UNIVERSITÄT DES SAARLANDES

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Noise Robust Speech Recognition for Search and Rescue Domain

Masters Thesis Colloquium, 5th April’23

Sangeet Sagar | M.Sc. Language Science & Technology

Supervisors

Prof. Dr. Josef van Genabith *(UdS, DFKI)*

Dipl. Inf. Bernd Kiefer *(DFKI)*

Advisor

Prof. Dr. Mirco Ravanelli *(Concordia Uni., Mila-Quebec AI Institute)*

Contents

Introduction

Introduction

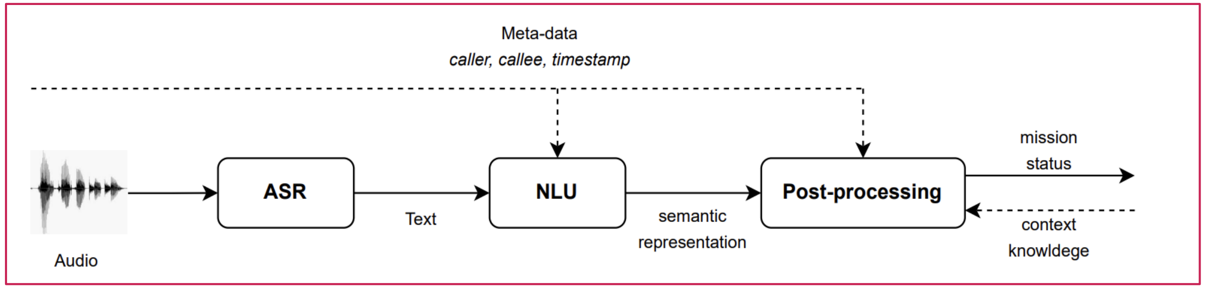
- Technology: keys to touch → touch to voice
- ASR- translation of spoken utterances
- Challenges
  - low resource languages
  - accent, dialect differences
  - domain mismatch
  - noisy surrounding

- This work focuses on noise-robust ASR for the German language in **Search and Rescue Domain**.

Motivaton & Contribution

Motivation

- Dedicated to "**A-DRZ**: Setting up the German Rescue Robotics Center" project
- **Objective**: Efficient disaster response with situational-aware robots.
- **Need**: High-risk scenarios exceed human capacity, require robot aid.
- **Solution**: Power robots with spoken language understanding (SLU).



A-DRZ: speech processing component [1]

Contribution

1. Lack of SAR speech data

2. Robustness to SAR noises
    - Multi-condition training approach & Speech enhancement module integration

- Release of ***RescueSpeech*** dataset
    - First publicly released audio dataset in the SAR domain
    - ~2 hours of annotated speech material

UNIVERSITÄT DES SAARLANDES

Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

Literature Survey
&
SpeechBrain toolkit

Literature Survey (ASR)

1952 "Audrey" → **1980s** Statistical modeling (HMMs) → **1990s** ASR with ANNs → **2000s** HMM-GMM

↓

2012+ HMM-DNN

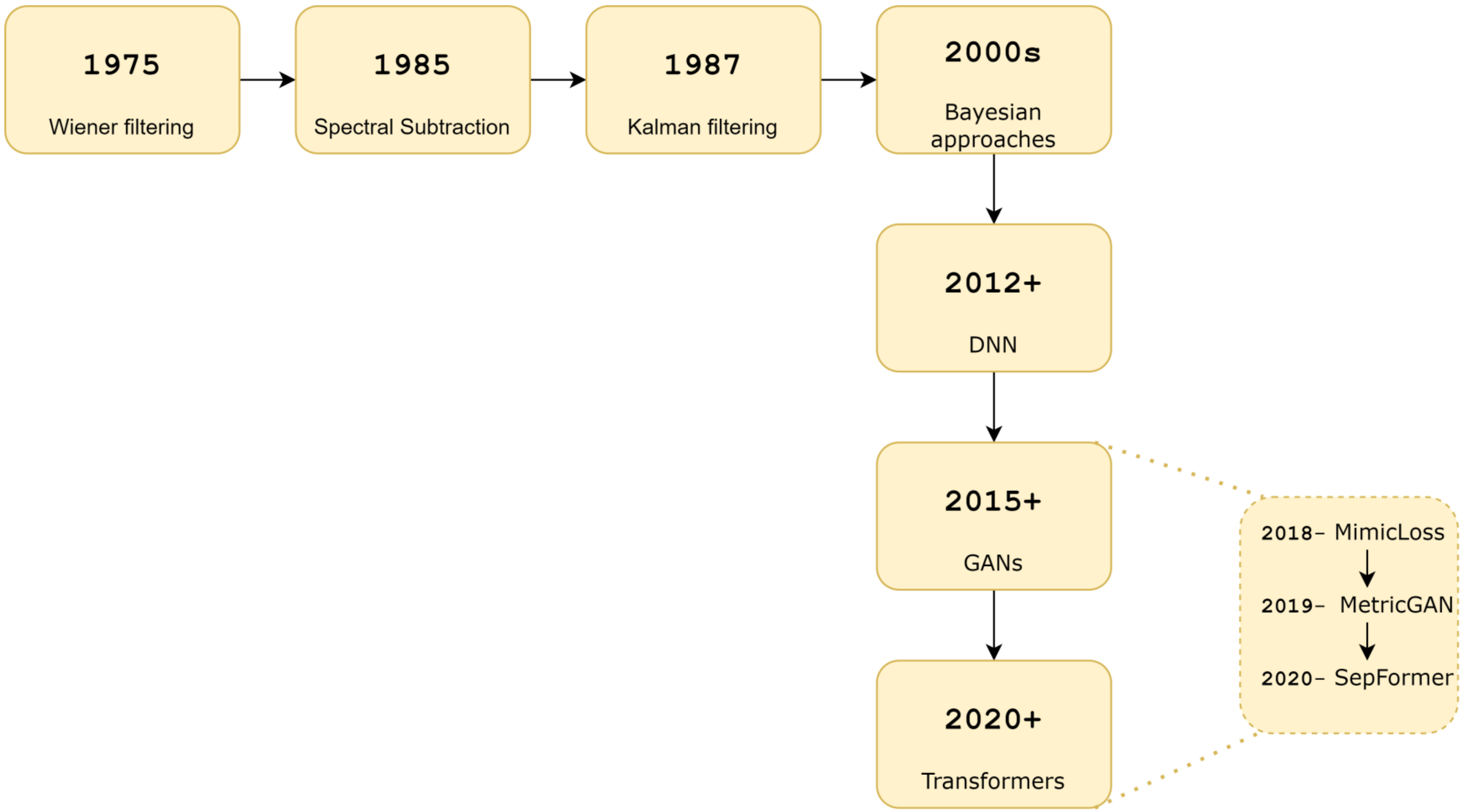
- 2012– RNN-Trandsucer
- ↓
- 2015– Attn. based Encoder-Decoder
- ↓
- 2016– Connectionist Temporal Classification (CTC)

