A PRIORI SNR COMPUTATION FOR SPEECH ENHANCEMENT BASED ON CEPSTRAL ENVELOPE ESTIMATION

Samy Elshamy, Nilesh Madhu†, Wouter Tirry° and Tim Fingscheidt**

*Institute for Communications Technology, Technische Universität Braunschweig
Schleinitzstr. 22, D–38106 Braunschweig, Germany
†Internet Technology and Data Science Lab, Universiteit Gent - imec, 9052 Gent, Belgium
°NXP Software, Interleuvenlaan 80, B–3001 Leuven, Belgium
{s.elshamy,t.fingscheidt}@tu-bs.de, nilesh.madhu@ugent.be, wouter.tirry@nxp.com

ABSTRACT

In this contribution we present our latest investigations and analysis on a novel *a priori* SNR estimator for speech enhancement applications. It is based on a clean spectral envelope estimation with a deep neural network (DNN) in the cepstral domain. Furthermore, by integrating our cepstral excitation manipulation (CEM) approach into this framework, we obtain not only a smooth and natural background noise experience, but also achieve noise reduction between harmonics which is not possible with low-order models. We investigate the performance of the proposed approach in conjunction with three different spectral weighting rules and show improvement of more than 3.5 dB noise attenuation vs. the well-known decision-directed (DD) approach without a significant trade-off in speech distortion.

Index Terms— a priori SNR, speech enhancement, cepstrum

1. INTRODUCTION

The broad field of speech enhancement comprises various applications that aim to facilitate the communication between human beings. Among them we find speech presence probability estimation, voice activity detection, and, e.g., noise reduction. The latter often uses a real-valued spectral weighting rule [1] in the frequency domain for a bin-wise noise suppression of a noisy microphone signal's amplitudes. These weighting rules are usually a function of the *a priori* signal-to-noise ratio (SNR) and oftentimes also of the *a posteriori* SNR.

The well-known decision-directed (DD) approach [2] defines an *a priori* SNR estimate that depends both on the past *a priori* SNR and the *a posteriori* SNR to obtain the estimate. Although the DD technique suffers from its incapability to track sudden changes of the true SNR, it is still regarded as classical state of the art.

Among the numerous more recent publications that investigate different *a priori* SNR estimation approaches [3, 4, 5, 6, 7], a generalized version of the DD approach has been proposed recently by Chinaev and Haeb-Umbach [8]. The method operates in a generalized spectral domain instead of the power domain. The authors show improved performance for high global SNR conditions for the generalized approach, while the original method, operating in the power domain, shows optimal behavior in low-SNR conditions.

Stahl and Mowlaee introduced a harmonic signal model for *a priori* SNR estimation in [9]. The model allows to interpolate between frequency bins and thus to smooth the *a priori* SNR according to harmonic trajectories. Thereby, the authors show improved noise attenuation capability without introducing additional speech distortion compared to the DD approach.

Furthermore, the incorporation of other models has been investigated in the recent past [10, 11, 12, 13], showing some improvement over the DD approach.

Very recently we proposed a novel *a priori* SNR estimator [12] based on cepstral excitation manipulation (CEM), which exploits the human speech production model. Its core features are the improvement of noise attenuation between harmonics and also the preservation of weak harmonic structures. Therein, we could show a more balanced and thus enhanced performance over the DD approach and also over two further, more recent *a priori* SNR estimators [4, 5]. Accordingly, both the DD and the CEM *a priori* SNR estimator serve as baselines for *this* work. Most recently we proposed a cepstral envelope estimation (CEE) approach [13] that nicely complements the CEM approach by not only enhancing the excitation signal but also the envelope. We described in detail how the proposed envelope estimator has been distilled from various investigated approaches.

In this paper we briefly revisit our findings from [13] and investigate the performance of the CEE approach for *a priori* SNR estimation alone, and in conjunction with CEM. We evaluate the estimators in a speech enhancement task together with three different weighting rules.

