Instantaneous *A Priori* SNR Estimation by Cepstral Excitation Manipulation

Samy Elshamy, Nilesh Madhu, Wouter Tirry, and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—As the *a priori* signal-to-noise ratio (SNR) contains crucial information about a signal's mixture of speech and noise, its estimation is subject to steady research. In this paper, we introduce a novel *a priori* SNR estimator based on synthesizing an idealized excitation signal in the cepstral domain. Our approach utilizes a source-filter decomposition in combination with a cepstral excitation manipulation in order to recreate an idealized excitation, which is subsequently shaped by an immanent envelope. In contrast to the well-known decision-directed approach by Ephraim and Malah, an *instantaneous* estimate is obtained, which is less prone to sudden acoustic environmental changes and musical noise. Additionally, the proposed estimator is able to preserve weak harmonic structures resulting in a spectrum that is more full-bodied. We present both a speaker-independent and a speaker-dependent variant of the new *a priori* SNR estimator, both showing more than 2 dB ΔSNR improvement versus state of the art, without any significant increase in speech distortion.**

***Index Terms*—*A priori* SNR, speech enhancement.**

I. Introduction

A priori SNR estimation has long been an important topic in speech enhancement. Having only a single mixture at hand most likely impedes enhancement tasks since no knowledge about the individual components of the observed mixture is available. Consequently, the need to estimate an *a priori* SNR arises and has been subject to research in several publications [1]–[7]. Algorithms such as voice activity detection [8], speech presence probability estimation [9] and, most importantly, spectral weighting rules for noise reduction algorithms [1], [2], [10], [11] take great profit from reliable *a priori* SNR estimates.

The decision-directed (DD) approach to estimate the *a priori* SNR by Ephraim and Malah [1] has been published along with a spectral amplitude estimator for noise reduction and is basically a weighted sum of two components. The first component is depicting the ratio of the previous frame's squared clean speech amplitude estimate and the provided noise power estimate also taken from the previous frame. The second component is an instantaneous estimate derived from the current frame's *a posteriori* SNR. The weights of both components sum up to unity and as proposed in [1] the weight for the first component is chosen close to unity. The approach has been thoroughly analyzed in [12] and [13], where the analysis of Cappé [12] has shown, that the DD *a priori* SNR estimate follows the *a posteriori* SNR[1] with a delay of one frame.

Cohen proposed a non-causal estimator which buffers a few frames and thus is capable of differentiating between onsets of speech and bursts of noise allowing less musical tones and distortion of transient speech regions [3]. The approach is also less sensitive to changes in the underlying speech model compared to the DD technique. In practice all this comes at the price of some frames of delay.

Breithaupt *et al.* proposed an estimator [5] which employs a quefrency-selective smoothing of a maximum likelihood (ML) speech power spectral density derived from the *a posteriori* SNR and the noise power estimate. In the cepstral domain the coefficients corresponding to the excitation and the envelope are smoothed differently. As a result they obtain an *a priori* SNR estimate that yields better results in a noise suppression framework than the one in [1] w.r.t. spectral distortion and musical tones, especially in non-stationary environments. However, the clean excitation is not directly modeled, still leaving potential for improvement.

A data-driven approach based on the DD formula has been published in [6]. The two components of the weighted sum are both input to two different neural networks, discriminating speech active and inactive frames, with the ideal *a priori* SNR as a target during the training process. In a practical system both networks are evaluated and a linear combination of the provided outputs yields the final *a priori* SNR estimate. The authors are able to show a reduction of speech distortion during speech onsets while maintaining a high noise attenuation during speech pause. As the training process requires noise signals the approach is not entirely independent of the noise type.

Our latest work [7] shows that a simple Gaussian mixture model (GMM), representing clean speech spectral amplitudes, is able to provide a ML clean speech amplitude estimate when preliminary denoising is applied to the observation and subsequently the GMM is evaluated. The provided estimate is then used as numerator for an intermediate *a priori* SNR estima-

Manuscript received October 20, 2016; revised March 15, 2017 and April 28, 2017; accepted April 30, 2017. Date of publication May 9, 2017; date of current version June 12, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Mads Græsbøll Christensen. *(Corresponding author: Tim Fingscheidt.)*

S. Elshamy and T. Fingscheidt are with the Institute for Communications Technology, Technische Universität Braunschweig, 38106 Braunschweig, Germany (e-mail: s.elshamy@tu-bs.de; t.fingscheidt@tu-bs.de).

N. Madhu and W. Tirry are with the NXP Software, 3001 Leuven, Belgium (e-mail: nilesh.madhu@nxp.com; wouter.tirry@nxp.com).

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/TASLP.2017.2702385

[1] Please note that the term "*a posteriori* SNR" in [12] differs from its use in mainstream literature as introduced in [1].

tion and continuously improved by repeated filtering and re-estimation.

Motivated by Cappé's observation, Plapous *et al.* propose a so-called two-step noise reduction (TSNR) technique [14] which is able to compensate for the one-frame delay. It is used as a preliminary noise reduction for their harmonic regeneration noise reduction (HRNR) introduced in [4]. The HRNR approach employs an improved *a priori* SNR estimator which applies a non-linear function to an enhanced time-domain signal in order to restore lost harmonics in the spectrum. The enhanced signal is subsequently mixed with the preliminary denoised signal, according to the calculated gains of the TSNR, and then used as numerator for the *a priori* SNR estimate. The applied non-linearity produces an unnatural harmonic and leads to audible artifacts in certain low-frequency noise types.

A recent analysis [15] deals with the over- and underestimation of estimated *a priori* SNR. The authors propose to use a correction term based on an empirically obtained distribution of the true bias in dependency of the *a priori* and *a posteriori* SNRs. The distribution is then subject to a vector quantizer which is later on used to estimate the bias on real data to compensate for the aberration. They show how to improve the DD and also the TSNR approach and additionally state that the proposed method could be used together with any spectral weighting rule.

In this paper we introduce a novel approach that consequently exploits the a priori knowledge that comes along with a model-based approach, while staying fully independent of noise types. The proposed method for *instantaneous a priori* SNR estimation without the need for lookahead is based on the *source-filter* model representing human speech production and also embraces the convenience a cepstral representation offers in terms of pitch estimation and cosine synthesis.

Furthermore, we also address a problem known to occur with approaches that model solely the spectral shape as they typically lack the fine structure of the speech spectrum and thus are not able to suppress noise between the harmonics [16].

In a first stage we employ a preliminary noise reduction driven by a noise power estimator such as [17]–[19], suitable state-of-the-art *a priori* SNR estimation [1], [5]–[7], and a weighting rule of choice, e.g., [1], [2], [10], [11]. In a second stage we utilize linear predictive coding (LPC) analysis to decompose the preliminary denoised signal into its spectral envelope and excitation followed by a transformation of the excitation signal to the cepstral domain. Subsequently, we detect the pitch, and, as a core of our approach, we synthesize an *idealized excitation* which is shaped by the spectral envelope of the preliminary denoised signal. The resulting spectrum is finally used as clean speech amplitude estimate for an instantaneous *a priori* SNR numerator. We then extend our approach to using two variants of excitation manipulation (synthetic and template-based) and show improvement of the template-based over the purely synthetically created excitation. Finally, we investigate the potential of a speaker-dependent (vs. a speaker-independent) setup of our estimator.

The structure of this paper is as follows: In Section II we introduce our mathematical notations and some baseline estimators, which serve as reference in the evaluation. Next, we present our cepstral processing methodology in Section III followed by the two proposed manipulation schemes in Section IV. In Section V we present the experimental results and discussion separately, and conclude the paper in Section VI.

II. Notations and Baselines

We assume an additive model for the microphone signal $y(n)$ in the time domain as

$$y(n) = s(n) + d(n), \tag{1}$$

with $s(n)$ being the clean speech signal we are interested in, and $d(n)$ being the noise signal we aim to suppress. The discrete-time sample index is n. The corresponding frequency-domain representation by applying the discrete Fourier transform (DFT) is

$$Y_\ell(k) = S_\ell(k) + D_\ell(k), \tag{2}$$

with frame index ℓ and frequency bin index k being restricted by the DFT size K to $0 \leq k \leq K-1$. Furthermore, as most approaches do, we assume that the speech and noise signals are zero-mean and statistically independent of one another.

A. Noise Power Estimation

An estimate of the noise power, which is denoted by $\hat{\sigma}_\ell^D(k)^2$, is required for noise reduction and can be obtained by several algorithms which have been published in the past. Among those is the minimum statistics (MS) approach [17], which is a commonly utilized estimator with good performance in stationary and non-stationary environments. Besides, there are further estimators such as the minima-controlled estimator proposed in [18], or estimators based on the minimum mean-square error (MMSE), e.g., [20].

B. Spectral Weighting Rules

The desired clean speech spectral estimate is generally obtained by applying a real-valued gain function, also referred to as spectral weighting rule $G_\ell(k)$, to the observed signal as follows

$$\hat{S}_\ell(k) = Y_\ell(k) \cdot G_\ell(k). \tag{3}$$

Thereby, the noisy phase is usually maintained as motivated in [1], [21], [22], although some more recent publications support phase-aware speech enhancement [23]–[25]. As the potential of amplitude-based speech enhancement seems not yet exhausted, we feel comfortable to focus on these in the following.

Amongst the most famous weighting rules utilized to calculate gain functions $G_\ell(k)$, we find the well-known Wiener filter (WF) [10], the MMSE short-time spectral amplitude estimator (MMSE-STSA) [1], the MMSE log-spectral amplitude estimator (MMSE-LSA) [2], and the super-Gaussian joint maximum a posteriori (SG-jMAP) estimator [11]. The aforementioned various frequency bin-selective gain functions $G_\ell(k)$ are mostly (nonlinear) functions $f(\cdot)$ of the *a priori* SNR

$$\xi_\ell(k) = \frac{\sigma_\ell^S(k)^2}{\sigma_\ell^D(k)^2} \tag{4}$$

and partly also of the *a posteriori* SNR

$$\gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma_\ell^D(k)^2} \tag{5}$$

allowing us to compute $G_\ell(k)$ as:

$$G_\ell(k) = f\left(\xi_\ell(k), \gamma_\ell(k)\right). \tag{6}$$

Since both entities require quantities that are not available in practice they need to be estimated (or at least components of them). We denote estimated entities with a hat $\hat{(\cdot)}$ as accent.

C. A Priori SNR Estimation

In this section we briefly sketch three baseline approaches which will later serve to compare our approach against.

1) Decision-Directed (DD)*:* The historic breakthrough to estimate the *a priori* SNR is the already mentioned DD approach by Ephraim and Malah [1]. In summary, the DD formula narrows down to

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

with β_{DD} and $(1-\beta_{\mathrm{DD}})$ being the weights of both components as mentioned in the introduction. Subsequently, as proposed in [12], the *a priori* SNR estimate is lower-bounded to a certain $\xi_{\min}$ to avoid musical tones.

