Data Augmentation and Loss Normalization for Deep Noise Suppression

Sebastian Braun(✉) and Ivan Tashev

Microsoft Research, Redmond, WA, USA
{sebastian.braun,ivantash}@microsoft.com
https://www.microsoft.com/en-us/research/group/audio-and-acoustics-research-group

Abstract. Speech enhancement using neural networks is recently receiving large attention in research and being integrated in commercial devices and applications. In this work, we investigate data augmentation techniques for supervised deep learning-based speech enhancement. We show that not only augmenting SNR values to a broader range and a continuous distribution helps to regularize training, but also augmenting the spectral and dynamic level diversity. However, to not degrade training by level augmentation, we propose a modification to signal-based loss functions by applying sequence level normalization. We show in experiments that this normalization overcomes the degradation caused by training on sequences with imbalanced signal levels, when using a level-dependent loss function.

Keywords: Data augmentation · Speech enhancement · Deep noise suppression

1 Introduction

Speech enhancement using neural networks has recently seen large attention and success in research [9,16] and is being implemented in commercial applications also targeting real-time communication. An exciting property of deep learning-based noise suppression is that it also reduces highly non-stationary noise and background sounds such as barking dogs, banging kitchen utensils, crying babies, construction or traffic noise, etc. This has not been possible so far using single-channel statistical model-driven speech enhancement techniques that often only reduce quasi-stationary noise [2,4,8]. Notable approaches towards real-time implementations have been proposed e. g. in [12–14,17,20].

The dataset is a key part of data-driven learning approaches, especially for supervised learning. It is a challenge to build a dataset that is large enough to generalize well, but still represents the expected real-world data sufficiently. Data augmentation techniques can not only help to control the amount of data, but is also necessary to synthesize training data that represents all effects encountered in practice.

A. Karpov and R. Potapova (Eds.): SPECOM 2020, LNAI 12335, pp. 79–86, 2020.
https://doi.org/10.1007/978-3-030-60276-5_8

While in many publications, data corpus generation is only roughly outlined due to lack of space, or often exclude several key practical aspects, we direct this paper on showing contributions on several augmentation techniques when synthesizing a noisy and target speech corpus for speech enhancement. In particular, we show the effects of increasing the SNR range and using a continuous instead of discrete distribution, spectral augmentation by applying random spectral shaping filters to speech and noise, and finally level augmentation to increase robustness of the network against varying input signal levels.

As we found that level augmentation can decrease the performance when using signal-level dependent losses, we propose a normalization technique for the loss computation that can be generalized to any other signal-based loss. We show in experiments on the CHIME-2 challenge dataset that the augmentation techniques and loss normalization substantially improve the training procedure and the results.

In this paper, we first introduce a the general noise suppression task in Sect. 2. In Sect. 3, we describe the used real-time noise suppression system based on a recurrent network, that works on a single frame in - single frame out basis, i. e. requires no look-ahead and memory buffer, and describe the training setup. In Sect. 4, we describe the used loss function and propose a normalization to remove the signal level dependency of the loss. In Sect. 5, we describe augmentation techniques for signal-to-noise ratio (SNR), spectral shaping, and sequence level dynamics. The experiments are shown in Sect. 6, and Sect. 7 concludes the paper.

2 Deep Learning Based Noise Suppression

In a pure noise reduction task, we assume that the observed signal is an additive mixture of the desired speech and noise. We denote the observed signal $X(k,n)$ directly in the short-time Fourier transform (STFT) domain, where k and n are the frequency and time frame indices as

$$X(k,n) = S(k,n) + N(k,n), \tag{1}$$

where $S(k,n)$ is the speech and $N(k,n)$ is the disturbing noise signal. Note that the speech signal $S(k,n)$ can be reverberant, and we only aim at reducing additive noise.

The objective is to recover an estimate $\widehat{S}(k,n)$ of the speech signal by applying a filter $G(k,n)$ to the observed signal by

$$\widehat{S}(k,n) = G(k,n)\,X(k,n). \tag{2}$$

The filter $G(k,n)$ can be either a real-valued suppression gain, or a complex-valued filter. While the former option (also known as *mask*) only recovers the speech amplitude, a complex filter could potentially also correct the signal phase. In this work, we use a suppression gain.

