DNN-Based Cepstral Excitation Manipulation for Speech Enhancement

Samy Elshamy and Tim Fingscheidt, *Senior Member, IEEE*

***Abstract*—This contribution aims at speech model-based speech enhancement by exploiting the source-filter model of human speech production. The proposed method enhances the excitation signal in the cepstral domain by making use of a deep neural network (DNN). We investigate two types of target representations along with the significant effects of their normalization. The new approach exceeds the performance of a formerly introduced classical signal processing-based cepstral excitation manipulation (CEM) method in terms of noise attenuation by about 1.5 dB. We show that this gain also holds true when comparing serial combinations of envelope and excitation enhancement. In the important low-SNR conditions, no significant trade-off for speech component quality or speech intelligibility is induced, while allowing for substantially higher noise attenuation. In total, a traditional purely statistical state-of-the-art speech enhancement system is outperformed by more than 3 dB noise attenuation.**

***Index Terms*—Speech enhancement, deep learning, cepstrum, *a priori* SNR.**

I. Introduction

SPEECH enhancement is still a very important and active field of research. Its primary aim is to improve speech quality and intelligibility, to facilitate the most natural way of communication. Speech signals might be corrupted by, e.g., bandwidth limitation, coupling of noise, echo, and reverberation. In order to combat such problems, various algorithms have been developed and improved over time.

Single-channel noise reduction is still a challenging task, which is addressed here. Even though traditional systems might be still considered as state of the art, recent advances in speech enhancement make more and more use of modern deep learning technologies and often end-to-end solutions are presented (e.g., [1]–[3]). As mentioned in [3], one issue of conventional DNN-based enhancement models is the discontinuity of the enhanced signals when processed in a frame-based manner. The authors resolve the problem by enhancing whole utterances on waveform level which requires the availability of complete recordings or at least a very large buffer. This is not applicable for telephony applications, where delay has to be as low as possible and frame-wise processing is essential. In the following, more recent advances will be presented briefly.

Manuscript received April 9, 2019; revised July 3, 2019 and August 2, 2019; accepted August 2, 2019. Date of publication August 8, 2019; date of current version August 21, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. S. Doclo. *(Corresponding author: Samy Elshamy.)*

The authors 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).

Digital Object Identifier 10.1109/TASLP.2019.2933698

A sketch of less holistic approaches, that in parts still respect traditional and statistical speech enhancement is shown in [4]. The publication nicely shows various levels of granularity that allow to move away from end-to-end solutions towards more traditional structures, still being able to benefit from modern technology. Following this, DNN-based learning of spectral weighting rules has been evaluated, e.g., for ideal binary masks and ideal ratio masks in [5], [6].

The spectral envelope codebook-based work by Srinivasan *et al.* [9]–[11] was brought from an autoregressive (AR) model to the cepstral domain by Rosenkranz *et al.* [12], and it has been picked up again recently in [13] and [14]. In [13], the authors combine the existing auto regressive-based approach with a noise estimator [15] to circumvent the dependency on a noise codebook. Additionally, they introduce an SPP estimator [16] to combat the lack of noise suppression between the harmonics, which is naturally not possible when only spectral envelopes are used for the estimation of the clean speech. This issue has been further addressed in our previous work [7] and is also investigated together with the preservation of harmonics in this publication by analyzing the effects of the normalization of targets during the training process. The authors in [14] aim to replace both codebooks by estimating the parameters of the AR models for speech and noise simultaneously with a single network that predicts line spectral pairs. In order to combat the inability to reduce noise between the harmonics, they also use the SPP estimator from [16]. In both cases the estimated entities are used to create a Wiener filter and for the latter approach it depicts a further step towards a more modular integration of DNNs into a statistical speech enhancement framework.

The *a priori* SNR represents a more generic entity, as it can be easily plugged into various statistical systems, also being a key factor in noise reduction. It has been subject to research not only through the past decades [17]–[19], but particularly in the recent past with quite some success [7], [8], [20]–[25]. While most approaches work in the frequency domain, Breithaupt *et al.* originally pioneered the way for *a priori* SNR estimation in the cepstral domain [20]. Stahl *et al.* pick up the original decision-directed (DD) approach by Ephraim and Malah [17] and propose to smooth the *a priori* SNR not over isolated frequency bins but with respect to harmonic trajectories [24]. This leads to higher noise attenuation without further speech distortion. Xu *et al.* make use of discriminative non-negative matrix factorization (DNMF) for *a priori* SNR estimation and present two different approaches [25]. One approach uses DNMF to estimate speech and noise power to directly calculate the *a priori* SNR, while

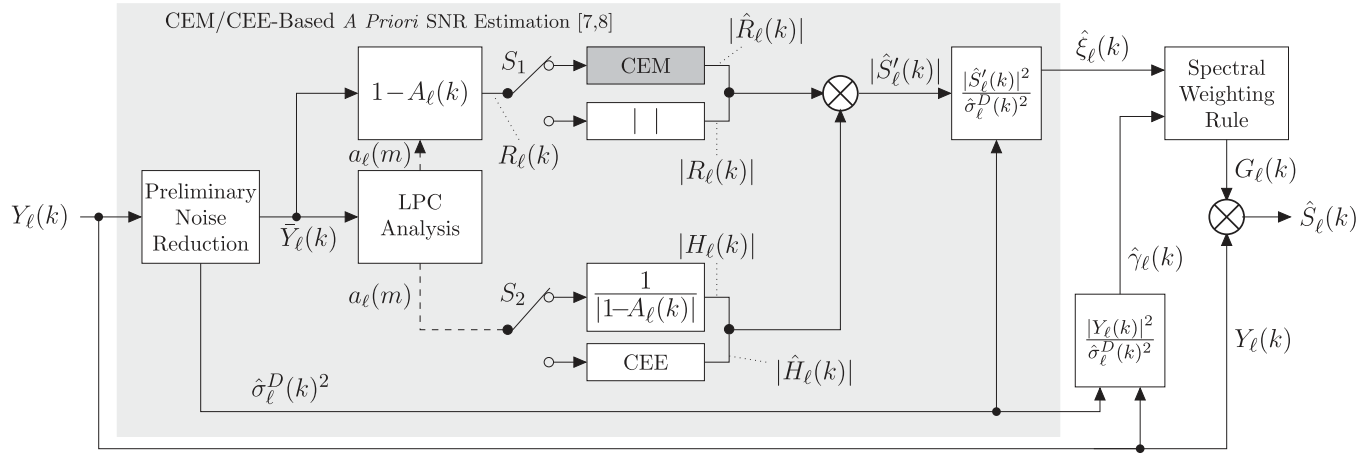

Fig. 1. Schematic of the speech enhancement framework with either **cepstral excitation manipulation (CEM)** [7] (switches S_1 and S_2 as shown) or **cepstral envelope estimation (CEE)** [8] (switches S_1 and S_2 in lower position) *a priori* SNR estimation. The CEM block is depicted in more detail in Fig. 2, as its replacement by a deep neural network is core novelty of this work.

the other uses DNMF to only estimate the noise power which is then used together with the DD approach. Both methods obtain better results than DNMF approaches that are commonly used to directly estimate the clean speech. However, they rely on noise codebooks which might limit the capability of generalization.

In this contribution, we aim to exploit the potential of the cepstral excitation manipulation (CEM) approach further, as the current state-of-the-art CEM solution [7] offers room for improvement, in terms of speech quality, speech intelligibility, and also noise attenuation. To do so, we incorporate deep neural network (DNN) models to enhance the residual signal for the purpose of *a priori* SNR estimation for speech enhancement. A particular aspect is that the explicit F_0 estimator as required by state-of-the-art CEM is not needed anymore for the core functionality of CEM in our new approach. We investigate two different lines of research for the *a priori* SNR numerator. The first aims to restore the clean speech residual signal from a noisy observation. The second is to restore the clean speech signal itself by estimating a residual signal which is also considering and compensating the degeneration of the spectral envelope in noisy conditions. The performance of the *a priori* SNR estimator is evaluated in a speech enhancement task—although its application is not limited to that—and measured by renowned metrics such as the PESQ score [26], [27], the short-time objective intelligibility measure (STOI) [28], and also the segmental noise attenuation (NA_{seg}) [29].

