Loss Functions for Deep Monaural Speech Enhancement

Jan Freiwald, Lea Schönherr, Christopher Schymura, Steffen Zeiler and Dorothea Kolossa
Ruhr University Bochum
Electrical Engineering and Information Technology
Cognitive Signal Processing Group
Germany
Email: [name].[surname]@rub.de

***Abstract*—Deep neural networks have proven highly effective at speech enhancement, which makes them attractive not just as front-ends for machine listening and speech recognition, but also as enhancement models for the benefit of human listeners. They are, however, usually being trained on loss functions that only assess quality in terms of a minimum mean squared error. This is neglecting the fact that human audio perception functions in a manner far better described by logarithmic measures than linear ones, that psychoacoustic hearing thresholds limit the perceptibility of many signal components in a mixture, and that a degree of continuity of signals may also be expected. Hence, sudden changes in the gain of a system may be detrimental. In the following, we cast these properties of human perception into a form that can aid the optimization of a deep neural network speech enhancement system. We explore their effects on a range of model topologies, showing the efficacy of the proposed modifications.**

***Index Terms*—speech enhancement, denoising, psychoacoustics, slow feature analysis**

I. Introduction

Speech enhancement and denoising have been around for a long time. Early works of spectral subtraction algorithms date back to 1979 [1], and although the algorithms have evolved, the problem has stayed the same; remove noise from speech recordings and make speech more intelligible.

The relevance of this topic, however, increased with the availability of mass-market telecommunication and especially mobile devices, as these are used in nearly all real-world situations and noise conditions. In recent years, this field of research has profited from the availability of massive and affordable computing power, together with an abundance of recorded or realistically generated data, e. g., [2]. This has enabled a proliferation of methods based on artificial neural networks, or more generally driven by the machine-learning idea of autonomously discovering optimal structures and parameter sets, e. g., [3]–[6]. These methods typically need to be trained on large-scale databases with respect to an optimization criterion, which is expressed as a task-appropriate loss function. One of the most commonly used examples of such a loss function is the mean squared error (MSE), the square of the Euclidean distance between the processed data and a clean reference. However, recent works have shown how more specialized functions can be utilized to optimize a neural network model. Hence, for the optimization of audio processing, it is interesting to include psychoacoustic principles and thus to take the human perception into account [7]. For example, the work in [8] optimized speech enhancement algorithms directly on a short-time objective intelligibility (STOI) [9] metric approximation. This approach works on par with, but could not outperform an MSE baseline and thus raises the question of what the best metric may be. Additionally, the framework introduced in [10] showed that incorporating psychoacoustic metrics into the loss function can be beneficial for neural network training. Lastly, a metric that evaluates the contributions of both signal-to-distortion-ratio (SDR) and perceptual-evaluation-of-speech-quality (PESQ) [11] was introduced in [12].

Recently, the idea of slow feature analysis (SFA) [13] has also taken hold in the deep learning community. The goal of SFA is to find a representation of a signal in such a way that this representation is slowly varying over time. References and examples of applications of SFA can be found in the field of computer vision [14], [15], as well as in audio processing [16]. In 2019, a system that uses an SFA loss to optimize for a slow representation of an audio signal was proposed in [17], but this resulted in a collapse of the latent space of the employed auto-encoder. In our work, we propose a normalization of the plain differential SFA loss, which solves this problem.

Not only the loss, but also the topology of the neural network is a pertinent point. A variety of possible model structures has been proposed and evaluated, e. g., a wave-net-based model [17], which optimizes network parameters directly on the time domain signal differences of clean and processed speech. Although wave-net encoders produce very natural sounding signals, they come with the drawback of requiring a large-scale training dataset. Therefore, we will incorporate the general idea of calculating parts of the loss directly on the signal in the time domain, but with a classical short-time Fourier transform instead of the wave-net architecture.

Concretely, we will evaluate a model topology that is based on classical frequency masking algorithms. As shown in [18],

This project has received funding from the European Regional Development Fund (ERDF) and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2092 CASA - 390781972 and under individual grant number KO3434/4-2

it is possible to use time-frequency masking to recover the amplitude spectrum of a speech signal, when it is superimposed by a number of other sources. This becomes possible due to the sparsity of speech, when represented in an appropriate time-frequency domain, a characteristic that is referred to as *W-disjoint orthogonality* by Yilmaz and Rickard.

Many works have exploited this property to separate a speech source of interest from interferers by estimating such a time-frequency mask, e.g. based on beamforming [19] or independent component analysis [20] for multi-channel enhancement, or on statistical properties of speech versus noise [21] or of speech over time [22] for the single-channel case. As we can see in [23], [24], it is also possible, and highly effective, to use a DNN to calculate such a soft mask for the amplitude spectrum of a single-channel input signal. This mask is then used to gate the input magnitude spectrum and it hence should encode the estimated proportion of the clean speech energy in all time-frequency bins.

