Xu *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2021) 2021:24
https://doi.org/10.1186/s13636-021-00207-6

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Components loss for neural networks in mask-based speech enhancement

Ziyi Xu*, Samy Elshamy, Ziyue Zhao and Tim Fingscheidt

Abstract

Estimating time-frequency domain masks for single-channel speech enhancement using deep learning methods has recently become a popular research field with promising results. In this paper, we propose a novel *components loss* (CL) for the training of neural networks for mask-based speech enhancement. During the training process, the proposed CL offers separate control over preservation of the speech component quality, suppression of the noise component, and preservation of a naturally sounding residual noise component. We illustrate the potential of the proposed CL by evaluating a standard convolutional neural network (CNN) for mask-based speech enhancement. The new CL is compared to several baseline losses, comprising the conventional mean squared error (MSE) loss w.r.t. speech spectral amplitudes or w.r.t. an ideal-ratio mask, auditory-related loss functions, such as the perceptual evaluation of speech quality (PESQ) loss and the perceptual weighting filter loss, and also the recently proposed SNR loss with two masks. Detailed analysis suggests that the proposed CL obtains a better or at least a more balanced performance across all employed instrumental quality metrics, including SNR improvement, speech component quality, enhanced total speech quality, and particularly also delivers a natural sounding residual noise component. For unseen noise types, we excel even perceptually motivated losses by an about 0.2 points higher PESQ score. The recently proposed so-called SNR loss with two masks not only requires a network with more parameters due to the two decoder heads, but also falls behind on PESQ and POLQA and particularly w.r.t. residual noise quality. Note that the proposed CL shows significantly more 1st ranks among the evaluation metrics than any other baseline. It is easy to implement, and code is provided at https://github.com/ifnspaml/Components-Loss.

Keywords: Mask-based speech enhancement, noise reduction, components loss, CNN

1 Introduction

Speech enhancement aims at improving the intelligibility and perceived quality of a speech signal that has been degraded, e.g., by additive noise. This task becomes very challenging when only a single-channel microphone mixture signal is available without any knowledge about the individual components. Single-channel speech enhancement has attracted a lot of research attention due to its importance in real-world applications, including telephony, hearing aids devices, and robust speech recognition. Numerous speech enhancement methods were proposed in the past decades. The classical method for single-channel speech enhancement is to estimate a time-frequency (TF) domain mask or, more specifically, to calculate a spectral weighting rule [1–5]. To obtain the TF domain coefficients for a spectral weighting rule, the estimation of the noise power, the a priori signal-to-noise ratio (SNR) [1, 6–10], and sometimes also the a posteriori SNR are required. Finally, the spectral weighting rule is applied to obtain the enhanced speech. Thereby, it is still common practice to enhance only the amplitudes and leave the noisy phase untouched. However, the performance of these classical methods degrades significantly in low-SNR conditions and also in the presence of non-stationary noise [11]. To mitigate this problem, e.g., a data-driven ideal mask-based approach has been proposed in [12, 13]. Therein, Fingscheidt et al. use a

*Correspondence: ziyi.xu@tu-bs.de
Institute for Communications Technology, Technische Universität Braunschweig, Schleinitzstr. 22, 38106 Braunschweig, Germany

simple regression for estimating the coefficients of the spectral weighting rules, which reduces the speech distortion while retaining a high noise attenuation. Interestingly, as with neural networks, this approach already allowed the definition of arbitrary loss functions. Note that Erkelens et al. published briefly afterwards on data-driven speech enhancement [14, 15].

In recent years, deep learning methods have been developed and used for weighting rule-based (now widely called mask-based) speech enhancement pushing performance limits even further, also in the presence of non-stationary noise [16–24]. The powerful modeling capability of deep learning enables the direct estimation of TF masks without any intermediate steps. Wang et al. [16, 25] illustrate that the ideal ratio mask-based approach, in general, performs significantly better than spectral envelope-based methods for supervised speech enhancement. Williamson et al. [21] propose to use a complex ratio mask which is estimated from the single-channel mixture to enhance both the amplitude spectrogram and also the phase of the speech. Different from other methods that directly estimate the TF mask, an approach that predicts the clean speech signal while estimating the TF mask inside the network is proposed in [17, 18]. Therein, the TF mask is applied to the noisy speech amplitude spectrum inside the network in an additional multiplication layer. Thus, the output of the network is already the enhanced speech spectrum, and not a mask which is instead learned implicitly. The authors in [17] demonstrate that the new method outperforms the conventional approach, where the TF mask is the training target and hence learned explicitly. In this paper, we estimate the mask implicitly by using convolutional neural networks (CNNs).

For the training of deep learning architectures for both, mask-based [16–21, 23] and regression-based [24, 26] speech enhancement, most networks use the mean squared error (MSE) as a loss function. The parameters of the deep learning architectures are then optimized by minimizing the MSE between the inferred results and their corresponding targets. In reality, optimization of the MSE loss in training does not guarantee any perceptual quality of the speech *component* and of the residual noise *component*, respectively, which leads to limited performance [27–36]. This effect is even more evident when the level of the noise component is significantly higher than that of the speech component in some regions of the noisy speech spectrum, which explains the bad performance at lower SNR conditions when training with MSE. To minimize the global MSE during training, the network may learn to completely attenuate such TF regions [27], a muting effect that is well-known from error concealment under bad channel SNR conditions [37, 38]. This can lead to insufficient quality of the speech component and very unnatural sounding residual noise. To keep more speech component details and to constrain the speech distortion to an acceptable level, Shivakumar et al. [27] assigned a high penalty against speech component removal in the conventional MSE loss function during training, which results in an improvement in speech quality metrics. A perceptually weighted loss function that emphasizes important TF regions has recently been proposed in [28, 29], improving speech intelligibility.

In fact, speech enhancement neural networks aim to improve the output SNR given the input noisy mixture. Thus, another straightforward direction is to use the SNR as a loss function as proposed in [39], which is optimized in the training phase. In this work, Erdogan et al. proposed a framework to implicitly estimate two separate masks for estimating the target speech and the target additive noise. During training, the frames with higher energy have more weight in the loss function, which is not desired and would limit the generalization ability to unseen signals. To mitigate this problem, a power-law compression is integrated into the SNR loss as proposed in [39], so that the loss function will not be affected by the power scaling of the training utterances. Nevertheless, the SNR loss [39] does not consider sophisticated perceptual aspects, which can again lead to an insufficient speech component quality and an unnatural-sounding residual noise component.

A more straightforward direction is to utilize the short-time objective intelligibility (STOI) [40] and the perceptual evaluation of speech quality (PESQ) [41] metrics as a loss function, which could be used to optimize for speech intelligibility and speech quality, respectively, during training [31–36]. Using STOI as an optimization criterion has been studied in [33, 35, 36]. Fu et al. [36] proposed a waveform-based utterance enhancement method to optimize the STOI score. They also show that combining STOI with the conventional MSE as an optimization criterion can further increase the speech intelligibility. Using PESQ as an optimization criterion is proposed and studied in [31, 32, 35]. In [31], the authors have amended the MSE loss by integrating parts of the PESQ metric. This proposed loss achieved a significant gain in speech perceptual quality compared to the conventional MSE loss. Zhang et al. [35] integrated both STOI and PESQ into the loss function, thereby improving speech separation performance.

However, both, original STOI and PESQ, are non-differentiable functions which cannot be used as an optimization criterion for gradient-based learning directly. A common solution is to use differentiable approximations for STOI or PESQ instead of the original expressions [31–33, 36]. Yet, how to find the best approximated expression is still an open question. In [35], the authors propose a gradient approximation method to estimate the gradients of

the original STOI and PESQ metrics. Still, these perceptual loss functions do not offer the flexibility of separate control over noise suppression and preservation of the speech component.

In this paper, we propose a novel so-called *components loss* (CL) for deep learning applications in speech enhancement. The newly proposed components loss is inspired by the merit of *separately measuring* the performance of speech enhancement systems on the speech component and the residual noise component, which is the so-called white-box approach [10, 42–44]. The white-box approach allows to measure the performance of mask-based speech enhancement w.r.t. three major aspects: (1) noise attenuation, (2) naturalness of residual noise, and (3) distortion of the speech component. Note that such component-wise quality metrics have also been adopted in ITU-T Rec. P.1100 [45], P.1110 [46], and P.1130 [47] to evaluate the performance of hands-free systems. We utilize a CNN structure adapted from [48] to illustrate the new components loss in the context of speech enhancement. However, the new loss function is not restricted to any specific network topology or application.

In contrast to the use of perceptual losses such as PESQ and STOI [31, 33], our proposed components loss (CL) is naturally differentiable for gradient-based learning and not a perceptual loss by design. In practice, the new loss function does not need any additional training material or extensive computational effort compared to explicitly auditory-related loss functions [27, 28], which makes it very easy to implement and also to integrate into existing systems. A further merit is that the new CL not only focuses on offering a strong noise attenuation and a good speech component quality, but also allows for a more natural residual noise, where the trade-off can be controlled directly. Note that highly distorted residual noise can be even more disturbing than the original unattenuated noise signal for human listeners [44]. Parts of this work, namely one of our two proposed losses, have been pre-published with limited analysis and evaluation in [49]. Following up our proposed CLs, Xia et al. proposed a modified components loss for speech enhancement in [50], and Strake et al. proposed a component loss for a joint denoising and dereverberation task in [51].

The rest of the paper is structured as follows: In Section 2, we describe the investigated speech enhancement task and introduce our mathematical notations. The baseline methods used as reference for evaluation are also introduced in this section. Next, we present our proposed components loss function for mask-based speech enhancement in Section 3. The experimental setup is provided in Section 4, followed by the results and discussion in Section 5. Our work is concluded in Section 6.

2 Notations and baselines

2.1 Notations

We assume an additive single-channel model for the time-domain microphone mixture $y(n) = s(n) + d(n)$ of the clean speech signal $s(n)$ and the added noise signal $d(n)$, with n being the discrete-time sample index. Since mask-based speech enhancement typically operates in the TF domain, we transfer all the signals to the frequency domain by applying a discrete Fourier transform (DFT). Note that this procedure is also often called short-time Fourier transform (STFT), and successive STFT frames overlap in time. Therefore, let $Y_\ell(k) = S_\ell(k)+D_\ell(k)$ be the respective DFT, and $|Y_\ell(k)|$, $|S_\ell(k)|$, and $|D_\ell(k)|$ be their DFT amplitudes, with frame index $\ell \in \mathcal{L} = \{1, 2, \ldots, L\}$ and frequency bin index $k \in \mathcal{K} = \{0, 1, \ldots, K-1\}$ with K being the DFT size. In this paper, we only estimate the real-valued mask $M_\ell(k) \in \mathbb{R}$ to enhance the amplitude spectrogram of the noisy speech and use the untouched noisy speech phase for reconstruction, obtaining the predicted enhanced speech spectrum

$$\hat{S}_\ell(k) = Y_\ell(k) \cdot M_\ell(k)\,. \tag{1}$$

It is then transformed back to the time domain signal $\hat{s}(n)$ with IDFT followed by overlap add (OLA).

2.2 Baseline network topology

As proposed in [17, 18], we predict the clean speech signal while estimating the TF mask inside the network as shown in Fig. 1. The NORM box in Fig. 1 represents a zero-mean and unit-variance normalization based on statistics collected on the training set. The CNNs used in this work have exactly the same structure as in [48, Fig. 6] but with different parameter settings, which will be explained later. This CNN topology has shown great success in coded speech enhancement [48] and is capable of improving speech intelligibility [52]. Although more complex deep learning architectures could be used, we choose this CNN structure for simple illustration. Note that any other network topology could be used instead.

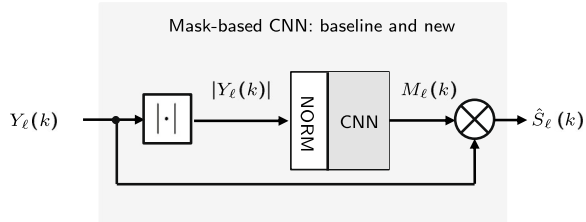

Fig. 1 Schematic of the **mask-based CNN** for speech spectrum enhancement, used for both the baseline CNN (baseline losses) and the new CNN (components loss). Details of the CNN can be seen in Fig. 2

The input of the CNN is a normalized noisy amplitude spectrogram matrix $\mathbf{Y}'_\ell$ with the dimensions $K_{\text{in}} \times L_{\text{in}}$ as shown in Fig. 2, where K_{in} represents the number of input and output frequency bins, and $L_{\text{in}} = 5$ being the number of normalized context frames centered around the normalized frame ℓ. Due to the conjugate symmetry of the DFT, it is not necessary to choose K_{in} equal to the DFT size K.

