A NOVEL A PRIORI SNR ESTIMATION APPROACH BASED ON SELECTIVE CEPSTRO-TEMPORAL SMOOTHING

Colin Breithaupt, Timo Gerkmann, and Rainer Martin

Institute of Communication Acoustics (IKA)
Ruhr-Universität Bochum, 44780 Bochum, Germany
{colin.breithaupt,timo.gerkmann,rainer.martin}@rub.de

ABSTRACT

While state-of-the-art approaches obtain an estimate of the *a priori* SNR by adaptively smoothing its maximum likelihood estimate in the frequency domain, we selectively smooth the maximum likelihood estimate in the cepstral domain. In the cepstral domain the noisy speech signal is decomposed into coefficients related mainly to the speech envelope, the excitation, and noise. As in the cepstral domain coefficients that represent speech can be robustly determined, we can apply little smoothing to speech coefficients and strong smoothing to noise coefficients. Thus, speech components are preserved and musical noise is suppressed. In speech enhancement experiments we obtain consistent improvements over the well known *decision-directed* approach.

Index Terms— Speech enhancement, decision-directed approach, SNR estimation, musical noise, cepstral analysis.

1. INTRODUCTION

Many of the most successful adaptive speech enhancement algorithms, *e.g.* those based on Wiener filtering, work in the short-time Fourier transform (STFT) domain. A drawback of STFT-based speech enhancement algorithms is that they yield unnatural sounding structured residual noise, often referred to as *musical noise* [1]. Musical noise can be avoided by trading off against noise suppression [2] or speech distortion [1]. Increasing the noise suppression without increasing musical noise or speech distortion remains a challenge especially in non-stationary noise.

The estimation of the *a priori* signal-to-noise (SNR) is a crucial part of speech enhancement algorithms [3]. An erroneous estimation of this parameter leads to speech distortion, musical noise, or reduced noise reduction. In non-stationary noise the estimation of the *a priori* SNR is particularly difficult.

In this paper we present an estimator for the *a priori* SNR that distinguishes sporadic narrowband noise bursts from speech by taking into account *a priori* knowledge about the speech production process. Recently, applying temporal smoothing in the cepstral domain was found to be a promising approach for speech enhancement in non-stationary noise environments [4]. In the cepstral domain the noisy speech signal is decomposed into coefficients related to the speech envelope, the excitation, and noise. While the speech envelope is always represented by the same small set of cepstral coefficients, the coefficients that represent the excitation can be found by searching for a cepstral peak in a defined range [5]. The remaining coefficients are dominated by noise. We can thus apply selective temporal smoothing to the cepstral representation of a maximum likelihood estimate of the speech power spectral density, *i.e.* strong smoothing to those coefficients that are dominated by noise, and only little smoothing to the coefficients representing speech. We show that the proposed method avoids spectral outliers in the residual noise signal, while the speech characteristics are preserved.

The paper is organized as follows: In Section 2 we review the *decision-directed* approach [3] which is most frequently used in state-of-the-art estimators. In Section 3 we present a novel estimation approach based on cepstral decomposition. In Section 4 we show that the proposed approach outperforms the *decision-directed* approach for non-stationary noise as well as for stationary noise in terms of several instrumental measures.

2. REVIEW OF A PRIORI SNR ESTIMATION

We assume an additive mixture of speech, $S(k,l)$, and noise, $N(k,l)$, in the STFT domain, where $S(k,l)$ and $N(k,l)$ are independent to each other. Here, k is the frequency index and l is the frame index. The noisy observation, $Y(k,l)$, is thus given by $Y(k,l) = S(k,l) + N(k,l)$. The *a priori* SNR, ξ, is defined as the ratio of the speech power, $\lambda_s(k) = \mathrm{E}\{|S(k)|^2\}$, and the noise power, $\lambda_n(k) = \mathrm{E}\{|N(k)|^2\}$. A maximum likelihood (ML) estimate, $\xi^{\mathrm{ml}}(k,l)$, of the *a priori* SNR given the *a posteriori* SNR, $\gamma(k,l) = \frac{|Y(k,l)|^2}{\lambda_n(k)}$, can be obtained as [6]:

