Cepstral Smoothing of Spectral Filter Gains for Speech Enhancement Without Musical Noise

Colin Breithaupt, Timo Gerkmann, and Rainer Martin, *Senior Member, IEEE*

***Abstract*—Many speech enhancement algorithms that modify short-term spectral magnitudes of the noisy signal by means of adaptive spectral gain functions are plagued by annoying spectral outliers. In this letter, we propose cepstral smoothing as a solution to this problem. We show that cepstral smoothing can effectively prevent spectral peaks of short duration that may be perceived as *musical noise*. At the same time, cepstral smoothing preserves speech onsets, plosives, and quasi-stationary narrowband structures like voiced speech. The proposed recursive temporal smoothing is applied to higher cepstral coefficients only, excluding those representing the pitch information. As the higher cepstral coefficients describe the finer spectral structure of the Fourier spectrum, smoothing them along time prevents single coefficients of the filter function from changing excessively and independently of their neighboring bins, thus suppressing *musical noise*. The proposed cepstral smoothing technique is very effective in nonstationary noise.**

***Index Terms*—Cepstral analysis, cepstral smoothing, musical noise, nonstationary noise, smoothing methods, speech enhancement.**

I. Introduction

FILTERS for the enhancement of noisy speech signals are often realized as a multiplicative gain in the discrete Fourier transform (DFT) domain. With a high spectral resolution, processing in the DFT domain allows for noise suppression in between pitch harmonics giving a relatively high auditive quality. A problem that comes along with this approach is the relatively large variance of spectral coefficients. In the adaptation of filter gains, spectral outliers may emerge that lead to an annoying auditive phenomenon called *musical noise* [1]. *Musical noise* and other artifacts are especially difficult to avoid under nonstationary noise conditions.

One way of preventing these artifacts is the soft-gain spectral weighting introduced in [2]. As this method relies on the estimates of the noise power spectral density and the signal-to-noise ratio (SNR), it is sensitive to estimation errors of these two parameters. As estimation errors are unavoidable in the statistical processing of noisy signals, different strategies have been presented that take these errors into account. In [3], a filter is described that has several parameters for adapting the spectral gain function to the noise condition. Another strategy is to search and remove spectral peaks in the filtered signal that lead to *musical noise* [4]. In [5], a recursive averaging is applied to the spectral gain function that smooths out fluctuations. As such a temporal smoothing would also severely affect speech components, the smoothing constant of this algorithm has to be carefully adapted.

In this contribution, we propose the smoothing of the filter gain function in the cepstral domain to suppress the tendency of adaptive spectral filters to produce *musical noise*. We show that this method also works well in nonstationary noises, to which the conventional approaches are particularly sensitive. The new method results in an effective smoothing of fine spectral variations that may be perceived as *musical noise*. At the same time, the spectral characteristics of speech are not affected.

This letter is structured as follows: In the following section, we introduce the conventional adaptation of filter gains and show how the amount of *musical noise* depends on the degree of temporal smoothing that is applied to the estimate of the SNR. Section III introduces the cepstral smoothing of the spectral gain function. The evaluation of the new method is detailed in Section IV.

II. Analysis of the Overall Filter Function

The observed noisy signal $y(t)$, where $t \in \mathbb{Z}$ is the discrete time index, is assumed to be a clean speech signal $s(t)$ perturbed by statistically independent additive noise $n(t)$, i.e., $y(t) = s(t) + n(t)$. The observed signal $y(t)$ is segmented into frames of length M, with frame overlap $M/2$, and weighted by a periodic Hann window $w_{\text{hann}}(\tau)$, $\tau = 0 \ldots M-1$. The weighted frames are transformed into the DFT domain resulting in the observed spectrum

$$\begin{aligned} Y(k,l) &= \text{DFT}\left\{ w_{\text{hann}}(\tau) y\left(l\frac{M}{2} + \tau \right) \right\} \\ &= S(k,l) + N(k,l) \end{aligned}$$

where $l \in \mathbb{Z}$ denotes the frame index, and $k = 0 \ldots M-1$ is the frequency bin index. Throughout this letter, the sampling rate of the signal is $f_s = 8$ kHz, and the DFT length is $M = 256$.