↓

2020+ End2End DNN

- 2015– CRDNN
- ↓
- 2020– Wav2vec, Wav2vec2.0
- ↓
- 2021– WavLM
- ↓
- 2022– OpenAI/Whisper

Literature Survey (Speech Enhc.)



SpeechBrain

- Open-source **conversational AI** toolkit based on PyTorch
- Flexible, replicable, and easy-to-use with well-documented features.
- Offers unique and flexible data-loading techniques- JSON & CSV
- Ease of convenience: `python train.py hparams.yaml`
- *All models trained and results evaluated are contributed to* **SpeechBrain.**

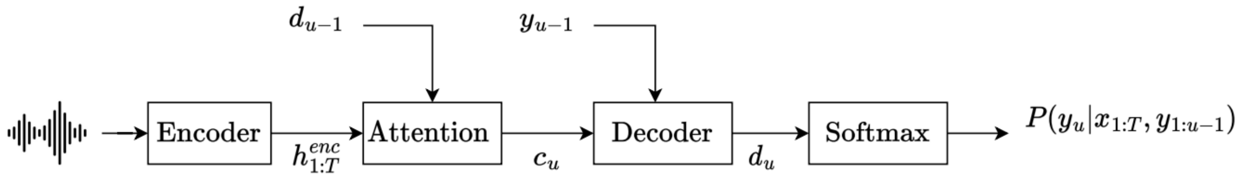
Technical Background

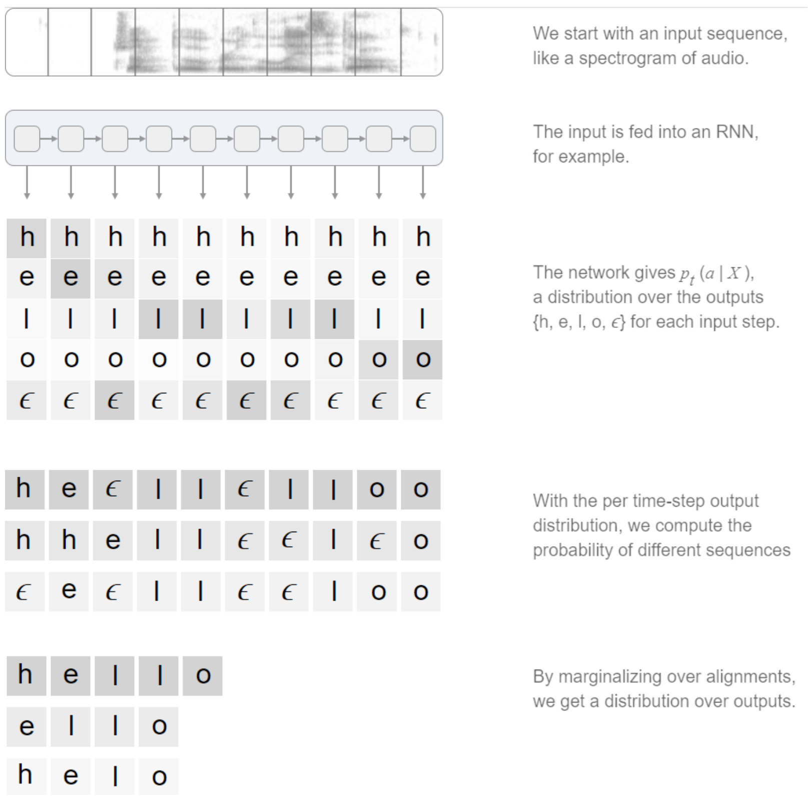
Technical Background (ASR)