The convolutional layers are represented by the Conv$(f, h \times w)$ operation in Fig. 2. The number of filter kernels is given by $f \in \{F, 2F\}$ and thus automatically defines also the number of output feature maps which are concatenated horizontally after each convolutional layer. The dimension of the filter kernel is defined by $h \times w$, where $h = H$ is the height and $w \in \{L_{\text{in}}, F, 2F\}$ is the width. The width of the kernel is always corresponding to the width of the respective input to that layer, so that the actual convolution is operating only in vertical (frequency) direction. In the convolution layers, the stride is set to 1, and zero-padding is implemented to guarantee that the first dimension of the layer output is the same as that for the layer input. The maxpooling and upsampling layers have a kernel size of (2×1). The stride of the maxpooling layers is set to 2. The number of the input and output frequency bins K_{in} must be compatible with the two times maxpooling and upsampling operations. All possible forward residual skip connections are added to the layers with matched dimensions to ease any vanishing gradient problems during training [53]. To estimate a real-valued mask $M_\ell(k) \in [0, 1]$, the activation function used in the last layer is sigmoid.

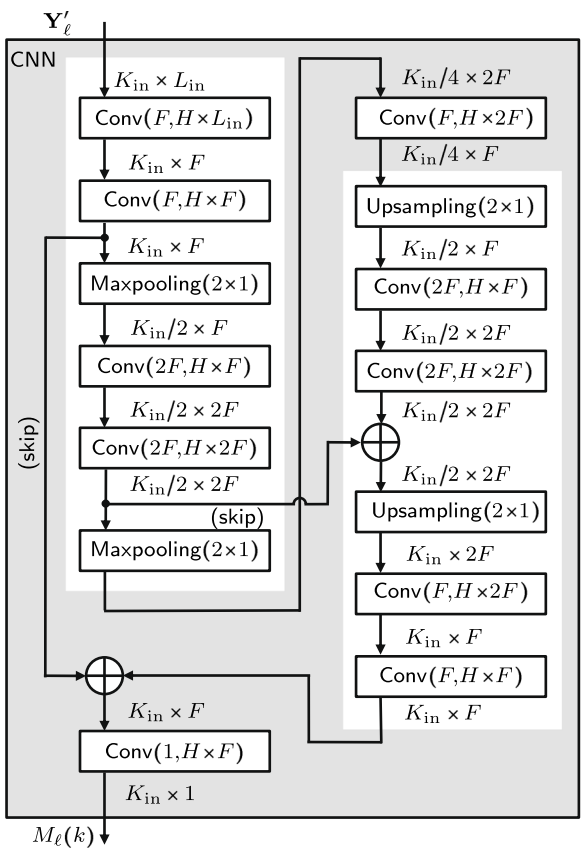

Fig. 2 Topology details of the **employed CNN** in Fig. 1 (adopted from [48, Fig. 6]). The operation Conv$(f, h \times w)$ stands for convolution, with F or $2F$ representing the number of filter kernels in each layer, and $(h \times w)$ representing the kernel size. The maxpooling and upsampling layers have a kernel size of (2×1). The stride of maxpooling layers is set to 2. The gray areas contain two symmetric procedures. All possible forward residual skip connections are added to the layers with matched dimensions

2.3 Baseline losses

2.3.1 Baseline MSE

The conventional approach to train a mask-based CNN for speech enhancement uses the MSE loss. In the training process, the input of the network is the normalized noisy amplitude spectrogram matrix $\mathbf{Y}'_\ell$ as above, and the training target is the corresponding amplitude spectrum of the clean speech $|S_\ell(k)|$ at frame ℓ, $k \in \mathcal{K}$. The implicitly estimated mask is applied to the noisy speech amplitude spectrum inside the network as shown in Fig. 1. The MSE loss function for each frame ℓ is measured between the clean and the predicted enhanced speech amplitude spectrum, and is defined as

$$J_\ell^{\text{MSE}} = \sum_{k \in \mathcal{K}} \left(|\hat{S}_\ell(k)| - |S_\ell(k)| \right)^2 . \tag{2}$$

As can be observed, all frequency bins have equal importance without any perceptual considerations, such as the masking property of the human ear [29], or the loudness difference [31]. Furthermore, as the MSE loss is optimized in a global fashion, the network may learn to completely attenuate some regions of the noisy spectrum, where the noise component is significantly higher compared to the speech component. This behavior can lead to insufficient performance at lower SNR conditions.

2.3.2 Baseline eIRM MSE

A further famous baseline known from Wang and Chen [25] is their "IRM" method, implementing an MSE loss directly in the mask domain, following

$$J_\ell^{\text{eIRM}} = \sum_{k \in \mathcal{K}} \left(M_\ell(k) - M_\ell^{\text{target}}(k) \right)^2 , \tag{3}$$

with $M_\ell(k)$ and $M_\ell^{\text{target}}(k)$ being the estimated mask and the target oracle IRM, respectively. We call this "baseline eIRM MSE," with "e" for "explicit."

2.3.3 Baseline PW-FILT

In order to obtain better perceptual quality of the enhanced speech, instead of the MSE loss, a so-called

perceptual weighting filter loss PW-FILT is used [29]. In this loss, the perceptual weighting filter from code-excited linear prediction (CELP) speech coding is applied to effectively weight the error between the network output and the target, which can be expressed as

$$J_\ell^{\text{PW-FILT}} = \sum_{k\in\mathcal{K}} \left|W_\ell(k)\right|^2 \cdot \left(\left|\hat{S}_\ell(k)\right| - \left|S_\ell(k)\right|\right)^2, \quad (4)$$

where $W_\ell(k)$ represents the weighting filter frequency response as explained in more detail in Appendix 1, Eq. 15. This loss has shown superior performance compared to the MSE loss in speech enhancement [29], as well as for quantized speech reconstruction [30]. Some more detail is given in Appendix 1.

2.3.4 Baseline PW-PESQ

Another option is to adapt PESQ [41], which is one of the best-known metrics for speech quality evaluation, to be used as a loss function. Since PESQ is a complex and non-differentiable function which cannot be directly used as an optimization criterion for gradient-based learning, a simplified and differentiable approximation of the standard PESQ has been derived and used as a loss function in [31]. The proposed PESQ loss is calculated frame-wise from the loudness spectra of the target and the enhanced speech signals. The finally used loss function, which considers both auditory masking and threshold effects, combines the PESQ loss with standard MSE to introduce the perceptual criteria, and is defined as

$$J_\ell^{\text{PW-PESQ}} = \lambda_1 \cdot J_\ell^{\text{MSE}} + \lambda_2 \cdot J_\ell^{\text{PESQ}}, \quad (5)$$

with J_ℓ^{MSE} directly calculated from (2), J_ℓ^{PESQ} being the proposed PESQ loss (see Appendix 1, Eq. 16). Hyperparameters $\lambda_1 \in [0,1]$ and $\lambda_2 \in [0,1]$ are the weighting factors for the MSE loss and the PESQ loss, respectively. The network trained by loss function (5) not only aims at a low MSE loss, but also at decreasing speech distortion. More details are given in Appendix 1.

2.3.5 Baseline PW-STOI

The maximization of STOI [40] during training is also the target in several publications [31–36]. In [33], Kolbcek et al. derive a differentiable approximation of STOI, which considers the frequency selectivity of the human ear, for the training of a mask-based speech enhancement DNN. Interestingly, the authors find that no improvement in STOI can be obtained by using the proposed loss function. They conclude in their work that "the traditional MSE-based speech enhancement networks may be close to optimal from an estimated speech intelligibility perspective" [33]. Note that PW-STOI is not calculated frame-wise compared to other baseline losses, which makes it very difficult to implement in our setup and to allow a fair comparison. In [33], the trained network needs to estimate 30 frames of enhanced speech at once. To meet this large output size, the input size can be quite large and unpractical in our implementation. Due to the above-cited conclusion from [33] and the large output size requirement, we will not implement the PW-STOI loss in our setup.

2.3.6 Baseline two-masks SNR

To compare to the two-masks estimation network trained with SNR loss as proposed in [39], we adopted it as a further baseline. The framework in [39] estimates the amplitude spectrum of the target clean speech and the additive noise by $\left|\hat{S}_\ell(k)\right| = \left|Y_\ell(k)\right| \cdot M_\ell^S(k)$ and $\left|\hat{D}_\ell(k)\right| = \left|Y_\ell(k)\right| \cdot M_\ell^D(k)$, with $M_\ell^S(k)$ and $M_\ell^D(k)$ being the implicitly estimated masks for the clean speech and the additive noise, respectively. Thus, we need to modify our network topology to implicitly estimate the two masks. Since our employed CNN has a symmetric encoder-decoder structure as shown in Fig. 2, we can simply add an additional decoder head, which is parallel to the existing one with exactly the same topology, to estimate the additional mask. So, the encoder is connected to two parallel decoder heads to estimate the two masks. However, the estimated clean speech and additive noise obtained by directly applying $M_\ell^S(k)$ and $M_\ell^D(k)$ from the two decoder heads may not meet the power conservation constraint $\left|\hat{S}_\ell(k)\right|^2 + \left|\hat{D}_\ell(k)\right|^2 = \left|Y_\ell(k)\right|^2$. To mitigate this problem, during the test phase, the estimated two masks are merged to one mask by $M_\ell(k) = 0.5 \cdot (1 + (M_\ell^S(k))^2 - (M_\ell^D(k))^2)$ as proposed in [39]. The final enhanced speech amplitude spectrum $\hat{S}_\ell(k)$ is obtained from (1). Details of the SNR loss are given in Appendix 1.

3 New components loss functions for mask-based speech enhancement

The newly proposed components loss (CL) is inspired by the so-called white-box approach [42], which utilizes the *filtered* clean speech spectrum $\tilde{S}_\ell(k)$ and the *filtered* noise component spectrum $\tilde{D}_\ell(k)$ to train the mask-based CNN for speech enhancement as shown in Fig. 3. We first motivate the use of the white-box approach in the following and then introduce the new components loss.

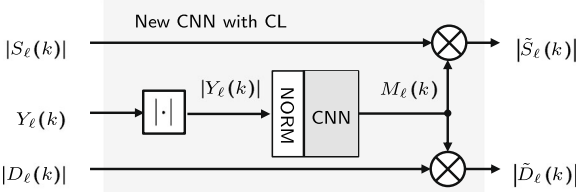

Fig. 3 Proposed **CNN training setup** for speech enhancement according to the **white-box approach**. The hereby applied components loss (CL) is given in (8) and (11)

3.1 White-box approach

Since our work is inspired by the so-called white-box approach ([42], see also [43, 44]), we introduce the *filtered* speech spectrum, which is obtained by

$$\tilde{S}_\ell(k) = S_\ell(k) \cdot M_\ell\,(k)\,, \tag{6}$$

while the *filtered* noise spectrum is estimated by

$$\tilde{D}_\ell(k) = D_\ell(k) \cdot M_\ell\,(k)\,. \tag{7}$$

The *filtered* speech component spectrum $\tilde{S}_\ell\,(k)$ and the *filtered* noise component spectrum $\tilde{D}_\ell\,(k)$ are transformed back to the time domain signals $\tilde{s}(n)$ and $\tilde{d}(n)$, respectively, with IDFT followed by overlap add (OLA).

Speech enhancement systems aim to provide a strong noise attenuation, a naturally sounding residual noise, and an undistorted speech component. Thus, the evaluation of a speech enhancement algorithm ideally needs to measure the performance w.r.t. all three aspects. The white-box approach, which allows to measure the performance based on the *filtered* speech component $\tilde{s}(n)$ and the *filtered* residual noise component $\tilde{d}(n)$, has been originally proposed in [42]. A white-box based measure *does not* employ the enhanced speech signal $\hat{s}(n)$, but only utilizes the *filtered* and *unfiltered* components with the *unfiltered* ones as a reference [42–44]. Due to its usefulness, this component-wise white-box measurement has been widely adopted in ITU-T Recs. P.1100 [45], P.1110 [46], and P.1130 [47] to evaluate the performance of hands-free systems. One might ask whether there is a price to pay with component-wise quality evaluation, since masking effects of human perception are not at all exploited. Accordingly, we will have to use also perceptual quality metrics in the evaluation Sections IV and V. Interestingly, supporting the adoption of components metrics in ITU-T recommendations, our newly proposed components loss (CL) turns out to be superior both in PESQ and POLQA (perceptual objective listening quality prediction).

3.2 New components loss with 2 components

3.2.1 New 2CL

The core innovative step of this work is as follows: Since we assume an additive single-channel model, *both* the amplitude spectrum of the clean speech $\left|S_\ell(k)\right|$ and the additive noise $\left|D_\ell(k)\right|$ are accessible during the training phase, and thus can be used as training targets. First, the *filtered* components $\left|\tilde{S}_\ell(k)\right|$ and $\left|\tilde{D}_\ell(k)\right|$ in Fig. 3 are obtained by (6) and (7), respectively. Then, we define our proposed components loss (CL) for each frame ℓ as

$$J_\ell^{\mathrm{2CL}} = (1-\alpha)\cdot\sum_{k\in\mathcal{K}}\left(\left|\tilde{S}_\ell\,(k)\,\right| - \left|S_\ell\,(k)\,\right|\right)^2 + \alpha\cdot\sum_{k\in\mathcal{K}}\left|\tilde{D}_\ell\,(k)\,\right|^2, \tag{8}$$

with $\alpha \in [0, 1]$ being the weighting factor that can be used to control the trade-off between noise suppression and speech component quality.

