How Does OpenAI's Whisper Interpret Dysarthric Speech?

An Analysis of Acoustic Feature Probing and Representation Layers for Dysarthic Speech

Orhan Agaoglu[1]

Supervisor(s): Zhengjun Yue[1]**, YuanYuan Zhang**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Orhan Agaoglu
Final project course: CSE3000 Research Project
Thesis committee: Zhengjun Yue, YuanYuan Zhang, Thomas Durieux

How Does OpenAI's Whisper Interpret Dysarthric Speech?

Orhan Agaoglu[1]

Abstract

This paper investigates how OpenAI's Whisper model processes dysarthric speech by probing its internal acoustic feature representations. Utilizing the TORGO database, we analyzed Whisper's capability to encode significant acoustic features specific to dysarthric speech across its encoding layers. Our findings reveal that initial layers are particularly effective in capturing distinct features, while deeper layers show generalized representations. Despite this, Whisper's zero-shot performance in distinguishing dysarthric speech severity levels is noteworthy. We employed a series of probing tasks tailored to dysarthric speech characteristics to pinpoint specific features and their transformation across the model's layers. This study highlights Whisper's potential in handling atypical speech patterns without fine-tuning, paving the way for further research into the interpretability and application of transformer-based models in medical and assistive technologies. We discuss the implications of these results for enhancing transparency, reliability, and safe AI integration in healthcare.

Index Terms: Interpretability, Whisper, Probing, Dysarthic Speech, Acoustic Features

1. Introduction

Since its release, OpenAI's Whisper models have emerged as a standout example of transformers their capabilities the field of Automatic Speech Recognition (ASR). Demonstrating great performance across recognition benchmarks, Whisper has excelled in converting raw audio into accurate textual representations [1]. Its transformer architecture not only meets the demanding benchmarks of ASR but also shows promising results in a variety of downstream tasks. These include enhancing speech translation, speaker identification, and various classification tasks underscoring the model's adaptability and broad applicability[2].

Whisper's architecture is rooted in the weakly supervised encoder-decoder structure, designed to handle complex, variable input patterns efficiently. At the core of Whisper's effectiveness is its multi-layered encoder, which processes audio signals through a series of encoding blocks. Each block applies self-attention mechanisms that adaptively focus on different aspects of the audio input, capturing nuances critical for accurate recognition[1]. This encoding blocks are so well-trained that they further enhance the capability of Whisper on atypical speech domain[3]. This training process enables the model's encoding layers to develop a robust understanding of non-standard speech patterns, such as those found in dysarthric speech[4].

Dysarthric speech is characterized by slurred or slow speech that can result from muscle weakness or neurological damage, presents significant challenges for ASR technologies[5]. Historically, dysarthric speech has been analyzed using a variety of acoustic features and machine learning models, exploring its prosodic, spectral, cepstral, and voice quality dimensions[6, 7]. Within these studies, certain acoustic features were extracted from the voice samples and a machine learning model was trained based on the hypothesis that these features are indicative of dysarthic speech. Many studies concluded well results in the recognition and classification of dysarthic speech through these acoustic features [8] yet Whisper is also considerably capable of competing with the studies[9].

Recent advancements have shown that applying transfer learning techniques to the Whisper's encoding layers is an effective way of recognizing dysarthric speech[4]. By fine-tuning Whisper with dysarthric speech datasets, researchers have been able to tailor the model's sensitivity to the specific acoustic features of speech impairments. The contrast between the use of transformers and the prior studies is that the features extracted from Whisper are more difficult to interpret. Therefore, these studies are based on the assumption that within the encoding layer's Whisper is capable of capturing high level acoustic features that containts the information related to dysarthic speech.

The competitive performance of Whisper in processing dysarthric speech raises important questions about the model's interpretability, particularly in handling complex and atypical speech patterns. This aspect remains largely unexplored in the existing literature, highlighting the need for further investigation. While previous voice foundation models such as wav2vec2 [10] and MockingJay [11] have been probed for their ability to process and interpret speech features [12], the probing of Whisper's encodings and its adaptability to atypical speech types like dysarthric speech remains an unstudied question.

The main motivation of interpreting "How Whisper processes dysarthric speech" is rooted in the potential benefits of applying explainable AI (XAI) techniques in medical domains [13]. By understanding the mechanisms through which Whisper processes dysarthric speech, we can further improve the model's reliability and trustworthiness, which is crucial for medical applications where decisions can significantly impact patient care. Interpreting such transformers, provides transparency, allowing clinicians to understand and trust the AI's decision-making process. Additionally, insights gained from such interpretability can drive improvements in model design, leading to more robust and accurate ASR systems tailored for dysarthric speech. The application of XAI in this domain supports the broader goal of integrating advanced AI technologies into healthcare, ensuring they are used safely and effectively[14, 13].

