



DNN-Supported Speech Enhancement With Cepstral Estimation of Both Excitation and Envelope

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

Abstract—In this paper, we propose and compare various techniques for the estimation of clean spectral envelopes in noisy conditions. The source-filter model of human speech production is employed in combination with a hidden Markov model and/or a deep neural network approach to estimate clean envelope-representing coefficients in the cepstral domain. The cepstral estimators for speech spectral *envelope*-based noise reduction are both evaluated alone and also in combination with the recently introduced cepstral *excitation* manipulation (CEM) technique for *a priori* SNR estimation in a noise reduction framework. Relative to the classical MMSE short time spectral amplitude estimator, we obtain more than 2 dB higher noise attenuation, and relative to our recent CEM technique still 0.5 dB more, in both cases maintaining the quality of the speech component and obtaining considerable SNR improvement.

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

I. INTRODUCTION

SPEECH enhancement is an important field of research to aid the most natural way of communication for human beings. It comprises a variety of applications among them dereverberation, acoustic echo cancellation, artificial bandwidth extension, voice activity detection, speech presence probability estimation, and also noise reduction algorithms. Many of these applications require the estimation of an *a priori* SNR which we are investigating in this publication in the context of a noise reduction framework. Furthermore, we focus on approaches exploiting the cepstral domain, since its properties and advantages have gained considerable attention in the recent past. For each component of a common noise reduction scheme, such as noise power estimator, *a priori* SNR estimator, and spectral weighting rule, approaches have been developed that exploit the cepstral domain.

Manuscript received April 10, 2018; revised July 11, 2018 and August 21, 2018; accepted August 22, 2018. Date of publication August 31, 2018; date of current version September 14, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Andy W. H. Khong. (*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 was with NXP Software, 3001 Leuven, Belgium. He is now with the Internet Technology and Data Science Lab, Universiteit Gent-imec, 9052 Gent, Belgium (e-mail: nilesh.madhu@ugent.be).

W. Tirry is with NXP Software, 3001 Leuven, Belgium (e-mail: 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.2018.2867947

A *noise power estimation* algorithm based on minimum mean-square error (MMSE) estimation originally proposed in [1] has been further improved by Gerkmann *et al.* in [2] by a bias compensation which is required due to the necessity of estimating intermediate entities and therewith arising aberrations. Finally, the estimator has been augmented with temporal cepstrum smoothing [3] to enhance the speech power estimation, resulting in higher noise attenuation and thus improving the signal-to-noise ratio (SNR) [4].

A cepstral *a priori* SNR estimation technique has been proposed in [5], where the ability to easily address the fine structure and the envelope of a speech power spectral density separately, has been successfully exploited by applying different smoothing factors to the corresponding regions in the cepstral domain. Here, also an improved bias compensation [6] can be employed to further increase the performance.

We presented a cepstral excitation manipulation (CEM) method in [7], [8] that benefits from the direct modeling of the excitation signal. It is obtained via linear predictive coding (LPC) analysis and is subsequently replaced by a pitch-dependent excitation template which has been extracted from clean speech prior to its application. The approach successfully conquers an often reported issue with low-order models considering the shortcomings of noise suppression between the spectral harmonics [9], [10]. Furthermore, it renders means such as a voicing-sensitive postfilter, spectral mask, or speech presence probability estimation [9], [11], needless.

The last component of common noise reduction schemes, a *spectral weighting rule*, has been published by Breithaupt *et al.* in [3], performing smoothing in the cepstral domain to finally suppress the noise in a noisy signal. It allows to successfully suppress spectral outliers that otherwise would cause musical tones. It is to say that the general concept of temporal cepstrum smoothing has found various applications in speech enhancement.

The source-filter model of human speech production, separating a speech signal into its excitation and envelope has also found its applications in speech enhancement and is used in various degrees. The usage of speech and noise spectral shapes as *a priori* information for speech enhancement has been suggested by Srinivasan *et al.* in [12]–[14] and was developed further over time. Two low-rank codebooks trained on speech and noise spectral shapes are employed and a maximum-likelihood (ML) estimate of the corresponding parameters, two indices for the codebook entries and two corresponding gain factors, given the noisy observation, are calculated. The obtained parameters are

used to estimate the spectra of speech and noise, and are finally used in a Wiener filter to calculate spectral weighting gains. A continuation of this work has been published by Rosenkranz *et al.* where cepstral modeling is preferred over autoregressive (AR) modeling [15].

A non-negative matrix factorization approach representing a source-filter system where separate dictionaries for the excitation and the envelopes are trained is proposed in [16]. During test it also requires a preliminary denoised signal as the algorithm needs additional information from the signal such as a pitch estimate. It seems to be quite complex and it is not entirely clear, whether it is a real-time-capable algorithm or not, at least the used pitch estimator [17] indicates that it is not suitable for telephony applications.

The approaches exploiting cepstro-temporal smoothing [3] address the source-filter model in a fashion that the cepstral coefficients are assigned to either part of the model, depending on their position in the cepstrum, and are treated differently. Please note that this kind of model is not subject to specific constraints as in LPC analysis, where a given order strictly defines the number of poles in the z -transform of the model.

A hidden Markov model (HMM) has been used for speech enhancement in [18], [19]. Therein, two HMMs are utilized to model the clean speech and the noise signal separately by AR processes. In both references, a Wiener filter is derived by incorporating the estimated spectral prototypes provided by the HMMs. Different from [18], [19] decouples the gain factors from the prototypes and introduces an explicit modeling of the gains, leading to a consistent improvement. The low-order modeling of the speech HMM suggests that the approaches also suffer from the same incapability to model the fine structure appropriately and thus leaves room for improvement.

With deep learning strategies on the rise, deep neural networks (DNNs) also find their way into speech enhancement and allow for a very broad variety of applications. Approaches range from directly estimating clean time-domain signals from the noisy observation [20] to mapping functions for extracted noisy features to clean features [21]. Those DNN techniques have in common that they completely disregard statistical speech enhancement approaches, which still are commonly utilized, and instead highly depend on their training material. However, it is also possible to incorporate DNNs into well-known statistical frameworks as, e.g., it has been shown in [22] that incorporating DNNs into a common noise reduction scheme and replacing certain estimators of the system yields better results than employing a simple regression DNN to estimate the clean speech directly. Source-filter model approaches for artificial speech bandwidth extension have been very successfully shown to take profit from DNN envelope modeling with or even without HMM [23]–[25]. Also, as known from automatic speech recognition, Gaussian mixture models (GMMs) have been successfully replaced by DNNs for the acoustic modeling [26].

In this publication, we investigate various approaches for the estimation of clean spectral envelopes based on noisy observations. In all cases, the actual estimation domain is the real-valued cepstrum, since it advantageously allows the minimum mean squared error (MMSE) as cost function. We evaluate the

performance with respect to their application in *a priori* SNR estimation for a noise reduction framework, as we expect quite some benefit from envelope enhancement in this field. We start with utilizing a classical HMM driven by GMMs, which are subsequently replaced by a DNN. Furthermore, we also investigate the replacement of the entire HMM by a single DNN, which is providing posterior probabilities instead of likelihoods for the HMM, or the use of a DNN to estimate clean coefficients directly in regression mode. Finally, we combine the enhanced spectral envelope with our recently proposed CEM approach and incorporate it into the *a priori* SNR estimator from [8]. Note, however, that the field of application for the proposed spectral envelope estimators is not limited to these specific use cases.

In the following, we briefly introduce the signal model in Section II and revisit the cepstral *excitation* manipulation technique in Section III, followed by the investigation of our various methods for clean spectral envelope estimation based on a preliminary denoised signal in Section IV, where we gradually replace the HMM by a DNN. We describe our experimental setup in Section V-A and subsequently provide our simulations and evaluation in Section VI. We finally conclude this article in Section VII.

II. SIGNAL MODEL

To model the microphone signal $y(n)$ we assume that the speech signal $s(n)$ and the noise signal $d(n)$ are superimposed in the time domain as

$$y(n) = s(n) + d(n), \quad (1)$$

where n is the discrete-time sample index. A corresponding frequency domain representation after a K -point discrete Fourier transform (DFT) is obtained as

$$Y_\ell(k) = S_\ell(k) + D_\ell(k), \quad (2)$$

with frame index ℓ and frequency bin index $0 \leq k \leq K - 1$. Also, we assume statistical independence of the speech and noise signal, and that they have zero mean.

III. CEPSTRAL EXCITATION MANIPULATION BASELINE

We choose to utilize our recently published *a priori* SNR estimation and noise reduction framework [8], as its modularity allows us to easily integrate the proposed estimators and evaluate their performance either alone (dubbed “solo”) or in interaction with the CEM approach (called “duo”).

As depicted in Fig. 1, a preliminary noise reduction stage is employed to get a more suitable signal for the proposed methods. This first noise reduction stage is a common noise reduction scheme with a noise power estimator such as minimum statistics (MS) [27], improved minima controlled averaging [28], or a more recent approach, the unbiased MMSE-based estimator [2]. Subsequently, it is followed by an *a priori* SNR estimator, e.g., the decision-directed (DD) approach [29], and finally, a spectral weighting rule to calculate the real-valued gain factors in the frequency domain for the noise suppression. Some quite often used spectral weighting rules are, e.g., the

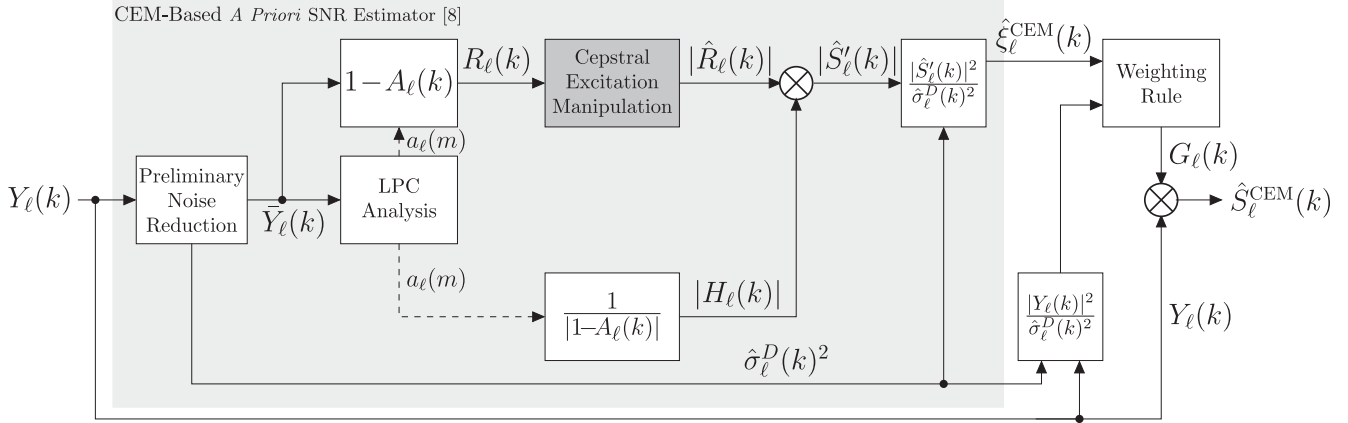


Fig. 1. High-level block diagram of the **cepstral excitation manipulation (CEM) noise reduction baseline**, incorporating a CEM-based *a priori* SNR estimator [8]. The proposed cepstral estimators for the spectral envelope are later on embedded into this approach (see Figs. 2 and 3).