This proposed CL dubbed as "2CL" is the combination of two independent loss contributions, where the first term represents the loss function for the *filtered* clean *speech* component, and the second term represents the power of the *filtered noise* component. Both of the two losses are calculated frame-wise. Minimizing the first term of the loss function is supposed to preserve detailed structures of the speech spectrum, so the perceptual quality of the speech component will be maintained. Any distortion or attenuation being present in the *filtered* speech spectrum will be punished by this loss term. The second term of 2CL representing the residual noise power should also be as low as possible. Thus, minimizing the second loss term is responsible for the actual noise attenuation (NA), which is not at all enforced by the first term.

The first and the second term in (8) are combined by the weighting factor α. Compared to conventional training using the standard MSE loss function as shown in Fig. 1, our newly proposed training with 2CL offers more information to the network to learn which part of the noisy spectrum belongs to the speech component that should be untouched, and which part is the added noise that should be attenuated. By tuning α close to 1, 2CL will penalize high residual noise power stronger than severe speech component distortion. Thus, the trained network tends to suppress more noise but maybe at the cost of more speech distortions. When α is close to 0, the trained network will behave conversely, so that it will offer better speech component quality and may not provide much noise attenuation. Controlling the trade-off between speech component quality and noise attenuation is impossible when using the conventional single-target MSE loss function (2). Note that the enhanced speech $\hat{S}_\ell(k)$ is not part of the loss anymore, only implicitly, keeping in mind that $\hat{S}_\ell(k) = \tilde{S}_\ell(k) + \tilde{D}_\ell(k)$. Furthermore, for a speech enhancement algorithm, a highly distorted residual noise can be even more disturbing than the original unattenuated noise signal for human listeners [44]. The conventional networks trained with MSE tend to have a strong noise distortion because of the TF bin attenuation behavior as mentioned in the "Introduction" section. Conversely, the network trained by the proposed 2CL may have less TF bin attenuation, because the TF bin attenuation is also harmful to the speech component and will be penalized by the first term of 2CL. As a consequence, the networks trained by the proposed 2CL are likely to offer more natural residual noise,

even though the residual noise *quality* is not explicitly considered in (8).

3.2.2 Reference iIRM MSE

For our proposed 2CL (8), the global optimal solution of the implicitly estimated mask can be derived as (see Appendix 2):

$$M_\ell^{\text{2CL-opt}}(k) = \frac{\left|S_\ell(k)\right|^2}{\left|S_\ell(k)\right|^2 + \frac{\alpha}{1-\alpha} \cdot \left|D_\ell(k)\right|^2}, \tag{9}$$

which is similar to an ideal ratio mask (IRM) proposed in [25] with $\beta = 1$, however, with the important difference that through our loss formulation in the signal domain (as opposed to an MSE on masks), (8) only *implicitly* trains a network with mask output, whereas the original IRM [25] (which we call "baseline eIRM MSE") aims at minimizing the MSE *explicitly* on the mask. Nevertheless, due to the similar global optimum (9), we compare our 2CL also to a network, which can perform an implicit IRM (iIRM) estimation. By doing this, we train a network with the same training setup as the baseline MSE shown in Fig. 1, using an MSE loss function. We construct an iIRM loss based on the standard MSE loss J_ℓ^{MSE} from (2) as

$$J_\ell^{\text{iIRM}} = \sum_{k \in \mathcal{K}} \left(\left|\hat{S}_\ell(k)\right| - \left|\hat{S}_\ell^{\text{target}}(k)\right| \right)^2, \tag{10}$$

with $\left|\hat{S}_\ell^{\text{target}}(k)\right| = \left|Y_\ell(k)\right| \cdot M_\ell^{\text{2CL-opt}}(k)$ being the enhanced speech amplitude spectrum obtained from an oracle IRM (9). Comparing to the baseline MSE loss (2), only the training target is different. This method is used as a reference (we do not call it baseline) to show the difference to our proposed 2CL, so we dubbed this method "Reference iIRM MSE."

Note that it can be shown that our new 2CL (8), and the reference iIRM MSE (10) have loss functions which are different by an important time- and frequency-dependent weighting factor. The interested reader is referred to Appendix 3.

3.3 New components loss with 3 components

3.3.1 New 3CL

Based on the proposed 2CL, to explicitly put the residual noise *quality* into consideration during training, we also propose an advanced CL, which is defined as

$$\begin{aligned} J_\ell^{\text{3CL}} = {} & (1-\alpha-\beta) \cdot \sum_{k \in \mathcal{K}} \left(\left|\tilde{S}_\ell(k)\right| - \left|S_\ell(k)\right| \right)^2 \\ & + \alpha \cdot \sum_{k \in \mathcal{K}} \left|\tilde{D}_\ell(k)\right|^2 \\ & + \beta \cdot \sum_{k \in \mathcal{K}} \left(\frac{\left|\tilde{D}_\ell(k)\right|}{\sqrt{\sum_{\kappa \in \mathcal{K}} \left|\tilde{D}_\ell(\kappa)\right|^2}} - \frac{\left|D_\ell(k)\right|}{\sqrt{\sum_{\kappa \in \mathcal{K}} \left|D_\ell(\kappa)\right|^2}} \right)^2, \end{aligned} \tag{11}$$

with $\alpha \in [0,1]$ and $\beta \in [0,1]$ being the weighting factors to control the speech component quality, the noise suppression, and now also the residual noise quality. In order to have stable training and not to enlarge the speech component MSE (first term in (11)) during training, we limit the tuning range of the weighting factors to $0 \leq \alpha + \beta \leq 1$. This CL with three terms (dubbed "3CL") is also used to train the speech enhancement neural network as shown in Fig. 3, without requiring any additional training material compared to when using 2CL.

The first two terms of the 3CL in (11) are the same as in (8), and the additional third term is the loss between the normalized spectra of the *filtered* and the *unfiltered* noise component and is supposed to preserve residual noise quality. In order to decouple noise attenuation and residual noise quality, firstly, this additional term is not directly calculated from the *filtered* and the *unfiltered* noise spectra, but utilizing the *normalized* ones. Secondly, both positive and negative differences between the *filtered* and the *unfiltered* noise spectra are punished equally, which means this loss should be non-negative. So, this additional term can have the form of the standard MSE, which is shown in (11). This additional loss aims to preserve the residual noise quality even more, enforcing a similarity of residual noise and the original noise component. Note that many alternative definitions of the residual noise quality loss term are possible; however, it should always be ensured that a fullband attenuation ($\tilde{D}_\ell(k) = \rho \cdot D_\ell(k), \rho < 1$) should lead to a zero loss contribution, since it perfectly preserves residual noise *quality*.

4 Experimental setup

4.1 Databases and experimental setup

4.1.1 Database

The used clean speech data in this work is taken from the Grid corpus [54]. The Grid corpus is particularly useful for our experiments, since it provides clean speech samples from many different speakers in a sufficient amount of data for our experiments, which is critical for *speaker-independent* training. To make our trained CNN *speaker-independent*, we randomly select 16 speakers, containing 8 male and 8 female speakers, and use 200 sentences per speaker for the CNN training. The duration of each sentence is exactly 3 seconds. The superimposed noises used in this paper are obtained from the CHiME-3 dataset [55]. Both the clean speech and the additive noise signals have a sampling rate of 16 kHz. To generalize the network and also to increase the amount of training data, the noisy speech always contains multiple SNR conditions and includes various noise types. We use pedestrian noise (PED), café noise (CAFE), and street noise (STR) to generate the training data. We simulate six SNR conditions

from −5 to 20 dB with a step size of 5 dB. The SNR level is adjusted according to ITU-T P.56 [56]. Thus, the training material consists of $16 \times 200 \times 3 \times 6 = 57,600$ sentences, in total 48 h. From the complete training material, 20% of the data is used for validation and 80% is used for actual training.

During the test phase, the clean speech data is taken from four further Grid speakers, two males and two females, with 10 sentences each neither seen during training nor during validation. The used test noise contains both *seen* and *unseen* noise *types*. The *seen* test noise includes PED and CAFE noise, but extracted from different files, which have not been used during training and validation. To perform a noise type-independent test, we additionally create noisy test data using *unseen* bus noise (BUS), which is also taken from CHiME-3 and is not seen during training and validation. The test data also contains the six SNR conditions.

4.1.2 Experimental setup

Speech and noise signals are subject to an FFT size of $K = 256$, using a periodic Hann window, and 50% overlap. We use the CNN illustrated in Fig. 2 for the mask estimation. Although more complex deep learning architectures could be used, we choose this CNN structure to illustrate our concept. The number of the input and output frequency bins K_{in} is set to $129 + 3 = 132$ for each frame's DFT, as shown in Fig. 2. The additional 3 frequency bins are taken from the redundant bins (from $k = 129$ to $k = 131$), which are used to make it compatible with the two times maxpooling and upsampling operation in the CNN. The input context is $L_{\text{in}} = 5$. The number of filters in each convolutional layer represented by F in Fig. 2 is set to 60. The used height of the filter kernels is $h = H = 15$. In the test phase, we only extract the first 129 frequency bins from the 132 output frequency bins to reconstruct the complete spectrum, which is used to obtain the time domain signal by IDFT with OLA. Furthermore, a minibatch size of 128 is used during training. The learning rate is initialized to $2{\cdot}10^{-4}$ and is halved once the validation loss does not decrease for two epochs. The CNN activation functions are exactly the same as used in [48].

In the baseline training for the perceptual weighting filter loss PW-FILT, the linear prediction order represented by N_{p} in (14) is set to 16. The perceptual weighting factors γ_1 and γ_2 in (14) are set to 0.92 and 0.6, respectively.

4.2 Quality measures

We use both the white-box approach [42] which provides the *filtered* clean speech component $\tilde{s}(n)$ and the *filtered* noise component $\tilde{d}(n)$, as well as standard measures operating on the predicted enhanced speech signal $\hat{s}(n)$. In this paper, we use the following measures [10]:

4.2.1 Delta SNR

$\Delta\text{SNR} = \text{SNR}_{\text{out}} - \text{SNR}_{\text{in}}$, measured in dB. SNR_{out} and SNR_{in} are the SNR levels of the enhanced speech and the noisy input speech, respectively, and are measured after ITU-T P.56 [56], based on $\tilde{s}(n)$, $\tilde{d}(n)$ and $s(n)$, $d(n)$, respectively. This measure should be as high as possible.

4.2.2 PESQ MOS-LQO

This measure uses $s(n)$ as reference signal and either the *filtered* clean speech component $\tilde{s}(n)$ or the enhanced speech $\hat{s}(n)$ as test signal according to [46, 57], being referred to as PESQ($\tilde{\mathbf{s}}$) and PESQ($\hat{\mathbf{s}}$), respectively. A high PESQ score indicates better speech (component) perceptual quality.

4.2.3 Perceptual objective listening quality prediction (POLQA)

This metric is one of the newest objective metrics for speech quality [58]. POLQA is measured between the reference signal $s(n)$ and the predicted clean speech $\hat{s}(n)$ according to [58] and is denoted as POLQA($\hat{\mathbf{s}}$). The same with PESQ, a higher POLQA score is favored.

4.2.4 Segmental speech-to-speech-distortion ratio

$$\text{SSDR} = \frac{1}{|\mathcal{L}_1|} \sum_{\ell \in \mathcal{L}_1} \text{SSDR}(\ell) \qquad [\,\text{dB}]$$

with $\mathcal{L}_1$ denoting the set of speech-active frames [10], and using

$$\text{SSDR}(\ell) = \max\left\{\min\left\{\text{SSDR}'(\ell), 30\,\text{dB}\right\}, -10\,\text{dB}\right\},$$

with

$$\text{SSDR}'(\ell) = 10\log_{10}\left[\frac{\sum\limits_{n\in\mathcal{N}_\ell} s^2(n)}{\sum\limits_{n\in\mathcal{N}_\ell}\left[\tilde{s}(n+\Delta) - s(n)\right]^2}\right], \qquad (12)$$

with $\mathcal{N}_\ell$ denoting the sample indices n in frame ℓ, and Δ being used to perform time alignment of the filtered signal $\tilde{s}(n)$. A low distortion of the filtered speech components leads to a high SSDR.

4.2.5 Segmental noise attenuation (NA_{seg})

$$\text{NA}_{\text{seg}} = 10\log_{10}\left[\frac{1}{|\mathcal{L}|}\sum_{\ell\in\mathcal{L}} \text{NA}_{\text{frame}}(\ell)\right], \qquad [\,\text{dB}] \qquad (13)$$

with

$$\text{NA}_{\text{frame}}(\ell) = \frac{\sum_{n\in\mathcal{N}_\ell} d^2(n)}{\sum_{n\in\mathcal{N}_\ell} \tilde{d}^2(n+\Delta)},$$

where $\mathcal{L}$ denotes the set of all frame indices. We measure NA_{seg} for the purpose of parameter optimization, so we can easily choose the weighting factors that offer a strong noise attenuation as well as a good speech component perceptual quality. In the test phase, we use the ΔSNR metric

to reflect the overall SNR improvement caused by noise suppression instead of using a single $\mathrm{NA}_{\mathrm{seg}}$ metric.