End2end Models

1. Connectionist Temporal Classification (**CTC**)
   - aligns input speech frames with text transcriptions

2. Attention-based Encoder-Decoder model

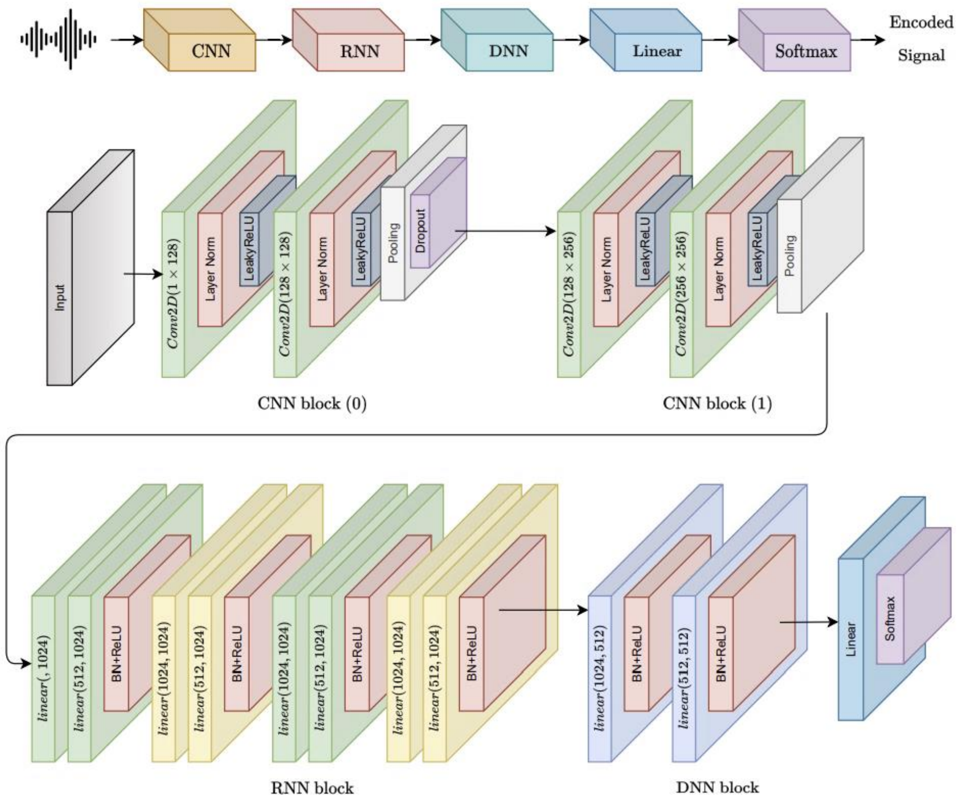




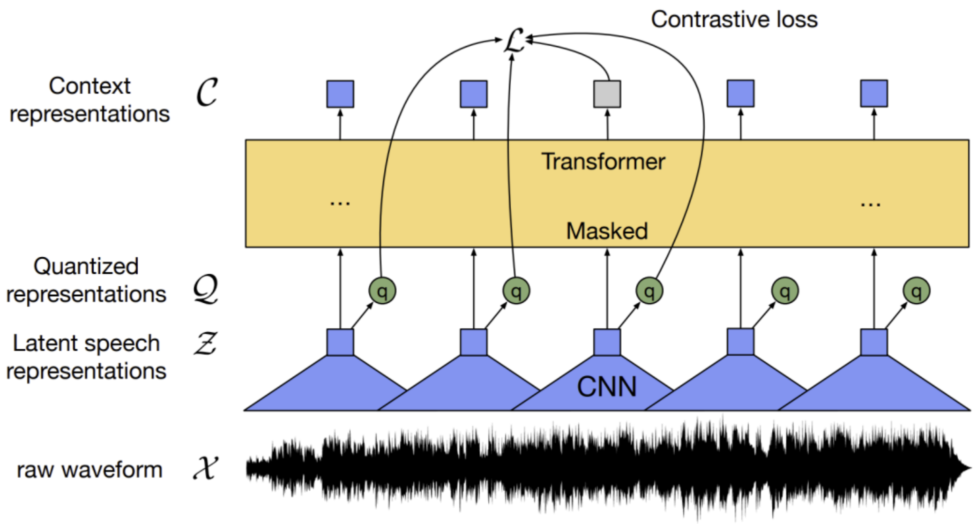
Technical Background (ASR)

