MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231129

Towards an Environmentally Robust Speech Assistant System for Emergency Medical Services

Zhenchuan ZHANG[a], Yu TIAN[b], Tianshu ZHOU[a], Yinghao ZHAO[a], Jungen ZHANG[c] and Jingsong LI[a,b,1]
[a]*Research Center for Healthcare Data Science, Zhejiang Lab, Hangzhou, China*
[b]*Engineering Research Center of EMR and Intelligent Expert System, Ministry of Education, College of Biomedical Engineering and Instrument Science, Zhejiang University, Hangzhou, China*
[c]*Hangzhou Emergency Medical Center of Zhejiang Province, China*

Abstract. Automated speech recognition technology with robust performance in various environments is highly needed by emergency clinicians, but there are few successful cases. One main challenge is the wide variety of environmental interference involved during a typical prehospital care emergency service such as background noises and overlapping speech. To solve this problem, we try to establish an environmentally robust speech assistant system with the help of the proposed personalized speech enhancement (PSE) method, which utilizes the target physician's voiceprint feature to suppress non-target signal components. We demonstrate its potential value using both general public test set and our real EMS test set by evaluating the objective speech quality metrics, DNSMOS, and the recognition accuracy. Hopefully, the proposed method will raise EMS efficiency and security against non-target speech.

Keywords. Prehospital care, emergence medical service, speech recognition, speech enhancement, personalized speech enhancement

1. Introduction

Prehospital care is an essential part of the health care system for every country. Upon an emergency call, the emergency medical center dispatches ambulance to the emergency site, and the emergency physician needs to triage, treat, and transport the patient(s) to the appropriate health care facility. Aside from these duties, emergency physicians are also generally required to do some other complex works such as documenting patient records and communicating with hospital emergency department. But physicians often won't be able to work on other things besides treatment. Therefore, a smart speech assistant is very much needed in emergency medical services. But until now there are few successful cases. On one hand, clinical speech recognition itself is quite challenging due to the nuanced terminology and speech patterns. On the other hand, prehospital care could happen in any places, which involve wide variety of acoustic scenarios and many of which are quite noisy. In [1], they focus on the first problem and propose an adaptive clinical transcription system for in-situ transcribing of patient encounter information for hospital emergency department. Unlike their work, prehospital care is our interest and

[1] Corresponding Author: Jingsong Li, email: ljs@zju.edu.cn.

we mainly focus on the second problem which are often eased with the help of speech enhancement (SE) techniques. But unlike other scenarios like meeting or daily conversation, in prehospital care scenario we usually only care about what the emergency physician says, and don't want other speech to interfere. To this end, a personalized speech enhancement (PSE) method capable of suppressing both the general background noise and the interfering speech is needed.

Unlike traditional SE methods, PSE approaches emerged in just few years ago and some pioneer works have been proposed like [2-4]. Generally, a PSE method consists of two major components: a speaker encoder and a speaker extractor. The speaker encoder extracts speaker embedding for some auxiliary speaker speech data, such as pre-recorded speaker enrollment utterances. Then this speaker embedding is used by the speaker extractor to suppress non-target signal components and retain the target-specific signal components for a query audio. Most PSE works rely on a pretrained speaker encoder, while some recent works [5,6] try to optimize the speaker encoder together with speaker extractor. Most speaker extractors work in T-F domain and some work in time domain. Voicefilter based methods [2,7] and SpEx based methods [4,5] are among the most influential studies.

In this work, we will report our proposed PSE model, P-Denoiser, and test its performance on some public test sets. Previously we have developed an android mobile app for Hangzhou Emergency Medical Center of Zhejiang Province to help their prehospital care services which includes a smart assistant module. We have rewritten the server-side to incorporate the P-Denoiser model and we collected a real EMS scenario test set to evaluate the model performance.

2. Methods

Our smart speech assistant for EMS consists of two parts: a mobile client app running on android devices such as tablets or smartphones, and a speech server backend. The mobile app keeps recording and streaming audio data to the server via WebSockets. The server-side detects voiced parts in the stream with the help of Silero VAD [8] which are then enhanced and transcribed, and the results are then sent back to the mobile client where corresponding actions will be taken. In this paper, we mainly focus on the speech enhancement (SE) module, since a versatile SE module is the key to environmental robustness. We propose P-Denoiser as our SE model which consists of two parts: a speaker encoder and a speaker extractor as shown in following sections.

2.1. Speaker Encoder

ECAPA-TDNN [9] is the state-of-the-art method for speaker recognition and therefore we use it as our speaker encoder network to extract speaker embeddings. The speaker encoder is first separately trained on speaker identification datasets, such as VoxCeleb2, and is frozen in all following steps. Our implementation is based on SpeechBrain, and 192-dim is chosen as the embedding size. For more details please refer to [9].