$$\xi^{\mathrm{ml}}(k,l) = \gamma(k,l) - 1\,. \tag{1}$$

Any deviation of $|N(k,l)|^2$ from its expected value, $\lambda_n(k)$, will cause fluctuations in the ML SNR estimate, $\xi^{\mathrm{ml}}(k,l)$. When employed in a speech enhancement framework, these fluctuations yield an unnatural sounding residual noise. In [3], before introducing the *decision-directed* approach, Ephraim and Malah derived an ML estimator based on consecutive analysis frames that results in a recursive smoothing of (1). This recursive smoothing can be interpreted as an approximation of the true *a priori* SNR $\xi(k,l) = \mathrm{E}\{\xi^{\mathrm{ml}}(k,l)\}$, assuming that the speech signal is ergodic. However, since speech is highly non-stationary (and hence not ergodic), recursive smoothing results in a poor trade-off between fluctuations in the residual noise and distortion of speech onsets and transients. If the recursive smoothing constant is chosen high enough to eliminate fluctuations in $\xi^{\mathrm{ml}}(k,l)$, it also distorts speech onsets and transitions, resulting in a reduced speech quality. Therefore, in state-of-the-art speech enhancement algorithms the *a priori* SNR is estimated in a *decision-directed* way [3, 7], *i.e.* based on a previous clean-speech estimate

The work of C. Breithaupt is funded by the German Research Foundation DFG.

$\widehat{S}(k,l-1)$:

$$\widehat{\xi}(k,l) = \max\left\{\alpha_{\text{dd}}\frac{|\widehat{S}(k,l-1)|^2}{\widehat{\lambda}_n(k,l-1)} + (1-\alpha_{\text{dd}})\widehat{\xi}^{\,\text{ml}}(k,l)\,,\xi_{\text{min}}\right\}, \tag{2}$$

where $\widehat{\lambda}_n$ is the estimated noise power, and $\widehat{\xi}^{\,\text{ml}}$ is $\xi^{\,\text{ml}}$ using $\widehat{\lambda}_n$. The parameters α_{dd} and ξ_{min} control the trade-off between noise reduction and distortion of speech transients in a speech enhancement framework [1]. The *decision-directed* procedure (2) allows for a fast tracking of increasing levels of the speech power, thus effectively resulting in an adaptive smoothing. Consequently, at speech onsets and transitions, less speech distortion is introduced. However, since the *decision-directed* SNR estimator is sensitive to rising spectral amplitudes, it does not only respond to speech onsets, but also to noise bursts that are not tracked by the noise power estimation algorithm. Therefore, noise bursts will cause a rising *a priori* SNR estimate, and thus outliers in the residual noise of the clean-speech estimate that are perceived as annoying musical tones. The SNR estimation approach proposed next is capable of avoiding these annoying outliers while preserving the speech characteristics.

3. PROPOSED A PRIORI SNR ESTIMATION

From the ML SNR estimate (1) we compute the speech power

$$\lambda_s^{\text{ml}}(k,l) = \lambda_n\ \max\left\{\xi^{\,\text{ml}}(k,l)\,,\xi_{\text{min}}^{\text{ml}}\right\}, \tag{3}$$

which is temporally smoothed in the cepstral domain. Here $\xi_{\text{min}}^{\text{ml}} > 0$ is a small lower bound which prevents $\xi^{\,\text{ml}}$ from taking negative values or values close to zero and thus avoids numerical difficulties in the following steps. A cepstral representation of $\lambda_s^{\text{ml}}(k,l)$ is calculated as

$$\lambda_s^{\text{ml,ceps}}(q,l) = \text{IDFT}\left\{\log\Big(\lambda_s^{\text{ml}}(k,l)\Big)\Big|_{k=0,\ldots,(M-1)}\right\}, \tag{4}$$

with $q = 0,...,(M-1)$ the cepstral bin index, M the length of the inverse discrete Fourier transform (IDFT), and $\log(\cdot)$ the natural logarithm. Note that the symmetry condition $\lambda_s^{\text{ml,ceps}}(M-q,l) = \lambda_s^{\text{ml,ceps}}(q,l)$ holds. Therefore, all further modifications applied to the cepstral bins $q = 0,...,M/2$ are applied accordingly to the symmetric counterpart $q = M/2+1,...,(M-1)$. An adaptive recursive smoothing is applied in the cepstral domain:

$$\lambda_s{}^{\text{ceps}}(q,l) = \alpha(q,l)\,\lambda_s{}^{\text{ceps}}(q,l-1) + \big(1-\alpha(q,l)\big)\,\lambda_s^{\text{ml,ceps}}(q,l)\,. \tag{5}$$

The smoothing factor, $\alpha(q,l)$, should be chosen so that only little smoothing is applied to the low cepstral coefficients which represent the speech envelope. Additionally, the cepstral bins which are likely to represent the fundamental frequency, f_0, should not be smoothed.

