ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-7605-5/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICASSP39728.2021.9413555

PERCEPTUAL LOSS BASED SPEECH DENOISING WITH AN ENSEMBLE OF AUDIO PATTERN RECOGNITION AND SELF-SUPERVISED MODELS

Saurabh Kataria, Jesús Villalba, Najim Dehak

Center for Language and Speech Processing, Human Language Technology Center of Excellence,
Johns Hopkins University, Baltimore, MD
{skatari1,jvillal7,ndehak3}@jhu.edu

ABSTRACT

Deep learning based speech denoising still suffers from the challenge of improving perceptual quality of enhanced signals. We introduce a generalized framework called Perceptual Ensemble Regularization Loss (PERL) built on the idea of *perceptual losses*. Perceptual loss discourages distortion to certain speech properties and we analyze it using six large-scale pre-trained models: speaker classification, acoustic model, speaker embedding, emotion classification, and two self-supervised speech encoders (PASE+, wav2vec 2.0). We first build a strong baseline (w/o PERL) using Conformer Transformer Networks on the popular enhancement benchmark called VCTK-DEMAND. Using auxiliary models one at a time, we find acoustic event and self-supervised model PASE+ to be most effective. Our best model (PERL-AE) only uses acoustic event model (utilizing AudioSet) to outperform state-of-the-art methods on major perceptual metrics. To explore if denoising can leverage full framework, we use all networks but find that our seven-loss formulation suffers from the challenges of Multi-Task Learning. Finally, we report a critical observation that state-of-the-art Multi-Task weight learning methods cannot outperform hand tuning, perhaps due to challenges of domain mismatch and *weak complementarity* of losses.

Index Terms— Speech Denoising, Perceptual Loss, Pre-trained Networks, Multi-Task Learning, Self-Supervised Features

1. INTRODUCTION

There is a growing focus on the perceptual and intelligibility quality of enhanced signals obtained from Deep Neural Network (DNN) based speech enhancement systems. Quantifying them through human Mean Opinion Score (MOS) is highly expensive and prone to error. Proxy objective metrics (reference/non-reference based) are used in enhancement like Perceptual Evaluation of Speech Quality (PESQ) but they are hard to improve upon. In [1], authors introduced *perceptual loss* (or Deep Feature Loss (DFL)) to speech denoising problem. This improves perceptual quality of audio and is successfully applied for Generative Adversarial Network (GAN) [2] and *task-specific* enhancement [3].

Alternatively, various differentiable metrics are proposed like Perceptual Metric for Speech Quality Evaluation (PMSQE) [4], Semi-supervised Speech Quality Assessment (SESQA) [5], and a metric based on Just Noticeable Differences (JNDs) [6]. In [5], authors proposed SESQA which is an eight-loss Multi-Task Learning (MTL) based semi-supervised framework for automated speech quality assessment. In addition to predicting MOS, they define various auxiliary tasks including predicting JND, pairwise comparison, degradation strength, etc. [6] took a different approach by constructing a large corpus based on (binary) human preference in audio pairs. By capturing JND, they are able to train a differentiable metric well correlated with human MOS rating.

Our aim is to improve perceptual and intelligibility metrics using *perceptual loss* on small-scale supervised speech denoising. To compensate for low resource training data, we investigate if large-scale pre-trained speech models can be utilized. Such models, when used frozen and in ensemble, can regularize enhancement training to minimize distortions to various speech properties. Furthermore, we explore if self-supervised models can help preserve speech representations since they are trained on generic *pretext tasks*. Our idea faces challenges of domain mismatch and weak complementarity of losses, which we explore in-depth with various MTL methods.

Our contributions are as follows. First, we establish a strong baseline based on Conformer Transformer [7] with a simple l_1 objective. Second, we propose a rich framework called Perceptual Ensemble Regularization Loss (PERL) which leverages six open-source large-scale pre-trained networks to analyze speech denoising with perceptual losses. Third, we rank various speech tasks in term of their effectiveness in PERL and, for the first time, we demonstrate the utility of self-supervised representations for perceptual loss training. Fourth, we show the importance of l_1 enhancement loss and all Conformer components to achieve state-of-the-art (SOTA) performance. Fifth, using all seven losses of PERL, we make an important finding that SOTA multi-task weight learning methods cannot outperform hand-tuned weights.