Therefore my thesis aims to fill this gap in literature by analyzing how Whisper adapts to and processes dysarthric speech,

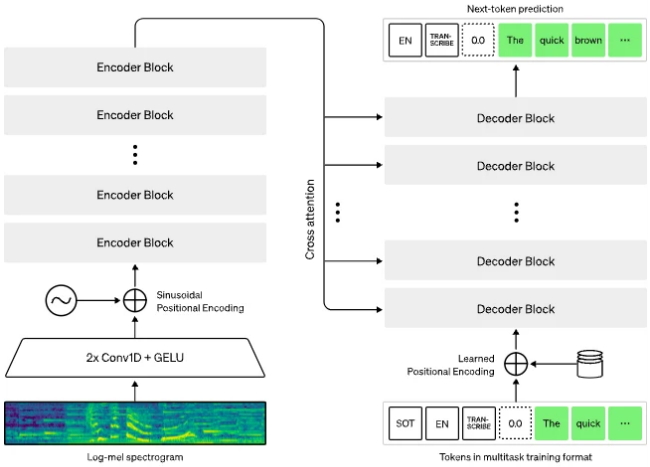

Figure 1: *Illustration of the Whisper model architecture*

utilizing its encoders to interpret what Whisper learn from the speech that undergoes through encoding layers and potential improvements in the field of ASR for impaired speech conditions. The specific contributions of this research are outlined as follows:

1. **Probing Tasks Specific to Dysarthric Speech:** To gain insights into the encoder's capabilities, 12 distinct probing tasks tailored specifically to the characteristics of dysarthric speech are chosen based on previous papers [6, 7]. These tasks are designed to test the encoder's ability to identify and differentiate between various severity levels of speech impairments, and to understand how the model's internal representations change in response to these variations.
2. **Detailed Analysis of Layer-wise Learning in the Encoder:** Each layer of Whisper's encoder potentially encodes different aspects of the speech input. A detailed analysis of what is learned at each layer aims to map the transformation of the acoustic signal from raw input to the model's final learnt representation.
3. **Comparative Analysis of Normal and Dysarthric Speech Processing:** A core aspect of this research involves comparing how Whisper processes normal speech versus various severity levels of dysarthric speech. This comparison will not only reveal the differences in how the model treats these two types of speech but will also shed light on its generalization capabilities across different speech conditions.

To the best of my knowledge, this is the first attempt towards interpreting Whisper's encodings within the atypical speech domain. This exploration aims to contribute significant insights into the adaptability and limitations of current transformer-based ASR technologies in handling diverse and challenging speech impairments.

2. Methodology

This section presents the methodology employed to investigate Whisper. It covers the technical aspects of the model's architecture, the data employed for analysis, and the specific probing tasks designed to evaluate the model's capability. The objective is to decode the internal mechanics of Whisper, thus providing insights into its ability to adapt and respond to the intricacies of speech affected by dysarthria. Each subsection will explain the entitled methodology followed.

2.1. Dataset

In this study, the TORGO database [15] will be used, which consists of high-quality audio recordings from individuals with various degrees and types of dysarthria, as well as age- and gender-matched controls. The variety of severity conditions and corresponding control group will allow us to perform a well-assessed analysis. We will use these speech files to extract both acoustic features that we are going to probe and internal representations from the Whisper model. These features and representations will then be analyzed through probing tasks, which are designed to assess how well Whisper captures the unique acoustic characteristics of dysarthric speech. The specifics of these probing tasks and their relevance to Whisper's performance will be detailed in the following paragraphs.

Speech samples are categorized into severity levels based on previous papers [16, 17] in Table 1, aiming a detailed analysis and evaluation of Whisper's performance on various severity of dysarthric speech. The dataset is divided into training and testing sets, with 80% allocated to training and 20% to testing. The training set is used to train the probing models, while the testing set is reserved for evaluating the representations of acoustic features within encodings of Whisper. This approach allows us to measure the variation of impairment levels affect on Whisper's capabilities.

Severity Level	# of Patients	# of Speech Files
Normal	7	6236
Very Low	2	864
Low	2	976
Medium	4	1341

Table 1: *Number of people and total number of speech files per severity level*

2.2. Whisper

The Whisper model processes audio inputs through a encoder-decoder pipeline that begins with the conversion of audio waveforms into log-mel spectrograms as seen in Figure 1. These spectrograms are then passed through two convolutional layers with GELU activation [18] to extract salient features. Following feature extraction, sinusoidal positional encoding is applied to incorporate sequence information. The core of the architecture consists of 24 encoder blocks [1](in the medium configuration of Whisper), each utilizing a multi-head self-attention mechanism followed by Multi Layer Perceptrons (MLP)[19]. This structure allows the model to capture and emphasize different aspects of the audio signal, generating a series of internal representations at each layer. These representations are key to our study; they will be extracted and analyzed to understand how Whisper processes and encodes features of dysarthric speech across its layers. This analysis will inform our probing tasks, aiming to reveal how the model discriminates between typical and atypical speech patterns and what features each encoding layer captures in dysarthric speech.

2.3. Probing

To interpret the process that input goes through, probing models [20] will be used. These regressors (g) are utilized to evaluate the internal representations ($\mathbf{r}$) generated by the Whisper model

Table 2: *Architecture details of the Whisper model family.*

Model	Layers	Width	Heads	Parameters
Tiny	4	384	6	39M
Base	6	512	8	74M
Small	12	768	12	244M
Medium	**24**	**1024**	**16**	**769M**k
Large	32	1280	20	1550M

across its layers. The performance of the regressors will be used to quantify Whisper's capability in encoding properties specific to dysarthric speech.In the following subsections, the methodology for probing will be explained.