In contrast to [23], we will use a multi-layered LSTM instead of explicit recurrent smoothing in order to train our setup end-to-end and let the network learn the correct transition rates between masks. Since we are primarily investigating the impact of different loss functions on enhancement quality, we did not try complex model topologies as described in [25]; also we did not look into generative adversarial network topologies [3]. Nevertheless it will be interesting for future research to investigate the impact of phase-aware algorithms such as [25]–[27] and to employ our suggested loss functions in a broader range of architectures.

II. System Design

In the following section, we describe our system architecture and training process as well as the high-level overview of the proposed objective functions.

A. Recurrent Soft Gating Filter

Our proposed models are based on classical speech enhancement structures. As shown in Figure 1, they utilize an STFT to transform the signal into the time-frequency domain, where they apply a point-wise gain to suppress non-speech. A corresponding inverse STFT (iSTFT) is applied to the resulting spectrogram, using the phase of the input signal for reconstruction.

The frame length of the STFT should not exceed the short-time stationarity of speech signals. Our choice of 16 ms frame length and 4 ms frame shift is motivated by this consideration and is also within the typical ranges of frame-lengths in similar approaches, cf. [5], [18]. Since we use a differentiable implementation of this spectral transformation, we can optimize the enhancement system parameters directly on losses that are calculated in the time domain.

All models utilize Long Short-Term Memory (LSTM) [28] cells to control this soft mask, similarly, e. g., to [12], which enables them to take temporal context into account. As shown in Figure 1, it fades out frequency content that is not speech by multiplying the amplitude spectrum with a mask that is

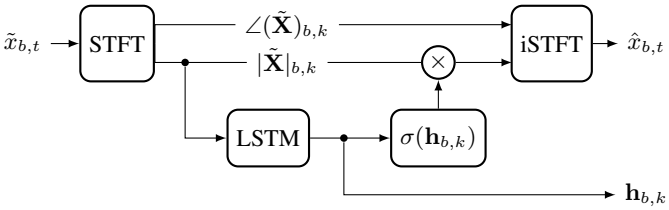

Fig. 1. In the proposed model, an STFT is applied to the input signal. The resulting magnitude spectrum $|\tilde{\mathbf{X}}|_{b,k}$ is multiplied point-wise with a soft mask $\sigma(\mathbf{h}_{b,k})$. We use the phase of the noisy signal $\angle(\tilde{\mathbf{X}})_{b,k}$ for reconstruction. During the training phase, the representation $\mathbf{h}_{b,k}$ is also passed to the loss function.

calculated from the noisy input amplitude spectrum. Moreover, the LSTM topology can be bidirectional, which means that future context can also be considered in the model. While this results in a loss of causality and real-time applicability of the speech enhancement model, it allows for an optimum enhancement.

Lastly, we can reconstruct the clean signal from the masked amplitude in conjunction with the input signal phase. Note that the noisy input phase, which is used to synthesize the signal, is only a stand-in for the clean speech phase, which is unavailable in this context. This has been done similarly in the majority of previous works, see, e. g., [20], [23], [29].

We tested our model with different numbers of LSTM layers and compared the results for uni- and bidirectional LSTMs. The number of LSTM output neurons is set to the number of frequency bins of the spectrogram, or twice this size for bidirectional topologies. The subsequent sigmoid layer generates a frequency-dependent soft-mask, which is used to filter the amplitude of the input signal. We chose the sigmoid function because it ensures a gain smaller than one. This is important since we are working with additive noise and the amplitude spectrum of a single input source is more sparse than the amplitude spectrum of the noisy superposition of multiple sources. Hence, the denoising algorithm should also be designed in such a way that it favors an increase of sparsity. Moreover, the sigmoid function ranges smoothly from 0 to 1, as is desired for the calculation of soft time-frequency masks.

B. Training Process Overview

Our framework is shown in Figure 2. As the primary input, we load a batch of clean speech signals with the indexing structure $x_{b,t}$, where b denotes the batch entry index and t the sample index in the discrete time domain, as encoded in the raw audio file. We use zero-padding to the length of the longest sequence in each batch. We add randomly chosen noise to every clean signal in the batch, which yields the noisy input signals denoted as $\tilde{x}_{b,t}$.

This degraded batch is fed into the denoising model described in Section II-A, which returns the processed signal $\hat{x}_{b,t}$ and a signal representation $\mathbf{h}_{b,k}$. Note that k is a frame index in the spectral domain. In contrast, t is the sample index of the raw audio file.

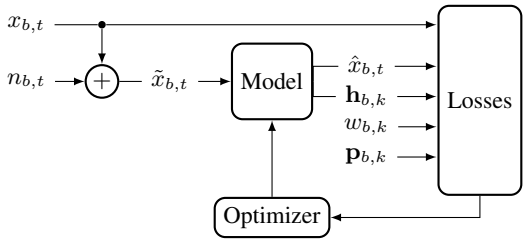

Fig. 2. Depiction of the overall training setup. Noise $n_{b,t}$ is added randomly. The two outputs of the model are the reconstructed signal $\hat{x}_{b,t}$ and the signal representation $\mathbf{h}_{b,k}$. The phoneme boundaries are encoded into the weights $w_{b,k}$, which are used in the loss function. The psychoacoustic hearing thresholds $\mathbf{p}_{b,k}$ are passed to the loss function as well. The optimizer tunes the model parameters.

