Speech enhancement augmentation for robust speech recognition in noisy environments

Rauf Nasretdinov, *Andrey* Lependin*, and *Ilya* Ilyashenko

Department of Information Security, Institute of Digital Technology, Electronics and Physics, Altai State University, 61 Lenina pr, Barnaul, 656049, Russia

Abstract. The use of augmentations as a data enrichment method has become an important element in improving the performance of speech recognition systems. To work effectively in noisy conditions, augmentation is usually used to simulate the presence of background noise. However, the quality of speech recognition on samples pre-processed by noise reduction models does not increase. This paper proposes a new approach to speech data augmentation when training ASR systems, intended for their joint use with models for speech enhancement. It was based on the creation of several additional data samples containing speech samples processed by the enhancement model. The proposed approach was tested on the E-Branchformer neural network model using data from the Librispeech set. The quality of speech samples was assessed using the DNSMOS metric. By means of a 100-hour sample of clean speech samples it was shown that the proposed augmentation allows for an improvement in the WER metric of more than 4% in absolute value compared to the generally accepted approach based on adding noisy speech samples. Experiments on 960-hour data demonstrated the robustness of this approach as the training set size increased.

1 Introduction

The development of speech recognition systems has attracted significant attention in recent years, especially in the context of the widespread adoption of voice assistants, smart city technologies, and automated video captioning. The improvement in the efficiency of these systems is largely due to advances in deep learning. The use of neural network models and approaches in the task of speech recognition has led to a significant increase in accuracy close to the level of human perception. However, the presence of additional distortions, such as background noise, reverberation and interference in the signal transmission channel, significantly reduces the effectiveness of recognition models.

Several approaches have been taken to address the distortion problem. speech enhancement models can be used to pre-process the recognized audio signal. Also, augmentation with artificially noisy records is widely used to enrich training data. This allows models to be trained on a variety of data, increases their generalization ability and stability to various conditions for recording a speech signal. Often, both approaches demonstrate

* Corresponding author: andrey.lependin@gmail.com

simultaneous improvement in recognition quality in noisy signal recording conditions and degradation in clean speech samples.

This article proposes a new approach to data augmentation for speech recognition models working in conjunction with speech enhancement systems. It uses a combination of several methods for processing speech signals of the training sample, simulating the use of noise reduction methods. The performance of the proposed approach is demonstrated both for the scenario of using a sample of limited size in the case of low-resource languages, and for a large amount of training data. An analysis and discussion of the proposed processing methods is carried out from the point of view of the perceptual quality of the received audio signals.

2 Review of existing methods

One of the first approaches to using data augmentation for automatic speech recognition was proposed in [1]. By applying random linear distortion in the frequency domain to transform spectrograms, the authors demonstrate significant improvement in phoneme error on a test subsample of the TIMIT dataset. These improvements were achieved without increasing the number of training epochs and highlight the potential of data transformation as a key element of training neural networks for speech recognition, especially in data-constrained projects. Another widely used augmentation method in speech recognition is the approach proposed in [2]. This method is based on changing the speed of playback of audio recordings in order to create variations in speech rate. The use of rate distortion provides robustness to changes in speech rate, which is an important aspect in real-world scenarios where speaking speed may vary depending on the context.

[3] presents the SpecAugment method, which introduces random distortion into a signal by masking successive segments in both the time and frequency domains. Application of the method to deep learning networks for speech recognition tasks allowed us to achieve advanced results on the LibriSpeech [4] and Swichboard data sets. In [5], the possibility of improving the performance of speech recognition systems using the Tacotron speech synthesis architecture and its variations was considered. These architectures have enabled the creation of natural-sounding, polyphonic synthesized speech, offering the prospect of replacing expensive manual transcription of training data for speech recognition.

