Speech Enhancement for Robust Automatic Speech Recognition

Sreeganesh Valathara Rajendran [1] S. Hamid Nawab [1]

BOSTON UNIVERSITY

Motivation

- **Automatic Speech Recognition (ASR)** is a technology that converts spoken words into text. It is a core component of systems that enable humans to interact with machines through speech.
- Advancements in ASR models, particularly in their **natural language understanding**, have enabled the creation of robust speech recognition systems.
- This robustness is limited to speech mixed with audio signals of **different statistical characteristics**, often failing with similar ones like babble or overlapping speech.
- We train a speech enhancement model to suppress the lower-volume speaker in **two-speaker** audio, improving transcription robustness in cluttered environments.

Setup

ASR Model: Whisper

- Whisper[2] is a general-purpose **speech recognition model**.
- It is trained on a large dataset of diverse audio and is also a multitasking model that can perform multilingual speech recognition, speech translation, and language identification.
- There are six model sizes offering speed and accuracy trade-offs: tiny, base, small, medium, turbo, **large**. We primarily experiment with "large".

Database: TIMIT (TIMIT Acoustic-Phonetic Continuous Speech Corpus)

- The TIMIT Database[1] consists of recordings of 630 speakers of 8 dialects of American English each reading 10 phonetically-rich sentences.
- It also comes with the word and phone-level transcriptions of the speech.
- Training Set: 4620, Test Set: 1680

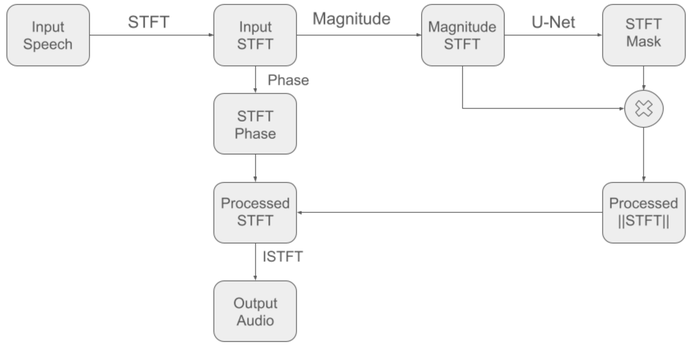
Audio Model: U-Net

- Takes in magnitude-STFT of noisy-speech dimensions [512, 128]
- 6 Convolutional Layers, 6 transposed-convolution layers.

Training

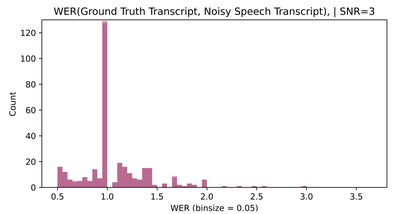
- Loss function = MAE($|STFT_{\text{Ideal}}| - |STFT_{\text{Processed}}|$) − MAE($|STFT_{\text{Processed}}| - |STFT_{\text{Input}}|$)
- Adversarial Training with two 1D Convolutional Neural Networks

Audio Processing Pipeline

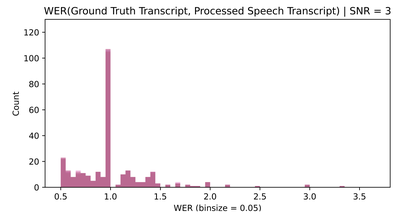


Results

WER Histogram for SNR = 3

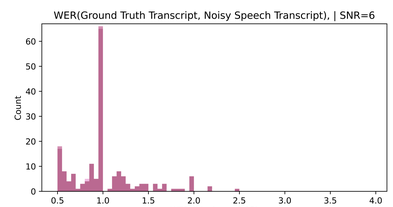


(a) Num Outliers with WER $\geq 50\%$ = 321

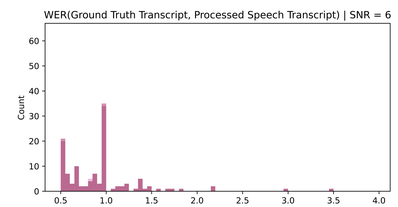


(b) Num Outliers with WER $\geq 50\%$ = 275

WER Histogram for SNR = 6

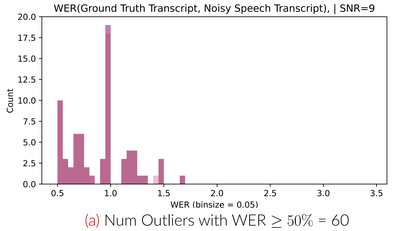


(a) Num Outliers with WER $\geq 50\%$ = 163

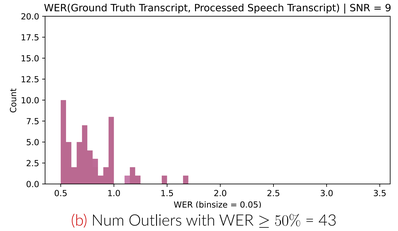


(b) Num Outliers with WER $\geq 50\%$ = 101

WER Histogram for SNR = 9



(a) Num Outliers with WER $\geq 50\%$ = 60



(b) Num Outliers with WER $\geq 50\%$ = 43

Outlier Reduction

For Whisper - Large

SNR (in dB)	Raw Input Audio	Processed Audio	Reduction in Outliers
3	321	275	46
6	163	101	62
9	60	43	17

Table 1. Outlier Reduction for Whisper - Large

Average Improvements in WER

Large

SNR	Raw Input	Processed Input	Δ WER
3	24.1%	17.3%	6.8%
6	13.9%	9.3%	4.6%
9	7.1%	6.1%	1.0%

Turbo

SNR	Raw Input	Processed Input	Δ WER
3	27.88%	20.24%	7.64%
6	14.52%	10.34%	4.18%
9	7.79%	6.26%	1.53%

Medium

SNR	Raw Input	Processed Input	Δ WER
3	30.21%	23.72%	6.49%
6	15.57%	11.77%	3.80%
9	10.13%	8.61%	1.52%

Small

SNR	Raw Input	Processed Input	Δ WER
3	34.40%	28.54%	5.85%
6	20.73%	16.50%	4.23%
9	13.96%	12.5%	1.46%

Transcription Samples

1. Sample 1
   - Ground-Truth: "Amoebas change shape constantly."
   - Noisy-Speech Transcription: "Don't ask me to carry an oily rag like that."
   - Processed-Speech Transcription: "Amoebas change shape constantly."
2. Sample 2
   - Ground-Truth: "They all agree that the essay is barely intelligible."
   - Noisy-Speech Transcription: "I took her word for it, but is she really going with you?"
   - Processed-Speech Transcription: "They all agree that the essay is barely intelligible."

Conclusion

- Developed audio processing model to minimize on-device transcription errors.
- Evaluated model's impact on transcription accuracy across ASR systems, achieving an average improvement of 6.7% at SNR = 3, 4.2% at SNR = 6, and 1.37% at SNR = 9.

Future Directions

- The key challenges lies in the inability to directly backpropagate transcription error as the black-box nature of ASR systems results in broken computation graphs.
- Such situations call for reinforcement learning methods. This task can be addressed using Policy Gradient methods by reformulating loss-function as reward function, input-speech as agent-state and output-speech as agent-action.

References

[1] John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, and Nancy L. Dahlgren.
Timit acoustic-phonetic continuous speech corpus.
Technical report, Linguistic Data Consortium, Philadelphia, PA, 1993.

[2] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever.
Robust speech recognition via large-scale weak supervision, 2022.

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox.
U-net: Convolutional networks for biomedical image segmentation, 2015.