3 Network and Training

We use a rather straightforward recurrent network architecture based on gated recurrent units (GRUs) [1] and feed forward (FF) layers, similar to the core architecture of [19] without convolutional encoder layers. Input features are the logarithmic power spectrum $P = \log_{10}(|X(k,n)|^2 + \epsilon)$, normalized by the global mean and variance of the training set. We use a STFT size of 512 with 32 ms square-root Hann windows and 16 ms frame shift, but feed only the relevant 255 frequency bins into the network, omitting 0th and highest (Nyquist) bins, which do not carry useful information. The network consists of a FF embedding layer, two GRUs, and three FF mapping layers. All FF layers use rectified linear unit (ReLU) activations, except for the last output layer. When estimating a real-valued suppression gain, a *Sigmoid* activation is used to ensure positive output. The network architecture is shown in Fig. 1, and has 2.8 M parameters.

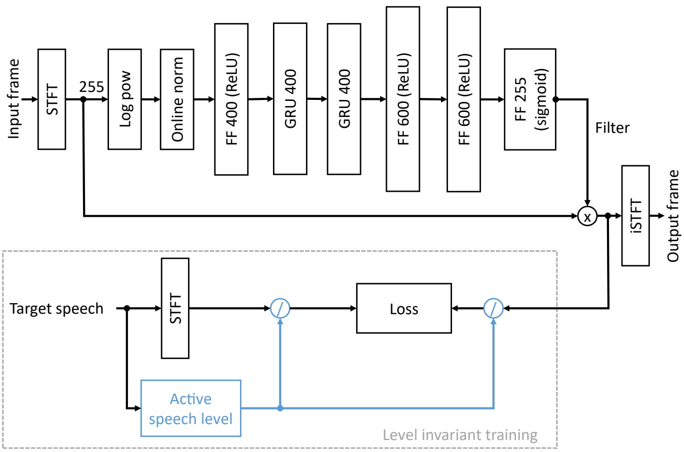

Fig. 1. Network architecture and enhancement system and training procedure.

The network was trained using the AdamW optimizer [7] with an initial learning rate of 10^{-4}, which was dropped by a factor of 0.9 if the loss plateaued for 5 epochs. The training was monitored every 10 epochs using a validation subset. The best model was chosen based on the highest perceptual evaluation of speech quality (PESQ) [6] on the validation set. All hyper-parameters were optimized by a grid search and choosing the best performing parameter for PESQ on the validation set.

4 Level Invariant Normalized Loss Function

The speech enhancement loss function is typically a distance metric between the enhanced and target spectral representations. The dynamically compressed loss

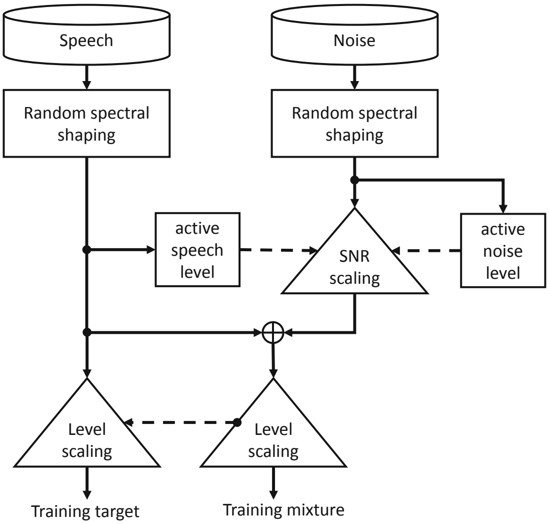

Fig. 2. On-the-fly training augmented data generation.

proposed in [3,18] has been shown to be very effective. A compression exponent of $0 < c \leq 1$ is applied to the magnitudes, while the compressed magnitudes are combined with the phase factors again. Furthermore, the magnitude only loss is blended with the complex loss with a factor $0 \leq \alpha \leq 1$.

$$\mathcal{L} = \alpha \sum_{k,n} \left| |S|^c e^{j\varphi_S} - |\widehat{S}|^c e^{j\varphi_{\widehat{S}}} \right|^2 + (1-\alpha) \sum_{k,n} \left| |S|^c - |\widehat{S}|^c \right|^2. \tag{3}$$

We chose $c = 0.3$ and $\alpha = 0.3$.

A common drawback of all similar signal-based loss functions is the dependency on the level of the signals S and $\widehat{S}$. This might have an impact on the loss when computing the loss over a batch of several sequences, which exhibit large dynamic differences. It could be that large signals dominate the loss, creating less balanced training.

Therefore, we propose to normalize the signals S and X by the active signal level of each utterance, before computing the loss. The normalized loss is computed as given by (3), but using the normalized signals $\tilde{S} = \frac{S}{\sigma_S}$ and $\tilde{X} = \frac{X}{\sigma_S}$, where σ_S is the active speech level per utterance, i.e. the target speech signal standard deviation computed only for active speech frames. Note that this normalization does not affect the input features of the network: they still exhibit the original dynamic levels.