3.1. Adaptive smoothing factors

As the speech envelope is always represented by a constant set of cepstral bins with low index (see Figure 1, $q < 20$), the smoothing factor $\alpha(q,l)$ always has low values for $q < 20$ in order to protect rapid changes in the speech spectral envelope. Nevertheless, $\alpha(q,l)$ is updated for every signal frame in order to adapt the smoothing of cepstral bins that possibly describe the fundamental frequency, f_0.

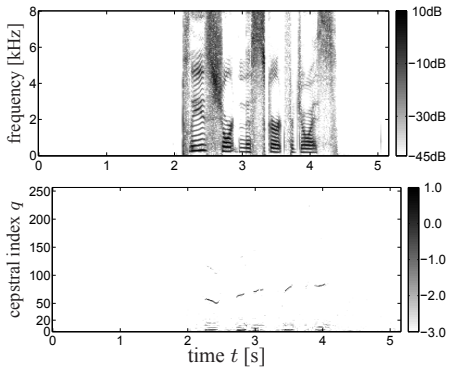

Fig. 1. Spectrogram of clean signal (top) and corresponding cepstrum (absolute values, logarithmic scale).

After a f_0 estimation (see Section 3.2), the smoothing factor $\alpha(q,l)$ in (5) is obtained as

$$\alpha(q,l) = \begin{cases}\alpha_{\text{pitch}} & \text{if } q \in \mathbb{Q}_{\text{pitch}}\,,\\ \overline{\alpha}(q,l) & \text{if } q \in \{0,...,M/2\}\setminus\mathbb{Q}_{\text{pitch}}\,,\end{cases} \tag{6}$$

where $\mathbb{Q}_{\text{pitch}}$ is a set of adjacent cepstral bins that are most likely to represent f_0, and α_{pitch} is the low smoothing constant for these bins. $\overline{\alpha}(q,l)$ contains information about the f_0 estimation from previous frames, and is gained as

$$\overline{\alpha}(q,l) = \beta\,\alpha(q,l-1) + (1-\beta)\,\overline{\alpha}^{\,\text{const}}(q)\,. \tag{7}$$

The smoothing constant β is the forgetting factor that determines how fast a value of $\alpha(q,l)$ adapts from α_{pitch} to $\overline{\alpha}^{\,\text{const}}(q)$, if it has been lowered in previous frames. Thus, an estimation error of f_0 in the current frame l does not lead to an immediate strong smoothing of the true current cepstral bin representing f_0 in (5). The stationary values $\overline{\alpha}^{\,\text{const}}(q)$ of the recursion are chosen so that little smoothing is applied to lower cepstral bins that represent the spectral envelope and a strong smoothing to higher cepstral coefficients except $q \in \mathbb{Q}_{\text{pitch}}$.

3.2. Fundamental frequency estimation

Because the cepstral coefficient that represents the fundamental frequency, f_0, varies over time (see Figure 1, $q > 50$), an f_0 estimation is employed. Here any f_0 estimation algorithm found in the literature may be considered. In our algorithm we use the method described next: Since the power of voiced sounds is less at high frequencies, the f_0 estimation algorithm is more robust, if only the spectrum up to a certain cut-off frequency is considered. This low-pass filtering of the log spectrum can be achieved by convolving the cepstral frame with a short Hamming window, $w_{\text{H}}(q)$, of length τ_{H} taps:

$$\bar{\lambda}_s^{\text{ml,ceps}}(q,l) = \lambda_s^{\text{ml,ceps}}(q,l) * w_{\text{H}}(q)\,. \tag{8}$$

The cepstral index $q_{\text{pitch}}(l)$ that most likely represents f_0 is found via a maximum search for a given frame, l, as [5]

$$q_{\text{pitch}}(l) = \underset{q}{\text{argmax}}\big\{\bar{\lambda}_s^{\text{ml,ceps}}(q,l)\big|q_{\text{low}} \le q \le q_{\text{high}}\big\}, \tag{9}$$

where the search is limited to possible fundamental frequencies between $f_{0,\text{low}}$ and $f_{0,\text{high}}$, resulting in the range $q_{\text{low}} = \lfloor f_s/f_{0,\text{high}}\rfloor$

to $q_{\text{high}} = \lfloor f_s / f_{0,\text{low}} \rfloor$, with f_s the sampling rate and $\lfloor \cdot \rfloor$ the flooring operator towards the nearest integer number. Note that (9) only yields meaningful results, if voiced speech is present.