4.2.6 The weighted log-average kurtosis ratio (WLAKR)

This metric measures the noise distortion (especially penalizing musical tones) using $d(n)$ as reference signal and the *filtered* noise component $\tilde{d}(n)$ as test signal according to ITU-T P.1130 [47]. A WLAKR score that is closer to 0 indicates less noise distortion [44, 59]. Accordingly, in our analysis, we will show averaged *absolute* WLAKR values.

4.2.7 STOI

We use STOI to measure the intelligibility of the enhanced speech, which has a value between zero and one [40]. A STOI score close to one indicates high intelligibility.

We group these measurements to noise *component* measures (ΔSNR and WLAKR), speech *component* measures (SSDR and PESQ($\tilde{\mathbf{s}}$)), and total performance measures (PESQ($\hat{\mathbf{s}}$), POLQA($\hat{\mathbf{s}}$), and STOI).

5 Results and discussion

5.1 Hyperparameter optimization

To allow for an efficient hyperparameter search, we optimize the weighting factors for our proposed components loss functions by using only 12.5% of the validation set data. The total performance measures PESQ($\hat{\mathbf{s}}$), POLQA($\hat{\mathbf{s}}$), and STOI are averaged over all training noise types and all SNR conditions.

5.1.1 2CL hyperparameter α

The performance for different weighting factors α for 2CL (8) is shown in Table 1. The baseline MSE in Table 1 represents the conventional mask-based CNN as shown in Fig. 1 and is trained using the MSE loss function. It becomes obvious that a choice of α in (8) being far away from 0.5 leads to either bad perceptual speech quality or low speech intelligibility. This behavior is expected, since speech enhancement requires a sufficiently strong noise attenuation as well as an almost untouched speech component. To choose the best weighting factor α from Table 1, we first discard all columns where at least one measure is below or equals the baseline MSE and subsequently select from the remaining values $\alpha \in \{0.45, 0.5, 0.55\}$ the best performing, which is $\alpha = 0.5$. The selected setting is gray-shaded as shown in Table 1.

In Fig. 4, we plot the obtained $\mathrm{NA}_{\mathrm{seg}}$ vs. PESQ($\tilde{\mathbf{s}}$) values for the various combinations of hyperparameters as shown in Table 1. Here, from top to bottom, each marker depicts a certain SNR condition varying from 20 to −5 dB in steps of 5 dB. The further a curve is to the right *and* to the top, the better the overall performance. We can see that the performance for the selected hyperparameter $\alpha = 0.5$ (dot-dashed pink line, circle markers) is a quite balanced choice. Furthermore, the settings with $\alpha = 0.45$ and 0.55, which are close to our chosen one, show quite similar performance. This can be seen both from the values in Table 1, and from the quite close lines in Fig. 4, indicating the sensitivity of the 2CL towards hyperparameter changes is not high.

Table 1 Optimization of hyperparameter α for the **new 2CL (8) on 12.5% of the validation set**. The selected setting is gray-shaded

	Baseline	New $J_{\ell}^{\mathrm{2CL}}(\alpha)$						
	MSE	$\alpha =$ 0	0.1	0.45	0.5	0.55	0.75	0.9
PESQ($\hat{\mathbf{s}}$)	2.21	1.78	2.10	2.48	2.50	2.41	2.60	2.59
POLQA($\hat{\mathbf{s}}$)	1.91	2.11	1.86	2.21	2.23	2.22	2.39	2.26
STOI	0.72	0.72	0.74	0.73	0.73	0.73	0.70	0.68

5.1.2 Reference iIRM MSE hyperparameter α

To further exploit the behavior of the reference iIRM network and to make a fair comparison to our proposed 2CL, we also optimized the hyperparameter α in the oracle IRM mask (9). We excluded the case of $\alpha = 0$, since when $\alpha = 0$, the training target becomes $|\hat{S}^{\mathrm{target}}| = |Y|$, which offers no noise suppression. The performance for different weighting factors α for the reference iIRM MSE methods is shown in Table 2. Similar to the hyperparameter optimization process for our 2CL, settings, where at least one measure is below the baseline MSE, are discarded. Among the remaining cases, we select the α which offers the highest overall perceptual quality measured by PESQ and POLQA. It turns out that the setting of $\alpha = 0.55$ offers the best results.

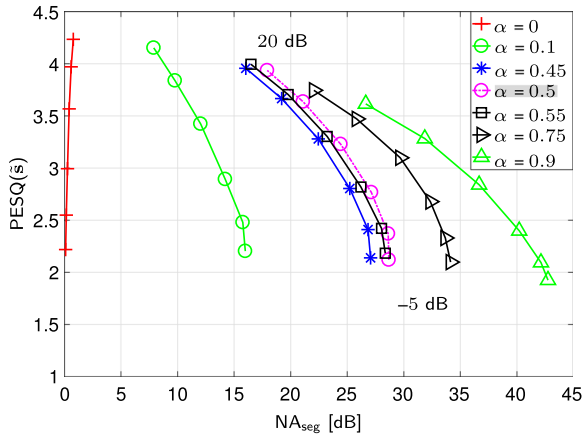

Fig. 4 Noise attenuation ($\mathrm{NA}_{\mathrm{seg}}$) vs. speech **component** quality (PESQ($\tilde{\mathbf{s}}$)) for different parameters α for the **new 2CL** (8) on **12.5% of the validation set**. From top to bottom, the markers are corresponding to six SNR conditions from 20 to −5 dB with a step size of 5 dB. The selected setting $\alpha = 0.5$ is gray-shaded in the legend

Table 2 Optimization of hyperparameter α for the **reference iIRM MSE** (10) on **12.5% of the validation set**. The selected setting is gray-shaded

	Baseline	New $J_\ell^{\text{iIRM}}(\alpha)$					
	MSE	$\alpha =$ 0.1	0.45	0.5	0.55	0.75	0.9
PESQ($\hat{\mathbf{s}}$)	2.21	2.25	2.43	2.47	2.49	2.56	2.62
POLQA($\hat{\mathbf{s}}$)	1.91	1.96	2.13	2.21	2.25	2.34	2.34
STOI	0.72	0.73	0.72	0.72	0.72	0.71	0.70

Similarly, we also plot the obtained NA_{seg} vs. PESQ($\tilde{\mathbf{s}}$) values for different hyperparameters from Table 2 as shown in Fig. 5. It can be seen that the selected hyperparameter $\alpha = 0.55$ (dot-dashed pink line, circle markers) is more to the top and to the right compared to other hyperparameter settings. So, the selected setting can offer a higher noise attenuation as well as a better speech component quality at the same time, which indicates a more balanced performance.

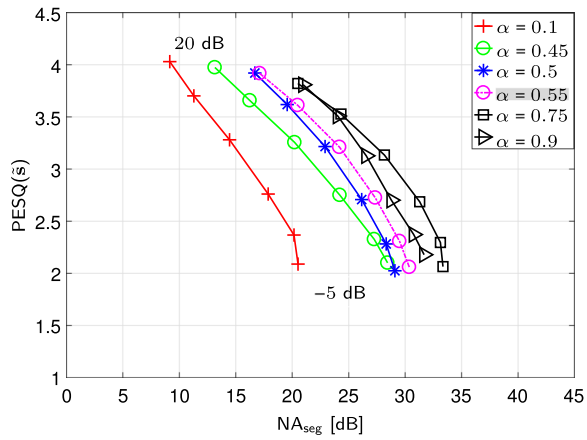

Fig. 5 Noise attenuation (NA_{seg}) vs. speech *component* quality (PESQ($\tilde{\mathbf{s}}$)) for different parameters α for the **reference iIRM MSE** (10) on **12.5% of the validation set**. From top to bottom, the markers are corresponding to six SNR conditions from 20 to −5 dB with a step size of 5 dB. The selected setting $\alpha = 0.55$ is gray-shaded in the legend

5.1.3 3CL hyperparameters α, β

We also optimize the combination of the weighting factors α and β for 3CL in (11) as shown in Table 3. The baseline MSE in Table 3 is the same as the one in Table 1. Interestingly, a good performance is achieved mostly[1] when the weighting factors for speech component quality $(1-\alpha-\beta)$ and noise attenuation (α) are equal or very close to each other—as is the case for our 2CL choice of $\alpha = 0.5$ in Table 1. This is the case for 3CL when $\alpha \in \{0.05, 0.1, 0.15, 0.2, 0.3, 0.4\}$ and the corresponding (in that order) $\beta \in \{0.9, 0.8, 0.7, 0.6, 0.4, 0.2\}$, as shown in Table 3 marked by ∗. Thus, tuning the weighting factors for speech component quality and noise attenuation in an unbalanced way will degrade the overall performance, especially for STOI or PESQ as shown in Table 3. As the best combination in Table 3, we select $\alpha = 0.1$ and $\beta = 0.8$, highlighted by a gray-shaded font. The additional term of 3CL (11), weighted with β, is supposed to preserve the residual noise quality. It can further improve the overall performance of PESQ and POLQA as can be seen when comparing the gray-shaded columns of Tables 1 and 3. The reason could be that PESQ and POLQA measures favor natural residual noise.

For the combinations of hyperparameters in Table 3, we also plot NA_{seg} vs. PESQ($\tilde{\mathbf{s}}$) as shown in Fig. 6. All curves marked by ∗ fulfill $\alpha = \frac{1-\beta}{2}$, meaning that the noise attenuation (α) and the speech distortion $(1-\alpha)$ contribute equally to the 3CL loss (11). Obviously, these curves show a comparably good speech component quality as well as a strong noise attenuation at the same time. The overall differences between these curves are very small, which is also reflected in Table 3. This also indicates that the sensitivity of the 3CL is very low, as long as we keep this condition. In Fig. 6, the curve for $\alpha = 0.8$ and $\beta = 0.1$ shows very strong noise attenuation, but with quite low PESQ($\tilde{\mathbf{s}}$). This is expected since the contribution of the noise attenuation in 3CL loss (11), which is controlled by α, is the strongest from the investigated values. On the contrary, when $\alpha = 0.1$ and the corresponding $\beta \in \{0.4, 0.6\}$, we obtain the highest PESQ($\tilde{\mathbf{s}}$) and the weakest noise attenuation. Our selected hyperparameter combination (dot-dashed pink line, circle markers) is among the curves marked by ∗ showing quite balanced performance.

5.1.4 Baseline eIRM MSE hyperparameter α

For the baseline eIRM MSE, we set the target oracle IRM $M_\ell^{\text{target}}(k)$ in (3) to the globally optimum mask of our 2CL shown in (9). Then, we optimized the hyperparameter α for eIRM MSE. The same hyperparameter settings are searched as for the reference iIRM MSE, and the results are shown in Table 4 and Fig. 7. The hyperparameter settings, where at least one measure is below or equal to the baseline MSE, are discarded. Among the remaining

Table 3 Optimization of hyperparameters α and β for the **new 3CL** (11) on **12.5% of the validation set**. The selected setting is gray-shaded

	Baseline	New $J_\ell^{\text{3CL}}(\alpha, \beta)$								
	MSE	$\alpha = 0.05$∗	0.1	0.1	0.1∗	0.15∗	0.2∗	0.3∗	0.4∗	0.6	0.8
		$\beta = 0.9$∗	0.4	0.6	0.8∗	0.7∗	0.6∗	0.4∗	0.2∗	0.2	0.1
PESQ($\hat{\mathbf{s}}$)	2.21	2.47	2.19	2.24	2.54	2.50	2.49	2.47	2.51	2.59	2.60
POLQA($\hat{\mathbf{s}}$)	1.91	2.25	1.88	1.97	2.28	2.23	2.23	2.20	2.26	2.34	2.24
STOI	0.72	0.73	0.74	0.74	0.73	0.73	0.73	0.73	0.73	0.71	0.68

[1] Note that the case of $\alpha = 0.6$ and $\beta = 0.2$ is also quite good on PESQ and POLQA, but performs poorly on STOI.