During training, we use different losses, which are described in Section II-C. Three of these losses measure the distance between $\hat{x}_{b,t}$ and $x_{b,t}$. The fourth loss is calculated on temporal information of the representation $\mathbf{h}_{b,k}$. We implemented the denoising framework with PyTorch [30] and the SciPy stack [31]. The models are trained with the Adam optimizer with decoupled weight decay (AdamW) [32], with a fixed learning rate of 0.0005.

To measure the quality of a denoising algorithm, different metrics are available. Here, we use PESQ and the short-time objective intelligibility measure (STOI) for evaluation. The PESQ metric is standardized by the ITU in recommendation P.862 and is designed to asses the quality of telecommunication lines perceived by human listeners [11].

C. Speech Enhancement Losses

We are interested in designing loss functions that are especially amenable to the goal of speech enhancement. Speech has numerous characteristics that can be utilized for the design of loss functions. Additionally, as we carry out speech enhancement for human listening, it is also important to improve perceptual quality, which implies a set of further desired characteristics, and hence, another set of loss functions. None of the loss functions can, however, be expected to work well in isolation. Therefore, in the course of this study, we conducted experiments with different combinations of optimization goals.

1) Mean Squared Error: The first loss, which we will consider as a baseline loss, is the well known and widely used mean squared error loss (MSE).

2) Mean Squared Logarithmic Error: The second loss is the mean squared logarithmic error (MSLE). Since the original clean signal, as well as the processed signal, are available in the time domain, this loss is calculated as the batch- and sample-wise mean over

$$\mathcal{L}_{\text{MSLE}_{b,t}} \propto \left(\log(|x_{b,t}|+1) - \log(|\hat{x}_{b,t}|+1)\right)^2. \qquad (1)$$

The parts of the sequences that are appended during zero padding are excluded in the computation.

3) Psychoacoustics: The psychoacoustic loss (PSY) approximates the human listening experience by utilizing hearing thresholds. Psychoacoustic hearing thresholds are an effective measure of audibility and describe how the dependencies between frequencies and across time lead to masking effects in human perception [33]–[35].

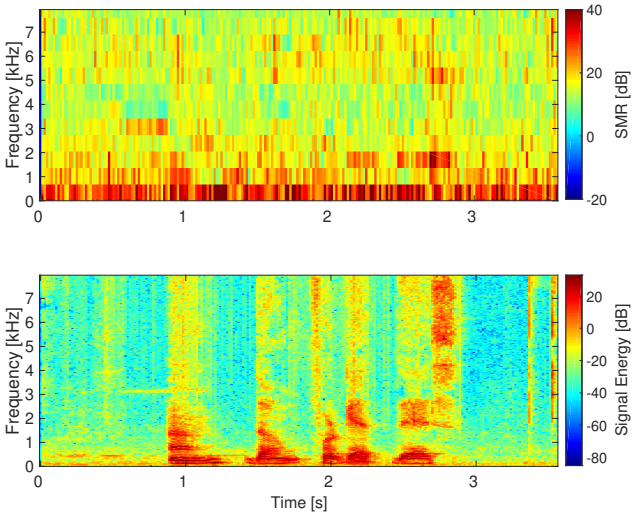

Fig. 3. An example of the signal-to-mask ratio (SMR) (top), shown together with the clean audio spectrogram (bottom), from which it was derived.

The hearing thresholds are therefore useful to identify and penalize those time-frequency bins, that deviate from the clean signal by a perceptible amount. To quantify the hearing thresholds, we calculate the signal-to-mask ratio (SMR) $\mathbf{p}_{b,k}$, which describes the masking effect via the logarithm of the ratio of the clean signal energy to the psychoacoustic masking threshold. Hence, the higher the value of the SMR, the more noise can be added to the signal without being perceivable by human listeners. An example of the SMR and its corresponding audio signal is given in Figure 3.

To obtain the SMR, we utilize the psychoacoustic model of the MP3 compression algorithm [36]. Our proposed loss is defined as the mean over

$$\mathcal{L}_{\text{Psy}_{b,k}} \propto \text{ReLU}\left(20\log_{10}\frac{|\hat{\mathbf{X}}_{b,k}|}{|\mathbf{X}_{b,k}|} - \mathbf{p}_{b,k}\right), \qquad (2)$$

where $\mathbf{X}_{b,k}$ is the target signal and $\hat{\mathbf{X}}_{b,k}$ the estimated signal, both in the frequency domain. Equation (2) thus assesses the degree to which the added noise exceeds the hearing thresholds, and sets the cost of all remaining time-frequency bins, which do not exceed the hearing thresholds, to zero.