We detect voiced speech sounds by means of two criteria. As voiced speech sounds have a comparatively high energy, the first criterion is the comparison of the cepstral peak value to a threshold, Λ^{thr}. This guarantees that $q_{\text{pitch}}(l)$ represents a considerable portion of the signal energy. The second criterion is derived from the fact that voiced speech is spectrally tilted, thus having more energy at low frequencies. This fact is reflected in the cepstral coefficient at $q = 1$ being positive. Thus, the set of cepstral bin indices associated with the fundamental frequency, $\mathbb{Q}_{\text{pitch}}$, is gained as

$$\mathbb{Q}_{\text{pitch}} = \begin{cases} \mathbb{Q}'_{\text{pitch}} & \text{if } \bar{\lambda}_s^{\text{ml,ceps}}(q_{\text{pitch}}, l) \geq \Lambda^{\text{thr}} \\ & \text{AND } \lambda_s^{\text{ml,ceps}}(1, l) > 0\,, \\ \emptyset & \text{else}\,, \end{cases} \tag{10}$$

where $\mathbb{Q}'_{\text{pitch}} = \{q_{\text{pitch}} - \Delta q_{\text{pitch}}, ..., q_{\text{pitch}} + \Delta q_{\text{pitch}}\}$ is the range of cepstral bins that are most likely to represent the fundamental frequency, Δq_{pitch} is a small margin, and $\emptyset$ is the empty set. A suitable value for the threshold Λ^{thr} is found from tests with representative noisy data. Note that Λ^{thr} is not a sensitive parameter, as long as it is chosen low enough so that no voiced speech is clipped.

3.3. log-bias correction and *a priori* SNR estimate

The recursive smoothing of the ML speech power estimate, λ_s^{ml}, can be seen as an approximation of the true speech power $\lambda_s = \lambda_n \mathrm{E}\{\gamma - 1\}$ (cf. Section 2). However, the recursive averaging in (5) is done in the log–domain, which results in a bias. This bias can be corrected via the correction κ [8], as :

$$\log\Big(\mathrm{E}\Big\{\lambda_s^{\text{ml}}\Big\}\Big) = \mathrm{E}\Big\{\log\left(\lambda_s^{\text{ml}}\right)\Big\} + \kappa\,. \tag{11}$$

Assuming zero-mean complex Gaussian distributed spectral coefficients, the logarithm of the correction factor, κ, equals the Euler constant, $\kappa^{\text{Gauss}} = 0.5772...$ [8]. Note that we only apply little smoothing to the actual speech coefficients. Thus, our estimate will be between the unbiased instantaneous values and the expected values biased according to (11). Consequently, we have to choose a lower bias correction. We found that $\kappa = 0.5\,\kappa^{\text{Gauss}}$ is sufficient to obtain a good quality of the processed speech.

A smoothed estimate $\widehat{\lambda}_s(k, l)$ of the speech power in the spectral domain is finally obtained by transforming ${\lambda_s}^{\text{ceps}}(q, l)$ back to the spectral domain, and by compensating for the bias:

$$\widehat{\lambda}_s(k, l) = \exp\left(\kappa + \mathrm{DFT}\left\{{\lambda_s}^{\text{ceps}}(q, l)\right\}\Big|_{q=0,\ldots,(M-1)}\right). \tag{12}$$

With the flooring, ξ_{min}, proposed in [1], the final *a priori* SNR estimate is computed as

$$\widehat{\xi}(k, l) = \max\left\{\frac{\widehat{\lambda}_s(k, l)}{\widehat{\lambda}_n(k, l)}, \xi_{\text{min}}\right\}. \tag{13}$$