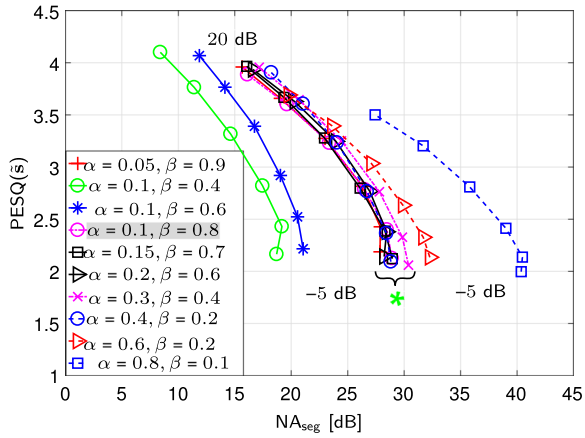

Fig. 6 Noise attenuation (NA_{seg}) vs. speech *component* quality ($\text{PESQ}(\tilde{\mathbf{s}})$) for different parameters α and β for the **new 3CL** (11) on **12.5% of the validation set**. From top to bottom, the markers are corresponding to six SNR conditions from 20 to −5 dB with a step size of 5 dB. The selected setting is gray-shaded in the legend. All curves marked by * fulfill $\alpha = \frac{1-\beta}{2}$, meaning that the noise attenuation and the speech distortion contribute equally to the 3CL loss (11)

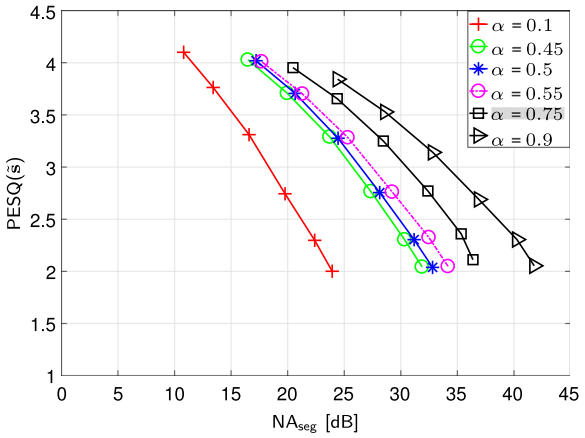

Fig. 7 Noise attenuation (NA_{seg}) vs. speech *component* quality ($\text{PESQ}(\tilde{\mathbf{s}})$) for different parameters α for the **baseline eIRM MSE** (3) on **12.5% of the validation set**. From top to bottom, the markers are corresponding to six SNR conditions from 20 to −5 dB with a step size of 5 dB. The selected setting $\alpha = 0.75$ is gray-shaded in the legend

settings, the hyperparameter α, which offers the highest overall PESQ and POLQA scores, is selected. It turns out that the setting of $\alpha = 0.75$ offers the best results and is gray-shaded. Note that for our selected setting $\alpha = 0.75$, the noise power in the target oracle IRM (9) is overestimated, which may increase the noise attenuation, but at the cost of stronger distortions to the residual noise and the speech components.

5.1.5 PW-PESQ hyperparameters λ_1, λ_2

For the baseline loss function $J_\ell^{\text{PW-PESQ}}$, to allow a fair comparison, the weighting factors λ_1 and λ_2 in (5) are also optimized, and the results are shown in Table 5. To limit the range of tuning parameters, we define $\lambda_1 + \lambda_2 \leqslant 1$. Since optimizing $J_\ell^{\text{PW-PESQ}}$ during training aims to improve the perceptual quality of the enhanced speech, we choose the optimal weighting factors, with which the best $\text{PESQ}(\hat{\mathbf{s}})$ is achieved. Furthermore, we discard the settings that offer a STOI lower than the baseline MSE. The selected setting $\lambda_1 = 0.2, \lambda_2 = 0.8$ in Table 5 provides a balanced performance and is also gray-shaded.

Table 4 Optimization of hyperparameter α for the **baseline eIRM MSE** (3) on **12.5% of the validation set**. The selected setting is gray-shaded

	Baseline	New $J_\ell^{\text{eIRM}}(\alpha)$						
	MSE	$\alpha =$	0.1	0.45	0.5	0.55	0.75	0.9
PESQ($\hat{\mathbf{s}}$)	2.21		2.36	2.56	2.58	2.61	2.70	2.76
POLQA($\hat{\mathbf{s}}$)	1.91		1.98	2.32	2.34	2.38	2.50	2.52
STOI	0.72		0.75	0.74	0.74	0.74	0.73	0.72

We plot NA_{seg} vs. $\text{PESQ}(\tilde{\mathbf{s}})$ for Table 5 as shown in Fig. 8. It can be seen that our selected hyperparameter combination (solid blue line, asterisk markers) offers mostly very good (among the two best) $\text{PESQ}(\tilde{\mathbf{s}})$ and a strong noise attenuation, yielding a balanced performance.

5.2 Experimental results and discussion

We report the experimental results on the test data for *seen* noises *types* (PED and CAFE) and *unseen* BUS noise separately. We investigate a CNN trained with the newly proposed 2CL and 3CL losses, and with the other baseline losses, which are the conventional MSE w.r.t. speech spectral amplitude and w.r.t. IRM, the auditory-related PW-PESQ and PW-FILT, and the recently proposed two-masks SNR. The measures on the *seen* noise types are shown in Tables 6 (all SNRs averaged) and 7 (−5 dB SNR); the results on *unseen* BUS noise are shown in Tables 8 (all SNRs averaged) and 9 (−5 dB SNR). The performance is averaged over all test speakers and if applicable all SNR conditions. In each column, the scheme offering the best performance is in bold font. For the CNN trained with

Table 5 Optimization of hyperparameters λ_1 and λ_2 for **baseline PW-PESQ** (5) on **12.5% of the validation set**. The selected setting is gray-shaded

	Baseline	Baseline $J_\ell^{\text{PW-PESQ}}(\lambda_1, \lambda_2)$							
	MSE	$\lambda_1 =$	0.01	0.1	0.2	0.4	0.5	0.6	0.8	0.9
		$\lambda_2 =$	0.99	0.9	0.8	0.6	0.5	0.4	0.2	0.1
PESQ($\hat{\mathbf{s}}$)	2.21		2.22	2.22	2.23	2.18	2.18	2.15	2.12	2.21
POLQA($\hat{\mathbf{s}}$)	1.91		1.90	1.87	1.89	1.84	1.87	1.81	1.84	1.85
STOI	0.72		0.71	0.72	0.72	0.73	0.73	0.72	0.72	0.72

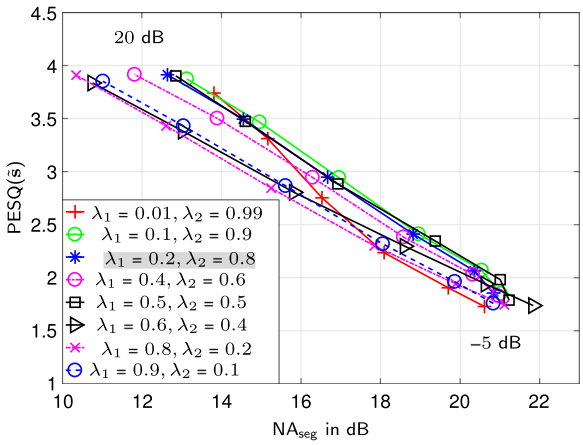

Fig. 8 Noise attenuation (NA_{seg}) vs. speech *component* quality (PESQ($\tilde{\mathbf{s}}$)) for different parameters λ_1 and λ_2 for the **baseline PW-PESQ** (5) on **12.5% of the validation set**. From top to bottom, the markers are corresponding to six SNR conditions from 20 to 5 dB with a step size of −5 dB. The selected setting is gray-shaded in the legend

2CL and 3CL, the selected settings are gray-shaded, as shown in Tables 1 and 3, respectively.

5.2.1 Seen noise types

First, we look at the performance on the *seen* noise types as shown in Table 6. It becomes obvious that the CNN trained by our proposed 2CL and 3CL mostly offers better SNR improvement than the CNN trained by the baseline MSE and the auditory-related losses, reflected by a higher ΔSNR. Among the CLs, 3CL offers the highest ΔSNR on average. This is supposed to be attributed to the second term of both 2CL (8) and 3CL (11) weighted by α, representing the *filtered* noise component power, which is explicitly forced to be low during the training process. The CNN trained by PW-FILT loss also offers quite good noise attenuation, but with a poor residual noise quality, which is reflected by a very high WLAKR score. Among the baseline methods, the CNN trained by PW-PESQ always shows the best residual noise quality. Surprisingly, the proposed 2CL also offers a better residual noise quality compared to the CNN trained with conventional MSE, even though the residual noise quality is not considered in the 2CL definition (8). The proposed 3CL offers a very good, for CAFE also the best residual noise quality, as well as the strongest noise attenuation at the same time. This is expected, and is likely from the contribution of the third term in 3CL (11), which is supposed to preserve residual noise quality. During training, this term is explicitly forced to be low to keep a naturally sounding residual noise, by enforcing a similarity of the residual noise and the original noise component.

The baseline eIRM MSE shows very strong noise attenuation reflected by a high ΔSNR score at the cost of limited residual noise and speech component qualities (high WLAKR and low PESQ($\tilde{\mathbf{s}}$) scores). This can be caused by the overestimated noise power in the target IRM for our selected $\alpha = 0.75$. As expected, the performance of the

Table 6 Performance for **seen noise types** (PED and CAFE) on the *test set*; **All SNRs** averaged. Best approaches from Tables 1 and 3 are gray-shaded; the best scheme is in boldface

Noise	Method	Noise component		Speech component		Total		
		ΔSNR	WLAKR	SSDR	PESQ($\tilde{\mathbf{s}}$)	PESQ($\hat{\mathbf{s}}$)	POLQA($\hat{\mathbf{s}}$)	STOI
PED	Baseline MSE	5.84	0.24	11.70	2.94	2.42	1.92	0.70
	Baseline eIRM MSE	6.89	0.33	11.43	2.94	2.65	**2.13**	0.69
	Baseline PW-FILT	6.37	0.45	11.52	2.87	2.53	1.91	0.70
	Baseline PW-PESQ	5.81	**0.16**	11.84	2.96	2.47	1.89	0.70
	Baseline two-masks SNR	6.92	0.37	12.27	2.88	2.59	2.08	0.70
	Reference iIRM MSE	6.90	0.21	12.24	2.97	2.64	2.10	0.69
	2CL ($\alpha = 0.5$)	6.18	0.22	**12.34**	**3.04**	**2.67**	2.11	**0.71**
	3CL ($\alpha = 0.1, \beta = 0.8$)	**7.05**	0.18	12.21	3.00	**2.67**	**2.13**	**0.71**
CAFE	Baseline MSE	5.76	0.26	11.44	2.87	2.33	1.90	0.69
	Baseline eIRM MSE	7.27	0.23	11.45	2.89	2.57	2.16	0.69
	Baseline PW-FILT	6.32	0.60	11.42	2.84	2.45	1.93	0.69
	Baseline PW-PESQ	5.78	0.21	11.57	2.90	2.35	1.90	0.69
	Baseline two-masks SNR	7.14	0.20	12.12	2.91	2.53	2.10	0.70
	Reference iIRM MSE	7.20	**0.13**	12.10	3.00	2.58	2.11	0.69
	2CL ($\alpha = 0.5$)	7.22	**0.13**	**12.20**	**3.05**	2.60	2.12	**0.70**
	3CL ($\alpha = 0.1, \beta = 0.8$)	**7.30**	**0.13**	12.10	3.03	**2.62**	**2.17**	**0.70**

Table 7 Performance for **seen noise types** (PED and CAFE) on the *test set*; **SNR= − 5 dB**. Best approaches from Tables 1 and 3 are gray-shaded; the best scheme is in boldface

Noise	Method	Noise component		Speech component		Total		
		ΔSNR	**WLAKR**	**SSDR**	**PESQ($\tilde{\mathbf{s}}$)**	**PESQ($\hat{\mathbf{s}}$)**	**POLQA($\hat{\mathbf{s}}$)**	**STOI**
PED	Baseline MSE	6.10	0.24	3.03	1.83	1.44	1.07	0.49
	Baseline eIRM MSE	8.46	0.50	2.50	2.00	1.53	1.12	0.48
	Baseline PW-FILT	8.17	0.30	2.73	1.76	1.46	1.14	0.50
	Baseline PW-PESQ	6.43	**0.11**	3.00	1.85	1.45	**1.35**	**0.51**
	Baseline two-masks SNR	**8.60**	0.44	2.99	1.74	1.46	1.12	0.49
	Reference iIRM MSE	8.19	0.31	3.04	2.02	1.52	1.18	0.48
	2CL ($\alpha = 0.5$)	8.25	0.21	**3.10**	**2.09**	**1.58**	1.28	0.50
	3CL ($\alpha = 0.1, \beta = 0.8$)	8.58	0.21	2.97	2.05	**1.58**	1.23	0.50
CAFE	Baseline MSE	7.56	0.13	2.74	1.76	1.39	**1.22**	0.49
	Baseline eIRM MSE	10.06	0.28	2.40	1.96	1.55	1.08	0.48
	Baseline PW-FILT	8.99	0.41	2.46	1.71	1.42	1.07	0.49
	Baseline PW-PESQ	7.74	0.12	2.75	1.81	1.42	**1.22**	**0.51**
	Baseline two-masks SNR	9.84	0.23	2.76	1.71	1.43	1.09	0.49
	Reference iIRM MSE	9.96	0.13	2.81	1.99	1.51	1.13	0.49
	2CL ($\alpha = 0.5$)	**10.15**	0.11	**2.88**	2.04	**1.57**	1.16	0.50
	3CL ($\alpha = 0.1, \beta = 0.8$)	9.93	**0.10**	2.84	**2.06**	1.54	1.17	0.50

eIRM MSE is more unbalanced compared to the reference iIRM MSE method, which has already been shown in [17, 18]. Since both POLQA and PESQ measured on the overall enhanced speech seem to favor high noise attenuation (high ΔSNR score), the baseline eIRM MSE still offers good results on these measures, on average, however with the worst intelligibility (reflected by the lowest STOI score).