The clean speech spectral coefficients given the observation $Y(k,l)$ are estimated as $\hat{S}(k,l) = G(k,l)\,Y(k,l)$. The spectral filter gain function $G(k,l)$ can be the Wiener filter

$$G(k,l) = \frac{\hat{\xi}(k,l)}{1 + \hat{\xi}(k,l)} \tag{1}$$

where $\hat{\xi}$ denotes an estimate of the *a priori* SNR. It is generally obtained using the "decision-directed" approach [1] as follows:

$$\hat{\xi}(k,l) = \alpha \frac{|\hat{S}(k,l-1)|^2}{\hat{P}_n(k,l-1)} + (1-\alpha)(\hat{\gamma}(k,l) - 1). \tag{2}$$

Manuscript received March 21, 2007; revised June 11, 2007. The work of C. Breithaupt was supported in part by the German Research Foundation DFG and in part by the HOARSE project (HPRN-CT-2002-00276) funded by the European Union. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Alan McCree.

The authors are with the Institute of Communication Acoustics (IKA) Ruhr-University Bochum, 44780 Bochum, Germany (e-mail: colin.breithaupt@rub.de; timo.gerkmann@rub.de; rainer.martin@rub.de).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2007.906208

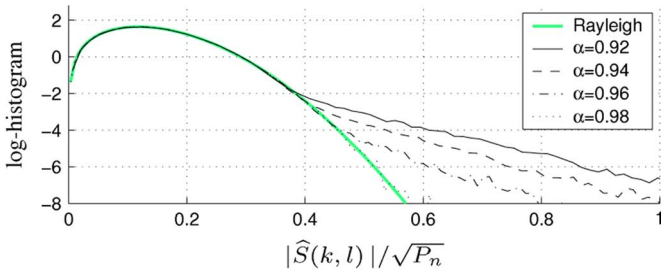

Fig. 1. Log-histogram of the *filtered* spectrum $|\hat{S}(k,l)|$ for Gaussian noise. For comparison, the Rayleigh pdf of the magnitude of Gaussian noise is also given. The lower the smoothing constant α, the more spectral outliers can be observed. The outliers of relatively large amplitude are perceived as *musical noise*.

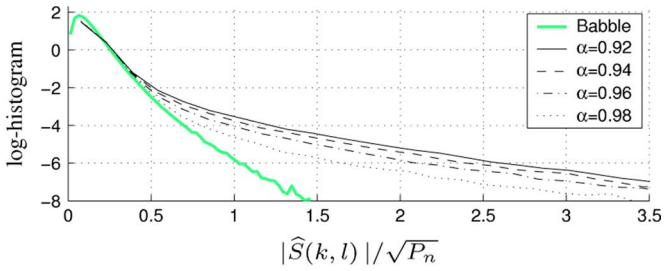

Fig. 2. Log-histogram of the filtered spectrum $|\hat{S}(k,l)|$ for babble noise. For comparison, the histogram of scaled unfiltered babble noise ($\hat{S}(k,l) = G_{\min} N(k,l)$) is also given. Due to high fluctuations in the noise, the number of outliers can hardly be controlled by the choice of α.

This estimate depends on an estimate $\hat{P}_n(k,l)$ of the noise power spectral density [6] and on an estimate $\hat{\gamma}(k,l) = |Y(k,l)|^2/\hat{P}_n(k,l)$ of the *a posteriori* SNR. Typical values for the smoothing factor α are in the range 0.92 to 0.98 [2]. A flooring of $\hat{\xi}(k,l)$ with respect to a minimum value $\xi_{\min}$ alleviates *musical noise* to some extent [2]. This is equivalent to defining a lower limit $G_{\min} = \xi_{\min}/(1+\xi_{\min})$ for (1). We set $\xi_{\min} = 0.2$ as in [2].

A drawback that comes with parameter estimators like (2) is *musical noise*. As reported in [2], a lower value of the smoothing constant α leads to an increased amount of these artifacts. In Fig. 1, we show the log-histogram of $|\hat{S}(k,l)|$ in the case of filtered white Gaussian noise. No speech is present. The log-histogram considers all spectral values of all frames l, excluding the dc and Nyquist frequency bin. For comparison, the Rayleigh probability density function (pdf) of the scaled magnitude of Gaussian noise ($|\hat{S}(k,l)| = G_{\min}|N(k,l)|$) is also given. In the graph, the filtered spectrum is normalized by $\sqrt{P_n}$. The filter uses (1) and (2) with $\alpha = 0.92 \ldots 0.98$ and $\xi_{\min} = 0.2$, and also the soft-gain of [2]. To avoid distortions of the speech signal, a lower value of α is desirable [2]. However, with lower values of α, the histogram has heavier tails caused by more outliers, as we see in Figs. 1 and 2. Correspondingly, the amount of *musical noise* perceivable in the filtered signal increases. Thus, the statistical analysis as seen in the histograms gives an indication of the *musical noise*. In this letter, in addition to listening tests, we therefore use log-histograms to assess the amount of *musical noise*.