This paper is structured as follows. We briefly describe the signal model and speech enhancement framework in Section II, followed by the introduction of the baseline approaches in Section III. Next, we present the new DNN-based CEM approach in Section IV, and subsequently depict our experimental setup in Section V. Finally, we evaluate, discuss, and conclude the paper in Section VI and Section VII, respectively.

II. Signal Model and Speech Enhancement Framework

In this section we briefly introduce our signal model and the speech enhancement framework which is used for some preliminary experiments and for the evaluation.

A. Signal Model

We model the noisy time-domain microphone observation as

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

where $s(n)$ is the clean speech component, $d(n)$ the noise component, and n the discrete-time sample index. Both components are superimposed to obtain the microphone signal $y(n)$. We apply a K-point discrete Fourier transform (DFT) to obtain the corresponding frequency domain representation as

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

with frequency bin index $0 \leq k \leq K-1$ and frame index ℓ. Furthermore, we assume zero-mean speech and noise signals.

B. Speech Enhancement Framework

The speech enhancement framework we are utilizing is depicted in Fig. 1. It is starting with a preliminary noise reduction which is intended to process the noisy microphone signal $Y_\ell(k)$ in a first stage to provide a more suitable input signal $\bar{Y}_\ell(k)$ for the following processing. This first noise reduction stage is not restricted to any specific configuration, however, one should assure matched conditions with any potential training algorithms that might be required for subsequent processing stages. We use the minimum statistics (MS) [30] noise power estimator together with decision-directed (DD) [17] *a priori* SNR estimation and as spectral weighting rule the minimum mean squared error log-spectral amplitude estimator (MMSE-LSA) [31]. This stage is followed by a linear predictive coding (LPC) analysis block which subsequently allows for separate enhancement of the excitation signal $R_\ell(k)$ (upper path) and of the spectral envelope $H_\ell(k)$ (lower path). Both enhancement methods are explained further in more detail in Sections III-A and III-B, respectively. The enhanced signals' spectral amplitudes ($|\hat{R}_\ell(k)|$ or $|\hat{H}_\ell(k)|$) are then mixed with the respective counterpart ($|H_\ell(k)|$ or $|R_\ell(k)|$), to obtain an intermediate clean speech spectral amplitude estimate $|\hat{S}'_\ell(k)|$. It is important to note that—along with the noise power estimate $\hat{\sigma}_\ell^D(k)^2$ from the preliminary noise reduction—this estimate is only used as the numerator for the *a*

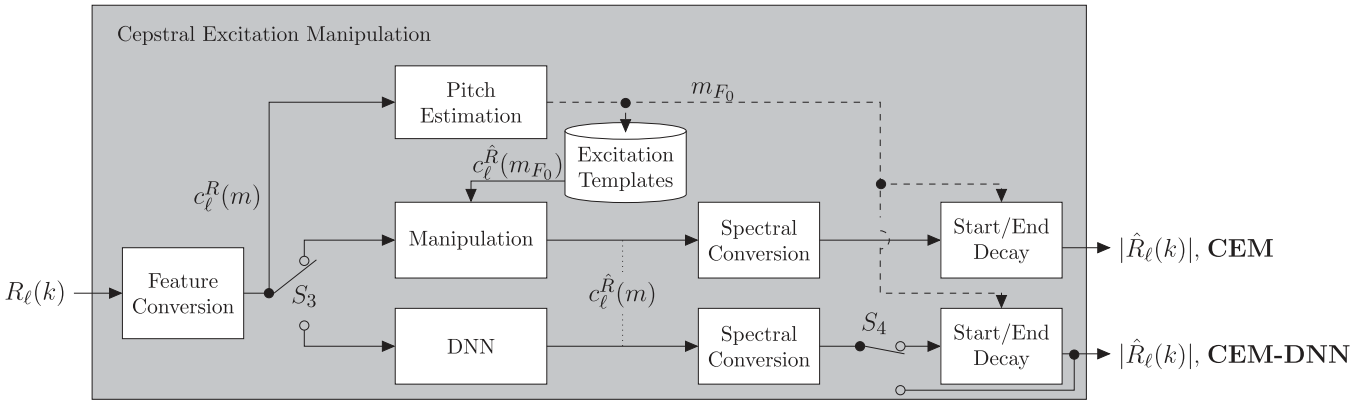

Fig. 2. Block diagram of the **CEM baseline approach** [7] and the **new proposed CEM-DNN approach** which is using a deep neural network (DNN). Here, switch S_3 determines the used algorithm. The **CEM-DNN** method is investigated with and without applied start/end decay which is determined by the position of switch S_4.

priori SNR estimate as

$$\hat{\xi}_\ell(k) = \frac{|\hat{S}'_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}. \tag{3}$$

It is then used jointly with the *a posteriori* SNR estimate $\hat{\gamma}_\ell(k)$ to calculate a spectral weighting rule

$$G_\ell(k) = f(\hat{\xi}_\ell(k), \hat{\gamma}_\ell(k)), \tag{4}$$

which is in our case again the MMSE-LSA estimator [31] for all traditional statistical-based approaches. Finally, the clean speech estimate $\hat{S}_\ell(k)$ is obtained by multiplying the real-valued gain function $G_\ell(k)$, which is limited to $G_{\min} = -15$ dB, with the microphone signal $Y_\ell(k)$ as

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

III. Baseline Approaches

As the proposed method builds upon the originally published CEM approach [7], we briefly revisit CEM as it has already shown to improve over common speech enhancement approaches. Among them are traditional statistical-based systems using e.g., the decision-directed *a priori* SNR estimator by Ephraim and Malah [17], the harmonic regeneration approach by Plapous *et al.* [19], and also the selective cepstro-temporal smoothing method proposed by Breithaupt *et al.* [20]. The superiority of CEM over these [7] is the reason why—for the sake of conciseness—we mostly concentrate on **CEM** as baseline in this work except for the final results, where we also present the results of a traditional speech enhancement system using the **DD** approach as *a priori* SNR estimator. As a more recent approach we also test against a DNN-based ideal ratio mask (**IRM**) solution. Furthermore, our recently proposed method [8], dealing with the enhancement of the spectral envelope, dubbed cepstral envelope estimation (**CEE**), is now also used as a baseline. It is the counterpart of the CEM approach and has shown to further improve CEM, when combined in a serial manner, where first CEE is applied followed by CEM. For more details we kindly refer to [8], where we also show that the baselines are able to compete with modern end-to-end speech enhancement techniques such as the ideal ratio mask [2], [5]. This serial combination is also used as further baseline, named **CEE → CEM**.

Both solo approaches, CEE and CEM, are depicted in Fig. 1, where switches S_1 and S_2, both in upper position, represent the CEM approach, and both in lower position, represent the CEE approach[1]. As can be seen in Fig. 1, both methods share a common pipeline up to the LPC analysis, where it branches to facilitate the enhancement of each component, excitation and envelope, separately. The use of the source-filter model allows to split the enhancement task into two sub-problems which are briefly revised as follows.

A. Cepstral Excitation Manipulation (CEM)

The baseline configuration of the **CEM** approach [7] is depicted in more detail in Fig. 2 with switch S_3 in upper position. The first block (Feature Conversion) represents a feature transformation from the spectral domain to the cepstral domain by applying a discrete cosine transform of type II (DCT-II), followed by a simple pitch estimation algorithm [32]. The quefrency bin index m_{F_0} corresponding to the pitch frequency is estimated by selecting the quefrency bin in a certain range of fundamental frequency-representing bins, that exposes the highest amplitude. Following, a pretrained clean speech excitation template $c_\ell^{\hat{R}}(m)$ that depends on the estimated fundamental frequency is selected from a storage and used further. The following processing aims to adjust the energy of the synthesized excitation signal by replacing the amplitude of the template's zeroth coefficient $c_\ell^{\hat{R}}(0)$ by the amplitude representing the energy of the preliminary enhanced residual signal $c_\ell^R(m)$ by

$$c_\ell^{\hat{R}}(0) = c_\ell^R(0). \tag{6}$$

[1]As a further option it is possible to apply CEM and CEE in parallel, when switch S_1 is in upper, and switch S_2 is in lower position. This parallel approach has been evaluated in [8] and shown to improve the noise attenuation. However, it also affects the speech component quality compared to the solo approaches CEM or CEE, and thus is disregarded here.

