Bias Compensation for Cepstro-Temporal Smoothing of Spectral Filter Gains

Timo Gerkmann, Colin Breithaupt, and Rainer Martin

Institut für Kommunikationsakustik (IKA), Ruhr-Universität Bochum
E-Mail: {timo.gerkmann,colin.breithaupt,rainer.martin}@rub.de
Web: www.rub.de/ika

Abstract

Many speech enhancement algorithms that modify short-term spectral magnitudes of the noisy signal are plagued by annoying spectral outliers that are perceived as musical noise. Recently, we presented techniques that reduce these outliers by means of a temporal smoothing in the cepstral domain. This cepstro-temporal smoothing increases the quality of the enhanced output signal, as it affects only spectral outliers caused by estimation errors, while the speech characteristics are well preserved. However, due to the cepstral transform, the temporal smoothing is done in the logarithmic domain rather than the linear domain, and hence results in a certain bias. In this paper we derive a general bias compensation for a cepstro-temporal smoothing of spectral filter gain functions that is only dependent on the lower limit on the spectral filter-gain function. We show that the proposed bias-compensation increases the performance in terms of instrumental measures.

1 Introduction

Many successful speech enhancement algorithms work in the short-time discrete Fourier transform (DFT) domain. A drawback of DFT based speech enhancement algorithms is that they yield unnatural sounding structured residual noise, often referred to as *musical noise.* Musical noise occurs, *e.g.* if in a noise-only signal frame single Fourier coefficients are not attenuated due to estimation errors, while all other coefficients are attenuated. The residual isolated spectral peaks in the processed spectrum correspond to sinusoids in the time domain and are perceived as tonal artifacts of one frame duration. Especially when the speech enhancement algorithms operate in non-stationary noise environments, unnatural sounding residual noise remains a challenge. Recently, a selective temporal smoothing of parameters of speech enhancement algorithms in the cepstral domain has been proposed [1, 2, 3] that reduces residual spectral peaks without affecting the speech signal. In [1, 3] the algorithms based on cepstro-temporal smoothing (CTS) are compared to state-of-the-art speech enhancement algorithms in terms of listening experiments. In [1] it is shown that CTS yields an output signal of higher quality especially in babble noise, and that the number of spectral outliers in the processed noise is less than with state-of-the-art algorithms. In [3] it is shown that CTS yields an output signal of increased quality when applied as a post processor in a speaker separation task. However, due to the non-linear log-transform inherent in the cepstral transform, a temporal smoothing yields a certain bias as compared to a smoothing in the linear domain. This bias results in an output signal with reduced power. While the reduced signal power has only a minor influence on the results of listening experiments, instrumental measures are often sensitive to a change in signal power. Thus, without a bias correction, instrumental measures may indicate a reduced signal quality if CTS is applied, while listening experiments indicate a clear increase in quality. In [2] CTS is applied to a maximum likelihood estimate of the speech power to replace the well-known decision-directed *a priori* signal-to-noise ratio (SNR) estimator [4]. It is shown that if a bias correction is applied, the speech power estimation based on CTS yields consistent improvements in terms of segmental SNR, noise reduction, *and* speech distortion. This can be attributed to the fact that in the cepstral domain speech specific properties can be taken into account. In this paper we derive a bias compensation for a CTS of arbitrary spectral filter gain functions, *e.g.* [1, 3]. We use the same setup as used for the listening experiments in [1] and show that without a bias correction instrumental measures indicate a decreased performance. Further we show that with the proposed bias correction instrumental measures indicate an increased performance especially in nonstationary noise environments.

In the next two sections we briefly introduce short-time DFT-domain speech enhancement and the concept of CTS. In Section 4 we present a bias correction for a cepstro-temporal smoothing of spectral filter gains. In Section 5 we show that the proposed correction successfully compensates for the bias introduced by a temporal smoothing in the cepstral domain for various input SNRs and noise-types.