MMSE log-spectral amplitude (MMSE-LSA) estimator [29], a more advanced gain function under a super-Gaussian assumption namely the super-Gaussian joint maximum a posteriori amplitude estimator [30], [31], or a simple Wiener filter [32]. In general, we do not restrict ourselves to a specific configuration, but have found a setup using MS noise power estimation along with the DD *a priori* SNR estimator and the MMSE-LSA spectral weighting rule as suitable for our method.

The preliminary denoised signal $\tilde{Y}_\ell(k)$ is subsequently subject to a source-filter decomposition block where LPC analysis is utilized to obtain an excitation signal $R_\ell(k)$ and the corresponding envelope $H_\ell(k)$, separately. They relate to the preliminary denoised signal as

$$\tilde{Y}_\ell(k) = R_\ell(k) \cdot H_\ell(k). \quad (3)$$

The CEM baseline as presented in [8] deals only with the enhancement of the excitation signal (Fig. 1, LPC analysis, upper path). As a first step of the CEM algorithm, the log-spectrum of the excitation signal is transformed into the cepstral domain by a discrete cosine transform of type II (DCT-II). Next, a (surprisingly) robust pitch estimation algorithm based on [33] provides the system with the corresponding cepstral pitch bin m_{F_0} by picking the maximum cepstral value within a queffrequency bin range representing typical pitch frequencies.

Consequently, a pitch bin-dependent cepstral excitation template $c_\ell^{\hat{R}}(m)$, with m being the cepstral bin index, is selected from a template codebook that has been trained on clean speech residual signals. The designated template vector is subject to two major manipulations: First, the template's cepstral energy coefficient $c_\ell^{\hat{R}}(0)$ is replaced by the corresponding value of the preliminary denoised signal's residual $c_\ell^R(0)$ as

$$c_\ell^{\hat{R}}(0) = c_\ell^R(0) \quad (4)$$

in order to receive a signal with a similar power level as the input signal. Second, the cepstral amplitude of the pitch bin $c_\ell^R(m_{F_0})$ is also transferred into the already power-adjusted excitation template and subsequently overestimated by a factor $\alpha > 1$ as

$$c_\ell^{\hat{R}}(m_{F_0}) = \alpha \cdot c_\ell^R(m_{F_0}). \quad (5)$$

Thereby, the harmonic structure of the excitation signal is overemphasized in both directions: The positive and also the negative half waves experience a boost or an attenuation, respectively. As a result, the algorithm is able to retain weak harmonics which might have been corrupted by the preliminary denoising stage and additionally, achieve a higher noise attenuation between the harmonics. Both characteristics are depicting the core features of the CEM algorithm. Until now, the manipulated template is transformed back into the spectral domain by an inverse DCT-II and used further with the spectral amplitudes of the preliminary denoised signal's envelope $|H_\ell(k)|$ to provide an improved clean speech amplitude estimate $|\hat{S}'_\ell(k)|$ by mixing the two components as

$$|\hat{S}'_\ell(k)| = |\hat{R}_\ell(k)| \cdot |H_\ell(k)|. \quad (6)$$

Finally, it is used as the numerator for a refined *a priori* SNR estimate along with the obtained noise power estimate from the preliminary noise reduction

$$\hat{\xi}_\ell^{\text{CEM}}(k) = \frac{|\hat{S}'_\ell(k)|^2}{\hat{\sigma}_\ell^D(k)^2}. \quad (7)$$

For more details about the CEM-based *a priori* SNR estimator, the interested reader may consult [8]. In *this* work, this estimator is embedded into a noise reduction framework as shown in Fig. 1, comprising also the computation of an *a posteriori* SNR

$$\gamma_\ell(k) = \frac{|Y_\ell(k)|^2}{\sigma_\ell^D(k)^2}, \quad (8)$$

as many gain functions $G_\ell(k)$ require either or both of the two SNRs for their calculation as

$$G_\ell(k) = f(\xi_\ell(k), \gamma_\ell(k)). \quad (9)$$

IV. CEPSTRAL ESTIMATION OF THE ENVELOPE

In this section we will now present our new methods of cepstral estimation to obtain speech spectral envelopes under noisy conditions. As outlined in Section I, we will embed these estimators into a noise reduction baseline which already performs cepstral estimation of the speech residual (see Fig. 1). Note that

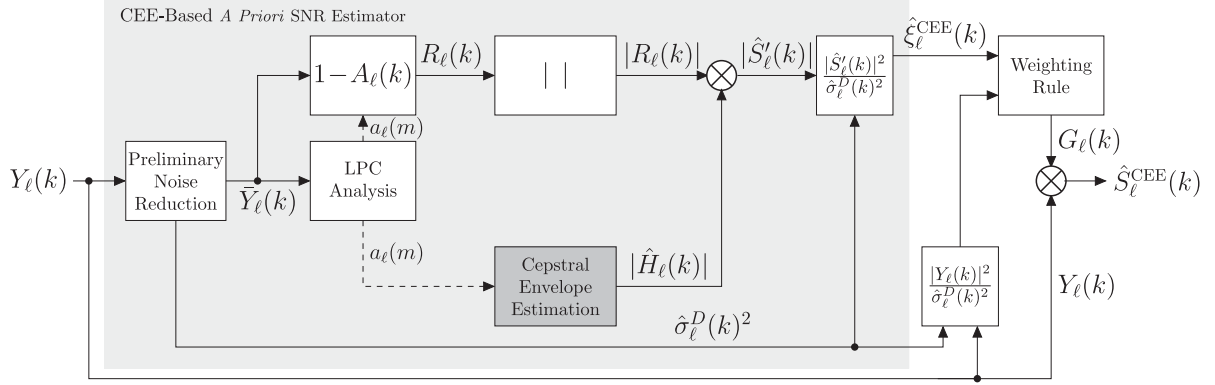


Fig. 2. High-level block diagram of the **proposed cepstral envelope estimation (CEE) noise reduction**, incorporating the new CEE-based *a priori* SNR estimator. For details of the CEE block please refer to Fig. 3.

this is only *one* of the many possibilities to employ our cepstral estimators of the speech spectral envelope. Our general approach advantageously uses a preliminary noise reduction, which provides an improved SNR for the subsequent envelope estimation. The spectral envelope of the preliminary denoised signal still suffers from distortions which tend to impede the speech quality, thus leaving room for further improvement. To our understanding it is reasonable to break down the noise reduction task for speech enhancement into smaller parts where possible. This is in line with divide-and-conquer strategies which have resulted in many useful solutions for various problems. As the production of speech can be modeled by two components, i.e., the source and the filter, it appears logical to attend each at a time which also has been done in, e.g., [13].

As a general framework we decided to employ a hidden Markov model (HMM) in order to estimate a clean spectral envelope, given the preliminary denoised observation. The motivation behind this is that we want to move from a bin-individual *a priori* SNR estimation (e.g., as the DD approach provides) to a more coherent and inter-frequency-dependent solution. Given the limited DFT length, this should be closer to the actual relationship between frequencies in speech, since they are not completely independent [34]. When dealing with spectral envelopes, this inter-frequency dependence becomes even more obvious. The application of a codebook that has been trained on clean speech spectral envelopes should be able to provide envelopes with a more realistic dependency between the *frequency* bins. In addition to that, we expect the HMM to capture the *temporal* context of envelopes which are usually smooth in transition.

The HMM in its classic form is using Gaussian mixture models (GMMs) to model the emission probabilities. As a second approach and along with the trend of deep learning we also employ a deep neural network (DNN) for classification to replace the GMMs. It has been shown in [26] that DNNs are capable of providing higher classification rates than GMMs, especially for acoustic models. A third variant we propose omits the HMM completely and solely uses the posterior distribution delivered by the classification DNN. As a fourth option we present a regression DNN in order to directly estimate clean envelope coefficients from the preliminary denoised observation.

In the following, we provide a generic recipe for our framework and the required training processes, while distinct parameters of our experimental setup will be provided in Section V-A.

A. Feature Conversion

As can be seen in Fig. 2, we also operate on the preliminary denoised signal $\tilde{Y}_\ell(k)$, which is decomposed into its source and filter by means of an LPC analysis. Up to this stage in the block diagram, both approaches CEM and also the now introduced cepstral envelope estimation (CEE) share the same processing structure. Now, the difference is that we operate on the LPC coefficients modeling the envelope (Fig. 2, LPC analysis, lower path) and not the excitation signal as before (Fig. 1, LPC analysis, upper path).

Since some training processes require the averaging of feature vectors, using the LPC coefficients directly could lead to instabilities. To obtain a representation of the envelope that has more suitable mathematical properties (Fig. 3, feature conversion block), we convert the N LPC coefficients to $N + 1$ cepstral coefficients by the following two formulae [35], which have been adjusted to our notation¹ of the LP analysis filter, here and also in (17), as (superscript H stands for the envelope filter)

$$c_\ell^H(m=0) = 0 = \log(P_p = 1). \quad (10)$$

The prediction error power P_p is set to an arbitrary fixed value to have envelopes with equal energy, allowing us to reduce the feature dimension to N since the zeroth coefficient is always zero. For $1 \leq m \leq N$ we calculate the cepstral coefficients recursively by

$$c_\ell^H(m) = a_\ell(m) + \frac{1}{m} \sum_{\mu=1}^{m-1} [(m-\mu) \cdot a_\ell(\mu) \cdot c_\ell^H(m-\mu)]. \quad (11)$$

We only compute the $N + 1$ non-redundant cepstral coefficients to maintain a small dimension, omitting $c_\ell^H(m=0)$ as explained and thus work with N features.

¹We denote the LPC analysis filter as $H_\ell(k) = 1 - A_\ell(k)$.