The approaches described above are not able to eliminate the degradation of speech recognition if the processed samples contain extraneous noise. As a solution to this problem, training of speech recognition systems using a large amount of noisy data was proposed in [6]. In [7], the authors point out the problem of using noise reduction methods for speech recognition, which lies in the difficulty of choosing the required level of noise reduction for signals with different signal-to-noise ratios. The authors proposed a new training error function for speech enhancement systems, which can be used to control the degree of noise reduction depending on the input signal. Another approach presented in [8] is based on the use of recurrent deep networks with attention to improve the quality of speech recordings in the time domain as signal pre-processing. The noise reduction and speech recognition models were trained separately.

3 Models and datasets used

3.1 Description of the speech recognition model

Recently, neural network models based on the Conformer network [9] have demonstrated the best quality when solving speech recognition problems. Models of this type contain components for extracting both global features from the entire input signal and local features

in its individual fragments. In the original work [9], self-attention layers [10] were responsible for extracting global features, and convolution-based layers were responsible for extracting local ones. This approach proved to be extremely effective and demonstrated the best quality on many data sets.

In this work, we used the E-Branchformer network [11], which develops the approach of [9]. In it, the same self-attention layers were responsible for extracting global signal features, but the part that extracted local features consisted of a block based on the Convolutional Spatial Gating Unit [9]. The main difference with E-Branchformer was that these parts were not calculated sequentially, as in the case of the Conformer network, but in parallel. This approach has been shown to perform better on a number of data sets. This determined our choice of this model for conducting numerical experiments.

3.2 Model for speech enhancement

Some of the audio signal augmentation methods considered in this work are based on the use of speech enhancement models. Most modern speech enhancement approaches are based on deep neural networks [12–14]. Unlike methods based on “classical” digital signal processing [15], neural network methods can work in a wider range of different types of non-stationary noise, while demonstrating better quality. They allow not only to clear the signal from noise and improve speech intelligibility, but also to preserve the timbre, intonation and individual speech characteristics of the speaker.

Recently, a large number of neural network methods for improving speech quality have been proposed, working both in real time [13, 16] and offline [14]. The first type mainly includes models based on recurrent or convolutional neural networks [17], the second type includes models using self-attention layers [18]. As a rule, models of the second type demonstrate better quality of the reconstructed speech signal, while sacrificing processing speed.

To construct additional training and test data sets, the work used the denoising model proposed in [19]. It was based on the architecture of a visual transformer, expanded to a pyramidal structure such as a U-Net network [19]. The use of self-attention to time-frequency fragments of the signal in this approach ensured a quality of noise reduction that was superior to existing analogues.

3.3 Librispeech dataset

To carry out numerical experiments, the Librispeech data set, standard for speech recognition tasks, was used [4]. It was collected from audio recordings of English-language audiobooks from the LibriVox corpus [4]. This set included several subsets of 100 hours (“clean_100”), 360 hours (“clean_360”) and 500 hours (“other_500”), which can be used individually or together. To form the first two parts, a method was used based on the use of a model neural network pretrained on Wall Street Journal data [4]. This sample network was used to recognize all examples of the Librispeech set. For them, the WER metric was assessed (see below, in section 4.1). Examples with the best results were selected into the “clean” subset, the rest were included in the “other”. Test examples were selected according to a similar principle. Note that the discrepancy in speech recognition quality was due to the quality of the original clean recordings and differences in speakers' accents, and not to the presence of additional external noise in the signal.

3.4 Dataset with noise samples

Samples of noise audio signals were selected from the DNS Challenge 2021 set [20]. The examples in this set were selected from several sources in such a way that there were at least 500 records for each noise class, thereby solving the problem of class imbalance. The sample included a total of 150 noise classes, with a total of more than 65,000 samples, each lasting at least 10 seconds.

4 Description of metrics

4.1 Metric for speech recognition

To assess the quality of speech recognition, the Word Error Rate (WER) metric [21] was used, which measured the accuracy of word recognition in test examples. The formula for calculating WER was determined as follows:

$$WER = (S + D + I)/N, \quad (1)$$

where S (substitutions) - number of words that were incorrectly recognized, D (deletions) - number of missing words, I (insertions) - number of extra (inserted) words, N - the total number of words in the reference text. The lower the WER, the better the quality of speech recognition.