2 Speech Enhancement in the short-time DFT-domain

For speech enhancement in the short-time DFT-domain, a noisy time domain speech signal is segmented into short frames, *e.g.* of length 32 ms. Each signal segment is windowed, *e.g.* with a Hann window, and transformed into the Fourier domain. The resulting complex spectral representation $Y_k(l)$ is a function of the spectral frequency index $k \in [0, K[$, and the segment index l. The spectral coefficients of the noise signal, $N_k(l)$, are assumed additive to the speech spectral coefficients $S_k(l)$, *i.e.* $Y_k(l) = S_k(l) + N_k(l)$. Note that the noise signal, $N_k(l)$, may be environmental noise as well as competing talkers as in the case of speaker separation. The aim of speech enhancement algorithms is to estimate the clean speech signal $S_k(l)$ given the noisy observation $Y_k(l)$. This is often achieved via a multiplicative gain function $G_k(l)$. An estimate of the clean speech spectral coefficients is thus computed as

$$\widehat{S}_k(l) = G_k(l)\, Y_k(l)\,. \tag{1}$$

3 Cepstro-Temporal Smoothing

CTS is based on the idea that in the cepstral domain, speech is represented by few coefficients, which can be robustly estimated. A cepstral transform of some positive, real valued spectral parameter $\Phi_k(l)$ of the speech enhancement algorithm (like the estimated speech peri-

odogram or the gain function) is given by

$$\phi_q(l) = \mathrm{IDFT}\{\log \Phi_k(l)\}\,, \tag{2}$$

where $q \in [0, K[$ is the cepstral quefrency index, and $\mathrm{IDFT}\{\cdot\}$ the inverse DFT. Note that as $\Phi_k(l)$ is real-valued, $\phi_q(l)$ is symmetric with respect to $q = K/2$. Therefore, in the following only the part $q \in [0, K/2]$ is discussed. The lower cepstral coefficients $q \in [0, q_{\mathrm{low}}]$ with, preferably, $q_{\mathrm{low}} \ll K/2$ represent the spectral envelope of $\Phi_k(l)$. For speech signals, the spectral envelope is determined by the transfer function of the vocal tract. The higher cepstral coefficients $q_{\mathrm{low}} < q < K/2$ represent the fine-structure of $\Phi_k(l)$. For speech signals, the fine-structure is caused by the excitation of the vocal tract. For voiced speech, the excitation is mainly represented by a dominant peak at $q_0 = f_s/f_0$, with f_0 the fundamental frequency. This fundamental frequency can be found by a maximum search in $q \in [q_{\mathrm{low}}, K/2]$ as proposed in [5]. Thus, in the cepstral domain voiced speech can be represented by the set

$$\mathbb{Q} = \{[0, q_{\mathrm{low}}], q_0\}\,. \tag{3}$$

If $\Phi_k(l)$ is an estimated parameter, like the estimated speech periodogram, or the spectral gain function, its fine-structure is also influenced by spectral outliers caused by estimation errors. Therefore, a recursive temporal smoothing is now applied on $\phi_q(l)$, such that only little smoothing is applied to those cepstral coefficients, $q \in \mathbb{Q}$, that are dominated by speech, and strong smoothing to all other coefficients:

$$\overline{\phi}_q(l) = \alpha_q \overline{\phi}_q(l-1) + (1-\alpha_q)\phi_q(l)\,, \tag{4}$$

with the smoothing factor

$$\alpha_q = \begin{cases} \ll 1 & \text{, for } q \in \mathbb{Q} \\ \rightarrow 1 & \text{, else} \end{cases} \tag{5}$$

After the recursive smoothing $\overline{\phi}_q(l)$ is transformed to the spectral domain to achieve the cepstro-temporally smoothed spectral parameter $\overline{\Phi}_k(l)$, as

$$\overline{\Phi}_k(l) = \exp\left(\mathrm{DFT}\left\{\overline{\phi}_q(l)\right\}\right)\,. \tag{6}$$

CTS allows for a reduction of spectral outliers due to estimation errors, while the speech characteristics are preserved. In the following cepstro-temporally smoothed parameters are marked by a bar, *e.g.* $\overline{G}$ for the cepstro-temporally smoothed spectral filter gain.