2.3.1. Extracting Representations

For each audio input x, Whisper produces a set of layer-specific representations:

$$\mathbf{r}_l = f_l(x),$$

where l indexes the layer, for layers $l = 1, \ldots, L$, and L is the total number of layers in the medium configuration of Whisper.

2.3.2. Designing Probing Tasks

Probing tasks are structured to predict distinct linguistic or acoustic properties z from the representations $\mathbf{r}_l$. Each task t is aimed at assessing a particular feature of dysarthric speech. From now on, *CV* will refer to the coefficient of variation. The specific features chosen for the probing tasks, based on their significance in detecting dysarthric speech as analyzed in previous studies, are as follows:

- **Loudness**: The average loudness of the speech signal.
- **50th Percentile Pitch (Semitone)**: The median pitch measured in semitones.
- **Mean Spectral Slope 500-1500 Voiced**: The average spectral slope in the 500-1500 Hz range for voiced segments.
- **Mean Harmonic Difference H1-H2**: The average difference in amplitude between the first and second harmonics.
- **Mean F2 Bandwidth**: The average bandwidth of the second formant.
- **Mean Harmonic Difference H1-A3**: The average difference in amplitude between the first harmonic and the third formant peak.
- **CV MFCC3**: The coefficient of variation of the third Mel-frequency cepstral coefficient.
- **CV HNR**: The coefficient of variation of the Harmonics-to-Noise Ratio.
- **CV Spectral Flux**: The coefficient of variation of the spectral flux.
- **CV F2**: The coefficient of variation of the second formant frequency.
- **CV F3 Bandwidth**: The coefficient of variation of the third formant bandwidth.
- **Log HNR**: The logarithm of the Harmonics-to-Noise Ratio.

These features are specifically chosen for their performance in previous studies and their ability to be used as discriminators for dysarthric speech.

2.3.3. Probing Model

Depending on the task the complexity of the probing model varies. In similar studies it is often chosen as a simple architecture like MLP's. By examining the performance of this probing model, we gain insights into how well different layers of Whisper encode the features of interest.

2.3.4. Probing Analysis

The described probing regressor g_t is applied to predict the property z from $\mathbf{r}_l$. The regressor's performance is evaluated by mean square error(MSE) loss:

$$\text{MSE}(g_t(\mathbf{r}_l), z),$$

which reflects how well the property z is represented within $\mathbf{r}_l$. Low loss indicates effective encoding of z, whereas high loss values suggests inadequate representation or extractability.

2.3.5. Interpretation of Results

Differential loss across layers and tasks will be examined to deduce the depth at which Whisper processes and encodes features present in dysarthric speech. In order to prove the validity of the information representation, a control task consisting on random vectors will be compared as a baseline probe [20].

2.4. SVM Classifier

To evaluate our findings about the features we employed a Support Vector Machine (SVM) [21]classifier to evaluate the effectiveness of Whisper's internal representations in the downstream task of in discriminating different severity levels of dysarthric speech. The SVM classifier provides a reference performance for Whisper encodings by assessing their capability to classify the severity levels at each layer of the model. This approach helps in understanding the discriminative power of the internal representation and encoded features.

3. Experiments

3.1. Data Preparation

The dataset was divided into four severity levels: Normal, Very Low, Low, and Medium, based on the speaker IDs outlined in Table 1. Then the dataset was used to extract acoustic features using the openSmile library [22] and additional scripts to calculate chosen features. Additionally, the voice samples are fed into the frozen Whisper model to extract the representations where Whisper is configured to ignore paddings. The extracted representations are time-averaged, preparing the dataset for SVM classifier and probing tasks.

3.2. SVM Classifier

To evaluate the effectiveness of Whisper's internal representations in discriminating between different severity levels of dysarthric speech, we employed a SVM classifier. The SVM classifier serves as a reference performance measure for Whisper encodings, assessing their capability to classify the severity levels at each layer of the model.

For each encoding layer of Whisper, the extracted representations were used to train an SVM classifier with a radial basis function (RBF) kernel. The classifier was configured with a regularization parameter $C = 5$. The dataset was split into training (80%) and testing (20%) sets, ensuring stratified sampling