2.2. Speaker Extractor

Our proposed speaker extractor is based on Denoiser, which consists of a speech encoder, a BLSTM speech enhancer, and a speech decoder. The BLSTM of Denoiser is replaced with a cross-attention conformer network like [10]. The network architecture is shown in the left side of Figure 1, and the cross-attention conformer block is given in the right side of Figure 1. A cross-attention conformer block is made up by 1 ×1 feed forward layer (FFN), temporal convolutional layer (Conv), multi-head cross-attention layer (MHCA), Feature-wise Linear Modulation layer (FiLM) [11] and layer normalization. It accepts the noisy representation $\hat{a}$ from the encoder network or the previous conformer block as the main input, and a speaker embedding e as side input, and outputs an enhanced representation.

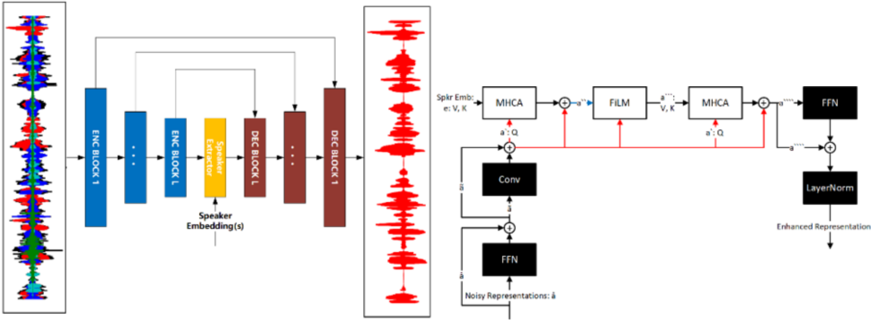

Figure 1. Speaker extractor architecture.

In EMS scenario, the interfering speakers are usually fixed, i.e., the stretcher-bearers, thus it should be beneficial to provide the interfering speaker embeddings also. Suppose the target speaker embedding is e_{tar} and the interfering speaker embedding is e_{itf}, we concatenate them to form e. Note that e_{itf} is optional. Zero vector is used as e_{itf} when interfering speaker information is not available. The speaker embedding is repeated along the temporal dimension to match the dimension of the output of speech encoder network. The conformer network consists of several stacked conformer block. In our work, we use two conformer blocks each having 8 heads. The 1d temporal convolution in conformer is of 31 kernel size. And the FFN layer has the same output channels with encoder input. For more details on the conformer, please refer to [12]

3. Results

In this paper, 16kHz is used as the sampling rate for all steps. To evaluate the effect of incorporating a PSE module, we compare the objective speech quality metric and the word error rate (WER) between raw noisy audio and the enhanced audio. We use DNSMOS [13] as the objective speech quality metric, since it’s recently proposed and better mimic the subjective metric Mean Opinion Score (MOS) compared to traditional objective metrics such as PESQ. It has three sub metric： SIG (speech quality), BAK (background noise quality) and OVRL (overall audio quality). All these metrics are on scale 1 to 5, the higher the better quality is.

The speaker extractor, P-Denoiser, is trained with the DNS4 training set along with a mandarin data set containing 40 speakers (20 males and 20 females) which has a total duration of about 20 hours. The DNS4 training set is clustered and merged since some of the speakers are actually the same person. The training data generation follows a similar approach to [14]. The hidden size of P-Denoiser is set to 64.

The test set of DNS4 development is used as the general-purposed test set. In collaboration with Hangzhou Emergency Medical Center, we gathered an EMS scenario test set consisting of 12 recordings during real prehospital care services, each last about 20 minutes to 1hours. In total, there are about 6.5 hours.

The results are shown in Table 1. The left half is the results on the general purposed test set DNS4 dev test set. The right half is the results on our EMS test set.

Table 1. DNSMOS and WER comparisons for raw noisy audio and PSE enhanced audio. These values are mean average over the entire test set.

	DNS4 dev testset			EMS test set			
	SIG	BAK	OVAL	SIG	BAK	OVAL	WER
Noisy	**3.81**	2.23	2.42	**3.89**	2.32	2.65	22.5%
Voicefilter	3.76	3.12	**3.28**	3.59	3.97	3.28	17.4%
Our Work	3.42	**3.33**	2.80	3.44	**3.99**	**3.35**	**14.2%**

4. Discussion

For both the general purposed test set and the EMS test set, the objective metrics follow the same pattern. The noisy speech has a better SIG metric, while the enhanced speech has better BAK and OVAL metrics. For DNS4 dev test set, our method has a 49.32% relative improvement on BAK, and a 15.70% relative improvement on OVAL. For EMS test set, our method has a 71.98% relative improvement on BAK, and a 26.42% relative improvement on OVAL, and the WER drops by 38.89%. For comparison, another PSE method, Voicefilter, was also tested. It better suits the DNS4 dev test set while our method is better for the EMS scenario.

We thus conclude that (1) a PSE module yields a large improvement on the word recognition accuracy in EMS scenario; (2) although the objective speech quality metric declines after the PSE process compared to the noisy original raw audio, the background noise and the overall speech quality could get great improvements; (3) our model better suits the EMS scenario than Voicefilter.

