How Does OpenAI's Whisper Interpret Dysarthric Speech?

Orhan Agaoglu | 5518725 | oagaoglu@tudelft.nl
Supervisors: Zhengjun Yue, YuanYuan Zhang
TU Delft Computer Science and Engineering

Introduction

Background

- OpenAI's Whisper has emerged as a state-of-the-art model in speech recognition, showcasing exceptional performance across diverse speech domains.
- Whisper's proficiency in transcribing both typical and atypical speech patterns underscores its significance in advancing speech recognition technology.
- However, the inner workings of Whisper's encoder blocks remain opaque, posing a challenge in understanding how it processes acoustic features, particularly in atypical speech.

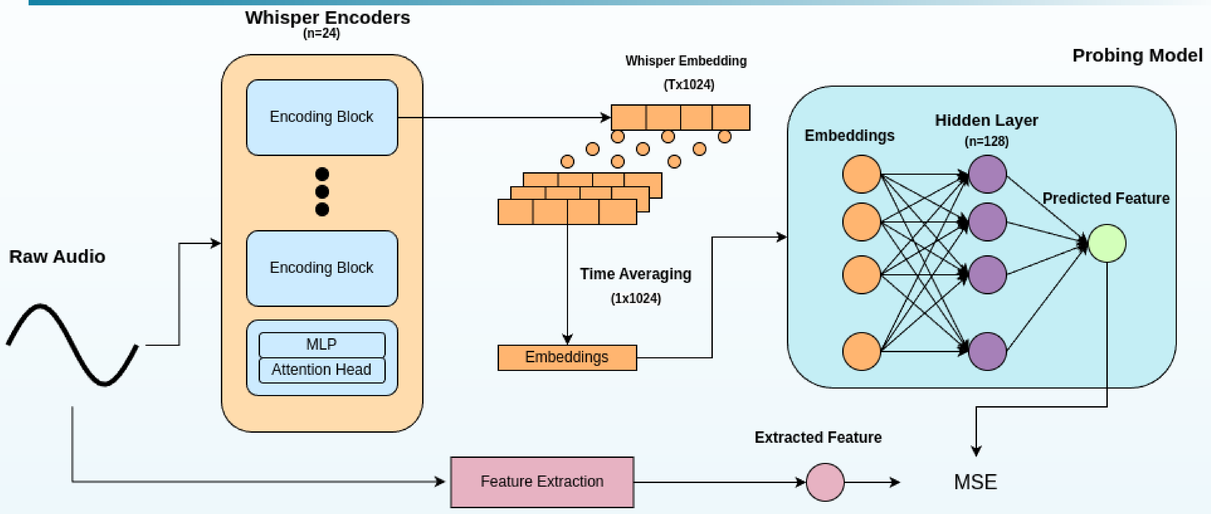
Research Question

- The primary research question guiding this study is:
  - **"How effectively does OpenAI's Whisper encode acoustic features of dysarthric speech?"**
- By probing Whisper's encoding layers, we aim to elucidate its treatment of acoustic parameters in dysarthric speech and assess its effectiveness in capturing the unique characteristics of atypical speech.
- This investigation holds promise for enhancing our understanding of Whisper's capabilities and advancing explainable AI in speech recognition, particularly in the context of dysarthric speech recognition.

Method

- **Data Extraction**: Extract Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) from TORGO datasets as labels for the audio files.
- **Embedding Extraction**: Utilize OpenAI's Whisper to extract embeddings from the speech samples. Whisper converts raw audio into log-mel spectrograms and processes them through a series of encoding blocks.
- **Probing Models**: Train Feedforward regressors using the extracted embeddings to predict specific acoustic features such as loudness, pitch, and spectral slopes.
- **Evaluation**: Evaluate the performance of the regressors across different encoding blocks of Whisper. Compare the extracted features against control tasks using random vectors to ensure meaningful capture of information.
- **Analysis:** Analyze the results to understand how Whisper's encoding layers treat acoustic features differently between dysarthric and normal speech domains.

Experiment

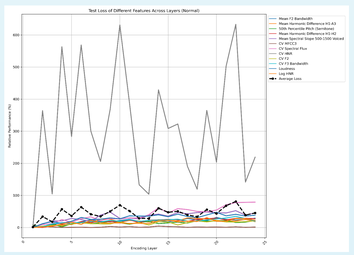


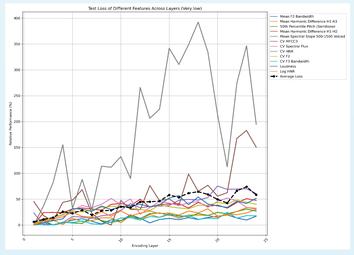
Feature Extraction: Key acoustic features such as loudness, pitch, and spectral slopes are extracted and analyzed. The probing tasks are designed to assess how well Whisper captures these features at various encoding layers.