A further step to enhance the excitation signal is that the incoming amplitude of the quefrency bin that represents the fundamental frequency $c_\ell^R(m_{F_0})$ is overestimated by a factor $\alpha > 1$ in order to boost the harmonic structure and simultaneously lower the energy between the harmonics to obtain a higher noise attenuation. It is then also inserted into the template as

$$c_\ell^{\hat{R}}(m_{F_0}) = \alpha \cdot c_\ell^R(m_{F_0}). \tag{7}$$

After these manipulation steps, the cepstral vector is transformed back into the spectral domain by an inverse DCT-II, yielding the manipulated residual spectral amplitude $|\hat{R}_\ell(k)|$. By using a cepstral representation of the excitation signal, one is able to address and manipulate all harmonics in the signal's spectral representation at a single cepstral bin.

Employing the F_0 estimate, finally some start/end decay to the spectral representation is applied, as this ensures a somewhat more natural rise and decay of the harmonic structure which might have been corrupted by the manipulations or is erroneous in the templates itself. The start decay is a simple linear continuation of the rising edge for the first harmonic while the end decay is applied in the same manner to the last fully representable harmonic, but in this case to the declining edge. Both measures lead to an attenuation of spectral content prior to the first and after the last harmonic, where no speech content is expected (further details in [7]).

B. Cepstral Envelope Estimation (CEE)

The counterpart of CEM is the enhancement of the spectral envelope which has been extensively investigated in [8], dubbed cepstral envelope estimation (**CEE**). We will briefly introduce the optimal solution in the following. The general idea (see also [33]–[35]) is to find a mapping between the spectral envelope of the preliminary denoised signal and a linear combination of pretrained N-dimensional prototypes $\tilde{\boldsymbol{c}}_i^H = [\tilde{c}_i^H(1), \ldots, \tilde{c}_i^H(m), \ldots, \tilde{c}_i^H(N)]^{\mathrm{T}}$, obtained from clean speech recordings which are stored in a codebook $\mathcal{C} = \{\tilde{\boldsymbol{c}}_i^H\}$. The prototypes are indexed by $i \in \mathcal{N}_S = \{0, 1, 2, \ldots, N_S\}$, where $i = 0$ represents a prototype for non-speech frames. The advantages of a cepstral representation are used once more, with the difference that not the DCT-II is used, but the LPC coefficients $a_\ell(m)$ are transformed directly by the recursive formula from [36], to obtain the cepstral representation $c_\ell^H(m)$. This allows to work safely with the coefficients without risking any instabilities of the filter as would be the case when working on LPC coefficients directly. A codebook size of $N_S + 1 = 65$ has proven to be optimal with dimensionality $N = 10$ and a simple feedforward classification DNN consisting of six hidden layers and 58 nodes each. It was shown, that the sigmoid activation functions have lead to slightly higher accuracies than rectified linear units and a softmax output layer. The network's input is the cepstral representation $c_\ell^H(m)$ and the output can be understood as a probability distribution over the prototypes in the codebook as

$$\mathrm{P}(s_\ell = i | \boldsymbol{x} = \boldsymbol{o}_\ell). \tag{8}$$

Hereby, s_ℓ represents a hidden state which is a proxy for the unknown truth behind the observation, i.e., the true clean spectral envelope, while the corresponding observation is defined as $\boldsymbol{o}_\ell = \left[c_\ell^H(1), \ldots, c_\ell^H(N)\right]$. Having obtained the probability distribution, MMSE estimation is performed by

$$c_\ell^{\hat{H}}(m) = \sum_{i \in \mathcal{N}_S} \mathrm{P}(s_\ell = i | \boldsymbol{x} = \boldsymbol{o}_\ell) \cdot \tilde{c}_i^H(m), \tag{9}$$

and the estimated cepstral vector $\boldsymbol{c}_\ell^{\hat{H}}$ is converted back to the estimated envelope spectral amplitudes $|\hat{H}_\ell(k)|$ by applying an IDCT-II. Further details can be found in [8].

C. Decision-Directed Approach (DD)

Originally proposed by Ephraim and Malah in [31], the decision-directed (**DD**) approach is still considered as an important baseline. Even though the previously mentioned baselines already outperform the DD approach, many researchers are also interested to see improvement vs. a speech enhancement system using the DD *a priori* SNR estimator. We use the DD estimator with $\beta_{\mathrm{DD}} = 0.975$ and $\xi_{\min} = -15$ dB to prevent too many musical tones.

D. Ideal Ratio Mask (IRM)

As a more recent approach we also test against an IRM approach based on a feedforward DNN which is in line with [2], [5]. The network consists of three hidden layers with 1024 nodes each and rectified linear units as activation functions. The total amount of parameters is 2,364,545. We are using log-spectral amplitude input features and calculate the target gains for training as

$$G_\ell^{\mathrm{IRM}}(k) = \left(\frac{|S_\ell(k)|^2}{|S_\ell(k)|^2 + |D_\ell(k)|^2} \right)^\beta, \tag{10}$$

with $\beta = 1.0$. In fact, this spectral weighting rule ($\beta = 1.0$) has been used for learning a lookup table with spectral gains based on the *a priori* and *a posteriori* SNR before [29].

IV. DNN-Supported Cepstral Excitation Manipulation

Incorporating the novel opportunities of deep learning we want to explore the potential of the CEM idea when it is realized by a neural network instead of the classical signal processing measures that have been applied until now (see Section III-A). We show both approaches in Fig. 2, where the classical baseline CEM is depicted in the upper path (switch S_3 in upper position) and the new proposed approach, dubbed **CEM-DNN**, in the lower path (switch S_3 in lower position). As further option a smooth start and end decay can be applied to the manipulated amplitude spectrum of the residual signal (S_4 in upper position), to ensure smooth transitions which was necessary for the template-based CEM approach. The start and end decay function still relies on the simple F_0 estimator proposed by [32], however, this is a less critical application compared to the former selection of templates based on the same estimate in state-of-the-art CEM. Following Fig. 2, the feature conversion block (see also Fig. 3) transforms the log-spectral amplitudes of the residual signal

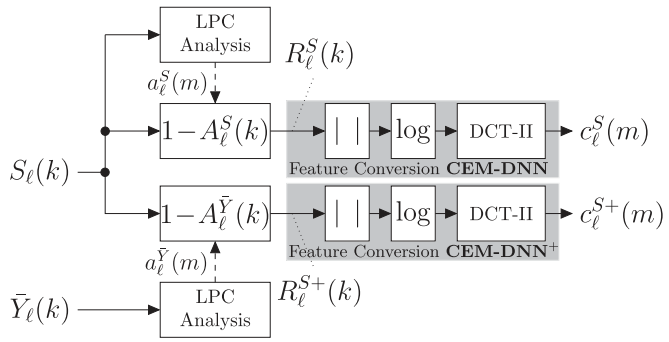

Fig. 3. Block diagram of the processing pipeline for two different representations of training targets for the **CEM-DNN** and **CEM-DNN**$^+$ approaches.

$R_\ell(k)$ into the cepstrum by applying the DCT-II, resulting in $c_\ell^R(m)$. When we apply normalization, all data is processed by bin-wise cepstral mean and variance normalization in order to remove potential channel mismatches. *Note that the core difference to the classical CEM approach is the replacement of the excitation templates $\tilde{\mathbf{c}}_i^H$ and MMSE estimation (9), as well as of the two manipulations (6) and (7) by a regression DNN.* In consequence, the core of CEM-DNN also does not need an F_0 estimator any more.

The output $c_\ell^{\hat{R}}(m)$ of the DNN is rescaled if necessary and subsequently transformed back into the spectral domain by the IDCT-II and optionally the start/end decay is applied. We finally obtain the estimated spectral amplitudes of the residual signal $|\hat{R}_\ell(k)|$. Rescaling of the DNN output is performed by using the mean and variance obtained from the respective data set. This translates to a practical system, as noise reduction is an uplink feature, which allows to calculate the required mean and variance of the signals after the preliminary noise reduction and LPC analysis for the input of the DNN, or during good SNR conditions to rescale the output of the DNN. In the following we will introduce our general setup for the DNN training and two different kinds of target representations.