The success of avoiding *musical noise* with the right choice of α is also limited by the statistical properties of the noise. In Fig. 2, it becomes evident that the log-histogram of filtered babble noise (many people speaking in the background) hardly changes with different choices of α. Short voiced babble-noise bursts in a frame l do not contribute to the noise power estimate immediately and are thus likely to increase the short-term *a priori* SNR estimate $\hat{\xi}(k,l)$. As a consequence of a larger value of $\hat{\xi}(k,l)$, the gain function (1) attenuates these fluctuations less, making them even more pronounced in the filtered spectrum.

In the following section, we therefore aim at the suppression of fluctuating peaks in the gain function $G(k,l)$ since they produce spectral outliers in the processed signal. We exploit the fact that these peaks are spectrally narrow and have a duration much shorter than the salient spectral features of speech.

III. Cepstral Smoothing

We propose a temporal smoothing of the cepstrum of the gain function $G(k,l)$ in order to avoid a peaked shape of $G(k,l)$ due to outliers in noise. The motivation for a cepstral representation of $G(k,l)$ is that speech characteristics and the unnatural noise artifacts are represented by a separate subset of coefficients in this domain. The cepstral bins describe different degrees of detail in the spectral structure. The coarse spectral envelope of speech is described by the first few cepstral coefficients. The pitch is represented by only one or by two consecutive higher coefficients. Spectral peaks in $G(k,l)$ caused by outliers will be represented by some of the remaining higher cepstral coefficients, because they belong to the fine structure of $G(k,l)$. Smoothing these higher cepstral coefficients will reduce their temporal dynamics. As the narrow spectral peaks of single outliers appear only for a duration of a single frame, they are strongly affected by such a cepstral smoothing.

Speech onsets and the spectral envelope of fricatives and plosives must not be distorted by the smoothing procedure. Therefore, the smoothing is not applied to low cepstral coefficients. This preserves the principal structure of the gain function in the case of speech presence. Additionally, less smoothing is used for the pitch-related coefficients. Due to the reduced smoothing of these cepstral coefficients and the relatively long duration of voiced speech sounds, the fine structure of speech—like pitch harmonics—is not affected severely.

A cepstral representation of $G(k,l)$ from (1) is calculated for each frame l as

$$G^{\text{cepst}}(k',l) = \text{IDFT}\{\log(G(k,l))\} \tag{3}$$

where $\text{IDFT}\{\cdot\}$ is the inverse DFT of length M resulting in cepstral bins k'. For reasons of symmetry, we only need to consider the first $D = M/2+1$ bins for the following description. Note that the IDFT can be replaced by the discrete cosine transform (DCT), as is common practice in feature-extraction frontends for speech recognition.

A smoothed version $G^{\text{cepst}}_{\text{smooth}}(k',l)$ is calculated as

$$G^{\text{cepst}}_{\text{smooth}}(k',l) = \beta G^{\text{cepst}}_{\text{smooth}}(k',l-1) + (1-\beta)G^{\text{cepst}}(k',l). \tag{4}$$

The smoothing is applied to cepstral bins $k' \in \{k'_{\text{low}} \ldots D-1\}\backslash\mathbb{K}'$. The set of cepstral indices $\mathbb{K}'$ consists of the cepstral index k'_{pitch} of the pitch and its two cepstral neighbors, i.e., $\mathbb{K}' = \{k'_{\text{pitch}}-1, k'_{\text{pitch}}, k'_{\text{pitch}}+1\}$. The pitch index k'_{pitch} is determined as the index of the maximum value of $G^{\text{cepst}}(k',l)$ in the range that corresponds to a pitch frequency between 70 and 500 Hz. For $f_s = 8$ kHz, the search interval thus is $k' =$