4) Slowness Loss: The slowness loss (SLOW) acts as a regularisation term. By penalizing rapid changes in the output nodes of the LSTM in the gating algorithm, we use the idea of slow feature analysis (SFA) under additional consideration of the phoneme annotation. The idea behind this is to allow the filters to change more easily near phoneme boundaries. Therefore, we use a forced alignment tool, which utilizes a trained automatic speech recognizer for the English language to calculate the phoneme transcription of our speech samples [37]. Those transcriptions are only needed in the training phase of the algorithm.

The main goal of SFA is to find a representation of an input signal which varies slowly over time. This optimization for slowness has shown its utility in recent publications and was introduced in [13], [38].

The slowness of the representation, in our case the output of the final LSTM layer $\mathbf{h}_{b,k}$, is calculated via

$$S(\mathbf{h}_b) \propto \sum_{k=1}^{K-1} w_{b,k} ||\mathbf{h}_{b,k+1} - \mathbf{h}_{b,k}||_2^2. \tag{3}$$

The weight w_k is set to zero if h_k is close to a phoneme boundary and to one if h_k is in the middle of a phoneme frame. Specifically, w_k is calculated in the following way: First, we start with the integer encoded phoneme annotation $b_{b,k}$ and generate a sequence of ones and zeros

$$\delta b_{b,k} = \begin{cases} 1 & \text{if } b_{b,k+1} - b_{b,k} \neq 0 \\ 0 & \text{otherwise} \end{cases}. \tag{4}$$

Subsequently, we convolve the resulting sequence $\delta b_{b,k}$ with a trapezoidal window

$$t_k = [0.25, 0.5, 0.75, 1, 1, 1, 0.75, 0.5, 0.25] \tag{5}$$

to smooth the phoneme boundary mask over time. The resulting sequence is called $\delta \hat{b}_{b,k}$. Finally, we set

$$w_{b,k} = 1 - \max\left(1, \delta \hat{b}_{b,k}\right). \tag{6}$$

This ensures that the representation is allowed to change when the spoken phoneme changes. In order to avoid a trivial solution, which results in the collapse of the representation space [17], we apply an additional scaling

$$\mathcal{L}_{\text{Slow}} \propto \sum_{\forall b} \left[\frac{S(\mathbf{h}_b)}{\sum_{k=1}^{K} ||\mathbf{h}_{b,k}||_2^2} \right]. \tag{7}$$

To perform a full SFA, it is common to apply a zero mean and unit variance constraint as well as to require decorrelation and ordering by slowness of the components of the representation. This is usually done by solving an eigenvalue problem. For this work, only slowness will be considered, since it is the most relevant property in the context of a denoising task. Because we are not enforcing unit variance, we are using normalization to prevent the trivial solution.

III. Evaluation

In this section we describe the training configuration and evaluation. We first introduce our baseline methods and database, followed by a description of the evaluation metrics.

A. Baseline Method

As a first baseline, we use the well-known improved minima-controlled recursive averaging (IMCRA) algorithm [39]. Furthermore, we use models that are trained on the MSE loss as a second baseline.

B. Dataset

The Mozilla Common Voice database (MCV) is a community-driven speech corpus that is permanently under development. It consists of over 1,965 hours of spoken and annotated recordings in 29 different languages. The MCV is published under a Creative Commons (CC-0) license and is publicly available. The English sub-corpus consists of 780 hours recorded from approximately 39,577 speakers, cf. [2]. Every utterance of this database is a separate file, which comprises a single sentence. Mozilla provides a word-level transcription for every utterance.

In order to validate our approach, we use a train-test-development split of this database. Cross-validation is not performed due to the excessive training time of the experiments. For this work, we use a subset of 56,843 files recorded from 1,577 different English speakers as our training set. Due to the crowd-sourcing nature of the database, these files vary in quality between studio-quality and low-quality recordings. This range of possible input devices closely resembles recordings in the field. Our development set consists of 11,499 files and our test set contains 100 files. The length of recordings in our database ranges from 2.3 to 7.2 seconds.

For the noise, we use two databases: the open-source DEMAND corpus [40] and the RSG-10 corpus [41]. The DEMAND corpus contains 16-channel recordings of different natural environments, like street or kitchen noise. We only used the first of the 16 channels. The RSG-10 corpus is another collection of noises, stemming from a wide variety of noise conditions.

Analogously to the clean signals, we applied a train-test-development split to the noises as well. Since we are performing noise-reduction speech enhancement, we excluded noise that contains intelligible speech. The noise level for training is chosen uniformly between -10 and 30 dB RMS SNR to cover a wide range of noise levels. Furthermore, the segment of the noise, which is added to the speech signal, is chosen randomly throughout the training.

C. Experimental Setup

We trained our models with different losses and parameter combinations for a maximum of 100 epochs. As combinations of loss functions, we considered MSE, MSLE, MSLE+PSY, MSLE+SLOW, and the combination of all, MSLE+PSY+SLOW. The MSE loss is not taking any psychoacoustics or temporal information into account. It is, therefore, considered as a baseline and not combined with the other metrics.