2) Selective Cepstro-Temporal Smoothing (CTS)*:* This method, proposed by Breithaupt *et al.* [5], is utilizing properties of the cepstral representation to obtain a more precise *a priori* SNR estimate. The core of this approach is an adaptive, first-order recursive smoothing of the cepstrum of the ML clean speech estimate $c_\ell^{\hat{S}^{\mathrm{ML}}}(m)$ according to

$$\bar{c}_\ell^{\hat{S}}(m) = \alpha_\ell(m) \cdot \bar{c}_{\ell-1}^{\hat{S}}(m) + (1-\alpha_\ell(m)) \cdot c_\ell^{\hat{S}^{\mathrm{ML}}}(m), \tag{8}$$

where $\bar{c}_\ell^{\hat{S}}(m)$ is the smoothed version of the cepstrum and $m \in \mathcal{M} = \{0, 1, \ldots, K-1\}$ is the cepstral bin index. The cepstrum of the ML clean speech estimate in this particular case is obtained as

$$\left(c_\ell^{\hat{S}^{\mathrm{ML}}}(m)\right)_{m=0}^{K-1} = \mathrm{IDFT}\left\{\left(\log|\hat{S}_\ell^{\mathrm{ML}}(k)|^2\right)_{k=0}^{K-1}\right\}, \tag{9}$$

with

$$|\hat{S}_\ell^{\mathrm{ML}}(k)|^2 = \hat{\sigma}_\ell^D(k)^2 \cdot \max\left\{\xi_\ell^{\mathrm{ML}}(k), \xi_{\min}^{\mathrm{ML}}\right\}. \tag{10}$$

The ML *a priori* SNR floor $\xi_{\min}^{\mathrm{ML}} > 0$ is a small number yielding numerical stability, while

$$\xi_\ell^{\mathrm{ML}}(k) = \gamma_\ell(k) - 1, \tag{11}$$

as shown in [5]. Parameter $\alpha_\ell(m)$ is not only time-variant, but also quefrency-selective. Cepstral coefficients with small indices controlling the shape of the spectral envelope are to be smoothed only slightly, whereas the higher-indexed cepstral coefficients are supposedly related to noise and thus heavily smoothed. An exception is made for bins related to the fundamental frequency as these are suggested to be smoothed even less than the envelope-related quefrencies. Therefore, this method relies on a cepstral pitch estimation. For the detailed smoothing scheme we refer to [5]. After a bias compensation required due to the smoothing in the logarithmic domain and inverse transformation the final *a priori* SNR estimate is obtained as

$$\hat{\xi}_\ell^{\mathrm{CTS}}(k) = \max\left\{\frac{|\hat{S}_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}, \xi_{\min}\right\}, \tag{12}$$

with

$$\left(|\hat{S}_\ell(k)|^2\right)_{k=0}^{K-1} = \exp\left(\kappa + \mathrm{DFT}\left\{\left(\bar{c}_\ell^{\hat{S}}(m)\right)_{m=0}^{K-1}\right\}\right), \tag{13}$$

and κ being a log-amplitude spectrum bias compensation. In our simulations an improved bias compensation term has been used as presented in [26]. The noise power estimation is not further restricted to any specific method. Finally, note that an instantaneous extension to CTS could be employed as being done in [27].

3) Harmonic Regeneration (HRNR)*:* The HRNR approach by Plapous *et al.* [4] is based on the DD estimator, but taking Cappé's observation into account to compensate for the one-frame delay of the *a priori* SNR, underlying some preliminary spectral weights $G_\ell^{\mathrm{DD}}(k)$. The authors employ a two-step noise reduction technique (TSNR) to accomplish the delay compensation. Therefore, they introduce a second gain function

$$G_\ell^{\mathrm{TSNR}}(k) = f(\hat{\xi}_\ell^{\mathrm{TSNR}}(k), \hat{\gamma}_\ell(k)) \tag{14}$$

with an updated *a priori* SNR

$$\hat{\xi}_\ell^{\mathrm{TSNR}}(k) = \frac{|Y_\ell(k) \cdot G_\ell^{\mathrm{DD}}(k)|^2}{\hat{\sigma}_\ell^D(k)^2} \tag{15}$$

being responsible for the actual compensation. A harmonic *spectral* regeneration method operates on the TSNR-enhanced signal $Y_\ell(k) \cdot G_\ell^{\mathrm{TSNR}}(k)$, applying a simple non-linear function in the time domain, here half-wave rectification, and thereby boosting the harmonics of voiced frames. After transformation the spectrum is depicted as $\check{S}_\ell(k)$, which is not directly used for clean speech estimation but for another *a priori* SNR estimate $\hat{\xi}_\ell^{\mathrm{HRNR}}(k)$. To obtain this estimate, $\check{S}_\ell(k)$ is mixed with the TSNR-enhanced signal according to the corresponding gain function as follows

$$\hat{\xi}_\ell^{\mathrm{HRNR}}(k) = \\ \frac{\alpha_\ell(k) \cdot |Y_\ell(k) \cdot G_\ell^{\mathrm{TSNR}}(k)|^2 + (1-\alpha_\ell(k)) \cdot |\check{S}_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2} \tag{16}$$

where the authors propose to use weights $\alpha_\ell(k) = G_\ell^{\mathrm{TSNR}}(k)$.

This constitutes the final *a priori* SNR estimate; again, the noise power estimator can be chosen from available literature for each of the proposed stages.

Throughout this paper we refer to a system that is composed of a noise power estimator (Section II-A), an *a priori* SNR estimation (Section II-C), and a spectral weighting rule (Section II-B) as either a common or a preliminary noise reduction.

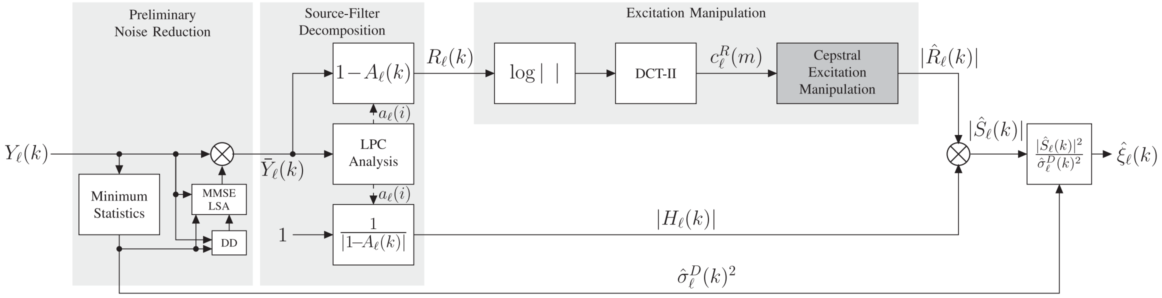

Fig. 1. Block diagram of the cepstral processing framework for our **proposed** *a priori* **SNR estimation**. The preliminary noise reduction consists of the MS noise power estimation algorithm, the DD *a priori* SNR estimation approach, and the MMSE-LSA spectral weighting rule.

III. New Cepstral Processing Framework for A Priori SNR Estimation

In this section we present the cepstral processing framework of our *a priori* SNR estimator, and provide some motivation for cepstral domain processing. Fig. 1 depicts its overall architecture.

A. Preliminary Noise Reduction

Similar to [4], we also employ a preliminary denoising stage before applying the actual approach. The motivation is to facilitate the extraction of required information for the proposed *a priori* SNR estimation. As it will rely on a pitch estimation, our approach benefits from the preliminary noise reduction rendering pitch estimation more robust, even in very low-SNR conditions. We target the preservation of harmonics which are often strongly attenuated, especially in adverse environments. In practice, this preliminary noise reduction is not limited to any specific components or approaches, but we propose to rely on a common noise reduction scheme being composed of some noise power estimation (e.g., minimum statistics [17]), the DD approach to *a priori* SNR estimation [1], and the MMSE-LSA spectral weighting rule [2] (referring to the left light gray block in Fig. 1).

B. Source-Filter Decomposition

Decomposing the preliminary denoised signal into its envelope and excitation (e.g., by LPC analysis) allows us to break down the enhancement task into two individual problems, thus enabling specific enhancement methods to be applied to each of the components, separately. However, in this paper we focus on the excitation only.

Our proposed method exploits knowledge about the process of human speech generation, especially of voiced speech. Therefore, it is important to have a reliable pitch estimation which is on the one hand supported by the preliminary denoising stage, and on the other hand by analyzing the excitation signal. For these reasons, we decompose the preliminary denoised signal $\bar{Y}_\ell(k)$ into the spectral excitation $R_\ell(k)$ and its spectral envelope $H_\ell(k)$, which is understood as the source and the filter, respectively. In each frame ℓ the spectral envelope $H_\ell(k)$ is obtained by first applying the K-point inverse discrete Fourier transform (IDFT) to the squared magnitude spectrum $|\bar{Y}_\ell(k)|^2$, resulting in the sequence of autocorrelation coefficients

$$\left(\varphi_\ell^{\bar{y},\bar{y}}(\nu)\right)_{\nu=0}^{K-1} = \text{IDFT}\left\{\left(|\bar{Y}_\ell(k)|^2\right)_{k=0}^{K-1}\right\}. \qquad (17)$$

The first $N+1<K$ elements $\varphi_\ell^{\bar{y},\bar{y}}(\nu),\ \nu \in \{0,1,\ldots,N\}$ are used to compute a set of N LPC coefficients $a_\ell(i)$, $i \in \{1,2,\ldots,N\}$ by the Levinson-Durbin recursion. The LP analysis filter in the DFT domain $(1-A_\ell(k))$ is then simply obtained by applying the K-point DFT to a sequence of the previously calculated N LPC coefficients, padded with $K-N-1$ zeros:

$$\left(A_\ell(k)\right)_{k=0}^{K-1} = \text{DFT}\left\{(0, a_\ell(1), \ldots, a_\ell(N), 0, \ldots, 0)\right\}. \quad (18)$$

The LP analysis filter is employed to process the preliminary denoised signal $\bar{Y}_\ell(k)$ to retrieve the respective residual signal as [28]:

$$R_\ell(k) = \bar{Y}_\ell(k) \cdot (1 - A_\ell(k)), \qquad (19)$$

while the spectral envelope is given by the inverse filter as

$$H_\ell(k) = \frac{1}{1-A_\ell(k)}. \qquad (20)$$

LPC analysis is an established method for source-filter decomposition [28]. An alternative to computing the envelope could have been simple liftering in the cepstral domain (i.e., taking only the lower part of the cepstrum), which, however, does not provide the exact same results as the Levinson-Durbin recursion of LPC analysis [29, Sec. 9.5.1].

Further investigations towards the processing framework as shown in Fig. 1 have also shown that a residual signal obtained via LPC analysis, being subsequently transformed into the cepstral domain, is better suited for later manipulation. This is further elaborated on in Section IV-B.

C. Cepstral Excitation Representation

Next, we obtain a cepstral representation of a signal by applying the discrete cosine transform of type II (DCT-II), but also an IFFT could have been chosen as in [5]. Additionally, we present some of its inherent, convenient properties we take advantage of. To further analyze the spectrum of the excitation signal in a first step we compute the cepstral coefficients upon the excitation signal's logarithmic magnitude spectrum as [30]

$$c_\ell^R(m) = \sum_{k=0}^{K-1} \log\left(|R_\ell(k)|\right) \cdot \cos\left[\pi m \left(k+0.5\right) \tfrac{1}{K}\right] \qquad (21)$$

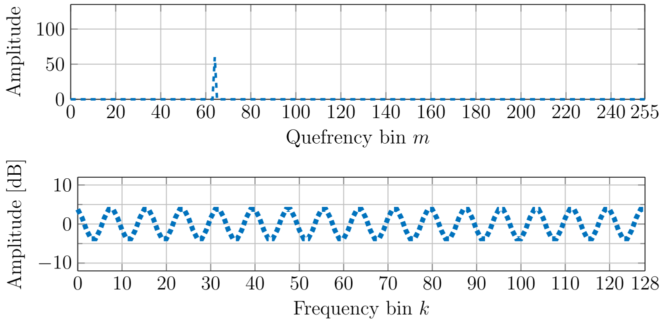

Fig. 2. Example of an **idealized synthetic excitation** for $K = 256$. Upper panel: Cepstrum $c_\ell^R(m)$ with $m_{F_0} = 64$ and $c_\ell^R(64) = 60$, representing a single zero-mean cosine in the log-spectral domain. Lower panel: Log-spectrum $20\log_{10}|R_\ell(k)|$ according to (22), showing only $\frac{K}{2}+1$ bins.

with $m \in \mathcal{M} = \{0, 1, \dots, K-1\}$. The obtained cepstrum has a doubled resolution since we compute it on the whole spectrum (and not only on $\frac{K}{2}+1$ bins). The inverse DCT-II (IDCT-II) will be required in a later stage at the end of the cepstral excitation manipulation (CEM) and is calculated as

$$|\hat{R}_\ell(k)| = \exp\left(\frac{c_\ell^{\hat{R}}(0)}{K} + \frac{2}{K}\sum_{m=1}^{K-1} c_\ell^{\hat{R}}(m)\cdot\cos\left[\pi m\,(k+0.5)\,\frac{1}{K}\right]\right). \tag{22}$$

After the manipulations, the residual signal is mixed with the spectral envelope of the preliminary denoised signal as

$$|\hat{S}_\ell(k)| = |\hat{R}_\ell(k)| \cdot |H_\ell(k)| \tag{23}$$

which is used as numerator for the final *a priori* SNR estimate in an *instantaneous* fashion as follows

$$\hat{\xi}_\ell(k) = \frac{|\hat{S}_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}. \tag{24}$$

One of the most important properties of the cepstrum is the possibility to find a quefrency corresponding to the pitch by simple peak picking [31]. Thus, we estimate the pitch bin index m_{F_0} in a very naïve way by a maximum search of the cepstrum in a defined range specified by naturally occurring pitch values. Our focus is restricted to pitch frequencies F_0 from about 50 Hz to 500 Hz. Using[2] $f = \frac{2f_s}{m}$, the resulting cepstral bin indices at sampling frequency $f_s = 8$ kHz are therefore in the range $m \in \mathcal{M}_{F_0} = \{m_{500} = 32, \dots, m_{50} = 320\}$. Pitch estimation on the basis of the residual signal after preliminary noise reduction is then simply performed according to

$$m_{F_0} = \arg\max_{\mu\in\mathcal{M}_{F_0}} \left(c_\ell^R(\mu)\right). \tag{25}$$

A further convenience is now the ability to easily synthesize a cosine in the log-spectral domain, by creating a cepstrum with only one non-zero bin. An example of such an *idealized* synthetic excitation is given in Fig. 2. The idea behind it is the fact that in voiced speech production harmonics occur at multiples of the fundamental frequency F_0, starting at F_0. A cosine in the log-spectral domain models this quite well already, as the maxima are located directly at the fundamental frequency and due to the periodicity also at the harmonics.

