Research on Improved Conv-TasNet of Speech Enhancement for Non-Stationary and Low SNR Noise During Aircraft Operating

Deyin Zhang[a], Wenxuan Hong[b], Juntong Li[c], Yuyao Zhang[d] and Li Wang[e]
Institute of Electronic and Electrical Engineering, Civil Aviation Flight University of China, Guanghan, China

Keywords: Civil Aviation, Noise, Non-Stationary, Speech Enhancement, Neural Network.

Abstract: A speech enhancement method based on improved Conv-TasNet (Convolution Time-Domain Audio Separation Network) is proposed in this paper so as to solve the problems of the high noise environment of the airport seriously affects the communication between airport ground staff. The traditional speech enhancement algorithm used by civil aviation is not effective in suppressing low SNR (signal-to-noise) ratio and non-stationary noise. The improved Conv-TasNet is based on the baseline Conv-TasNet, and fused the multi-head-attention module and the Efficient Channel Attention Network channel attention module. The ablation experiment is carried out by using four neural networks to deal with the noisy speech of five kinds SNR: 10dB, 5dB, 0dB, -5dB respectively. The performance of the neural network is analyzed by four subjective and objective speech evaluation indicators including MOS (Mean Opinion Score), segSNR (Segment Signal-to-Noise Ratio), PESQ (Perceptual Evaluation of Speech Quality) and STOI (Short-Time Objective Intelligibility). The experiment results show that, the improved Conv-TasNet has an average increase of 1.4984 in MOS, 11.9261 in segSNR, 0.5868 in PESQ, and 0.0455 in STOI compared with the baseline Conv-TasNet. The improved neural network has better speech quality and intelligibility, which can solve the problem of used in baseline Conv-TasNet has poor effect on speech enhancement with low SNR and non-stationary environmental noise during aircraft operating.

1 INTRODUCTION

The working environment of civil aviation staff has many sources of noise, mainly including wind noise from aircraft take-off and landing, engine running noise (DING Cong et al.,2021), etc. Its characteristics are high sound level, wide influence range, three-dimensional spatial diffusion, and non-stationary, etc. When the aircraft is taking off and landing, the noise can reach 100~180 decibels (DING Cong et al., 2021), which seriously affects the speech communication between civil aviation staff (HE Liqing, 2020) and will lead to errors in communication between the two parties. Therefore, in a high-noise environment, how to reduce the harm of high noise to civil aviation staff and ensure the high quality and high efficiency of speech information exchange between relevant civil aviation staff is an urgent problem to be solved. Speech enhancement refers to extracting the original speech signal from a noisy speech signal as undistorted as possible through signal processing (Khattak Muhammad Irfan et al., 2022). The methods of speech enhancement are mainly divided into unsupervised and supervised methods (Ribas Dayana et al., 2022). Among them, unsupervised methods are mainly divided into time-domain methods and frequency-domain methods, such as spectral subtraction, wiener filtering, etc (G Thimmaraja Yadava et al., 2022; Jaiswal Rahul Kumar et al., 2022). Supervised method is mainly divided into artificial neural network (Wang Youming et al., 2021), Hidden Markov Model (E. Golrasan et al., 2016) and non-negative matrix (Tank Vanita Raj et al., 2022). Most of the speech enhancement technologies used in the civil aviation field are based on traditional field are unsupervised methods. An adaptive Kalman filtering algorithm based on wavelet analysis was used to the research of VHF speech enhancement technology (LU Yong,

[a] https://orcid.org/0000-0003-0763-4690
[b] https://orcid.org/0000-0003-4126-1189
[c] https://orcid.org/0000-0002-4985-5929
[d] https://orcid.org/0000-0002-3820-1436
[e] https://orcid.org/0000-0003-1017-8171

Zhang, D., Hong, W., Li, J., Zhang, Y. and Wang, L.
Research on Improved Conv-TasNet of Speech Enhancement for Non-Stationary and Low SNR Noise During Aircraft Operating.
DOI: 10.5220/0012047300003612
In *Proceedings of the 3rd International Symposium on Automation, Information and Computing (ISAIC 2022)*, pages 776-781
ISBN: 978-989-758-622-4; ISSN: 2975-9463