In a first experiment, we assessed the impact of specific model parameter settings, testing LSTM layers counts from 3 to 4 and comparing bidirectional and unidirectional LSTM topologies on a slightly larger version of our database. Here, we selected the weights for the losses as $w_{\text{MSLE}} = 1, w_{\text{PSY}} = w_{\text{SLOW}} = 0.0001$ and the learning rate of the optimizer as $\alpha = 0.0005$. The batch size was selected to be 24. In Table I we list the corresponding results.

For a thorough evaluation of the optimized model topology, we ran experiments using all considered combinations of loss functions, with a layer count of 4 and a bidirectional LSTM. To also understand the impact of the stationarity of noise on the performance of the system, we randomly added noises from the RSG10 database, grouped into stationary and non-stationary types. For this purpose, we considered the noise files *hfchannel*, *pink*, *volvo*, *white*, *buccaneer2*, *f16*, *leopard* and *m109* as stationary; whereas the files *babble*, *factory1*, *factory2*, *buccaneer1*, *destroyerengine*, *destroyerops* and *machinegun* contain transient noises as well as babble noise and thus are considered non-stationary. We used the first $60\,\%$ of each file for training, the last $20\,\%$ of each file for testing; the remaining $20\,\%$ of the files were used in the development set.

Furthermore, we used four different noise levels, $[0, 5, 10, 15]$ dB ITU-T P.56 speech signal-to-noise-ratio (SNR), which we calculated using the Maracas package [42]. This leads to a test set size of 100 files per group and SNR. The same noisy test set is used for every model and parameter set.

Since neither STOI nor PESQ values of our result files are normally distributed, we used a non-parametric Mann–Whitney–Wilcoxon statistical test [43] to check the significance of the improvement of our method relative to the baselines. The p-value of the hypothesis test is indicated by the number of stars, with the attribution $\star\star\star\star : p \leq 0.0001$, $\star\star\star : 0.0001 < p \leq 0.001$, $\star\star : 0.001 < p \leq 0.01$ and no line indicating not significant.

IV. Results

An overview of the STOI metric can be found in Figure 4, which shows that our model outperformed IMCRA with a high significance regardless of the optimization criterion w.r.t. to the STOI metric. From this plot, we can also conclude that the STOI metric is more suited to test the general functioning of a speech enhancement model than to compare the fine-tuning of models. Due to the upper bound of the STOI metric, the values saturate above a certain model quality, which makes it hard to distinguish between different high quality models.

In the following, we will look at the PESQ values in more detail: The PESQ results of our method, as well as our IMCRA and MSE baseline, are depicted in Figure 5 for non-stationary noise and in Figure 6 for stationary noise,

TABLE I
Preliminary PESQ scores of stationary noise for different model topologies of the soft gating filter (SGF), using the loss combination PSY+SLOW+MSLE. The layer count of the LSTM and its bidirectionality are given in parentheses.

Model topology	Signal-to-noise ratio [dB]			
	0	5	10	15
SGF(l=3, b=0)	1.55	1.83	2.19	2.59
SGF(l=3, b=1)	1.60	1.89	2.20	2.53
SGF(l=4, b=0)	1.55	1.81	2.15	2.50
SGF(l=4, b=1)	**1.63**	**1.94**	**2.28**	**2.68**

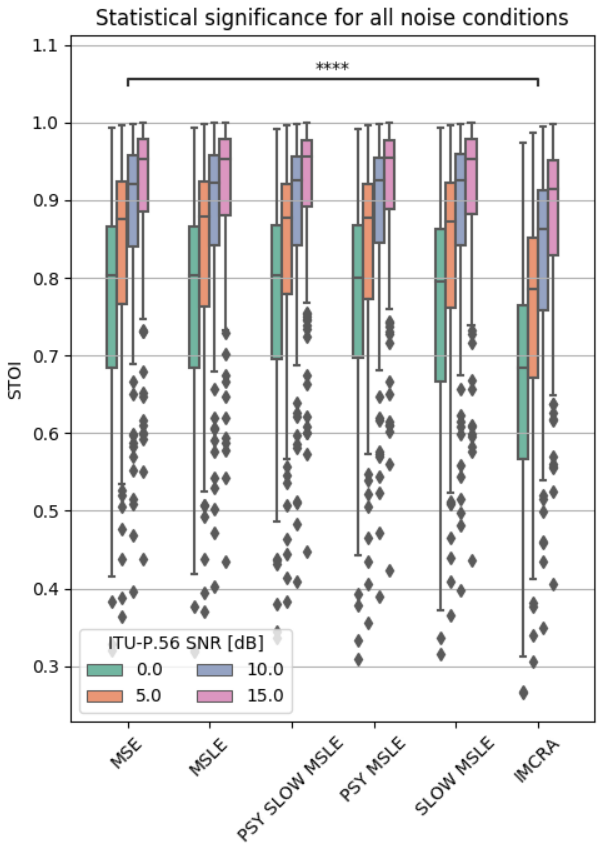

Fig. 4. The boxplot shows the STOI values after speech enhancement for different loss combinations on both stationary and non-stationary noises. Each box is based on 100 files. The p-value of the hypothesis test is indicated by the number of stars, with the attribution $\star\star\star\star : p \leq 0.0001$ and no line indicating not significant.

respectively. From the figures, we can conclude that our MSE baseline outperforms IMCRA regardless of noise type and SNR. Also, there is no relevant difference between MSE and MSLE. Therefore, further comparisons will be made in relation to the MSLE loss.