Compared to our proposed CLs, the baseline two-masks SNR loss offers a similarly strong ΔSNR on average. This is expected, since the goal of the SNR loss is to maximize the output SNR of the trained network. Similar to the MSE loss, however, the SNR loss leads to an unnatural-sounding residual noise, reflected by a high WLAKR score, and a poor speech component quality, reflected by a low PESQ($\tilde{\mathbf{s}}$) score. This behavior is particularly obvious for PED noise. Since PESQ($\hat{\mathbf{s}}$) favors high noise attenuation (reflected by high ΔSNR), the two-masks SNR loss offers a good PESQ($\hat{\mathbf{s}}$) score among the baselines. However, comparing to our proposed CLs, the two-masks SNR loss clearly falls behind both in PESQ($\hat{\mathbf{s}}$) and in POLQA($\hat{\mathbf{s}}$). Note that the network trained by the two-masks SNR loss also has more parameters due to the two decoder heads, compared to the network trained by our proposed CLs.

As introduced before, the CNN trained with the baseline MSE tends to attenuate regions with very low SNR to optimize the global MSE [27], which may lead to strong noise distortion and speech component distortion. The proposed 2CL penalizes this speech component distortion by the first term of (8), weighted by $1-\alpha$, which is not only good for preserving the speech component quality, but also for maintaining a naturally sounding residual noise. The CNNs trained by our proposed 2CL and 3CL by far provide the best speech component quality, which is reflected by a higher SSDR and about 0.1 higher PESQ($\tilde{\mathbf{s}}$) on average, compared to all the baseline losses. This is attributed to the first term of 2CL (8) and 3CL (11), which is the loss function for the *filtered* speech component, and is supposed to preserve detailed structures of the speech signal and punishes the attenuation of the speech component. Among the CNNs trained by the components losses, 2CL offers slightly better PESQ($\tilde{\mathbf{s}}$) and about 0.1 dB higher SSDR compared to 3CL. One possible reason is that the weight for speech distortion in 3CL (11) represented by 1–0.1–0.8 = 0.1 is less compared to the one in 2CL (8) represented by $1-0.5=0.5$. *Our proposed 2CL and 3CL losses provide the best overall enhanced speech quality, which is reflected by obtaining the highest PESQ($\hat{\mathbf{s}}$) and POLQA($\hat{\mathbf{s}}$) scores.* In addition to that, 2CL and 3CL obtain slightly better speech intelligibility reflected by 0.01 higher STOI score for *seen* noise types on average. Among the CL-based CNNs, 3CL is better by offering a stronger noise attenuation, a more natural residual noise, and the best enhanced speech quality, yielding a more balanced performance.

The performance on the *seen* noise types at SNR= −5 dB is shown in Table 7. Again, the baseline losses two-masks SNR, eIRM MSE, and the PW-FILT can offer very good SNR improvement comparable to our proposed CLs,

Table 8 Performance for **unseen noise** (BUS) on the *test set*; **All SNRs** averaged. Best approaches from Tables 1 and 3 are gray-shaded; the best scheme is in boldface

Noise	Method	Noise component		Speech component		Total		
		ΔSNR	**WLAKR**	**SSDR**	**PESQ($\tilde{\mathbf{s}}$)**	**PESQ($\hat{\mathbf{s}}$)**	**POLQA($\hat{\mathbf{s}}$)**	**STOI**
BUS	Baseline MSE	4.50	0.20	13.14	3.03	2.39	2.39	0.71
	Baseline eIRM MSE	6.19	0.24	14.00	3.31	2.63	2.62	0.75
	Baseline PW-FILT	5.56	0.39	13.43	3.12	2.50	2.35	**0.75**
	Baseline PW-PESQ	4.73	0.19	13.27	3.00	2.42	2.34	**0.75**
	Baseline two-masks SNR	5.96	0.23	14.28	**3.38**	2.64	2.57	**0.75**
	Reference iIRM MSE	6.08	0.19	14.25	3.34	2.61	2.61	0.74
	2CL ($\alpha = 0.5$)	5.60	**0.18**	**14.30**	**3.38**	2.63	**2.64**	0.74
	3CL ($\alpha = 0.1, \beta = 0.8$)	**6.28**	**0.18**	14.23	3.35	**2.68**	**2.64**	**0.75**

however, with worse residual noise and speech component qualities, reflected by very high WLAKR and very low PESQ($\tilde{\mathbf{s}}$) scores. For CAFE noise, the proposed 2CL shows higher ΔSNR compared to 3CL. The same as in Table 6, the proposed 3CL and the baseline PW-PESQ provide the best residual noise quality for CAFE and PED noise, respectively. The proposed 2CL and 3CL offer the best speech component quality (PESQ($\tilde{\mathbf{s}}$)) and overall enhanced speech quality (PESQ($\hat{\mathbf{s}}$)). At SNR= −5 dB, the CNN trained by the PW-PESQ loss offers slightly better speech intelligibility reflected by 0.01 higher STOI score compared to the CNNs trained by other losses. The two-masks SNR loss offers 0.07…0.16 points lower POLQA($\hat{\mathbf{s}}$) and PESQ($\hat{\mathbf{s}}$) scores compared to our proposed CLs, due to the strong distortions of the residual noise and the speech components. The performance of the baseline eIRM MSE becomes even more unbalanced in very harsh SNR conditions, which leads to a poor POLQA($\hat{\mathbf{s}}$) score and the lowest STOI score.

Now, let us have a look into the reference iIRM MSE in Tables 6 and 7. We observe that it offers better performance compared to the baseline MSE for most of the measures. This shows that the implicit constraint of the ideal mask (9) is advantageous for reference iIRM MSE, although on high level, it employs the same loss (2) as the baseline MSE. Comparing the performance of the reference iIRM MSE method with our proposed 2CL on the *seen* noise types, the reference iIRM MSE averaged over SNRs offers similar ΔSNR and WLAKR scores compared to our proposed 2CL. In low SNR, the reference iIRM MSE shows a worse residual noise quality compared to the 2CL, as can be seen in Table 7. This is particularly prominent in PED noise, reflected by a 0.1 points higher WLAKR score. In both Tables 6 and 7, our proposed 2CL offers better speech component quality than the reference iIRM MSE, reflected by higher SSDR and PESQ($\tilde{\mathbf{s}}$) scores. Even more, 2CL and 3CL offer consistently better overall enhanced speech quality compared to the reference iIRM MSE.

5.2.2 Unseen noise

The performance on the *unseen* BUS noise is shown in Tables 8 and 9. Again, among the baseline losses, the eIRM MSE, the PW-FILT, and the two-masks SNR always provide quite strong ΔSNR, however, with very low residual noise quality (high WLAKR score). The same as before,

Table 9 Performance for **unseen noise** (BUS) on the *test set*; **SNR= − 5 dB**. Best approaches from Tables 1 and 3 are gray-shaded; the best scheme is in boldface

Noise	Method	Noise component		Speech component		Total		
		ΔSNR	**WLAKR**	**SSDR**	**PESQ($\tilde{\mathbf{s}}$)**	**PESQ($\hat{\mathbf{s}}$)**	**POLQA($\hat{\mathbf{s}}$)**	**STOI**
BUS	Baseline MSE	6.99	0.22	4.92	2.11	1.55	1.59	0.61
	Baseline eIRM MSE	**9.09**	0.37	5.24	2.54	1.75	1.57	0.63
	Baseline PW-FILT	8.03	0.24	5.07	2.17	1.58	1.40	**0.64**
	Baseline PW-PESQ	7.05	**0.16**	4.86	2.05	1.55	1.58	**0.64**
	Baseline two-masks SNR	8.54	0.33	5.66	2.50	1.70	1.54	**0.64**
	Reference iIRM MSE	8.55	0.27	5.67	2.55	1.73	1.58	0.63
	2CL ($\alpha = 0.5$)	8.31	0.20	5.66	2.57	1.73	**1.63**	0.63
	3CL ($\alpha = 0.1, \beta = 0.8$)	8.56	0.21	**5.67**	**2.60**	**1.77**	1.61	0.63

2CL and 3CL provide both very good noise attenuation and residual noise quality. On average, the CNN trained by 3CL offers the highest ΔSNR compared to the ones trained by other losses. The PW-PESQ loss offers very good residual noise quality, sometimes even ranking best. Again, the proposed CLs can offer the best speech component quality (SSDR, PESQ($\tilde{\mathbf{s}}$)) and total enhanced speech quality (PESQ($\hat{\mathbf{s}}$), POLQA($\hat{\mathbf{s}}$)). Averaged over SNR conditions (Table 8), the proposed 3CL provides better overall enhanced speech quality reflected by an about 0.2 points higher PESQ($\hat{\mathbf{s}}$) compared to the baseline MSE and the auditory-related losses. Unlike the performance on the *seen* noise types, the two-masks SNR can offer a comparable overall enhanced speech quality to our 2CL, however, still falls behind our 3CL. Except for the baseline MSE, the remaining baseline losses and our proposed CLs provide very comparable speech intelligibility as shown in the last column of Tables 8 and 9. As before, the proposed 3CL performs best by offering good and balanced results.

Similarly, the reference iIRM MSE offers lower residual noise quality (particularly in low-SNR condition) and speech component quality (averaged over SNR conditions) compared to our proposed 2CL. Therefore, although it is somehow better in ΔSNR, it slightly falls behind in the total enhanced speech quality measured by PESQ($\hat{\mathbf{s}}$) and POLQA($\hat{\mathbf{s}}$). To sum up, the network trained by our proposed 2CL is better than the reference iIRM MSE. We see by experimental evidence that in gradient-based learning, the two different loss formulations (10) and (8) offer different performance, even though they share the same global optimum (9).

5.2.3 Summary of results and discussion

In total, the CNN trained by our proposed components loss offers the best speech component quality for both *seen* and *unseen* noise types, in both averaged and very harsh noise conditions. At the same time, the two proposed CLs also mostly offer the highest ΔSNR, the best speech component quality, as well as a very good, in some cases even the best residual noise quality. So, the CNN trained by our CLs show both a strong and a balanced performance by not only providing a strong noise attenuation, but also providing a naturally sounding residual noise, and a less distorted speech component. Likely from the contribution of all these aspects, our proposed CLs also provide the best enhanced speech quality and speech intelligibility in almost all experiments. Meanwhile, the investigated baselines have problems with at least one of the employed quality measures. Surprisingly, compared to the 2CL results, the additional third term in 3CL (11), which is supposed to preserve good residual noise quality, not only provides the same, sometimes even a better residual quality, but also indirectly increases noise attenuation during training. In total, the CNN trained by our 3CL offers the best and the most balanced performance.

In Table 10, we provide a final overview of the methods over all metrics averaged, by simply showing how often

Table 10 Test set performance ranking for both **seen noise types** (PED and CAFE) and **unseen noise types** (BUS), based on how many times each loss function provides the best scores in all employed instrumental metrics. Best approaches from Tables 1 and 3 are gray-shaded; the best scheme is in boldface

Noise	Method	Times of best scores in employed metrics		
		All SNRs averaged	SNR= −5 dB	In total
Seen noise types	Baseline MSE	0	1	1
	Baseline eIRM MSE	1	0	1
	Baseline PW-FILT	0	0	0
	Baseline PW-PESQ	1	5	6
	Baseline two-masks SNR	0	1	1
	Reference iIRM MSE	1	0	1
	2CL ($\alpha = 0.5$)	8	**6**	**14**
	3CL ($\alpha = 0.1, \beta = 0.8$)	**9**	3	12
Unseen noise types	Baseline MSE	0	0	0
	Baseline eIRM MSE	0	1	1
	Baseline PW-FILT	1	1	2
	Baseline PW-PESQ	1	2	3
	Baseline two-masks SNR	2	1	3
	Reference iIRM MSE	0	0	0
	2CL ($\alpha = 0.5$)	4	1	5
	3CL ($\alpha = 0.1, \beta = 0.8$)	**5**	**3**	**8**

each method scores best (boldface numbers) in Tables 6 and 7 (seen noise types) and Tables 8 and 9 (unseen noise types). We easily see the dominance and balanced performance of both 2CL and 3CL for all noise types in the SNR averaged case. In very low SNR, 2CL is best on seen noise types, while 3CL generalizes better and is best on unseen noise types. The best among the baseline methods in this analysis seems to be PW-PESQ, however, with mediocre PESQ and POLQA performance as can be seen in Tables 6 to 9.