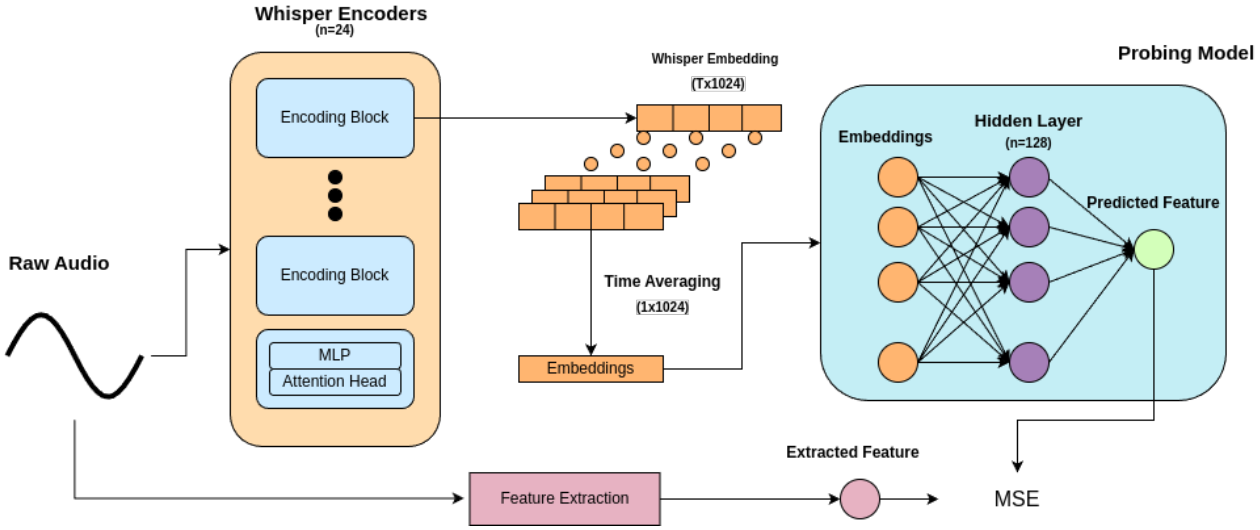

Figure 2: *Illustration of the model architecture*

to maintain equal representation of severity levels in both sets. The use of stratified sampling is crucial in this context, as it allows us to assess the effectiveness of Whisper's representations in distinguishing between severity levels, rather than optimizing for the best accuracy. This approach ensures an equal distribution of severity levels, providing a more accurate evaluation of whether Whisper's internal representations carry severity information through the layers. The SVM classifier was trained on the scaled representations, and its performance was evaluated based on accuracy and confusion matrix metrics.

3.3. Model

We trained separate probing regressors for each severity level, each layer, and each acoustic feature. The regressors were configured as MLP with one hidden layer consisting of 128 neurons, using a ReLU activation function [23] and dropout regularization [24]. The training process involved optimizing the MSE loss using the ADAM optimizer [25] with a learning rate of 0.001. The entire pipeline, including data preparation and probing, can be seen in Figure 2. This approach ensures a detailed analysis of Whisper's internal representations and their effectiveness in capturing dysarthric speech features.

In addition to training probing models on the actual representations, a control task was conducted using randomly generated vectors. These random vectors were matched in size to the original representations and trained on the same acoustic feature labels. This control task allows for a baseline comparison to validate that the probing models are indeed capturing meaningful information from Whisper's representations, rather than merely learning from the labels. By comparing the performance of probing models on true representations against the random baseline, we can have a better understanding of the selectivity of the probing model.

3.4. Evaluation of Results

The results were evaluated in multiple ways:

1. **Across Severity Levels:** We analyzed how well Whisper's internal representations could discriminate between different severity levels of dysarthric speech. This was done using both the probing models and the SVM classifier, comparing their performance across the severity levels.
2. **Across Features:** We examined how different acoustic features were encoded across the layers of Whisper. The probing models' performance on each feature provided insights into which layers of Whisper were most effective in capturing specific acoustic characteristics of dysarthric speech.
3. **Comparison with Control Task:** By comparing the performance of probing models on true representations against the probe based on random vectors, we assessed the selectivity and effectiveness our probes of representing Whisper's capability in encoding meaningful information.

4. Results Discussions

4.1. PCA Visualization of Representations

To understand how Whisper processes and represents dysarthric speech across its encoding layers, we performed principal component analysis (PCA) [26] on the extracted representations. Plots in Figure 3, illustrate the distribution of representations for different severity levels at Encoding Layers 1, 14, and 23, respectively.

In the initial layers (e.g., Encoding Layer 1), the representations of different severity levels are more dispersed. This indicates that these layers capture distinctive features, making it easier to differentiate between the severity levels. As we move deeper into the network (e.g., Encoding Layer 23), the representations converge and become more centralized around the "Normal" severity level. This centralization suggests that deeper layers encode more generalized features, which may reduce the classifier's ability to distinguish between different severity levels.

4.2. Probe Analysis Across Severity Levels

4.2.1. Normal Severity

The plot (Figure 4) shows the test loss for each feature across the layers for Normal severity. Most features demonstrate consistent trends across the layers with minimal fluctuations in test loss, suggesting stable representations. Features like "Loud-