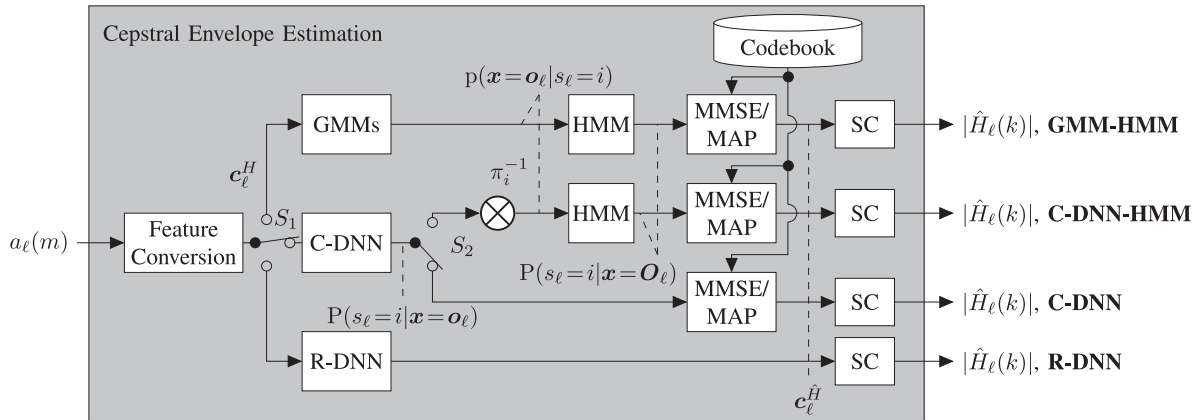


Fig. 3. Block diagram of the four different **proposed approaches for cepstral envelope estimation (CEE)** using either an HMM in combination with GMMs (first path) or alternatively a classification DNN to model the emissions (second path), or to model the posterior distribution of a classification DNN directly (third path). All these three approaches work with an LBG codebook for clean cepstral envelopes. Another option is a DNN trained as regressor (fourth path), estimating the clean cepstral coefficients directly from the input features. Since each approach yields enhanced cepstral coefficients, a required conversion to the spectral domain takes place in the spectral conversion (SC) boxes. Any of the four methods shown on the right side of the figure is determined by the setting of the switches S_1 and S_2 (shown: **C-DNN**).

In order to remove channel mismatches, we normalize all data in a bin-wise manner by cepstral mean subtraction with the mean obtained from the corresponding data set. In the following, we aim at estimating the corresponding clean envelope $c_\ell^{\hat{H}}(m)$ on basis of the preliminary denoised coefficients $c_\ell^H(m)$ from (11). Next, we provide our method to obtain a codebook for clean spectral envelopes, which is the backbone for the first three classification-based approaches as depicted in Fig. 3.

B. Codebook

The codebook $\mathcal{C} = \{\tilde{c}_i^H\}$ consists of N_S envelope templates obtained from clean speech. Each template is represented by an N -dimensional vector of cepstral coefficients $\tilde{c}_i^H = [\tilde{c}_i^H(1), \dots, \tilde{c}_i^H(m), \dots, \tilde{c}_i^H(N)]^T$. Each entry of the codebook is representing a hidden state of the HMM, which is indexed by $i \in \{1, 2, \dots, N_S\}$. We utilize the unsupervised Linde-Buzo-Gray (LBG) algorithm [36] to generate the codebook. We use an unsupervised method, since we are not interested in specific labels like, e.g., phonemes, but to obtain a good representation of many different envelopes. For training the codebook, any clean speech database is suitable. We use zero-mean clean speech envelope features (see Section IV-A) from frames identified by a simple energy threshold-based voice activity detection (VAD) as input to the LBG algorithm. The remainder of the clean speech training material is assigned an extra index $i = 0$, denoting non-speech frames, and is represented by an all-zero vector $\tilde{c}_0^H = \mathbf{0}$ in the codebook. Accordingly, there are $N_S + 1$ states indexed by $i, j \in \mathcal{S} = \{0, 1, \dots, N_S\}$. These states are to be estimated, e.g., by the HMM, which is introduced in the next subsection.

C. Hidden Markov Model

For the first two proposed approaches we will utilize a continuous density HMM to find a sequence of the hidden states s_1, s_2, \dots, s_ℓ , with $\lambda = \{\pi, \mathbf{A}, b_j(\mathbf{x})\}$ being the set of parameters defining the HMM. Here, $\pi = \{\pi_i\}$ denotes the

initial state probability vector, \mathbf{A} the state transition probability matrix with entries $a_{j,i} = P(s_\ell = i | s_{\ell-1} = j)$ representing the probability to go from state $j \in \mathcal{S}$ into state $i \in \mathcal{S}$, and $b_i(\mathbf{x})$ the corresponding continuous emission probability density function for each hidden state. An observation is defined as $\mathbf{o}_\ell = [c_\ell^H(1), \dots, c_\ell^H(N)]$, with $\mathbf{O}_\ell = \mathbf{o}_1, \mathbf{o}_2, \dots, \mathbf{o}_\ell$ being a sequence of observations. The posterior distribution of the state probabilities given all the observations up to the current frame ℓ , $P(s_\ell = i | \mathbf{O}_\ell)$, is obtained by applying the forward algorithm [37] as

$$\alpha_\ell(i) = b_i(\mathbf{x} = \mathbf{o}_\ell) \cdot \sum_{j \in \mathcal{S}} a_{j,i} \cdot \alpha_{\ell-1}(j), \quad (12)$$

followed by a normalization

$$P(s_\ell = i | \mathbf{O}_\ell) = \frac{\alpha_\ell(i)}{\sum_{j \in \mathcal{S}} \alpha_\ell(j)}. \quad (13)$$

The first frame is initialized with $\alpha_1(i) = \pi_i \cdot b_i(\mathbf{x} = \mathbf{o}_1)$. In order to stay capable of realtime processing, we use the forward algorithm instead of, e.g., the forward-backward algorithm which would calculate the posterior distribution with even higher precision.

Having obtained the posterior distribution, we calculate the MMSE estimate

$$c_\ell^{\hat{H}}(m) = \sum_{i \in \mathcal{S}} P(s_\ell = i | \mathbf{O}_\ell) \cdot \tilde{c}_i^H(m), \quad (14)$$

which represents a weighted average over all entries in the clean envelope codebook according to their respective probabilities. Alternatively, we use the maximum a posteriori (MAP) estimate

$$c_\ell^{\hat{H}}(m) = \tilde{c}_{i_\ell^*}^H(m) \quad (15)$$

with

$$i_\ell^* = \arg \max_{i \in \mathcal{S}} \alpha_\ell(i), \quad (16)$$

which simply selects the envelope with the highest posterior probability from the codebook. Here, the normalization from

(13) can be omitted, since it does not influence the $\arg \max$ operator. Note that for numerical stability we implemented our algorithms in the logarithmic domain.

The resulting zero-mean estimate of the clean envelope is required to maintain the channel properties, which is resolved by adding the corresponding cepstral mean. Finally, we calculate the spectral representation of the envelope as depicted by the SC blocks in Fig. 3. To accomplish this, we transform the estimated envelope back into N LPC coefficients by applying the following formula [35]:

$$\hat{a}_\ell(m) = c_\ell^{\hat{H}}(m) - \frac{1}{m} \sum_{\mu=0}^{m-1} \left[(m-\mu) \cdot c_\ell^{\hat{H}}(m-\mu) \cdot \hat{a}_\ell(\mu) \right] \quad (17)$$

for $1 \leq m \leq N$. Its spectral representation $|\hat{H}_\ell(k)|$ is received by first applying a K -point DFT to the LPC coefficients, padded with $K - N - 1$ zeros. This results in

$$\left(\hat{A}_\ell(k) \right)_{k=0}^{K-1} = \text{DFT} \{ (0, \hat{a}_\ell(1), \dots, \hat{a}_\ell(N), 0, \dots, 0) \}, \quad (18)$$

followed by

$$|\hat{H}_\ell(k)| = \frac{1}{|1 - \hat{A}_\ell(k)|}. \quad (19)$$

The initial state distribution vector π is assuming a uniform distribution (and is therefore not effective in Fig. 3), while the required state transition matrix \mathbf{A} is generated by counting transitions between the states in the clean training material followed by a normalization to calculate the conditional probabilities.

In the following, we will present two different methods to model the observations in order to obtain emission probabilities $b_i(\mathbf{x})$ by using either GMMs, or a classification DNN with prior division. We further investigate using the posterior distribution from a classification DNN directly, or a regression DNN (Section IV-E), directly estimating clean coefficients from the preliminary denoised observations. We then provide a generic description of the DNN training mechanisms in Section IV-F and will finally show, how these four CEE schemes can be combined with CEM, if desired.

D. HMM With GMM or Classification DNNs

Now, with the hidden states obtained from the LBG algorithm as described in Section IV-B, we generate quite some training material that represents typical observations for the HMM. To accomplish that, we simulate various SNR and noise conditions with the same clean speech data that has been used to retrieve the hidden states. This noisy speech data is subsequently processed by a preliminary noise reduction scheme running with the same parameterization as will be used for testing. It is followed by source-filter decomposition via LPC analysis, where only the envelope is used further for the GMM/DNN training. During this process it is important to keep track of the corresponding hidden state for each processed frame by knowing the quantization index i of its equivalent in the clean envelope codebook \mathcal{C} . This is required in order to obtain an assignment between a clean

envelope and all its corresponding denoised observations. With the aid of this information we are able to train models which represent the denoised observations for each of the states. We introduce GMMs and DNNs as such models in the following two subsections and also show how to replace the HMM completely by a DNN in the third.

1) *GMM-Based HMM (GMM-HMM)*: For each state i and its corresponding training material (representing observations) we use the expectation maximization (EM) algorithm [38] to train all parameters of a GMM with G modes, separately. The GMM is representing statistics of the preliminary denoised envelope observations which is later on mapped to a hidden state. In that fashion we receive the required models for the emission probabilities $b_i(\mathbf{x})$, $i \in \mathcal{S}$.

The observation probabilities for a certain input $b_i(\mathbf{x} = \mathbf{o}_\ell)$ are obtained by evaluating each GMM as follows

$$b_i(\mathbf{x} = \mathbf{o}_\ell) = \sum_{g \in \mathcal{G}} c_{i,g} \cdot \mathcal{N}(\mathbf{x} = \mathbf{o}_\ell; \boldsymbol{\mu}_{i,g}, \boldsymbol{\Sigma}_{i,g}), \quad (20)$$

with $g \in \mathcal{G} = \{1, \dots, G\}$ being the mode index, weights $c_{i,g}$ constrained to $\sum_{g \in \mathcal{G}} c_{i,g} = 1$, $\boldsymbol{\mu}_{i,g}$ as mean vectors, and $\boldsymbol{\Sigma}_{i,g}$ being the (in our case diagonal) covariance matrix for each corresponding mode g and state i . It plugs directly into (12) and is representing the GMM block in the upper path (S_1 in upper position) in Fig. 3.

2) *DNN-Based HMM (C-DNN-HMM)*: An alternative to GMMs as observation models is a feedforward DNN trained as classifier. The output of the classification DNN, the posterior probabilities for each of the hidden states given the current observation, is defined as $P(s_\ell = i | \mathbf{x} = \mathbf{o}_\ell)$. To use the output of the DNN in the HMM framework (12) (Fig. 3, second path from top, S_1 in center position, C-DNN block, and S_2 in upper position) we actually need to divide it by the prior state probability to obtain the likelihood as

$$b_i(\mathbf{x} = \mathbf{o}_\ell) = p(\mathbf{x} = \mathbf{o}_\ell | s_\ell = i) \propto \frac{P(s_\ell = i | \mathbf{x} = \mathbf{o}_\ell)}{P(s = i)}, \quad (21)$$

with $P(s = i) = \pi_i$. We omit the evidence $p(\mathbf{x})$ as it has only a normalizing function.