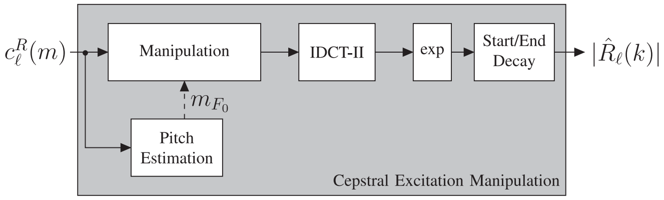

Fig. 3. Block diagram of the **proposed cepstral excitation manipulation** based on an **idealized synthetic excitation**.

IV. Cepstral Excitation Manipulation (CEM)

In the following, we introduce ways to manipulate the excitation in the cepstral domain (referring to the upper right gray block in Fig. 1). These methods form the core of our proposed approach. First, a manipulation towards an idealized *synthetic* excitation is introduced and second, a *template-based* alternative is presented.

A. *Idealized Synthetic Excitation* (CEM$_{\text{ID}}$)

The first option we propose to manipulate the excitation in the cepstral domain is to completely replace it by an idealized synthetic one, followed by the IDCT-II (22) and some final manipulation of the start and the end of the log-spectrum (see Fig. 3). Having found the index m_{F_0} of the cepstral peak amplitude which corresponds to the pitch according to (25), we overestimate its amplitude and transfer it into our synthetic cepstrum

$$c_\ell^{\hat{R}}(m_{F_0}) = c_\ell^R(m_{F_0}) \cdot \alpha_\ell(m_{F_0}), \tag{26}$$

while the remaining quefrencies, except for ($m = 0$), are assigned a zero amplitude:

$$c_\ell^{\hat{R}}(m) = 0, \qquad \forall m \notin \{0, m_{F_0}\}. \tag{27}$$

In order to retain the energy of the preliminary denoised signal's residual, we preserve the cepstral energy coefficient ($m = 0$):

$$c_\ell^{\hat{R}}(0) = c_\ell^R(0). \tag{28}$$

The proposed overestimation factor $\alpha_\ell(m) \geq 1$ could be time-variant and cepstral bin-dependent. While Fig. 2, upper panel, shows an example cepstrum, Fig. 4, upper panel, depicts the same cepstrum with applied cepstral overestimation factor. The resulting effect on the log-spectrum can be seen in Fig. 4, center panel, where a one-view comparison of both log-spectra is provided. Now, the benefit of the directed amplitude overestimation becomes obvious: The overestimation allows a narrower modeling of the harmonics (positive half waves), and also a correspondingly strong emphasis of the valleys (negative half waves) resulting in an increased attenuation between the harmonics. Note that this effect would not be obtained when boosting the harmonics in a shaped or already power-adjusted spectrum with a simple overestimation factor, as this would result only in a shift of the spectrum leaving the negative half

[2] The factor 2 stems from the doubled resolution of our cepstrum definition (21).

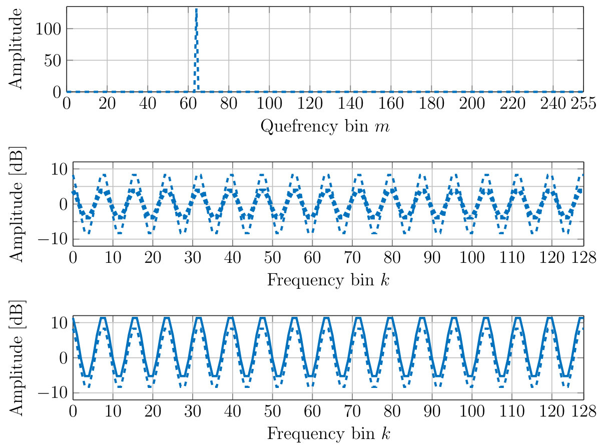

Fig. 4. Example of an **idealized synthetic excitation with overestimation** factor $\alpha_\ell(m_{F_0}) = 2.2$, DFT length $K = 256$, and optionally with preserved cepstral energy coefficient. Upper Panel: Cepstrum $c_\ell^{\hat{R}}(m)$ with $m_{F_0} = 64$ and $c_\ell^{\hat{R}}(64) = 60 \cdot 2.2 = 132$, representing a single zero-mean cosine in the log-spectral domain. Center Panel: Log-spectra $20\log_{10}|R_\ell(k)|$ (bold, dotted line from Fig. 2, lower panel) and $20\log_{10}|\hat{R}_\ell(k)|$ (dashed line, using (26)). Lower panel: Log-spectra $20\log_{10}|\hat{R}_\ell(k)|$ from center panel (dashed line), and power-adjusted log-spectra with additional $c_\ell^{\hat{R}}(0) = 90$ (solid line). All log-spectra show only $\frac{K}{2} + 1$ bins.

waves unmodified. Besides, since our manipulation is in the cepstral domain, our approach translates consistently to all harmonics. This effect is difficult to achieve when operating in the spectral domain.

An overestimation of the energy coefficient would result in a scaling of the whole spectrum which is not desired here, as explained above. An example spectrum depicting the effect of (28) is shown in Fig. 4, lower panel, solid plot.

Naturally, spectral content of voiced human speech starts to occur at the fundamental frequency (after one period of the cosine), then being followed by the corresponding harmonics at multiples of the fundamental frequency, but there should be no spectral content prior to the fundamental frequency. Motivated by our observations during the training of excitation templates, we assume a similar effect at high frequencies (see Fig. 5, upper panel). Thus, a continuation of the cosine beyond the highest, fully representable harmonic is also not desired.

Similarly, in the HRNR approach [4, Fig. 8] a continuous harmonic log-amplitude spectrum is obtained. A problem that has been left unattended there is the falsely introduced half period at low frequencies, which is caused by the non-linear function in [4], applied in order to regenerate harmonics.

To tackle this issue, we propose a simple continuation of the decay of the cosine at low and high frequencies (instead of Fig. 4, lower panel, solid line, now Fig. 5, lower panel). To identify the first local minimum prior to the fundamental frequency we utilize $f = \frac{2f_s}{m}$ to obtain the corresponding pitch frequency F_0 based on the estimated cepstral bin index m_{F_0} from (25). We now convert the pitch frequency F_0 to its corresponding real-valued frequency-domain "bin index" as $\tilde{k}_{F_0} = F_0 \cdot \frac{K}{f_s}$, on basis of the simple relation that $\frac{K}{2}$ corresponds to $\frac{f_s}{2}$ and every frequency bin index containing a related frequency can be obtained by linear interpolation. Due to the periodicity of the cosine we compute the integer bin index of the first local minimum according to

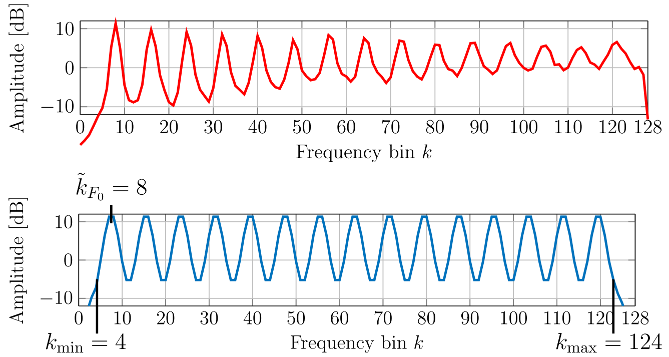

Fig. 5. Upper panel: Example of a log-spectrum excitation template for $m_{F_0} = 64$, obtained as described in Section IV-B. Lower panel: Example of an idealized synthetic excitation log-spectrum with preserved cepstral energy coefficient, overestimation factor $\alpha_\ell(m_{F_0}) = 1.5$, and applied start and end decay to remove the two false half periods.

$$k_{\min} = \left\lceil \frac{\tilde{k}_{F_0}}{2} \right\rceil, \qquad k_{\min} \in \{0, 1, \ldots, K-1\}. \qquad (29)$$

The maximum for the high frequencies is found by analyzing whether the highest possible harmonic frequency and thus the corresponding period of the cosine fits into the non-redundant frequency range as limited by $\frac{f_s}{2}$, or not, according to (30) shown at the bottom of the page.

Here, we have to distinguish between two cases: either, the highest depictable harmonic frequency $(F_0 \cdot \lfloor \frac{f_s}{2F_0} \rfloor)$ including its falling edge $(+\frac{F_0}{2})$ fits into the non-redundant frequency range $(\leq \frac{f_s}{2})$ or it overlaps $(> \frac{f_s}{2})$. For the former we simply calculate the frequency of the last local minimum at the end of the falling edge of the last harmonic $(F_0 \cdot \lfloor \frac{f_s}{2F_0} \rfloor + \frac{F_0}{2})$ and calculate its corresponding frequency bin index $(\cdot \frac{K}{f_s})$. For the latter, since this frequency would be outside of the non-redundant frequency range, we calculate it for the last but one harmonic $(\lfloor \frac{f_s}{2F_0} \rfloor - 1)$, accordingly. For more clarity we refer to the lower panel of Fig. 5, depicting $k_{\min}$, $\tilde{k}_{F_0}$, and also $k_{\max}$.

For all $k < k_{\min}$ and $k > k_{\max}$ the real-valued cosine in the log-spectral domain is discarded and simply to be extended linearly with the slope around $k = k_{\min}$ and $k = k_{\max}$, respectively. The proposed mechanism is *one* possibility to solve the issue quite well already, as a comparison of Fig. 5, upper and lower panel, suggests. Alternatively, different monotonically

$$k_{\max} = \begin{cases} \left\lceil \left(F_0 \cdot \left\lfloor \frac{f_s}{2F_0} \right\rfloor + \frac{F_0}{2} \right) \cdot \frac{K}{f_s} \right\rceil, & \text{for } F_0 \cdot \left\lfloor \frac{f_s}{2F_0} \right\rfloor + \frac{F_0}{2} \leq \frac{f_s}{2} \\ \left\lceil \left(F_0 \cdot \left(\left\lfloor \frac{f_s}{2F_0} \right\rfloor - 1 \right) + \frac{F_0}{2} \right) \cdot \frac{K}{f_s} \right\rceil, & \text{for } F_0 \cdot \left\lfloor \frac{f_s}{2F_0} \right\rfloor + \frac{F_0}{2} > \frac{f_s}{2} \end{cases} \qquad (30)$$

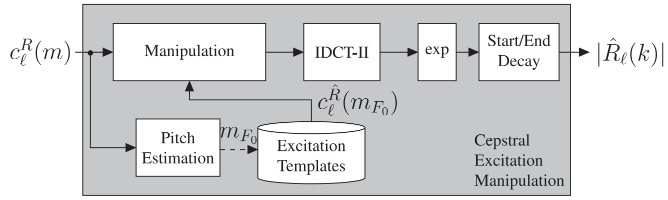

Fig. 6. Block diagram of the **proposed cepstral excitation manipulation** based on **excitation templates**.

increasing and decreasing functions could be applied to model the start and end decay as required.