2. SPEECH DENOISING WITH PERCEPTUAL ENSEMBLE REGULARIZATION LOSS (PERL)

We consider a simple signal model or Noise Corruption Process (NCP): $y(t) = x(t) + n(t)$, where $x(t)$ is clean (time-domain) signal, $n(t)$ is noise signal, and $y(t)$ is the resultant noisy signal. In general, $x(t)$ suffers distortions to its acoustic event, speaker characteristics, and acoustic content. In DNN based speech enhancement, a simple l_1 loss is not able to restore such distortions [3]. Moreover, enhancement may introduce such distortions [3]. To counter this, [1, 3] used *perceptual loss* or *deep feature loss*. Such loss compares enhanced signals with reference clean signals in the activation space of a pre-trained auxiliary network (trained for task $\mathcal{T}$). Depending on the choice of the auxiliary network, we accomplish speech enhancement with minimal distortions to representations required to solve task $\mathcal{T}$. For e.g., [3] did DFL based speech enhancement while minimizing distortions to speaker identities.

To avoid distortions to various speech characteristics during enhancement and/or NCP, we propose PERL: an ensemble of varied types of pre-trained (and frozen) speech models for perceptual

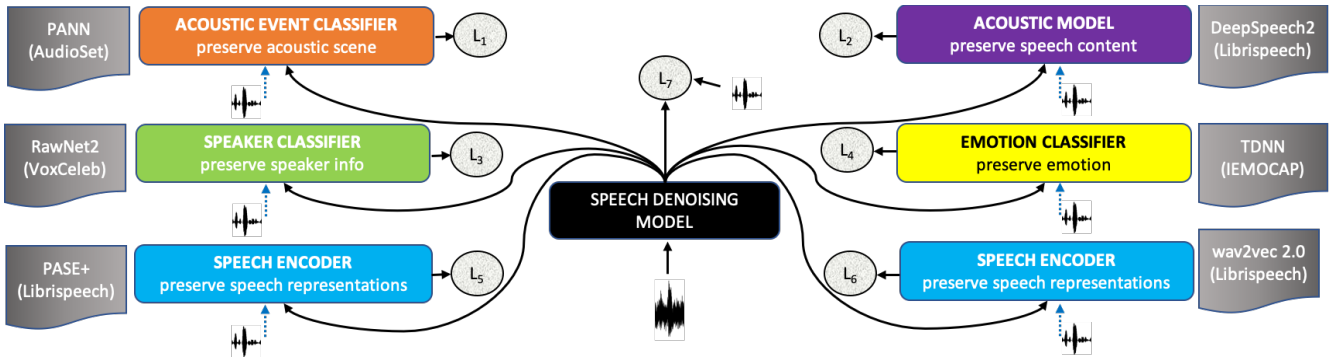

Fig. 1: Illustration of supervised framework called Perceptual Ensemble Regularization Loss (PERL) using seven losses: $L_1, \ldots, L_7$. To extract perceptual losses, partial forward pass of enhanced and reference clean signals is done as illustrated by using temporal signals.

loss training. Specifically, we use acoustic event classification, speaker embedding, acoustic model, and Speech Emotion Recognition (SER). We also incorporate self-supervised speech encoders since they model general speech representations [8]. Our scheme is illustrated in Fig. 1. When using all components, loss function is

$$\begin{aligned}\mathcal{L} = \lambda_{\text{event}}\mathcal{L}_{\text{event},n_1} + \lambda_{\text{acoustic}}\mathcal{L}_{\text{acoustic},n_2} + \lambda_{\text{speaker}}\mathcal{L}_{\text{speaker},n_3} \\ +\lambda_{\text{emotion}}\mathcal{L}_{\text{emotion},n_4} + \lambda_{\text{pase}}\mathcal{L}_{\text{pase},n_5} + \lambda_{\text{wav2vec}}\mathcal{L}_{\text{wav2vec},n_6} \\ +\lambda_{l_1}\mathcal{L}_{l_1},\end{aligned} \quad (1)$$

where the last term refers to the simple Short-Time Fourier Transform (STFT) feature-domain l_1 enhancement loss and n_i refers to the number of internal layers of the corresponding auxiliary network used for deriving perceptual loss. The perceptual losses derived from i-th network are summed and divided by n_i. We term PERL as a *regularization loss* since the auxiliary networks are frozen and only help constrain the output space. Note that this loss is different from usual formulations of transfer learning, knowledge distillation, and even multi-task learning. [9] appropriately terms such setting as *single-task multi-loss*. Our framework is flexible and, instead of speech denoising, it can be used for voice conversion and domain adaptation as well. In terms of framework richness, our work is close to SESQA [5] and PASE+ [10].

