RescueSpeech: A German Corpus for Speech Recognition in Search and Rescue Domain

Sangeet Sagar [1,3], Mirco Ravanelli [2], Bernd Kiefer [1,3],
Ivana KruijffKorbayová [3], Josef van Genabith [1,3]

[1]Saarland University, Germany [2]Concordia University, Mila-Quebec AI Institute, Canada
[3]German Research Center for Artificial Intelligence (DFKI), Germany

Introduction

- **ASR crucial in search and rescue missions**: vital for quick, accurate decision-making in hostile conditions.
- **Challenges for ASR in SAR**: fast, emotional, and stressful speech, extreme noise, unpredictable disturbances.
- **Limited data availability**: privacy restrictions, difficulty in collecting SAR-specific speech data.
- **Our contribution**: released **RescueSpeech**, a 2-hour annotated speech dataset from the German SAR domain, marking the first public release in this domain.
- **Experimental focus**: noise-robust German speech recognition, combining speech enhancement methods.

The RescueSpeech Dataset

- Consists of mixture of microphone and radio-recorded speech from simulated SAR exercises involving robot-assisted emergency response teams.
- Recorded by native German speakers in radio-style dialogues among team members, radio operators, and team leaders.
- Applications extend beyond robot control such as using speech recognition to support decision-makers and process monitors in disaster situations.

RescueSpeech clean set

The recordings, initially captured at a 44.1 kHz sampling rate, underwent down-sampling to 16 kHz. Following this, segmentation is performed to extract mono-speaker single-channel audio recordings, all of which are manually transcribed.

RescueSpeech noisy set

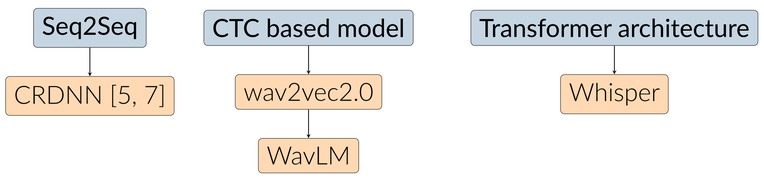
- Noises are injected into clean utterances to address low-noise profile of clean RescueSpeech. This helps mimic SAR condition.
- Noise types– emergency vehicle siren, breathing, engine, chopper, static radio noise.
- **Limitation**: Real-world noisy data often involves complex non-linear relationships between noise and speech, not easily replicated by artificially adding noise to clean speech.

Table 1. Distribution of utterances and hours in the RescueSpeech clean and noisy dataset.

	Clean		Noisy	
	Mins	#Utts.	HRS	#Utts.
Train	61.86	1591	7.20	4500
Valid	9.61	245	2.16	1350
Test	24.68	576	2.16	1350

Pre-training

ASR training



Speech enhancement training: SepFormer

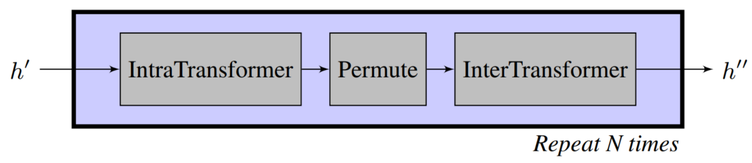


Figure 1. SepFormer block diagram that combines IntraTransformer and InterTransformer to model short-term and long-term dependencies.

Fine-tuning strategies

1. Clean training

Pretrain ASR and language models, then fine-tune on RescueSpeech clean dataset for domain adaptation.

2. Multi-condition training

Use pretrained model for training on an equal mix of clean and noisy audio from RescueSpeech dataset.

3. Model-combination I: Independent training

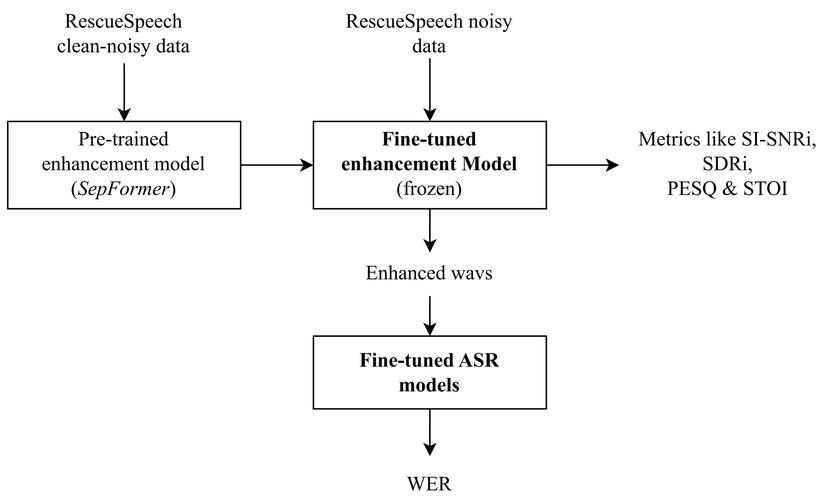


Figure 2. Training schema for independent model training strategy.