A. DNN Training

The general setup of our DNN training process is based on the `KERAS` toolkit [37] together with the `TensorFlow` [38] backend. We normalize all *input features* by cepstral mean and variance normalization and in some cases we also normalize the target representation. The normalization is important to provide similar data ranges to the network which can ensure convergence and stability during training [39]. A similar argument holds for *target* normalization when regression networks are used: We explore the benefits of target normalization in more detail in Section VI-B2, however, it is not always applicable. Each input layer has the same amount of nodes as the input feature dimension $N = 256$. The subsequent N_H hidden layers each have N_N nodes. As we have experienced before, the difference between sigmoid and rectified linear units as activation function can be very marginal [8]. Since we did not encounter any problems with vanishing gradients so far, but obtained slightly better results with sigmoid activation functions, we decided to only investigate sigmoid activation functions in this case. The final output layer has also $N = 256$ nodes and uses linear activation functions since we only investigate regression DNNs. The parameters of the network, the biases and weights, are all initialized as proposed by Glorot *et al.* [40]. We employ the mean squared error (MSE) loss function in order to make the network learn the mapping between input and output representations. The training data is randomly accessed by the sequencing mechanism and provides batches of $L = 2048$ input and target frames at a time. For the gradient-based optimization we use the adaptive moment estimator (Adam) [41] with default parameterization, including a learning rate of $\eta = 0.001$. The networks are trained straight for 300 epochs from which the best model on some development set is selected and used further. In the following, we describe the two types of target representations and their advantages and disadvantages.

B. Target Representations

Since we aim to improve the excitation signal, the intuitive way is to simply extract excitation signals $R_\ell^S(k)$ from clean speech data $S_\ell(k)$ as targets for the training process of the DNN. The corresponding input features are the noisy, or in our case the already preliminary denoised, residual signals obtained from multiple simulated SNR and noise conditions. The pipeline for the target extraction is shown in Fig. 3 at the top. The frequency-domain representation of the clean speech data $S_\ell(k)$ is used for LPC analysis and subsequently filtered with the corresponding analysis filter $1 - A_\ell^S(k)$. The resulting spectral representation of the residual signal $R_\ell^S(k)$ is then subject to feature conversion, i.e., conversion to the log-amplitude spectrum, followed by the DCT-II to obtain the cepstral coefficients $c_\ell^S(m)$. The advantage of this target representation is that it is possible to obtain mean and variance data of $c_\ell^S(m)$ for the rescaling of the DNN output during inference (it is sufficient to collect these statistics from time to time during good SNR conditions), even in a practical application. Note that in such a practical implementation the input $S_\ell(k)$ for *both* the LPC analysis and the LPC analysis filtering in the upper path of Fig. 3 would have to be replaced by $\bar{Y}_\ell(k)$. In Fig. 1 it can be seen that the estimated amplitudes of the residual signal $|\hat{R}_\ell(k)|$ are mixed with the envelope of the preliminary denoised signal $|H_\ell(k)|$ (switches S_1, S_2 in upper positions). Thus, there will be still some mismatch between residual and envelope. We refer to the CEM method trained with these particular targets in the following as **CEM-DNN**.

Better targets for the training can be obtained by also considering the preliminary denoised signal's envelope. This is shown in Fig. 3 at the bottom, where the LPC coefficients are obtained from the preliminary denoised signal $\bar{Y}_\ell(k)$. The clean speech signal $S_\ell(k)$ is then filtered with the corresponding analysis filter $1 - A_\ell^{\bar{Y}}(k)$ which yields, after the usual feature conversion, the cepstral coefficients $c_\ell^{S+}(m)$ of our other target features. Those features allow, theoretically, the reconstruction of the clean speech signal even with a preliminary denoised signal's envelope during inference. However, the required mean and variance data of $c_\ell^{S+}(m)$ for the rescaling of the network's output can only be obtained in lab conditions, since the core idea of this approach consists of the discrimination between $S_\ell(k)$ and $\bar{Y}_\ell(k)$, and the use of both. This prohibits target

normalization in practice, or target normalization is done on some static precalculated mean and variance from, e.g., the training data. The corresponding CEM approach, using mean and variance obtained in lab conditions, is dubbed **CEM-DNN$^+$**.

V. Experimental Setup

In the following, we describe the used databases for the development process of our system and also the instrumental quality measures which are used for the final evaluation of the baselines and the proposed approach.

A. Databases

In order to ease comparison to our earlier works [8], we use the same database setup for training, development, and testing. The training and development sets are based on the TIMIT database [42], where the training set is used as training set and the test set of the TIMIT database as development set for our experiments. We finally report results on the NTT super wideband database [43] (only British and American English speakers) which serves as a test set and allows us to also report cross-database results. The clean databases are corrupted by noises from the QUT [44] and the ETSI [45] databases. Please note that all data is downsampled to 8 kHz for our experiments. Except for the male single voice distractor noise file from the ETSI database, all files are used. Among them we find, e.g., babble, road, car, office, aircraft, and also kitchen noise. Four noise files are reserved for a special test set with unseen noise files, which is important to show how well results of data-driven algorithms generalize to unseen data. We generate noisy data at 8 kHz sample rate for six SNR conditions, i.e., –5 dB to 20 dB in steps of 5 dB. The noise files are split up into non-overlapping sections, where 60% are used for training, 20% for development, and the remaining 20% for testing. Each file from the two speech databases is mixed with a random part of each noise file's respective section (four noise files held out for test with unseen noise files, as said above). To accomplish this, both clean speech signal and noise part, are level-adjusted according to ITU-T P.56 [46] and subsequently superimposed. In total we generate 6 (SNRs) $\times$ 53 (noise files) $=$ 318 conditions, represented by 318 $\times$ 4620 $=$ 1,469,160 (training set) and 318 $\times$ 1680 $=$ 534,240 (development set) noisy speech files[2]. Last, our framing setup is using a periodic square root Hann window, along with a frame shift of 50% and a frame length of $K = 256$ samples.

[2]This is a multitude of files that forces us to develop strategies to successfully cope with a huge amount of data for the training, and also the development process. Due to the large amount of data, i.e., the input features $c_\ell^R(m)$ and the targets $c_\ell^{S+}(m)$ for all 318 conditions, consuming together around 532 GB of disk space when stored as single-precision floating-point values, we decided to take two measures: First, we store all data as half-precision floating-point values resulting in a reduction to 266 GB and second, for our development process of the network structure, we optimize on the -5 dB SNR condition for all noise types only which reduces the amount of data further to roughly 44 GB. This allows us to be more flexible and we finally show that the loss-optimized topology found by single SNR condition training is also optimal for the multi-condition training which takes much more time.

B. Instrumental Quality Measures

As basis for our evaluation we employ the white-box approach [47], which allows us to assess the speech and noise component quality separately (see also ITU-T P.1100 Section 8 [48]). This is achieved by applying the gain function $G_\ell(k)$ not only to the microphone signal $Y_\ell(k)$, in order to obtain the enhanced signal $\hat{S}_\ell(k)$, but also to the separate components. This yields the *filtered* clean speech component $\tilde{S}_\ell(k) = S_\ell(k) \cdot G_\ell(k)$, and the *filtered* noise component $\tilde{D}_\ell(k) = D_\ell(k) \cdot G_\ell(k)$.

Both are subsequently transformed into the time domain by applying an inverse DFT followed by overlap-add synthesis, resulting in $\tilde{s}(n)$ and $\tilde{d}(n)$, respectively.

For the instrumental evaluation of the approaches, we use measures of two different categories in order to assess the amount of noise attenuation on the one hand and speech quality and intelligibility on the other hand. For the first, we use the segmental noise attenuation (NA_{seg}) measure [29] which can be obtained as

$$\text{NA}_{\text{seg}} = 10 \log_{10} \left[\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{NA}(\ell) \right], \tag{11}$$

with

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

The measure depicts the logarithmic average over the noise attenuation of all frames $\ell \in \mathcal{L}$. Each frame contains $N = 256$ samples and Δ compensates potential processing delay. A high value indicates good performance.