This contribution is structured as follows: In Section 2 we briefly introduce the signal model along with the CEE technique and provide insight into our investigations on the different methods. This is followed by a short introduction of the speech enhancement framework and the three weighting rules in Section 3. Subsequently, the experimental setup and the evaluation of our results is presented in Section 4. We finally conclude the paper in Section 5.

2. CEPSTRAL ENVELOPE ESTIMATION (CEE)

The microphone signal $y(n)$ is modeled as the superposition of the time-domain speech signal $s(n)$ and the noise signal $d(n)$ as $y(n) = s(n) + d(n)$, with n as discrete-time sample index. The frequency-domain entities are obtained by applying a K-point discrete Fourier transform as $Y_\ell(k) = S_\ell(k) + D_\ell(k)$, where ℓ represents the frame index and $0 \leq k \leq K-1$ the frequency bin index. Furthermore, we assume that both signals, noise and speech, have zero mean and that they are statistically independent.

The basic idea of our approach (see Figure 1) is to split a preliminary denoised microphone signal $\bar{Y}_\ell(k)$ into its envelope (Figure 1,

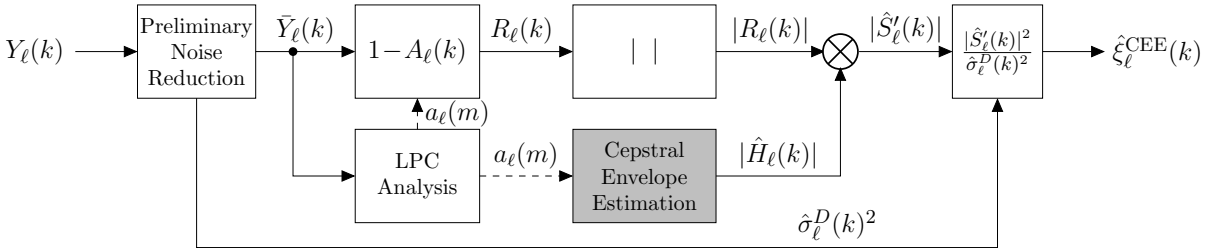

Fig. 1. Block diagram of the **proposed *a priori* SNR estimator** based on **cepstral envelope estimation (CEE)**.

LPC analysis, lower path) and its excitation $R_\ell(k)$ by LPC analysis. The denoised envelope is subsequently replaced by a clean envelope estimate $|\hat{H}_\ell(k)|$ and mixed with the excitation signal. It is used further with the noise power estimate $\hat{\sigma}_\ell^D(k)^2$ from the preliminary noise reduction to calculate the *a priori* SNR $\hat{\xi}_\ell^{\mathrm{CEE}}(k)$. The estimation is done in the cepstral domain by converting (Figure 2, feature conversion block) the $N = 10$ LPC coefficients to $N + 1 = 11$ cepstral coefficients using [14] as

$$c_\ell^H(m) = a_\ell(m) + \frac{1}{m}\sum_{\mu=1}^{m-1}\Big[(m-\mu)\cdot a_\ell(\mu)\cdot c_\ell^H(m-\mu)\Big] \quad (1)$$

for $1 \le m \le N$ and

$$c_\ell^H(m=0) = 0 = \log(P_p = 1) \quad (2)$$

for $m = 0$. We set the prediction error power P_p to a fixed value to obtain envelopes with a comparable energy level. This allows us to work with N coefficients only, as the first coefficient has the same value (zero) for all vectors. After estimating the clean envelope representing coefficients $c_\ell^{\hat{H}}(m)$ (see Figure 2, bold face for vector notation) we convert them back to LPC coefficients with [14]

$$\hat{a}_\ell(m) = c_\ell^{\hat{H}}(m) - \frac{1}{m}\sum_{\mu=0}^{m-1}\Big[(m-\mu)\cdot c_\ell^{\hat{H}}(m-\mu)\cdot \hat{a}_\ell(\mu)\Big] \quad (3)$$

for $1 \le m \le N$. The spectral representation $|\hat{H}_\ell(k)|$ is obtained from

$$|\hat{H}_\ell(k)| = \frac{1}{|1-\hat{A}_\ell(k)|}, \quad (4)$$

where $\hat{A}_\ell(k)$ is calculated by applying a K-point DFT to the zero-padded LPC coefficients $\hat{a}_\ell(m)$. This is done in the spectral conversion block in Figure 2.