4. EVALUATION

For the evaluation we implement the proposed *a priori* SNR estimator in a speech enhancement filter. The estimator uses parameter values as in Table 1. For comparison, we alternatively use the *decision-directed* estimation approach (2) with $\alpha_{\text{dd}} = 0.98$ as proposed in [3]. The estimate $\widehat{\lambda}_n(k, l)$ of the noise power is obtained with the method from [9]. The sampling rate of the system is $f_s = 16\,\text{kHz}$, the frame-length is $M = 512$. The frameshift is $M/2$. Each frame is weighted with a M-tap Hann window and transformed with a DFT of length M. An estmate $\widehat{S}(k, l) = G(k, l)\, Y(k, l)$ of the clean speech spectral coefficient $S(k, l)$ is obtained by the Wiener filter gain function $G^{\text{wiener}}(k, l) = \widehat{\xi}(k, l)\,/(1 + \widehat{\xi}(k, l))$. The gain function is floored to G_{min} as proposed in [2] in order to prevent musical noise in stationary noises: $G(k, l) = \max\{G^{\text{wiener}}(k, l)\,, G_{\text{min}}\}$. The enhanced time signal is finally obtained using the overlap-add method.

f_s	$= 16\,\text{kHz}$	M	$= 512$	Δq_{pitch}	$= 2$
$10\log_{10}(\xi_{\text{min}})$	$= -25\text{dB}$	Λ^{thr}	$= 0.2$	α_{pitch}	$= 0.2$
$10\log_{10}(\xi_{\text{min}}^{\text{ml}})$	$= -27\text{dB}$	$f_{0,\text{low}}$	$= 70\text{Hz}$	β	$= 0.96$
$20\log_{10}(G_{\text{min}})$	$= -17\text{dB}$	$f_{0,\text{high}}$	$= 300\text{Hz}$	τ_{H}	$= 8$

$$\overline{\alpha}^{\,\text{const}}(q) = \begin{cases} 0.5 & \text{if } q \in \{0, ..., 2\} \\ 0.7 & \text{if } q \in \{3, ..., 19\} \\ 0.97 & \text{if } q \in \{20, ..., 256\} \end{cases}$$

Table 1. Parameter values for the evaluated system.

We process 320 speech samples of [10, 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 average per speech-sample values of the improvement of the segmental SNR, the segmental speech SNR [11], and the segmental noise reduction [11] are given in Figure 2. With the proposed method consistent improvements of the segmental SNR are obtained. The noise suppression is virtually identical for the stationary noises, as in low SNR conditions the limiting constant G_{min} comes into effect rendering $G^{\text{wiener}}(k, l)$ without effect. Nevertheless, in non–stationary babble noise, noise bursts result in estimation errors in the case of the *decision-directed* approach, while the proposed method prevents such outliers, thus obtaining a better noise reduction. In terms of the speech distortion measure, *i.e.* speech SNR, the cepstral approach yields better results for all noise types. Note that in the case of babble noise, an improvement for both speech distortion and noise reduction is obtained simultaneously. In Figures 1 and 3, spectrograms of the sentence "Please shorten this skirt for Joyce." (female voice) are shown for a stationary and a non-stationary noise. The proposed method notably preserves plosives (time $t = 2.1\,\text{s}$), vowels (*e.g.* $t = 2.3\,\text{s}$) and the envelope of fricatives (*e.g.* $t = 4.2\,\text{s}$ for white noise). At the same time, musical noise due to narrow-band bursts in non-stationary noise – like in babble noise – is effectively prevented.

Informal subjective listening reveals that signals processed with the filter using the proposed *a priori* SNR estimator sound clearer in the case of white Gaussian and speech shaped noise, as more low–energy speech components are preserved. As for the residual noise, in the case of white and speech shaped noise, neither approach produces musical noise. For babble noise, the speech signals of both approaches sound similar, but the new approach is able to suppress musical noise even during speech presence.

5. CONCLUSION

For speech enhancement frameworks, smoothing the maximum likelihood estimate of the signal-to-noise ratio is indispensable for the suppression of musical noise. We show that a temporal smoothing