4.2 Metric for assessing the quality of the speech signal

The standard method for assessing the quality of speech signals is the average value of the opinion of expert listeners (Mean Opinion Score, MOS) [22]. MOS is calculated by averaging the ratings of audio recordings given by 20-60 experts on a scale from 1 (unacceptable quality) to 5 (excellent quality). As stated in [22], a MOS score of 4.0 or higher corresponds to a high-quality speech signal.

In practice, using expert judgments for sufficiently large speech data sets is difficult. Therefore, at present, the DNSMOS metric has begun to be used, which is an estimate of the MOS value using a pre-trained neural network [23]. An important feature of the DNSMOS metric is the ability to assess the quality of a speech signal without having an exemplary "clean" version, which allows it to be used in a wide range of tasks.

5 Experiments

5.1 Generation of augmented speech samples

To conduct numerical experiments, samples of speech samples were generated based on the Librispeech dataset with various types of augmentations applied. Samples were created for two series of experiments: using the 100-hour subset "clean_100" and using the 960-hour subset, which was a union of the "clean_100", "clean_360" and "other_500" subsets described above in section 3.3. A brief description of all transformations applied to the samples is given in Table 1. Similar transformations were also applied to the corresponding test samples of the Librispeech set (not shown in Table 1).

Noise was added to the "Noisy" sample as follows. For each example, random records were selected from the "Clean" sample from a set of noise samples [20]. In the case where the noise sample exceeded the pure one in duration, it was trimmed to the length of the pure

one. Otherwise, the next randomly selected sample was added to this noise sample. Next, the collected noise additive was amplified or attenuated in amplitude so that the signal-to-noise ratio for the sum of the pure signal and the additive corresponded to a randomly selected value from a uniform distribution over the interval from 0 dB to 40 dB. No additional artificial distortions were introduced into the audio signal (reverberation, modeling of the recording path, etc.).

The transformation to obtain the "Enhanced" and "Denoised" samples consisted of applying a pre-trained neural network model of speech enhancement [19] to the corresponding sample ("Clean" or "Noisy"). It should be noted that the application of this model to clean data ("Clean") is justified, since when removing distortions in the speech signal, the model not only clears noisy time-frequency fragments, but also, as shown in [19], restores lost fragments in these fragments. formant components and noise-like consonants. Thus, the clean speech audio signal after quality enhancement was applied typically did not exactly match the original.

Table 1 also shows the durations of the corresponding samples for both series of experiments and the average values of the DNSMOS quality metric based on the data from the 100-hour experiment. The latter quite accurately characterize the nature of how much the corresponding augmentation changes the perceived quality of the speech signal. The use of standard recording acceleration/deceleration does not have a significant impact on expert quality assessments. Using the speech enhancement model increases the average DNSMOS value even for the "Clean" sample, which is likely due to the imperfection of the original recordings of the "clean" subset "clean_100" of the Librispeech set and the presence of signal distortions and extraneous noise in them. As expected, the noisy data had the lowest average value of the quality metric. Applying a speech enhancement model to them made it possible to obtain a selection of speech samples simulating the use of noise reduction in real conditions, when a speech recognition system works with low-quality signals.

Table 1. Description of samples used when training a speech recognition model.

Sample designation	Brief description of the applied transformation	Duration, h	DNSMOS_100
Clean	No transformations. Original samples from the Librispeech dataset	100 / 960	4.02
Clean_sp	Applying speech rate change augmentation with coefficients 0.9, 1.0, 1.1 to the "Clean" sample	300 / 2880	4.02
Enhanced	Applying the Speech Enhancement Model to Clean Sample Samples	100 / 960	4.06
Noisy	Adding Noise to the Clean Subsample Using Noise Samples	100 / 960	2.72
Denoised	Applying the Speech Enhancement model to samples from the Noisy sample	100 / 960	3.59

5.2 Training ASR models on generated samples

The acoustic features used were mel-frequency spectral coefficients with a dimension of 80, calculated on the basis of a windowed Fourier transform with a frame duration of 25 ms, an overlap of successive frames of 10 ms, and a Hann window function. SpecAugment [3] was also applied during training. BPE (Byte-Pair Encoding) partition [24] of size 5000 was used as output tokens.