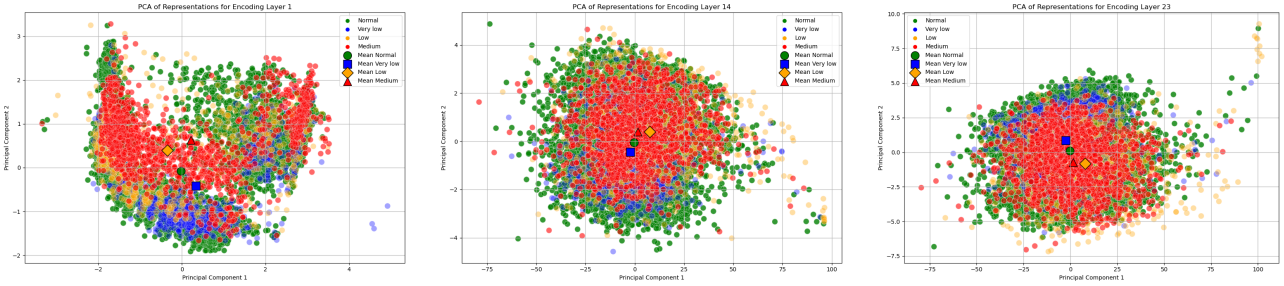

Figure 3: *PCA of Representations for Encoding Layers 1, 14, and 23.*

ness" and "50th Percentile Pitch (Semitone)" show low variability, indicating effective capture of these characteristics. Some features, such as "Mean Spectral Slope 500-1500 Voiced" and "Mean Harmonic Difference H1-H2," exhibit higher fluctuations, indicating challenges in consistent capture across layers. The average loss, represented by the black dashed line, increases slightly in deeper layers, suggesting these layers focus more on generalized features rather than specific acoustic characteristics.

The initial layers (e.g., Encoding Layer 1) effectively capture distinct features, making it easier to differentiate between severity levels, as indicated by the dispersed PCA plots and low test losses. However, as representations progress through deeper layers (e.g., Encoding Layer 23), they become more generalized, converging around the "Normal" severity level and resulting in higher test losses.

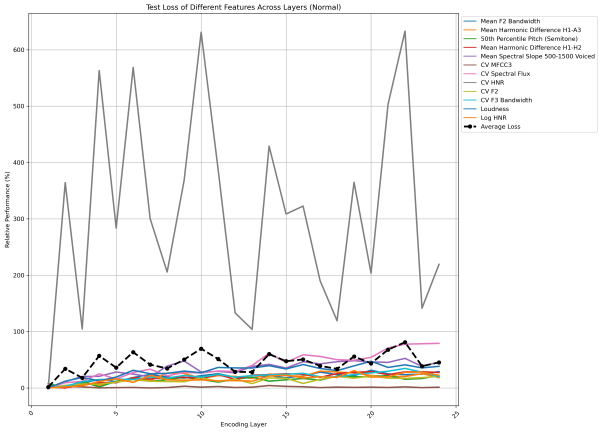

Figure 4: *Test Loss of Features Across Layers (Normal).*

To validate the probing results, we compared the minimum test loss of each feature against a random baseline, as shown in Figure 5. The baseline was generated using randomly assigned vectors, providing a control to ensure that the probes capture meaningful information from Whisper's representations.

From the comparison, it is evident that the actual probes perform significantly better than the random baseline for most features. For instance, feature "CV_HNR" has much lower test losses compared to the random baseline, indicating that Whisper's internal representations contain valuable information for these features. However, feature "CV_MFCC3" show worst performance compared to the random baseline, suggesting that these features are not effectively encoded or more challenging to learn from the representations.

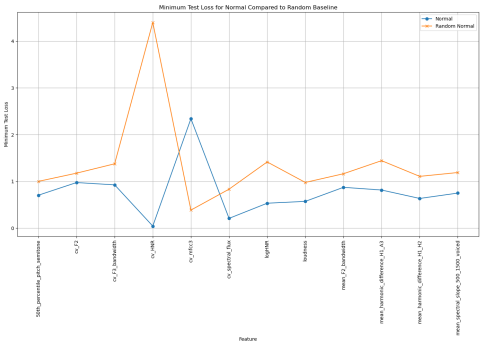

Figure 5: *Minimum Test Loss for Normal Compared to Random Baseline*

4.2.2. Very Low Severity

The analysis for the "Very Low" severity level, illustrated in Figure 6, demonstrates varied performance across different acoustic features throughout Whisper's encoding layers.

For most of the features the minimal loss is achieved in the first couple layers. For the feature *CV MFCC3*, the performance is minimal in the 9th layer, indicative of a characteristic to the feature that leads to be encoded in deeper layers. Contrastingly, *CV HNR* show significant spikes in test loss throughout layers. Notably, it experiences a sharp increase around Encoding Layer 12 and then again around Encoding Layer 18. This observation might be indicative of Whisper's process of normalizing the speech characteristic and capturing the distinctive properties again.

The average loss generally increases in the later layers, particularly noticeable after Encoding Layer 15. This trend is indicative of a reduction in the model's ability to discriminate between different severities as it moves to deeper layers, correlating with the PCA analysis.

The comparison between the minimum test loss for Very Low severity and the random baseline probe is presented in Figure 7. The random baseline, shows higher test losses compared to the actual Very Low severity probes in all features except *Mean F2 Bandwith*. This indicates either a lack of learning from the representations or the absence of this information in the encodings. For the rest of the features, this graph underscores that, while the probes for Very Low severity might not capture the acoustic features as effectively as those for Normal severity, they still outperform the random baseline, indicating some level of learning from the representations.