B. *Cepstral Excitation Templates* ($\mathrm{CEM}_{\mathrm{SI}}$, $\mathrm{CEM}_{\mathrm{SD}}$)

In the cepstral excitation manipulation as described before, the idealized synthetic excitation carries no specific information except for the location of harmonics and a proper energy coefficient. However, further investigations have shown that a residual signal that is obtained via LPC analysis and subsequently transformed into the cepstral domain holds non-negligible information in the remaining bins. As a consequence, the synthetic excitation lacks attributes of the human vocal chords and lungs, being responsible for the naturalness of the modeled excitation signal. Thus, with the now described alternative approach we aim at modeling these components in more detail by creating excitation templates based on excitation signals originating from LPC analysis. A high-level diagram of the cepstral excitation manipulation is depicted in Fig. 6. We obtain cepstral excitation templates in two different ways, depending on whether they are going to be speaker-independent (SI) or speaker-dependent (SD). In the following we describe a general way which is used for both, SI and SD templates, where for the latter it is just a first stage towards a more speaker-specific modeling. In general, the idea is to have a cepstral excitation template for each detectable pitch bin value $m \in \mathcal{M}_{F_0}$. For this, clean speech training material is analyzed and subsequently the DFT spectrum $S_\ell(k)$ of each frame ℓ is separated into spectral envelope and excitation signal (see Fig. 1 and assume $\bar{Y}_\ell(k) = S_\ell(k)$). The DCT-II is employed to transform the excitation signal into the cepstral domain and the pitch bin index m_{F_0} is estimated as explained in Section III-C. Accordingly, we collect the set

$$\mathcal{C}_{m_{F_0}} = \left\{ \mathbf{c}_\ell^R | c_\ell^R(m_{F_0}) \geq c_\ell^R(\mu) \, \forall \mu \in \mathcal{M}_{F_0} \right\} \tag{31}$$

of all cepstral vectors in the training material belonging to each particular pitch bin index $m_{F_0} \in \mathcal{M}_{F_0}$, with $\mathbf{c}_\ell^R = \big(c_\ell^R(0), \ldots, c_\ell^R(m), \ldots, c_\ell^R(K-1)\big)$. Now, the cepstral representation allows us to average per bin over all cepstral excitations in a given set and to obtain a representative excitation template for each pitch bin index m_{F_0} as

$$\bar{\mathbf{c}}^R(m_{F_0}) = \frac{1}{|\mathcal{C}_{m_{F_0}}|} \sum_{\mathbf{c}^R \in \mathcal{C}_{m_{F_0}}} \mathbf{c}^R, \quad \forall m_{F_0} : |\mathcal{C}_{m_{F_0}}| > 0, \tag{32}$$

where $|\cdot|$ is the cardinality of a certain set. If a set is empty, the codebook entry is assigned an all-zero vector. Furthermore, we drop the frame index ℓ as it is only required during the collection of the training material in (31).

The templates can be obtained either in an SI or an SD fashion, only depending on the training data. We propose to use the following adaptation scheme to create SD templates on basis of SI templates. At first, we separately obtain the SI templates $\bar{\mathbf{c}}^R(m_{F_0})$ and preliminary SD templates $\check{\mathbf{c}}^R(m_{F_0})$ stemming from much less data of the target speaker, both according to (31) and (32), differing only in the training material. The actual adaptation is a weighted mixture of both, SI and SD templates for each given pitch bin index m_{F_0} as

$$\tilde{\mathbf{c}}^R(m_{F_0}) = \beta(m_{F_0}) \cdot \bar{\mathbf{c}}^R(m_{F_0}) + (1 - \beta(m_{F_0})) \cdot \check{\mathbf{c}}^R(m_{F_0}) \tag{33}$$

with

$$\beta(m_{F_0}) = \frac{|\bar{\mathcal{C}}_{m_{F_0}}|}{|\bar{\mathcal{C}}_{m_{F_0}}| + \delta \cdot |\check{\mathcal{C}}_{m_{F_0}}|} \tag{34}$$

where $\delta \geq 1$ allows to compensate for the typical lack of SD training material and to artificially emphasize the SD material.

Having obtained and stored the cepstral excitation templates, their application is very similar to the scheme in Section IV-A. After having detected the pitch bin index m_{F_0} according to (25), the SD (33) or SI (32) cepstral excitation template addressed by m_{F_0} is taken. Here, e.g., for SI templates:

$$c_\ell^{\hat{R}}(m) = \bar{c}^R(m_{F_0}, m) \qquad \forall m \notin \{0, m_{F_0}\}\,. \tag{35}$$

The subsequent manipulations from (26) to (28) are applied as before in order to obtain a level consistent with the preliminary denoised signal where (27) is replaced by (35). The proposed start and end decay from (29) and (30) can optionally be applied to compensate for aberrations due to noise in the training material.

If an empty template (originating from $|\mathcal{C}_{m_{F_0}}| = 0$ during training) has been selected by the detected pitch bin index m_{F_0}, we do *not* apply the manipulations from (26) to (28). Instead, we continue with the all-zero cepstral template $\bar{\mathbf{c}}^R(m_{F_0})$ which results in a flat spectrum with unity amplitude. Thus, in such a situation of uncertainty, we do not harm the signal nor do we necessarily enhance it since the clean speech estimate $|\hat{S}_\ell(k)|$ then reduces to solely the envelope $|H_\ell(k)|$. Alternatively, one could also employ the idealized approach in such cases or learn missing templates in an adaptive manner. We comment on the amount of empty templates and their selection frequency during test at the end of Section V-D.

V. Experimental Evaluation

We embed the proposed and the baseline *a priori* SNR estimators in a common noise reduction algorithm to evaluate their performance and analyze their behavior in four different noise types, six different SNR conditions and with two commonly employed spectral weighting rules. Four different quality measures are utilized to compare the different approaches.

A. *Experimental Setup*

Throughout the whole experimental section of this contribution we employ a sample rate $f_s = 8$ kHz with a frame size of 32 ms, corresponding to $K = 256$ samples, and a 50% frame shift by 128 samples. As analysis and overlap-add synthesis window we utilize a periodic square root Hann window and for the source-filter decomposition we compute $N = 10$ LPC coefficients.

The NTT super wideband database [32] is used as a basis and thus downsampled to 8 kHz. We only use the American and British English sets which consist of eight and six speakers, respectively, where each set offers an equal number of speakers per gender. The database comes with 120 utterances for each American English and 100 for each British English speaker. Thus, we decided to artificially decrease the amount to 100 files for American English speakers by random picking, amounting to a total of 14 speakers and 1400 utterances. Next, we use 80% of each speaker's material for training and the remaining 20% for SI and SD testing. For our SI experiments, we decided to use a leave-one-out method to increase the amount of training material. For this, we generate a training set for each speaker separately containing the training material of the 13 other speakers consisting of $13 \times 80 = 1040$ utterances. The training material for the SD adaptation is represented by the 80 utterances of each speaker which have been left out during the SI training. The training itself of the SI and SD templates is conducted according to Section IV-B with applied start and end decay.

The four different noise types are taken from the ETSI [33] database and represent road, car, office, and pub noise. Each segment used to generate the microphone signal is randomly extracted matching the length of the clean speech file. We process the files at six different SNR conditions ranging from -5 dB up to 20 dB in steps of 5 dB. The level of the clean speech and the noise is measured by the active speech level and the root-mean-square level, respectively, according to ITU-T P.56 [34] and both adjusted also according to P.56 prior to superposition. In total we process $14 \times 20 \times 4 \times 6 = 6720$ files for each *a priori* SNR estimator under test. Please note that for the SD experiments, we switch the SD templates corresponding to each speaker being processed, accordingly.

The evaluation of the different *a priori* SNR estimators is placed in a common noise reduction system with MS noise power estimation, one of the *a priori* SNR estimators under test (DD, HRNR, CTS, $\mathrm{CEM_{ID}}$, $\mathrm{CEM_{SI}}$, $\mathrm{CEM_{SD}}$) and the two spectral weighting rules (MMSE-LSA and SG-jMAP) used to calculate the final gains $G_\ell(k)$ which are limited to $G_{\min} = -15$ dB.

The DD approach is tuned with optimal parameters[3] adopted from [35] for each of the weighting rules.

For the HRNR approach a preliminary noise reduction is required for which we also use the MS noise power estimation, DD *a priori* SNR estimation with $\beta_{\mathrm{DD}} = 0.985$ and $\xi_{\min} = -15$ dB as it is just an intermediate step. Furthermore, $G_\ell^{\mathrm{DD}}(k)$ and $G_\ell^{\mathrm{TSNR}}(k)$ are calculated using the WF as proposed and we follow the author's suggestion and utilize $\alpha_\ell(k) = G_\ell^{\mathrm{TSNR}}(k)$ as weights for the mixing in (16) .

The CTS implementation was kindly provided by the authors and thus the parameters left as originally initialized.

Our three proposed estimators $\mathrm{CEM_{ID}}$, $\mathrm{CEM_{SI}}$, and $\mathrm{CEM_{SD}}$, share the same preliminary noise reduction with the HRNR approach except for the weighting rule being MMSE-LSA (instead of the WF) as mentioned in Section III-A. The overestimation factor for (26) is empirically determined and set to $\alpha_\ell(m_{F_0}) = 2$. The required parameter to compensate the lack of speaker-dependent data is found for this particular training set with $\delta = 30$.

B. Quality Measures

To measure the quality of our proposed *a priori* SNR estimator in two example noise reduction contexts, we employ the so-called white-box approach [36], i.e., we calculate the gains $G_\ell(k)$ and subsequently apply it not only to the microphone signal $Y_\ell(k)$ in order to obtain the enhanced signal, but also to the clean speech component $S_\ell(k)$ and the noise component $D_\ell(k)$, separately. The obtained components after IDFT and overlap-add are called the *filtered* clean speech component $\tilde{s}(n)$ and the *filtered* noise component $\tilde{d}(n)$, respectively, which is applicable by assuming (2). The measures are operating *not* on the enhanced signal $\hat{s}(n)$, but only on the filtered and unfiltered *components* with the latter as a reference.

The segmental noise attenuation (NA) [37] is calculated as

$$\mathrm{NA_{seg}} = 10 \log_{10} \left[\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \mathrm{NA}(\ell) \right], \tag{36}$$

with

$$\mathrm{NA}(\ell) = \frac{\sum_{\nu=0}^{N-1} d(\nu + \ell N)^2}{\sum_{\nu=0}^{N-1} \tilde{d}(\nu + \ell N + \Delta)^2},$$

where ℓ defines a segment of length $N = 256$ samples, Δ is compensating the sample delay of the filtered signal, and $\frac{1}{|\mathcal{L}|}$ is a normalization factor since $|\mathcal{L}|$ is the cardinality of the set $\mathcal{L}$, containing all frames. The segmental NA depicts the average of a local frame-wise ratio of the noise component and the corresponding filtered noise component and is sought to be high.

Different from that we define a global measure

$$\Delta\mathrm{SNR} = \mathrm{SNR_{out}} - \mathrm{SNR_{in}}, \tag{37}$$

where $\mathrm{SNR_{in}}$ is the SNR of the clean speech and noise component measured according to ITU P.56 [34], and $\mathrm{SNR_{out}}$ correspondingly for the *filtered* components. This measure gives a more general information of the achieved noise suppression over the whole file compared to the segmental NA. A positive ΔSNR indicates an improved SNR after processing.

Please note that the segmental NA and the ΔSNR are not directly related due to their different scopes (local and global) and have to be interpreted separately, as a high segmental NA is not necessarily indicating great SNR improvement and vice versa.