2021). A VHF speech enhancement algorithm based on improved power spectrum subtraction was proposed (YI Xue, 2020). The MMSE-LSA method for speech enhancement was adopted in the noise environment of the aircraft cockpit (HE Liqing, 2020). The traditional unsupervised speech enhancement technology is suitable for stationary noise with low sound level, and has a better suppression effect for stationary environmental noise. However, due to the limitations of the traditional algorithm, it is very insufficient for the non-stationary high noise such as airport environment, and the neural network can make up for this problem. Therefore, in order to solve the problem that the noise in the working environment of the civil aviation staff seriously affects the speech communication between personnel and damages the hearing of the personnel. An improved Conv-TasNet is proposed in this paper. It is based on the baseline Conv-TasNet (LUO Y, 2019), and fused the multi-head-attention module and the ECA-Net channel attention network. It solves the problem that the baseline Conv-TasNet has poor performance of speech enhancement in dealing with low SNR speech and non-stationary noise.

2 NEURAL NETWORK SPEECH ENHANCEMENT MODEL

2.1 Baseline Conv-TasNet

The Conv-TasNet is a time-domain based end-to-end speech separation model, which adopts an encoder-decoder structure, including an encoder, a separator, and a decoder, as shown in Figure 1.

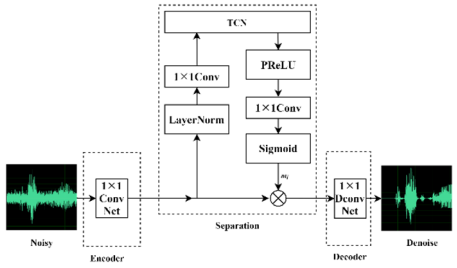

Figure 1: The structure of baseline Conv-TasNet.

Among them, the encoder consists of a single one-bit convolutional network and an activation function PReLU(Parametric Rectified Linear Unit), in which a one-dimensional convolution with a fixed convolution kernel size is used for feature extraction. The input noisy speech signal is first cut into multiple audio segments with partial overlap and fixed length.

The separator generates a weight mask, shielding the noise speech to extract the pure speech signal, and realizes the function of speech enhancement. The encoded noisy speech signal first goes through a normalization layer to keep its input distribution consistent during the training process, solving the problem of internal covariate shift caused by updating parameters, enhancing the generalization ability of the model, and avoid gradient disappearance and gradient explosion. Then use a 1×1 convolutional layer to keep the minimum number of input feature channels. The processed sequence enters the TCN layer (Y. A. Farha and J. Gall, 2019), and after the output from the TCN module, the Parametric Rectified Linear Unit is used as the activation function. Then use 1×1 convolution to restore the number of feature channels. Finally, through the Sigmoid activation function, the time domain mask of the signal source is obtained.
The decoder is similar in structure to the encoder but with opposite function. Its function is to reconstruct the separated audio signal to obtain the time domain waveform of the original signal.

2.2 Multi-Head-Attention Module

The multi-head-attention is a variant of the attention mechanism (YANG Lei et al., 2022), which is essentially an integration of several self-attention mechanisms. The structure is shown in Figure 2. Use multiple queries to splice different groups of information from input information in parallel. The advantage is that relevant information can be obtained from different subspaces.

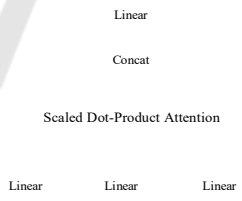

Figure 2: The structure of multi-head-attention.

2.3 Channel Attention Network

Efficient Channel Attention Network (ECA-Net) is based on the shortcomings of the SENet network (HU J et al., 2020), such as poor versatility and an additional large amount of data, by using a fast 1D convolution of size K instead of the dimension reduction operation, avoiding the side effects of the dimension reduction operation to obtain the information of

cross-channel interaction, and the effect is to obtain a greater performance improvement on the basis of adding very few parameters. The structure of ECA-Net is shown in Figure 3.

