2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Speech signal enhancement in ADOS recordings

N. Tsur, M. Eni and Y. Zigel, *Senior Member, IEEE*

***Abstract*— This work examines the utility of Speech Enhancement (SE) in estimating autism severity in children based on audio recordings of clinical sessions. We implemented a DNN-based algorithm using U-Net model to perform SE in 204 Autism Diagnostic Observation Schedule (ADOS) session recordings. This approach provided performance gain in terms of Signal-to-Noise Ratio (SNR) in the recordings and preserved a good performance in estimating autism severity with reduced overfitting.**

***Clinical Relevance*— This work shows that the use of SE method based on U-Net architecture has the potential to improve automatic estimation of autism severity.**

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is characterized, among others, by speech abnormalities, which potentially carry important diagnostic information. We currently developing a system for ASD severity estimation in children. This system is based on the analysis of audio recordings of clinical assessments using the Autism Diagnostic Observation Schedule (ADOS) [1]. The ADOS recording room includes noisy devices such as a server computer and an air conditioner. The noisy clinical environment may affect the characteristics of the speech signal and can pose a challenge to the system. Therefore, we utilized Speech Enhancement (SE) algorithm to reduce the impact of background noises on the system and examine the utility of SE on the system's performance.

II. METHODS

In this work, we implemented a DNN-based SE algorithm using U-Net architecture [2], a Convolutional Neural Network (CNN) that as well applied to audio source separation [3]. We used recordings of 226 ADOS sessions that were performed at the National Autism Research Center of Israel, Soroka University Medical Center (SUMC). The sessions are ~40 minutes long and recorded at a sampling rate of 44.1 kHz (downsampled to 16 kHz) with a single microphone (CHM99, AKG, Vienna) which was located 1-2 meters from the child and connected to a sound card (US-16x08, TASCAM, California). The U-Net model was first trained on 10 hours training set of clean speech signals from LibriSpeech [4] dataset with White Gaussian Noise (WGN). Following this, we used the trained model and additional 2.5 hours of clean speech signals with background noises from 22 ADOS session recordings to perform transfer learning.

* This research was supported by the Israeli Ministry of Science and Technology, Israel, grant no. 3-17422.

N. Tsur (e-mail: netatsu@post.bgu.ac.il), M. Eni, Y. Zigel are with the Negev Autism Center and Department of Biomedical Engineering, Ben-Gurion University of the Negev, Beer–Sheva, Israel.

III. RESULTS

We analyzed the SNR in the recordings before and after performing SE. The results were obtained from the remaining 204 ADOS recordings and are shown in Fig 1.

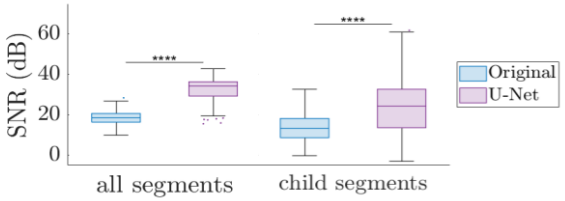

Figure 1. SNR in all speech segments (left) and child segments (right) in the recording before and after performing SE using the U-Net model.

After extracting 48 features from 140 recordings, in which there are at least 50 child's speech segments, we trained and tested our ASD severity estimation system detailed in [1] and analyzed the RMSE and correlation of the train and test set. The results are shown in Table I.

TABLE I. RMSE AND CORRELATION IN TRAIN AND TEST SET

Recording	Set	RMSE	R
Original	Train	4.975 ± 0.180 (4.975)	0.804 ± 0.017 (0.804)
	Test	6.182 ± 0.679 (6.279)	0.640 ± 0.096 (0.638)
U-Net	Train	5.504 ± 0.368 (5.448)	0.772 ± 0.020 (0.772)
	Test	6.461 ± 0.658 (6.475)	0.612 ± 0.098 (0.611)

IV. DISCUSSION & CONCLUSION

The results show that the U-Net model provides significantly improved SNR along with a good performance of the ASD severity estimation system. In addition, the results indicate that SE using U-Net model produces less overfitting in the system, as the difference of RMSE score in train and test sets reduced after performing SE (p-value<0.0001). We suggest that this method for SE may increase signal quality, especially in distant microphone recordings of ADOS sessions, reduce the effect of background noises, and can lead to improvement in autism severity estimation.

REFERENCES

[1] M. Eni, I. Dinstein, M. Ilan, I. Menashe, G. Meiri, and Y. Zigel, "Estimating Autism Severity in Young Children from Speech Signals Using a Deep Neural Network," *IEEE Access*, 2020.

[2] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2015.

[3] E. M. Grais and M. D. Plumbley, "Single channel audio source separation using convolutional denoising autoencoders," in *Proc. Glob. Conf. Signal, Inform. Processing (GlobalSIP),* 2017.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. Int. Conf. Acoust. Speech Signal Process (ICASSP),* 2015.