3) *DNN Without HMM (C-DNN)*: A further option to obtain posterior state probabilities is to use the output of a classification DNN directly (Fig. 3, third path from top, S_1 in center position, C-DNN block, and S_2 in lower position) and to omit the HMM framework, thereby losing the advantage of the temporal modeling from the HMM. Since we can understand the output of the DNN as $P(s_\ell = i | \mathbf{x} = \mathbf{o}_\ell)$, we can use it directly for either MMSE estimation as shown in (14) or MAP estimation in (16), where it is necessary to replace $\alpha_\ell(i)$ with the DNN output.

Next, we introduce a solution that is independent of a codebook or an HMM and estimates the clean envelope representing coefficients directly.

E. Regression DNN (R-DNN)

Instead of using DNNs as a classifier, it is also possible to directly estimate enhanced coefficients $c_\ell^{\hat{H}}(m)$ from a denoised input vector by means of regression. The output plugs directly

into the spectral conversion (SC) block in Fig. 3 and renders a codebook needless (Fig. 3, lowest path, S_1 in lower position, R-DNN block). In this particular case, the temporal context is also lost, unless the input layer of the DNN supports multiple input frames.

The coming subsection gives a brief overview of the training procedures required for the introduced DNN approaches.

F. DNN Training

To maintain comparability, we use the same zero-mean input features for the DNN as for the GMMs, and for regression also zero-mean targets. The number of nodes for the input layer is corresponding to the feature vector dimension N and the number of nodes for the output layer corresponding either to the amount of classes $N_S + 1$, or also to the feature dimension N (regression training). We understand hidden layers as every layer between the input and the output layer and their number is N_H , where each hidden layer has N_N nodes. The initialization of the network's parameter set, comprising the weights and biases, is done as proposed by Glorot *et al.* in [39]. In order to obtain posterior class probabilities we use the negative log-likelihood (NLL) error criterion during training with the backpropagation algorithm [40] and a softmax output layer. The difference for a regression-based DNN is mainly the final layer, which is a linear output layer in this case. Also, the used error criterion during the training is the mean squared error (MSE) instead of NLL. As activation functions in the other layers we employ sigmoid functions or rectified linear units (ReLUs). The latter are resolving the vanishing gradient issue [41], known to occur with sigmoid functions. After network initialization, the training material is randomly assigned to batches containing L input frames each. Then, according to the error criterion, the gradients of the loss function between the outputs of the network and the corresponding targets are calculated for each batch, and are subsequently backpropagated through the network. The deltas of the parameters are accumulated and finally the network's weights and biases are updated. We train each network with $L = 1024$ samples (frames) per batch and a fixed learning rate of $\eta = 0.001$ for 100 epochs. Finally, we select the model with the best performance on the development set for speech active frames (H_1), as experiments with adaptive learning rate decay have shown to perform only as good as but not better. Also, the investigation of L2 regularization did not lead to improvement, even worse, we could witness some configurations, where the networks deteriorate and classify every input as speech inactive (H_0).

Next, we provide instructions on how to apply or combine some of the introduced approaches.

G. Applications With CEM

The CEE scheme can be combined with CEM in two different ways: A parallel structure, where CEM and CEE are applied simultaneously, meaning that the CEE block from Fig. 2 is placed into the lower path of the LPC analysis in Fig. 1, or a serial structure where the systems from Fig. 1 and Fig. 2 are cascaded in either way. Here, cascading means that the output of the first system (being either $\hat{S}_\ell^{\text{CEM}}(k)$ or $\hat{S}_\ell^{\text{CEE}}(k)$) is used as

input for the LPC analysis block of the second system, thereby replacing $\tilde{Y}_\ell(k)$. Hence, the preliminary noise reduction of the second system is omitted and the noise power estimate $\sigma_\ell^D(k)^2$ of the first system is used throughout. The final gain function is also applied to the original microphone signal $Y_\ell(k)$.

V. EXPERIMENTAL SETUP

A. CEM and DD Baselines (CEM_{SI} and DD)

As we have already shown in [8], our baseline CEM algorithm outperforms several state of the art *a priori* SNR estimation algorithms. As motivated before, our experiments aim at further enhancing the CEM algorithm by employing our various envelope estimators, and compare the new approach to the speaker-independent CEM baseline (CEM_{SI}) and also the DD estimator (DD) which is parameterized with $\xi_{\min} = -15$ dB and $\beta_{\text{DD}} = 0.975$. For a detailed setup of the training for the CEM_{SI} approach, we kindly refer to [8].

B. Ideal Ratio Mask Baseline (IRM)

In addition, we simulate a data-driven baseline using a feed-forward neural network that predicts the ideal ratio mask (IRM). This baseline DNN has 2,364,545 parameters and is mostly in line with Wang's work ([42] and [20]). We use non-redundant amplitude features compressed by the natural logarithm as input features, while the IRM targets are calculated as

$$G_\ell^{\text{IRM}}(k) = \left(\frac{|S_\ell(k)|^2}{|S_\ell(k)|^2 + |D_\ell(k)|^2} \right)^\beta, \quad (22)$$

with $\beta = 1.0$. By interpreting the IRM as a gain function, we are able to integrate this baseline into our evaluation methodology (we require separately processed speech and noise *components*, as will be outlined at the end of this section). As some of our introduced approaches are based on the CEM_{SI} baseline, and thus indirectly on the DD baseline, we will first report the performance of our approaches w.r.t. the two baselines for our development process. However, for the final evaluation on the test data, we will also compare our approaches with the data-driven IRM baseline.

C. Databases and Preprocessing

We evaluate the algorithms in a noise reduction framework and analyze the performance in a total of 318 different conditions, embracing six different SNRs from -5 dB up to 20 dB in steps of 5 dB, and 53 different noise files where we use all 20 files from the QUT [43] database and 33 out of 38 files from the ETSI [44] database. Among them we find noise types such as babble, car, street, aircraft, train, work, and more. We leave out the male single voice distractor noise file and hold out four further noise files from the ETSI database for an extra test set with noise files which have not been seen during training. We split each noise file into three non-overlapping parts, where 60% are used for training, 20% for the development set, and another 20% for the test set. As clean speech databases we utilize the TIMIT [45] and also the NTT super wideband database [46] (American and British English only), both downsampled to 8 kHz.

The designated training set of the TIMIT database is used for training while the test set is used as development set and the NTT database is used for testing only. We decided to utilize the databases in that way since the training process requires a lot of data which the TIMIT database delivers and also we are able to show performance across different databases. For evaluation of the training and development set with our CEM_{SI} approach, we use one speaker-independent codebook based on the NTT database and for the test sets we use the speaker-independent codebooks as obtained in [8].

The various SNR conditions are obtained by measuring and adjusting the levels of the randomly selected noise portions and clean speech files after ITU-T P.56 [47], followed by their superposition. The framing (analysis and also overlap-add synthesis) is done with a periodic square root Hann window and a 50% frame shift, where one frame embraces $K = 256$ samples. The LPC analysis calculates $N=10$ LPC coefficients. Furthermore, we conduct the DNN training with the `Torch` toolkit [48] on CUDA-capable GPUs.

D. Instrumental Quality Assessment

For the quality assessment we employ the white-box approach [49], which means that we apply the calculated gains $G_\ell(k)$ not only to the microphone signal $Y_\ell(k)$ to obtain the clean speech estimate $\hat{S}_\ell(k)$, but also to the components $S_\ell(k)$ and $D_\ell(k)$, separately. We refer to the resulting entities after IDFT and overlap-add as the *filtered* clean speech component $\tilde{s}(n)$ and the *filtered* noise component $\tilde{d}(n)$, respectively. As instrumental measures we use the segmental noise attenuation (NA_{seg}) [50] which is calculated as

$$\text{NA}_{\text{seg}} = 10 \log_{10} \left[\frac{1}{|\mathcal{L}|} \sum_{\ell \in \mathcal{L}} \text{NA}(\ell) \right], \quad (23)$$

where

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

with $\ell \in \mathcal{L}$ defining a segment of $N = 256$ samples, Δ being the compensation term for potential delay due to filtering, and a normalizing factor $\frac{1}{|\mathcal{L}|}$, taking into account the number of all frames. Furthermore, we also evaluate the delta SNR as

$$\Delta \text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}. \quad (24)$$

Here, SNR_{in} depicts the SNR of the clean speech and noise component while SNR_{out} depicts the SNR of the *filtered* speech and noise components, after processing. This measure allows to draw conclusions on the actual improvement of the SNR, since a high noise attenuation might also affect the speech component.

We also employ the PESQ score (mean opinion score, listening quality objective (MOS-LQO)) [51], [52], on the *filtered* clean speech component $\tilde{s}(n)$ with $s(n)$ as reference. Thereby, we are able to evaluate the noise and also the speech *components* separately. We do *not* measure PESQ on the enhanced signal $\hat{s}(n)$, since PESQ has not been validated for artifacts caused by noise reduction techniques. In line with P.1100 [53, Sect. 8] and using [49] to obtain the processed clean speech component, we instead measure the distortion of the clean speech compo-

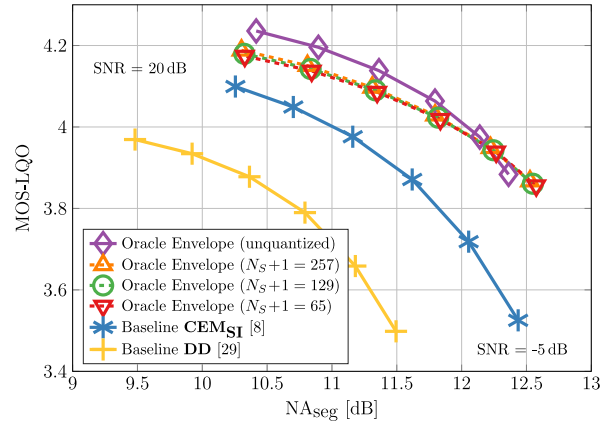


Fig. 4. Evaluation of the *speech component* MOS-LQO and segmental NA for all noise and SNR conditions of the unquantized and quantized **oracle experiments** and the **two baseline estimators** showing the potential of the proposed method on the **development set**.

nent, thereby being also compliant to the intended use case of P.862 [51]. Since PESQ is somewhat level-agnostic and thus not penalizing broadband attenuation of a signal, it is important to report the ΔSNR . This allows to draw conclusions on speech attenuation which would not be possible otherwise.