For the non-stationary noise, we find no significant difference between the MSE, MSLE, and MSLE+SLOW. Nevertheless, adding a slowness loss term to MSLE+PSY does yield small improvements that are significant when compared to the MSLE. Furthermore, we recognize a significant improvement if our psychoacoustic loss PSY is used for the training. Both cases MLSE→MLSE+PSY, as well as MLSE+SLOW→MLSE+SLOW+PSY gained from the additional loss term. Further, the additive combination of all three losses—MSLE+PSY+SLOW—yields the highest mean values in all SNR groups, which is significantly better in comparison to the MSLE and the MSE loss, respectively. The behavior under stationary noise is only slightly different when comparing the efficacy of the suggested optimization criteria to that of the baseline loss functions. The overall PESQ values are higher, hence indicating a somewhat easier problem. Additionally, in the case of stationary noise, we find stronger improvements

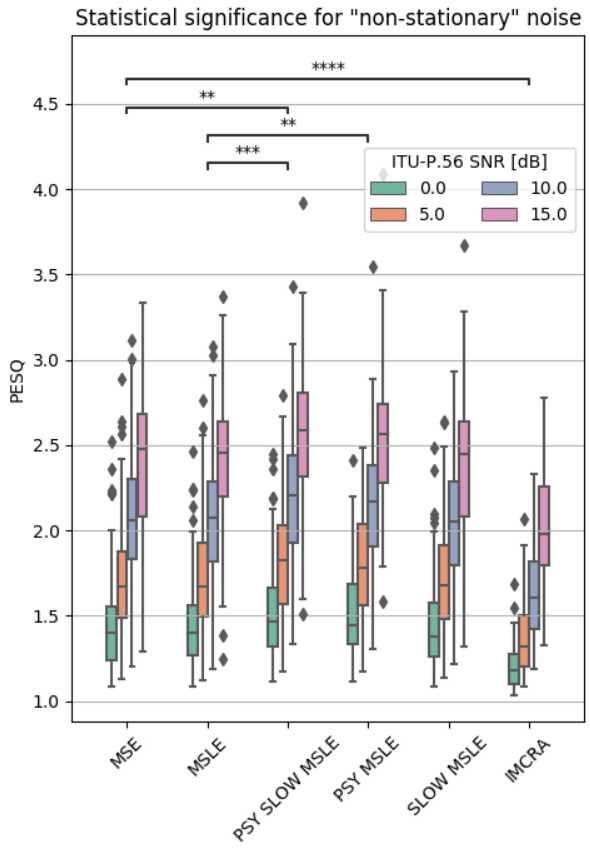

Fig. 5. PESQ values after speech enhancement for different loss combinations on non-stationary noises. Each box is based on 100 files. The p-value of the hypothesis test is indicated by the number of stars above the line, with the attribution $\star\star\star\star : p \leq 0.0001$, $\star\star\star : 0.0001 < p \leq 0.001$, $\star\star : 0.001 < p \leq 0.01$ and no line indicating not significant.

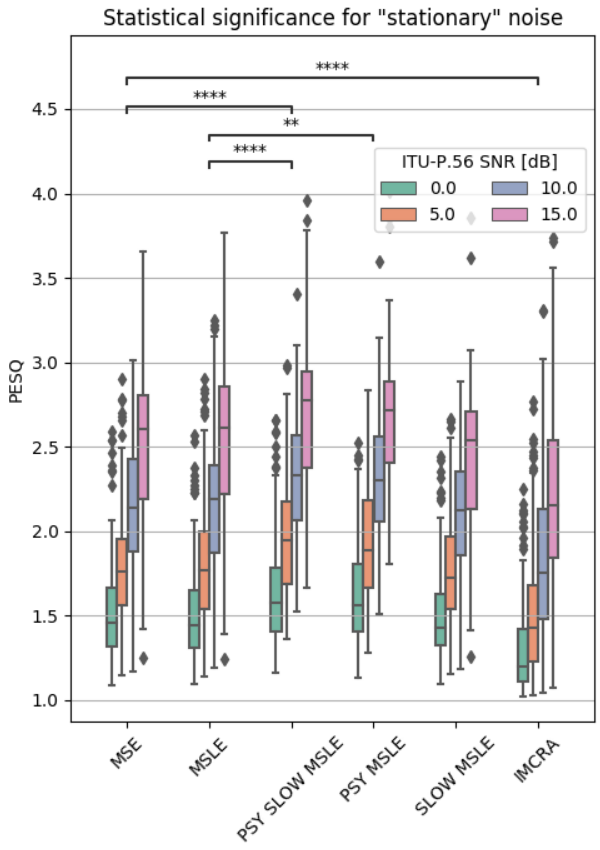

Fig. 6. PESQ values after speech enhancement for all considered loss combinations on stationary noises. Each box is calculated from 100 utterances. The p-value of the hypothesis test is indicated by the number of stars, with the attribution $\star\star\star\star : p \leq 0.0001$, $\star\star : 0.001 < p \leq 0.01$ and no line indicating not significant.

in significance for MLSE+PSY→MSLE+PSY+SLOW, indicating that our slowness loss function has a more significant benefit for this noise type.

