# Robust Automatic Speech Recognition and Transcription with Wav2Vec 2.0 and Whisper

Data Science Institute
COLUMBIA UNIVERSITY

**Authors: Sivan Ding, Alexandria Guo, Anh-Vu Nguyen, Antonin Vidon, Julia Wang, and Maxwell Zhou**
**Industry Mentors: Bhushan Jagyasi, Priyanka Pandey, and Surajit Sen**

**Data Science Capstone Project with Accenture AI**

## Background and Motivation

**Whisper and Wav2Vec2.0 are large open source-based ASR (Automatic Speech Recognition) models developed respectively by OpenAI and Facebook, allowing users to transcribe audio to text. The architectures of both models are summarized in Figure 1.**

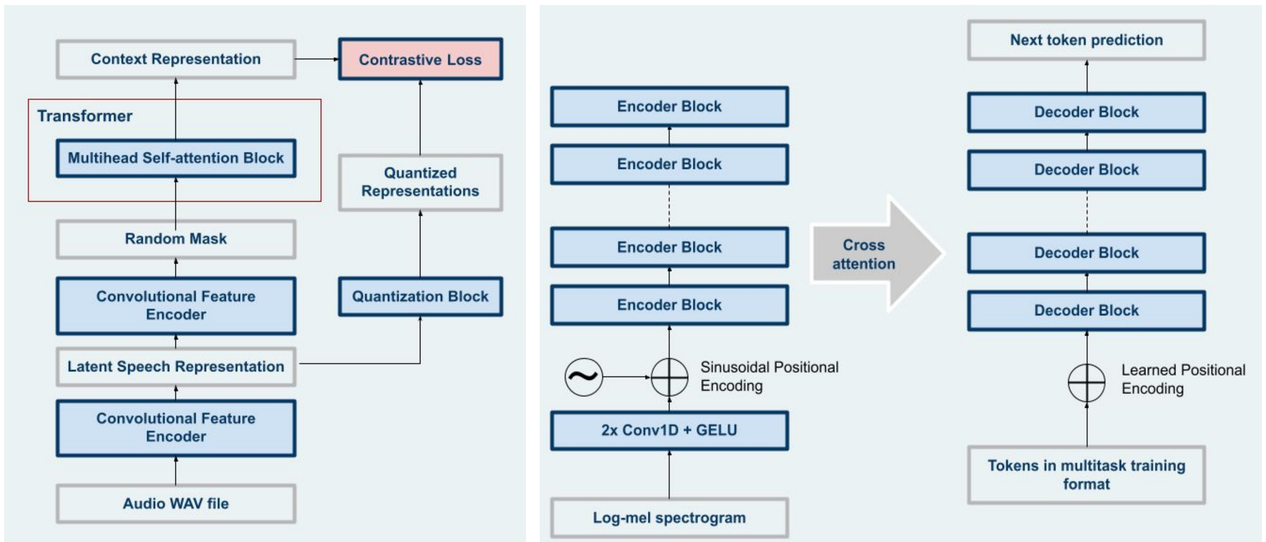



**Figure 1. Wav2Vec 2.0 pre-training (left) and Whisper (right) diagrams**

**The goal of our capstone project is to evaluate and compare the performance of each model across various complex scenarios, in order to test their robustness and determine the strengths and weaknesses of each model.**

## Preliminary Results

**We examined the ASR frameworks in transcription tasks under more challenging conditions, where we downsampled and added noise to Librispeech audio to compare the robustness of Wav2Vec2.0 and Whisper. The metric we use for evaluation is Word Error Rate (WER) %.**