In order to assess the intelligibility of the enhanced speech, we employ the short-time objective intelligibility measure (STOI) [54]. STOI is an intrusive metric that is operating on the clean speech signal $s(n)$ which serves as a reference and the enhanced signal $\hat{s}(n)$. This metric provides values in the range $[0, 1]$, where high values represent high intelligibility.

VI. SIMULATIONS AND DISCUSSION

A. Solo: Cepstral Envelope Estimation (CEE)

Number of HMM States: At first, we perform two different oracle experiments in order to analyze the potential of our approach and to figure out how many states are providing good performance. In Fig. 4, we evaluate MOS-LQO by PESQ and also NA_{seg} , both measured on the separately filtered components. Here, each marker depicts a certain SNR condition, with -5 dB in the lower right and 20 dB in the upper left corner. The solid purple plot (with diamond markers) shows the performance of the proposed method when instead of the applied CEE (see Fig. 2, grey box), the oracle envelope from the clean speech is injected and mixed with the denoised residual signal (referred to as Oracle Envelope). Accordingly, this plot depicts the upper performance limit of the CEE technique in our noise reduction framework. Now, the first choice we need to make is on the amount of states the HMM should be able to estimate. Therefore, we train three different codebooks (see Section IV-B) for $N_S \in \{64, 128, 256\}$ with the LBG algorithm [36] on the extracted envelopes of the TIMIT training set. Subsequently, we run our framework, again replacing the CEE block by quantizing the oracle envelopes obtained from the corresponding clean speech files with our trained LBG codebooks (three dashed lines, triangle and circle markers). Comparing both oracle experiments to the **DD** (solid yellow line, plus markers) and **CEM_{SI}** (solid blue line, asterisk markers) baselines, shows that there

TABLE I

ANALYSIS OF THE **GMM-HMM** APPROACH WITH $N_S + 1 = 65$ STATES, G BEING THE NUMBER OF MODES: **POSTERIOR STATE PROBABILITY ACCURACY** DELIVERED BY THE HMM. SPEECH ACTIVE (H_1) AND INACTIVE (H_0) FRAMES ARE EVALUATED SEPARATELY

G	$H_0 \cup H_1$			H_0			H_1		
	4	8	16	4	8	16	4	8	16
Train	0.4358	0.4421	0.4460	0.6376	0.6455	0.6433	0.3648	0.3705	0.3766
Dev	0.4286	0.4352	0.4376	0.6332	0.6454	0.6415	0.3587	0.3635	0.3680
Test	0.4993	0.5034	0.5014	0.6415	0.6474	0.6397	0.3419	0.3440	0.3484

is good potential of the approach, especially in terms of speech component quality. One can also see that the quantization causes a slightly higher NA_{seg} in the lower SNR conditions compared to the oracle envelope, where in the other SNR conditions it is more a loss in speech component quality only. The three dashed lines representing the different quantization levels show a very similar performance with only a slight preference for the larger codebooks. However, since it is only a marginal benefit, we decide to use $N_S + 1 = 64 + 1$, as the trade-off between lower complexity and higher quality clearly favors the former in this case.

Number of GMM Modes: Next, we investigate the number of modes G which represent the denoised observations. Therefore, we train GMMs with $G \in \{4, 8, 16\}$ and evaluate the posterior state probabilities of the HMM by measuring the accuracy. The results are shown in Table I and are depicted for speech active (H_1), speech inactive (H_0), and both kinds of frames together ($H_0 \cup H_1$). The H_0/H_1 distinction is performed by a simple VAD on the clean speech material with a dynamic threshold which tests if a frame's energy is above the average frame energy of the corresponding file. The rationale behind this is that the prior distribution of the state representing speech inactive frames differs between the three sets, being roughly 25% for the training and development set, and 50% for the test set. This, if only regarding the accuracy of all frames jointly, would raise questions as to why the accuracy on the test set is higher than on the training and development set. Considering both classes separately gives a more consistent view on the performance, showing an expectedly higher accuracy with increasing number of modes on the speech active frames. The gain is rather small compared to the rising complexity with increasing G , making us comfortable with the choice of $G = 16$ (grey-shaded), delivering the best accuracy for speech active frames on the development set, without exploring the effects of more modes which we assume would lead to overfitting at some point and also to a lack of training data. Fortunately, this coincides with the best H_1 performance on the test set as well, which is not taken for granted.

GMM-HMM Envelope Estimation: Thus, having found a suitable configuration we evaluate the performance of the **GMM-HMM** approach with $N_S + 1 = 65$ states each represented by $G = 16$ modes with either MAP (16) or MMSE (14) estimation of the clean envelope in Fig. 5. On top, the unquantized and also quantized oracle experiments with $G = 16$ are shown. Compared to the **DD** baseline (solid yellow line, plus markers) the MAP approach (dashed green line, square markers) is able to show consistent improvement in terms of both measures, MOS-LQO and NA_{seg} . Especially the low-SNR conditions benefit from the enhanced envelope in terms of speech component

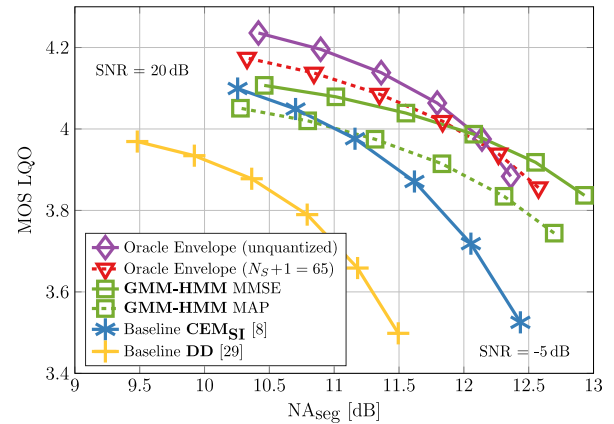


Fig. 5. Evaluation of the *speech component* MOS-LQO and segmental NA for all noise and SNR conditions of the **two optimized GMM-HMM approaches** using MAP and MMSE ($N_S + 1 = 65$, $G = 16$) compared to the two corresponding **oracle experiments** and **baseline estimators** showing the performance of the optimized **GMM-HMM** approaches on the **development set**.

TABLE II

ABERRATION OF PARAMETERS FOR SEVERAL NUMBERS OF HIDDEN LAYERS N_H WITH NUMBER OF NODES N_N IN EACH LAYER COMPARED TO THE DESIGNATED GMM CONFIGURATION WITH 21840 PARAMETERS

N_H	1	2	3	4	5	6
N_N	286	114	86	73	64	58
# Parameters	21801	21839	21565	21819	21569	21583
Aberration %	0.18	0.01	1.26	0.10	1.24	1.18

quality. The proposed approach also exceeds the **CEM_{S1}** baseline (solid blue line, asterisk markers) in the SNR conditions from -5 dB up to 10 dB quite clearly. Only the two best SNR conditions enable the CEM approach to obtain better speech component quality, which gives hope that a combination of both approaches might be able to mitigate the drawbacks of either method. When evaluated against the corresponding oracle envelope experiment (dashed red line, triangle markers) a more or less constant gap of around 0.05 MOS points remains. To circumvent the limitation of using a single entry of the codebook only, as done by the MAP estimation, we also calculate the MMSE estimate (solid green line, square markers), allowing us to consistently exceed the performance of the MAP approach by up to 0.09 MOS points for the -5 dB SNR condition. Even the oracle envelope experiment can be outperformed in terms of NA_{seg} , however, with a slightly lower MOS-LQO. This depicts nicely the benefit of the MMSE over the MAP estimate, being able to exploit the codebook space to a larger extent. The experiment using the unquantized oracle envelope performs clearly better than the **GMM-HMM** with MMSE estimation in the 20 dB to 10 dB SNR conditions, while in the remaining SNR conditions the MMSE approach obtains a much higher NA_{seg} which might be caused by a less accurate state estimation due to the SNR, being reflected by the lower MOS-LQO values.

C-DNN Envelope Estimation Approaches: The GMMs with $G = 16$ and $N_S + 1 = 65$ embrace a total of 21,840 parameters, which we target also for the training of DNNs to ensure a fair comparison. In Table II we depict several basic network configurations with up to six hidden layers, trying to keep a comparable amount of parameters as used for GMM training and we

TABLE III

EVALUATION OF VARIOUS C-DNN TRAININGS WITH COMPARABLE AMOUNT OF PARAMETERS RELATED TO THE BEST GMM CONFIGURATION IN TERMS OF THE POSTERIOR STATE PROBABILITY ACCURACY DELIVERED BY THE RESPECTIVE DNN. SPEECH ACTIVE (H_1) AND INACTIVE (H_0) FRAMES ARE EVALUATED SEPARATELY. THE EPOCH #E OF THE BEST PERFORMING NETWORK WITH RESPECT TO ACCURACY ON H_1 ON THE DEVELOPMENT SET IS ALSO REPORTED

N_H	N_N	Activation	#E	Training Set			Development Set			Test Set		
				$H_0 \cup H_1$	H_0	H_1	$H_0 \cup H_1$	H_0	H_1	$H_0 \cup H_1$	H_0	H_1
1	286	Sigmoid	55	0.5573	0.8788	0.4442	0.5452	0.8762	0.4321	0.6309	0.8529	0.3851
		ReLU	99	0.5552	0.8786	0.4415	0.5435	0.8768	0.4296	0.6312	0.8525	0.3863
2	114	Sigmoid	100	0.5596	0.8816	0.4463	0.5461	0.8775	0.4329	0.6337	0.8565	0.3871
		ReLU	6	0.5571	0.8813	0.4430	0.5439	0.8779	0.4299	0.6312	0.8514	0.3874
3	86	Sigmoid	68	0.5605	0.8803	0.4481	0.5467	0.8758	0.4344	0.6364	0.8592	0.3898
		ReLU	100	0.5588	0.8840	0.4444	0.5450	0.8801	0.4306	0.6351	0.8621	0.3838
4	73	Sigmoid	50	0.5612	0.8813	0.4485	0.5472	0.8757	0.4349	0.6385	0.8642	0.3887
		ReLU	55	0.5583	0.8863	0.4429	0.5451	0.8823	0.4299	0.6372	0.8668	0.3830
5	64	Sigmoid	89	0.5616	0.8809	0.4493	0.5472	0.8759	0.4349	0.6360	0.8616	0.3864
		ReLU	93	0.5581	0.8819	0.4443	0.5444	0.8770	0.4309	0.6358	0.8602	0.3873
6	58	Sigmoid	33	0.5608	0.8782	0.4492	0.5469	0.8734	0.4354	0.6367	0.8617	0.3876
		ReLU	41	0.5577	0.8828	0.4433	0.5445	0.8786	0.4304	0.6353	0.8607	0.3857

also depict the aberration of parameters in percent. Thereby, we make sure that we do not use more parameters than the GMM baseline does. With this setup we are able to analyze the influence of depth independently. Following, in Table III we present the posterior state accuracies of the various C-DNN configurations which we have examined. In general, there is to say that the differences between the configurations are rather small so that no network performs significantly better than any other. Judging from the development set, the networks with sigmoid activation obtain slightly better accuracies on speech active frames than the ones with ReLUs. Another observation is that with growing depth we can see a slight but steady increase of the accuracy on the H_1 frames of the development set when the sigmoid function is employed. For the subsequent C-DNN-HMM and also C-DNN approaches, we use the network with $N_H = 6$, $N_N = 58$, and sigmoid activation, as it performs best on speech active frames on the development set (grey-shaded). Note that it just does not match the best results on the test set which the network with $N_H = 3$ and $N_N = 86$ yields. When compared to the best GMM-HMM result in Table I (43.8%), the superiority of the DNN (54.7%) becomes obvious, as the accuracy gain on the development set ($H_0 \cup H_1$) is better than 10% absolute, and also on the speech active frames the accuracy increases by more than 6%. For the test set, the overall accuracy is more than 13% higher, while the gain for speech active frames of the test set melts down to about 4%.