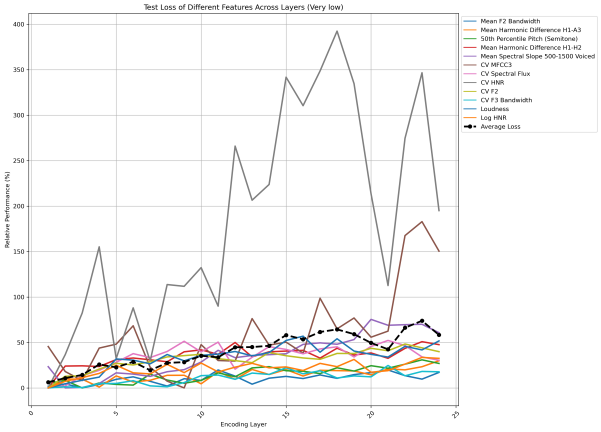

Figure 6: *Test Loss of Different Features Across Layers (Very Low Severity)*

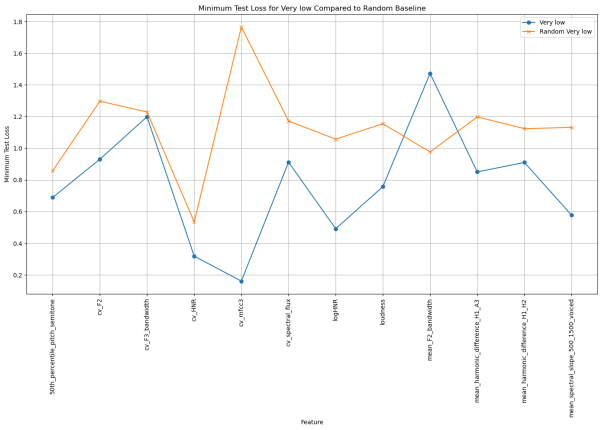

Figure 7: *Minimum Test Loss for Very Low Compared to Random Baseline*

4.2.3. Low Severity

The analysis of probing tasks for the Low severity level, as shown in Figure 8, indicates varied performance across different features. The relative performance percentage for different features across the encoding layers shows that while some features like "Mean F2 Bandwidth" demonstrate increasing test loss as we go deeper into the layers, others maintain a more stable loss profile. "Mean F2 Bandwidth" shows a peaking increase in loss particularly around Encoding Layers 10 and 20, suggesting that Whisper struggles to encode or loses this feature in these regions.

Interestingly, the average loss, represented by the black dashed line, increases steadily in the deeper layers. This trend suggests that Whisper's deeper layers capture more generalized features, potentially losing specific information necessary for distinguishing Low severity dysarthric speech.

To further validate these findings, the minimum test loss for Low severity probes against a random baseline is depicted in Figure 9. The results indicate that while most features show a noticeable difference between the actual and random probes, "50th Percentile Pitch Semitone," "CV HNR," and "CV MFCC3" could not surpass the baseline. These results suggest that Whisper's representation for these features does not capture meaningful information for Low severity speech or our probes struggle to learn effectively from the representations.

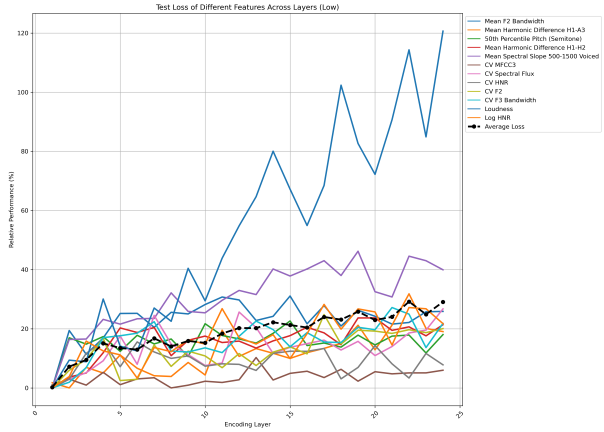

Figure 8: *Test Loss of Different Features Across Layers (Low)*

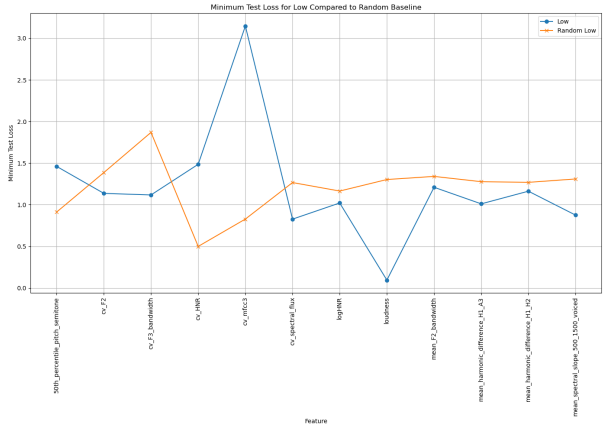

Figure 9: *Minimum Test Loss for Low Compared to Random Baseline*

4.2.4. Medium Severity

The analysis of probing tasks for the Medium severity level, as shown in Figure 10, indicates notable performance fluctuations across different features throughout Whisper's encoding layers. In the initial layers, the probes generally demonstrate better performance, with lower test loss values, indicating that Whisper's early layers capture more specific information relevant to severity dysarthric speech, correlating with the rest of the severity levels.