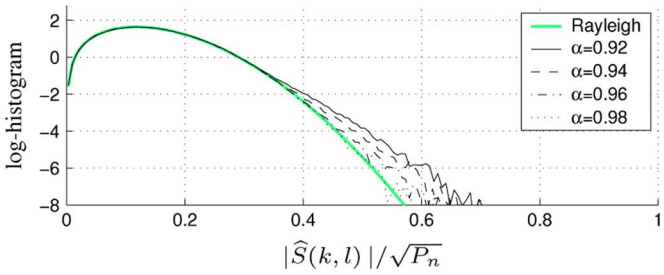

Fig. 3. Log-histogram of the filtered spectrum $|\hat{S}(k,l)|$ for Gaussian noise as in Fig. 1 but with cepstral smoothing applied. Cepstral smoothing reduces the number of outliers considerably. The residual noise sounds like Gaussian noise.

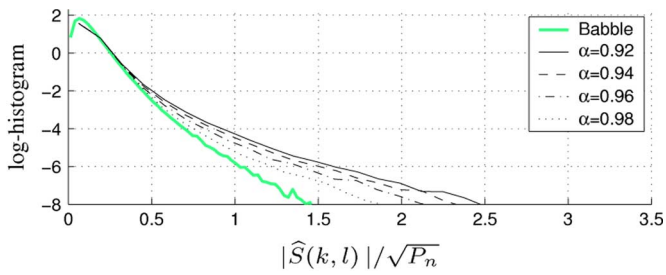

Fig. 4. Log-histogram of the filtered spectrum $|\hat{S}(k,l)|$ for babble noise as in Fig. 2 but with cepstral smoothing applied. The residual noise sounds more natural, as spectral outliers are less pronounced.

$16 \ldots 114$. k'_{pitch} is only determined when speech presence is signaled by a voice activity detection (VAD) [7]. Otherwise, $\mathbb{K}'$ is the empty set. For the cepstral coefficients $k' \in \mathbb{K}'$, we use a lower smoothing constant β_{pitch} in (4). For $k' \in \{0 \ldots k'_{\text{low}} - 1\}$, no smoothing is applied at all; therefore, $G^{\text{cepst}}_{\text{smooth}}(k',l) = G^{\text{cepst}}(k',l)$. All of the above smoothing operations have to be applied accordingly to the remaining symmetric half of the cepstrum, $k' = D \ldots M - 1$.

Note that the log-function in (3) is not essential for the selective smoothing just described. Nevertheless, we found that this nonlinear transform of $G(k,l)$ considerably reduces noise shaping caused by (4) in stationary Gaussian noise.

The final smoothed spectral gain function is obtained by a transform inverse to (3) as follows:

$$G_{\text{smooth}}(k,l) = \exp\left(\text{DFT}\left\{G^{\text{cepst}}_{\text{smooth}}(k',l)\right\}\right) \qquad (5)$$

where $G_{\text{smooth}}(k,l)$ is additionally constrained to values below or equal to one. The resulting filter gain can then be applied instead of (1).

Although a VAD is used for finding k'_{pitch}, we found that false alarms do not have a large effect. For background noises or unvoiced sounds, the maximum cepstral bin in the pitch range does not contribute as significantly to the filter result as in the case of voiced speech.

IV. Evaluation

We now compare the above algorithm with a conventional approach that does not use cepstral smoothing. The evaluation of the algorithms is done in three steps. First, we analyze the statistics of Gaussian noise and babble noise after processing. Then, a comparison of spectrograms demonstrates how cepstral smoothing reduces noise fluctuations and at the same time preserves the speech characteristics. Finally, the results of listening tests are presented.

Figs. 3 and 4 depict the log-histograms corresponding to Figs. 1 and 2 but with cepstral smoothing (4) applied. From the figures, it becomes clear that the number of outliers is dramatically reduced. Even for $\alpha = 0.92$, no *musical noise* is perceivable in Gaussian noise. The filtered babble noise also sounds more natural than in the case of the unmodified filter. The amplification of voiced bursts is alleviated so that tonal residuals occur to a much lesser degree.

For the analysis of the spectra and the listening test, we used the following configuration of the filters. The conventional reference algorithm uses $\alpha = 0.97$ as the smoothing factor in (2). This gave a good trade-off between the suppression of noise fluctuations in *stationary* noise and the audible distortions of speech. Additionally, we applied the multiplicative soft-gain [2] that considers speech presence uncertainty. We denote the conventional filter as "soft-gain." As recommended in [2], we also replace the filter gain $G(k,l)$ by the constant $G_{\min}$, whenever the VAD signals speech absence.