R-DNN Envelope Estimation: Results of the second training process for the R-DNN are shown in Table IV. Again, we made sure that the amount of parameters relates closely to the best performing GMM-HMM configuration. The DNNs trained with sigmoid activation function slightly outperform the ones with ReLU activation function, as before. However, the latter tend to converge a bit faster for some topologies but with a higher loss. In general, the differences across all configurations are rather marginal. Nevertheless, we find the best configuration for $N_H = 6$ and $N_N = 58$ combined with the sigmoid activation function (grey-shaded). This is the same configuration as we found to be optimal for the C-DNN approaches. Also, this network shows only second best performance on the test set.

TABLE IV

EVALUATION OF VARIOUS R-DNN TRAININGS WITH COMPARABLE AMOUNT OF PARAMETERS RELATED TO THE BEST GMM CONFIGURATION IN TERMS OF THE MSE LOSS. THE EPOCH #E OF THE BEST PERFORMING NETWORK WITH RESPECT TO THE MINIMAL MSE LOSS ON THE DEVELOPMENT SET IS ALSO REPORTED

N_H	N_N	Activation	#E	Training Set	Development Set	Test Set
				MSE loss	MSE loss	MSE loss
1	286	Sigmoid	100	0.0480	0.0495	0.0506
		ReLU	76	0.0482	0.0498	0.0508
2	114	Sigmoid	100	0.0468	0.0485	0.0504
		ReLU	100	0.0471	0.0487	0.0507
3	86	Sigmoid	100	0.0465	0.0483	0.0515
		ReLU	95	0.0468	0.0485	0.0502
4	73	Sigmoid	100	0.0464	0.0481	0.0501
		ReLU	100	0.0468	0.0485	0.0501
5	64	Sigmoid	93	0.0464	0.0481	0.0496
		ReLU	50	0.0468	0.0485	0.0498
6	58	Sigmoid	91	0.0463	0.0480	0.0498
		ReLU	59	0.0468	0.0485	0.0499

All Approaches: Now, we evaluate the performance of the optimal networks in our system for the MAP estimation, as shown in Fig. 6. Comparing the GMM-HMM approach (dashed green line, square markers) to the C-DNN-HMM configuration (dashed orange line, triangle markers), results in an unchanged performance, which is surprising, since the accuracy of the C-DNN alone is significantly higher. A gain is seen, however, for the C-DNN (dashed orange line, circle markers), where the HMM is omitted and the posterior distribution of the network is used directly. An analysis of the state posterior distribution accuracy on the development set shows that the reported 54.7% ($H_0 \cup H_1$) and 43.5% (H_1) of the C-DNN (both Table III) correspond to only 45.2% ($H_0 \cup H_1$) and 38.0% (H_1) for the C-DNN-HMM approach, which is still higher by more than 1% compared to the GMM-HMM method (cf. Table I). However, this latter only small accuracy improvement explains the comparable performance of C-DNN-HMM and GMM-HMM in Fig. 6. The C-DNN consistently outperforms the two HMM-based systems in both quality dimensions by up to 0.05 MOS

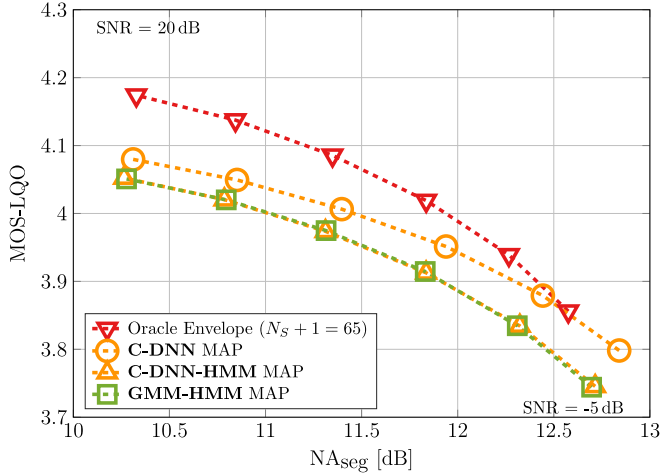


Fig. 6. Comparing the performance of the **GMM-HMM** system ($N_S + 1 = 65$, $G = 16$) and the various **DNN-supported approaches** with MAP estimation in terms of the speech component quality measured by MOS-LQO and NA_{seg} on the **development set**. The upper limit is depicted by the respective oracle experiment.

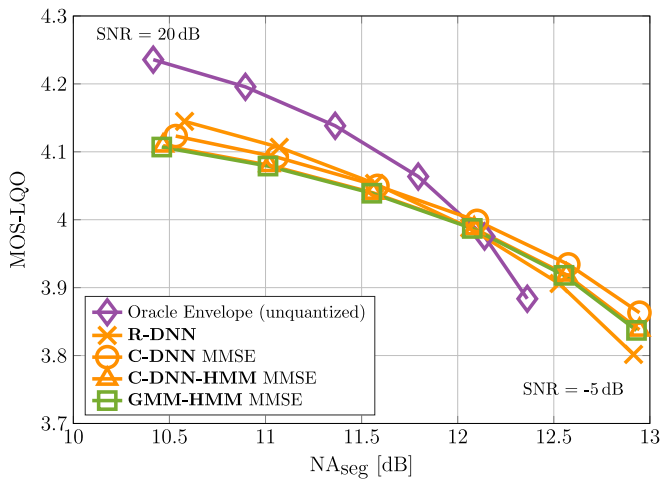


Fig. 7. Comparing the performance of the **GMM-HMM** system ($N_S + 1 = 65$, $G = 16$) and the various **DNN-supported approaches** with MMSE estimation in terms of the speech component quality measured by MOS-LQO and NA_{seg} on the **development set**. The upper limit is depicted by the respective oracle experiment.

points and 0.1 dB NA (-5 dB SNR condition), showing improved performance especially in the low-SNR conditions. This indicates that the HMM seems to be a limiting factor here, which could be caused by the temporal context, since it is the remaining factor that is able to overrule the network's decision.

The results for the MMSE estimation are reported in Fig. 7. Again, we can see that replacing the GMMs by a DNN (solid green line, square markers: **GMM-HMM** vs. solid orange line, triangle markers: **C-DNN-HMM**) has very little effect due to the limiting HMM. The performance of the **C-DNN** (solid orange line, circle markers) again shows consistent improvement over the HMM results, which indicates that the overall estimation of the posterior probability distribution is more accurate. Given the 10% accuracy improvement of **C-DNN** vs. **GMM-HMM**, and the 56.2% accuracy improvement of the oracle vs.

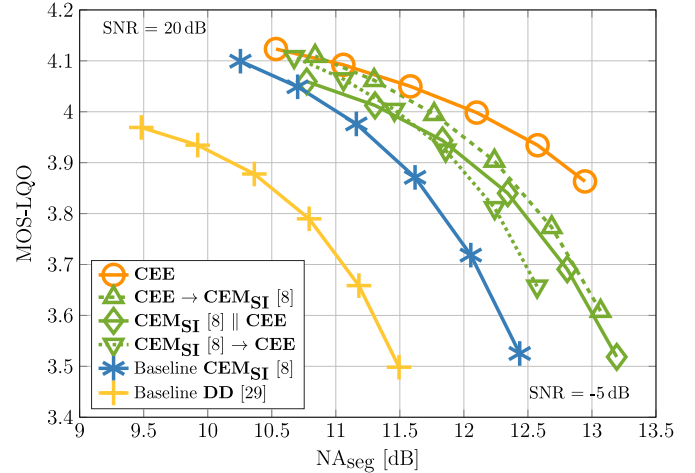


Fig. 8. Comparing the performance of the **CEE** system, the **baselines CEM_{S1}** and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech component quality measured by MOS-LQO and NA_{seg} on the **development set**.

GMM-HMM, **C-DNN** performs better than expected. This is visible, e.g., in SNR = -5 dB, where its MOS-LQO is about half way between **GMM-HMM** and oracle, while it exceeds the oracle NA_{seg} by more than 0.5 dB. Finally, the **R-DNN** (solid orange line, cross markers) shows an imbalanced behavior as it exceeds the performance of the **C-DNN** for the 15 and 20 dB SNR conditions but deteriorates with decreasing SNR. This results in the worst performance among the depicted methods for the two lowest SNR conditions. This is an interesting result as this shortcoming could not be observed for the classification DNNs. It could be due to the rather small amount of parameters, preventing the network to cover all SNR conditions equally as the regression task is more complex than classification. Consequently, we favor the **C-DNN** approach with MMSE estimation, as it performs best in the important low-SNR conditions. The approach still leaves space for improvement, especially for the higher SNR conditions, when compared to the oracle experiment.

B. Duo: CEM With Cepstral Envelope Estimation (CEE)

Having successfully identified the best performing envelope estimator, namely the **C-DNN** approach with MMSE estimation, which we will simply dub **CEE** in the following, we will now combine **CEE** with **CEM_{S1}** by replacing the preliminary denoised envelope in Fig. 1 (lower LPC analysis path, white box) with the proposed **C-DNN** cepstral envelope estimation method. This is referred to as parallel approach (symbol ||). Alternatively, we will also investigate using either **CEM_{S1}** or the **C-DNN** approach as preliminary noise reduction for the other, referred to as serial approaches (symbol →).