There are some inherent challenges with PERL. One, since the auxiliary networks are trained on different features and domain, there can be significant domain mismatch which hinder training. To avoid this, we prefer to use initial layers since they tend to be more generic. Two, our test set is not equipped with labels for all auxiliary tasks and hence evaluation is restricted to enhancement metrics only.

3. MODEL DESCRIPTIONS

3.1. Baseline Conformer Transformer Enhancement Network

Following the success of Transformer-based modeling of speech features [7, 11], we choose *convolution-augmented Transformer* or Conformer [7] for the denoising network. For modeling long-term and short-term patterns, it relies on self-attention mechanism and specially designed *convolution modules* respectively. Moreover, it combines the power of relative Position Encoding (PE) scheme and Macaron-style half-step Feed-Forward Networks (FFNs) [7]. We additionally include Squeeze-and-Excitation [12] module (squeeze factor of 8) after the *1D Depthwise Convolution* inside the *convolution module* of Conformer. To avoid down-sampling in time, for the first layer of our Conformer, we use a Batch Normalization (BN) layer followed by a linear transformation instead of the usual convolutional sub-sampling layer. The attention dimension of the network is 240, the number of conformer blocks are 4, and total parameters are 10M. The denoising network predicts a mask which is then multiplied with the noisy spectra to predict the clean spectra.

3.2. Brief Overview of Classification Models

Acoustic Event Classification: In [13], authors train a large-scale audio event classification network called Pre-trained Audio Neural Network (PANN) and prove its transferability to six audio pattern recognition tasks. It is trained with 1.9M audio clips from AudioSet [14] (5000hrs, 527 sound classes). We choose the wide-band 14-layer version which has 81M parameters. They also address data imbalance problem of AudioSet and do online data augmentation with MixUp and SpecAugment [15]. We choose $n_{\text{event}} = 4$ by using the first four convolution blocks (or first eight convolution layers).

Acoustic Model: Deep Speech 2 (DS2) is a multi-lingual end-to-end Automatic Speech Recognition (ASR) model proposed in [16]. We choose the English model trained on LibriSpeech [17] consisting of 13M parameters. The architecture follows the Recurrent Neural Network (RNN)-Connectionist Temporal Classification (CTC) style. It has a convolutional front-end which is followed by five Bi-directional Long Short Term Memory (BLSTM) layers and final classification/CTC layer for English graphemes. We choose $n_{\text{acoustic}} = 3$ by using the output of the first two convolutional layers and first RNN layer for DFL.

Speaker Embedding: In [18], authors propose RawNet2. It is a time-domain speaker embedding network trained with Voxceleb [18], which contains over 1M utterances from 6112 speakers. The architecture followed is a Convolutional Neural Network (CNN) made of six residual blocks followed by Gated Recurrent Units (GRU). The first layer is a sinc-convolution layer of SincNet [19]. Total number of parameters are 87M. We choose $n_{\text{speaker}} = 3$ by using the output of the first sinc layer and the next two convolutional blocks.

Speech Emotion Classification: [20] trains a time-domain emotion identification network on a 12hr emotion corpus called IEMOCAP. The architecture follows a Time-Delay Neural Network (TDNN)-LSTM style and has 1M parameters. The first layer is a 1-D convolution based data dependent layer. It also consists of five LSTM layers followed by four TDNN layers and a time-restricted self-attention based pooling of embeddings. Each frame of speech is classified in four emotions: happy, sad, neutral, and angry. We train this network using the same training scheme used for Conformer as described in Sec. 4. We choose $n_{\text{emotion}} = 3$ by using the output of the first three layers.