We have evaluated and optimized different approaches for the cepstral envelope estimation task in [13]. We started with a classic hidden Markov model (HMM) with Gaussian mixture models (GMMs) as acoustic backend (GMM-HMM) where the hidden states represent clean, and the observations denoised envelopes. We found out by means of an oracle experiment that a codebook size of $64 + 1 = 65$ is sufficient. The codebook entries are obtained by using the Linde-Buzo-Gray [15] algorithm and the extra entry is exclusively representing non-speech envelopes. Best posterior state probability accuracy was obtained by choosing 16 modes for the GMMs. Furthermore, we investigated maximum a posteriori (MAP) and also minimum mean-square error (MMSE) estimation, resulting in superior performance of the latter in our noise reduction task.

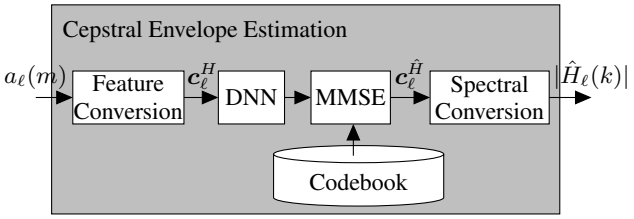

Fig. 2. Block diagram of the **preferred cepstral envelope estimation (CEE) method** using a classification DNN together with MMSE estimation.

Hence, we fixed the number of parameters and investigated the replacement of the GMMs by a classification deep neural network (DNN). We trained differently configured networks with up to six hidden layers, making sure that the aberration of parameters is always less than 2% by adjusting the number of nodes per layer, accordingly. Based on the best state posterior probability accuracy on speech active frames we found a network with six hidden layers, 58 nodes per layer, and sigmoid activation function to be optimal for classification. Thereby, the overall accuracy could be increased by 10 % absolute on the development set compared to the GMM-HMM approach. However, the incorporation of the DNN into the HMM yielded only comparable performance of the DNN-HMM vs. the GMM-HMM. Subsequently, we replaced the whole HMM structure by the classification DNN and could further improve the performance, now being able to fully benefit from the additional 10 % accuracy on the development set.

In Figure 2 we depict the processing structure of our favored estimator and refer to this method as CEE throughout the remainder of this paper. We have also investigated the performance of a DNN trained in regression mode, to directly estimate clean envelope-representing coefficients from the denoised observation. An optimal configuration was found for six hidden layers, 58 nodes, and also sigmoid activation function, but the performance in our noise reduction task was imbalanced, showing some detriments in the low-SNR conditions. So our proposal is to use the aforementioned classification DNN with subsequent codebook-supported MMSE estimation, as shown in Figure 2.

3. SPEECH ENHANCEMENT FRAMEWORK

Later evaluation of the *a priori* SNR estimators will be conducted in a speech enhancement framework consisting of a minimum statistics

noise power estimator [16], the *a priori* SNR estimator under test, and a spectral weighting rule to obtain the enhanced speech signal as

$$\hat{S}_\ell(k) = G_\ell(k) \cdot Y_\ell(k). \tag{5}$$

The spectral weighting rules $G_\ell(k) = f(\hat{\xi}_\ell(k), \gamma_\ell(k))$ are the minimum mean square error log-spectral amplitude (MMSE-LSA) estimator [17], the Wiener filter (WF) [18], and the super-Gaussian joint maximum a posteriori (SG-jMAP) estimator [19]. An *a posteriori* SNR

$$\gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma_\ell(k)^2} \tag{6}$$

is required for the MMSE-LSA and the SG-jMAP spectral weighting rule, and also for the DD *a priori* SNR baseline estimator according to [2]

$$\hat{\xi}_\ell^{\mathrm{DD}}(k) = (1-\beta_{\mathrm{DD}}) \cdot \max\{\hat{\gamma}_\ell(k)-1, 0\} + \beta_{\mathrm{DD}} \frac{|\hat{S}_{\ell-1}(k)|^2}{\hat{\sigma}^D_{\ell-1}(k)^2}. \tag{7}$$