At the same time, we are not quite there yet. Firstly, it could be seen that even with the help of a PSE module, the WER is still not acceptable for practical use in prehospital care scenario. Secondly, our speech services work on an AMD64 Linux server. The mobile client app needs to first record and send audio data to server and wait for results. Generally, this process takes about less than 2s in our system when the network status is well. But still, this delay is unacceptable for some EMS cases. We will try to embed the whole speech services into mobile ends.

5. Conclusions

Our work pursues the ultimate goal of an environmentally robust speech assistant system for prehospital care services that works in real-time and that can be used in various noisy EMS scenarios.

The work presented in this paper contributes to this goal by (a) proposing a simple-structured open PSE model that can suppress both general background noises and unwanted interfering speech (b) incorporating the proposed PSE model to an EMS app and evaluating its performance with real EMS data. The results demonstrate the potential application value of PSE in EMS scenario, which could lead to lives to be saved timely.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 82172069), the Key Research Project of Zhejiang Lab (No.2022ND0AC01), the Key Program of Science and Technology Development Program of Hangzhou, China (No. 2020ZDSJ0885).

References

[1] Van Woensel W, Taylor B, Abidi SSR, editors. Towards an Adaptive clinical transcription system for in-situ transcribing of patient encounter information. Stud Health Technol Inform. 2022 Jun, doi: 10.3233/SHTI220052.

[2] Wang Q, Muckenhirn H, Wilson K, Sridhar P, Wu Z, Hershey JR, Saurous RA, Weiss RJ, Jia Y, Moreno IL, editors. VoiceFilter: targeted voice separation by speaker-conditioned spectrogram masking. Interspeech 2019; 2019, doi: 10.48550/arXiv.1810.04826.

[3] Žmolíková K, Delcroix M, Kinoshita K, Ochiai T, Nakatani T, Burget L, Černocký J. Speakerbeam: speaker aware neural network for target speaker extraction in speech mixtures. IEEE J Sel Top Signal Process. 2019 Jun;13(4):800-14, doi: 10.1109/jstsp.2019.2922820.

[4] Xu C, Rao W, Chng ES, Li H. Spex: multi-scale time domain speaker extraction network. IEEE/ACM Trans Audio, Speech, Language Process. 2020 Apr;28:1370-84, doi: 10.1109/taslp.2020.2987429.

[5] Ge M, Xu C, Wang L, Chng ES, Dang J, Li H, editors. SpEx+: a complete time domain speaker extraction network. Interspeech 2020; 2020, doi: 10.48550/arXiv.2005.04686.

[6] Ge M, Xu C, Wang L, Chng ES, Dang J, Li H, editors. Multi-stage speaker extraction with utterance and frame-level reference signals. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); Toronto, Canada: IEEE; 2021. p. 6109-13, doi: 10.1109/ICASSP39728.2021.9413359.

[7] Wang Q, Moreno IL, Saglam M, Wilson K, Chiao A, Liu R, He Y, Li W, Pelecanos J, Nika M, Gruenstein A, editors. VoiceFilter-Lite: streaming targeted voice separation for on-device speech recognition. interspeech 2020; 2020, doi: 10.48550/arXiv.2009.04323.

[8] Team S. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier GitHub repository: GitHub; 2021. Available from: https://github.com/snakers4/silero-vad.

[9] Desplanques B, Thienpondt J, Demuynck K, editors. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. Interspeech 2020; 2020, doi: 10.48550/arXiv.2005.07143.

[10] Narayanan A, Chiu CC, Malley TO, Wang Q, He Y, editors. Cross-attention conformer for context modeling in speech enhancement for ASR. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); Cartagena, Colombia: IEEE; 2021, p. 312-9, doi: 10.1109/ASRU51503.2021.9688173.

[11] Perez E, Strub F, De Vries H, Dumoulin V, Courville A. Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2018 Apr 29; 32(1), doi: 10.1609/aaai.v32i1.11671.

[12] Malley TO, Narayanan A, Wang Q, Park A, Walker J, Howard N, editors. A conformer-based ASR frontend for joint acoustic echo cancellation, speech enhancement and speech separation. In: 2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU); 2021 Dec 13-17; Cartagena, Colombia: IEEE; 2021, p. 304-311, doi: 10.1109/ASRU51503.2021.9687942.

[13] Reddy CKA, Gopal V, Cutler R, editors. Dnsmos P.835: a non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23-27; Singapore: IEEE; 2022, p. 886-90, doi: 10.1109/ICASSP43922.2022.9746108.

[14] Ju Y, Rao W, Yan X, Fu Y, Lv S, Cheng L, Wang Y, Xie L, Shang S, editors. TEA-PSE: tencent-ethereal-audio-lab personalized speech enhancement system for ICASSP 2022 DNS challenge. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23-27; Singapore: IEEE; 2022, p. 9291-9295, doi: 10.1109/ICASSP43922.2022.9747765.