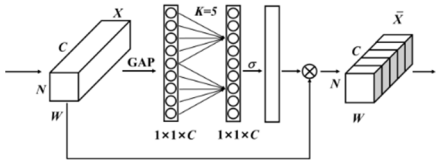

Figure 3: The structure of ECA-Net.

2.4 Improved Conv-TasNet

Compared with the general environmental noise, the airport environmental noise has more non-stationary characteristics and lower SNR. Although the TCN (Temporal Convolutional Network) used in the separator by the baseline Conv-TasNet also improves the spatial level of the CNN, it does not consider the correlation between speech channels, especially for the low SNR noisy speech. Therefore, an improved Conv-TasNet is proposed in this paper. It is based on the baseline Conv-TasNet, and fuses the multi-head-attention module and the ECA-Net channel attention module. The multi-head-attention module is placed after the LN in the bottleneck layer, and the ECA-Net is placed after the depth-wise convolution in the TCN. The network structure is shown in Figure 4.

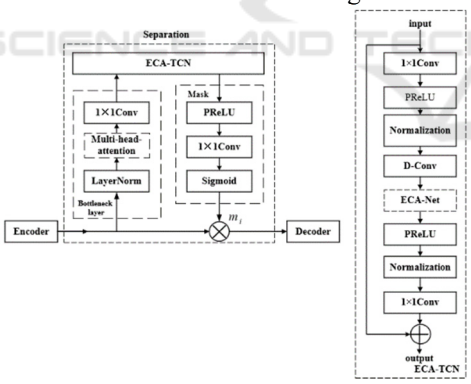

Figure 4: The structure of improved Conv-TasNet.

Multi-head-attention module can use efficient matrix operation, which improves the efficiency of parallel computing efficiency of the network. At the same time, the multi-head-attention module can calculate the similarity between the input speech signal feature of each frame and the adjacent frame. For the speech signal containing non-stationary environmental noise, it can more obviously distinguish the pure speech part and the noise part of the input feature. And the operation is based on independent frames at the same time, without flattening the input features, and avoid the problem of destroying the voice signal structure. ECA-Net channel attention module can make the network to fully consider the correlation between the channels. For low noise ratio of noise speech, its correlation between channels is complex, and the introduction of ECA-Net can further strengthen the importance of input information, improving the effective use of each information, strengthening the subsequent convolution layer of channel relationship mapping, and achieve the purpose of improving network speech enhancement performance.

3 EXPERIMENT AND RESULT ANALYSIS

3.1 Dataset Establishment

The noise sample set is partly derived from the domestic classic airport recording and the network public airport environment noise audio; the pure voice sample set is derived from the network public TIMIT voice library; and the noisy voice is derived from the open space real-time dialogue recording and additive synthesis. Among them, the additive synthesis of noisy speech refers to the additive superposition of the noise sample set and the pure speech sample set according to a certain SNR. Finally, 9530 noise-containing speech items with a sampling rate of 16kHz and a bit depth of 32bit are synthesized.

3.2 Comparison Method

Ablation experiments were designed to test the effectiveness of the baseline Conv-TasNet fused the modules. In this paper, the baseline Conv-TasNet is named BCTN, and baseline Conv-TasNet fused multi-head-attention modules is named MCTN, and baseline Conv-TasNet fused ECA-Net is named ECTN, and baseline Conv-TasNet fused both two modules is named MECTN. All models are trained in the same environment and the same sample set.

3.3 Performance Analysis

3.3.1 Baseline Conv-TasNet Performance

Figure 5 shows the denoising effect of five airport noise speech with different ambient noise ratio, namely the Mel spectrogram of pure speech, noisy speech and speech processed by four networks under each SNR.