Notably, *"CV MFCC3"* exhibit dramatic peaks and valleys in test loss across the layers. It displays fluctuations with drastic peaks around Encoding Layers 5 and 15, and a sharp decline at Layer 21 where it reaches its minimal test loss. This suggests that while certain layers struggle to encode this feature effectively, others, such as Layer 21, capture it very well.

The average loss, represented by the black dashed line, also exhibits an increasing trend in the deeper layers, similar to other severity levels. As expected from the PCA analysis, the deeper layers show a reduction in the model's ability to discriminate between different severities. While these layers capture more abstract and generalized features, they seem to lose some of the detailed information crucial for distinguishing the severity of dysarthric speech.

The comparison between the minimum test loss for Medium severity probes and the random baseline is depicted

in Figure 11. Most features show a clear distinction between the actual probes and the random baseline, underscoring the effectiveness of Whisper's representations in encoding relevant information. However, *"CV F3 Bandwith"* the probes perform worse than the random baseline, indicating either inadequate learning from the representations or the absence of useful information in the encodings for these specific features.

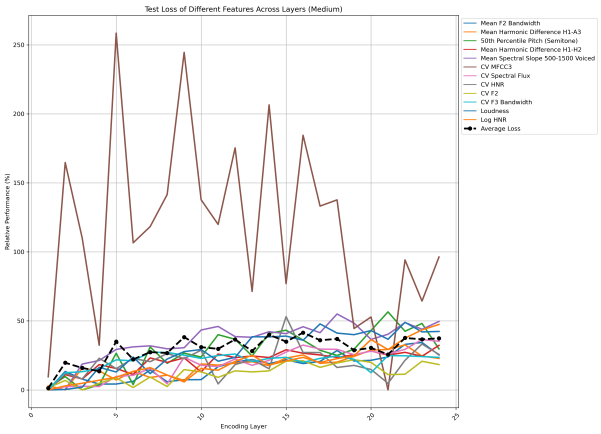

Figure 10: *Test Loss of Different Features Across Layers (Medium)*

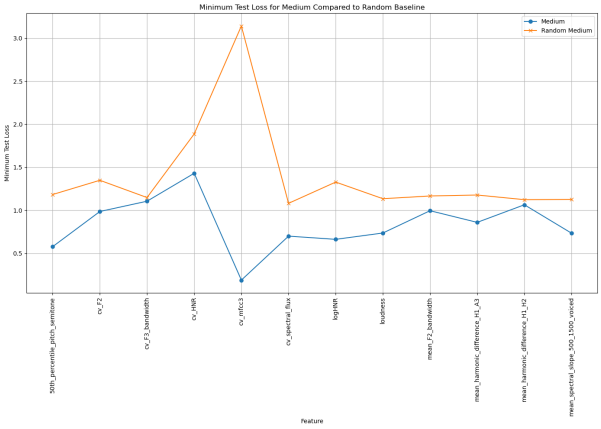

Figure 11: *Minimum Test Loss for Medium Compared to Random Baseline*

4.3. Evaluation and Discussion

To validate our probing analysis, we employed an SVM classifier to evaluate Whisper's internal representations for distinguishing dysarthric speech severity levels. The SVM results provide a downstream task perspective to see if our probe-based insights align with actual classification performance.

The SVM classification accuracy across different encoding layers is presented in Figure 12. The highest accuracy of approximately 0.988 was observed at Encoding Layer 3, while the lowest accuracy of around 0.906 was observed at Encoding Layer 22. This suggests that the initial layers capture more distinctive features relevant to classifying the severity of dysarthric speech.

The PCA visualizations and SVM accuracy results together illustrate that Whisper's initial layers maintain distinct and dispersed representations for each severity level, capturing detailed

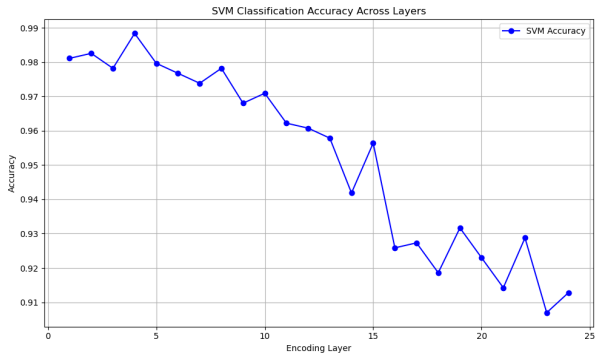

Figure 12: *SVM Classification Accuracy Across Layers.*

acoustic features that facilitate differentiation. However, as the representations move through deeper layers, they become more generalized, centralizing around "Normal" speech, and resulting in a decline in classification accuracy.

This trend concludes that deeper layers, while effective for general speech recognition, may lose specific information crucial for distinguishing different severities of dysarthric speech. The observed increase in average test loss for probing tasks in deeper layers further supports this, suggesting a shift from specific acoustic details to more abstract, high-level features.