6 Conclusions

In this paper, we illustrated the benefits of a components loss (CL) for mask-based speech enhancement. We introduced the 2-components loss (2CL), which controls speech component distortion and noise suppression separately, and also the 3-components loss (3CL), which includes an additional term to control the residual noise quality. Our proposed 2CL and 3CL are naturally differentiable for gradient-based learning and do not need any additional training material or extensive computational effort compared to, e.g., auditory-related loss functions. Furthermore, we point out that these new loss functions are not limited to any specific network topology or application. In the context of a speech enhancement framework that uses a convolutional neural network (CNN) to estimate a spectral mask, our new CL formulations provide better or at least more balanced performance across various instrumental quality measures than the investigated baselines. For *unseen* noise types, we excel even perceptually motivated losses by an about 0.2 points higher PESQ score. Averaged over all SNR conditions and all metrics combined, both 2CL and 3CL show significantly more 1st rank results than any of the baseline losses. The recently proposed so-called SNR loss with two masks not only requires a network with more parameters due to the two decoder heads, but also falls behind on PESQ and POLQA and particularly w.r.t. residual noise quality. The new 2CL and 3CL loss functions are easy to implement, and example code is provided at https://github.com/ifnspaml/Components-Loss.

Appendix 1

Baseline PW-FILT: The perceptual weighting filter applied in this loss function is borrowed from CELP speech coding, e.g., the adaptive multi-rate (AMR) codec [60], in order to shape the coding noise / quantization error to be less audible by the human ear. This weighting filter is calculated according to [60] as

$$W_\ell(z) = \frac{1 - A_\ell(z/\gamma_1)}{1 - A_\ell(z/\gamma_2)}, \tag{14}$$

with the predictor polynomial $A_\ell(z/\gamma) = \sum_{i=1}^{N_\mathrm{p}} a_\ell(i)\gamma^i z^{-i}$, $a_\ell(i)$ are the linear prediction (LP) coefficients of frame ℓ, N_p is the prediction order, and γ_1, γ_2 are the perceptual weighting factors. During the search of the codebooks in CELP encoding, the error between the clean speech and the coded speech is weighted by the weighting filter and subsequently minimized. As a result, the weighted error becomes spectrally white, meaning that the final (unweighted) quantization error has a frequency distribution that is proportional to the frequency characteristics of the *inverse* weighting filter $1/W_\ell(z)$, which has similarities to the shape of the clean speech spectral envelope. This property of the weighting filter allows to exploit the masking effect of the human ear: More energy of the quantization error will be placed in the speech formant region, where $1/W_\ell(z)$ is at some level below the spectral envelope [29].

After the original CELP weighting filter has been revisited, the corresponding perceptual weighting filter loss is now straightforward, as shown in (4), where both $\hat{S}_\ell(k)$ and $S_\ell(k)$ are effectively weighted by the weighting filter frequency response

$$W_\ell(k) = W_\ell(z)\Big|_{z=e^{j\frac{2\pi k}{K}}}. \tag{15}$$

Similar to the original application of the weighting filter in speech coding, where the quantization error becomes less audible, the residual noise is also expected to be less audible compared to using the MSE loss. As a result, improved perceptual quality of the enhanced speech has been reported in [29].

Baseline PW-PESQ: As with the standard PESQ, the PESQ loss as proposed in [31] consists of a symmetrical and an asymmetrical distortion, both are computed frame-by-frame in the loudness spectrum domain, which is closer to human perception [31]. The authors of [31] adopt the transformation operations from the amplitude spectrum domain to the loudness spectrum domain for the target and enhanced speech signals directly from the PESQ standard [41]. The symmetrical distortion $L_\ell^{(s)}$ for frame ℓ is obtained directly from the difference between the target and enhanced speech loudness spectra. Auditory masking effects should also be considered in calculating $L_\ell^{(s)}$. The corresponding asymmetrical distortion $L_\ell^{(a)}$ is computed based on the symmetrical distortion $L_\ell^{(s)}$, but weighting the positive and negative loudness differences differently. Since human perceptions of the positive and negative loudness differences are not the same, different auditory masking effects must be considered, respectively. Then, the PESQ loss is defined as:

$$J_\ell^{\mathrm{PESQ}} = \theta_1 \cdot L_\ell^{(s)} + \theta_2 \cdot L_\ell^{(a)}, \tag{16}$$

where θ_1 and θ_2 are the weighting factors, and are set to 0.1 and 0.0309, respectively [31]. Since J_ℓ^{PESQ} is highly non-

linear and not fully differentiable, the authors propose to combine the PESQ loss with the conventional MSE that is fully differentiable as the final loss to make the gradient-based learning more stable. Thus, the used loss function for training is defined as (5) as proposed in [31].

Baseline two-masks SNR: During the training process, the proposed SNR loss J_ℓ^{SNR} is calculated for each output decoder head separately and is summed up as the final loss [39]. We use the J_ℓ^{SNR} calculated on the speech output decoder head as an example. To mitigate the power scaling problem as mentioned before, the power-law compression is applied to the speech amplitude spectrum of the predictions and the targets separately [39]. Then, the negative SNR loss is calculated by:

$$-J_\ell^{\text{SNR}} = -10 \cdot \Bigg[\log_{10}\left(\sum_{k\in\mathcal{K}}\left(\left|S_\ell(k)\right|^{1/2}\right)^2\right) - \log_{10}\left(\sum_{k\in\mathcal{K}}\left(\left|\hat{S}_\ell(k)\right|^{1/2} - \left|S_\ell(k)\right|^{1/2}\right)^2\right)\Bigg], \tag{17}$$

which is minimized during the training process. By replacing $\hat{S}_\ell(k)$ and $S_\ell(k)$ with $\hat{D}_\ell(k)$ and $D_\ell(k)$, we can obtain the negative SNR loss for the noise output decoder head.

Please note that the SNR as used in (17) is different from the SNR definition in Section IV.B, which is measured after ITU-T P.56 [56], based on $\tilde{s}(n)$, $\tilde{d}(n)$ and $s(n)$, $d(n)$, respectively. The value range of J_ℓ^{SNR} is in between $-\infty$ and $+\infty$, which can be problematic in stabilizing the training process. To mitigate this problem, a compression function $20 \cdot \tanh\left(J_\ell^{\text{SNR}}\right)$ is used in [39] for optimization to limit the SNR value between -20 and 20 dB.

Appendix 2

Gradient-based optimization requires differentiation. To obtain the differentiation of the proposed 2CL (8) w.r.t. $M_\ell(k)$, we need to replace $\tilde{S}_\ell(k)$ and $\tilde{D}_\ell(k)$ in (8) by (6) and (7), respectively, resulting in:

$$J_\ell^{\text{2CL}} = (1-\alpha)\cdot\sum_{k\in\mathcal{K}}\left(\left|M_\ell(k)\cdot S_\ell(k)\right| - \left|S_\ell(k)\right|\right)^2 + \alpha\cdot\sum_{k\in\mathcal{K}}\left(\left|M_\ell(k)\cdot D_\ell(k)\right|\right)^2. \tag{18}$$

Since a sigmoid activation function is used for the output layer of the employed CNN, the estimated mask $M_\ell(k)$ has a value between 0 and 1, so $\left|M_\ell(k)\right|$ and $M_\ell(k)$ are the same. Then, we obtain:

$$\begin{aligned}\frac{\partial J_\ell^{\text{2CL}}}{\partial M_\ell(k)} &= 2\cdot(1-\alpha)\cdot\left(M_\ell(k)\cdot\left|S_\ell(k)\right| - \left|S_\ell(k)\right|\right)\cdot\left|S_\ell(k)\right| \\ &\quad + 2\cdot\alpha\cdot M_\ell(k)\cdot\left|D_\ell(k)\right|^2 \\ &= 2\cdot M_\ell(k)\cdot\left[(1-\alpha)\cdot\left|S_\ell(k)\right|^2 + \alpha\cdot\left|D_\ell(k)\right|^2\right] \\ &\quad - 2\cdot(1-\alpha)\cdot\left|S_\ell(k)\right|^2.\end{aligned} \tag{19}$$

By setting (19) to 0, we obtain the optimal mask for our proposed 2CL as:

$$M_\ell^{\text{2CL-opt}}(k) = \frac{(1-\alpha)\cdot\left|S_\ell(k)\right|^2}{(1-\alpha)\cdot\left|S_\ell(k)\right|^2 + \alpha\cdot\left|D_\ell(k)\right|^2}, \tag{20}$$

which results in (9).

Appendix 3

In the reference iIRM MSE loss (10), we can replace $\left|\hat{S}_\ell(k)\right|$ and $\left|\hat{S}_\ell^{\text{target}}(k)\right|$ by $\left|Y_\ell(k)\right|\cdot M_\ell(k)$ and $\left|Y_\ell(k)\right|\cdot M_\ell^{\text{2CL-opt}}(k)$, respectively, with $M_\ell^{\text{2CL-opt}}(k)$ obtained from (9), and rewrite (10) to an equivalent loss formulation as:

$$J_\ell^{\text{iIRM}} = \sum_{k\in\mathcal{K}}\left|Y_\ell(k)\right|^2 \cdot\left(M_\ell(k) - \frac{\left|S_\ell(k)\right|^2}{\left|S_\ell(k)\right|^2 + \frac{\alpha}{1-\alpha}\cdot\left|D_\ell(k)\right|^2}\right)^2. \tag{21}$$

Similarly, we can derive the equivalent formulation of our proposed 2CL (8) with the help of (6) and (7) as:

$$\begin{aligned}J^{\text{2CL}} &= \sum_{k\in\mathcal{K}}(1-\alpha)\cdot\left(\left|S_\ell(k)\right|\cdot M_\ell(k) - \left|S_\ell(k)\right|\right)^2 \\ &\quad + \sum_{k\in\mathcal{K}}\alpha\cdot\left(\left|D_\ell(k)\right|\cdot M_\ell(k)\right)^2 \\ &= \sum_{k\in\mathcal{K}}\left((1-\alpha)\cdot\left|S_\ell(k)\right|^2 + \alpha\cdot\left|D_\ell(k)\right|^2\right) \\ &\quad\cdot\left(M_\ell(k) - \frac{\left|S_\ell(k)\right|^2}{\left|S_\ell(k)\right|^2 + \frac{\alpha}{1-\alpha}\cdot\left|D_\ell(k)\right|^2}\right)^2 + \text{constant}.\end{aligned} \tag{22}$$

Note, for our selected $\alpha = 0.5$, (21) becomes equivalent to:

$$J_\ell^{\text{iIRM}} = \sum_{k\in\mathcal{K}}\left|Y_\ell(k)\right|^2\cdot\left(M_\ell(k) - \frac{\left|S_\ell(k)\right|^2}{\left|S_\ell(k)\right|^2 + \left|D_\ell(k)\right|^2}\right)^2, \tag{23}$$

while (22)—apart from a factor $\frac{1}{2}$—becomes equivalent to:

$$J^{\mathrm{2CL}} = \sum_{k\in\mathcal{K}} \left(\left|S_\ell(k)\right|^2 + \left|D_\ell(k)\right|^2\right) \cdot \left(M_\ell(k) - \frac{\left|S_\ell(k)\right|^2}{\left|S_\ell(k)\right|^2 + \left|D_\ell(k)\right|^2}\right)^2 + \text{constant}. \tag{24}$$

Since the constant term will vanish in the gradient computation, the "only" difference between (23) and (24) is the weights of the squared mask error, which are $\left|Y_\ell(k)\right|^2$ and $\left(\left|S_\ell(k)\right|^2 + \left|D_\ell(k)\right|^2\right)$, respectively. At this point, it is very important to note that this difference is highly relevant, particularly in the important low-SNR region. In case $S_\ell(k) = -D_\ell(k)$ (complex numbers!), for a certain time-frequency bin (ℓ, k), we obtain $Y_\ell(k) = 0$ in (23), leading to a zero weight in the loss J_ℓ^{iIRM} (23). Such situation at least approximately often occurs in low-SNR conditions. In consequence, no backpropagation and accordingly no learning takes place, while with our 2CL (24), the weight is even higher, the lower the local SNR in that time-frequency bin gets. This seems to be the crucial advantage of our proposed 2CL vs. the so-called "Reference iIRM MSE." This drawback of (23) can be observed by the typically lower residual noise quality of a noise reduction at low-SNR conditions, while the 2CL ((24) or (22)) reveals quite good residual noise quality.