As additional measure to assess the SNR improvement on a global level we introduce the delta SNR which is calculated as

$$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}. \tag{12}$$

SNR_{out} represents the SNR of the *filtered* speech and noise component after processing and SNR_{in} the corresponding SNR of the clean speech and noise signals.

The speech quality of the *filtered* speech component $\tilde{s}(n)$ is measured by the PESQ score (mean opinion score, listening quality objective (MOS-LQO)) [26], [27] with $s(n)$ as the reference signal.

As fourth measure, we use the short-time objective intelligibility measure (STOI) [49] to rate the intelligibility of the enhanced speech signal $\hat{s}(n)$ compared to the clean speech signal $s(n)$. The closer the value is to unity, the better.

VI. Evaluation and Discussion

A. Oracle Experiments and Motivation

First of all, we conduct two oracle experiments which serve as motivation for our research. In Figs. 4 and 5, both oracle experiments show the performance of an *a priori* SNR estimator with different use of partial oracle knowledge, set in the context of the noise reduction framework as described in Section II-B. They use the same noise power estimate obtained by MS, along with an adjusted numerator as follows: The *oracle excitation* experiment (solid purple line, diamond markers) mixes the denoised

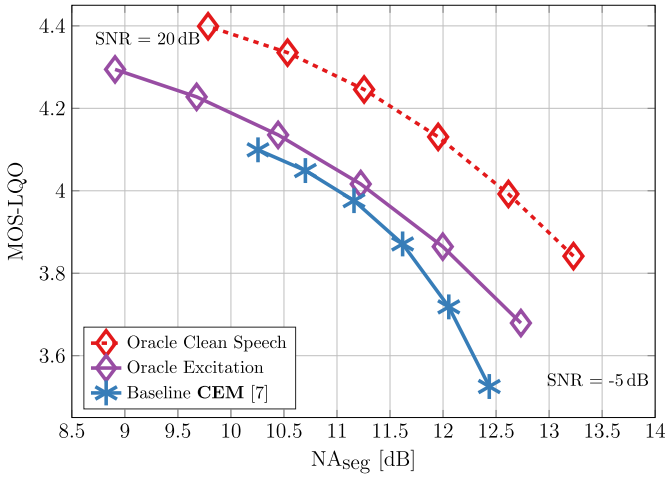

Fig. 4. Two oracle experiments showing the motivation and the unexhausted potential of the baseline **CEM** approach in terms of NA_{seg} and speech component quality measured by speech component **MOS-LQO**. All results are obtained on the **development set**.

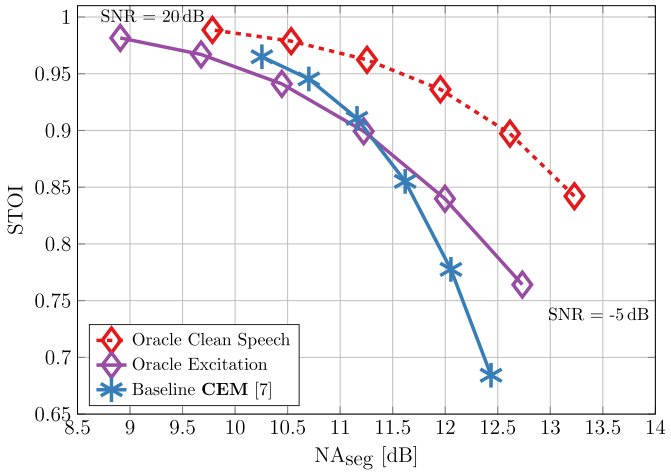

Fig. 5. Two oracle experiments showing the motivation and the unexhausted potential of the baseline **CEM** approach in terms of NA_{seg} and speech intelligibility measured by **STOI**. All results are obtained on the **development set**.

envelope $|H_\ell(k)|$ (see Fig. 1) with the oracle excitation signal obtained from clean speech. A more advanced oracle experiment (dashed red line, diamond markers) uses the *oracle clean speech* in the numerator for the *a priori* SNR, which assumes to know not only the clean speech excitation but also the corresponding clean speech envelope. The results show quite expected behavior, as with increasing oracle knowledge the potential gain in NA_{seg}, MOS-LQO, and also STOI increases, compared to the baseline **CEM** approach (solid blue line, asterisk markers). In the figures, each marker depicts a certain SNR condition from -5 dB at the bottom, in steps of 5 dB, up to 20 dB at the top. Using the oracle excitation signal shows less potential in terms of NA_{seg} compared to using the oracle clean speech signal. However, the potential gain in speech component quality and intelligibility (the vertical in both figures) is still worth pursuing, especially when considering the low-SNR conditions.

TABLE I
EVALUATION OF THE **MSE** LOSS FOR VARIOUS NETWORK TOPOLOGIES BASED ON THE **−5 DB SNR** CONDITION WITH $c_\ell^S(m)$ TARGETS FOR THE **DEVELOPMENT SET**

N_H	$N_N = 64$	$N_N = 128$	$N_N = 256$	$N_N = 512$	$N_N = 1024$
1	0.839	0.811	0.796	0.787	0.782
2	0.830	0.794	0.771	0.758	0.753
3	0.823	0.785	0.761	0.746	0.742
4	0.816	0.779	0.754	0.742	0.742
5	0.815	0.776	0.752	0.741	0.742
6	0.813	0.774	0.750	**0.739**	0.744

TABLE II
EVALUATION OF THE **MSE** LOSS FOR VARIOUS NETWORK TOPOLOGIES BASED ON **ALL SNR** CONDITIONS WITH $c_\ell^S(m)$ TARGETS FOR THE **DEVELOPMENT SET**

N_H	$N_N = 64$	$N_N = 128$	$N_N = 256$	$N_N = 512$	$N_N = 1024$
1	0.744	0.679	0.643	0.654	0.648
2	0.732	0.661	0.632	0.622	0.615
3	0.726	0.654	0.631	0.608	0.604
4	0.721	0.648	0.632	0.603	**0.600**
5	0.719	0.645	0.633	0.601	0.601
6	0.718	0.643	0.636	**0.600**	0.603

B. Cepstral Excitation Manipulation With DNN

In order to tap the potential of the cepstral excitation manipulation approach we decide to integrate a regression DNN. We briefly scanned on the development set through various parameters and ended up with the configuration as given in Section IV-A, as results stayed quite comparable. However, the topology of the network had quite some impact on the quality of the network's output. In Table I we show the MSE loss for several configurations of hidden layers N_H and their number of nodes N_N for the -5 dB SNR condition of the development set. It was necessary to make optimizations on a small set of data as the training process with all SNR conditions is quite time-consuming. In Table II (all SNR conditions), the MSE loss appears to be comparable for $N_N \in \{512, 1024\}$, which is natural since due to mean and variance normalization of the targets the number range of the loss also decreases. Solving the tie in Table II, we feel comfortable to put focus on the -5 dB condition (Table I) and decide for a configuration of $N_H = 6$ and $N_N = 512$ resulting in a total amount of 1,576,192 parameters. It might be possible that with increasing number of hidden layers the loss would drop further, which we expect to be rather marginal in this case. Note that the trainings have been conducted with $c_\ell^S(m)$ targets and we assume that the results translate also to $c_\ell^{S+}(m)$ targets without significant aberrations.

Now, we investigate the influence of the applied start and end decay as depicted in Fig. 2, the effects of target normalization, and the two different types of target representations as shown in Fig. 3.