For the proposed approach, we set $\alpha = 0.94$ in (2). The smoothing factor for the cepstral smoothing is set to $\beta = 0.8$. With $k'_{\text{low}} = 4$, a sufficient protection of the speech envelope is achieved. For $k' \in \mathbb{K}'$, we choose $\beta_{\text{pitch}} = 0.4$. The filter with cepstral smoothing neither uses the soft-gain [2] nor the VAD-based substitution $G(k,l) \rightarrow G_{\min}$ during speech pauses, because it has no audible effect.

Fig. 5 shows the spectrograms of a noisy speech sample filtered with the conventional approach and the cepstrally smoothed filter, respectively. The noisy signal is perturbed by babble noise, which has been recorded inside a cafeteria. It consists of nonstationary speech bursts and a relatively stationary floor due to reverberation. While the conventional filter is not able to suppress voiced bursts in babble, it suppresses the stationary portion. This enhances the spectral contrast in the residual noise. The filtered babble contains unnatural tonal residuals. The cepstrally smoothed filter suppresses the stationary portions of the babble noise to the same degree. Additionally, it hinders tonal contents from being emphasized.

The cepstrally smoothed filter also better preserves important structures of speech. In the spectrograms, it can be seen that spectrally broad sounds are less distorted, e.g., the marked plosive /k/. As pitch harmonics are less smoothed, low energy pitch harmonics are also better preserved.

The choice of parameters $\alpha = 0.94$ and $\beta = 0.8$ for our approach results in a slight noise shaping at the end of words for white and pink Gaussian noise, which makes the speech sound slightly reverberant. This does not occur for $\beta = 0.7$. However, many participants of the listening test indicated increased listening comfort when the reverberation effect was present, as low energy syllables at the end of words are less attenuated.

In listening tests, we compared the performance of the conventional method to the new cepstral smoothing approach in four nonstationary noisy environments (babble, subway, street, and white noise bursts) and two stationary noisy environments (pink and white). For each noise type, ten different speech samples from [8] were presented, five spoken by male, five by female speakers. In order to allow the subjects to get an impression of the residual noise by itself, the speech samples were preceded

