doi:10.20944/preprints202101.0081.v1

# What all do audio transformer models hear? Probing Acoustic Representations for Language Delivery and its Structure

**Jui Shah**[*,1] **Yaman Kumar Singla**[*,1,2,3] **Changyou Chen**[3] **Rajiv Ratn Shah**[1]
[1]IIIT-Delhi [2]Adobe [3]State University of New York at Buffalo
jui.shah@midas.center, yamank@iiitd.ac.in, changyou@buffalo.edu, rajivratn@iiitd.ac.in

## Abstract

In recent times, BERT based transformer models have become an inseparable part of the 'tech stack' of text processing models. Similar progress is being observed in the speech domain with a multitude of models observing state-of-the-art results by using audio transformer models to encode speech. This begs the question of what are these audio transformer models learning. Moreover, although the standard methodology is to choose the last layer embeddings for any downstream task, but is it the optimal choice? We try to answer these questions for the two recent audio transformer models, Mockingjay and wave2vec2.0 . We compare them on a comprehensive set of language delivery and structure features including audio, fluency and pronunciation features. Additionally, we probe the audio models' understanding of textual surface, syntax, and semantic features and compare them to BERT. We do this over exhaustive settings for native, non-native, synthetic, read and spontaneous speech datasets.

## 1 Introduction

Since the advent of transformers in the computational linguistics field in 2017 (Vaswani et al., 2017), they have gained great attention across various domains for a wide variety of tasks. Tay et al. (2020) survey prominent transformer models, which have now become a formidable force in the tech stack in Natural Language Processing (NLP), Computer Vision and Reinforcement Learning. Their inherent property to facilitate parallel training makes it easier to train models on large datasets. These pre-trained models are then fine-tuned on a variety of user-specific downstream tasks, achieving state-of-the-art results. At the same time, recent research has started focusing on interpreting what these models learn that helps such a wide variety of downstream tasks. Besides, as more and more applications start relying on such models, it becomes all the more important to explain what these embeddings capture to check for potential flaws and biases that can affect a large number of applications. To this end, different research studies started probing language model embeddings for particular linguistic properties of interest. Belinkov et al. (2017) probed for part-of-speech language understanding, Hewitt and Manning (2019) for syntax, Peters et al. (2018) on morphology, Zhang et al. (2020) for scales and numbers, *etc.* However, progress in the audio domain has been very limited with only a few works (Raj et al., 2019; Alishahi et al., 2017; Belinkov and Glass, 2017).

Transformers have predominantly addressed the discrete data domain. Hence, NLP and vision fields oversaw a tremendous amount of work on transformer-based modelling. Speech being in the continuous domain, lagged behind. As one of the first models for this problem, vq-wav2vec (Baevski et al., 2019) proposed a 2 stage pipeline. It discretizes an input speech to a K-way quantized embedding space. This is similar to word tokens in NLP tasks. The embeddings are then extracted from a BERT-based model. However, this technique does not capture the context representation and dependencies across the time domain essential for continuous speech. Wav2vec2.0 (Baevski et al., 2020) addresses this issue by designing three subunits - the feature encoder, the transformer, and the quantization module (discussed in Section 5.2). These units convert the input audio to latent space embeddings via a contrastive task. This task involves selecting the correct quantized latent representation of the masked time steps from a distractor set.

Mockingjay (Liu et al., 2020) and AudioAL-

[*]Authors (listed in alphabetical order) contributed equally.

BERT (Chi et al., 2020) are other such transformer models. These are modified versions of BERT for the audio domain. They do not have an inbuilt feature extractor module. Hence, the former takes the input of 160-dim mel features and the latter takes in 160-dim fbank features. They share the same architecture with the difference being that AudioALBERT has shared parameters across the 12 encoder units, while for Mockingjay they are different.

These audio transformers have been applied over many diverse downstream speech-language processing (SLP) tasks with state-of-the-art results. Tasks such as phoneme classification (Graves and Schmidhuber, 2005), speaker recognition (Tian et al., 2020), automatic scoring (Grover et al., 2020), and sentiment classification (Tang et al., 2020) have shown promising results even with pre-trained transformers. This also begs the question as to what these transformer models are able to learn during the pretraining phase for the various evaluation tasks. The sentiment of the above inquiry is also conveyed by Prof. Ray Mooney's quip that the meaning of a whole sentence cannot be captured by a $&!#* vector (Conneau et al., 2018; Mooney, 2014).

In this paper, we make the following contributions. (1) We propose a detailed analysis of what the two recent transformer-based semisupervised audio encoder models, Mockingjay and wav2vec2.0, learn. We do this by implementing post hoc probing on the embeddings extracted from each of the intermediate units of the transformer models. We probe those embeddings using an extensive number of features (46 in total), each categorized by the linguistic property they probe. We do this for text-based, audio-based, vocabulary-based, fluency-based, and suprasegmental pronunciation-based features. The results help us lay out a map of the layers where a particular feature or category of features are learnt while also providing a metric of comparison between the two models. This measures what the models are learning on various linguistic tasks, which can then inform downstream tasks to use these models.

(2) We test the models for their representative effectiveness on different types of speech settings: native-read, native-spontaneous, and non-native-read. We find that for the most part native-spontaneous and non-native speech settings follow the result patterns for native-read dataset albeit with a worse performance. We find that in general, type of speakers matter lesser than the type of speech. Therefore, for both the models, we observe that non-native read speech performs better than spontaneous speech in general.

(3) Additionally, we identify the role of the feature extractor module in wav2vec2.0, which enables it to process raw input audio of $16Hz$ without any preprocessing. We find that the subsequent layers of the feature encoder can encode all features into increasingly dense and informative representation vectors without any “intelligent processing” on them.

(4) We compare the performance of the representations obtained by audio models and BERT on text features. This is the first to check the representative capacity of audio representations for the text captured by audio. We find that despite having no text-specific error metrics, the audio models are able to encode text well and are comparable to BERT on several parameters. We find that the dataset used to pre-train the audio models has a significant effect on the downstream performance. Surprisingly, while both wave2vec2.0 and Mockingjay outperform BERT on LibriSpeech (the dataset they were trained on), they underperform in other settings. Additionally, both models seem to learn surface-level text features (such as the number of nouns and pronouns) comparable to BERT.

We release our code, datasets and tools used to perform the experiments and inferences. To the best of our knowledge, this is the first attempt towards interpreting audio transformer models. The conclusion points out that the transformers are able to learn a holistic range of features which enable them to perform with great accuracy on various downstream tasks even while training solely on unlabeled speech.

## 2 Problem Definition and Data

Given the latent representations of a model's intermediate layers, we define the problem of probing that model for the knowledge of different linguistic features as a regression task. With the input as the intermediate-layer embeddings, the probing model is trained to map them to normalized feature values which are extracted from the data. The probe is a 3 layer fully connected neural network with the hidden layer having a ReLU activation and dropout to avoid over-fitting. The model dimensions are $(768, 128, 1)$ for all the intermediate layers of transformers and $(512, 128, 1)$ for the feature extractor. Adam optimizer and a learning rate of $0.0001$ is used. We compare the representative capacity of different embeddings on the basis of the loss values reported by the prober. Further, we take a randomly-initialized vector as a baseline to compare against all the 'intelligent' embeddings.

We test Mockingjay and wav2vec2.0 models on the following linguistic features: audio features (§3.1), fluency features (§3.2), pronunciation features (§3.3). Since spoken language can be considered as a combination of spoken words and language delivery, we compare representations from both models with BERT embeddings (Devlin et al., 2019) as well We compare them on the basis of text features (including syntactic, semantic and surface-level features) (§3.4). For comparing on text features, we perform two kinds of experiments: probing the models on text features extracted from original transcripts of all the audio datasets we have considered and probing on text features extracted from Wikipedia articles. For the latter experiments, we convert Wikipedia articles to speech using the Google text-to-speech model (Durette and Contributors, 2020) and use that to feed the generated audio to the audio transformer models.

Since the pre-trained audio transformer models are used in a variety of domains, we tested them on: LibriSpeech dataset, which is a dataset comprising of read speech by native English speakers (Panayotov et al., 2015), spontaneous native English speech by using Mozilla Common Voice dataset (Ardila et al., 2019), and speakers with English as their second language (L2 English speakers) by using L2-Arctic (Zhao et al., 2018).

In addition, we explore the role of the feature extractor in wav2vec2.0. We probe the multiple convolutional layers within the model to find out what the extractor learns that helps the model in learning from the raw input audio. We also include the post-projection layer, which is after the feature encoder but before the transformer that maps the dimensions of extracted features $(512)$ to that of the transformer $(768)$.

## 3 Features Probed

We assess the models on the knowledge of language delivery and content features. This set of features assesses both *what* was spoken and *how* it was spoken. Next, we describe each of the individual features considered.

### 3.1 Audio Features

Acoustic analysis of speech includes temporal and spectral analysis of audio waveforms. Hence, we measure the following features in this category: *Total duration, zero-crossing rate, energy entropy, spectral centroid, mean pitch, local jitter, local shimmer, and voiced to unvoiced ratio*. *Total duration* is a characteristic feature of the audio length that tells us about the temporal shape. The temporal feature *zero crossing rate* measures the rate at which a signal moves from positive to a negative value or vice-versa. It is widely used as a key feature in speech recognition and music information retrieval (Neumayer and Rauber, 2007; Simonetta et al., 2019). Energy features of audio are an important component that characterizes audio signals. We use *energy entropy* and the standard deviation of energy (*std_dev*) to evaluate the energy. *Spectral centroid* is used to characterise the spectrum by its centre of mass. Additionally, to estimate the quality of speech as perceived by the ear, we measure *mean pitch*. We also probe for frequency instability (*localJitter*), amplitude instability (*localShimmer*), and *voiced to unvoiced ratio*.