[3] Optimal parameters for DD *a priori* SNR estimation for the two weighting rules:

$$\text{MMSE-LSA: } \beta_{\mathrm{DD}} = 0.975, \xi_{\min} = -15 \text{ dB}$$

$$\text{SG-jMAP: } \beta_{\mathrm{DD}} = 0.993, \xi_{\min} = -14 \text{ dB}.$$

TABLE I
DETAILED EVALUATION OF SEGMENTAL NA, ΔSNR, PESQ MOS-LQO, AND SEGMENTAL SSDR, FOR THE FOUR DIFFERENT NOISE TYPES, FIVE SNR CONDITIONS, THE BASELINES VS. THE PROPOSED *A PRIORI* SNR ESTIMATORS, AND THE **MMSE-LSA** SPECTRAL WEIGHTING RULE

		NA_seg [dB]						ΔSNR [dB]						PESQ MOS-LQO						SSDR_seg [dB]					
	SNR [dB]	-5	0	5	10	15	mean	-5	0	5	10	15	mean	-5	0	5	10	15	mean	-5	0	5	10	15	mean
ROAD	DD	12.43	12.24	12.02	11.80	11.60	11.91	9.96	10.21	10.12	9.81	9.42	9.74	3.63	3.83	3.98	4.10	4.21	4.01	10.69	14.21	18.19	22.09	25.24	19.63
	HRNR	**14.00**	**13.66**	13.23	12.83	12.46	13.05	11.34	11.57	10.94	9.86	8.51	9.85	3.15	3.39	3.57	3.75	3.92	3.64	6.46	9.19	12.55	16.22	20.02	14.65
	CTS	12.15	11.98	11.86	11.77	11.67	11.83	8.25	8.73	8.69	8.40	7.94	8.23	**3.68**	**3.89**	**4.02**	**4.13**	4.21	**4.04**	9.54	13.47	17.69	21.67	24.87	19.06
	CEM_ID	13.81	13.53	13.24	12.94	12.65	13.08	11.83	11.92	11.52	10.75	9.66	10.66	3.60	3.80	3.97	4.11	**4.22**	4.00	9.56	13.20	17.38	21.55	25.00	19.00
	CEM_SI	13.80	13.55	**13.29**	**13.03**	12.78	**13.16**	12.05	12.08	11.73	11.14	10.39	11.14	3.60	3.79	3.95	4.09	4.21	3.99	10.51	14.32	18.46	22.24	25.22	19.66
	CEM_SD	13.79	13.54	13.28	**13.03**	**12.79**	**13.16**	**12.06**	**12.09**	**11.74**	**11.17**	**10.45**	**11.18**	3.60	3.79	3.95	4.09	4.21	3.99	**10.71**	**14.51**	**18.60**	**22.34**	**25.28**	**19.78**
CAR	DD	11.82	11.80	11.74	11.66	11.54	11.65	7.04	7.28	7.35	7.38	7.36	7.29	**4.19**	4.28	4.34	4.39	4.43	**4.35**	20.54	23.90	26.53	28.17	**28.99**	26.25
	HRNR	12.86	12.39	11.96	11.61	11.29	11.85	6.99	6.68	5.72	4.65	3.78	5.15	3.78	3.96	4.13	4.25	4.32	4.13	13.41	18.03	22.22	25.09	27.08	22.36
	CTS	12.23	12.27	12.25	12.20	12.12	12.18	6.99	6.91	6.72	6.49	6.25	6.56	**4.19**	4.28	4.34	4.39	4.42	4.34	20.09	23.47	26.14	27.94	28.93	25.99
	CEM_ID	14.54	14.41	14.26	14.05	13.78	14.08	14.22	14.22	14.07	13.73	13.18	13.62	4.17	**4.29**	**4.35**	**4.40**	**4.44**	**4.35**	20.66	24.27	**26.74**	**28.19**	28.97	**26.37**
	CEM_SI	**14.60**	**14.51**	**14.41**	14.27	14.09	14.29	**14.40**	14.41	**14.34**	14.16	13.85	14.08	4.15	4.27	4.34	**4.40**	4.43	4.34	21.53	24.32	26.40	27.78	28.61	26.29
	CEM_SD	**14.60**	**14.51**	**14.41**	**14.28**	**14.11**	**14.30**	**14.40**	**14.42**	**14.34**	**14.17**	**13.88**	**14.10**	4.16	4.27	4.34	**4.40**	4.43	4.34	**21.64**	**24.40**	26.47	27.84	28.67	26.36
OFFICE	DD	8.11	7.99	7.87	7.76	7.66	7.83	1.47	2.03	2.32	2.44	2.47	2.20	3.56	3.78	3.94	4.07	**4.18**	3.97	**11.36**	**14.85**	**18.60**	**22.21**	25.17	**19.92**
	HRNR	9.71	9.43	9.20	9.03	8.90	9.18	0.86	1.91	2.48	2.68	2.64	2.18	3.01	3.28	3.50	3.68	3.85	3.55	6.21	8.56	11.78	15.65	19.63	14.17
	CTS	9.42	9.28	9.20	9.14	9.11	9.21	1.44	2.27	2.67	2.85	2.88	2.49	**3.61**	**3.83**	**3.98**	**4.09**	4.17	**3.99**	10.27	13.85	17.81	21.57	24.67	19.18
	CEM_ID	**10.21**	**9.98**	9.79	9.64	9.52	9.76	3.02	3.74	4.06	4.11	4.01	3.80	3.48	3.74	3.92	4.07	**4.18**	3.94	9.52	13.20	17.57	21.89	**25.34**	19.18
	CEM_SI	10.16	9.97	9.81	9.67	9.55	**9.77**	3.31	4.02	4.32	4.38	4.29	4.08	3.47	3.73	3.91	4.05	**4.18**	3.94	9.91	13.81	18.02	21.93	25.00	19.30
	CEM_SD	10.16	9.97	**9.82**	**9.68**	**9.56**	**9.77**	**3.37**	**4.05**	**4.35**	**4.40**	**4.32**	**4.11**	3.48	3.74	3.91	4.06	**4.18**	3.94	10.11	14.01	18.19	22.04	25.06	19.43
PUB	DD	7.72	7.59	7.46	7.37	7.31	7.46	1.49	2.42	2.90	3.06	3.04	2.64	3.16	3.51	3.76	3.94	4.07	3.77	**8.19**	**11.32**	**14.83**	**18.62**	22.33	**16.79**
	HRNR	**10.29**	**10.06**	**9.83**	**9.66**	**9.55**	**9.81**	1.22	3.14	**4.31**	**4.83**	**4.86**	**3.84**	2.65	2.85	3.15	3.44	3.67	3.27	4.29	6.18	9.01	12.80	16.71	11.58
	CTS	9.44	9.25	9.11	9.04	9.02	9.15	1.55	2.90	3.60	3.86	3.84	3.23	**3.26**	**3.58**	**3.80**	**3.96**	**4.08**	**3.81**	6.80	9.86	13.74	17.98	22.05	15.95
	CEM_ID	9.39	9.12	8.89	8.74	8.62	8.88	1.37	3.04	3.75	3.88	3.72	3.20	2.90	3.33	3.70	3.92	4.07	3.68	5.62	9.05	13.55	18.36	**22.81**	15.92
	CEM_SI	9.22	9.00	8.80	8.66	8.54	8.78	1.81	3.35	3.99	4.09	3.92	3.47	2.91	3.35	3.69	3.90	4.06	3.68	6.14	9.70	14.06	18.50	22.57	16.12
	CEM_SD	9.21	8.99	8.80	8.66	8.54	8.77	**1.91**	**3.41**	4.03	4.12	3.95	3.52	2.92	3.36	3.69	3.90	4.06	3.69	6.29	9.89	14.24	18.61	22.62	16.23
Means	DD	10.02	9.90	9.77	9.65	9.53	9.71	4.99	5.49	5.67	5.67	5.57	5.47	3.63	3.85	4.01	4.13	4.22	4.02	**12.69**	**16.07**	**19.53**	**22.77**	25.43	**20.65**
	HRNR	11.72	11.38	11.06	10.78	10.55	10.97	5.10	5.83	5.86	5.51	4.95	5.25	3.15	3.37	3.59	3.78	3.94	3.65	7.59	10.49	13.89	17.44	20.86	15.69
	CTS	10.81	10.70	10.60	10.54	10.48	10.59	4.56	5.20	5.42	5.40	5.23	5.13	**3.68**	**3.89**	**4.04**	**4.14**	4.22	**4.04**	11.68	15.16	18.85	22.29	25.13	20.05
	CEM_ID	**11.99**	**11.76**	11.55	11.34	11.14	11.45	7.61	8.23	8.35	8.12	7.64	7.82	3.54	3.79	3.99	4.12	**4.23**	4.00	11.34	14.93	18.81	22.50	**25.53**	20.12
	CEM_SI	11.95	**11.76**	**11.58**	**11.41**	11.24	**11.50**	7.89	8.47	8.59	8.44	8.11	8.19	3.53	3.78	3.97	4.11	4.22	3.99	12.02	15.54	19.24	22.61	25.35	20.34
	CEM_SD	11.94	11.75	**11.58**	**11.41**	**11.25**	**11.50**	**7.94**	**8.49**	**8.61**	**8.47**	**8.15**	**8.23**	3.54	3.79	3.98	4.11	4.22	3.99	12.19	15.70	19.38	22.71	25.41	20.45

The first measure to assess the quality of the *filtered* clean speech component is the segmental speech-to-speech-distortion ratio (SSDR) [37] calculated as

$$\text{SSDR}_\text{seg} = \frac{1}{|\mathcal{L}_1|} \sum_{\ell \in \mathcal{L}_1} \text{SSDR}(\ell) \tag{38}$$

where $\mathcal{L}_1$ depicts the set of speech active frames obtained by a simple energy threshold-based voice activity detection operating on the clean speech signal $s(n)$. Additionally, $\text{SSDR}(\ell)$ is limited to values between -10 dB and 30 dB by

$$\text{SSDR}(\ell) = \max\left\{\min\left\{\text{SSDR}'(\ell), R_\text{max}\right\}, R_\text{min}\right\}.$$

The actual ratio necessary for computation is obtained by

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

where the speech distortion is

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

A high segmental SSDR indicates low speech distortion and thus good preservation of the speech component.

Furthermore, we employ the PESQ mean opinion score (MOS-LQO) [38] to obtain another measure for the quality of the *filtered* clean speech component. Please note that in line with P.1100 [39, Sec. 8] we do not utilize the enhanced speech $\hat{s}(n)$ in the PESQ measure but the separately processed speech component $\tilde{s}(n)$ as PESQ has not been validated for potential artifacts caused by noise reduction algorithms. We aim at being more compliant to P.862 [38] by doing so.

C. Experimental Results: Details

We provide a detailed evaluation for both the MMSE-LSA and the SG-jMAP spectral weighting rule in the following Tables I and II. Each table depicts the four quality measures for all of the noise types in the SNR conditions from -5 dB to 15 dB averaged over the whole test set, where the best scores are highlighted in boldface. For the computation of the mean over SNRs also the 20 dB SNR condition has been included, which is, however, left out as separate column simply due to space restrictions. This allows for a very extensive analysis of the tested *a priori* SNR estimators for each condition separately.

In Table I the performance results for the MMSE-LSA spectral weighting rule are shown. In terms of noise suppression (measures NA_seg and ΔSNR) the CEM approaches clearly show the strongest performance for each SNR condition, averaged over all four noise types. Both the DD and the CTS baselines show poor performance, and have only few SNR/noise type conditions with convincing performance. The HRNR approach is on average in many cases the best of the baseline approaches w.r.t. noise suppression, showing particularly good performance in pub noise (in NA_seg, and for medium to high SNRs also best

TABLE II
DETAILED EVALUATION OF SEGMENTAL NA, ΔSNR, PESQ MOS-LQO, AND SEGMENTAL SSDR, FOR THE FOUR DIFFERENT NOISE TYPES, FIVE SNR CONDITIONS, THE BASELINES VS. THE PROPOSED *A Priori* SNR ESTIMATORS, AND THE **SG-JMAP** SPECTRAL WEIGHTING RULE