From both PESQ evaluations, we conclude that the application of only the slowness loss is not beneficial, but together with the psychoacoustic loss, we see a significant improvement. Both losses consider different aspects of the optimization. The psychoacoustic hearing thresholds emphasize on changes, which are exceeding a certain level. Therefore, this loss can not be applied without MSLE because the optimization would stop too early. Because of its regulatory nature, the same holds true for the slowness loss, and as we see from the results, the slowness term is better suited to filter stationary noises. However, it can improve the PESQ values in both noise conditions, when used together with the psychoacoustic loss.

V. Conclusions

In this work, novel loss functions for speech enhancement based on slow feature analysis and psychoacoustic insights into the human perception of speech have been presented. The proposed slowness-based loss exploits phoneme boundaries during the training process to account for variations in speech production. Additionally, the incorporation of hearing thresholds in the psychoacoustic loss function allows us to model masking effects occurring in human perception of speech to further refine the perceptual quality of enhanced speech.

An evaluation corpus based on the Mozilla Common Voice database was used to evaluate the proposed loss functions. The underlying speech enhancement network utilizes a standard neural time-frequency masking. The evaluation was conducted by comparing STOI and PESQ scores obtained for the proposed psychoacoustic and slowness-based loss functions with those for more classical losses based on the mean squared error, and comparing all of these learning-based approaches to the well-known IMCRA speech enhancement framework. The results indicate a small but consistent and statistically significant improvement in the achieved PESQ scores of the combined psychoacoustic and slowness-based loss function over the classical MSE and MSLE losses. Additionally, all DNN-based models significantly outperformed the IMCRA baseline under all noise conditions.

Future investigations should focus on extending the proposed framework to include an estimated signal phase into the

enhancement process. Additionally, a complex-valued variant of the system that directly utilizes the complex spectrum provides a promising direction for further research.

REFERENCES

[1] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, p. 113–120, April 1979.

[2] Mozilla, "Common Voice," https://voice.mozilla.org/en, Sep. 2019.

[3] S. Pascual, A. B., and J. Serrà, "Segan: Speech enhancement generative adversarial network," *arXiv preprint arXiv:1703.09452*, 2017.

[4] Q. Wang, J. Du, L. Dai, and C. Lee, "Joint noise and mask aware training for dnn-based speech enhancement with sub-band features," in *2017 Hands-free Speech Communications and Microphone Arrays (HSCMA)*, March 2017, pp. 101–105.

[5] Y. Xu, J. Du, L. Dai, and C. Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *INTERSPEECH*, 2014.

[6] C. Févotte, E. Vincent, and A. Ozerov, *Single-Channel Audio Source Separation with NMF: Divergences, Constraints and Algorithms*. Cham: Springer International Publishing, 2018, pp. 1–24. [Online]. Available: https://doi.org/10.1007/978-3-319-73031-8_1

[7] E. Vincent, "Improved perceptual metrics for the evaluation of audio source separation," in *10th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA)*, Tel Aviv, Israel, Mar. 2012, pp. 430–437. [Online]. Available: https://hal.inria.fr/hal-00653196

[8] M. Kolbæk, Z. Tan, and J. Jensen, "Monaural Speech Enhancement using Deep Neural Networks by Maximizing a Short-Time Objective Intelligibility Measure," Feb. 2018. [Online]. Available: http://arxiv.org/abs/1802.00604v1;http://arxiv.org/pdf/1802.00604v1

[9] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4214–4217.

[10] Y. Zhao, B. Xu, R. Giri, and T. Zhang, "Perceptually guided speech enhancement using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 5074–5078.

[11] ITU-T, "Perceptual evaluation of speech quality (PESQ) version 2.0," *Recommendations P.862, P.862.1, P.862.2*, Oct. 2005.