3.3. Brief Overview of Self-Supervised Speech Encoder Models

PASE+: PASE+ [10] is a self-supervised speech encoder trained with 12 self-supervised tasks. For *regression tasks*, the model learns to predict various known speech transformations. For *binary tasks*, a contrastive loss tries to bring close representations of speech chunks belonging to same utterance while pushing apart representations of chunks belonging to different utterances. PASE+ can model speech content and speaker properties, which are robust to noise perturbations and have good transferability. The architecture has a convolutional front-end with a Quasi-Recurrent Neural Network (QRNN) backbone. It is trained with 50hrs of Librispeech [17] and number of trainable parameters are 8M. We interchangeably refer to this model as PASE. We use $n_{\text{pase}} = 6$ by using the output of the first six convolutional blocks.

wav2vec 2.0 In [8], authors propose wav2vec 2.0, a self-supervised model which learns generic speech representations by solving a contrastive loss in the (quantized) latent space using a masked Transformer. It is trained on 960hrs of LibriSpeech without transcriptions. The architecture consists of 12 transformer blocks, model dimension of 768, inner dimension (Feed-Forward Network (FFN)) of 3,072, eight attention heads, and 96M parameters. We interchangeably refer to this model as wav2vec. We choose $n_{\text{wav2vec}} = 5$ by using the output of the convolutional front-end and the next four Transformer blocks.

3.4. Brief Overview of Multi-Task Learning Methods

We use four MTL methods off-the-shelf whose some aspects are as follows. *Uncertainty Weighting* [21] uses the notion of an inherent *task-dependent uncertainty* or *homoscedastic uncertainty*. Using a Gaussian likelihood formulation, it defines a loss function consisting of learnable variance parameters for each loss. Hence, it accounts for the dynamic range of loss terms in a principled way. *Coefficient of Variation* [9] states that a loss term is satisfied when its variance has decreased to zero. Hence, it assigns weight for a loss term as its coefficient of variation, i.e., standard deviation divided by mean. This quantity can compensate for the different dynamic range of loss terms. For mean estimate, a robust formulation of the current loss is used. For variance estimate, an online statistics tracking algorithm is used. Finally, all weights are normalized to make sum equals to one. *Dynamic Weight Averaging* [22] keeps track of instantaneous ratio of change of loss value. Using that and a temperature parameter T (fixed to 2), it assigns weight to each task via a softmax classification. *GradCosine* [23] treats main and auxiliary loss separately. To update the shared parameters (in our case, full denoising network), cosine similarity of gradient w.r.t. auxiliary loss is checked against the gradient w.r.t. main loss. When auxiliary gradients align, they are used to update shared parameters, otherwise they are rejected.

4. EXPERIMENTAL SETUP

We evaluate speech denoising on VCTK-DEMAND [24]. It is a 16KHz small corpus with fixed training (10hr, 30 speakers) and validation data (30m, 2 speakers). We use the predicted clean spectra to compute Signal Approximation (SA) loss or, simply, the l_1 loss (in STFT domain). We experimented with several popular loss functions but found l_1 loss to be the best and the most resilient to change in experimental setup. For compatibility with auxiliary networks, we do appropriate on-the-fly transformations. To convert to time-domain, we re-use the phase of the noisy signal. Due to the large number of experiments, instead of subjective evaluation, we aim to improve upon standardized perceptual and intelligibility metrics like PESQ, CSIG, CBAK, COVL, and STOI. PESQ $\in [-0.5, 4.5]$, STOI (in %) $\in [0, 100]$, while other metrics $\in [0, 5]$. Our seven-network based framework is memory intensive, and hence, we do gradient accumulation. We train for 10 epochs using Adam optimizer, batch size of 32, a simple learning rate scheduler, and an initial learning rate of 0.00075. Best scores on validation data are reported.