1) *Evaluation on the Development Set:* The results are depicted in Fig. 8, where the **CEM_{S1}** (solid blue line, asterisk markers) benefits especially in the important low-SNR conditions from incorporating the **CEE** (solid orange line, circle markers) in a parallel manner (solid green line, diamond markers) by obtaining a higher NA_{seg} while maintaining a

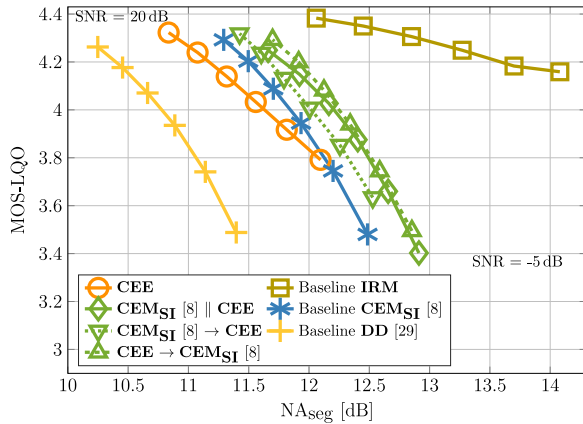


Fig. 9. Comparing the performance of the **CEE** system, the **baselines** CEM_{SI} , **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech component quality measured by MOS-LQO and NA_{seg} on the **test set**.

comparable speech component quality. Also the serial approaches (green lines, triangle markers) both outperform the CEM_{SI} baseline consistently in both quality dimensions, gaining up to 0.63 dB higher NA_{seg} and 0.13 MOS points. Applying CEM_{SI} first (dotted green line, inverted triangle markers) followed by the **CEE** yields a slightly higher speech component quality at the cost of a little less NA_{seg} compared to the other serial setup (dashed green line, triangle markers). The CEM_{SI} approaches, solo and duo, have one important advantage over the solo **CEE** approach: They are able to restore harmonics and to suppress noise between them, where the latter is a shortcoming of all approaches which only estimate the envelope. However, we expected a more consistent improvement by applying both techniques in parallel and suspect that some mismatch between the enhanced excitation and envelope could prevent further improvement, which could be subject to future research. This mismatch seems to be eased by the sequential application of both approaches, where we manipulate one component of the estimated clean speech amplitude spectrum at a time.

2) *Evaluation on the Test Set*: Until now, all results and optimizations have been analyzed and taken out on the development set. In Fig. 9 we report the test set performance of the three baseline approaches, **DD** (solid yellow line, plus markers), CEM_{SI} (solid blue line, asterisk markers), and **IRM** (solid sand line, square markers). We also report on our best cepstral envelope estimator **C-DNN** with MMSE estimation, i.e., **CEE** solo (solid orange line, circle markers), and also in conjunction (green lines) with the CEM_{SI} baseline. When the solo **CEE** approach is applied, a consistent improvement of the speech component quality over the **DD** and CEM_{SI} baselines is obtained, but the NA_{seg} now falls behind the CEM_{SI} method. This probably reflects the detriment of the **CEE** approach being a data-driven technique, since this was not the case on the development set. Interestingly, the two baseline approaches (**DD**, CEM_{SI}) yield lower PESQ scores on the development set than on the test set (compare Figs. 8 and 9). This is most likely due to the choice of two different databases which have quite different recording characteristics and settings. Thus, the one seems to be easier to be processed by noise reduction algorithms than the other.

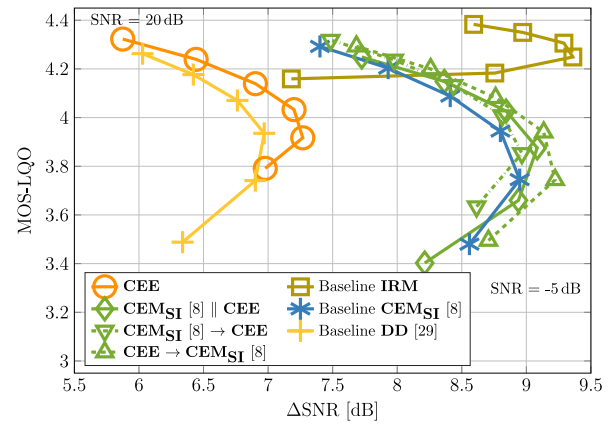


Fig. 10. Comparing the performance of the **CEE** system, the **baselines** CEM_{SI} , **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech component quality measured by MOS-LQO and ΔSNR on the **test set**.

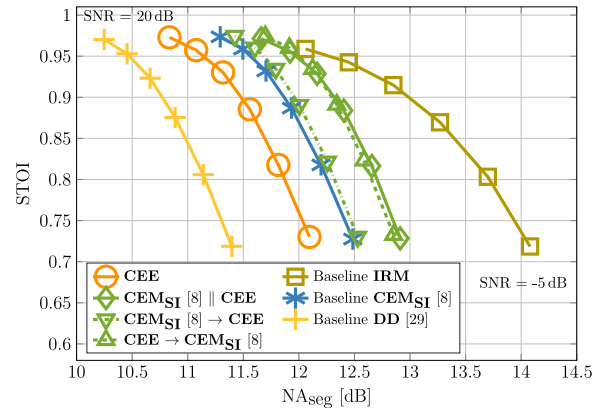


Fig. 11. Comparing the performance of the **CEE** system, the **baselines** CEM_{SI} , **IRM**, and **DD**, and the **parallel/serial combinations of both approaches** in terms of the speech intelligibility measured by **STOI** and NA_{seg} on the **test set**.

As the other approaches (green lines) are heavily influenced by the data-driven **CEE** approach, which has been trained on data stemming from the same database (but disjoint data sets) as the development set, the decreasing performance is quite expected when changing to a different database. However, the combination with CEM_{SI} seems to mitigate the drawback of the **CEE** approach caused by its data dependency to quite some extent. In parallel with the CEM_{SI} (solid green line, diamond markers) a gain of up to 0.4 dB NA_{seg} can be obtained, resulting in a slight shift of the trade-off point for speech component quality and noise attenuation compared to CEM_{SI} . Both serial approaches manage to consistently mitigate this drawback, where applying CEM_{SI} first (dotted green line, inverted triangle markers) is able to further improve CEM_{SI} by up to 0.15 MOS points at an additionally slightly higher noise attenuation. Alternatively, when applying the envelope enhancement first (dashed green line, triangle markers), the CEM_{SI} baseline can be improved by an average of 0.4 dB NA_{seg} , while maintaining a comparable speech component quality.

The data-driven **IRM** baseline shows a surprisingly high speech component quality that is exceeding the performance of all other approaches. However, when further analyzing the

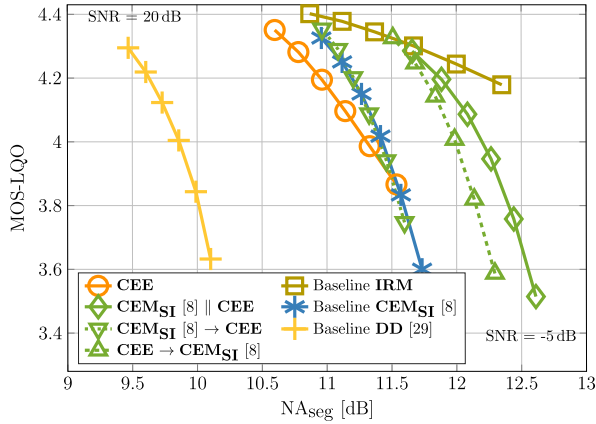


Fig. 12. Comparing the performance of the CEE system, the baselines CEM_{SI}, IRM, and DD, and the parallel/serial combinations of both approaches in terms of the speech component quality measured by MOS-LQO and NA_{seg} on the test set with unseen noise files.

Δ SNR as shown in Fig. 10, the approach shows the lowest Δ SNR improvement, especially in the important low-SNR conditions. This indicates that the IRM approach causes a broadband attenuation of noise *and* speech which is not penalized by PESQ as mentioned in Section V-D. Only in the (not so important) high-SNR conditions the IRM approach outperforms the other approaches also in terms of Δ SNR. A further issue with IRM is that the residual background noise shows a fluctuating temporal evolution and thus results in an unsettled subjective listening experience.² The IRM approach seems to be unable to generate coherent residual background noise which is not surprising, as the neural network has no recurrent modules or any memory which would allow it to produce coherent output w.r.t. previously processed frames. Even though it obtains high NA_{seg} results, the CEE approach also shows that the Δ SNR improvement is quite limited. Nonetheless, an improvement over the DD baseline, except for the 20 dB condition, is obtained. The proposed serial approach (CEE first) takes most profit from the combination of both methods and shows a small but consistent improvement over CEM_{SI}.

In Fig. 11 we present the intelligibility results measured with STOI for the different approaches. All methods perform similar on STOI, with IRM being best in NA_{seg}—with the known Δ SNR issue and residual noise quality issue² as discussed before.

Furthermore, we have investigated the performance of all the seven depicted approaches on the clean speech data of the test set without noise. Hence, it is not possible to report NA_{seg}, but PESQ scores are higher or equal than 4.43 MOS points and STOI is higher or equal than 0.981 for all approaches. This shows that the approaches do not significantly degrade speech quality or intelligibility in clean conditions.

Informal expert listening tests and spectrogram analyses³ have shown that the parallel and serial (CEE first) approaches

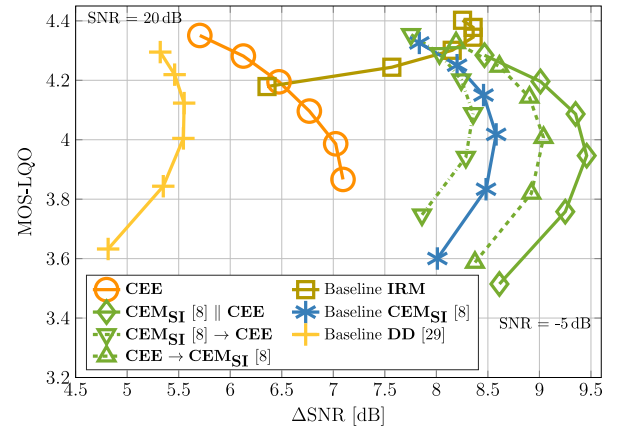


Fig. 13. Comparing the performance of the CEE system, the baselines CEM_{SI}, IRM, and DD, and the parallel/serial combinations of both approaches in terms of the speech component quality measured by MOS-LQO and Δ SNR on the test set with unseen noise files.

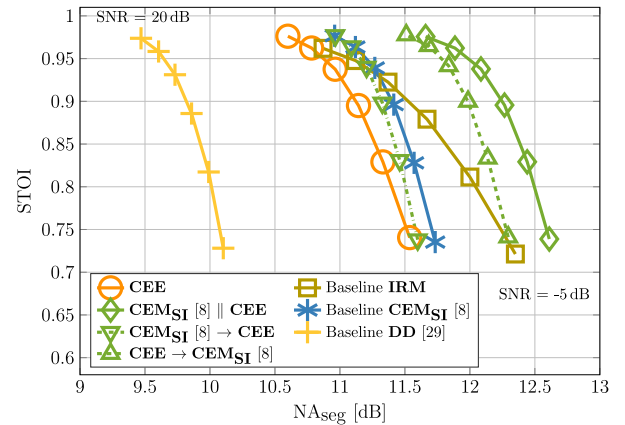


Fig. 14. Comparing the performance of the CEE system, the baselines CEM_{SI}, IRM, and DD, and the parallel/serial combinations of both approaches in terms of the speech intelligibility measured by STOI and NA_{seg} on the test set with unseen noise files.

result in a much smoother and more natural background noise, even in babble noise, owing to the introduced CEE method. The approaches also manage to reduce the noise between harmonics facilitated by the integration of the CEM_{SI} method.