4. Model-combination II: Joint training

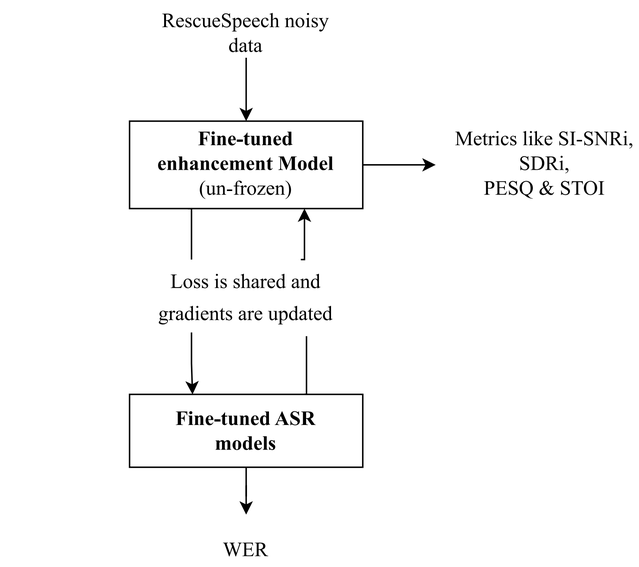


Figure 3. Training schema for independent model training strategy.

Results

ASR Performance

Table 2. Comparison of test WERs using different training strategies on clean and noisy speech inputs from the RescueSpeech dataset.

	ASR Model	clean	noisy
Pre-training	CRDNN	52.03	81.14
	Wav2vec2	47.92	76.98
	WavLM	46.28	73.84
	Whisper	27.01	50.85
Clean training	CRDNN	31.18	60.10
	Wav2vec2	27.69	62.60
	WavLM	23.93	58.28
	Whisper	**23.14**	46.70
Multi-cond. training	CRDNN	33.22	58.95
	Wav2vec2	29.89	57.98
	WavLM	25.22	52.75
	Whisper	24.11	**45.84**

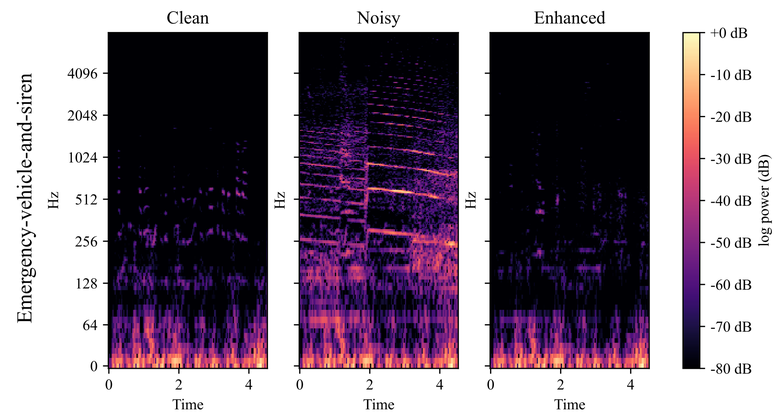


Figure 4. Log-power spectrogram of clean, noisy, and SepFormer-enhanced utterances for *emergency vehicle siren* noise type at -5 dB SNR.

Results

Combining ASR and Speech Enhancement

Table 3. Speech enhancement performance on the RescueSpeech noisy test inputs.

Metric	Model Comb. I	Model Comb. II			
		CRDNN	wav2vec2	WavLM	Whisper
SI-SNRi	6.516	6.618	7.205	7.140	7.482
SDRi	7.439	7.490	7.765	7.694	8.011
PESQ	2.008	2.010	2.060	2.064	2.083
STOI	0.842	0.844	0.854	0.854	0.859

Table 4. Word-Error-Rate (WER%) achieved with independent training (Model Comb. I) and joint training (Model Comb. II).

ASR Model	Model Comb. I	Model Comb. II
CRDNN	54.98	54.55
Wav2vec2	50.68	49.24
WavLM	48.24	46.04
Whisper	48.04	**45.29**

Conclusion

- Addressing challenges in SAR domain: limited speech data, SAR noise robustness, and conversational speech.
- Introduced RescueSpeech– search and rescue domain audio dataset.
- Despite leveraging advanced models, the best performance achieved WER of only 45.29%.

SpeechBrain

(a) Dataset (b) Source code

References

[1] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, 2020.

[2] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, oct 2022.

[3] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio Set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision, 2022.

[5] Tara N. Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. Convolutional, Long Short-Term Memory, fully connected Deep Neural Networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4580–4584, 2015.

[6] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is All You Need in Speech Separation, 2020.

[7] Yusheng Xiang, Tian Tang, Tianqing Su, Christine Brach, Libo Liu, Samuel S. Mao, and Marcus Geimer. Fast CRDNN: Towards on Site Training of Mobile Construction Machines. *IEEE Access*, 9:124253–124267, 2021.