1) Influence of Start and End Decay Function: In Figs. 6 and 7, we depict the **CEM-DNN** approach (square markers) which aims at estimating the clean excitation signal and the **CEM-DNN**$^+$ approach (plus markers) which aims at compensating also for the denoised spectral envelope, and thus to obtain the

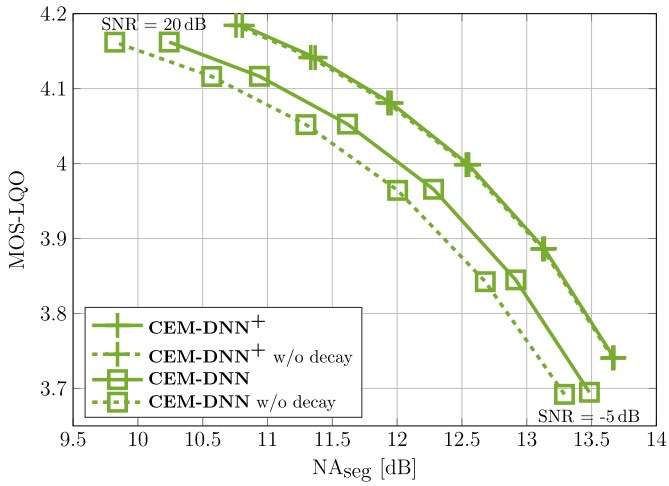

Fig. 6. The effect of applying start and end decay to either **CEM-DNN** or **CEM-DNN**$^+$ measured by NA_{seg} and speech component **MOS-LQO** on the **development set**.

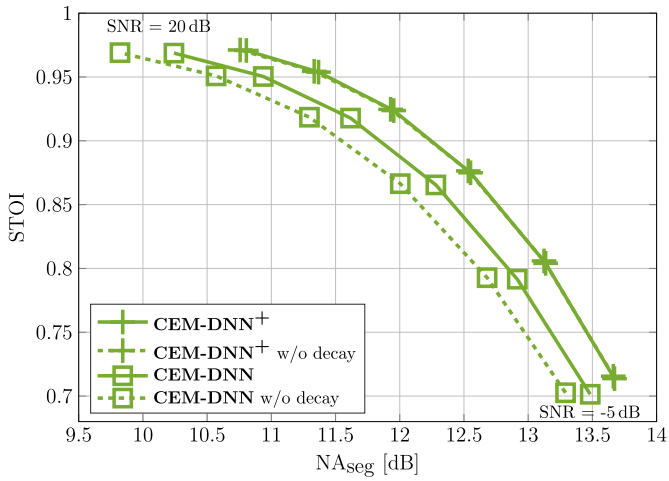

Fig. 7. The effect of applying start and end decay to either **CEM-DNN** or **CEM-DNN**$^+$ measured by NA_{seg} and **STOI** on the **development set**.

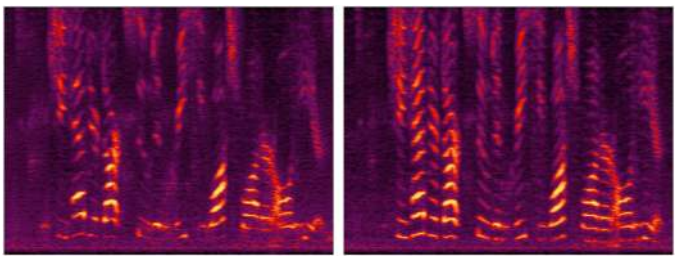

Fig. 8. **Spectrograms** of an enhanced microphone signal from the **development set** at 10 dB SNR with `CAFE-CAFE-1` noise processed by **CEM-DNN** trained **without** (left) and **with mean/variance normalization** (right) of the targets.

clean speech signal. Both approaches are depicted with applied start and end decay (solid green lines) and without (dashed green lines). The results show that the start and end decay has only an effect on **CEM-DNN** while the effect on **CEM-DNN**$^+$ is

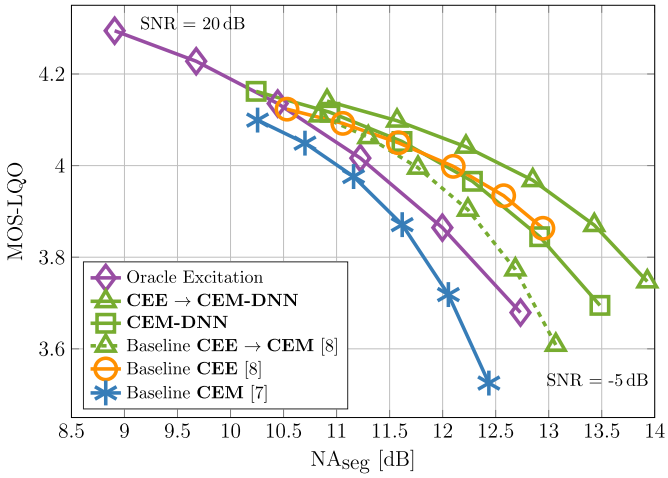

Fig. 9. Showing the performance (NA_{seg} and speech component **MOS-LQO**) on the **development set** for the baseline approaches, the new **CEM-DNN** method with applied decay, its serial concatenation with **CEE**, and the oracle experiment depicting the upper limit of the CEM approach.

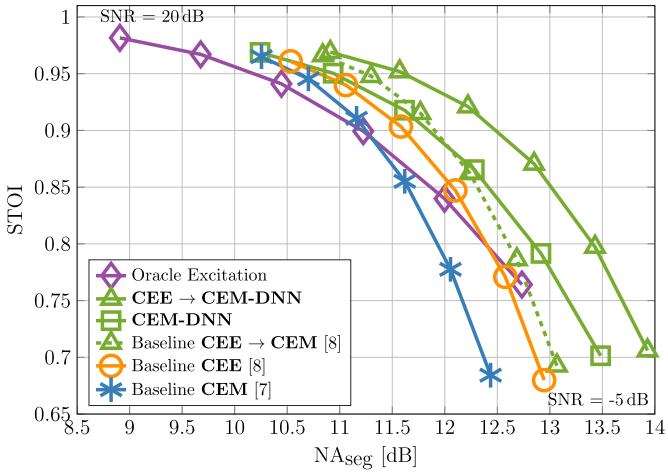

Fig. 10. Showing the performance (NA_{seg} and **STOI**) on the **development set** for the baseline approaches, the new **CEM-DNN** method with applied decay, its serial concatenation with **CEE**, and the oracle experiment depicting the upper limit of the CEM approach.

negligible. This is quite interesting, as it indicates that the application of the start and end decay might be naturally attributed to the envelope and is automatically compensated for by the DNN. Furthermore, the results show that **CEM-DNN** is able to benefit from the application of the start and end decay as NA_{seg} is consistently improved without significant impact on MOS-LQO and STOI. From here on all experiments are shown with applied start and end decay function.

2) Influence of Target Normalization: Next, we investigate the effect of target normalization in Fig. 8, showing the spectrograms of an enhanced microphone signal from the development set with `CAFE-CAFE-1` noise and 10 dB SNR condition. The microphone signal is then processed by **CEM-DNN** with applied start and end decay, once for a network trained without (left spectrogram), and once for a network trained with (right spectrogram) target normalization. The richness of the spectrogram

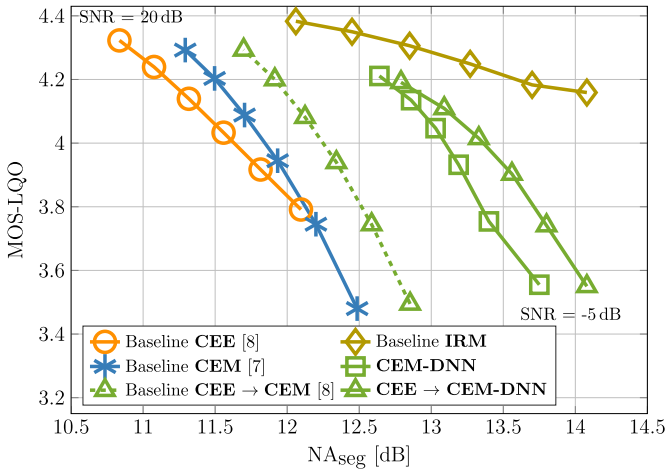

Fig. 11. Showing the performance (NA_{seg} and speech component **MOS-LQO**) on the **test set** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