### 3.2 Fluency Features

The key features of fluency are: rate of speech, pauses, and length of runs between pauses. For measuring the rate of speech, we measure the speech rate (number of words per second in the total response duration) (*speaking_rate*) and articulation rate (number of words per second in the total articulation time, *i.e.,* the resulting duration after subtracting the time of silences and filled pauses from the total response duration) (*articulation_rate*) (Wood, 2001). Apart from these rates, pauses in speech are the second most observable feature to indicate disfluency (Igras-Cybulska et al., 2016). Therefore, we measure the duration, location and frequency of pauses as prototypical features. For this, we measure the number of filled pauses per second - (*filled_pause_rate*), *silence deviation* (absolute difference from the mean of silence durations) which along with the total duration of the audio helps to indicate the length of runs between the pauses (Möhle, 1984). This also serves an important indicator for fluency. Other such features include total number of silences (*general silence*), mean duration of silences (*mean_silence*), average silence per word (*SilenceRate1*), average silence per second (*SilenceRate2*) and number of long silence per word (*longpfreq*).

Furthermore, conversational fillers are a major source of disfluency. Sounds like *uh, um, okay, you know, etc*. are used by speakers to bring naturalness and fluency to their speech. The extent of fillers is an important feature to check for speech fluency. We use the average number of syllables in a word (*average_syllables_in_word*), the number of words with syllables greater than 2 (*wordsyll2*) and the repetition frequency (*repetition_freq*), for measuring this.

### 3.3 Pronunciation Features

The intelligibility, perceived comprehensibility, and accentedness of speech are impacted by phonemic errors (Derwing and Munro, 1997). Segmental pronunciation is judged based on the amount of listener effort with lower being the better. Hence, we probe the models for the following pronunciation characteristic features - the percentage, standard deviation, duration and Normalized Pairwise Variability Index (PVI) for vowels (*vowelPercentage, vowelDurationSD, vowelSDNorm, vowelIPVINorm*), consonants (*consontantPercentage, consontantDurationSD, consonantSDNorm, consonantIPVINorm*), and syllables (*syllableDurationSD, syllableSDNorm, syllablePVINorm*) . Moreover, we also study the presence of stress in the speech with the characteristic features of stress syllables distance mean (*stressDistanceMean*), and stress distance mean (*stressDistanceSyllMean*).

### 3.4 Text Features

We further divide the text features into three categories: Surface-level, Syntax-level, and Semantic-level text features (Conneau et al., 2018; Jawahar et al., 2019).

**Surface-Level:** The lexical diversity of spoken speech is an important metric to evaluate its quality (Read et al., 2006; Kumar et al., 2019). Surface features can be inferred by simply looking at the content spoken. It is basically a measure of how many different words are used in speech such as *total nouns, pronouns, adverbs, adjectives, verbs, conjunction, and determiners* along with *unique word count* and the average word complexity (*Word Complexity*).

**Syntax-Level:** Syntax is the key component of the grammatical structure of a sentence, which in turn is a key component of the communicative competence (Canale and Swain, 1980). We use the *depth of the syntax tree* constructed from the sentences spoken in each sound clip as a feature to evaluate the syntax content (Conneau et al., 2018; Jawahar et al., 2019).

**Semantic-Level:** The relationship between the words spoken and our comprehension of that spoken content falls into the domain of semantics. To produce meaning in a sentence, it is almost necessary for it to have a subject and a direct object that the subject addresses. The *number of subjects* and *number of direct objects* are hence in our set of features to evaluate the spoken content (Conneau et al., 2018; Jawahar et al., 2019).

## 4 Results and Discussion

We discuss the results obtained with mean squared error (MSE) as the comparison metric. The regression task of predicting the normalized features from time-averaged latent embeddings makes MSE a suitable metric of evaluation. By comparing the losses, we identify the learning patterns of the two representational learning-based audio transformer models - wave2Vec2.0 and Mockingjay for different audio and text features. We also draw out a comparison between the two. Furthermore, we study the role of the feature encoder module in wav2vec2.0 and include the post projection layer in it which transforms the 512 dimension to the encoder input dim of 768.

We first present the results for read speech of native English speakers on LibriSpeech. Then, we conduct experiments to see how the different layers of the models pretrained on native English speakers behave in various settings. For this, we then present the results for spontaneous speech of native English speakers using Mozilla Common Voice dataset and non-native read speech using the CMU L2-Arctic dataset.

### 4.1 Audio Features

**Native Read Speech:** Figures 1(a), and (e)[1] show the results obtained for audio features probed on wav2vec2.0 and Mockingjay respectively for the Librispeech dataset. It can be seen that the lowest loss for these features is obtained in the initial two layers for wav2vec2.0, whereas for Mockingjay, it

[1] Refer Tables 2 and 6 of Appendix for loss values

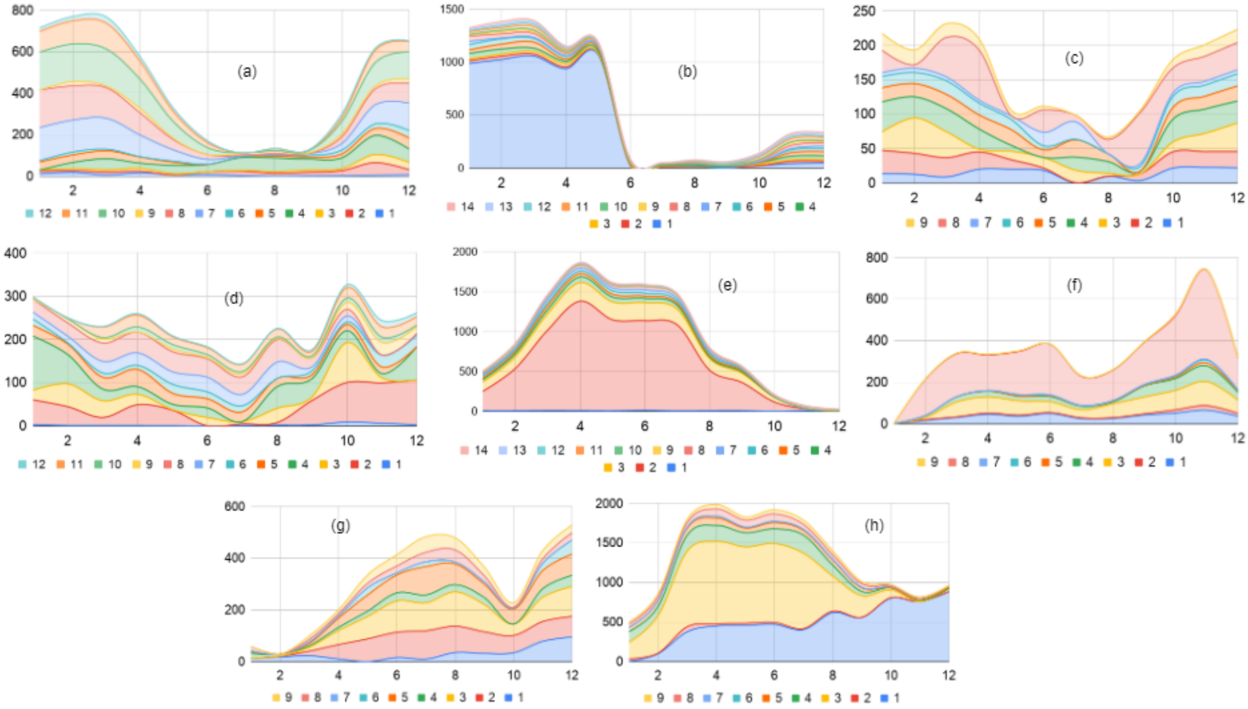

Figure 1: Performance of wave2vec2.0 on: (a) Fluency features (b) Pronunciation features (c) Surface features (g) Audio Features, Performance of Mockingjay on: (d) Fluency features (e) Pronunciation features (f) Surface features (h) Audio Features The graphs are the stacked area charts with the x-axis being the layers of the model and y-axis - the relative performance of each layer with respect to the maximum loss for each feature *((loss - min_loss)*100%/min_loss)*. Hence, higher the value, higher the loss, lower the performance. Different colours symbolize the curves followed by the different features across the transformers for **Audio features** (*total duration, stdev energy, mean pitch, voiced to unvoiced ratio, zero crossing rate, energy entropy, spectral centroid, localJitter, localShimmer*), **Fluency features** (*filled pause rate, general silence, mean silence, silence abs deviation, SilenceRate1, SilenceRate2, speaking rate, articulation rate, longpfreq, average syllables in words, wordsyll2, repetition freq*), **Pronunciation features** (*StressedSyllPercent, StressDistanceSyllMean, StressDistanceMean, vowelPercentage, consonantPercentage, vowelDurationSD, consonantDurationSD, syllableDurationSD, vowelSDNorm, consonantSDNorm, syllableSDNorm, vowelPVINorm, consonantPVINorm, syllablePVINorm*) and **Text features** (*Total adjectives, Total adverbs, Total nouns, Total verbs, Total pronoun, Total conjunction, Total determiners, Unique Word count, Word Complexity*).

is the final layers. We can see a clear ascent in the losses as we traverse the table for wav2vec2.0 from left to right, *i.e.*, from lower layers to the higher layers. This suggests that as we go deeper into the 12 block transformer model the audio features are diluted by wav2vec2.0. Mockingjay on the other hand, follows a negative slope (meaning an increasing line) for its losses as we traverse through the layers. Hence, the audio features are best captured in the final layers of the Mockingjay model.