5 Data Augmentation Techniques

Especially for small and medium-scale datasets, augmentation is a powerful tool to improve the results. In the case of supervised speech enhancement training,

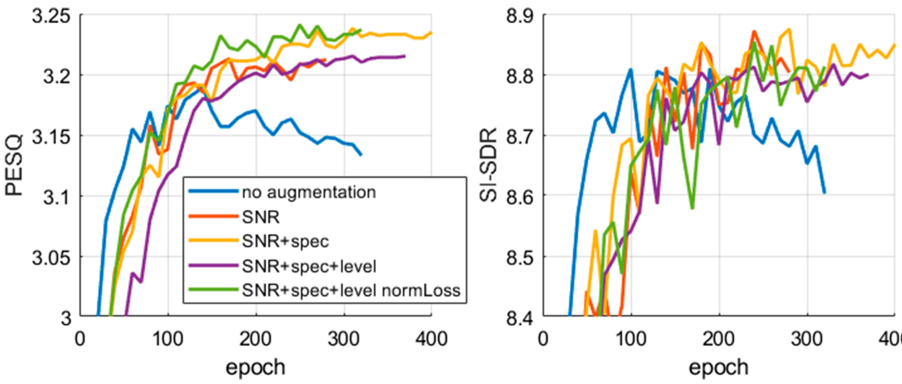

Fig. 3. Validation metrics for various augmentation techniques. (Color figure online)

where the actual noisy audio training data is generated synthetically by mixing speech and noise, there are some augmentation steps, which are essential to mimic effects on data encountered in the wild. Disregarding reverberation, we need to be able to deal with different SNRs, audio levels, and filtering effects that can be caused e.g. by acoustics (room, occlusion), or the recording device (microphone, electronic transfer functions).

Our augmentation pipeline is shown in Fig. 2. Before mixing speech with noise, we applied random biquad filters [14] to each noise sequence and speech sequences separately to mimic different acoustic transmission effects. From these signals, active speech and noise levels are computed using a level threshold-based voice activity detector (VAD). Speech and noise sequences are then mixed with a given SNR on-the-fly during training. After mixing, the mixture is scaled using a given level distribution. The clean speech target is scaled by the same factor as the mixture. The data generation and augmentation procedure is depicted in Fig. 2.

6 Experiments

6.1 Dataset and Experimental Setup

We used the CHIME-2 WSJ-20k dataset [15], which is currently, while only being of medium size, the only realistic self-contained public dataset including matching reverberant speech and noise conditions. The dataset contains 7138, 2418, and 1998 utterances for training, validation and testing, respectively. The utterances are reverberant using binaural room impulse responses, and noise from the same rooms was added with SNRs in the range of -6 to 9 dB in the validation and test sets. We used only the left channel for our single-channel experiments.

The spectral augmentation filters are designed as proposed in [14] by

$$H(z) = \frac{1 + r_1 z^{-1} + r_2 z^{-2}}{1 + r_3 z^{-1} + r_4 z^{-2}} \quad (4)$$

with r_i being uniformly distributed in $[-\frac{3}{8}, \frac{3}{8}]$. For SNR augmentation, the mixing SNRs were drawn from a Gaussian distribution on the logarithmic scale with mean 5 dB and standard deviation 10 dB. The signal levels for dynamic range augmentation were drawn from a Gaussian distribution with mean −28 dBFS and variance 10 dB.

We evaluate our experiments on the development and test set from the CHIME-2 challenge. As objective metrics, we use PESQ [6], short-time objective intelligibility (STOI) [11], scale-invariant signal-to-distortion ratio (SI-SDR) [10], cepstral distance (CD), and frequency-weighted segmental SNR (fwSegSNR) [5].

6.2 Results

Figure 3 shows the training progression in terms of PESQ and SI-SDR on the validation set. The blue curve shows training on the original data without any augmentation, mixed the 6 different SNR levels between −6 and 9 dB. We can see that the validation metrics decrease after 150 epochs. When applying SNR augmentation (red curve) with a broader and continuous distribution, we prevent the early validation decrease and can train for 280 epochs. Further, adding spectral augmentation increases the validation PESQ slightly. However, the level augmentation when training with standard loss (3) (purple curve) decreases the performance compared to SNR and spectral augmentation only. We attribute this effect to the large level imbalance per batch, which affects the standard level-dependent loss function. When computing the loss from normalized signals (green curve), this drawback is overcome and we obtain similar or even slightly better results than the yellow curve, but making the system robust to varying input signal levels. The validation SI-SDR shows similar behavior as PESQ.