4 Bias Compensation for Cepstro-Temporal Smoothing of Spectral Filter Gain Functions

In [1] and [3] CTS of the spectral gain function is proposed (*i.e.* $\Phi_k(l) = G_k(l)$ in (2)) to reduce spectral outliers that do not correspond to speech but to estimation errors. Smoothing the gain function for reducing spectral outliers is a very flexible technique. It can be applied to any speech enhancement algorithm where the output signal is gained via a multiplicative gain function as in (1). This includes noise reduction [1] and source separation [3]. In speech enhancement algorithms the gain function is usually bound to be larger than a certain value G_{min} [6]. Therefore, after the derivation of a gain function G', a constrained gain G is computed as $G = \max\{G', G_{\mathrm{min}}\}$. The choice of G_{min} is a trade-off between speech distortion, musical noise and noise reduction. A large G_{min} masks musical noise and reduces speech distortions at the cost of less noise reduction. The aim of this work is to derive a general bias correction for CTS of arbitrary gain functions. We thus assume a uniform distribution of G' between 0 and 1, independent of its derivation and the underlying distribution of the speech and noise spectral coefficients. To construct the probability density function (PDF) of the constrained G we map $\int_0^{G_{\mathrm{min}}} p(G')\mathrm{d}G'$ onto $p(G = G_{\mathrm{min}})$ (cf. Figure 1).

Since the values of the gain function are limited in their dynamic range ($G_{\mathrm{min}} \leq G \leq 1$), the non-linear compression via the log-function in (2) is not mandatory, *i.e.* the principle behavior of the cepstral coefficients stays the same with or without the log-function. However, in [1] it is noted, that incorporating the log-function may help reducing noise shaping effects that may arise due to the temporal smoothing. We argue that the recursive averaging (4) can be interpreted as an approximation of the expected value operator. However, if the log-function is applied in (2), an arithmetic mean of ϕ_q corresponds to a geometric mean of $\Phi_k = G_k$. Therefore, CTS changes the mean of the gain function, as in general $\mathrm{E}\{G\} \neq \exp(\mathrm{E}\{\log(G)\})$, with $\mathrm{E}\{\cdot\}$ the expected value operator. If the distribution of G is known the difference

$$\kappa_G = \log(\mathrm{E}\{G\}) - \mathrm{E}\{\log(G)\} \tag{7}$$

can be determined and accounted for. For the distribution in Figure 1, the expected value of the gain function can be determined as:

$$\mathrm{E}\{G\} = G_{\mathrm{min}}^2 + \int_{G_{\mathrm{min}}}^{1} G\,dG = \frac{1}{2}\left(1 + G_{\mathrm{min}}^2\right)\,, \tag{8}$$

and the expected value of the log-gain function results in

$$\begin{aligned} \mathrm{E}\{\log G\} &= G_{\mathrm{min}} \log G_{\mathrm{min}} + \int_{G_{\mathrm{min}}}^{1} \log G\,dG \\ &= G_{\mathrm{min}} - 1\,. \end{aligned} \tag{9}$$

With (7) the bias correction κ_G thus results in:

$$\kappa_G(G_{\mathrm{min}}) = \log\left(\frac{1}{2} + \frac{1}{2}G_{\mathrm{min}}^2\right) - G_{\mathrm{min}} + 1\,. \tag{10}$$

We can now apply a bias correction to a cepstro-temporally smoothed gain function $\overline{G}_k(l)$, as

$$\widetilde{G}_k(l) = \overline{G}_k(l)\exp(\kappa_G)\,. \tag{11}$$

In Figure 2 κ_G is plotted as a function of G_{min}. Note that, as small values of G have a strong influence on the difference between geometric and arithmetic mean, the bias correction κ_G is larger the smaller G_{min}. The cepstro-temporally smoothed and bias compensated spectral gain $\widetilde{G}_k(l)$ can now be applied to the noisy speech spectrum as in (1).

5 Evaluation

In this section we evaluate the bias correction derived in Section 4. As in [1] we compare CTS to the softgain

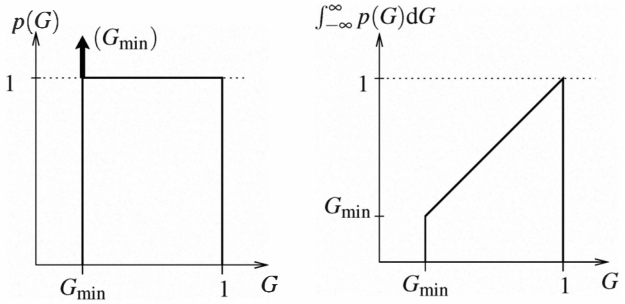

Figure 1: The assumed PDF $p(G)$ of the gain function (left) and its cumulative distribution (right).