The $\mathrm{CEM_{SI}}$ baseline [12] is refining a clean speech amplitude estimate in an instantaneous fashion by modifying the excitation signal based on pre-trained templates. Is is subsequently used with the noise power estimate from the preliminary noise reduction to obtain $\hat{\xi}_\ell^{\mathrm{CEM}}(k)$.

If our proposed CEE-based *a priori* SNR estimator according to Figure 2 is employed, the minimum statistics noise power estimator is executed as part of the preliminary noise reduction, which internally also contains a DD *a priori* SNR estimation and an MMSE-LSA weighting rule. The rest of our CEE *a priori* SNR estimator is shown in Figure 1 with Figure 2 as discussed.

We will also investigate an approach that concatenates the CEE *a priori* SNR estimator with the CEM technique from [12]. For that purpose the preliminary noise reduction as it is required for the CEM approach consists of the complete Figure 1, including a subsequent MMSE-LSA spectral weighting rule which is applied to the microphone signal. The further processing according to [12] provides then the final *a priori* SNR estimate $\hat{\xi}_\ell^{\mathrm{CEE}\rightarrow\mathrm{CEM}}(k)$ — for details please be referred to [12, 13]. Note that in this serial approach the noise power estimate that is used throughout is the one of the aforementioned preliminary noise reduction in the CEE approach, see Figure 1.

4. EXPERIMENTAL EVALUATION

4.1. Experimental Setup

The DD estimator, wherever it is employed, is tuned with optimal parameters[1] [20] for each weighting rule. The DD estimator as part of the preliminary noise reduction in Figure 1 uses parameters as shown[1] for MMSE-LSA, since this is the weighting rule of the preliminary noise reduction. We work with a sample rate of 8 kHz, a frame size of $K = 256$ samples with a frame shift of 50 %. For analysis and overlap-add synthesis we utilize a periodic square root Hann window. The training and development sets for the investigations in Section 2 are taken from the TIMIT database [21]. The clean speech is mixed at six different SNR conditions ranging from -5 dB to 20 dB in 5 dB steps together with disjoint portions of 53 noise files taken from the ETSI [22] and the QUT [23] noise databases. For a test set with unseen noise files we use four files[2] from the ETSI database exclusively which are not used for training or development. However, similar noise types have been used also for the training process. Signal levels are adjusted according to ITU-T P.56 [24] and subsequently superimposed.

[1] Optimal parameters for the DD estimator and each weighting rule:

MMSE-LSA: $\beta_{\mathrm{DD}} = 0.975, \xi_{\mathrm{min}} = -15$ dB
SG-jMAP: $\beta_{\mathrm{DD}} = 0.993, \xi_{\mathrm{min}} = -14$ dB
WF: $\beta_{\mathrm{DD}} = 0.99, \xi_{\mathrm{min}} = -14$ dB.