Table 1. Evaluation metrics on CHIME-2 test set.

Augmentation	Loss	PESQ	STOI	CD	SI-SDR	fwSegSNR
–	Noisy	2.29	81.39	5.46	1.92	16.96
None	Standard	3.27	91.20	2.90	9.48	23.57
SNR	Standard	3.31	91.40	**2.85**	9.45	23.30
SNR+spec	Standard	**3.32**	91.57	2.89	9.55	23.30
SNR+spec+level	Standard	3.30	**91.68**	2.87	**9.57**	**23.48**
SNR+spec+level	Normalized	3.31	91.55	2.89	9.52	23.41

In Table 1 shows the results on the CHIME-2 test set. The enhancement systems improve all results substantially over the noisy input. Adding SNR augmentation adds a gain of 0.05 PESQ. As in the development set, spectral augmentation adds an additional minor improvement. Interestingly, on the test set, the normalized loss shows no influence on the results. We assume this is due to that fact that the given test set does not exhibit largely varying signal levels.

7 Conclusion

We have shown the effectivity of data augmentation techniques for supervised deep learning-based speech enhancement by augmenting the SNR, spectral shapes, and signal levels. As level augmentation degrades the performance of the learning algorithm when using level-dependent losses, we proposed a normalization technique for the loss, which is shown to overcome this issue. The experiments were conducted using a real-time capable recurrent neural network on the reverberant CHIME-2 dataset. Future work will also investigate augmentation for acoustic conditions with reverberant impulse responses.

References

1. Cho, K., Merriënboer, B.V., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: encoder-decoder approaches. In: Proceedings of the Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8) (2014)
2. Ephraim, Y., Malah, D.: Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. **32**(6), 1109–1121 (1984)
3. Ephrat, A., et al.: Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. ACM Trans. Graph. **37**(4), 112:1–112:11 (2018)
4. Gerkmann, T., Hendriks, R.C.: Noise power estimation based on the probability of speech presence. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 145–148, October 2011
5. Hu, K., Divenyi, P., Ellis, D., Jin, Z., Shinn-Cunningham, B.G., Wang, D.: Preliminary intelligibility tests of a monaural speech segregation system. In: Proceedings of the Workshop on Statistical and Perceptual Audition, Brisbane, September 2008
6. ITU-T: Recommendation P.862: Perceptual evaluation of speech quality (PESQ), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs, February 2001
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=Bkg6RiCqY7
8. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**, 504–512 (2001)
9. Reddy, C.K.A., et al.: The INTERSPEECH 2020 deep noise suppression challenge: datasets, subjective speech quality and testing framework. In: Proceedings of the INTERSPEECH 2020 (2020, to appear)
10. Roux, J.L., Wisdom, S., Erdogan, H., Hershey, J.R.: SDR - half-baked or well done? In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 626–630, May 2019
11. Taal, C.H., Hendriks, R.C., Heusdens, R., Jensen, J.: An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Trans. Audio Speech Lang. Process. **19**(7), 2125–2136 (2011)
12. Tan, K., Wang, D.: A convolutional recurrent neural network for real-time speech enhancement. In: Proceedings of the Interspeech, pp. 3229–3233 (2018)

13. Tu, Y.H., Tashev, I., Zarar, S., Lee, C.: A hybrid approach to combining conventional and deep learning techniques for single-channel speech enhancement and recognition. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2531–2535, April 2018
14. Valin, J.: A hybrid DSP/deep learning approach to real-time full-band speech enhancement. In: 20th International Workshop on Multimedia Signal Processing (MMSP), pp. 1–5, August 2018
15. Vincent, E., Barker, J., Watanabe, S., Nesta, F.: The second 'CHIME' speech separation and recognition challenge: datasets, tasks and baselines. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2012
16. Wang, D., Chen, J.: Supervised speech separation based on deep learning: an overview. IEEE/ACM Trans. Audio Speech Lang. Process. **26**(10), 1702–1726 (2018)
17. Wichern, G., Lukin, A.: Low-latency approximation of bidirectional recurrent networks for speech denoising. In: Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 66–70, October 2017
18. Wilson, K., et al.: Exploring tradeoffs in models for low-latency speech enhancement. In: Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC), pp. 366–370, September 2018
19. Wisdom, S., et al.: Differentiable consistency constraints for improved deep speech enhancement. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 900–904, May 2019
20. Xia, R., Braun, S., Reddy, C., Dubey, H., Cutler, R., Tahev, I.: Weighted speech distortion losses for neural-network-based real-time speech enhancement. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2020)