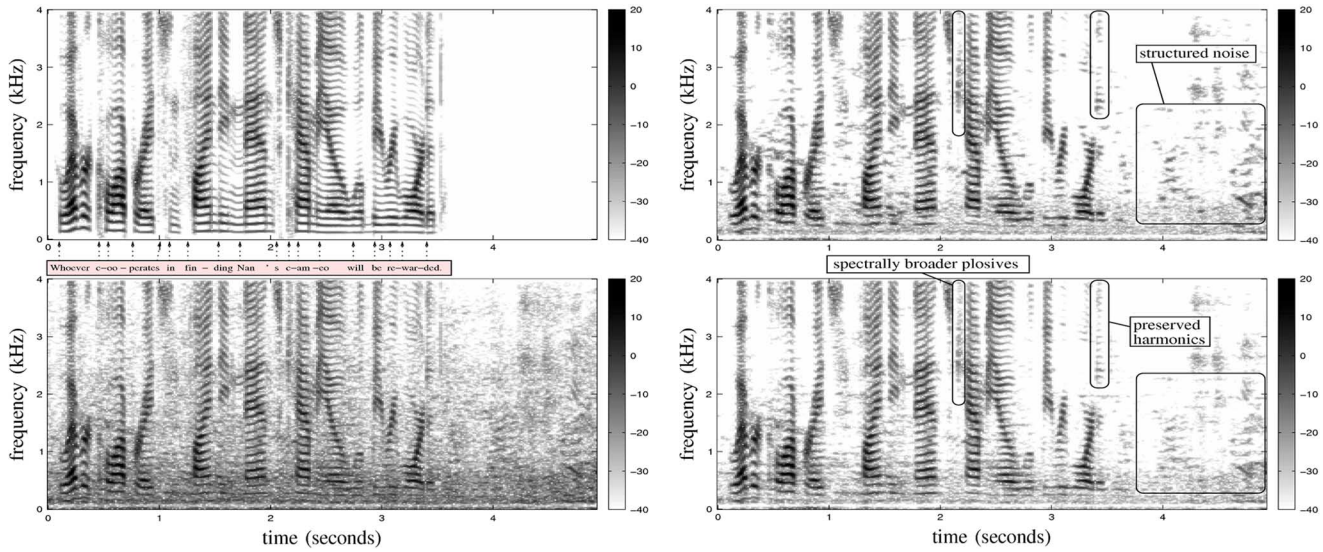

Fig. 5. Comparison of spectrograms. The sentence is "Whoever cooperates in finding Nan's cameo will be rewarded." On the left, the clean (top) and the unfiltered signal (bottom) are shown. The noise is babble noise at an SNR of 0 dB. On the right, the filtered signals are given. The conventional filter (top) causes more noise fluctuations than the filter with cepstral smoothing (bottom). At the same time, cepstral smoothing preserves the structure of speech better.

TABLE I
RESULTS OF THE LISTENING TEST FOR BABBLE AND PINK NOISE. THE NUMBERS STATE THE PERCENTAGE OF VOTES IN FAVOR OF ONE OF THE FILTERS. THE CHOICE "EQUALLY SUITED" WAS ALSO POSSIBLE

Noise	Category	Ceps. Sm.	Soft-gain [2]	Equally Suited
Babble	Backgr.	68%	5%	27%
	Speech	54%	8%	38%
	Overall	75%	7%	18%
Pink	Backgr.	18%	18%	64%
	Speech	52%	23%	25%
	Overall	50%	22%	28%

and followed by speech pauses of 3 s overall duration. The average duration of the resulting samples was about 7 s. The noise was scaled and added such that the noisy samples had an average segmental SNR of 0 dB in frames where speech is present. Each of the noisy samples was filtered by the conventional and proposed approach, respectively, resulting in ten pairs of enhanced samples per noise type. The participants were asked to select the file in each pair they preferred in terms of speech quality, naturalness of the background, and overall quality, respectively. The comparison was done blindly and in randomized order. The participants were divided into two groups: experts and nonexperts. While the seven expert listeners clearly favored the proposed cepstral smoothing approach, we would like to present detailed results only for the 12 nonexpert listeners for babble and pink noise in Table I. It may be seen that in nonstationary environments, the participants favored the cepstral approach. This is because the background noise sounds less tonal and thus more natural with the proposed approach. This is achieved without affecting the speech quality. On the contrary: for stationary noise sources, where both algorithms perform equally well in terms of background quality (no *musical noise*), a preference for our approach in terms of speech and overall quality may be seen. Note that the parameter settings for all noisy environments were the same. Audio examples are available at [9].

V. CONCLUSION

Cepstral smoothing is a useful amendment to speech enhancement filters operating in real noise environments. Annoying noise fluctuations are prevented even in the case of babble noise. As opposed to conventional methods, cepstral smoothing allows for a selective smoothing of different spectral structures represented by the respective cepstral coefficients. This makes the protection of the characteristics of speech possible while *musical noise* is suppressed.

REFERENCES

[1] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[2] D. Malah, R. V. Cox, and A. J. Accardi, "Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments," in *Proc. IEEE ICASSP*, 1999, vol. 2, pp. 789–792.
[3] K. Linhard and T. Haulick, "Noise subtraction with parametric recursive gain curves," in *Proc. Eurospeech—Eur. Conf. Speech Communication and Technology*, Sep. 1999, pp. 2611–2614.
[4] Z. Goh, K.-C. Tan, and B. Tan, "Postprocessing method for suppressing musical noise generated by spectral subtraction," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 3, pp. 287–292, May 1998.
[5] H. Gustafsson, S. E. Nordholm, and I. Claesson, "Spectral subtraction using reduced delay convolution and adaptive averaging," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 8, pp. 799–807, Nov. 2001.
[6] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 5, pp. 504–512, Jul. 2001.
[7] C. Breithaupt and R. Martin, "Voice activity detection in the DFT domain based on a parametric noise model," in *Proc. Int. Workshop Acoustic Echo and Noise Control (IWAENC)*, 2006.
[8] J. S. Garofolo, "DARPA TIMIT acoustic-phonetic speech database (prototype distribution)," National Institute of Standards and Technology (NIST), 1988.
[9] C. Breithaupt and T. Gerkmann, Cepstal Smoothing: Audio Examples. [Online]. Available: http://www2.ika.rub.de/audioexamples/csmooth/csmooth.html.