When we compare the minimum losses across both models, the average learning of these features for wave2vec is better than that of Mockingjay by $28.59\%$. Even when we compare the final layer embedding performance, wav2vec2.0 performs better than Mockingjay by $24.53\%$. This is interesting given that the final layer of wav2vec2.0 contains the most diluted version of the learnt features and Mockingjay has its best version (in the final layers).

**Native Spontaneous Speech:** Comparing the performance of native read and spontaneous speeches (Figure 2[2]), we observe that wav2vec2.0 performs better than Mockingjay for both. Wav2vec2.0, on average, performs better by $41.69\%$ when compared across the best performing layers and $51.12\%$ when end layer losses are compared. The pattern of the best performing layer also remains the same as the case of native read speech. Mockingjay learns these features best in the second last (11th layer) for native spontaneous speech, while it was the last two layers for native read speech. For wav2vec2.0, native read speech was best captured in the initial 2 layers, but for spontaneous speech, the layers are a bit more spread out across the initial half of the transformer model. We also observe that the loss values on native spontaneous speech are higher than the ones for native read and non-native read corpora.

**Non-native Speech:** When tested on L2 speakers (Figure 2[3]), wav2vec2.0 outperforms Mockingjay by $9.53\%$ and $12.51\%$ on minimum and end layer loss respectively. Additionally, similar to the

[2] Refer Tables 18 and 20 of Appendix for loss values
[3] Refer Tables 10, 14 of Appendix for loss values

doi:10.20944/preprints202101.0081.v1

case of native read speech, Mockingjay learns the audio features best in the final layers (11th and 12th layers). As for wav2vec2.0, the layers learning the audio features are spread out with the initial half of the model learning them more accurately than the later half.

**Feature Extractor:** The feature extractor of wav2vec2.0 shows a considerable decrease in loss with each passing layer for the audio features. Hence, we can infer that the feature extractor extracts the audio-based features to a great extent. Moreover, it is this propagation of high quality learnt features to the transformer module which makes it possible for the minimum loss in the wav2vec transformer module to appear in the initial 2 layers itself. In comparison, Mockingjay needs to go deeper to extract the same features.

### 4.2 Fluency features

**Native Read Speech:** For fluency based features on native read speech, we see that similar to audio features, wave2vec2.0 performs better than Mockingjay (Figures 1 (b), and (f)[4]). While the fluency features are not layer specific but are spread across the model for Mockingjay, they tend to show the best performance in the middle layers for wav2vec2.0. While comparing the final layer embeddings for both models, wav2vec2.0 performs better than Mockingjay by 12.23%. The performance gap increases by four folds to 42.37% when compared on the minimum losses (among all observed for the intermediate layers) learnt by both models.

**Non-native Speech:** For the L2 Arctic dataset (Figure 2[5]), the learning of fluency features is concentrated in the middle layers for wav2vec2.0. Moreover, here we see a definite pattern that Mockingjay is learning better in the final layers compared to the no pattern observed in the case of Librispeech. Overall, wav2vec2.0 outperforms Mockingjay by 5.06% on the minimum loss layers but by 105.62% for the final layers. Thus, wave2vec2.0 heavily outperforms Mockingjay on non-native speech settings.

**Feature Extractor:** Results from the feature extractor module of wav2vec2.0 show the minimum loss on the post-projection layer *i.e.*, the layer giving input to the transformer located after the feature extractor. The fluency features are extracted a better way as we go deeper in the extractor. This decreasing trend is carried on in the transformer block as well (as explained above).

### 4.3 Pronunciation features

**Native Read Speech:** Figures 1(c) and (g)[6] show the results for probing pronunciation features on wav2vec2.0 and Mockingjay with the Librispeech data. These features are learnt best by the last layers in Mockingjay. Wav2vec2.0 learns these features the most in the 6th to 8th layers amongst its 12 layers. Mockingjay performs better for pronunciation-based features than wav2vec2.0 by 30.4% in the final layer embeddings. Comparing the minimum loss layers for both models, the difference is 16.19% in favor of Mockingjay.

**Non-native Speech:** Mockingjay follows the same pattern for L2 Arctic dataset as for the Librispeech dataset. It learns these features better in the last layers. However, for wav2vec2.0, the layers learning each of these pronunciation features are more spread out across the initial layers of the second half of the model. Wav2vec2.0 outperforms Mockingjay but the differences here are reduced to 8.9% in the end layer and 2.20% in the best performing layer. This pattern follows the non-native speech performance of wave2vec2.0 and Mockingjay seen with audio and fluency features. Here too, the performance difference between wave2vec2.0 and Mockingjay widens when compared to the native speech scenario.

**Feature Extractor:** While the slope of the loss graph is negative for the extractor module of wav2vec2.0, the minimum loss observed is after the last convolutional unit of the extractor and not in the post-projection layer. However, here too we observe a similar pattern as audio and fluency features, showing that the feature extractor module indeed is performing the way it is supposed to.

### 4.4 Text based features

As discussed in the Section 3.4, the quality of spoken content can be broken down into its performance into three subcategories of text-based features. We evaluate the model's learning in these categories to identify which aspects of the conveyed text do these audio transformers learn. This helps us to evaluate the speech content knowledge of speech transformer models. Here, apart from comparing Mockingjay and wave2vec2.0's perfor-

[4] Refer Tables 3 and 7 of Appendix for loss values

[5] Refer Tables 11 and 15 of Appendix for loss values

[6] Refer Tables 4 and 8 of Appendix for the loss values

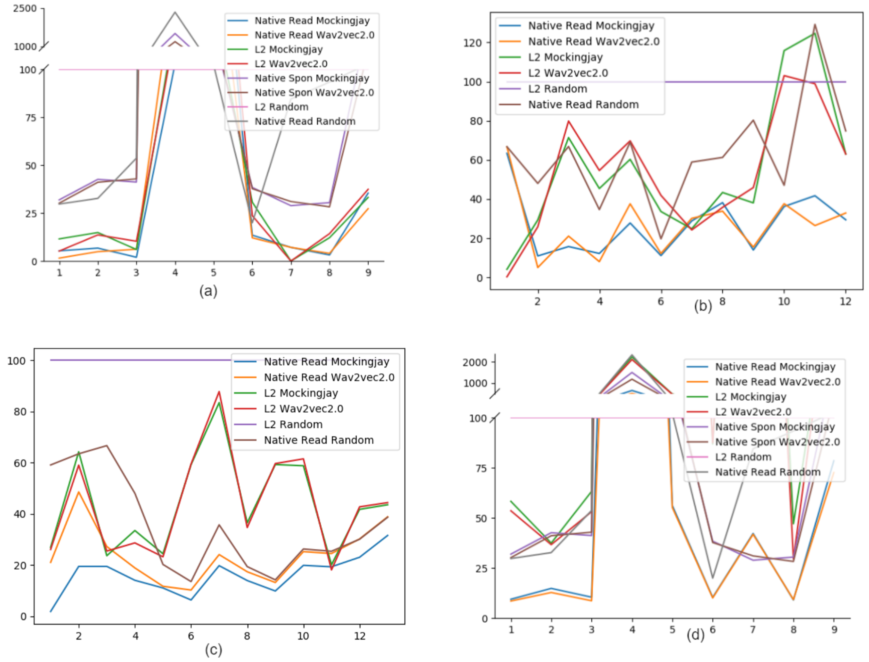

Figure 2: Performance of (a) Audio Features (b) Fluency features (c) Pronunciation features (d) Surface features across all datasets by both Mockingjay and wave2vec2.0 . The graphs show the performance of each feature of a particular category (on the x-axis) wrt to the the performance of random embeddings on L2 Arctic data features *(loss*100/l2_random_loss)* on the y-axis. All the features for a particular category follow the same order and label as that of 1 with the exception of Pronunciation, which has one less feature (the first one - *StressedSyllPercent*).

mance, we also compare them with the text transformer model, BERT. We do this to check how does the content knowledge of the speech transformer models compare to the text ones.

**Surface level features:** When compared on LibriSpeech, vocabulary-based features are learnt better by wav2vec2.0 than Mockingjay by 7.83% Figures 1(d) and (h)[7]. These features are learnt best in the intermediate layers in wav2vec2.0 and initial layers in Mockingjay. From the results, we observe that the text understanding of both models becomes increasingly diffused as we go towards the later layers. Wav2vec2.0 outperforms Mockingjay by 9.70% in the final layer. For L2 arctic data, the difference widens to 10.23% on the end layers (Figure 2). However, the difference reduces to just 2.49% for native spontaneous speech.

**Semantic features:** Surprisingly, despite performing worse on surface features, Mockingjay performs better than wav2vec2.0 in this setting by 16.83% (Figure 2). Additionally, while wav2vec2.0 outperforms Mockingjay be 9.35% on minimum layer loss for L2 speech, but Mockingjay performs better by 3.17% on the end layer.