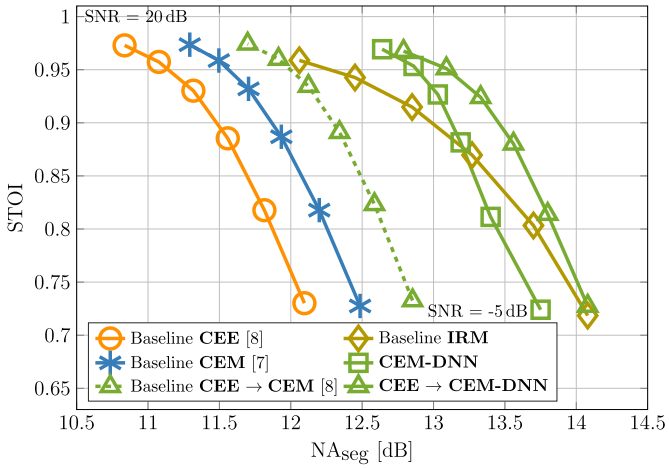

Fig. 12. Showing the performance (NA_{seg} and **STOI**) on the **test set** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

on the right shows the importance of target normalization which results in a much better preservation, especially in the lower frequency regions, of harmonic structures compared to the left spectrogram. This is a problem for the **CEM-DNN**$^+$ approach, as rescaling of the DNN output, as mentioned in Section IV-B, during inference would only be possible with pre-trained statistics, without any possibility of adaptation. Hence, we will continue only with **CEM-DNN**, with start and end decay, and with target normalization.

3) Results for the Development Set: In Figs. 9 and 10 we show the performance of the baselines **CEM** (solid blue line, asterisk markers), **CEE** (solid orange line, circle markers), and the serial concatenation of the former two approaches **CEE** → **CEM** (dashed green line, triangle markers) on the development set. Furthermore, we show the upper limit of the **CEM** approach by using the oracle excitation (solid purple line, diamond markers), the new approach **CEM-DNN** with start and end decay,

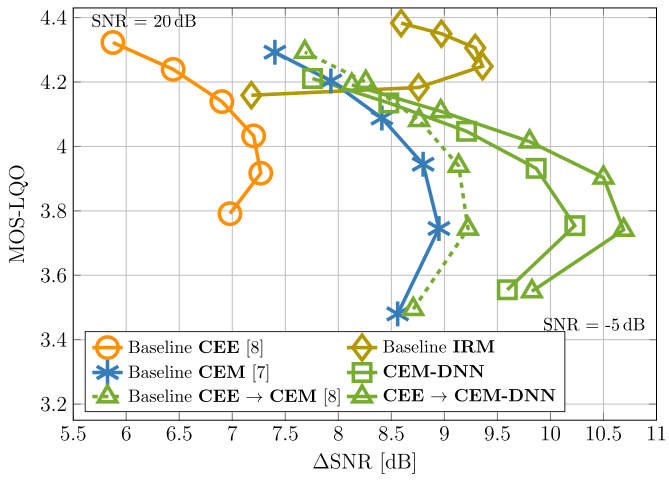

Fig. 13. Showing the performance (ΔSNR and speech component **MOS-LQO**) on the **test set** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

and its serial concatenation with the baseline **CEE**, labelled as **CEE** → **CEM-DNN** (solid green line, triangle markers). The noise attenuation of **CEM-DNN** improves over **CEM** by up to 1 dB for the −5 dB SNR condition, while increasing MOS-LQO by more than 0.1 points and also slightly improving STOI. This is an absolute improvement for the worst and most important SNR condition. The approach is even able to outperform **CEE** → **CEM** consistently up to and including the 5 dB SNR condition. The **CEE** approach still shows superior speech component quality measured by MOS-LQO, however, it is unable to remove noise between the harmonics and falls behind in most conditions for NA_{seg} and also slightly for STOI. Surprisingly, compared to the oracle excitation experiment, **CEM-DNN** obtains higher NA_{seg}, and in some cases comparable MOS-LQO, but does not match in speech intelligibility. In serial combination with the **CEE** approach, **CEE** → **CEM-DNN** yields further absolute improvement in terms of NA_{seg} by up to more than 0.5 dB with comparable MOS-LQO and STOI values.

4) Results for the Test Set: On the test set, which evaluates a different database, shown in Figs. 11–13, the behavior is quite similar. **CEM-DNN** and also **CEE** → **CEM-DNN** obtain higher NA_{seg} by more than 1 dB over their corresponding baseline. Thereby, MOS-LQO is slightly improving for the −5 dB SNR condition and STOI stays about the same. Only in high-SNR conditions the proposed approaches drop slightly in speech component quality, which is, however, uncritical as the quality still remains very high and STOI also reports no significant loss of intelligibility.

In addition to that, we also show the **IRM** baseline (solid sand line, diamond markers) which shows exceedingly high speech component quality. However, in terms of NA_{seg} and STOI the approach falls behind **CEE** → **CEM-DNN** with increasing SNR. In Fig. 13, for low and medium SNRs, the SNR improvement (ΔSNR) of the IRM approach falls far behind the proposed approach which outperforms all other approaches consistently. This also indicates that the attenuation characteristic of IRM is

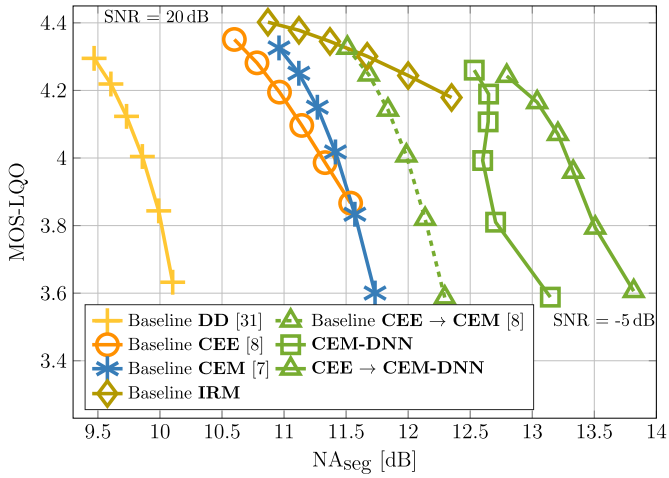

Fig. 14. Showing the performance (NA_{seg} and speech component **MOS-LQO**) on the **test set with unseen noise files** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

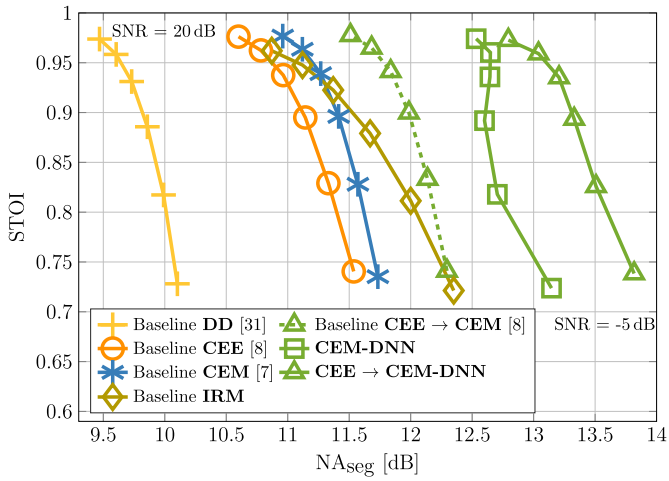

Fig. 15. Showing the performance (NA_{seg} and **STOI**) on the **test set with unseen noise files** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

more broadband and thus affecting speech and noise simultaneously, which explains the high MOS-LQO as PESQ is internally adjusting the level. Another issue with the **IRM** approach is the mentioned discontinuity problem as detailed in [3], and also observed in [8, Section VI-B2].