All experiments were carried out using the ESPnet neural network speech signal processing library [25]. The E-Branchformer architecture from the standard ESPNet recipe for Librispeech for the “clean_100” subset was used as a neural network model. The number of model parameters was 3.847·107. The training was carried out on two Nvidia 1080 Ti video cards. All experiments on the 100-hour subset went through the same number of training steps of the order of 3.55·105, for 960 hours – 1.7·106. To train the models, the adam optimizer was used with a warmup learning rate schedule, the peak was reached at 15% of training with a maximum value of 0.001.

During training, the CTC-attention loss function [26] was used with a CTC weight of 0.3. When testing the models, beam search decoding was used with the beam size parameter of 20 without using language models.

5.3 Results and discussion

Table 2 shows the results of numerical experiments using the 100-hour “clean_100” subset and augmented samples generated on its basis. Each row of Table 2 corresponds to a separately trained ASR model. All versions of the model differed only in the structure of the training set.

From the results given in rows 1-2 of Table 2, it is clear that when adding a 100-hour “Enhanced” sample containing “improved” versions of clean speech signals, no changes were observed in the quality of recognition both in the pure tests (“Test Clean” and “Test Other"), and on noisy (“Test Clean Noisy” and “Test Other Noisy”) and cleaned noisy (“Test Clean Denoised” and “Test Other Denoised”) test samples. The addition of speech rate change augmentation (rows 3–4 of Table 2) resulted in some improvement in WER, apparently unrelated to the addition of the “Enhanced” sample. From Fig. 1 shows that a possible explanation for the weak impact of adding this training sample on the results is associated with the small difference in the distributions of the DNSMOS metric for the “Clean” and “Enhanced” samples.

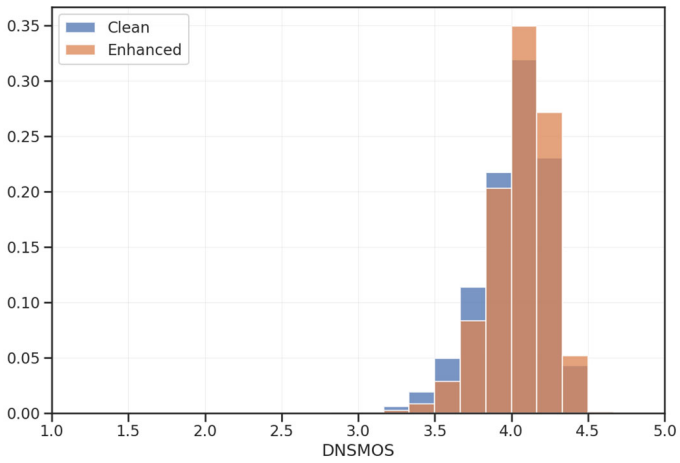

Fig. 1. Comparison of DNSMOS speech quality metric distributions on Clean and Enhanced speech samples.

An obvious step to improve WER on test samples containing denoised speech was to add a sample with appropriate augmentation to the clean training set (row 5 of Table 2). This resulted in a decrease in WER on the cleaned noisy test samples by approximately 6–7% (in absolute value). However, the quality of recognition on clean test examples deteriorated

somewhat. As can be seen from Fig. 2, the distribution of the DNSMOS metric for samples cleared of noise captured a wider range of values compared to the "Clean" sample.

An unexpected result was that the addition of the "Enhanced" training sample (line 6) made it possible to improve the quality of recognition both on clean samples (0.5–0.6% absolute reduction in WER), and additionally on all test cleaned and test cleaned noise samples (0.7–1.3 %).

To compensate for the small lag from the experiment on the "Clean_sp" data (line 4), the model was trained on data expanded compared to experiment 6 by adding samples with changes in speed (line 7). This made it possible to eliminate the lag on "clean" speech examples, while maintaining high quality on test subsets cleared of noise.