In a broader scope for acoustic features, "CV MFCC3" shows minimal loss around layer 21 in Medium severity and layer 9 in Very Low severity. This indicates that "CV MFCC3" is better captured in the later encoding layers, highlighting its importance in representing severe dysarthric speech, which is less prevalent in normal speech. The variation in loss for "CV MFCC3" across severity levels suggests it captures unique characteristics of dysarthric speech that are less apparent in typical speech patterns. This feature's effective capture in deeper layers may indicate Whisper's ability to progressively refine and highlight these unique characteristics, crucial for distinguishing more severe dysarthric conditions. Conversely, the normalizing behavior of Whisper and fluctuating acoustic feature representations across severe speech display the need for an in-depth analysis of the reasons behind the variations.

Overall, the results indicate that Whisper's initial encoding layers are more effective at capturing specific acoustic features crucial for distinguishing between severity levels. This effectiveness diminishes in the deeper layers, where representations become more generalized. While this generalization benefits overall speech recognition, it poses challenges for tasks requiring fine-grained discrimination, such as identifying dysarthric speech severity. Despite this, the SVM's ability to distinguish severity levels even in deeper layers indicates that some discriminatory power is maintained, highlighting the intricate capabilities of Whisper's internal representations. Notably, Whisper demonstrates a strong zero-shot capability in encoding significant acoustic features pertinent to dysarthric speech, despite not being fine-tuned specifically for this task. This suggests that with fine-tuning, Whisper's performance in recognizing and differentiating the severity levels of dysarthric speech could see substantial improvements, making it a valuable tool for applications in medical settings.

5. Conclusion

Within this study we tried to demonstrate an interpretation for Whisper's performance of effectively capturing properties of dysarthric speech. With probing, we analyzed the presence of discriminative features spesific for dysarthic speech. This analysis revealed that Whisper's initial encoding layers are particularly adept at capturing acoustic features yet it still better represents some features like "CV MFCC3" in deeper layers. Despite the generalization in deeper layers that causes the incrasing trend in loss values, Whisper maintains some discriminatory power, as evidenced by the SVM classification results.

By diving into the internal workings of Whisper, we aimed to demystify the "black box" nature of transformer models, promoting transparency and safety in AI applications, especially in medical settings. Understanding how Whisper processes atypical speech enhances its reliability and trustworthiness, paving the way for its broader adoption in healthcare. These insights are crucial for advancing the field of ASR, making sophisticated AI technologies more interpretable and beneficial for real-world, critical applications.

6. Responsible Research

This research adheres to ethical standards and promotes transparency. All datasets used, specifically the TORGO database, are publicly available, ensuring no privacy or copyright issues. Our methodologies and experimental procedures are thoroughly documented to facilitate reproducibility. The Python code and scripts used in this study are available at [https://github.com/Oagaoglu/Probing-Whisper-for-Dysarthic-Speech/], allowing others to replicate and build upon our work. Proper citations are provided for all referenced studies. By sharing our findings and tools, we aim to contribute to ethical and transparent AI research, particularly in medical applications.

7. References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

[6] B. A. Al-Qatab and M. B. Mustafa, "Classification of dysarthric speech according to the severity of impairment: an analysis of acoustic features," *IEEE Access*, vol. 9, pp. 18 183–18 194, 2021.

[7] E. Castillo Guerra and D. Lovey, "A modern approach to dysarthria classification," in *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, vol. 3, 2003, pp. 2257–2260 Vol.3.

[8] Z. Qian and K. Xiao, "A survey of automatic speech recognition for dysarthric speech," *Electronics*, vol. 12, no. 20, p. 4278, 2023.

[9] S. Rathod, M. Charola, and H. A. Patil, "Noise robust whisper features for dysarthric severity-level classification," in *International Conference on Pattern Recognition and Machine Intelligence*. Springer, 2023, pp. 708–715.

[10] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020.

[11] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2020. [Online]. Available: http://dx.doi.org/10.1109/ICASSP40776.2020.9054458

[12] J. Shah, Y. K. Singla, C. Chen, and R. R. Shah, "What all do audio transformer models hear? probing acoustic representations for language delivery and its structure," *arXiv preprint arXiv:2101.00387*, 2021.

[13] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, "What do we need to build explainable ai systems for the medical domain?" *arXiv preprint arXiv:1712.09923*, 2017.

[14] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "Survey of explainable ai techniques in healthcare," *Sensors*, vol. 23, no. 2, p. 634, 2023.

[15] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language resources and evaluation*, vol. 46, pp. 523–541, 2012.

[16] K. T. Mengistu and F. Rudzicz, "Adapting acoustic and lexical models to dysarthric speech," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4924–4927.

[17] M. Soleymanpour, M. T. Johnson, R. Soleymanpour, and J. Berry, "Accurate synthesis of dysarthric speech for asr data augmentation," *arXiv preprint arXiv:2308.08438*, 2023.

[18] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," 2023.

[19] S. Haykin, *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994.

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

[21] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 1459–1462.

[23] A. F. Agarap, "Deep learning using rectified linear units (relu)," 2019.

[24] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," 2012.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017.

[26] K. P. F.R.S., "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.