and the CP ratio is \( R = 2 \).

The dash-dotted line shows the MAP solution without CP [5] as given in (15). The dashed line represents the same estimate when subsequently modified by a dynamic CP (cf. Figure 1 a)). A transition can be observed at \( \tilde{A}_k = 1 \) which is caused by the dynamic CP which becomes active for amplitudes that exceed the CP threshold (cf. (1)). The lower dotted line shows the solution that would be obtained if the CP was permanently active.

The result for the proposed solution of the MAP estimator that is optimized with respect to the compressed clean speech (cf. Figure 1 c)) is plotted with a solid line and shows the same asymptotic behavior as the MAP estimator with serially concatenated dynamic CP (dashed line). However, we observe that the compression starts at slightly higher input amplitudes than in the case of concatenated NR and CP. The thick solid line is obtained by assembling the solution without CP and the solution given in (20) at their intersection point. (The upper dotted line shows the solution of (20) with the CP permanently active.) The differences to the solution with serial NR and CP are small and are in general not perceivable in a processed signal.

Figure 4 presents the plots for an \textit{a priori} SNR of \( \xi_k = 10 \) dB. The differences between the serial solution (dashed) and the proposed solution (solid) are even smaller than in Figure 3 and are hence not perceivable.
NR without CP, (15)
NR with integrated CP, (20), Fig. 1 c
NR and subsequent CP, (15), (1), Fig. 1 a

Figure 3: Input-output characteristics of the MAP estimator for the clean speech (dash-dotted), for the compressed speech (solid), and for the serial concatenation of MAP estimation of the uncompressed speech and subsequent dynamic CP (dashed). The a priori SNR is $\xi = 0$ dB, the CP ratio $R = 2$, the CP threshold $c_{th} = 1$ at a constant power, $\sigma_{S,k}^2 + \sigma_{N,k}^2 = 2$. The thick lines represent the curves including a switching from the uncompressed to a compressed solution (cf. (1)) while the dotted lines show the solutions for combined NR and CP if no switching was performed.

When combining spectral domain noise reduction (NR) algorithms with a dynamic compression (CP), undesired interaction effects, like an SNR deterioration, are observed. We developed new MMSE- and MAP-optimal estimators that incorporate the CP task in the derivation of the NR algorithm by estimating the compressed clean speech spectral magnitude. For a CP ratio $R = 1$ (i.e. no CP) all estimators reduce to the known solutions. We showed that in case of the MMSE LSA estimation the new approach yields the same solution as obtained with the serial concatenation of a NR and CP in either order. The modified MAP- and MMSE-STSA-approaches yield results that differ analytically from a concatenation of NR and CP. However, we showed that the differences are small and are in general not perceivable in a processed signal. Hence, in particular in the field of hearing aids, where high linear amplification is required, an efficient NR remains a challenging task.

The approaches result in solutions where NR and CP are inseparably combined. Hence, realizing CP attack and release times [2] by temporally smoothing the filter gain would inevitably also affect the NR properties. This drawback can be avoided if the attack and release characteristics are realized by a time variant CP ratio, $R$.

4.1. Acknowledgment

This work was in part supported by grants from the European Union FP6, Project 004171 HEARCOM.

5. References