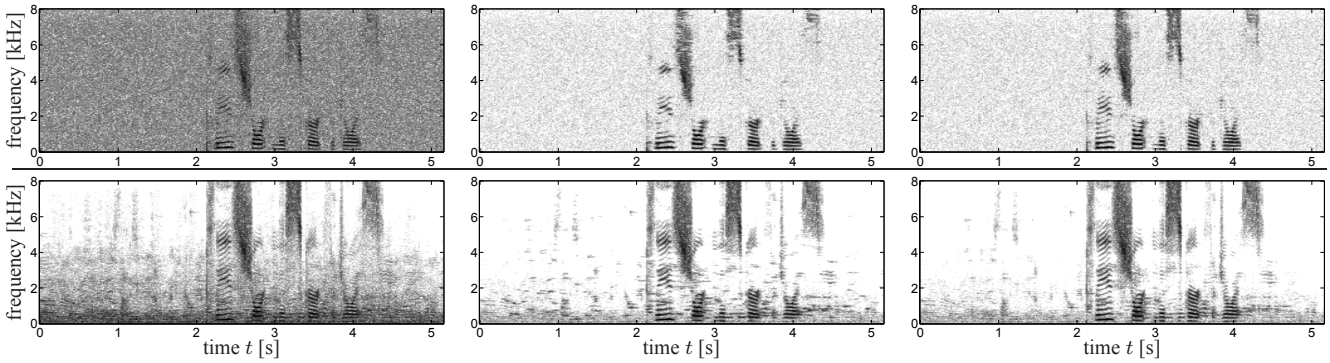

Fig. 3. Spectrograms of noisy signals at 0dB segmental SNR (left), of signals filtered using the *decision-directed* approach (center), and of signals filtered using the cepstral approach (right). The noises are stationary white Gaussian noise (top row) and babble noise (bottom row). The color coding is the same as in Figure 1 (top) which shows the spectrogram of the clean signal.

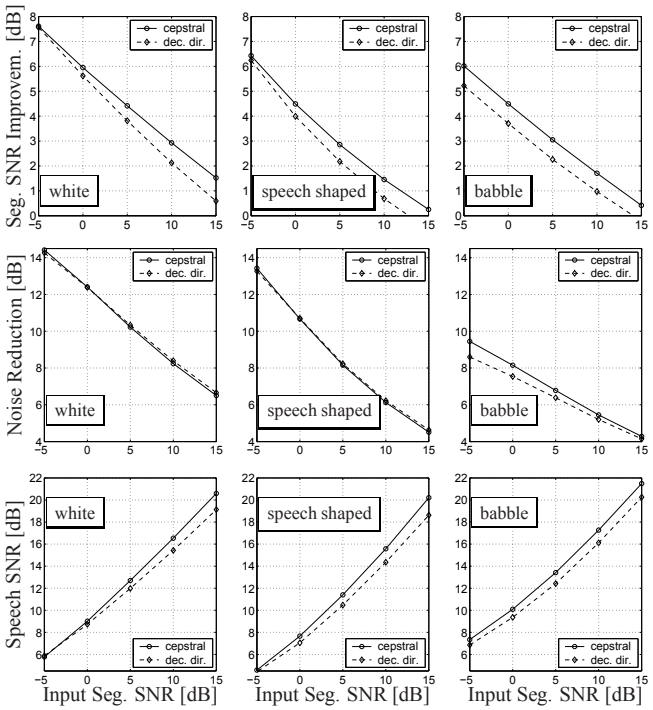

Fig. 2. Averages of segmental SNR improvement (top), noise reduction (middle), and speech SNR (bottom) for 320 TIMIT sentences and different noise types. The noises are white stationary Gaussian noise, speech shaped noise, and babble noise.

in the cepstral domain is superior to a temporal smoothing in the frequency domain. In the cepstral domain we can exploit *a priori* knowledge about speech production and thus selectively smooth the coefficients that most likely represent noise and those that represent speech. The proposed estimator consistently outperforms the well known *decision-directed* approach for *a priori* signal-to-noise ratio estimation in terms of output segmental signal-to-noise ratio, spectral distortion and noise reduction in non–stationary noise. Informal listening shows that the proposed estimator yields a clearer speech signal and, especially in non-stationary noise environments, a more natural sounding residual noise without musical noise.

6. REFERENCES

[1] O. Cappé, "Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor," *IEEE TSAP*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

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

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

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

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

[6] R. J. McAulay and M. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE TASSP*, vol. 28, no. 2, pp. 137–145, 1980.

[7] Y. Ephraim and I. Cohen, "Recent advancements in speech enhancement," in *The Electrical Engineering Handbook*, R.C. Dorf, Ed. CRC Press, 2006.

[8] Y. Ephraim and M. Rahim, "On second-order statistics and linear estimation of cepstral coefficients," *IEEE TSAP*, vol. 7, no. 2, pp. 162–176, Mar. 1999.

[9] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE TSAP*, vol. 9, no. 5, pp. 504–512, 2001.

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

[11] 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.