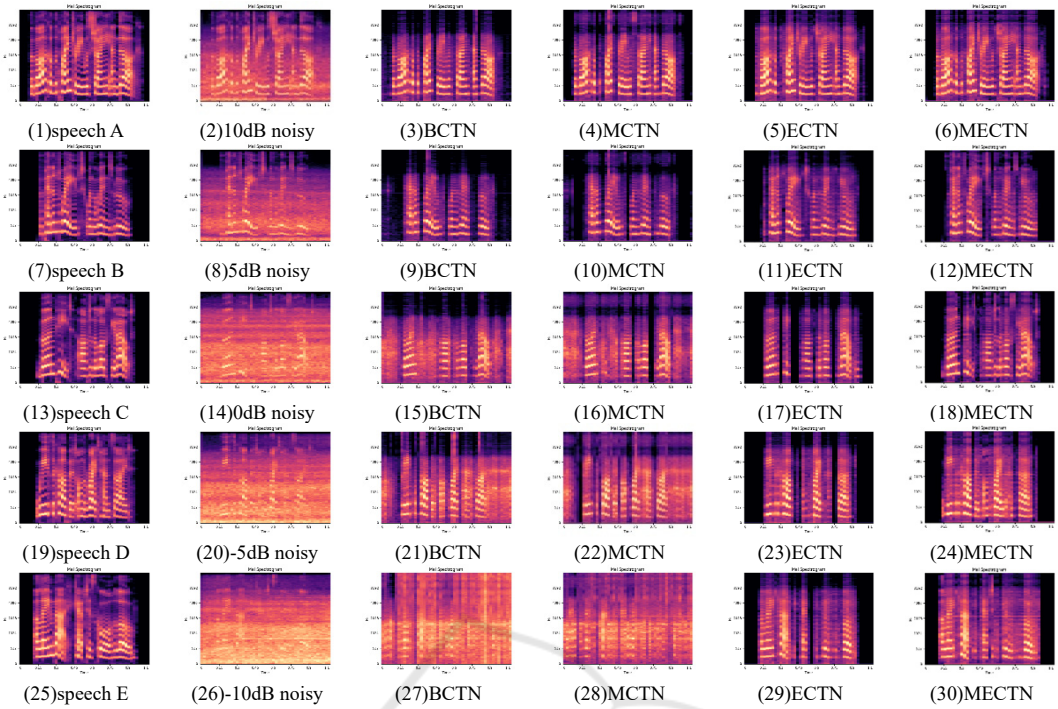

Figure 5: Speech enhancement effects of noisy speech processed by different models under different SNR.

As can be seen from Figure 5(3)(9)(15)(21)(27), after the baseline Conv-TasNet processing, it can basically restore the timing and energy intensity of pure speech. However, for the silent segment of the front and the last part of the pure voice signal, part of the environmental noise is still there. Moreover, with the decrease of SNR, the high-frequency harmonic loss is gradually serious, and the energy intensity is obviously low with the serious loss of speech signal. From figure 5(4)(10)(16)(22)(28), after the fusion of multiple attention module, the high frequency harmonic loss is significantly reduced, and the energy intensity of each frequency is closer to the pure voice, visible for the prediction of non-stationary noise is more accurate, but for the pure voice signal and the last expansion of silent segment, still retain part of environmental noise, and for the middle part of the voice signal has many obvious eliminations. As can be seen from figure 5(5)(11)(17)(23)(29), after integrating the ECA-Net module, the environmental noise at the expanded silent segment of the first and last parts of the pure speech signal is significantly eliminated, and the middle part of the voice signal is not significantly eliminated. Therefore, after the introduction of ECA-Net, the network can better restore the timing of pure speech, especially for the low energy intensity and low SNR. From figure 5(6)(12)(18)(24)(30), after the fusion of the multiple attention module and the ECA-Net channel attention module, the network combines the advantages of the two modules,not only the harmonic loss of the high-frequency part is significantly reduced, the energy intensity of each frequency is closer to the pure speech, but also the ambient noise at the first and last part of the extended silent segment of the pure speech signal and the noise in the noisy speech with low SNR are significantly eliminated. It can be seen that the network is more accurate for non-stationary noise prediction, and the ability of pure speech at low SNR is significantly improved.