**Syntax features:** In this setting as well, Mockingjay performs better than wav2vec2.0 on end layer by 21.5% (Figure 2). But for L2 arctic data, both perform almost similar with a difference of 2.8% on in favor of wave2vec2.0 on the last layer.

**Feature Extractor:** The pattern observed in the feature extractor module for these surface level features is the same as that of audio features with minimum losses seen in the post projection layer. However, the noticeable fact that the value of the minimum loss in this layer is less than that of the minimum loss in the transformer module of wav2vec2.0. This gives some intuition for the better performance of Mockingjay over wav2vec2.0 since the transformer is unable to capture the features or unlearns the presented vocabulary features.

### 4.4.1 Comparison with BERT

When we compare the performance of audio-transformer models with BERT (Table 1) on the native read speech, we observe that on an average, both wave2vec2.0 and Mockingjay perform better than BERT by 41.54% and 48.77%% on surface features, 56.90% and 73.57% on syntactic features and 36.35% and 53.25% on semantic features respectively. These results are surprising since none of the speech transformer models was trained with text objective functions. We hypothesize that this could be due to differences in the train set of the three models. LibriSpeech is the train-set for both the speech-transformer models

[7] Refer Tables 5 and 9 of Appendix for the loss values

where as Wikipedia is the train-set for BERT. To confirm this, we test the performance of the three models on text features extracted from Wikipedia and native spontaneous speech datasets. These datasets provide us with a comprehensive comparison. While on one hand, Wikipedia is the train-set for BERT, and the text features from Wikipedia articles are very different from LibriSpeech, on the other, native spontaneous speech dataset can be considered out-of-domain for both the speech transformer models and BERT.

For the first part, we convert 2000 random sentences from Wikipedia articles to speech by using Google's text-to-speech API (Durette and Contributors, 2020). We made sure that the audios constructed had similar lengths as those of LibriSpeech. The audios so obtained were then passed through both the speech transformer models and the layers were then probed. On this synthetic dataset, for the semantic features, BERT outperforms both the models by more than 50% when compared on minimum loss across all the layers. However, by the end layers, both the models learn the features well and the performance difference between BERT and audio-transformer models reduces greatly ($0.04\%$ and $0.92\%$ difference for surface features, $4.41\%$ and $7.68\%$ for syntax and $7.58\%$ and $15.56\%$ for semantic features). These results are motivating since this means that audio transformer models' embeddings do not just capture audio, fluency and pronunciation features but also textual features to a large extent.

For the second part of the experiments, we use the CMU L2 Arctic dataset. Table 1 presents the results for all the experiments. Here the results are the most different from the previous ones. While the performance difference between BERT and the audio models is still comparable for surface features, for the syntax and semantical features, BERT outperforms both the models by more than $15\%$. This result when compared with Wikipedia TTS and native read speech implies that the audio models capture text features for native speakers in 'cleaner settings' but they are not able to work in not-so controlled environments. Therefore, in a general setting, BERT text embeddings combined with audio embeddings can capture all the speech features adequately.

# 5 Related Work

## 5.1 Audio Probing

In the domain of speech processing, probes have been carried out on feature vectors (Raj et al., 2019), neural networks like RNN (Alishahi et al., 2017), end-to-end ASR systems (Belinkov and Glass, 2017) and audio-visual models (Drexler and Glass, 2017). In Raj et al. (2019), probing on x-vectors trained solely to predict the speaker label, revealed they also contain incidental information about the transcription, channel, or meta-information about the utterance. Probing the Music Information Retrieval (MIR) prediction through Local Interpretable Model-Agnostic Explanations (LIME) by using AudioLIME (Haunschmid et al., 2020) helped interpret MIR for the first time. They demonstrated that the proposed AudioLIME produces listenable explanations that creates trustworthy predictions for music tagging systems. Nagamine et al. (2015) analyse a DNN for phoneme recognition, both at single node and population levels. Further research on interpretation of the role of non-linear activation of the nodes of a sigmoid DNN built for phoneme recognition task is done in (Nagamine et al., 2016). Research has also been done to address why LSTMs work well as a sequence model for statistical parametric speech synthesis (Wu and King, 2016). Several other studies have been conducted to interpret the correlation between audio and image structures for audio-visual tasks (Drexler and Glass, 2017; Harwath and Glass, 2017). Even for Deep ASR models, efforts have been made to comprehend the hidden and learned representations (Belinkov and Glass, 2017; Elloumi et al., 2018). However, probing of representation learning audio transformers is yet unexplored.

## 5.2 Wav2vec2.0

Wav2vec2.0[8] (Baevski et al., 2020) is a transformer architecture which learns the latent speech representations jointly via a defined contrastive task over their quantization. It is comprised of 3 major components - the feature encoder, the transformer and the quantization module. The feature encoder consists of a multi-layer convolutional network which converts the raw input audio $X$ to latent representation $Z_1, Z_2, .., Z_T$ which are fed

[8] We use the terms wave2vec and wave2vec2.0 interchangeably in the text

| Dataset | Model Comparison | Surface | Syntax | Semantics |
|---|---|---|---|---|
| Native Read Speech | Wave2Vec2.0 | 48.77%, 40.91% | 56.90%, 67.15% | 36.35%, 52.73% |
| | Mockingjay | 41.54%, 36.73% | 73.57%, 74.21% | 53.25%, 60.68% |
| Non-native Read Speech | Wave2Vec2.0 | -28.72%, -2.51% | -59.05%, -30.33% | -20.49%, -17.83% |
| | Mockingjay | -44.37%, -12.75% | -79.72%, -33.97% | -31.39%, -14.07% |
| Wikipedia TTS | Wave2Vec2.0 | -0.56%, -0.04% | -34.55%, 4.41% | -52.36%, -7.58% |
| | Mockingjay | 0.73%, 0.92% | -47.70%, 7.68% | -63.18%, -15.56% |

Table 1: Table for comparison of the performance of BERT with wave2vec2.0 and Mockingjay on text features. The two values mentioned per cell indicate the relative minimum loss across all the model layers and the relative end layer losses when compared with the corresponding values for BERT. The values shown are an average across all features of a particular category with the relative performance calculated as $(bert\_loss - model\_loss) * 100\% / bert\_loss$

into the transformer to build the representations $C_1, C_2, ...C_n$. The targets in the self-supervised objective $q_t$ are built by passing the output of feature encoder to the quantizater. The training is similar to that of BERT (Devlin et al., 2019). Certain time-steps in the latent feature representation are masked and the contrastive task requires finding the correct quantized latent audio representation for those masked steps from a set of distractors. The model is pretrained on unlabeled Librispeech data (Panayotov et al., 2015) and then finetuned on the TIMIT (Garofolo et al., 1993) dataset.

It achieves a 1.8/3.3 WER on the clean/noisy test sets on experiments using all labeled data of Librispeech and 5.2/8.6 WER on the noisy/clean test sets of Librispeech using just ten minutes of labeled data. The authors claim that even while lowering the amount of labeled data to one hour, wav2vec 2.0 outperforms the previous state of the art on the 100 hour subset while using 100 times less labeled data.

All our experiments are based on the wav2vec 2.0 base model in which the feature encoder contains 7 blocks having a temporal convolution of 512 channels with strides $(5, 2, 2, 2, 2, 2, 2)$ and kernel widths $(10, 3, 3, 3, 3, 2, 2)$ respectively and the there are 12 transformer blocks with a model dimension of 768, inner dimension (FFN) $3,072$ and 8 attention heads.

### 5.3 Mockingjay

Mockingjay (Liu et al., 2020) is a bidirectional transformer model which allows representation learning by joint conditioning on past and future frames. It accepts input as 160 dimension log-Mel spectral features and has outperformed it for phoneme classification, speaker recognition and sentiment discrimination accuracy on a spoken content dataset by 35.2%, 28.0% and 6.4% respectively. The authors claim the model is capable of improving supervised training in real world scenarios with low resource transcribed speech by presenting that the model outperforms other existing methods while training on 0.1% of transcribed speech as opposed to their 100%.

For our experiments, we use the MelBase-libri model. The architecture comprises of 12 encoder layers and each unit has 0he same output dimension of 768 and comprises of sub-layers which include a feed-forward layer of size 3072 and 12 self-attention heads. We probe each of the 12 transformer blocks of both models and the feature encoder of wav2vec2.0 to check if they learn the features of audio, fluency, suprasegmental pronunciation and text.

## 6 Conclusion

Transformer models across multiple domains such as natural language processing and computer vision now form an unavoidable part of the tech stack. Audio transformers that exploit representational learning to train on unlabeled speech have recently been suggested. It is observed that when these pre-trained models are fine-tuned on labelled speech, they perform highly efficiently for different downstream tasks such as phoneme recognition, character prediction, etc. In this paper, we interpret two such models, wav2vec2.0 and Mockingjay, to understand what each of their unit focuses on learning in their architecture. We show that the models are capable of significantly capturing a wide range of characteristics such as audio, fluency, suprasegmental pronunciation, and text-based characteristics (further divided into semantic, syntax, and surface characteristics). For each category of characteristics, we identify a learning pattern for each framework, and conclude which model and which layer of that model is better for a specific category of feature to choose for downstream tasks. We find that wave2vec2.0 outperforms Mockingjay on audio and fluency features but underperforms on pronunciation features. Further, we compare text-BERT with the audio models on text features and find that the audio models surprisingly outperform BERT in cleaner, controlled settings (native read speech and synthetic speech in native voice) but are not able to perform in uncontrolled environment such as of spontaneous speech or even non-native speech. We show our results on a variety of settings including native, non-native, read, spontaneous and synthetic speech datasets.