[12] J. Kim, M. El-Khamy, and J. Lee, "End-to-End Multi-Task Denoising for joint SDR and PESQ Optimization," *CoRR*, vol. abs/1901.09146, 2019. [Online]. Available: http://arxiv.org/abs/1901.09146

[13] L. Wiskott, "Learning invariance manifolds," in *Proc. 8th Intl. Conf. on Artificial Neural Networks (ICANN'98), Skövde, Sweden*, ser. Perspectives in Neural Computing, L. Niklasson, M. Bodén, and T. Ziemke, Eds. London: Springer, Sep. 1998, p. 555–560. [Online]. Available: https://link.springer.com/chapter/10.1007/978-1-4471-1599-1_83

[14] M. Schüler, H. H. Davíð, and L. Wiskott, "Gradient-based training of slow feature analysis by differentiable approximate whitening," *CoRR*, vol. abs/1808.08833, 2018. [Online]. Available: http://arxiv.org/abs/1808.08833

[15] J. Freiwald, M. Karbasi, S. Zeiler, J. Melchior, V. Kompella, L. Wiskott, and D. Kolossa, "Utilizing Slow Feature Analysis for Lipreading," in *Speech Communication; 13th ITG-Symposium*. VDE Verlag GmbH, 10 2018, p. 191–195. [Online]. Available: https://ieeexplore.ieee.org/document/8578021

[16] T. Blaschke and L. Wiskott, "Independent slow feature analysis and nonlinear blind source separation," in *Independent Component Analysis and Blind Signal Separation*, C. G. Puntonet and A. Prieto, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 742–749.

[17] J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord, "Unsupervised speech representation learning using WaveNet autoencoders," Sep. 2019. [Online]. Available: http://arxiv.org/abs/1901.08810v2;http://arxiv.org/pdf/1901.08810v2

[18] O. Yilmaz and S. Rickard, "Blind separation of speech mixtures via time-frequency masking," *IEEE Transactions on Signal Processing*, vol. 52, no. 7, pp. 1830–1847, July 2004.

[19] N. Roman and D. Wang, "Binaural sound segregation for multisource reverberant environments," in *Proceedings ICASSP 2004, Volume 2*, 2004, pp. 373 – 376.

[20] D. Kolossa and R. Orglmeister, "Nonlinear postprocessing for blind speech separation." in *ICA*, 2004, pp. 832–839.

[21] M. Seltzer, B. Raj, and R. Stern, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.

[22] S. Roweis, "Factorial models and refiltering for speech separation and denoising," in *Proc. Eurospeech*, 2003. [Online]. Available: citeseer.ist.psu.edu/roweis03factorial.html

[23] H. Yang, S. Choe, K. Kim, and H. Kang, "Deep learning-based speech presence probability estimation for noise PSD estimation in single-channel speech enhancement," *2018 International Conference on Signals and Systems (ICSigSys)*, pp. 267–270, 2018.

[24] Y. Zhao, Z. Wang, and D. Wang, "A two-stage algorithm for noisy and reverberant speech enhancement," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 5580–5584.

[25] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, p. 483–492, 2015.

[26] N. Zheng and X. Zhang, "Phase-aware speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 1, pp. 63–76, Jan 2019.

[27] T. Gerkmann, M. Krawczyk-Becker, and J. Le Roux, "Phase processing for single-channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, no. 2, pp. 55–66, March 2015.

[28] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, p. 1735–80, 12 1997.

[29] R. Martin, "Spectral subtraction based on minimum statistics," *Proc. Eur. Signal Processing Conf.*, vol. 6, 01 1994.

[30] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[31] P. Virtanen and SciPy 1.0 Contributors, "SciPy 1.0–Fundamental Algorithms for Scientific Computing in Python," *arXiv e-prints*, p. arXiv:1907.10121, Jul 2019.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017.

[33] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, 3rd ed. Springer, 2007.

[34] L. Schönherr, K. Kohls, S. Zeiler, T. Holz, and D. Kolossa, "Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding," in *Network and Distributed System Security Symposium (NDSS)*, 2019.

[35] L. Schönherr, T. Eisenhofer, S. Zeiler, T. Holz, and D. Kolossa, "Imperio: Robust over-the-air adversarial examples against automatic speech recognition systems," *arXiv preprint arXiv:1908.01551*, 2019.

[36] ISO, "Information Technology – Coding of moving pictures and associated audio for digital storage media at up to 1.5 Mbits/s – Part3: Audio," International Organization for Standardization, ISO 11172-3, 1993.

[37] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," 08 2017, pp. 498–502.

[38] L. Wiskott and T. J. Sejnowski, "Slow Feature Analysis: Unsupervised Learning of Invariances," *Neural Computation*, vol. 14, no. 4, p. 715–770, 2002.

[39] I. Cohen, "Noise spectrum estimation in adverse environments: Improved minima controlled recursive averaging," *Speech and Audio Processing, IEEE Transactions on*, vol. 11, p. 466–475, 10 2003.

[40] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (DEMAND): A database of multichannel environmental noise recordings," *The Journal of the Acoustical Society of America*, vol. 133, p. 3591, 05 2013.

[41] H. J. M. Steeneken and F. W. M. Geurtsen, "Description of the RSG. 10 noise database," 1988.

[42] J. F. Santos, "Maracas," 2016. [Online]. Available: https://github.com/jfsantos/maracas

[43] H. B. Mann and D. R. Whitney, "On a test of whether one of two random variables is stochastically larger than the other," *Ann. Math. Statist.*, vol. 18, no. 1, pp. 50–60, 03 1947. [Online]. Available: https://doi.org/10.1214/aoms/1177730491