4.2. Quality Measures

To evaluate the estimators in a speech enhancement task we use the white-box approach [25] which allows us to evaluate the *filtered* speech component $\tilde{s}(n)$ and the *filtered* noise component $\tilde{d}(n)$ of the enhanced signal $\hat{s}(n)$, separately. This is done by applying the final gain function $G_\ell(k)$ not only to the microphone signal $Y_\ell(k)$ in order to obtain the enhanced speech $\hat{S}_\ell(k)$, but also to the separate speech component $S_\ell(k)$ and noise component $D_\ell(k)$, followed by inverse DFT and overlap-add synthesis. As objective measures we use the segmental speech-to-speech-distortion ratio (SSDR) [26] which is calculated as

$$\mathrm{SSDR_{seg}} = \frac{1}{|\mathcal{L}_1|} \sum_{\ell \in \mathcal{L}_1} \mathrm{SSDR}(\ell) \tag{8}$$

with $\mathcal{L}_1$ being the set of speech active frames,

$$\mathrm{SSDR}(\ell) = \max\left\{\min\left\{\mathrm{SSDR}'(\ell), R_{\mathrm{max}}\right\}, R_{\mathrm{min}}\right\} \tag{9}$$

where R_{max} and R_{min} limit the values to 30 dB and -10 dB, respectively. The frame-wise ratio is obtained as

$$\mathrm{SSDR}'(\ell) = 10\log_{10}\left[\frac{\sum_{\nu=0}^{N-1} s(\nu+\ell N)^2}{\sum_{\nu=0}^{N-1} e(\nu+\ell N)^2}\right] \tag{10}$$

with the error signal being

$$e(\nu+\ell N) = \tilde{s}(\nu+\ell N+\Delta) - s(\nu+\ell N). \tag{11}$$

The term Δ is accounting for potential processing delay and ℓ is depicting a segment of length $N = 256$ samples. A high $\mathrm{SSDR_{seg}}$ indicates a strong similarity of the speech component with respect to the clean reference signal.

To account for the noise attenuation we additionally report the ΔSNR which is a global measure and calculated as

$$\Delta\mathrm{SNR} = \mathrm{SNR_{out}} - \mathrm{SNR_{in}}, \tag{12}$$

where $\mathrm{SNR_{out}}$ is the SNR of the *filtered* components $\tilde{s}(n)$ and $\tilde{d}(n)$, and $\mathrm{SNR_{in}}$ is the SNR of the unprocessed components $s(n)$ and $d(n)$. Both SNRs are measured in line with ITU-T P.56 [24] where for the speech signals only speech active portions are considered. The ΔSNR gives information on the global SNR improvement by considering both components simultaneously.

4.3. Discussion

In Figure 3 we depict the results for the different *a priori* SNR estimators under test with the three weighting rules MMSE-LSA, SG-jMAP, and WF. We plot the $\mathrm{SSDR_{seg}}$ vs. the ΔSNR and each marker represents one SNR condition, where -5 dB is at the bottom and

[2] `Fullsize_Car1_80Kmh`, `Outside_Traffic_Crossroads`, `Pub_Noise_Binaural_V2`, `Work_Noise_Office_Callcenter`

20 dB is at the top in steps of 5 dB. In general, the WF seems to achieve the highest ΔSNR for each approach, while the speech component quality suffers, which is quite obvious especially for the **DD** approach. The most recent weighting rule SG-jMAP provides best speech component quality among the analyzed estimators, however, offering less noise attenuation as a typical trade-off. The MMSE-LSA estimator settles somewhere in between showing a balanced performance of the *a priori* SNR estimators.

The **CEE** *a priori* SNR estimator (solid orange line, asterisk markers) outperforms the **DD** baseline (solid yellow line, plus markers) by about 2 dB ΔSNR for the MMSE-LSA and SG-jMAP weighting rules in the -5 dB SNR condition. Using the SG-jMAP weighting rule, **CEE** exceeds the performance of the **DD** approach also consistently in terms of SSDR_{seg}. When used with the WF, only the important low-SNR conditions show reasonable performance gain for **CEE**.