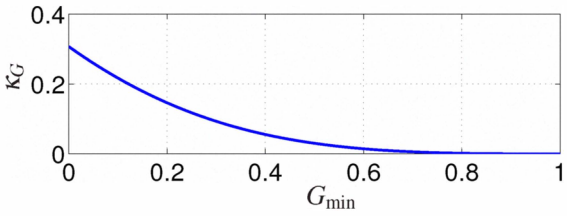

Figure 2: The bias correction $\kappa_G(G_{\min})$ for a CTS of the filter gain G, as a function of the lower limit $G_{\min}$ of the gain function.

method of [6]. We use the same smoothing constants for the softgain method and CTS as used for the listening tests in [1]. There, the smoothing constants were chosen so that both methods do not produce musical noise in stationary noise. As in [1] we set the lower limit on the gain function to $20\log 10(G_{\min}) = -15\,\mathrm{dB}$. In [1] listening tests indicated a clear preference for CTS. In this paper we evaluate the algorithms in terms of instrumental measures. We measure the SNR in terms of the frequency weighted segmental SNR (FW-SNR) [7], speech distortion in terms of the Itakura-Saito distance [7], and noise reduction according to [8]. We process 320 speech samples of [9, dialect region 6] that sum up to approximately 15 minutes of fluent, phonetically balanced conversational speech of both male and female speakers. The speech samples are disturbed by several noise types. The results are presented in Figure 3 for input segmental SNRs between -5 and 15 dB. For CTS we present results without a bias-correction (CTS-noCorr), with the bias correction (CTS-corr), and when the cepstrum is computed without the log function in (2) (CTS-noLog). As for CTS-noLog the temporal smoothing is done in the linear domain, a bias-correction is not necessary. The results are given in Figure 3. The FW-SNR and the Itakura-Saito distance indicate a decreased performance when comparing CTS-noCorr to the softgain method. This decrease of performance can be attributed to the bias that occurs due to the temporal smoothing in the log-domain. We see, that the decrease in performance is compensated with the proposed bias correction of (10), as CTS-noLog, CTS-corr, and the softgain method yield similar results in terms of FW-SNR, Itakura-Saito measure, and, for stationary noise, noise reduction. Further it can be seen that CTS is very effective in non-stationary noise. For babble noise CTS-corr and CTS-noLog achieve a higher noise reduction than the softgain method while the SNR and the speech distortion are virtually the same. This can be attributed to a successful elimination of spectral outliers caused by babble noise. Thus, even in babble noise, CTS yields an output signal without musical noise. In [1] the successful elimination of spectral outliers has been shown via statistical analyses, and listening tests indicated a residual noise of higher perceived quality.

6 Conclusion

In this paper we present a bias-compensation for a cepstro-temporal smoothing of spectral filter gain functions. We showed that in a speech enhancement system the bias introduced by a temporal smoothing in the cepstral domain yields a degradation of the output SNR and an increased speech distortion. The proposed bias compensation method is shown to successfully compensate for the introduced bias. Furthermore, compared to state-of-the-art single channel speech enhancement algorithms, cepstro-temporal smoothing is shown to yield higher noise reduction in babble noise, without an increase in speech distortion.

References

[1] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007.

[2] ——, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," *IEEE ICASSP*, pp. 4897–4900, Apr. 2008.

[3] N. Madhu, C. Breithaupt, and R. Martin, "Temporal smoothing of spectral masks in the cepstral domain for speech separation," *IEEE ICASSP*, pp. 45–48, Apr. 2008.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on Acoustics, Speech and Signal Proc.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[5] A. M. Noll, "Cepstrum pitch estimation," *Journal of the Acoustical Society of America*, vol. 41, pp. 293–309, Feb. 1967.

[6] D. Malah, R. Cox, and A. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," *IEEE ICASSP*, vol. 2, pp. 789–792, 1999.

[7] P. C. Loizou, *Speech Enhancement - Theory and Practice.* CRC Press, 2007.

[8] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP Journal of Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[9] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database," *National Institute of Standards and Technology (NIST)*, 1988.