5) Results for the Test Set With Unseen Noise Files: The final evaluation on the test set with completely unseen noise files[3] during training is shown in Figs. 14–16. The results show that the performance transfers quite nicely to (also non-stationary) noise files that have not been seen during training, which is closest to a real-world scenario. Except for the already explained high speech component MOS-LQO, the proposed approach outperforms IRM clearly. Analyzing MOS-LQO and STOI at −5 dB

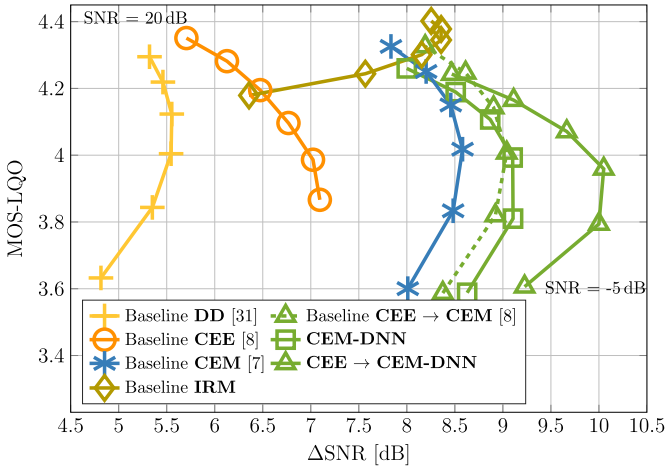

Fig. 16. Showing the performance (ΔSNR and speech component **MOS-LQO**) on the **test set with unseen noise files** for the baseline approaches, the new **CEM-DNN** method with applied decay, and its serial concatenation with **CEE**.

SNR in Figs. 14 and 15, we observe an 1.5 dB NA_{seg} advantage of **CEE** → **CEM-DNN** vs. **IRM**, which is not the case in Figs. 11 and 12 (seen noises). This shows that baseline **IRM** generalizes not as good w.r.t. background noises. Compared to the respective baselines, there is no significant trade-off for speech intelligibility, and for the speech component quality only minor drawbacks in the high-SNR conditions, where the absolute speech component quality is already very high (above 4 MOS-LQO points).

For similar MOS-LQO (Fig. 14) and STOI (Fig. 15) we can also report a gain in NA_{seg} of approximately 1.5 dB for the −5 dB SNR condition (lowest marker) when comparing the new **CEM-DNN** to the baseline **CEM**, and also when comparing **CEE** → **CEM-DNN** to the baseline **CEE** → **CEM**. Compared to the **DD** approach, the proposed **CEE** → **CEM-DNN** obtains more than 3 dB NA_{seg} while maintaining comparable speech component quality and speech intelligibility[4].

VII. Conclusion

In this work we have investigated the application of a deep neural network (DNN) to the cepstral excitation manipulation (CEM) approach for *a priori* SNR estimation in a speech enhancement task. We have investigated two target representations, where one is not applicable to practical systems and the other shows convincing performance. Furthermore, we could verify the benefit of applying some start and end decay to the estimated residual signal and have shown the importance of target normalization. Thus, we have successfully enhanced the classical signal processing-based CEM approach by introducing a simple feedforward DNN which has lead to an *improvement on unseen and non-stationary noise files by up to 1.5 dB of segmental noise attenuation without sacrificing speech component quality and speech intelligibility*. Compared to a traditional speech

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

[4] Audio samples can be found under: https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2019-taslp-cem/

enhancement system with the decision-directed *a priori* SNR approach, an *improvement of even more than 3 dB segmental noise attenuation with comparable speech intelligibility is achieved on the same data.*

REFERENCES

[1] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.

[2] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 4390–4394.

[3] S.-W. Fu, T.-W. Wang, Y. Taso, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 9, pp. 1570–1584, Sep. 2018.

[4] S. Mirsamadi and I. Tashev, "Causal speech enhancement combining data-driven learning and suppression rule estimation," in *Proc. Interspeech*, San Francisco, CA, USA, Sep. 2016, pp. 2870–2874.

[5] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[6] F. Bao and W. H. Abdulla, "A new ratio mask representation for CASA-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 27, no. 1, pp. 7–19, Jan. 2019.

[7] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Instantaneous a priori SNR estimation by cepstral excitation manipulation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 8, pp. 1592–1605, Aug. 2017.

[8] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "DNN-supported speech enhancement with cepstral estimation of both excitation and envelope," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 12, pp. 2460–2474, Dec. 2018.

[9] S. Srinivasan and J. Samuelsson, "Speech enhancement using a-priori information," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1405–1408.

[10] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 163–176, Jan. 2006.

[11] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook-based Bayesian speech enhancement for nonstationary environments," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 2, pp. 441–452, Feb. 2007.

[12] T. Rosenkranz, "Noise codebook adaptation for codebook-based noise reduction," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Tel Aviv, Israel, Aug. 2010.

[13] Q. He, F. Bao, and C. Bao, "Multiplicative update of auto-regressive gains for codebook-based speech enhancement," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 3, pp. 457–468, Mar. 2017.

[14] Y. Yang and C. Bao, "DNN-based AR-wiener filtering for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 2901–2905.

[15] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.

[16] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC Press, 2007.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] L. Nahma, P. C. Yong, H. H. Dam, and S. Nordholm, "Convex combination framework for a priori SNR estimation in speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, New Orleans, LA, USA, Mar. 2017, pp. 4975–4979.

[24] J. Stahl and P. Mowlaee, "A simple and effective framework for a priori SNR estimation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 5644–5648.

[25] Z. Xu, S. Elshamy, and T. Fingscheidt, "A priori SNR estimation using discriminative non-negative matrix factorization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Calgary, AB, Canada, Apr. 2018, pp. 661–665.

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

[27] ITU, *Rec. P.862.1: Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*. International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Nov. 2003.

[28] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 4214–4217.

[29] 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.

[30] 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.

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

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

[33] J. Abel, M. Strake, and T. Fingscheidt, "Artificial bandwidth extension using deep neural networks for spectral envelope estimation," in *Proc. Int. Workshop Acoust. Echo Noise Control*, Xi'an, China, Sep. 2016, pp. 1–5.

[34] J. Abel and T. Fingscheidt, "A DNN regression approach to speech enhancement by artificial bandwidth extension," in *Proc. Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Dec. 2017, pp. 219–223.

[35] J. Abel and T. Fingscheidt, "Artificial speech bandwidth extension using deep neural networks for wideband spectral envelope estimation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 26, no. 1, pp. 71–83, Jan. 2018.

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

[37] F. Chollet *et al.*, Keras. (2015). Available: https://keras.io

[38] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software. [Online]. Available: https://www.tensorflow.org/

[39] C. M. Bishop, *Neural Networks for Pattern Recognition*. New York, NY, USA: Oxford University Press, Inc., 1995.

[40] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, vol. 9, Sardinia, Italy, May 2010, pp. 249–256.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014. [Online]. Available: http://arxiv.org/abs/1412.6980

[42] J. S. Garofolo *et al.*, *TIMIT aAcoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium (LDC), 1993.

[43] *Super Wideband Stereo Speech Database*. NTT Advanced Technology Corporation (NTT-AT).

[44] D. Dean, S. Sridharan, R. Vogt, and M. Mason, "The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010, pp. 3110–3113.

[45] 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, Sep. 2008.

[46] ITU, *Rec. P.56: Objective Measurement of Active Speech Level*. International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Dec. 2011.
[47] S. Gustafsson, R. Martin, and P. Vary, "On the optimization of speech enhancement systems using instrumental measures," in *Proc. Workshop Quality Assessment Speech Audio Image Commun.*, Darmstadt, Germany, Mar. 1996, pp. 36–40.
[48] ITU, *Rec. P.1100: Narrow-Band Hands-Free Communication in Motor Vehicles*. International Telecommunication Union, Telecommunication Standardization Sector (ITU-T), Jan. 2015.
[49] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.

Samy Elshamy received the B.Sc. degree in bioinformatics from the Friedrich-Schiller-Universität Jena, Germany, in 2011, and the M.Sc. degree in computer science from the Technische Universität 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, Germany.

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 the 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 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 Full Professor with the Institute for Communications Technology, Technische Universität Braunschweig, Germany. His research interests include speech and audio signal processing, enhancement, transmission, recognition, and instrumental quality measures.

Dr. Fingscheidt received several awards; among them are 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). In 2017, he co-authored the ITG award-winning publication, which is awarded only once in a life time. He has been a speaker of the Speech Acoustics Committee ITG AT3 since 2015. From 2008 to 2010, he was an Associate Editor for the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, and since 2011, he serves as a member of the IEEE Speech and Language Processing Technical Committee.