Table 1: Comparison of perceptual losses of various speech models with simple l_1 loss. n_p refers to #parameters of auxiliary network.

	n_p	PESQ	CSIG	CBAK	COVL	STOI
l_1	-	3.01	4.27	3.48	3.65	**94.9**
AcousticEvent	81	**3.09**	**4.38**	**3.43**	**3.75**	94.8
AcousticModel	87	2.97	4.15	3.43	3.55	94.6
SpeakerEmbedding	13	2.86	3.63	3.32	3.23	94.5
SpeechEmotion	1	2.41	3.49	3.07	2.93	92.8
PASE	8	3.04	4.23	3.42	3.63	94.7
wav2vec	95	2.93	3.67	3.33	3.29	93.9

5. RESULTS

5.1. Comparison of perceptual losses of various speech models

In Table 1, we first note that using a simple l_1 loss in STFT domain with Conformers, we get close to SOTA performance (Table 3). In [1], authors used acoustic event classifier to achieve PESQ score of 2.57. Using Conformers and AudioSet based auxiliary network, we drastically improve that score (3.09). We find other speech models also give competitive performance except for the emotion model, perhaps due to its low resource training data. Here, n_p refers to the number of parameters in auxiliary network (in M). PASE+ is the best model considering performance as well as parameter efficiency. We experimented with various popular loss functions based on the ideas of multi-resolution spectra, Time-Domain Reconstruction (TDR), Signal-to-Distortion Ratio (SDR), etc. but found them much inferior to the l_1 loss. It is important to state that we expect acoustic event loss to be the best since during *noise corruption process* (and consequently during enhancement), for a simple test set like ours, *acoustic event information of audio is adversely affected the most*. However, its comparison with other speech models is novel.

5.2. Combining l_1 enhancement loss with perceptual losses

Previous work [1] did not investigate the benefit of combining feature-domain enhancement loss with perceptual losses. In Table 2, we combine l_1 loss with six auxiliary losses individually as well in certain combinations. Hand-tuning of loss terms gave $(\lambda_{\text{event}}, \lambda_{\text{acoustic}}, \lambda_{\text{speaker}}, \lambda_{\text{emotion}}, \lambda_{\text{pase}}, \lambda_{\text{wav2vec}}, \lambda_{l_1})$ = (5e-03, 1e-04, 1.25e-04, 4e-05, 1.7e-04, 3.5e-05, 1.1e-01). Combining l_1 with acoustic event (PERL-AE) yields better results than baseline and outperforms SOTA models (Table 3). Some enhanced samples are available online[1]. Note that we can further improve by predicting phase too by using a time-domain model like DEMUCS [25]. For all cases, combining l_1 loss results in better performance than without it (Refer Table 1). Second last row in Table 2 shows good performance by combining best three models: PASE, acoustic event, and speaker embedding. This suggests that if we use such networks with more data and less domain mismatch w.r.t. enhancement training data, further improvements can be observed. The final row uses all losses

[1] https://github.com/saurabh-kataria/PERL-samples

Table 2: Effect of adding *perceptual losses* (individually and in certain combinations) to l_1 loss. Last row uses all seven loss terms.

	PESQ	CSIG	CBAK	COVL	STOI
l_1	3.01	4.27	3.48	3.65	94.9
l_1+AcousticEvent	**3.17**	**4.43**	**3.53**	**3.83**	**95.0**
l_1+AcousticModel	2.97	4.15	3.43	3.55	94.6
l_1+SpkEmbedding	2.91	4.18	3.39	3.55	94.4
l_1+SpkEmotion	2.86	4.12	3.38	3.5	94.2
l_1+PASE	3.03	4.24	3.43	3.63	94.8
l_1+wav2vec	2.92	4.16	3.37	3.54	94.5
l_1+AcousticEvent +AcousticModel +SpkEmbedding +SpkEmotion	2.88	4.15	3.40	3.52	94.5
l_1+PASE+wav2vec	2.95	4.2	3.4	3.58	94.6
l_1+PASE +AcousticEvent +SpkEmbedding	3.05	4.31	3.5	3.69	94.8
l_1+PASE +AcousticEvent	3.04	4.37	3.47	3.72	94.9
l_1+ALL (**PERL**)	2.89	4.11	3.41	3.5	94.6

Table 3: Comparison of the best denoising system (PERL-AE or l_1+AcousticEvent) with the state-of-the-art methods