End2end Architectures

- **CRDNN**

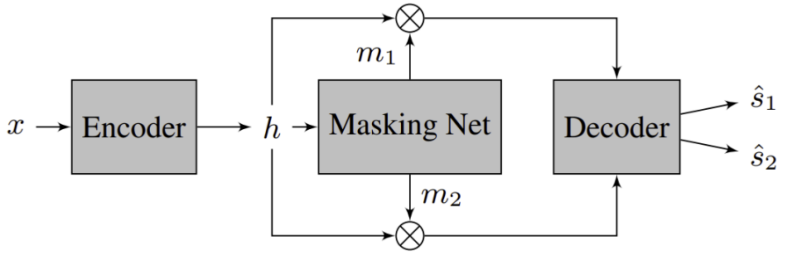


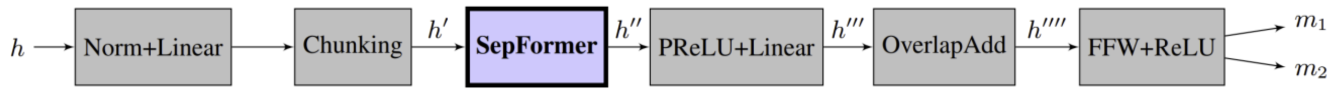
- **Wav2Vec2.0**



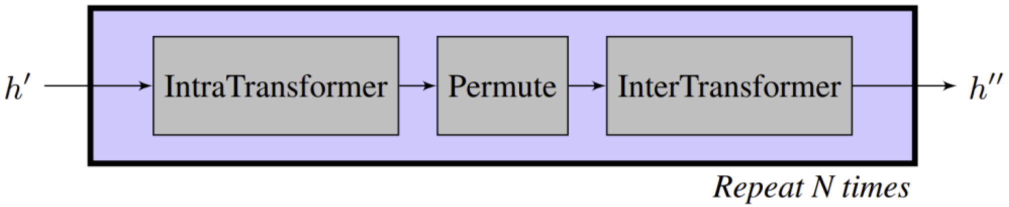
Technical Background (Speech Enhc.)

- ***SepFormer***: Transformer-based neural network
  - Encoder
  - Masking network
  - Decoder





Masking network



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

Evaluation Metrics

- **Speech recognition**
  - $WER = \frac{I+S+D}{N}$
  - where,
    - I : # insertion
    - S: # substitution
    - D: # deletions
    - N: # words in ref. text

- **Speech enhancement**
  - Intrusive
    - SI-SNRi (scale-invariant signal-to-noise ratio)
    - SI-SDRi (scale-invariant signal to distortion ratio)
    - PESQ (Perceptual Evaluation of Speech Quality) [-0.5 to 4.5]
    - STOI (Short-time objective intelligibility) [0-1]
  - Non-intrusive
    - DNSMOS (Deep noise suppression- Mean Opinion Score)
      - SIG (speech quality)
      - BAK (background noise quality)
      - OVRL (overall quality)

      1 - 5

Dataset

Dataset Description

Pre-training

- Purpose
  - Pre-train ASR and enhancement models
- CRDNN, Wav2vec2.0, WavLM
  - CommonVoice (DE)
- SepFormer
  - DNS4
  - 150 noise types (-5, 15 dB)

	CommonVoice10.0		**DNS4**	
	HRS	#Utts.	HRS	#Utts.
Train	739.17	466189	1317	1186019
Valid	26.97	16067	6.67	5965
Test	27.15	16067	5.17	921

Dataset Description

The **RescueSpeech** Dataset

- Purpose
  - Fine-tune ASR and enhancement models
- Simulated SAR exercises
- Noise types:
  - emergency vehicle siren
  - breathing
  - engine
  - chopper
  - static radio noise

	Clean		Noisy	
	HRS	#Utts.	HRS	#Utts.
Train	1.02	1543	4.84	3000
Valid	0.26	387	1.43	900
Test	0.32	484	1.40	900