## References

Afra Alishahi, Marie Barking, and Grzegorz Chrupała. 2017. Encoding of phonology in a recurrent neural model of grounded speech. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 368–378.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Steffen Schneider, and Michael Auli. 2019. vq-wav2vec: Self-supervised learning of discrete speech representations. In *International Conference on Learning Representations*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? *arXiv preprint arXiv:1704.03471*.

Yonatan Belinkov and James Glass. 2017. Analyzing hidden representations in end-to-end automatic speech recognition systems. In *Advances in Neural Information Processing Systems*, pages 2441–2451.

Michael Canale and Merrill Swain. 1980. Theoretical bases of communicative approaches to second language teaching and testing. *Applied linguistics*, 1(1):1–47.

Po-Han Chi, Pei-Hung Chung, Tsung-Han Wu, Chun-Cheng Hsieh, Shang-Wen Li, and Hung-yi Lee. 2020. Audio albert: A lite bert for self-supervised learning of audio representation. *arXiv preprint arXiv:2005.08575*.

Alexis Conneau, Germán Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single vector: Probing sentence embeddings for linguistic properties. *arXiv preprint arXiv:1805.01070*.

Tracey M Derwing and Murray J Munro. 1997. Accent, comprehensibility and intelligibility: Evidence from four l1s. *Studies in Second Language Acquisition*, 19(1):1–16.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Drexler and James Glass. 2017. Analysis of audio-visual features for unsupervised speech recognition. In *Grounded Language Understanding Workshop*.

Pierre Nicolas Durette and Contributors. 2020. Google text to speech model. `https://pypi.org/project/gTTS/`.

Zied Elloumi, Laurent Besacier, Olivier Galibert, and Benjamin Lecouteux. 2018. Analyzing learned representations of a deep asr performance prediction model. In *Blackbox NLP Workshop and EMLP 2018*.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. 1993. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *STIN*, 93:27403.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610.

Manraj Singh Grover, Yaman Kumar, Sumit Sarin, Payman Vafaee, Mika Hama, and Rajiv Ratn Shah. 2020. Multi-modal automated speech scoring using attention fusion. *arXiv preprint arXiv:2005.08182*.

David Harwath and James Glass. 2017. Learning word-like units from joint audio-visual analysis. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 506–517.

Verena Haunschmid, Ethan Manilow, and Gerhard Widmer. 2020. audiolime: Listenable explanations using source separation. *arXiv preprint arXiv:2008.00582*.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Magdalena Igras-Cybulska, Bartosz Ziółko, Piotr Żelasko, and Marcin Witkowski. 2016. Structure of pauses in speech in the context of speaker verification and classification of speech type. *EURASIP Journal on Audio, Speech, and Music Processing*, 2016(1):18.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Yaman Kumar, Swati Aggarwal, Debanjan Mahata, Rajiv Ratn Shah, Ponnurangam Kumaraguru, and Roger Zimmermann. 2019. Get it scored using autosas—an automated system for scoring short answers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9662–9669.

Andy T Liu, Shu-wen Yang, Po-Han Chi, Po-chun Hsu, and Hung-yi Lee. 2020. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6419–6423. IEEE.

Dorothea Möhle. 1984. A comparison of the second language speech production of different native speakers. *Second language productions*, 26:49.

Ray Mooney. 2014. You can't cram the meaning of a whole %&!$# sentence into a single $&!#* vector! `https://www.cs.utexas.edu/~mooney/cramming.html`.

Tasha Nagamine, Michael L Seltzer, and Nima Mesgarani. 2015. Exploring how deep neural networks form phonemic categories. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Tasha Nagamine, Michael L Seltzer, and Nima Mesgarani. 2016. On the role of nonlinear transformations in deep neural network acoustic models. In *Interspeech*, pages 803–807.

Robert Neumayer and Andreas Rauber. 2007. Integration of text and audio features for genre classification in music information retrieval. In *European Conference on Information Retrieval*, pages 724–727. Springer.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur. 2019. Probing the information encoded in x-vectors. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 726–733. IEEE.

John Read, Paul Nation, et al. 2006. An investigation of the lexical dimension of the ielts speaking test. *IELTS research reports*, 6:207–231.

Federico Simonetta, Stavros Ntalampiras, and Federico Avanzini. 2019. Multimodal music information processing and retrieval: Survey and future challenges. In *2019 International Workshop on Multilayer Music Representation and Processing (MMRP)*, pages 10–18. IEEE.

Hao Tang, Donghong Ji, Chenliang Li, and Qiji Zhou. 2020. Dependency graph enhanced dual-transformer structure for aspect-based sentiment classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6578–6588.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2020. Efficient transformers: A survey.

Zhengkun Tian, Jiangyan Yi, Ye Bai, Jianhua Tao, Shuai Zhang, and Zhengqi Wen. 2020. Synchronous transformers for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7884–7888. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

David Wood. 2001. In search of fluency: What is it and how can we teach it? *Canadian Modern Language Review*, 57(4):573–589.

Zhizheng Wu and Simon King. 2016. Investigating gated recurrent networks for speech synthesis. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5140–5144. IEEE.

Xikun Zhang, Deepak Ramachandran, Ian Tenney, Yanai Elazar, and Dan Roth. 2020. Do language embeddings capture scales? *arXiv preprint arXiv:2010.05345*.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. 2018. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech*, page 2783–2787.

| Audio | 0th layer | 1stlayer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| total_duration | 0.001000995591 | 0.001052616374 | 0.001098873668 | 0.0009882489029 | **0.0008932195598** | 0.001036955389 | 0.0009853900297 | 0.00120506479 | 0.001173329026 | 0.001198166304 | 0.001597253794 | 0.001742114511 |
| stdev_energy | **0.002306532136** | 0.002380363624 | 0.002714319057 | 0.003584466838 | 0.00434291883 | 0.004559425422 | 0.004828923042 | 0.00466799069 | 0.004243278617 | 0.003855743693 | 0.004065452462 | 0.004142789993 |
| mean_pitch | **0.001811751374** | 0.001820682815 | 0.002083195027 | 0.002789952571 | 0.00337429036 | 0.003992556246 | 0.003757578298 | 0.004210399257 | 0.003714940709 | 0.00260877705 | 0.003510799603 | 0.003896026832 |
| voiced_to_unvoiced_ratio | 0.00223168878 | 0.002014200847 | 0.001971397127 | 0.002092458664 | 0.00224987847 | 0.002456693423 | 0.002439862854 | 0.002375320708 | 0.002359549305 | **0.00186251703**7 | 0.002476975146 | 0.002676547201 |
| zero_crossing_rate | 0.004400080311 | **0.00415155569**7 | 0.004536247542 | 0.005637626985 | 0.006736799315 | 0.007006201082 | 0.008680683359 | 0.007472786243 | 0.006308502927 | 0.006545616986 | 0.007058227951 | 0.007532593417 |
| energy_entropy | 0.004003145208 | **0.003851908827** | 0.004065196172 | 0.004143885762 | 0.004913005053 | 0.004165694571 | 0.004565524077 | 0.004139930451 | 0.004200284971 | 0.004021279313 | 0.00492770492 | 0.00593517518 |
| spectral_centroid | 0.0003349381924 | **0.0003348789101** | 0.0003349927652 | 0.0003349526704 | 0.0003349562986 | 0.0003349558744 | 0.0003350423341 | 0.0003352064503 | 0.0003349179998 | 0.0003349463851 | 0.0003350426625 | 0.0003350045659 |
| localJitter | 0.002260860691 | **0.001933041139** | 0.002130917682 | 0.002110966278 | 0.002272641959 | 0.002445947651 | 0.00263751574 | 0.002834154751 | 0.002378949972 | 0.001995716744 | 0.002267349635 | 0.002488864739 |
| localShimmer | **0.003058716087** | 0.00305748341 | 0.003449485296 | 0.003517129689 | 0.004046810921 | 0.004389293538 | 0.004901103827 | 0.00447761415 | 0.004076987467 | 0.003549565677 | 0.004036656624 | 0.00396044435 |