The recently published $\textbf{CEM}_{\textbf{SI}}$ baseline [12] (solid green line, square markers) exceeds clearly the **DD** baseline, and also **CEE** when operating alone, in terms of noise attenuation for every weighting rule owing to its ability to effectively reduce noise between the harmonics. The highest performance gain obtained over **DD** amounts to more than 3 dB ΔSNR for $\textbf{CEM}_{\textbf{SI}}$ when either using MMSE-LSA or SG-jMAP. This gain can be further enlarged by concatenating (symbol →) the **CEE** approach with the $\textbf{CEM}_{\textbf{SI}}$ baseline (dashed green line, triangle markers). Thereby, we obtain a ΔSNR that is higher by more than 3.5 dB compared to the DD approach for the MMSE-LSA weighting rule.

The investigated approaches (**CEE** and $\textbf{CEE}\rightarrow\textbf{CEM}_{\textbf{SI}}$) appear to be more robust compared to **DD** as the speech component quality remains comparable for the respective approach when exchanging MMSE-LSA by the WF, while simultaneously also showing higher ΔSNR. Here, the **DD** approach experiences quite some negative effects on the speech component quality due to the increase of noise attenuation. Hence, we recommend the serial approach $\textbf{CEE}\rightarrow\textbf{CEM}_{\textbf{SI}}$ as it offers robustness across various weighting rules while being able to mitigate the classical trade-off between speech component quality and noise attenuation. Informal expert analysis and listening tests[3] have shown that the approach results in a very smooth and also natural sound of the remaining low-level background noise. This is an advantage over both baselines, **DD** and also $\textbf{CEM}_{\textbf{SI}}$.

5. CONCLUSIONS

We investigated the performance of a novel *a priori* SNR estimator in a noise reduction environment with three different spectral weighting rules. We could show that the proposed serial estimator, which uses cepstral envelope estimation (CEE) in conjunction with cepstral excitation manipulation (CEM), exceeds CEM consistently by up to 0.4 dB ΔSNR, even in non-stationary noise, and improves by more than 3.5 dB vs. the decision-directed (DD) approach. At the same time, no significant trade-off in speech distortion is observed.

6. REFERENCES

[1] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*, Springer, Berlin, 2008.

[3] Audio samples can be found under:
https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2018-iwaenc-cee/

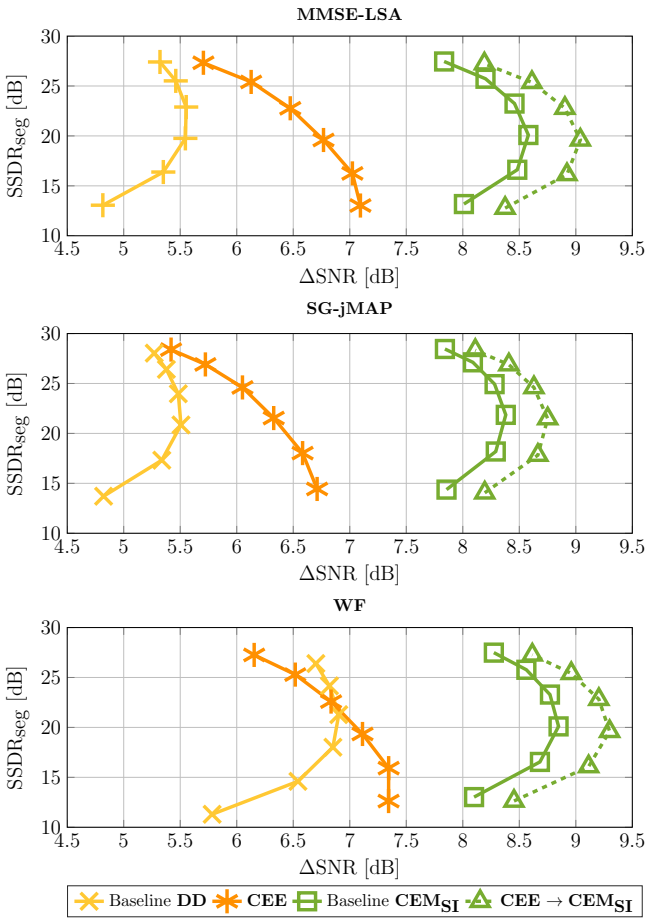

Fig. 3. Evaluation of SSDR_{seg} and ΔSNR for the *a priori* SNR estimators under test in non-stationary and unseen noises, together with three different spectral weighting rules.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[3] I. Cohen, "Relaxed Statistical Model for Speech Enhancement and A Priori SNR Estimation," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 870–881, Sep 2005.