3.3.2 Subjective and Objective Evaluation

In order to evaluate the network performance more comprehensively, this paper uses both subjective and objective evaluation. The subjective evaluation adopts MOS (Randhir Singh et al., 2017), the objective evaluation used segSNR (Rashmirekha Ram et al., 2019), PESQ (RIX ANTONY W et al., 2002) and STOI (TAAL C H et al., 2011). Four network models were used for 200 pieces of five different SNR speech with noise for speech enhancement, and average of the evaluation scores of each SNR was calculated, as shown in Table 1.

Table 1: Evaluation index of network model for speech enhancement with different SNR.

Evaluation	Model	SNR				
		10dB	5dB	0dB	-5dB	-10dB
MOS	Noisy	2.713	1.916	1.641	1.873	1.477
	BCTN	2.880	2.891	2.984	3.067	2.989
	MCTN	3.083	3.051	3.207	3.287	3.112
	ECTN	3.170	3.170	3.255	3.310	3.205
	MECTN	3.317	3.495	3.362	3.479	3.459
segSNR	Noisy	3.7957	-1.8659	-5.0278	-7.1134	-8.5055
	BCTN	8.7555	4.6158	3.2732	2.0801	0.3003
	MCTN	9.8841	6.3587	5.1664	4.8988	2.3299
	ECTN	10.3367	6.6742	5.7543	4.2594	2.0641
	MECTN	13.8277	8.4318	7.1765	6.8463	4.6311
PESQ	Noisy	1.4424	1.1470	1.0525	1.0367	1.0250
	BCTN	1.7582	1.5849	1.3874	1.2565	1.1545
	MCTN	1.8488	1.6761	1.4847	1.3999	1.2555
	ECTN	1.9628	1.7876	1.5669	1.4702	1.3510
	MECTN	2.1786	1.8548	1.7617	1.5846	1.4579
STOI	Noisy	0.9425	0.8696	0.7380	0.7429	0.5104
	BCTN	0.9595	0.8722	0.7561	0.7566	0.5256
	MCTN	0.9707	0.8910	0.7762	0.7783	0.5486
	ECTN	0.9796	0.9037	0.7797	0.7817	0.5465
	MECTN	0.9923	0.9177	0.750	0.7974	0.5735

It can be seen from the above table, after the baseline Conv-TasNet and the improved Conv-TasNet enhance the noise-containing speech, the speech is significantly improved in all the four evaluation indicators. At the five SNRs, increased the MOS by 1.0382, segSNR by 7.5484, PESQ by 0.2876, and STOI by 0.0133. Conv-TasNet only fused multi-head-attention module and only fused ECA-Net are both better than baseline Conv-TasNet for high SNR or low SNR speech, increased MOS by 17.90% and 25.02% respectively, segSNR by 1.923 and 2.013 respectively, PESQ by 17.63% and 25.50% respectively, STOI by 0.019dB and 0.024dB respectively. However, Improved Conv-TasNet, which simultaneously fused the multi-head-attention and ECA-Net, increased the MOS by 1.4984, segSNR by 11.9261,0. PESQ by 0.5868, and STOI by 0.0455. In conclusion, the improved Conv-TasNet speech enhancement algorithm proposed here is able to handle the low SNR ratio and non-stationary noise-containing speech more effectively than the baseline Conv-TasNet.

4 CONCLUSIONS

A speech enhancement method based on improved Conv-TasNet is proposed in this paper, it fused multi-head-attention module and ECA-Net network based on the baseline Conv-TasNet. The experimental results show that after the baseline network integrates the two modules, the prediction of non-stationary noise and the speech signal at low SNR have a higher completeness. Compared with the baseline Conv-TasNet, the proposed improved Conv-TasNet increases MOS by 1.4984, segSNR by 11.9261, PESQ by 0.5868 and STOI by 0.0455. Therefore, the improved Conv-TasNet has better speech quality and understanding, which can solve the problem of the traditional unsupervised speech enhancement algorithm on such low signal to noise ratio and non-stationary environment noise speech enhancement in airports.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the financial support of Research Fund of Civil Aviation Flight University of China (Grant No. J2019-88 and Grant No. ZJ2022-007).