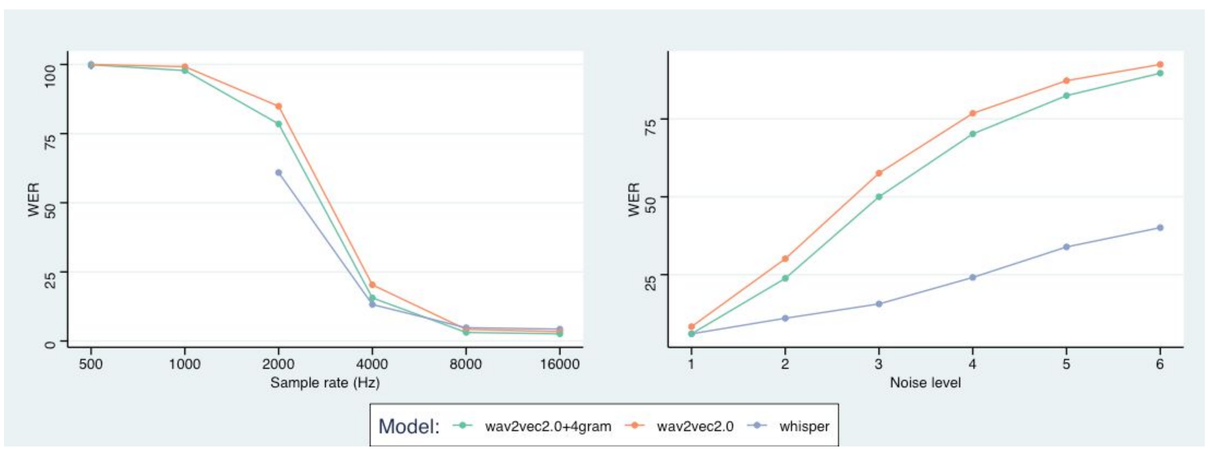



**Figure 2. Robustness of Wav2Vec2.0 and Whisper to downsampling (left) and noise (right)**

## Confidence Level Results

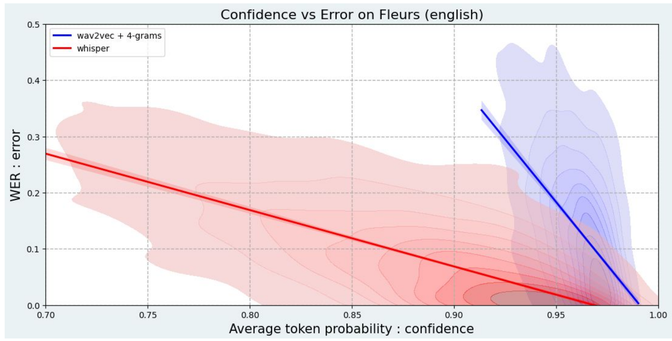



**Figure 3. Correlation b/w confidence level and error**

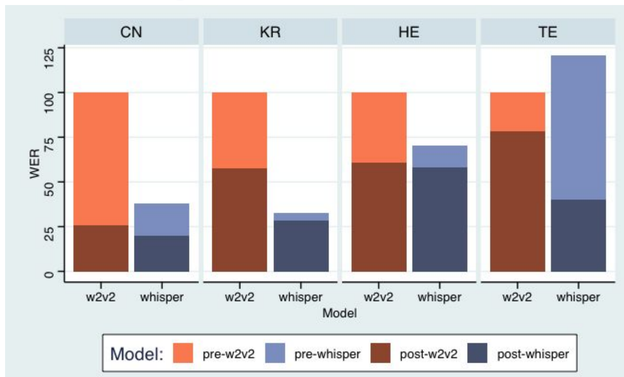
```
w2v2+4g : the sport is primarily played in a dual form at one fencer duly another
whisper  : the sport is primarily played in a dual format one fencer dualing another
g.truth  : the sport is primarily played in a duel format one fencer dueling another
WER w2v : 28.6%  - confidence range : [84.4%, 100.0%]
WER wsp : 15.4%  - confidence range : [51.9%, 99.6%]

w2v2+4g : this might require filling out a form with the local police or visit to emigration offices
whisper  : this might require filling out a form with the local police or revisit to the immigration offices
g.truth  : this might require filling out a form with the local police or a visit to the immigration offices
WER w2v : 5.6%  - confidence range : [93.4%, 100.0%]
WER wsp : 11.8%  - confidence range : [54.1%, 99.8%]

w2v2+4g : police said they suspect an alleged dash i s l militant of responsibility for the attack
whisper  : police said they suspect an alleged dish isil militant of responsibility for the attack
g.truth  : police said they suspect an alleged daesh isil militant of responsibility for the attack
WER w2v : 29.4%  - confidence range : [92.0%, 100.0%]
WER wsp : 7.1%  - confidence range : [15.0%, 99.8%]
```

**Figure 4. Sample predictions colored by confidence level of respective token: Wav2Vec2.0 + 4-grams vs. Whisper**

## Fine-tuning Results




| Training Data | | | |
|---|---|---|---|
| Language | Whisper Pre-trained (hrs) | Wav2Vec2.0-XLSR Pretrained (hrs) | Fleurs-Train (hrs) |
| Chinese | 23446 | 90 | 10 |
| Korean | 7993 | 61 | 8 |
| Hebrew | 688 | 77 | 10 |
| Telugu | 4 | 62 | 8 |

**Table 1. Summary of Training Data used to train and finetune each model**

**From table 1, we see the Whisper model is pretrained on many more hours of data compared to Wav2Vec2.0 with the exception of Telugu.**

**Figure 5. WER of Wav2Vec2.0 and Whisper on Fleurs Multilingual Test Data**

**In Figure 5, we see that across the board Whisper performs better than Wav2Vec2.0 in terms of WER, after fine-tuning. For Telugu and Korean, Whisper has roughly half the WER, while Chinese and Hebrew are more comparable. Overall, fine-tuning has a significant impact on the performance of both Wav2Vec2.0 and Whisper, with the largest improvement in Telugu for Whisper and in Chinese for Wav2Vec2.0.**

## Conclusion and Next Steps

**Testing has shown that Whisper overall outperforms Wav2Vec2.0, especially in terms of noise and downsampling robustness. While Whisper gives a better WER over Wav2Vec2.0, fine-tuning has shown consistent improvement over the baseline models.**

## Acknowledgments

**We thank Accenture, Bhushan Jagyasi, Surajit Sen, and Priyanka Pandey for advising our team on this project, and providing good insights and guidance on our project. We would also like to thank Meta, OpenAI, and HuggingFace for their ASR open source contributions.**

## References

**Baevski, A., Zhou, H., Mohamed, A., and Auli, M. 2020. "wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations"**

**Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. 2022, Preprint. "Robust Speech Recognition via Large-Scale Weak Supervision"**