[4] C. Plapous, C. Marro, and P. Scalart, "Improved Signal-to-Noise Ratio Estimation for Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 2098–2108, Nov. 2006.

[5] C. Breithaupt, T. Gerkmann, and R. Martin, "A Novel A Priori SNR Estimation Approach Based on Selective Cepstro-Temporal Smoothing," in *Proc. of ICASSP*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.

[6] S. Suhadi, C. Last, and T. Fingscheidt, "A Data-Driven Approach to A Priori SNR Estimation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[7] H. S. Shin, T. Fingscheidt, and H.-G. Kang, "A Priori SNR Estimation Using Air and Bone-Conduction Microphones,"

IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 2015–2025, Nov. 2015.

[8] A. Chinaev and R. Haeb-Umbach, "A Priori SNR Estimation Using a Generalized Decision Directed Approach," in *Proc. of Interspeech*, San Francisco, CA, USA, Sept. 2016, pp. 3758–3762.

[9] J. Stahl and P. Mowlaee, "A Simple and Effective Framework for A Priori SNR Estimation," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 5644–5648.

[10] S. Elshamy, N. Madhu, W. J. Tirry, and T. Fingscheidt, "An Iterative Speech Model-Based A Priori SNR Estimator," in *Proc. of Interspeech*, Dresden, Germany, Sept. 2015, pp. 1740–1744.

[11] A. Chinaev, J. Heitkaemper, and R. Haeb-Umbach, "A Priori SNR Estimation Using Weibull Mixture Model," in *Proc. of ITG Conference on Speech Communication*, Paderborn, Germany, Oct. 2016, pp. 297–301.

[12] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous A Priori SNR Estimation by Cepstral Excitation Manipulation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[13] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope," *Submitted to IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

[14] P. E. Papamichalis, *Practical Approaches to Speech Coding*, Prentice Hall, Inc., Upper Saddle River, NJ, USA, 1987.

[15] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," in *IEEE Transactions on Communications*, Jan. 1980, vol. 28, pp. 84–95.

[16] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, Jul. 2001.

[17] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[18] P. Scalart and J. V. Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation," in *Proc. of ICASSP*, Atlanta, GA, USA, May 1996, pp. 629–632.

[19] T. Lotter and P. Vary, "Speech Enhancement by MAP Spectral Amplitude Estimation Using a Super-Gaussian Speech Model," *EURASIP Journal on Applied Signal Processing*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[20] H. Yu, *Post-Filter Optimization for Multichannel Automotive Speech Enhancement*, Ph.D. thesis, Institute for Communications Technology, Technische Universität Braunschweig, 2013.

[21] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT Acoustic-Phonetic Continuous Speech Corpus," Linguistic Data Consortium (LDC), 1993.

[22] ETSI, *EG 202 396-1: Speech Processing, Transmission and Quality Aspects (STQ); Speech Quality Performance in the Presence of Background Noise; Part 1: Background Noise Simulation Technique and Background Noise Database*, European Telecommunications Standards Institute, Sept. 2008.

[23] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT Corpus for the Evaluation of Voice Activity Detection Algorithms," in *Proc. of Interspeech*, Makuhari, Japan, Sept. 2010, pp. 3110–3113.

[24] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*, International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.

[25] S. Gustafsson, R. Martin, and P. Vary, "On the Optimization of Speech Enhancement Systems Using Instrumental Measures," in *Proc. of Workshop on Quality Assessment in Speech, Audio, and Image Communication*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[26] T. Fingscheidt, S. Suhadi, and S. Stan, "Environment-Optimized Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 4, pp. 825–834, May 2008.