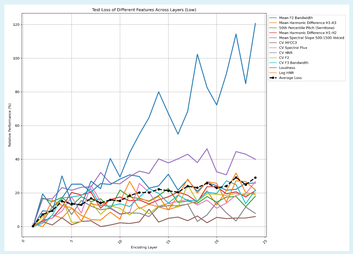
Whisper Encoders: Whisper processes audio inputs through multiple encoding blocks, generating embeddings that capture different aspects of the audio signal.

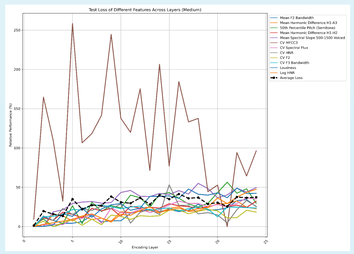
Probing Model: Feedforward regressors are trained on these embeddings to predict specific acoustic features.

Results









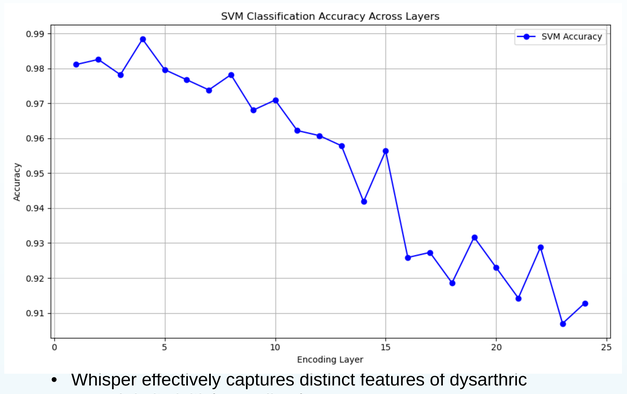
Normal Severity: Initial layers effectively capture distinct features like "Loudness" and "50th Percentile Pitch (Semitone)" with low variability. Features such as "Mean Spectral Slope 500-1500 Voiced" exhibit higher fluctuations. As we move to deeper layers, representations become more generalized, leading to higher test losses and reduced discriminative power. Comparison with a random baseline shows features like "CV HNR" are well captured, while "CV MFCC3" performs worse.

Very Low Severity: Features like "CV MFCC3" show minimal loss in the 9th layer, while "CV HNR" displays significant spikes. Deeper layers exhibit increased average loss, indicating reduced effectiveness. Most probes outperform the random baseline, except for "Mean F2 Bandwidth".

Low Severity: "Mean F2 Bandwidth" shows increasing test loss in deeper layers, while other features remain stable. Deeper layers show increased average loss. Features such as "50th Percentile Pitch Semitone" and "CV HNR" fail to surpass the random baseline.

Medium Severity: Initial layers capture specific information well, while "CV MFCC3" shows fluctuations with minimal loss at Layer 21. Deeper layers have increased average loss. Most features outperform the random baseline, except "CV F3 Bandwidth".

Conclusion / Discussion



- Whisper effectively captures distinct features of dysarthric speech in its initial encoding layers.
- Deeper layers tend to generalize features, reducing their discriminatory power for specific speech impairments.
- Despite generalization, Whisper shows strong zero-shot performance in distinguishing severity levels of dysarthric speech.
- Probing analysis highlights that features like "Loudness" and "50th Percentile Pitch (Semitone)" are well encoded.
- Challenges remain in consistently capturing features like "Mean Spectral Slope" and "Harmonic Differences" across layers.
- Understanding Whisper's processing of atypical speech can enhance its application in healthcare.
- This study promotes the transparency and trustworthiness of AI-driven diagnostic tools.
- Future work should focus on fine-tuning Whisper for better handling of dysarthric speech.
- Further exploration of Whisper's clinical applications is needed to maximize its potential in medical settings.

References

A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in International Conference on Machine Learning. PMLR, 2023, pp. 28 492–28 518.

V. Chemudupati, M. Tahaei, H. Guimaraes, A. Pimentel, A. Avila, M. Rezagholizadeh, B. Chen, and T. Falk, "On the transferability of whisper-based representations for 'in-the-wild' cross-task downstream speech applications," 2023.

C. Graham and N. Roll, "Evaluating openai's whisper asr: Performance analysis across diverse accents and speaker traits," JASA Express Letters, vol. 4, no. 2, 2024.

S. Rathod, M. Charola, and H. A. Patil, "Transfer learning using whisper for dysarthric automatic speech recognition," in International Conference on Speech and Computer. Springer, 2023, pp. 579–589.

P. Lieberman, "Primate vocalizations and human linguistic ability," The Journal of the Acoustical Society of America, vol. 44, no. 6, pp. 1574–1584, 1968.

Y. Belinkov, "Probing classifiers: Promises, shortcomings, and advances," 2021.