Table 2: Results (MSE) for audio features on wav2vec2.0 for native read speech corpus (Librispeech)

| Fluency | 0th layer | 1st layer | 2nd layer | 3rd layer | **4rth layer** | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| filled_pause_rate | 0.0008765513876 | 0.0009303065029 | 0.0008303899183 | 0.0009168531955 | 0.0008311720524 | 0.0008384440081 | 0.0008098078354 | **0.0007824271691** | 0.000794454555 | 0.0008019173777 | 0.0008153088127 | 0.0008293819283 |
| general_silence | 0.001895778607 | 0.00180526613 | 0.0019404146 | 0.001794739728 | **0.001683760422** | 0.001923729788 | 0.00203110742 | 0.001936706927 | 0.001979602653 | 0.002097983946 | 0.002721516396 | 0.002111513689 |
| mean_silence | 0.001807076578 | 0.001907845938 | 0.001890766051 | 0.00195921772 | 0.001820755066 | **0.001723015355** | 0.001787426496 | 0.001845146334 | 0.001906279179 | 0.001886272296 | 0.002394365884 | 0.002327627502 |
| silence_abs_deviation | **0.0009753903965** | 0.001266490641 | 0.001489715954 | 0.001221203025 | 0.001371391982 | 0.001315605468 | 0.00158018127 | 0.001617509273 | 0.001493427956 | 0.001483640437 | 0.001869316284 | 0.001599149844 |
| SilenceRate1 | 0.005096142174 | 0.004996513698 | 0.005074212019 | 0.004758236756 | 0.004217072683 | **0.003675997698** | 0.00383899521 | 0.004035006432 | 0.004022787603 | 0.00436030715 | 0.004927401867 | 0.005626966203 |
| SilenceRate2 | 0.00516018685 | 0.005519745856 | 0.005247613608 | 0.004932789173 | 0.005084989392 | 0.004940673442 | **0.00484461307** | 0.005111577654 | 0.005173505513 | 0.005739863698 | 0.005895325256 | 0.006605277493 |
| speaking_rate | 0.013042829 | 0.01278397844 | 0.01249306561 | 0.01023916275 | 0.007733273967 | 0.006184341667 | 0.005215829595 | **0.005029336262** | 0.005487236958 | 0.006622599568 | 0.009678859003 | 0.01164027247 |
| articulation_rate | 0.01682406134 | 0.01586561525 | 0.0147932579 | 0.01239362625 | 0.008916872238 | 0.00737402545 | 0.006135157085 | 0.006320548084 | **0.006001218728** | 0.007957721356 | 0.01058859492 | 0.01172303029 |
| longpfreq | 0.001641537146 | 0.001979431651 | 0.00177438887 | 0.00198161383 | 0.001731098501 | 0.001646489692 | 0.001797643877 | 0.001868383728 | 0.001697789939 | 0.001848310577 | 0.001862609234 | 0.001994732835 |
| average_syllables_in_words | 0.01831333714 | 0.0182145931 | 0.01801639705 | 0.01556206624 | 0.01166978505 | 0.008615066344 | 0.00665224039 | **0.006485827068** | 0.007123317853 | 0.01045775345 | 0.01383351046 | 0.01486869956 |
| wordsyll2 | 0.01010911856 | 0.01072634095 | 0.0113568104 | 0.009438352727 | 0.00750566223 | 0.006293300612 | 0.005057610628 | **0.00555890078** | 0.005261181218 | 0.005987681059 | 0.008059176219 | 0.007614122738 |
| repetition_freq | 0.01558559961 | 0.01603584763 | 0.01725062927 | 0.01613746001 | 0.01541195257 | 0.01444718251 | 0.013343901 | 0.01352003816 | **0.01331795615** | 0.01555349831 | 0.01470347635 | 0.01371891228 |

Table 3: Results (MSE) for fluency features on wav2vec2.0 for native read speech corpus (Librispeech)

# Appendices

## A Details

For LibriSpeech, we take the default 'train-clean-100' set for training the probing model and the test-clean set for testing it. For the L2 dataset, we take 500 audios of 4 speakers each from the CMU L2 ARTIC dataset for training. Testing is done on 50 audios of each of those speakers. The 4 speakers are selected in such a way that there is 1 male and 1 female speaker with their L1 language as Hindi and Spanish. For Mozilla Common Voice database, we use a subset of 2000 random audios for training and 200 for testing. For evaluating. Similar to previous work on text-probing (Jawahar et al., 2019), we evaluate the performance of embeddings on the mean-square error for each encoder unit of both the transformer models.

## B Detailed Results

| **Pronunciation** | 0th layer | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | **6th layer** | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| StressedSyllPercent | 0.01890501628 | 0.01956094468 | 0.02011212084 | 0.01804626824 | 0.02036499647 | 0.002052113536 | 0.001922684484 | 0.00194589348 | **0.001740619761** | 0.002095186832 | 0.002428855659 | 0.002546980326 |
| StressDistanceSyllMean | 0.009509657533 | 0.009830785929 | 0.009026296587 | 0.00883039541 | 0.008502671769 | 0.008189299609 | 0.007759167168 | 0.007837277387 | 0.008493753841 | **0.007519809215** | 0.009220257061 | 0.008767722483 |
| StressDistanceMean | 0.01203977135 | 0.01253746532 | 0.01293151754 | 0.0134521405 | **0.01042984904** | 0.01061420965 | 0.01109855297 | 0.01077771859 | 0.01051226946 | 0.01190284735 | 0.0119963279 | 0.01198325226 |
| vowelPercentage | 0.007321653116 | 0.007084223609 | 0.006277669612 | 0.005988899252 | 0.005835871794 | **0.005384963073** | 0.005204501774 | 0.005007754601 | **0.004806478696** | 0.005449700467 | 0.00639412457 | 0.006526244781 |
| consonantPercentage | 0.005969989639 | 0.006529447789 | 0.007061653695 | 0.005323137319 | 0.005464023667 | 0.004764846733 | 0.004961187878 | 0.005012049382 | **0.004823568541** | **0.004472242492** | 0.005979049553 | 0.006222096301 |
| vowelDurationSD | 0.002791774429 | 0.002813237081 | 0.002516468629 | 0.002352531096 | 0.00215940007 | 0.002075787549 | 0.001875098799 | **0.001893849925** | **0.001865638044** | 0.001997030019 | 0.002378806411 | 0.002422903978 |
| consonantDurationSD | 0.001644715121 | 0.001361352002 | 0.001378395709 | 0.001351638796 | 0.001318852895 | 0.00127390065 | **0.001253660388** | **0.001391661375** | 0.001345806587 | 0.001281701203 | 0.001411203498 | 0.00154006419 |
| syllableDurationSD | 0.005843084217 | 0.005443600054 | 0.005412509827 | 0.004812982093 | **0.004506568165** | 0.003986704464 | **0.003907546667** | 0.003954436508 | 0.004033446013 | 0.004483370001 | 0.005114432883 | 0.005242870329 |
| vowelSDNorm | 0.003603924935 | 0.003876086222 | 0.003826305281 | 0.003511226036 | 0.003457750157 | 0.003276613134 | **0.003310506462** | 0.003345796281 | **0.00323037547** | 0.003448930384 | 0.003970837289 | 0.003586144238 |
| consonantSDNorm | 0.002453842467 | 0.0025985138 | 0.002593293276 | 0.002413579199 | 0.00239902707 | 0.002300432955 | **0.00227182392** | 0.002299094841 | 0.002388701117 | 0.002295757807 | 0.002519879923 | 0.00246978861 |
| syllableSDNorm | 0.006878603268 | 0.006955421732 | 0.007363689281 | 0.006514014327 | 0.005890450824 | 0.005689484998 | **0.005542829456** | 0.00561948376 | 0.005712793922 | 0.006977381009 | 0.007220585786 | 0.007268757285 |
| vowelPVINorm | 0.008082672033 | 0.008476856275 | 0.008792739705 | 0.008540074177 | **0.007293748419** | 0.007871793747 | 0.007359564478 | 0.008042923709 | 0.007456988762 | 0.008196797107 | 0.008580631534 | 0.008054205767 |
| consonantPVINorm | 0.007040520918 | 0.007458248667 | 0.007527642767 | 0.006909237687 | **0.006418862915** | 0.006852352593 | 0.006777341343 | 0.006804729463 | 0.007341054301 | 0.006959356205 | 0.007224888278 | 0.007434453143 |
| syllablePVINorm | 0.01270211817 | 0.01366898045 | 0.01309619151 | 0.01206185004 | 0.01207675386 | 0.01115063976 | **0.01108818391** | 0.01142518361 | 0.01109130918 | 0.01258012968 | 0.01311898603 | 0.01304500855 |

Table 4: Results (MSE) for pronunciation features on wav2vec2.0 for native read speech corpus (Librispeech)

| **Vocabulary_regularize** | 0th layer | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total adjectives | 0.008716142488 | 0.008636044114 | 0.008344734763 | 0.009196954348 | 0.00920171865 | 0.00906208174 | **0.00767217035** | 0.008421085408 | 0.007952393079 | 0.009323148797 | 0.009411014633 | 0.009365854858 |
| Total adverbs | 0.01164788316 | 0.01136059262 | 0.01114977867 | 0.01089577453 | 0.009900781466 | 0.00900951438 | **0.008756045924** | 0.008709974339 | 0.00956855771 | 0.01077304163 | 0.01070268154 | 0.01076030249 |
| Total nouns | 0.004878784877 | 0.005831316795 | 0.005285657415 | 0.004013603487 | 0.004325847282 | 0.004438708249 | 0.004539393746 | 0.004013435025 | **0.003849645621** | 0.004391862111 | 0.004873595781 | 0.00544428236 |
| Total verbs | 0.00974838165 | 0.008873088062 | 0.009064525829 | 0.008776794947 | 0.007376144148 | **0.00679188574** | 0.008080296239 | 0.007943765053 | 0.007044202868 | 0.00904671477 | 0.009181167671 | 0.008955392854 |
| Total pronoun | 0.002277963775 | 0.00225066888 | 0.002279728323 | 0.002273580031 | 0.002317937319 | 0.0021001447 | 0.002364496845 | 0.002101460836 | **0.001888682704** | 0.0022107569 | 0.002240079234 | 0.002303594478 |
| Total conjunction | 0.004890780031 | 0.004881876443 | 0.005034004704 | 0.00492905184 | 0.004821617171 | 0.004451050328 | 0.004238064357 | **0.004200762846** | 0.004504997002 | 0.004869517252 | 0.004884660971 | 0.004949440463 |
| Total determiners | 0.001953601612 | 0.001965513096 | 0.001956661394 | 0.001929882795 | 0.001953829346 | 0.002218593656 | 0.002319763435 | **0.001853604799** | 0.001930536822 | 0.00194624582 | 0.001955567774 | 0.001953491279 |
| Unique Word count | 0.005403100472 | 0.004279057349 | 0.006433682515 | 0.007045152681 | **0.004074146837** | 0.005392239981 | 0.004464366672 | 0.004927875533 | 0.007097396162 | 0.005556035332 | 0.005560704814 | 0.005695657893 |
| Word Complexity | 0.01100121507 | 0.01072912385 | 0.01059699809 | 0.01025654061 | 0.009527121818 | 0.009293687084 | **0.008829665164** | 0.009122303688 | 0.009117919263 | 0.009887773664 | 0.01030740389 | 0.01051470681 |