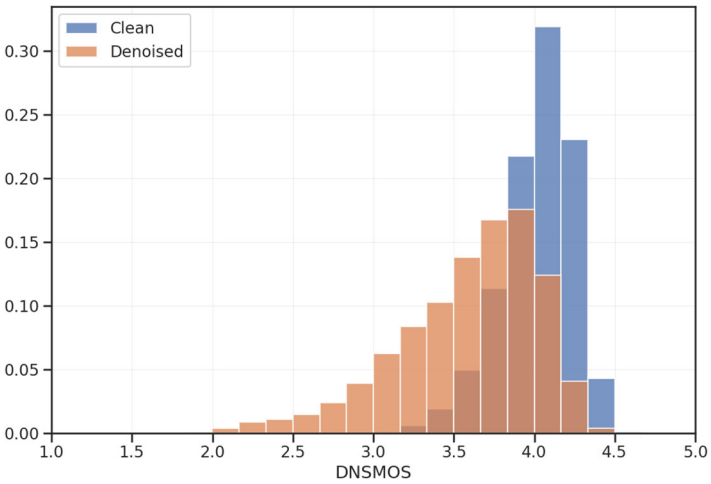

Fig. 2. Comparison of distributions of the DNSMOS speech quality metric on samples of clean (Clean) and improved noisy (Denoised) speech samples.

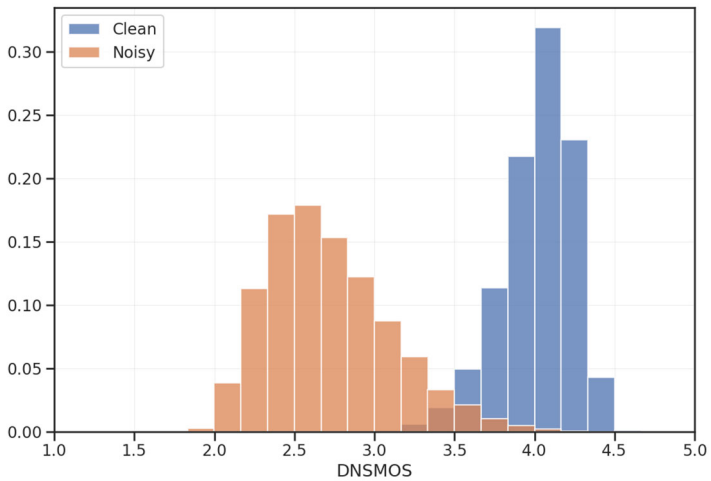

Fig. 3. Comparison of distributions of the DNSMOS speech quality metric on samples of clean (Clean) and noisy (Noisy) speech samples.

Next, an experiment was carried out with the addition of augmentations with noisy data (line 8 of Table 2). In this case, the quality on clean test samples degraded, and on tests

cleared of noise ("Test Clean Denoised" and "Test Other Denoised") it turned out to be slightly lower than in experiments 5–7. The ASR model appeared to be trained to better recognize the extremes of clean and noisy signals, which had a significant difference in speech quality distribution from each other (Figure 3) and did not capture the average DNSMOS speech samples obtained by cleaning the noisy audio signals.

When adding "Denoised" and "Enhanced" data (line 9) to the training samples of the previous experiment, the quality on "clean" tests became comparable to experiment 3. At the same time, a significant improvement was observed (4–4.2% absolute change in the WER metric) on clean tests compared to experiment 8. Thus, the combined use of all types of augmentations led to a significant improvement in the quality of ASR when recognizing both noisy and denoised speech signals, without accompanying degradation on clean samples.

Table 2. Test results of the ASR model trained on 100-hour samples.