Experimental Setup

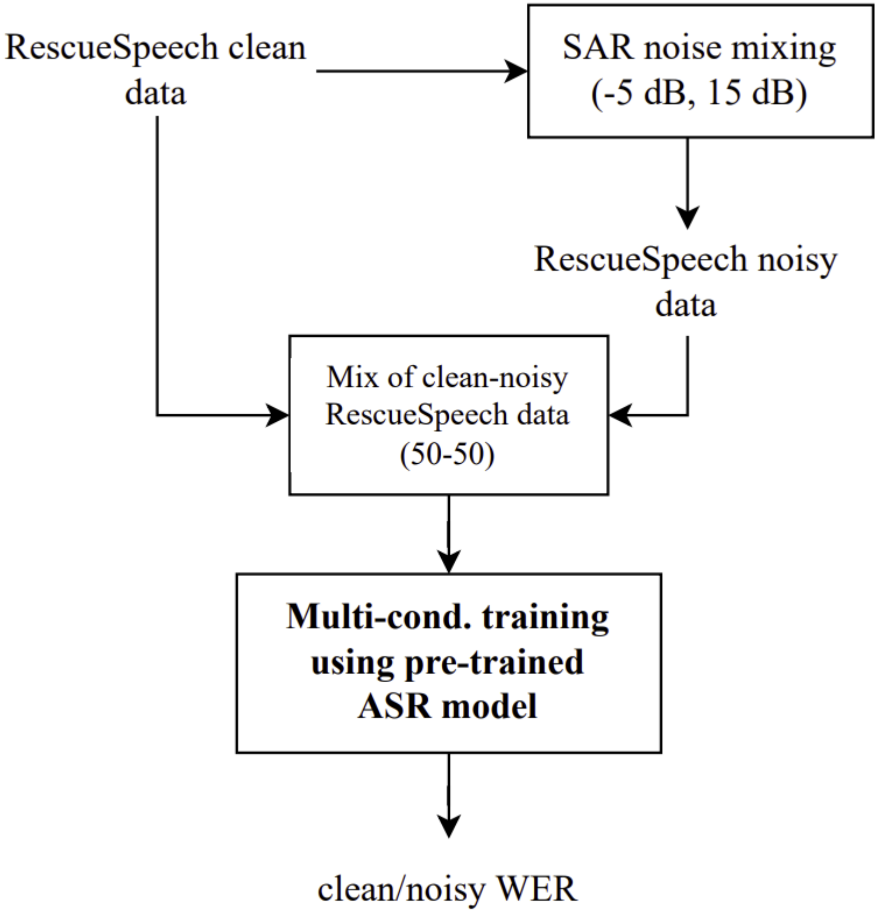
Experimental Setup

1. ASR training
   a. CRDNN seq2seq with beam-search + LM
   b. Wav2vec2.0 CTC with greedy decoder
   c. WavLM CTC with greedy decoder
   d. Whisper *(no pre-training needed)*

2. Speech enhancement training
   a. SepFormer

Experimental Setup

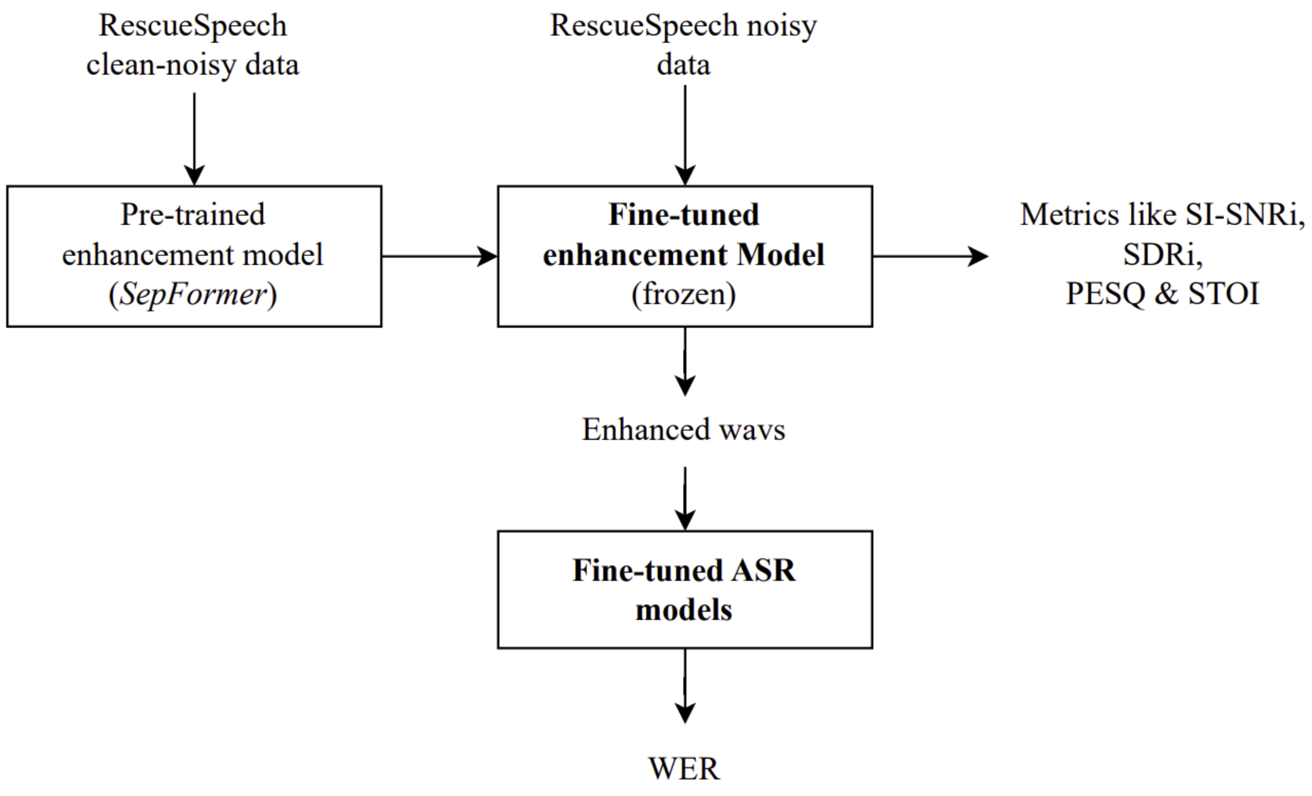
3. Training strategies
   a) Multi-condition training