		NA$_{\text{seg}}$ [dB]						ΔSNR [dB]						PESQ MOS-LQO						SSDR$_{\text{seg}}$ [dB]					
	SNR [dB]	-5	0	5	10	15	mean	-5	0	5	10	15	mean	-5	0	5	10	15	mean	-5	0	5	10	15	mean
ROAD	DD	13.25	13.06	12.84	12.64	12.45	12.75	10.61	10.80	10.66	10.39	9.91	10.27	3.63	3.75	3.90	4.05	4.18	3.97	11.22	15.08	19.31	23.27	26.18	20.50
	HRNR	13.36	13.08	12.73	12.40	12.12	12.59	10.44	10.23	9.36	8.86	7.59	8.72	3.41	3.58	3.80	4.01	4.17	3.87	11.36	15.65	20.22	24.23	26.91	21.14
	CTS	12.57	12.40	12.25	12.16	12.07	12.24	8.98	9.19	9.05	8.82	8.34	8.69	**3.76**	**3.86**	**4.00**	**4.14**	**4.24**	**4.05**	**11.85**	**16.22**	**20.66**	**24.44**	**27.03**	**21.45**
	CEM$_{\text{ID}}$	**13.78**	13.53	13.26	13.00	12.76	13.14	11.98	12.03	11.71	11.04	10.18	11.01	3.69	3.82	3.97	4.11	4.23	4.02	11.06	15.21	19.73	23.77	26.67	20.80
	CEM$_{\text{SI}}$	**13.78**	**13.55**	**13.31**	13.07	12.86	13.20	**12.11**	**12.14**	**11.87**	11.34	10.70	11.35	3.68	3.80	3.95	4.10	4.23	4.01	11.63	15.83	20.19	23.93	26.61	21.06
	CEM$_{\text{SD}}$	13.77	**13.55**	**13.31**	**13.08**	**12.87**	**13.21**	**12.11**	**12.14**	**11.87**	**11.35**	**10.74**	**11.37**	3.68	3.80	3.95	4.10	4.23	4.02	11.76	15.94	20.27	23.99	26.64	21.13
CAR	DD	13.00	12.97	12.89	12.82	12.73	12.84	6.66	6.90	6.94	7.01	6.95	6.89	4.15	4.25	4.33	4.39	4.43	4.33	21.66	24.94	27.34	28.60	29.19	26.87
	HRNR	12.94	12.55	12.09	11.72	11.40	11.98	5.91	5.57	5.05	3.93	3.18	4.35	3.99	4.15	4.27	4.34	4.39	4.26	22.26	25.53	27.76	28.90	**29.46**	27.26
	CTS	12.56	12.59	12.56	12.51	12.45	12.51	6.21	6.29	6.19	6.12	5.97	6.10	**4.21**	**4.30**	**4.37**	**4.41**	**4.44**	**4.36**	22.73	25.88	**27.94**	**29.00**	**29.46**	**27.44**
	CEM$_{\text{ID}}$	14.65	14.54	14.42	14.26	14.06	14.29	14.12	14.12	14.01	13.78	13.39	13.70	**4.21**	**4.30**	4.36	**4.41**	**4.44**	**4.36**	22.83	**25.96**	27.87	28.87	29.38	27.42
	CEM$_{\text{SI}}$	**14.69**	**14.61**	**14.52**	**14.41**	14.28	**14.44**	**14.36**	**14.39**	**14.32**	14.18	13.96	**14.14**	4.19	4.29	4.36	4.40	4.43	4.35	22.96	25.74	27.59	28.64	29.20	27.27
	CEM$_{\text{SD}}$	**14.69**	**14.61**	**14.52**	**14.41**	**14.29**	**14.44**	**14.36**	14.38	**14.32**	**14.19**	**13.97**	**14.14**	4.19	4.29	4.36	4.40	4.43	4.35	**23.01**	25.79	27.63	28.66	29.22	27.30
OFFICE	DD	8.45	8.35	8.26	8.19	8.16	8.26	1.49	1.96	2.21	2.31	2.33	2.10	3.57	3.75	3.89	4.02	4.15	3.94	12.02	15.97	19.99	23.60	26.29	20.99
	HRNR	8.16	7.98	7.82	7.72	7.68	7.85	1.25	1.60	1.51	1.56	1.17	1.32	3.38	3.63	3.81	3.98	4.13	3.86	11.95	16.53	21.05	**24.76**	**27.23**	21.70
	CTS	8.82	8.71	8.65	8.62	8.65	8.70	1.44	1.98	2.24	2.32	2.36	2.09	**3.73**	**3.87**	**3.99**	**4.10**	**4.21**	**4.03**	**12.57**	**16.90**	**21.18**	24.70	27.14	**21.85**
	CEM$_{\text{ID}}$	9.72	9.55	9.41	9.32	9.27	9.42	2.87	3.44	3.68	3.72	3.68	3.50	3.65	3.83	3.97	**4.10**	**4.21**	4.01	11.47	15.92	20.63	24.58	27.20	21.40
	CEM$_{\text{SI}}$	**9.75**	9.60	**9.49**	**9.40**	9.35	**9.49**	3.05	3.62	3.86	3.92	3.89	3.69	3.63	3.81	3.96	4.09	4.20	4.00	11.61	16.06	20.54	24.27	26.86	21.29
	CEM$_{\text{SD}}$	**9.75**	**9.61**	**9.49**	**9.40**	**9.36**	**9.49**	**3.08**	**3.64**	**3.87**	**3.93**	**3.90**	**3.71**	3.64	3.81	3.96	4.09	4.20	4.00	11.76	16.19	20.63	24.32	26.87	21.36
PUB	DD	7.85	7.74	7.65	7.61	7.60	7.68	1.35	2.31	2.75	2.91	2.93	2.52	3.26	3.54	3.72	3.87	4.02	3.76	**8.47**	12.10	16.03	20.07	23.83	17.86
	HRNR	7.67	7.53	7.44	7.40	7.39	7.48	1.20	2.09	2.35	2.46	2.48	2.15	3.01	3.36	3.64	3.83	4.00	3.66	7.97	12.27	16.94	21.49	**25.35**	18.64
	CTS	8.54	8.40	**8.32**	**8.30**	**8.34**	**8.39**	1.25	2.39	2.90	3.06	3.08	2.62	**3.51**	**3.71**	**3.85**	**3.98**	**4.10**	**3.89**	8.38	**12.58**	**17.21**	**21.63**	25.30	**18.80**
	CEM$_{\text{ID}}$	**8.66**	**8.46**	8.31	8.23	8.19	8.34	1.30	2.62	3.12	3.22	3.14	2.73	3.24	3.59	3.80	3.95	**4.10**	3.82	7.10	11.52	16.57	21.40	25.24	18.25
	CEM$_{\text{SI}}$	8.59	8.42	8.29	8.22	8.18	8.31	1.51	2.77	3.26	3.36	3.28	2.89	3.24	3.58	3.78	3.94	4.09	3.80	7.33	11.72	16.56	21.13	24.90	18.17
	CEM$_{\text{SD}}$	8.59	8.42	8.29	8.22	8.18	8.31	**1.56**	**2.80**	**3.28**	**3.38**	**3.30**	**2.91**	3.24	3.58	3.78	3.94	4.09	3.81	7.45	11.84	16.63	21.15	24.90	18.23
Means	DD	10.64	10.53	10.41	10.31	10.24	10.38	5.02	5.49	5.64	5.66	5.53	5.44	3.65	3.82	3.96	4.08	4.19	4.00	13.34	17.02	20.67	23.89	26.37	21.55
	HRNR	10.53	10.28	10.02	9.81	9.65	9.97	4.70	4.87	4.57	4.20	3.61	4.13	3.45	3.68	3.88	4.04	4.17	3.91	13.38	17.49	21.49	24.85	**27.24**	22.19
	CTS	10.62	10.52	10.44	10.40	10.38	10.46	4.47	4.96	5.09	5.08	4.94	4.87	**3.80**	**3.93**	**4.05**	**4.16**	**4.25**	**4.09**	**13.88**	**17.90**	**21.75**	**24.94**	27.23	**22.39**
	CEM$_{\text{ID}}$	**11.70**	11.52	11.35	11.20	11.07	11.30	7.57	8.05	8.13	7.94	7.60	7.73	3.70	3.89	4.03	4.14	4.24	4.05	13.12	17.15	21.20	24.66	27.12	21.97
	CEM$_{\text{SI}}$	**11.70**	**11.55**	**11.40**	11.27	11.16	**11.36**	7.76	8.23	8.33	8.20	7.96	8.02	3.68	3.87	4.01	4.13	4.24	4.04	13.38	17.34	21.22	24.49	26.89	21.95
	CEM$_{\text{SD}}$	**11.70**	**11.55**	**11.40**	**11.28**	**11.17**	**11.36**	**7.78**	**8.24**	**8.34**	**8.21**	**7.98**	**8.03**	3.69	3.87	4.01	4.13	4.24	4.04	13.49	17.44	21.29	24.53	26.91	22.01

in ΔSNR). On the contrary, in the SNR = −5 dB condition, HRNR's ΔSNR in pub noise is worst among all schemes, while it yields the best NA$_{\text{seg}}$ in road noise at that SNR. The proposed CEM schemes are much more consistent in terms of NA$_{\text{seg}}$ and ΔSNR over SNR conditions and noises: The speaker-dependent CEM$_{\text{SD}}$ is best in all cases, except only for NA$_{\text{seg}}$ in very low SNR, where CEM$_{\text{ID}}$ is slightly ahead.

In terms of the speech component quality (PESQ, SSDR$_{\text{seg}}$) the picture is partly different: On average over the noise types CTS performs best in PESQ in most SNRs, being ahead up to 0.15 MOS points vs. the worst CEM approach. Interestingly, however, in car noise, CEM$_{\text{ID}}$ is slightly better than CTS in most SNR conditions. The classical DD approach performs on a par with other approaches with regard to the PESQ metric, while HRNR consistently fails to provide an acceptable speech component quality, both in PESQ and SSDR$_{\text{seg}}$. Surprisingly, DD delivers very good SSDR$_{\text{seg}}$ performance in office and pub noise, while the CEM approaches perform best in car and road noise.

In Table II the performance results for the SG-jMAP spectral weighting rule are shown. In terms of noise suppression (measures NA$_{\text{seg}}$ and ΔSNR) the CEM approaches, especially the speaker-dependent variant CEM$_{\text{SD}}$, clearly perform best in each SNR condition, averaged over all four noise types. Both the DD and here the HRNR baselines show poor performance, and have no single SNR/noise type condition with convincing NA$_{\text{seg}}$ performance. Considering ΔSNR, *none* of the three baselines has a single SNR/noise type condition with superior performance. The CTS approach is on average in most cases the best of the baseline approaches w.r.t. NA$_{\text{seg}}$, showing particularly good performance in pub noise for medium to high SNRs. Interestingly, for the SG-jMAP, the DD approach is on average the best baseline w.r.t. ΔSNR, showing that sophisticated spectral weighting rules are able to heal some shortcomings of earlier processing stages such as the SNR estimation. The proposed CEM schemes are much more consistent in terms of NA$_{\text{seg}}$ and ΔSNR over SNR conditions and noises: The speaker-dependent CEM$_{\text{SD}}$ is best in all cases.

In terms of the speech component quality (PESQ, SSDR$_{\text{seg}}$) the picture again is partly different: On average over the noise types, CTS performs best in PESQ for all SNRs, being ahead up to 0.12 MOS points vs. the worst CEM approach. In car noise, however, CEM$_{\text{ID}}$ is on a par with CTS in most SNR conditions. On average, the DD approach performs quite well in PESQ, while HRNR consistently settles for the worst score. The SSDR$_{\text{seg}}$ is also mostly in favor of the CTS approach, and opposite to PESQ, the HRNR approach is found to be slightly ahead of the DD estimator on average.

Please note that the advantage of a speaker-dependent approach vs. all other speaker-independent approaches is of course somehow expected, yet CEM$_{\text{SD}}$ is only slightly ahead of our speaker-independent method CEM$_{\text{SI}}$.

We can summarize for both weighting rules, that on average over the noise types the CEM approaches perform best in

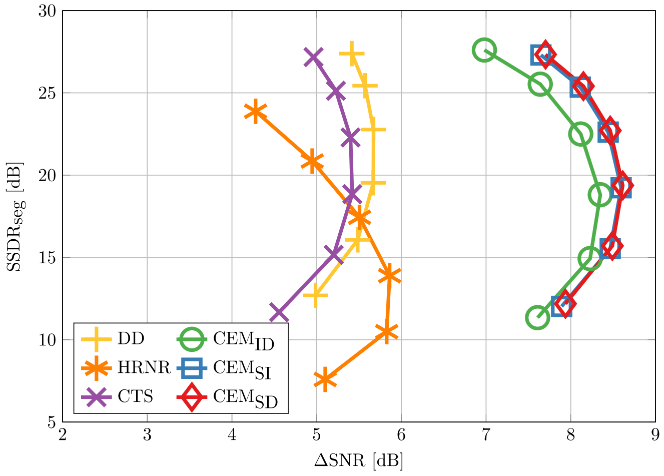

Fig. 7. Segmental SSDR and **ΔSNR** averaged over the four different noise types for the different *a priori* SNR estimators under test with the **MMSE-LSA** spectral weighting rule.