N	Samples for training	WER, %					
		Test Clean	Test Other	Test Clean Denoised	Test Clean Noisy	Test Other Denoised	Test Other Noisy
1	Clean	7.0	18.4	31.4	57.8	51.5	74.2
2	Clean + Enhanced	6.9	18.3	31.1	57.9	51.4	74.1
3	Clean_sp	6.7	17.5	30.9	57.4	50.8	73.2
4	Clean_sp + Enhanced	6.7	17.2	30.7	56.0	49.8	71.9
5	Clean + Denoised	7.2	18.5	24.7	52.7	43.9	70.4
6	Clean + Denoised + Enhanced	6.8	17.9	24.0	51.4	43.2	68.5
7	Clean_sp + Denoised + Enhanced	6.6	17.5	23.1	49.8	42.3	68.0
8	Clean_sp + Noisy	7.5	17.9	26.5	26.6	45.8	46.7
9	Clean_sp + Noisy + Denoised + Enhanced	6.9	17.3	22.5	27.3	41.6	47.7

Table 3 presents the results for testing the approach on a larger data set using a 960-hour pool of all subsets of the Librispeech set. The table shows that the simultaneous use of several training samples ("Clean_sp", "Noisy", "Denoised" and "Enhanced") made it possible to obtain results qualitatively similar to the experiments described above for 100 hours. This indicates the promise and sustainability of the approach to augmentation developed in this work.

Table 3. Test results of the ASR model trained on 960-hour samples.

N	Samples for training	WER, %					
		Test Clean	Test Other	Test Clean Denoised	Test Clean Noisy	Test Other Denoised	Test Other Noisy
1	Clean_sp	2.8	6.8	21.9	40.6	37	56.4
2	Clean_sp + Noisy + Denoised + Enhanced	2.9	6.9	12.8	14.8	25.9	28.9

6 Conclusion

This article proposes a new approach to the augmentation of speech samples for training a neural network model of speech recognition. Its essence is to add signals processed by a speech enhancement model to the training sample. It was shown that adding a sample cleared of artificial model noise made it possible to reduce the WER error in the scenario of combined speech enhancement and ASR. Possible degradation in the quality of speech recognition on clean samples was compensated by adding an additional augmented sample, consisting of samples of original speech processed by the denoising model. The proposed approach to augmentation can be used in training speech recognition systems designed to work in difficult conditions of a noisy environment.

The research was supported by the Russian Science Foundation grant No. 22–21–00199, https://rscf.ru/en/project/22-21-00199/.

References

1. N. Jaitly, G. E. Hinton, *Vocal tract length perturbation (VTLP) improves speech recognition*, in Proceedings of the International Conference on Machine Learning, ICML, Workshop on Deep Learning for Audio, Speech, and Language Processing, 20-21 June 2013, Atlanta, USA (2013)
2. T. Ko, V. Peddinti, D. Povey, S. Khudanpur, *Audio Augmentation for Speech Recognition*, in Proceedings of the Interspeech, 6-10 September 2015, Dresden, Germany (2015)
3. D. S. Park, W. Chan, Y. Zhang, C. Chiu, B. Zoph, E. D. Cubuk, *Specaugment: A simple data augmentation method for automatic speech recognition*, in Proceedings of the Interspeech, 15-19 September 2019, Graz, Austria (2019)
4. V. Panayotov, G. Chen, D. Povey, S. Khudanpur, *LibriSpeech: An ASR corpus based on public domain audio books,* in Proceedings of the International Conference on Acoustics, Speech and Signal Processing, ICASSP, 19-24 April 2015, Brisbane, Queensland (2015)
5. A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. Moreno, Y. Wu, Z. Wu, *Speech recognition with augmented synthesized speech*, in Proceedings of the IEEE automatic speech recognition and understanding workshop, ASRU, 14-18 December 2019, Sentosa, Singapore (2019)
6. A. Hannun, C. Case, J. Casper, B. Catanzaro, *Deep Speech: Scaling up end-to-end speech recognition,* arXiv preprint arXiv:1412.5567
7. Y. Koizumi, S. Karita, A. Narayanan, *SNRi Target Training for Joint Speech Enhancement and Recognition*, in Proceedings of the Interspeech, 18-22 September 2022, Incheon, Korea (2022)
8. Y. Yang, A. Pandey, D. Wang, *Time-Domain Speech Enhancement for Robust Automatic Speech Recognition*, in Proceedings of the Interspeech, 20-24 August 2023, Dublin, Ireland (2023)
9. A. Gulati, J. Qin, C. Chiu, N. Parmar, *Conformer: Convolution-augmented Transformer for Speech Recognition,* in Proceedings of the Interspeech, 25-29 October 2020, Shanghai, China (2020)
10. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, *Attention Is All You Need*, in Proceedings of the Neural Information Processing Systems, NIPS, 4-9 December 2017, Los Angeles, USA (2017)