Experimental Setup

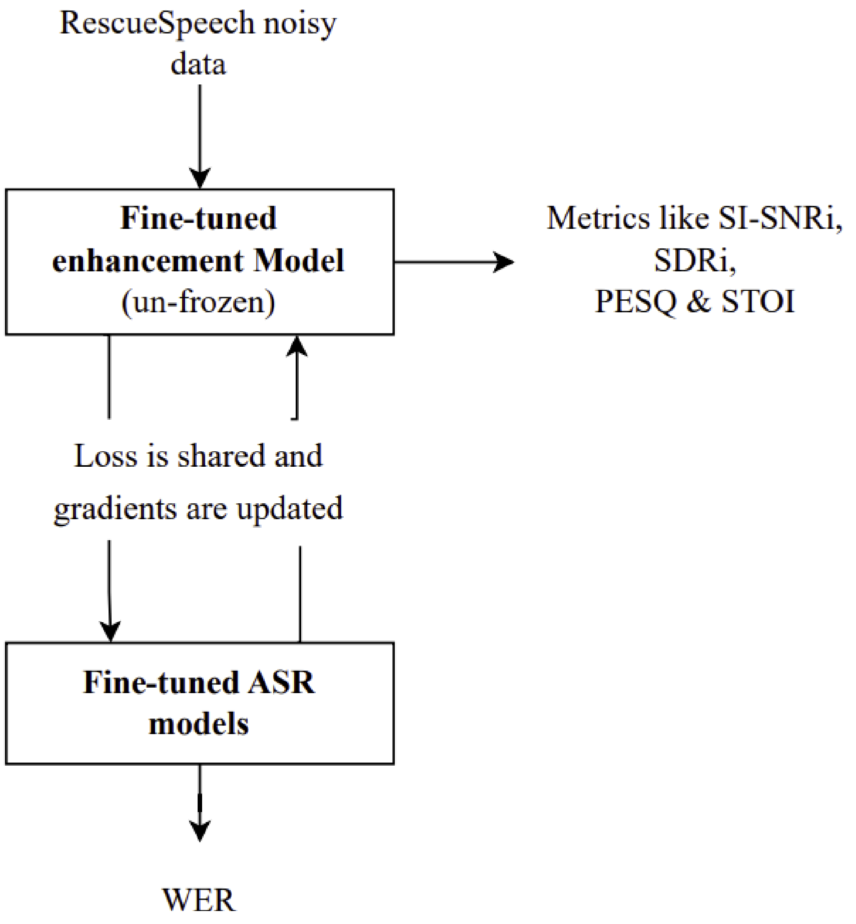
3. Training strategies

   (b) Model-combination I: Independent training



Experimental Setup

3. Training strategies
   (c) Model-combination I: Joint training



Results & Discussion

Results

UNIVERSITÄT DES SAARLANDES
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- **Pre-training Performance**
  - **ASR** dataset used- German CommonVoice (1200h)
  - **Speech Enhancement** dataset used- DNS4 (1300h)

Comparison of WER on CommonVoice test set

ASR Model	**WER**
CRDNN	7.92
Wav2vec2	9.54
WavLM	**8.98**

Evaluation on DNS4 2022 baseline dev set using DNSMOS

Model	**SIG**	**BAK**	**OVRL**
Noisy	2.984	2.560	2.205
NSNet2 [77]	3.014	3.942	2.712
SepFormer	2.999	3.076	2.437

Baseline model

Results

ASR Performance

- 1st attempt to noise robust speech recognition
  - Dataset used- RescueSpeech

WER comparison on RescueSpeech dataset

	ASR Model	clean	noisy
Pre-training	CRDNN	57.05	86.48
	Wav2vec2	50.03	86.45
	WavLM	49.81	83.82
	Whisper	28.41	61.86
Clean training	CRDNN	24.47	59.52
	Wav2vec2	22.16	65.65
	WavLM	**21.67**	61.13
	Whisper	28.39	56.60
Multi-cond. training	CRDNN	27.45	57.95
	Wav2vec2	23.91	60.61
	WavLM	22.48	**55.53**
	Whisper	29.75	62.53

Results

Combining ASR and Speech Enhancement

- 2^{nd} & 3^{rd} attempt to noise robust speech recognition
  - Dataset used- RescueSpeech

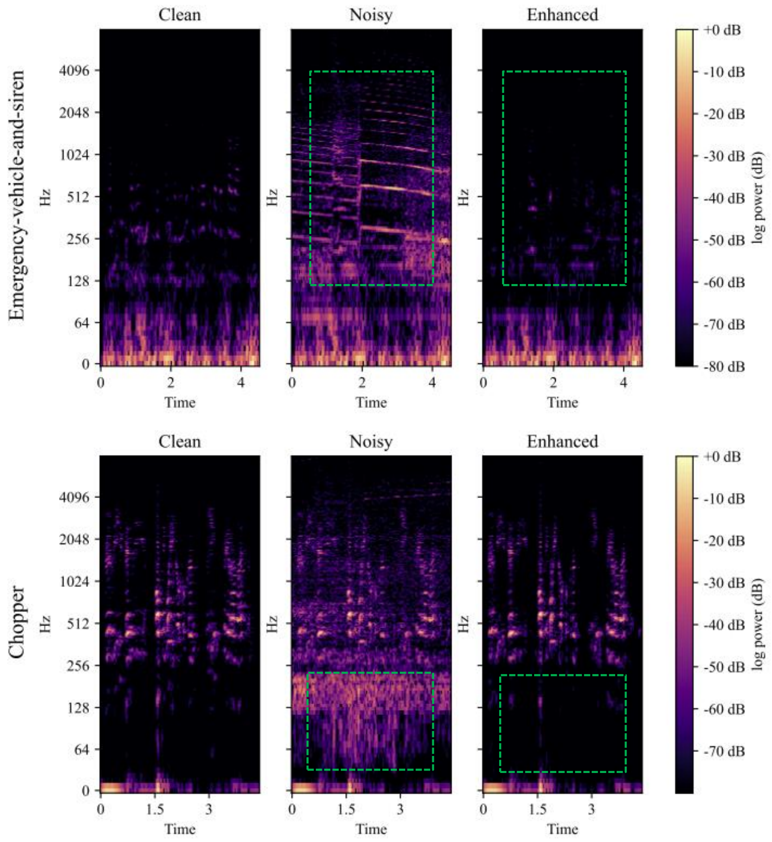
Speech enhancement performance on the RescueSpeech noisy test set