REFERENCES

DING Cong, ZENG Wei-li, WEI Wen-bin, YANG Ai-wen. Review of Civil Airport Noise Assessment [J]. *Aeronautical Computing Technique*,2021,51(05):130-134.

E. Golrasan, H. Sameti. Speech enhancement based on hid den Markov model using sparse code shrinkage[J]. *Jou rnal of Artificial Intelligence and Data Mining*,2016,4(2).

G Thimmaraja Yadava,G Nagaraja B,S Jayanna H. A spatial procedure to spectral subtraction for speech enhancement[J]. *Multimedia Tools and Applications*,2022,81(17).

HE Liqing. Study of Artificial Intelligence Flight Co-Pilot Speech Recognition Technology[D]. *Civil Aviation Flight University of China*, 2020.DOI: 10.27722/d.cnki.gzgmh.2020.000026.

HU J, SHEN L, ALBANIE S, et al. Squeeze-and-excitation networks[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*,2020, 42(8): 2011-2023.

Jaiswal Rahul Kumar, Yeduri Sreenivasa Reddy, Cenkeramaddi Linga Reddy. Single-channel speech enhancement using implicit Wiener filter for high-quality speech communication[J]. *International Journal of Speech Technology*,2022,25(3).

Khattak Muhammad Irfan, Saleem Nasir, Gao Jiechao, Verdu Elena, Fuente Javier Parra. Regularized sparse features for noisy speech enhancement using deep neural networks[J]. *Computers and Electrical Engineering*,2022,100.

LU Yong. Research on speech enhancement algorithm based on VHF[J]. *Electronic Measurement Technology*,2021,44(02):65-70.

LUO Y, MESGARANI N. Conv-TasNet: surpassing ideal time- frequency magnitude masking for speech separation[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019, 27(8):1256-1266.

Randhir Singh, Ajay Kumar, Parveen Kumar Lehana. Effect of bandwidth modifications on the quality of speech imitated by Alexandrine and Indian Ringneck parrots[J]. *International Journal of Speech Technology*,2017,20(3):659-672.

Rashmirekha Ram, Mihir Narayan Mohanty. Use of radial basis function network with discrete wavelet transform for speech enhancement[J]. *Int. J. of Computational Vision and Robotics*,2019,9(2):207-223.

Ribas Dayana, Miguel Antonio, Ortega Alfonso, Lleida Eduardo. Wiener Filter and Deep Neural Networks: A Well-Balanced Pair for Speech Enhancement[J]. *Applied Sciences*,2022,12(18).

RIX ANTONY W, BEERENDS JOHN G , HOLLIER MICHAEL P,et al. Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs[C]//*Proceedings of the 2001 IEEE International Conference on Acoustics , Speech, and Signal Processing. Piscataway: IEEE Press*,2002: 749-752.

TAAL C H, HENDRIKS R C, HEUSDENS R, et al. An algorithm for intelligibility prediction of time–frequency weighted noisy speech[J].*IEEE Transactions on Audio Speech & Language Processing*, 2011,19(7): 2125-2136.

Tank Vanita Raj, Mahajan Shrinivas Padmakar. Adaptive recurrent nonnegative matrix factorization with phase compensation for Single-Channel speech enhancement[J]. *Multimedia Tools and Applications*,2022,81(20).

Wang Youming, Han Jiali, Zhang Tianqi, Qing Didi. Speech enhancement from fused features based on deep neural network and gated recurrent unit network[J]. *EURASIP Journal on Advances in Signal Processing*,2021,2021(1).

Y. A. Farha and J. Gall, "MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3570-3579, doi: 10.1109/CVPR.2019.00369.

YANG Lei, ZHAO Hongdong, YU Kuaiguai. End-to-end speech emotion recognition based on multi-head attention[J]. *Journal of Computer Applications*,2022,42(06):1869-1875.

YI Xue. VHF Speech Enhancement Based on Short-time Spectrum Estimation[D]. *Civil Aviation University of China*, 2020.000078.DOI: 10.27627/d.cnki.gzmhy.2020.000078.