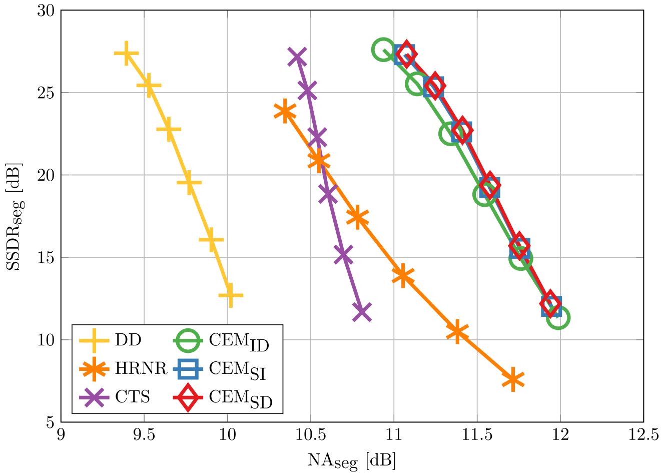

Fig. 9. Segmental SSDR and **segmental NA** averaged over the four different noise types for the different *a priori* SNR estimators under test with the **MMSE-LSA** spectral weighting rule.

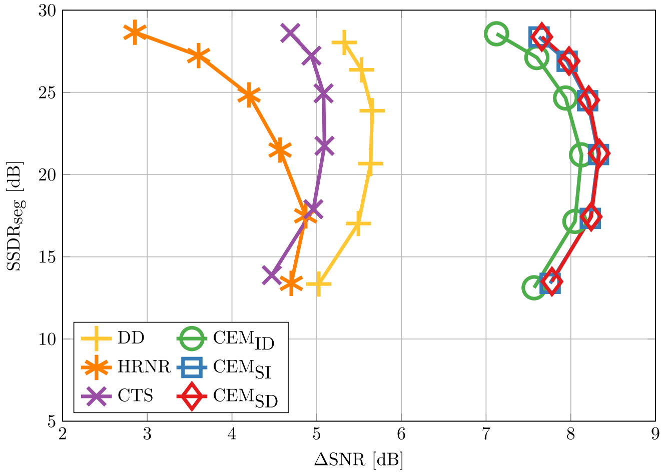

Fig. 8. Segmental SSDR and **ΔSNR** averaged over the four different noise types for the different *a priori* SNR estimators under test with the **SG-jMAP** spectral weighting rule.

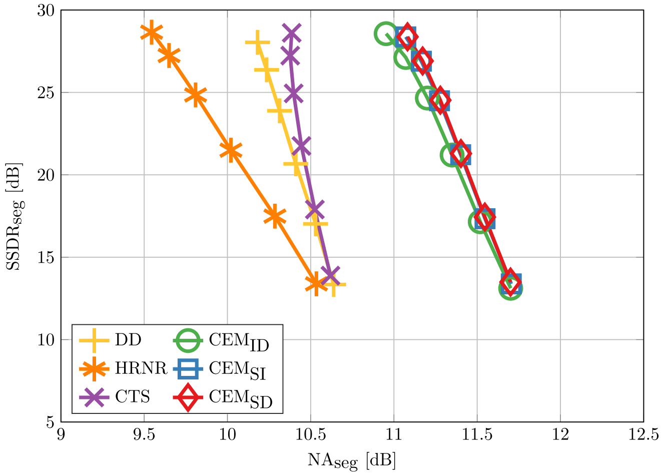

Fig. 10. Segmental SSDR and **segmental NA** averaged over the four different noise types for the different *a priori* SNR estimators under test with the **SG-jMAP** spectral weighting rule.

terms of $\mathrm{NA}_{\mathrm{seg}}$ and ΔSNR, and almost on a par with the best performing method w.r.t. PESQ and $\mathrm{SSDR}_{\mathrm{seg}}$. The baselines all show an imbalanced performance being inferior in one of the two main categories: For both weighting rules, they are inferior w.r.t. $\mathrm{NA}_{\mathrm{seg}}$ and ΔSNR (DD, HRNR and CTS). For the MMSE-LSA weighting rule, HRNR shows poor performance w.r.t. PESQ and $\mathrm{SSDR}_{\mathrm{seg}}$, where for the SG-jMAP weighting rule HRNR performs poorly w.r.t. PESQ, while DD is only slightly inferior in $\mathrm{SSDR}_{\mathrm{seg}}$.

D. Discussion

To enable further analysis of the results, we plot the $\mathrm{SSDR}_{\mathrm{seg}}$ for the two spectral weighting rules over the ΔSNR (Figs. 7 and 8) and the $\mathrm{NA}_{\mathrm{seg}}$ (Figs. 9 and 10), respectively. The plots simplify the interpretation on a more global level compared to the tables as only two dimensions are considered at a time. Each plot is a visualization of the mean section from the corresponding table. In each figure the different estimators are specifiable by their respective marker, being $+$ for DD, $*$ for HRNR, and $\times$ for CTS, the three illustrating the baseline algorithms. The proposed techniques are distinguishable by $\circ$ for the idealized synthetic approach $\mathrm{CEM}_{\mathrm{ID}}$, $\square$ for the template-based variant $\mathrm{CEM}_{\mathrm{SI}}$, and finally $\diamond$ for the speaker-dependent template-based implementation $\mathrm{CEM}_{\mathrm{SD}}$. Each marker depicts one of the SNR conditions from -5 dB to 20 dB in steps of 5 dB, where the lowest corresponds to the worst and the highest to the best condition, respectively. The further a marker is located to the top right hand corner of each plot, the better is the performance. The range and scaling of each axis showing its respective measure is the same to ensure comparability across the two spectral weighting rules.

The MMSE-LSA estimator is depicted in Fig. 7 showing how close DD and CTS are. The HRNR approach exhibits quite an unbalanced behavior as ΔSNR performance is similar to the

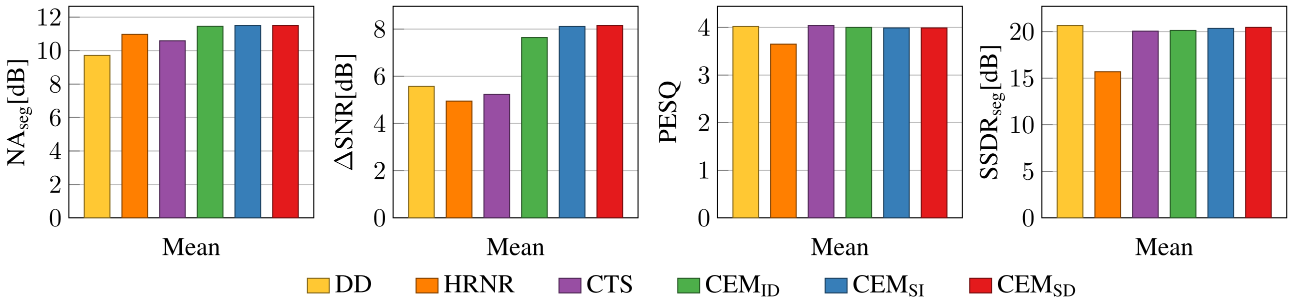

Fig. 11. Evaluation of segmental NA, ΔSNR, PESQ MOS-LQO, and segmental SSDR, averaged over the four different noise types and six SNR conditions showing the baselines vs. the proposed *a priori* SNR estimators, and the **MMSE-LSA** spectral weighting rule.

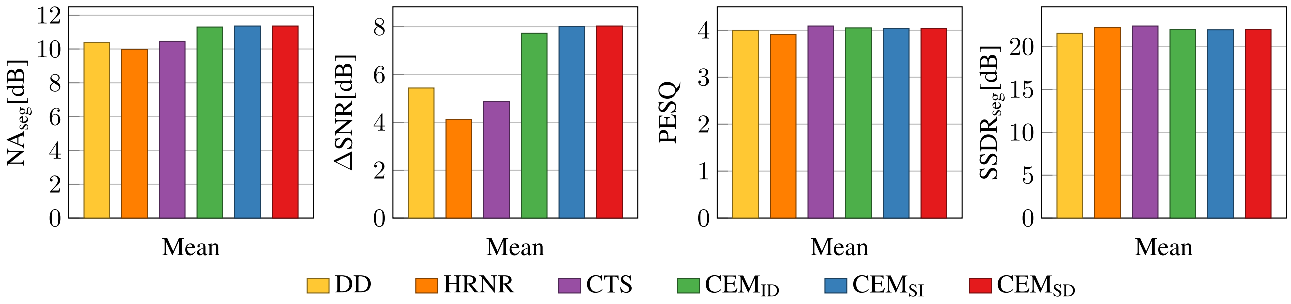

Fig. 12. Evaluation of segmental NA, ΔSNR, PESQ MOS-LQO, and segmental SSDR, averaged over the four different noise types and six SNR conditions showing the baselines vs. the proposed *a priori* SNR estimators, and the **SG-jMAP** spectral weighting rule.

other baselines, but $SSDR_{seg}$ is significantly lower. The CEM approaches show highest ΔSNR in all conditions with comparable or better speech component quality w.r.t. the baselines. The flexion of the three curves also indicates a balanced working point in different SNR conditions. Using SG-jMAP (Fig. 8), the relationships across the tested algorithms are quite similar, except that HRNR improves in $SSDR_{seg}$, and gets worse in ΔSNR. The CEM algorithms obtain the best ΔSNR at very comparable speech component quality. For both spectral weighting rules, the CEM implementations among themselves show a very consistent rank order such that CEM_{ID} is marginally outperformed by the two template-based algorithms, which are almost equivalent.

Interestingly, all approaches obtain an increased preservation of the speech component quality with the SG-jMAP weighting rule compared to the MMSE-LSA spectral weighting rule at comparable (DD, CEM) or slightly lower (HRNR, CTS) ΔSNR values. *CEM is ahead of the baseline approaches by a ΔSNR of at least 2.35 dB (MMSE-LSA) and 2.29 dB (SG-jMAP) on total average.*

Figs. 9 and 10 provide the same analysis but for the $SSDR_{seg}$ and the NA_{seg}. Fig. 9 depicting MMSE-LSA shows that CTS outperforms DD in terms of noise attenuation, however, with comparable quality of the speech component. The performance of HRNR is more difficult to interpret than before as the working point clearly is shifted since the single SNR conditions do not even roughly line up horizontally with the other baseline approaches under test. This results in a decreased speech component quality at substantially higher NA_{seg} values for each condition. Still, the proposed CEM approaches manage to show exceeding performance in all SNRs. The SG-jMAP spectral weighting rule is shown in Fig. 10. DD improves in NA_{seg} at similar $SSDR_{seg}$ compared to Fig. 9 such that DD and CTS are much closer now. However, CTS is still able to consistently show a superior performance compared to the DD approach. Using SG-jMAP, HRNR loses performance in NA_{seg}, and becomes better in $SSDR_{seg}$, resulting in a clearer picture as opposed to Fig. 9. Best performing is again the CEM group, showing a similar behavior amongst themselves as in the other figures. In general, the range of NA_{seg} is the most compressed, considering the other two figures. *CEM is ahead of the baseline approaches by an NA_{seg} of up to 1.79 dB (MMSE-LSA) and 1.39 dB (SG-jMAP) on total average.*

We attribute the strong increase of NA_{seg} and ΔSNR in car noise mainly to the applied start decay as seen in Fig. 5, as it causes a good suppression in low frequencies typical for car noise. Moreover, for negative SNR conditions in pub noise we encounter cases where some other approach provides the best results with regard to the noise attenuation and speech component quality metrics. This is most likely due to F_0 estimation errors caused by the naïve pitch estimation which is unable to track a target speaker due to the presence of other speakers. We assume that the overall increase in NA_{seg} and also ΔSNR is caused by the introduced overestimation of the harmonics and the simultaneous attenuation in between them as shown in Fig. 4, center panel. Specifically, the attenuation between the harmonics should account for the increased overall noise attenuation. As there is usually a trade-off between noise attenuation and speech component quality [40], the CEM approach seems to mitigate this effect and allows us to be nearly on a par with the best baseline on average in terms of $SSDR_{seg}$ and PESQ, while maintaining a higher NA_{seg} and ΔSNR.