Table 5: Results (MSE) for vocabulary features on wav2vec2.0 for native read speech corpus (Librispeech)

| Audio | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| total_duration | **0.000600071184** | 0.001194973024 | 0.002856690456 | 0.003293383617 | 0.003333996603 | 0.003412567622 | 0.003016885431 | 0.004304298778 | 0.003942841525 | 0.005396684544 | 0.005147869638 | 0.005897252446 |
| stdev_energy | 0.00723100334 | 0.005820590444 | 0.008878210752 | 0.006919622558 | 0.00687860724 | 0.006747412762 | 0.006155843421 | 0.006520227762 | 0.005613004126 | 0.005844482009 | **0.005424464209** | 0.005705473624 |
| mean_pitch | 0.002863807692 | 0.005313305857 | 0.009792676723 | 0.01058728287 | 0.009882777699 | 0.01016859527 | 0.009742434661 | 0.005038250845 | 0.003442760592 | 0.001832193298 | **0.0009265430398** | 0.001271571409 |
| voiced_to_unvoiced_ratio | 0.004340692376 | 0.003959067798 | 0.005516976871 | 0.00553126727 | 0.005118835903 | 0.005269143517 | 0.00604222624 | 0.004361297257 | 0.002816091147 | 0.002328551165 | 0.002052913026 | **0.001833819867** |
| zero_crossing_rate | 0.011891957 | 0.01371445914 | 0.01484485415 | 0.01525584922 | 0.01248049667 | 0.01397136767 | 0.01256750925 | 0.01301585488 | 0.01143560999 | **0.007881070712** | 0.008747631008 | 0.00927834384 |
| energy_entropy | 0.005735883236 | **0.005620545311** | 0.006332269671 | 0.006844853533 | 0.006356723115 | 0.006517795374 | 0.006331183471 | 0.006070982313 | 0.005918655801 | 0.006350301985 | 0.006840320168 | 0.006620443672 |
| spectral_centroid | 0.0003346101311 | 0.0003346023506 | 0.0003346405374 | 0.0003347533921 | **0.0003345458164** | 0.000334732861 | 0.000334659656 | 0.000335508554 | 0.0003477160163 | 0.000334919021 | 0.0003347013857 | 0.0003349827391 |
| localJitter | 0.002801504119 | 0.002951995575 | 0.003129463739 | 0.003784385363 | 0.003684017212 | 0.003826605739 | 0.003191981708 | 0.003100164045 | 0.002830896401 | 0.002180713756 | 0.002076462643 | **0.0019662347** |
| localShimmer | 0.006433841691 | 0.006934908596 | 0.007137632993 | 0.007962424604 | 0.007199847233 | 0.00775156746 | 0.007866452774 | 0.007676613417 | 0.00645506422 | 0.005965135014 | 0.0054241555 | **0.005134716676** |

Table 6: Results (MSE) for audio features on Mockingjay for native read speech corpus (Librispeech)

| **Fluency** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| filled_pause_rate | 0.0007896804519 | 0.0007753406579 | 0.0007756551652 | 0.0007743559126 | 0.0007731516167 | **0.0007686477696** | 0.0007754823766 | 0.0007801994088 | 0.0007851040799 | 0.0008353136358 | 0.0008146844383 | 0.0007885025868 |
| general_silence | 0.003559262719 | 0.003243917261 | 0.002682329097 | 0.003353263319 | 0.003066778417 | **0.002264005409** | 0.002346107626 | 0.002401035414 | 0.003543556918 | 0.004335656832 | 0.004375657099 | 0.004599670043 |
| mean_silence | 0.002123982004 | 0.002660733886 | 0.002410833628 | 0.002158139771 | 0.00176415953 | 0.002073197064 | 0.001832566403 | 0.002312672945 | 0.001882997119 | 0.003345401533 | 0.001941584076 | **0.001739722123** |
| silence_absolute_deviation | 0.003095026406 | 0.002298633902 | 0.001742826631 | 0.001615163456 | 0.00154309816 | 0.001675014342 | **0.001372048238** | 0.002061204291 | 0.001954002011 | 0.001749570452 | 0.001508719356 | 0.002431766388 |
| SilenceRate1 | 0.00518266935 | 0.005170917284 | 0.005297491523 | 0.005822366144 | 0.005220723398 | 0.005045141669 | 0.005097903655 | 0.004941406869 | 0.004532841564 | 0.004755333236 | 0.004788232997 | **0.0041621159** |
| SilenceRate2 | 0.005450757446 | 0.00487912989 | 0.005217385559 | 0.005187273815 | 0.005730677094 | 0.0054971075 | 0.005416637539 | **0.004746003002** | 0.005169486386 | 0.005023079466 | 0.006061529806 | 0.006017881528 |
| speaking_rate | 0.01280338028 | 0.0127671412 | 0.01403781692 | 0.01425169368 | 0.01424678066 | 0.01481147138 | 0.01395719384 | 0.01526570594 | 0.01208603786 | 0.01258988822 | **0.01105901185** | 0.01113718211 |
| articulation_rate | 0.01670899878 | 0.01673615793 | 0.01822332645 | 0.01893648004 | 0.01875100032 | 0.01826913523 | 0.01799609967 | 0.01945555888 | 0.01551157328 | 0.01481122948 | **0.01279541887** | 0.01321809932 |
| longpfreq | **0.001566089048** | 0.001712909641 | 0.001707112282 | 0.001574314056 | 0.001660662408 | 0.00160382734 | 0.001672355608 | 0.001607164656 | 0.00157055707 | 0.001813219451 | 0.002025115399 | 0.001800152604 |
| average_syllables_in_words | 0.01432082053 | **0.01373734453** | 0.01423033558 | 0.01529792868 | 0.01519453918 | 0.01470058786 | 0.01448966154 | 0.01417792526 | 0.01428191168 | 0.01510727035 | 0.01540889251 | 0.01430874493 |
| wordsyll2 | **0.01001236552** | 0.01024585443 | 0.01244183527 | 0.0127584636 | 0.01183449561 | 0.01163715513 | 0.01168768375 | 0.0118791085 | 0.0110287354 | 0.01243767997 | 0.01239873538 | 0.01199892545 |
| repetition_freq | 0.01128155206 | **0.01127694702** | 0.01144730346 | 0.01167238078 | 0.01150220612 | 0.01149450501 | 0.01144426619 | 0.01142765812 | 0.01173432472 | 0.01203966654 | 0.01292588468 | 0.01227801511 |

Table 7: Results (MSE) for fluency features on Mockingjay for native read speech corpus (Librispeech)