Abbreviations

CL: Components loss; CNN: Convolutional neural networks; MSE: Mean squared error; PESQ: Perceptual evaluation of speech quality; SNR: Signal-to-noise ratio; TF: Time-frequency; STOI: Short-time objective intelligibility; OLA: Overlap add; IRM: Ideal ratio mask; eIRM: Explicit ideal ratio mask; iIRM: Implicit ideal ratio mask; PW-FILT: Perceptual weighting filter loss; PW-PESQ: Perceptual evaluation of speech quality (PESQ) loss; PW-STOI: Short-time objective intelligibility (STOI) loss; CELP: Code-excited linear prediction; POLQA: Perceptual objective listening quality prediction; FFT: Fast Fourier transformation; DFT: Discrete Fourier transform; IFFT: Inverse fast Fourier transformation; SSDR: Segmental speech-to-speech-distortion ratio; NA: Noise attenuation; WLAKR: Weighted log-average kurtosis ratio; IDFT: Inverse discrete Fourier transform; STFT: Short-time Fourier transform

Acknowledgements

Not applicable.

Authors' contributions

All authors contributed to the conception and design of the experiments and the interpretation of the simulation results. ZX had the initial idea of the components loss, wrote the software, performed the experiments and data analysis, and wrote the first draft of the manuscript. ZZ provided the code and supported the implementation of the baseline PW-FILT. TF substantially revised both experiments and the manuscript. SE contributed additional revisions of the text. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Availability of data and materials

The data supporting the findings of this study are the Grid corpus [54] and the CHiME-3 dataset [55]. The Grid corpus is free for research, and the website of the complete corpus can be frond in [54]. The CHiME-3 dataset is available from the ChiME-3 Speech Separation and Recognition Challenge. Restrictions apply to the availability of these data, which were used under license for the current study and so is not publicly available. Data is however available from the authors upon reasonable request and with permission of ChiME-3 Speech Separation and Recognition Challenge organizers. The code for the newly proposed CL is provided at https://github.com/ifnspaml/Components-Loss.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 20 November 2020 Accepted: 15 March 2021
Published online: 02 July 2021

References

1. Y. Ephraim, D. Malah, Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
2. Y. Ephraim, D. Malah, Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **33**(2), 443–445 (1985)
3. P. Scalart, J. V. Filho, in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. Speech enhancement based on a priori signal to noise estimations (IEEE, Atlanta, GA, USA, 1996), pp. 629–632
4. T. Lotter, P. Vary, Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. EURASIP J. Appl. Signal Process. **2005**(7), 1110–1126 (2005)
5. B. Fodor, T. Fingscheidt, in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Speech enhancement using a joint map estimator with Gaussian mixture model for (non-) stationary noise (IEEE, Prague, Czech Republic, 2011), pp. 4768–4771
6. I. Cohen, Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation. Speech Commun. **47**(3), 336–350 (2005)
7. T. Gerkmann, C. Breithaupt, R. Martin, Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors. IEEE Trans. Audio Speech Lang. Process. **16**(5), 910–919 (2008)
8. S. Suhadi, C. Last, T. Fingscheidt, A data-driven approach to a priori SNR estimation. IEEE Trans. Audio Speech Lang. Process. **19**(1), 186–195 (2011)
9. S. Elshamy, N. Madhu, W. J. Tirry, T. Fingscheidt, in *Sixteenth Annual Conference of the International Speech Communication Association*. An iterative speech model-based a priori SNR estimator (ISCA, Dresden, Germany, 2015), pp. 1740–1744
10. S. Elshamy, N. Madhu, W. Tirry, T. Fingscheidt, Instantaneous a priori SNR estimation by cepstral excitation manipulation. IEEE/ACM Trans. Audio Speech Lang. Process. **25**(8), 1592–1605 (2017)
11. D. Malah, R. V. Cox, A. J. Accardi, in *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments (IEEE, Phoenix, AZ, USA, 1999), pp. 789–792
12. T. Fingscheidt, S. Suhadi, in *Proc. of ITG Conf. on Speech Communication*. Data-driven speech enhancement (ITG, Kiel, Germany, 2006), pp. 1–4
13. T. Fingscheidt, S. Suhadi, S. Stan, Environment-optimized speech enhancement. IEEE Trans. Audio Speech Lang. Process. **16**(4), 825–834 (2008)
14. J. Erkelens, J. Jensen, R. Heusdens, in *2006 14th European Signal Processing Conference*. A general optimization procedure for spectral speech enhancement methods (EURASIP, Florence, Italy, 2006), pp. 1–5
15. J. Erkelens, J. Jensen, R. Heusdens, A data-driven approach to optimizing spectral speech enhancement methods for various error criteria. Speech Commun. **49**(7-8), 530–541 (2007)
16. Y. Wang, A. Narayanan, D. L. Wang, On training targets for supervised speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **22**(12), 1849–1858 (2014)
17. F. Weninger, J. R. Hershey, J. Le Roux, B. Schuller, in *2014 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. Discriminatively trained recurrent neural networks for single-channel speech separation (IEEE, Atlanta, GA, USA, 2014), pp. 577–581

18. P. S. Huang, M. Kim, M. H. Johnson, P. Smaragdis, in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Deep learning for monaural speech separation (IEEE, Florence, Italy, 2014), pp. 1562–1566
19. H. Erdogan, J. R. Hershey, S. Watanabe, J. Le Roux, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks (IEEE, Brisbane, QLD, Australia, 2015), pp. 708–712
20. Y. Wang, D. L. Wang, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. A deep neural network for time-domain signal reconstruction (IEEE, Brisbane, QLD, Australia, 2015), pp. 4390–4394
21. D. S. Williamson, Y. Wang, D. L. Wang, Complex ratio masking for monaural speech separation. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(3), 483–492 (2016)
22. F. Bao, W. H. Abdulla, A new ratio mask representation for CASA-based speech enhancement. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(1), 7–19 (2019)
23. M. Strake, B. Defraene, K. Fluyt, W. Tirry, T. Fingscheidt, in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. Separated noise suppression and speech restoration: LSTM-based speech enhancement in two stages (IEEE, New Paltz, NY, USA, 2019), pp. 239–243
24. M. Strake, B. Defraene, K. Fluyt, W. Tirry, T. Fingscheidt, in *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Fully convolutional recurrent networks for speech enhancement (IEEE, Barcelona, Spain, 2020), pp. 6674–6678
25. D. L. Wang, J. T. Chen, Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018)
26. J. Du, Y. Tu, L. R. Dai, C. H. Lee, A regression approach to single-channel speech separation via high-resolution deep neural networks. IEEE/ACM Trans. Audio Speech Lang. Process. **24**(8), 1424–1437 (2016)
27. P. G. Shivakumar, P. G. Georgiou, in *Interspeech*. Perception optimized deep denoising autoencoders for speech enhancement (ISCA, San Francisco, CA, USA, 2016), pp. 3743–3747
28. Q. J. Liu, W. Wang, P. J. B. Jackson, Y. Tang, in *2017 25th European Signal Processing Conference (EUSIPCO)*. A perceptually-weighted deep neural network for monaural speech enhancement in various background noise conditions (EURASIP, Kos, Greece, 2017), pp. 1270–1274
29. Z. Zhao, S. Elshamy, T. Fingscheidt, in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. A perceptual weighting filter loss for DNN training in speech enhancement, (2019), pp. 229–233
30. C. Brauer, Z. Zhao, D. Lorenz, T. Fingscheidt, in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Learning to dequantize speech signals by primal-dual networks: an approach for acoustic sensor networks (IEEE, Brighton, UK, 2019), pp. 7000–7004
31. J. M. Martín-Doñas, A. M. Gomez, J. A. Gonzalez, A. M. Peinado, A deep learning loss function based on the perceptual evaluation of the speech quality. IEEE Signal Process. Lett. **25**(11), 1680–1684 (2018)
32. Y. Koizumi, K. Niwa, Y. Hioka, K. Kobayashi, Y. Haneda, DNN-based source enhancement to increase objective sound quality assessment score. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1780–1792 (2018)
33. M. Kolbcek, Z. H. Tan, J. Jensen, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Monaural speech enhancement using deep neural networks by maximizing a short-time objective intelligibility measure (IEEE, Calgary, AB, Canada, 2018), pp. 5059–5063
34. G. Naithani, J. Nikunen, L. Bramslow, T. Virtanen, in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*. Deep neural network based speech separation optimizing an objective estimator of intelligibility for low latency applications (IEEE, Tokyo, Japan, 2018), pp. 386–390
35. H. Zhang, X. L. Zhang, G. L. Gao, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Training supervised speech separation system to improve STOI and PESQ directly (IEEE, Calgary, AB, Canada, 2018), pp. 5374–5378
36. S. W. Fu, T. W. Wang, Y. Tsao, X. Lu, 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. **26**(9), 1570–1584 (2018)
37. T. Fingscheidt, P. Vary, in *Proc. of ITG-Fachtagung "Sprachkommunikation"*. Error concealment by softbit speech decoding (ITG, Frankfurt a.M., Germany, 1996), pp. 7–10
38. T. Fingscheidt, P. Vary, Softbit speech decoding: a new approach to error concealment. IEEE Trans. Speech Audio Process. **9**(3), 240–251 (2001)
39. H. Erdogan, T. Yoshioka, in *Interspeech*. Investigations on data augmentation and loss functions for deep learning based speech-background separation (ISCA, Hyderabad, India, 2018), pp. 3499–3503
40. C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. A short-time objective intelligibility measure for time-frequency weighted noisy speech (IEEE, Dallas, TX, USA, 2010), pp. 4214–4217
41. 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 Standardization Sector (ITU-T)*. (ITU-T, 2001)
42. S. Gustafsson, R. Martin, P. Vary, in *Proc. Workshop on Quality Assessment in Speech, Audio, and Image Communication*. On the optimization of speech enhancement systems using instrumental measures (ITG/EURASIP, Darmstadt, Germany, 1996), pp. 36–40
43. T. Fingscheidt, S. Suhadi, in *Interspeech*. Quality assessment of speech enhancement systems by separation of enhanced speech, noise, and echo (ISCA, Antwerp, Belgium, 2007), pp. 818–821
44. H. Yu, T. Fingscheidt, in *Proc. of 5th Biennial Workshop on DSP for In-Vehicle Systems*. A figure of merit for instrumental optimization of noise reduction algorithms (Springer, Kiel, Germany, 2011), pp. 1–8
45. ITU, Rec. P.1100: narrowband hands-free communication in motor vehicles. International Telecommunication Standardization Sector (ITU-T) (2019)
46. ITU, Rec. P.1110: wideband hands-free communication in motor vehicles. International Telecommunication Standardization Sector (ITU-T) (2015)
47. ITU, Rec. P.1130: subsystem requirements for automotive speech services. International Telecommunication Standardization Sector (ITU-T) (2015)
48. Z. Zhao, H. J. Liu, T. Fingscheidt, Convolutional neural networks to enhance coded speech. IEEE/ACM Trans. Audio Speech Lang. Process. **27**(4), 663–678 (2019)
49. Z. Xu, S. Elshamy, T. Fingscheidt, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Using separate losses for speech and noise in mask-based speech enhancement (IEEE, Barcelona, Spain, 2020), pp. 7519–7523
50. Y. Xia, S. Braun, C. K. A. Reddy, H. Dubey, R. Cutler, I. Tashev, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Weighted speech distortion losses for neural-network-based real-time speech enhancement (IEEE, Barcelona, Spain, 2020), pp. 871–875
51. M. Strake, B. Defraene, K. Fluyt, W. Tirry, T. Fingscheidt, in *Proc. Interspeech*. INTERSPEECH 2020 deep noise suppression challenge: a fully convolutional recurrent network (FCRN) for joint dereverberation and denoising (ISCA, Shanghai, China, 2020), pp. 2467–2471
52. S. Z. Fu, C. F. Liao, Y. Tsao, S. D. Lin, in *International Conference on Machine Learning*. MetricGAN: generative adversarial networks based black-box metric scores optimization for speech enhancement, (2019), pp. 2031–2041
53. A. Veit, M. J. Wilber, S. Belongie, in *Proc. of NIPS*. Residual networks behave like ensembles of relatively shallow networks (Curran Associates Inc., Barcelona, Spain, 2016), pp. 550–558
54. M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. **120**(5), 2421–2424 (2006)
55. J. Barker, R. Marxer, E. Vincent, S. Watanabe, in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. The third 'CHiME'speech separation and recognition challenge: dataset, task and baselines (IEEE, Scottsdale, AZ, USA, 2015), pp. 504–511
56. ITU, Rec. P.56: objective measurement of active speech level. International Telecommunication Standardization Sector (ITU-T) (2011)
57. ITU, Rec. P.862.2: corrigendum 1, wideband extension to recommendation P.862 for the assessment of wideband telephone

networks and speech codecs. International Telecommunication Standardization Sector (ITU-T) (2017)
58. ITU, Rec. P.863: perceptual objective listening quality prediction (POLQA). International Telecommunication Union, Telecommunication Standardization Sector (ITU-T) (2018)
59. H. Yu, T. Fingscheidt, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Black box measurement of musical tones produced by noise reduction systems (IEEE, Kyoto, Japan, 2012), pp. 4573–4576
60. 3GPP, Mandatory speech codec speech processing functions; adaptive multi-rate (AMR) speech codec; transcoding functions (3GPP TS 26.090, Rel. 14). 3GPP; TSG SA (2017)

Publisher's Note