11. K. Kim, F. Wu, Y. Peng, *E-Branchformer: Branchformer with Enhanced Merging for Speech Recognition*, in Proceedings of the IEEE Spoken Language Technology Workshop (SLT), 9-12 January 2023, Doha, Qatar (2023)
12. Y. Luo, N. Mesgarani, *Conv-TasNet: surpassing ideal time-frequency magnitude masking for speech separation*, IEEE/ACM Trans. Audio Speech Lang. Process. **27**, 8, 1256–1266 (2019)
13. X. Hao, X. Su, R. Horaud, X. Li, *FullSubNet: a full-band and sub-band fusion model for real-time single-channel speech enhancement,* in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 6–11 June 2021, Toronto, Canada (2021)
14. A. Tan, D.A. Wang, *Convolutional recurrent neural network for real-time speech enhancement*, in Proceedings of the Interspeech, 2-6 September 2018, Hyderabad, India (2018)
15. P. Loizou, *Speech Enhancemen: Theory and Practice, Second Edition* (CRC Press, 2013)
16. R. Nasretdinov, I. Ilyashenko, A. Lependin, *Two-Stage Method of Speech Denoising by Long Short-Term Memory Neural Network,* C.C.I.S, **1526**, Springer, Cham (2022)
17. Y. Hu, Y. Liu, S. Lv, M. Xing, *DCCRN: deep complex convolution recurrent network for phase-aware speech enhancement*, in Proceedings of the Interspeech, 25–29 October 2020, Shanghai, China (2020)
18. C. Zheng, X. Peng, Y. Zhang, *Interactive Speech and Noise Modeling for Speech Enhancement*, arXiv preprint arXiv:2012.09408
19. R. Nasretdinov, I. Ilyashenko, J. Filin, A. Lependin, *Hierarchical Encoder-Decoder Neural Network with Self-Attention for Single-Channel Speech Denoising,* CCIS, **1733**, Springer, Cham (2023)
20. C. Reddy, H. Dubey, K. Koishida, A. Nair, *INTERSPEECH 2021 Deep Noise Suppression Challenge*, in Proceedings of the Interspeech, 30 August – 3 September 2021, Bron, Czechia (2021)
21. A. Ali, S. Renals, *Word Error Rate Estimation for Speech Recognition: e-WER*, in Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 15-20 July 2018, Melbourne, Australia (2018)
22. *Itu-t recommendation P.800: Methods for subjective determination of transmission quality*, (1998)
23. R. Chandan, V. Gopal, R. Cutler, *Dnsmos: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors,* in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, 6-11 *June 2021,* Toronto, Canada (2021)
24. R. Sennrich, B. Haddow, A. Birch, *Neural Machine Translation of Rare Words with Subword Units,* in Proceedings of the of the 54th Annual Meeting of the Association for Computational Linguistics, ACL, Berlin, Germany (2016)
25. S. Watanabe, T. Hori, S. Karita, T. Hayashi, *ESPnet: End-to-End Speech Processing Toolkit*, in Proceedings of the Interspeech, 2-6 September 2018, Hyderabad, India (2018)
26. S. Kim, T. Hori, S. Watanabe, *Joint CTC-attention based end-to-end speech recognition using multi-task learning*, in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processings, ICASSP, 5-9 March 2017, New Orleans, USA (2017)