	PESQ	CSIG	CBAK	COVL	STOI
Noisy	1.97	3.35	2.44	2.63	-
Weiner filtering	2.22	3.23	2.68	2.67	-
Deep Feature Loss [1]	2.57	-	-	-	-
Hi-Fi GAN [2]	2.94	4.07	3.07	3.49	-
Self-Adapt MHSA [26]	2.99	4.15	3.46	3.51	-
T-GSA [11]	3.06	4.18	3.59	3.62	-
DEMUCS [25]	3.07	4.31	3.4	3.63	**95**
PERL-AE (ours) (l_1+AcousticEvent)	**3.17**	**4.43**	**3.53**	**3.83**	**95**

and, hence, is termed PERL. Using full framework, we get inferior results w.r.t. baseline. This is perhaps due to sub-optimal choice of loss weights (selected via greedy hand-tuning for best performance per task) or simply due to detrimental nature of some loss terms (investigated in Sec. 5.3).

To emphasize the role of Conformer architecture, we do an ablation study similar to [7]. In Table 4, we progressively (1) replace SWISH activation by ReLU; (2) remove *convolution module* inside Conformer blocks; (3) replace Macaron-style FFN pairs with a single FFN layer; (4) replace relative position embedding with absolute. The trend of performance degradation shows that all components of Conformer are important, especially the *convolution module*.

5.3. Automatic weight learning versus hand-tuning of weights

PERL faces challenges of MTL. For e.g., loss terms can have different dynamic range, speed of convergence (for fixed learning rate), and complementarity with main loss. In this section, we investigate if SOTA automatic weight learning methods can outperform hand-tuned PERL system i.e. the system trained with all seven losses. We also want to investigate if detrimental losses can be automatically rejected, especially by GradCosine [23]. In Table 5, second row refers to PERL system fine-tuned with the best loss discovered in Table 2 i.e. "l_1+AcousticEvent". Note that it is able to (almost) recover fully. Results with equal weighting shows that the choice of weights is paramount. We find GradCosine to be the worst and Uncertainty

Table 4: Reproducing ablation study of [7] by removing features from Conformer network to converge to vanilla Transformer.

	PESQ	CSIG	CBAK	COVL	STOI
Full Conformer	**3.17**	**4.43**	**3.53**	**3.83**	**95**
- SWISH + ReLU	3.15	4.38	3.52	3.78	94.9
- Convolution Block	2.81	4.21	3.36	3.51	94.7
- Macaron FFN	2.74	4.14	3.31	3.44	94.8
- Relative Pos. Emb.	2.65	4.01	3.30	3.36	94.4

Table 5: Comparison of fixed hand-tuned weights and multi-task weights learning methods for training PERL model (seven losses)

	PESQ	CSIG	CBAK	COVL	STOI
Hand Tuning (**PERL**)	2.89	4.11	3.41	3.5	94.6
Fine-Tuning on l_1+AcousticEvent	3.12	4.4	3.51	3.78	95
Equal weights	2.74	4.02	3.31	3.37	94.3
Coefficient of Variation [9]	2.74	4.08	3.31	3.31	94.3
Uncertainty Weighting [21]	**2.85**	4.12	**3.35**	**3.48**	**94.5**
Dynamic Weight Average [22]	2.8	**4.14**	3.34	3.47	94.3
GradCosine [23]	2.54	3.92	3.15	3.22	93.1

Weighting to be the best among four MTL methods. Our important finding is that learning weights or even dynamically adjusting (non-learnable) weights cannot outperform well-tuned (by hand) results. We suspect this because of two reasons. One, the domain mismatch of enhancement training data with auxiliary network data is hard to overcome. By leveraging more training data and learning a transformation layer between auxiliary networks and the main (denoising) network, this problem can perhaps be quelled. Two, the fundamental assumption of multi-task learning of *complementarity of losses* does not hold and even SOTA MTL methods cannot overcome it by, for example, (soft/hard) rejection of terms unaligned with *main loss*.