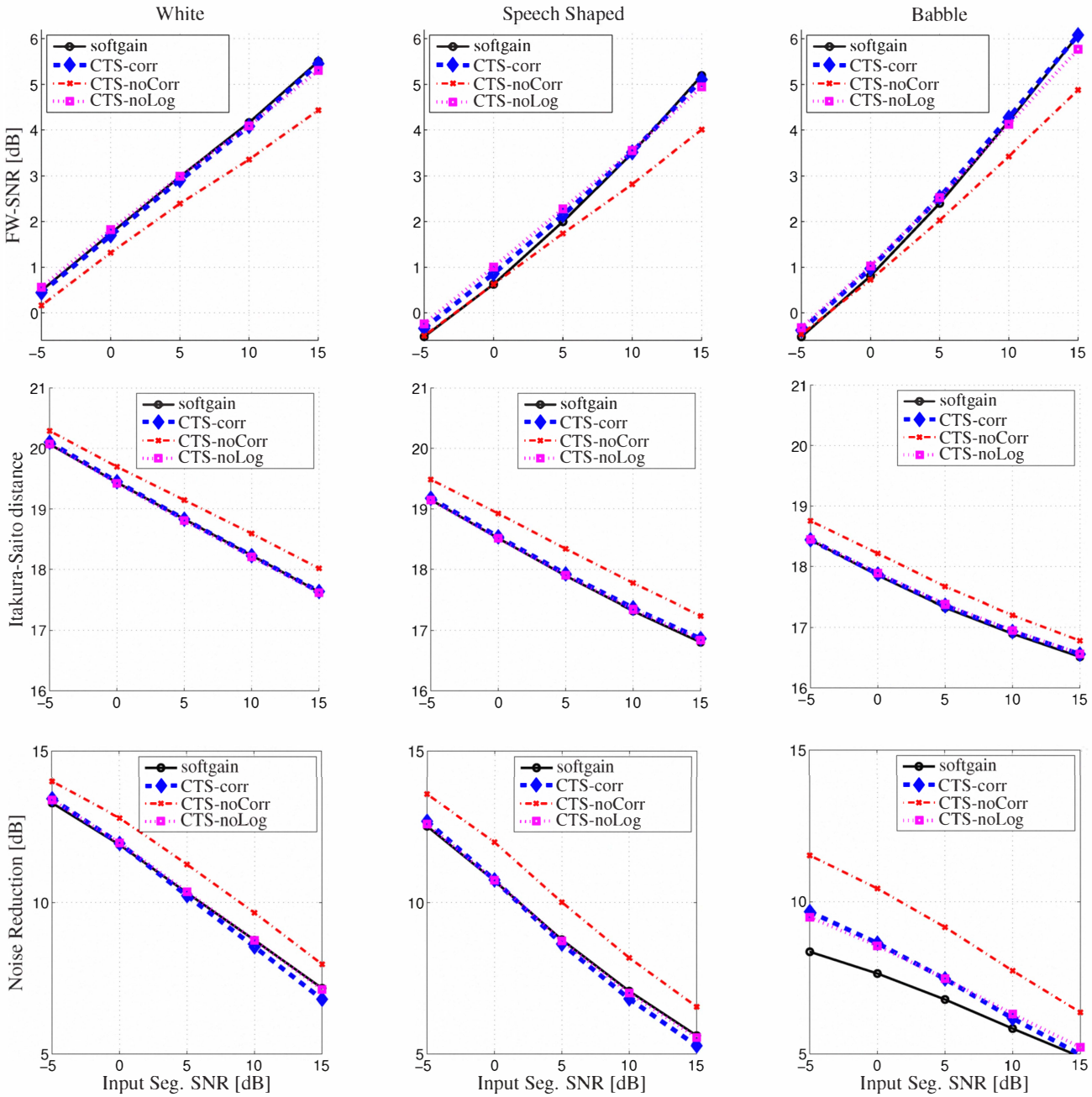

Figure 3: Averages of segmental frequency weighted SNR (top), Itakura-Saito distance (middle), and noise reduction (bottom) for 320 TIMIT sentences and white stationary Gaussian noise (left), speech shaped noise (middle), and babble noise (right). We present the results for the "softgain" approach [6], CTS of the gain function with (CTS-corr) and without (CTS-noCorr) the bias correction (10), and when no log function is used in (2) (CTS-noLog). For the FW-SNR (top) and noise reduction (bottom) larger values indicate increased performance. For the Itakura-Saito distance (middle) smaller values indicate better performance.