3) *Evaluation on the Test Set with Unseen Noise Files:* Finally, in Fig. 12 we evaluate the performance on the test set with four unseen noise files, where three are quite non-stationary. The files⁴ are taken from the ETSI noise database [44]. Here, the solo CEE approach (solid orange line, circle markers) obtains up to 1.4 dB higher NA_{seg}, compared to the DD baseline and also improves the speech component quality significantly. The performance of the parallel approach (solid green line, diamond markers) is comparable to Fig. 9, where the NA_{seg} is increased at the cost of a lower speech component quality. This is also a general difference between Figs. 9 and 12, since the NA_{seg} in Fig. 12 is consistently lower and thus allows to obtain a higher PESQ score as the classical trade-off. This can be dedicated to

²Audio samples for the IRM baseline can be found under: <https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2018-taslp-cee/>

³Audio samples can be found under: <https://www.ifn.ing.tu-bs.de/en/ifn/sp/elshamy/2018-taslp-cee/>

⁴Fullsize_Car1_80Kmh, Outside_Traffic_Crossroads, Pub_Noise_Binaural_V2, Work_Noise_Office_Callcenter

the different noise types, as for Fig. 9 more stationary noise files have been included in the evaluation, which are naturally easier to process than non-stationary noise types, which are predominant in the data for Fig. 12.

The **IRM** baseline shows less improvement w.r.t. NA_{seg} compared to the test set with seen noise files. However, the speech component quality is still quite high, while showing clear detriments in the SNR improvement, as can be seen in Fig. 13. This indicates again that there is also quite some speech attenuation, which is also reflected in STOI (Fig. 14). Here, the **IRM** baseline is outperformed by our serial approach (**CEE** first) and also our parallel approach, where both also show convincing performance in Fig. 13 by improving the SNR consistently.

The serial approach with **CEM_{SI}** first (dotted green line, inverted triangle markers) also shows only limited improvement over the **CEM_{SI}** baseline (solid blue line, asterisk markers), mainly resulting in an improved speech component quality with a comparable NA_{seg} . However, when applying **CEE** first (dashed green line, triangle markers), we again consistently outperform the **CEM_{SI}** baseline by up to more than 0.5 dB NA_{seg} , while obtaining all its benefits even in non-stationary and unseen noise files. Thus, from the various schemes we have proposed in this paper, this is the strongest approach.

VII. CONCLUSIONS

We investigated several methods of spectral envelope estimation in the cepstral domain for *a priori* SNR estimation and evaluated their performance in a speech enhancement task with MMSE spectral amplitude estimation. Replacing a hidden Markov model by a deep neural network improves the state accuracy by more than 13% absolute. Evaluated on non-stationary and unseen noise files, the cepstral envelope estimation (**CEE**) approach alone shows significant improvement over the decision-directed (**DD**) estimator by up to 1.4 dB noise attenuation (**NA**), also significantly improving the speech component quality.

The combination with cepstral excitation manipulation (**CEM** with **CEE** first) provides a gain of 0.5 dB over **CEM** and of up to 2 dB over **DD** in terms of **NA**, without degrading the speech component quality or intelligibility. The proposed combination also obtains considerable SNR improvement over the baselines in the important low-SNR conditions.

There is still some room for improvement, as shown by the difference in the performance obtained with oracle envelopes and estimated envelopes. Future work will comprise the investigation of how to further reduce this gap, e.g., by more advanced topologies of neural networks which could lead to higher classification accuracies.

REFERENCES

- [1] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise PSD tracking with low complexity," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 4266–4269.
- [2] 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.
- [3] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Process. Lett.*, vol. 14, no. 12, pp. 1036–1039, Dec. 2007.
- [4] T. Gerkmann and R. C. Hendriks, "Improved MMSE-based noise PSD tracking using temporal cepstrum smoothing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, May 2012, pp. 105–108.
- [5] C. Breithaupt, T. Gerkmann, and R. Martin, "A novel a priori SNR estimation approach based on selective cepstro-temporal smoothing," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Las Vegas, NV, USA, Mar. 2008, pp. 4897–4900.
- [6] 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.
- [7] S. Elshamy, N. Madhu, W. Tirry, and T. Fingscheidt, "Two-stage speech enhancement with manipulation of the cepstral excitation," in *Proc. 5th Joint Workshop Hands-Free Speech Commun. Microphone Arrays*, San Francisco, CA, USA, Mar. 2017, pp. 106–110.
- [8] 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.
- [9] F. Deng and C. Bao, "Speech enhancement based on ar model parameters estimation," *Speech Commun.*, vol. 79, pp. 30–46, May 2016.
- [10] R. Rehr and T. Gerkmann, "A combination of pre-trained approaches and generic methods for an improved speech enhancement," in *Proc. ITG Conf. Speech Commun.*, Paderborn, Germany, Oct. 2016, pp. 51–55.
- [11] T. Yoshioka and T. Nakatani, "Speech enhancement based on log spectral envelope model and harmonicity-derived spectral mask, and its coupling with feature compensation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Prague, Czech Republic, May 2011, pp. 5064–5067.
- [12] S. Srinivasan and J. Samuelsson, "Speech enhancement using a-priori information," in *Proc. Eurospeech*, Geneva, Switzerland, Sep. 2003, pp. 1405–1408.
- [13] 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.
- [14] 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.
- [15] T. Rosenkranz, "Noise codebook adaptation for codebook-based noise reduction," in *Proc. 12th Int. Workshop Acoust. Echo Noise Control*, Tel Aviv, Israel, Aug. 2010.
- [16] U. Şimşekli, J. Le Roux, and J. R. Hershey, "Non-negative source-filter dynamical system for speech enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 6206–6210.
- [17] D. Talkin, "A robust algorithm for pitch tracking (RAPT)," in *Speech Coding and Synthesis*. New York, NY, USA: Elsevier Science, 1995, pp. 495–518.
- [18] Y. Ephraim, "A Bayesian estimation approach for speech enhancement using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 40, no. 4, pp. 725–735, Apr. 1992.
- [19] D. Y. Zhao and W. B. Kleijn, "HMM-based gain modeling for enhancement of speech in noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 3, pp. 882–892, Mar. 2007.
- [20] Y. Wang and D. L. Wang, "A deep neural network for time-domain signal reconstruction," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Brisbane, Australia, Apr. 2015, pp. 4390–4394.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] J. Abel and T. Fingscheidt, "A DNN regression approach to speech enhancement by artificial bandwidth extension," in *Proc. IEEE Workshop Appl. Signal Process. Audio Acoust.*, New Paltz, NY, USA, Dec. 2017, pp. 219–223.
- [25] 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.

- [26] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
- [27] 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.
- [28] 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.
- [29] 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.
- [30] 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.
- [31] B. Fodor and T. Fingscheidt, "MMSE speech enhancement under speech presence uncertainty assuming (generalized) gamma priors throughout," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, Mar. 2012, pp. 4033–4036.
- [32] P. Scalart and J. V. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Atlanta, GA, USA, May 1996, pp. 629–632.
- [33] A. M. Noll, "Cepstrum pitch determination," *J. Acoust. Soc. Amer.*, vol. 41, no. 2, pp. 293–309, Feb. 1967.
- [34] T. Fingscheidt, C. Beaugeant, and S. Suhadi, "Overcoming the statistical independence assumption w.r.t. frequency in speech enhancement," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, Philadelphia, PA, USA, Mar. 2005, pp. 1081–1084.
- [35] P. E. Papamichalis, *Practical Approaches to Speech Coding*. Upper Saddle River, NJ, USA: Prentice-Hall, 1987.
- [36] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. 28, no. 1, pp. 84–95, Jan. 1980.
- [37] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
- [38] G. McLachlan and D. Peel, *Finite Mixture Models*. Hoboken, NJ, USA: Wiley, 2000.
- [39] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, Sardinia, Italy, May 2010, vol. 9, pp. 249–256.
- [40] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. Hoboken, NJ, USA: Wiley-Interscience, 2000.
- [41] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Trans. Neural Netw.*, vol. 5, no. 2, pp. 157–166, Mar. 1994.
- [42] 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.
- [43] 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.
- [44] *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.
- [45] J. S. Garofolo *et al.*, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*. Philadelphia, PA, USA: Linguistic Data Consortium, 1993.
- [46] *Super Wideband Stereo Speech Database*. San Jose, CA, USA: NTT Advanced Technology Corporation.
- [47] *Objective Measurement of Active Speech Level*, ITU-T Rec. P.56, Dec. 2011.
- [48] R. Collobert, K. Kavukcuoglu, and C. Farabet, "Torch7: A MATLAB-like environment for machine learning," in *Proc. BigLearn, NIPS Workshop*, Sierra Nevada, Spain, Dec. 2011, pp. 1–6.
- [49] 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.
- [50] 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.
- [51] *Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs*, ITU-T Rec. P.862, Feb. 2001.
- [52] *Mapping Function for Transforming P.862 Raw Result Scores to MOS-LQO*, ITU-T Rec. P.862.1, Nov. 2003.
- [53] *Narrow-Band Hands-Free Communication in Motor Vehicles*, ITU-T Rec. P.1100, Jan. 2015.
- [54] 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.



Niles Madhu received the Dr.-Ing. degree from the Faculty of Electrical Engineering and Information Sciences, Ruhr-Universität Bochum, Bochum, Germany, in 2009. Following this he received a Marie-Curie experienced researcher 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 and biomedical signal analysis. From 2011 to 2017, he was with NXP Semiconductors, Belgium, where he held the position of a Principal Scientist within the product line Voice and Audio Solutions. During this period, he and his team worked on developing innovative algorithms for audio and speech enhancement for mobile communications devices. Since December 2017, he has been a Professor for audio and speech processing with Ghent University and imec, Belgium. He is passionate about signal processing and is especially interested in signal detection and enhancement for various applications in the fields of healthcare, automation, and communications.



leading the speech technology development activities.



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 Post-doc, 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 as a Research Engineer and System Architect. He is currently a Senior Principal with the Product Line Voice and Audio Solutions, NXP, Leuven, Belgium,

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 with the Speech Processing Software and Technology Research Department, AT&T Labs, Florham Park, NJ, USA. In 1999, he joined the Signal Processing Department, Siemens AG (COM Mobile Devices), Munich, Germany, and contributed to speech codec standardization in ETSI, 3GPP, and ITU-T. In 2005, he joined Siemens Corporate Technology, Munich, Germany, leading the speech technology development activities in recognition, synthesis, and speaker verification. Since 2006, he has been a Full Professor with 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 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, but the ITG prize is only awarded 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.