To facilitate a conclusive interpretation of Tables I and II we provide bar charts depicting the overall mean values (last column of the mean section of Tables I and II for each measure) in Figs. 11 and 12 for the MMSE-LSA and the SG-jMAP weighting rule, respectively. Both figures show the advantage of the proposed approaches on average over the baselines in terms of $\mathrm{NA}_{\mathrm{seg}}$ and particularly ΔSNR. The CEM approaches are ahead of the baselines by at least 2 dB w.r.t. ΔSNR, while maintaining a very comparable speech component quality in both other measures, PESQ and $\mathrm{SSDR}_{\mathrm{seg}}$. However, the quality improvement from $\mathrm{CEM}_{\mathrm{ID}}$ to $\mathrm{CEM}_{\mathrm{SI}}$ or even $\mathrm{CEM}_{\mathrm{SD}}$ is only marginal. Nevertheless, $\mathrm{CEM}_{\mathrm{SD}}$ on average performs best among the CEM approaches. The HRNR approach seems to deliver a better speech component quality when used together with SG-jMAP (as compared to MMSE-LSA), which is strongly reflected in the $\mathrm{SSDR}_{\mathrm{seg}}$ measure at the cost of only a minor decrease in noise attenuation. Again, this shows how an advanced weighting rule can mend estimation flaws of earlier components in a noise reduction scheme. The DD and CTS baselines show a quite consistent performance regardless of the applied weighting rule.

In our experiments we encountered some empty templates during the training, which is caused by a lack of training material. A brief analysis has shown that mostly for lower cepstral bin indices we find every other set being empty, indicating that for some higher pitch frequencies ($F_0 > 400\,\mathrm{Hz}$) no material has been seen during training. However, we also obtain some coherent clusters for the lower frequencies. One way to avoid this could be to reduce the resolution of the cepstrum as we would not have seen any empty sets with the normal resolution but also would not have had the gain of the additional precision reflected by the coherent clusters. Furthermore, we could verify that an excitation template has been applied to 99.99% of the frames processed by the template-based methods, showing that empty templates do not have any significant relevance at this point. In addition to that, informal listening tests have shown that the proposed CEM methods also allow for almost musical tone-free noise suppression due to the instantaneous nature of *a priori* SNR estimation.[4]

VI. Conclusion

In this paper we have introduced three novel methods for instantaneous *a priori* SNR estimation utilizing the source-filter model for speech production. A preliminary denoised signal is decomposed into its source and corresponding filter, allowing to impose an idealized excitation on the degenerated source. The cepstral domain is exploited to manipulate the excitation signal at hand. We further enhance the technique by obtaining excitation templates from clean speech in either a speaker-independent or speaker-dependent fashion, where the latter is the slightly superior approach on average. However, the idealized technique shows some advantages over the codebook-based approaches, especially in SNR conditions ≥ 10 dB where it achieves equal or even better speech component quality at the cost of slightly lower noise suppression. We tested our algorithms and three baseline estimators in a common noise reduction algorithm with two different spectral weighting rules and managed to show a ΔSNR improvement of more than 2 dB, while the amount of speech distortion is largely kept on a constant level. Future work will include an enhancement of not only the excitation but also the envelope which is still taken from the preliminary denoised signal. Also a more sophisticated approach to F_0 estimation or tracking could improve our approaches in low-SNR conditions with multiple speakers.

Acknowledgment

The authors would like to thank T. Gerkmann for providing an executable to simulate the CTS baseline and also the anonymous reviewers for their constructive feedback helping to improve the quality of this paper.

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] Y. Ephraim and D. Malah, “Speech enhancement using a minimum mean-square error log-spectral amplitude estimator,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.

[3] I. Cohen, “Speech enhancement using super-Gaussian speech models and noncausal *a priori* SNR estimation,” *Speech Commun.*, vol. 47, no. 3, pp. 336–350, Nov. 2005.

[4] C. Plapous, C. Marro, and P. Scalart, “Improved signal-to-noise ratio estimation for speech enhancement,” *IEEE Trans. Audio, Speech, Lang. Process.*, 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. IEEE Int. Conf. Acoust., Speech Signal Process.*, 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 Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 1, pp. 186–195, Jan. 2011.

[7] S. Elshamy, N. Madhu, W. J. Tirry, and T. Fingscheidt, “An iterative speech model-based *a priori* SNR estimator,” in *Proc. INTERSPEECH*, Dresden, Germany, Sep. 2015, pp. 1740–1744.

[8] J. Sohn, N. S. Kim, and W. Sung, “A statistical model-based voice activity detection,” *IEEE Signal Process. Lett.*, vol. 6, no. 1, pp. 1–3, Jan. 1999.

[9] T. Gerkmann, C. Breithaupt, and R. Martin, “Improved *a posteriori* speech presence probability estimation based on a likelihood ratio with fixed priors,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 910–919, Jul. 2008.

[10] P. Scalart and J. V. Filho, “Speech enhancement based on *a priori* signal to noise estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Atlanta, GA, USA, May 1996, pp. 629–632.

[11] T. Lotter and P. Vary, “Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model,” *EURASIP J. Appl. Signal Process.*, vol. 2005, no. 7, pp. 1110–1126, 2005.

[12] O. Cappé, “Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor,” *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.

[13] C. Breithaupt and R. Martin, “Analysis of the decision-directed SNR estimator for speech enhancement with respect to low-SNR and transient conditions,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 2, pp. 277–289, Feb. 2011.

[14] C. Plapous, C. Marro, L. Mauuary, and P. Scalart, “A two-step noise reduction technique,” in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Montreal, QC, Canada, May 2004, pp. 289–292.

[15] M. Djendi and P. Scalart, “Reducing over- and under-estimation of the *a priori* SNR in speech enhancement techniques,” *Digit. Signal Process.*, vol. 32, pp. 124–136, Sep. 2014.

[16] F. Deng and C. Bao, “Speech enhancement based on AR model parameters estimation,” *Speech Commun.*, vol. 79, pp. 30–46, May 2016.

[4] Audio samples can be found under: https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2017-taslp-cem/

[17] 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.

[18] I. Cohen, “Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging,” *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[19] S. Rangachari and P. C. Loizou, “A noise-estimation algorithm for highly non-stationary environments,” *Speech Commun.*, vol. 48, no. 2, pp. 220–231, Feb. 2006.

[20] T. Gerkmann and R. C. Hendriks, “Unbiased MMSE-based noise power estimation with low complexity and low tracking delay,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383–1393, May 2012.

[21] D. L. Wang and J. S. Lim, “The unimportance of phase in speech enhancement,” *IEEE Trans. Acoust., Speech Signal Process.*, vol. ASSP-30, no. 4, pp. 679–681, Aug. 1982.

[22] P. Vary, “Noise suppression by spectral magnitude estimation—Mechanism and theoretical limits—,” *Signal Process.*, vol. 8, no. 4, pp. 387–400, Jul. 1985.

[23] K. Paliwal, K. Wójcicki, and B. Shannon, “The importance of phase in speech enhancement,” *Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011.

[24] P. Mowlaee, R. Saeidi, and Y. Stylianou, “INTERSPEECH 2014 special session: Phase importance in speech processing,” in *Proc. INTERSPEECH*, Singapore, Sep. 2014, pp. 1623–1627.

[25] J. Kulmer and P. Mowlaee, “Phase estimation in single channel speech enhancement using phase decomposition,” *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 598–602, May 2015.

[26] T. Gerkmann and R. Martin, “On the statistics of spectral amplitudes after variance reduction by temporal cepstrum smoothing and cepstral nulling,” *IEEE Trans. Signal Process.*, vol. 57, no. 11, pp. 4165–4174, Nov. 2009.

[27] T. Gerkmann, “Cepstral weighting for speech dereverberation without musical noise,” in *Proc. 19th Eur. Signal Process. Conf.*, Barcelona, Spain, Sep. 2011, pp. 2309–2313.

[28] J. E. Markel and A. H. Gray, *Linear Prediction of Speech*. Berlin, Germany: Springer-Verlag, 1976.

[29] J. Benesty, M. M. Sondhi, and Y. Huang, Eds., *Springer Handbook of Speech Processing*. Berlin, Germany: Springer-Verlag, 2008.

[30] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, 1st ed. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[31] A. M. Noll, “Cepstrum pitch determination,” *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.

[32] “Super Wideband Stereo Speech Database,” NTT Advanced Technology Corporation (NTT-AT). [Online]. Available: http://www.ntt-at.com/product/widebandspeech/

[33] European Telecommunications Standards Institute, *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*, ETSI EG 202 396-1, Sep. 2008.

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

[35] H. Yu, “Post-filter optimization for multichannel automotive speech enhancement,” Ph.D. dissertation, Inst. Commun. Tech., Technische Univ. Braunschweig, Braunschweig, Germany, 2013.

[36] S. Gustafsson, R. Martin, and P. Vary, “On the optimization of speech enhancement systems using instrumental measures,” in *Proc. Workshop Qual. Assessment Speech, Audio, Image Commun.*, Darmstadt, Germany, Mar. 1996, pp. 36–40.

[37] T. Fingscheidt, S. Suhadi, and S. Stan, “Environment-optimized speech enhancement,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 16, no. 4, pp. 825–834, May 2008.

[38] International Telecommunication Union, *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, Telecommunication Standardization Sector (ITU-T) Rec. P.862, Feb. 2001.

[39] International Telecommunication Union, *Narrow-Band Hands-Free Communication in Motor Vehicles*, Telecommunication Standardization Sector (ITU-T) Rec. P.1100, Jan. 2015.

[40] J. Chen, J. Benesty, Y. Huang, and S. Doclo, “New insights into the noise reduction wiener filter,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 14, no. 4, pp. 1218–1234, Jul. 2006.

Samy Elshamy received the B.Sc. degree in bioinformatics from Friedrich-Schiller-Universität Jena, Jena, Germany, in 2011 and the M.Sc. degree in computer science from Technische Universität Braunschweig, Braunschweig, Germany, in 2013. He is currently working toward the Ph.D. degree in the field of speech enhancement at the Institute for Communications Technology, Technische Universität Braunschweig.

Nilesh Madhu received the Dr.-Ing. degree in electrical engineering and information technology from the Ruhr-Universität Bochum, Bochum, Germany, in 2009. Following this, he received a Marie-Curie fellowship for a two-year postdoctoral stay at the KU Leuven, Belgium, where he successfully applied his signal processing knowledge to the field of hearing prostheses. Since 2011, he has been with NXP, Leuven, Belgium, and is currently a Principal Scientist within the Product Line Mobile Audio Solutions, where he and his team work on developing innovative algorithms for audio and speech enhancement for mobile devices. He is passionate about signal processing, and is especially interested in the field of signal detection and enhancement for various applications, not just audio.

Wouter Tirry received the M.Sc. degree in physics and the Ph.D. degree in solar physics from the University of Leuven, Leuven, Belgium, in 1994 and 1998, respectively. As a Postdoc, he further pursued his research at the National Centre for Atmospheric Research, Boulder, CO, USA. Since 1999, he has been building up expertise in the domain of speech enhancement for mobile devices at Philips and NXP, Leuven, Belgium, as a Research Engineer and a System Architect. He is currently a Senior Principal at the Product Line Mobile Audio Solutions, NXP, leading the speech technology development activities.

Tim Fingscheidt (S’93–M’98–SM’04) received the Dipl.-Ing. degree in electrical engineering in 1993 and the Ph.D. degree in 1998 from RWTH Aachen University, Aachen, Germany. He further pursued his work on joint speech and channel coding as a Consultant in the Speech Processing Software and Technology Research Department, AT&T Labs, Florham Park, NJ, USA. In 1999, he entered the Signal Processing Department of Siemens AG (COM Mobile Devices) in Munich, Germany, and contributed to speech codec standardization in ETSI, 3GPP, and ITU-T. In 2005, he joined Siemens Corporate Technology in Munich, Germany, leading the speech technology development activities in recognition, synthesis, and speaker verification. Since 2006, he is a Full Professor in the Institute for Communications Technology, Technische Universität Braunschweig, Braunschweig, Germany. His research interests include speech and audio signal processing, enhancement, transmission, recognition, and instrumental quality measures. He received several awards, among them a prize of the Vodafone Mobile Communications Foundation in 1999, and the 2002 prize of the Information Technology branch of the Association of German Electrical Engineers (VDE ITG), where he is leading the Speech Acoustics Committee ITG FA4.3 since 2015. From 2008 to 2010, he served as an Associate Editor for IEEE Transactions on Audio, Speech, and Language Processing, and since 2011 as a member of the IEEE Speech and Language Processing Technical Committee.