	Model Comb. I	Model Comb. II			
		CRDNN	wav2vec2	WavLM	Whisper
SI-SNRi	5.624	6.145	5.913	5.959	6.137
SDRi	5.278	5.668	5.465	5.475	5.686
PESQ	2.249	2.304	2.259	2.270	2.296
STOI	0.816	0.823	0.822	0.820	0.822

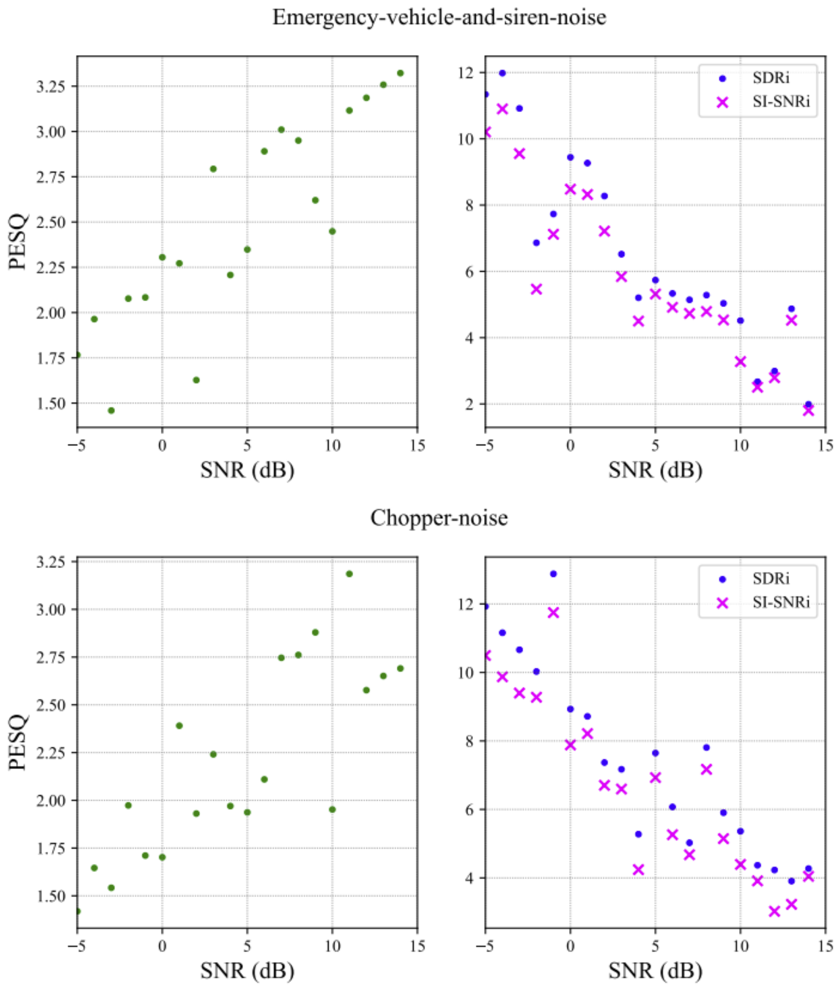
WER achieved with independent training (Model Comb. I) and joint training (Model Comb. II)

ASR Model	Model Comb. I	Model Comb. II
CRDNN	56.62	56.02
Wav2vec2	50.39	51.58
WavLM	48.25	50.04
Whisper	**29.97**	33.19

Results



Log-power spectrogram of clean, noisy, and SepFormer-enhanced utterances (-5 dB)



PESQ, SDRi, SI-SNRi vs SNR of SepFormer enhanced utterances

Conclusion

- Addressed challenges: **lack of speech data**, **robustness to SAR noises**, and conversational speech
- Introduced **RescueSpeech**: a new German speech dataset for robust speech recognition in noisy environments
- Proposed **multiple training strategies** involving fine-tuning pretrained models on in-domain data
- Tested cutting-edge self-supervised models (Wav2Vec2, WavLM, and Whisper) but best model- **Whisper** only achieved WER of 29.97%, highlighting the need for further research in this domain.

Future work

Universität des Saarlandes
Deutsches Forschungszentrum für Künstliche Intelligenz GmbH

- Consider channel characteristics in speech recognition model design and training.
- Extend dataset to include other languages (English, French, Italian, Spanish).
- Use data augmentation techniques to generate more SAR data.
- Test and compare with other Speech Enhancement models.
- Address issues with highly emotional speech.
- Address additional noise types (foot stomping, structural, interference noises).