6. CONCLUSION

To improve perceptual metrics for speech denoising, we propose Perceptual Ensemble Regularization Loss (PERL). PERL leverages an ensemble of variety of large-scale pre-trained speech models for deriving perceptual loss. We show the efficacy of all models individually as well as in certain combinations. Acoustic event and self-supervised PASE+ models are found to be most effective. We also find combining regular feature-domain enhancement loss to be complementary. With an ablation study, we highlight the critical role of components of Conformer Transformer. When using all components of PERL, we find that our seven-loss framework suffers from the challenges of Multi-Task Learning (MTL). To recover, we experiment with various MTL methods to conclude that state-of-the-art MTL methods cannot outperform greedy hand-tuning based weight selection but, interestingly, this system can be fine-tuned on the best loss (l_1+AcousticEvent) to restore lost performance. In future, we can (1) build towards universal enhancement by analyzing PERL on a richer test setup, (2) incorporate automatic speech quality quantification networks like SESQA [5], and (3) learn bridges/adaptors between main and auxiliary networks to alleviate domain mismatch.

7. REFERENCES

[1] Francois G Germain, Qifeng Chen, and Vladlen Koltun, "Speech denoising with deep feature losses," *arXiv preprint arXiv:1806.10522*, 2018.

[2] Jiaqi Su, Zeyu Jin, and Adam Finkelstein, "Hifi-gan: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *arXiv preprint arXiv:2006.05694*, 2020.

[3] Saurabh Kataria, Phani Sankar Nidadavolu, Jesús Villalba, and Najim Dehak, "Analysis of deep feature loss based enhancement for speaker verification," *arXiv preprint arXiv:2002.00139*, 2020.

[4] Juan Manuel Martín-Doñas, Angel Manuel Gomez, Jose A Gonzalez, and Antonio M Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal processing letters*, vol. 25, no. 11, pp. 1680–1684, 2018.

[5] Joan Serrà, Jordi Pons, and Santiago Pascual, "Sesqa: semi-supervised learning for speech quality assessment," *arXiv preprint arXiv:2010.00368*, 2020.

[6] Pranay Manocha, Adam Finkelstein, Zeyu Jin, Nicholas J Bryan, Richard Zhang, and Gautham J Mysore, "A differentiable perceptual audio metric learned from just noticeable differences," *arXiv preprint arXiv:2001.04460*, 2020.

[7] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

[8] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[9] Rick Groenendijk, Sezer Karaoglu, Theo Gevers, and Thomas Mensink, "Multi-loss weighting with coefficient of variations," *arXiv preprint arXiv:2009.01717*, 2020.

[10] Mirco Ravanelli, Jianyuan Zhong, Santiago Pascual, Pawel Swietojanski, Joao Monteiro, Jan Trmal, and Yoshua Bengio, "Multi-task self-supervised learning for robust speech recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6989–6993.

[11] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6649–6653.

[12] Jie Hu, Li Shen, and Gang Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[13] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *arXiv preprint arXiv:1912.10211*, 2019.

[14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[15] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[16] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al., "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*, 2016, pp. 173–182.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] Jee-weon Jung, Seung-bin Kim, Hye-jin Shim, Ju-ho Kim, and Ha-Jin Yu, "Improved rawnet with filter-wise rescaling for text-independent speaker verification using raw waveforms," *arXiv preprint arXiv:2004.00526*, 2020.

[19] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[20] Mousmita Sarma, Pegah Ghahremani, Daniel Povey, Nagendra Kumar Goel, Kandarpa Kumar Sarma, and Najim Dehak, "Emotion identification from raw speech signals using dnns.," in *Interspeech*, 2018, pp. 3097–3101.

[21] Alex Kendall, Yarin Gal, and Roberto Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7482–7491.

[22] Shikun Liu, Edward Johns, and Andrew J Davison, "End-to-end multi-task learning with attention," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1871–1880.

[23] Yunshu Du, Wojciech M Czarnecki, Siddhant M Jayakumar, Razvan Pascanu, and Balaji Lakshminarayanan, "Adapting auxiliary losses using gradient similarity," *arXiv preprint arXiv:1812.02224*, 2018.

[24] Cassia Valentini-Botinhao, Xin Wang, Shinji Takaki, and Junichi Yamagishi, "Investigating rnn-based speech enhancement methods for noise-robust text-to-speech.," in *SSW*, 2016, pp. 146–152.

[25] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi, "Real time speech enhancement in the waveform domain," *arXiv preprint arXiv:2006.12847*, 2020.

[26] Yuma Koizumi, Kohei Yaiabe, Marc Delcroix, Yoshiki Maxuxama, and Daiki Takeuchi, "Speech enhancement using self-adaptation and multi-head self-attention," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 181–185.