Project Demo

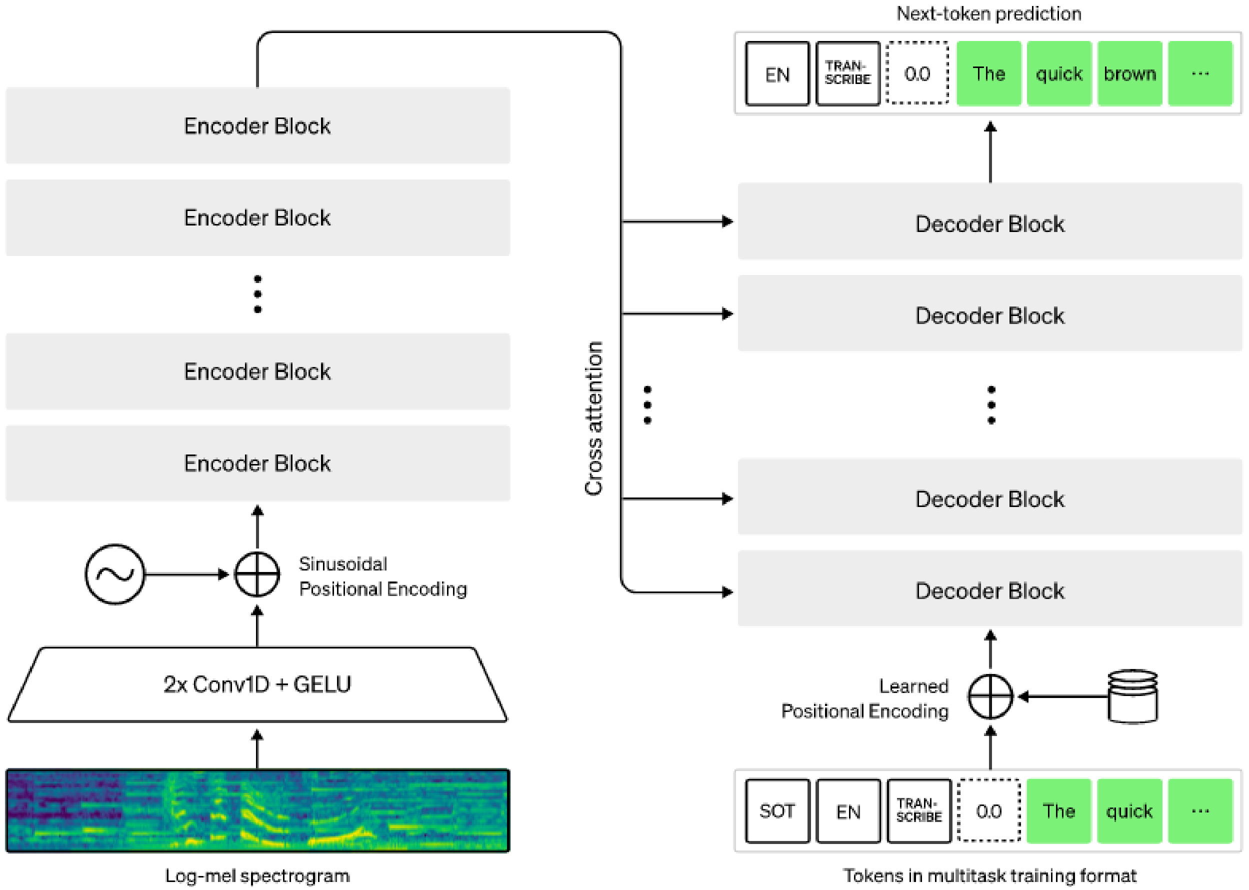
https://sangeet2020.github.io/

References

1. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2019). Common Voice: A Massively-Multilingual Speech Corpus. arXiv. https://doi.org/10.48550/arXiv.1912.06670
2. Dubey, H., Gopal, V., Cutler, R., Aazami, A., Matusevych, S., Braun, S., Eskimez, S. E., Thakker, M., Yoshioka, T., Gamper, H., & Aichner, R. (2022). ICASSP 2022 Deep Noise Suppression Challenge. arXiv. https://doi.org/10.48550/arXiv.2202.13288
3. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J., Yeh, S., Fu, S., Liao, C., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., . . . Bengio, Y. (2021). SpeechBrain: A General-Purpose Speech Toolkit. arXiv. https://doi.org/10.48550/arXiv.2106.04624
4. Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J., Yeh, S., Fu, S., Liao, C., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., . . . Bengio, Y. (2021). SpeechBrain: A General-Purpose Speech Toolkit. arXiv. https://doi.org/10.48550/arXiv.2106.04624
5. Titouan Parcollet, Mirco Ravanelli. The Energy and Carbon Footprint of Training End-to-End Speech Recognizers. 2021. ffhal-03190119
6. Subakan, C., Ravanelli, M., Cornell, S., Bronzi, M., & Zhong, J. (2020). Attention is All You Need in Speech Separation. arXiv. https://doi.org/10.48550/arXiv.2010.13154

Appendix- Whisper ASR

- Trained on 680K hrs of multilingual data.
- Directly learns mapping between utterances and transcriptions
- Model: encoder-decoder transformer



Appendix- WavLM

- Trained on 94K hrs of English data.

- Model: Transformer based
  - CNN as feature encoder