| **Suprasegmental_Pronounciation** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| StressedSyllPercent | 0.01677603851 | 0.01677690267 | 0.01713654871 | 0.01702071686 | 0.01675987583 | 0.01747081832 | 0.01663555327 | 0.01688359415 | 0.01691505183 | **0.01596088778** | 0.01695140029 | 0.01643004819 |
| StressDistanceSyllMean | 0.002738108792 | 0.005020415113 | 0.008836323638 | 0.0117448189 | 0.0099045551 | 0.009777836222 | 0.00939542788 | 0.004853118264 | 0.003611929217 | 0.001711291598 | 0.0009603263773 | **0.0007985681893** |
| StressDistanceMean | 0.01044764089 | 0.01186773043 | 0.01470363532 | 0.01604556385 | 0.01557287461 | 0.01560927067 | 0.01430394109 | 0.01183633506 | 0.01037475399 | 0.007646403158 | 0.005007342329 | **0.004816823865** |
| vowelPercentage | 0.006114930849 | 0.006362555379 | 0.007021313219 | 0.007936547978 | 0.007495822035 | 0.007167205806 | 0.006865992613 | 0.005512591575 | 0.005455678074 | 0.00481291271 | **0.004648090234** | 0.004672305081 |
| consonantPercentage | 0.005198686002 | 0.005548506413 | 0.006473757052 | 0.00613964942 | 0.00595527057 | 0.005664949475 | 0.005455310576 | 0.005207239287 | 0.005086207487 | 0.004533730105 | **0.004377456692** | 0.004634832106 |
| vowelDurationSD | 0.002703244139 | 0.002683064792 | 0.002921476773 | 0.002859975682 | 0.0030806755 | 0.002907091215 | 0.002782790701 | 0.002687664197 | 0.002456641901 | 0.002193780611 | **0.002132015797** | 0.002287623415 |
| consonantDurationSD | 0.001127580104 | 0.001202319546 | 0.001373357087 | 0.001322325721 | 0.001312409267 | 0.001374819467 | 0.001263774466 | 0.001206966207 | 0.001088689489 | 0.001026378124 | 0.001073778125 | **0.0009591924034** |
| syllableDurationSD | 0.004989195368 | 0.005502979374 | 0.005784374275 | 0.005818281663 | 0.005727545156 | 0.005974870626 | 0.005903584651 | 0.005595582375 | 0.004834336158 | 0.004513104287 | **0.004303701421** | 0.00430450747 |
| vowelSDNorm | 0.003145973789 | 0.002862659282 | 0.002959128833 | 0.002968847753 | 0.003009786252 | 0.002889282253 | 0.00294655775 | 0.002856702744 | 0.002926424974 | 0.002931777451 | 0.002936682594 | **0.00287968754** |
| consonantSDNorm | 0.001863017581 | 0.002026477828 | 0.0018358383 | 0.001914912937 | 0.001887930823 | 0.001875639279 | 0.001864685231 | 0.001870850405 | 0.00193912884 | 0.001928171101 | 0.001947828614 | **0.001850899233** |
| syllableSDNorm | 0.00598585168 | 0.005941390994 | 0.005993520023 | 0.005847209069 | 0.005853014254 | 0.00590510027 | 0.00600419214 | 0.00583421369 | 0.005788959916 | 0.005879298497 | 0.00582767144 | **0.00572723638** |
| vowelPVINorm | 0.006571093213 | 0.006465900116 | 0.006615874495 | 0.006876691473 | 0.006913420618 | 0.006888030944 | 0.006578881004 | 0.006727194324 | 0.006547975926 | 0.006447218204 | 0.006748358005 | **0.006355829911** |
| consonantPVINorm | 0.005579530263 | 0.005593982887 | 0.005785705544 | 0.005956501091 | 0.005809107271 | 0.005824047072 | 0.005664172926 | 0.005536649725 | 0.005373282505 | **0.005286989802** | 0.005390780173 | 0.005677257282 |
| syllablePVINorm | 0.01068316305 | 0.01074624356 | 0.01102650851 | 0.01067957145 | 0.01056679301 | 0.01083935643 | 0.01079975097 | 0.01071715446 | **0.01055280899** | 0.01065924382 | 0.0110233876 | 0.01060212574 |

Table 8: Results (MSE) for pronunciation features on Mockingjay for native read speech corpus (Librispeech)

| **Vocabulary** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total adjectives | **0.007570225304** | 0.00882034699 | 0.009862320957 | 0.01103550483 | 0.01032327913 | 0.01129100272 | 0.009432048004 | 0.009538001196 | 0.01074250833 | 0.0114706295 | 0.01251287625 | 0.01034303427 |
| Total adverbs | **0.01099107328** | 0.01133155755 | 0.01131501089 | 0.01153728418 | 0.01149840418 | 0.0114583776 | 0.01138668751 | 0.01135164256 | 0.01159329082 | 0.01265736261 | 0.01334193369 | 0.01246985235 |
| Total nouns | **0.004010554653** | 0.004432963924 | 0.006881253289 | 0.007136661454 | 0.006792365692 | 0.006114054424 | 0.005584657589 | 0.006672900855 | 0.00726183225 | 0.007767768879 | 0.00868267008 | 0.006561221521 |
| Total verbs | **0.008358126713** | 0.009058740265 | 0.01034887386 | 0.01068432699 | 0.01019642499 | 0.01044021179 | 0.009786931247 | 0.00914714026 | 0.012780693 | 0.01314768836 | 0.01465605013 | 0.01154828264 |
| Total pronoun | 0.002238017096 | 0.002216677495 | **0.002210232125** | 0.002230531061 | 0.002242119138 | 0.002281428065 | 0.002264127223 | 0.002303299702 | 0.00225891565 | 0.002373602284 | 0.002524744261 | 0.002345227557 |
| Total conjunction | 0.004934738222 | **0.004901665396** | 0.004965355505 | 0.005017159723 | 0.005184474139 | 0.005050440986 | 0.004963267165 | 0.00500932937 | 0.005076171444 | 0.005169624677 | 0.005550043994 | 0.005081462702 |
| Total determiners | **0.001940148071** | 0.001973972214 | 0.001943869767 | 0.001952394604 | 0.001968247053 | 0.001976830932 | 0.001941036299 | 0.001941974822 | 0.00197671487 | 0.002005681096 | 0.002023148284 | 0.001973492972 |
| Unique Word count | **0.002199487996** | 0.005995549278 | 0.006729647211 | 0.005918513972 | 0.006723115879 | 0.007532181416 | 0.005242927665 | 0.005412766723 | 0.006634656724 | 0.008468982434 | 0.01162688466 | 0.005578399092 |
| Word Complexity | **0.01112771632** | 0.01140624251 | 0.01153482002 | 0.01159091894 | 0.01167012734 | 0.01132541804 | 0.01152729342 | 0.01143041 | 0.01146659636 | 0.01201426225 | 0.01146897922 | 0.0113627403 |

Table 9: Results (MSE) for vocabulary features on Mockingjay for native read speech corpus (Librispeech)

| Audio | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| total_duration | 0.003036852584 | **0.002240838744** | 0.002386579194 | 0.002948529764 | 0.002987011009 | 0.002997231913 | 0.002993126693 | 0.003980276745 | 0.003188442568 | 0.003082025403 | 0.004429940143 | 0.005752840747 |
| stdev_energy | 0.01324670535 | **0.01077789761** | 0.01322383268 | 0.01168904504 | 0.01125114423 | 0.01118140825 | 0.01116425438 | 0.01122983216 | 0.01251280088 | 0.01179641799 | 0.01188905848 | 0.01145454932 |
| mean_pitch | 0.004493347796 | **0.003569331866** | 0.003842852411 | 0.004897000657 | 0.005505153375 | 0.004684224744 | 0.005197434331 | 0.005699304544 | 0.005733258858 | 0.004194434231 | 0.008189400288 | 0.006506099881 |
| voiced_to_unvoiced_ratio | 0.002073641644 | 0.002023948256 | 0.001661073744 | 0.002073183115 | 0.002288006682 | **0.001632213221** | 0.001988362484 | 0.001961196856 | 0.001981582041 | 0.001903942197 | 0.002232590702 | 0.002125300737 |
| zero_crossing_rate | 0.01051946283 | 0.007791790176 | 0.006900655476 | 0.007061863494 | 0.00820783994 | 0.00678975414 | 0.007180293173 | 0.006586508343 | 0.006429331852 | **0.006368979636** | 0.01055886017 | 0.009978544089 |
| energy_entropy | 0.0131660124 | 0.01051870928 | **0.008691528516** | 0.01035972697 | 0.01041408811 | 0.01094031447 | 0.01078624351 | 0.01015224294 | 0.01352512007 | 0.009773955154 | 0.01058837599 | 0.01159235612 |
| spectral_centroid | 0.000004050640549 | 0.000002884069191 | 0.0000026139057 | 0.00000268593849 | 0.000002993052206 | 0.000003158745487 | **0.000002540464208** | 0.000003608490735 | 0.000004299552994 | 0.000002586428523 | 0.000002568550171 | 0.000002869018177 |
| localJitter | 0.008843279339 | 0.00763015628 | 0.01008887258 | 0.007923614721 | **0.0070344993** | 0.008898885073 | 0.007445504953 | 0.01006001376 | 0.00802474246 | 0.007730700765 | 0.00864250852 | 0.008793759531 |
| localShimmer | 0.005665909947 | 0.004940269417 | 0.005272669217 | 0.004515491684 | 0.006484832943 | **0.004504353798** | 0.004658858808 | 0.005152755834 | 0.004648094578 | 0.005181036245 | 0.006094569912 | 0.005418662421 |

Table 10: Results (MSE) for audio features on wave2vec2.0 for non-native read speech corpus (L2 Arctic)

| **Fluency** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| filled_pause_rate | 0.000006387970224 | 0.000008682296089 | 0.000004741688422 | 0.000004237409567 | 0.000009187543289 | 0.000003579812709 | 0.000005561210126 | 0.00002653749623 | 0.00002342882729 | 0.000003693411822 | **0.00003357185108** | 0.000003749156653 |
| general_silence | 0.01033555951 | 0.009149169464 | 0.009780546845 | 0.009448548738 | 0.01021406095 | 0.009556660139 | 0.01046188957 | 0.01013046589 | 0.009625767908 | 0.01069982609 | **0.00914006391** | 0.01085762649 |
| mean_silence | 0.008366614439 | 0.008180335582 | 0.008528918032 | **0.007217843789** | 0.007962025364 | 0.008569399553 | 0.007367689096 | 0.008397009404 | 0.0086400927 | 0.008453370616 | 0.009267107046 | 0.00882008465 |
| silence_abs_deviation | 0.008834311295 | 0.008650845385 | 0.008100582414 | 0.008430736154 | 0.008082393993 | 0.008105375897 | 0.008495559228 | 0.007867558302 | **0.007671361145** | 0.008465820279 | 0.009120660437 | 0.0108708525 |
| SilenceRate1 | 0.01044123743 | 0.009092348424 | 0.01043301035 | 0.009221869468 | 0.009357026617 | 0.008964186501 | 0.009186406205 | 0.009737917622 | 0.01007781563 | 0.009489678463 | **0.008924994621** | 0.01044440624 |
| SilenceRate2 | 0.01916863316 | **0.01771614741** | 0.01913049109 | 0.018712528 | 0.01860373252 | 0.01792320075 | 0.01849100455 | 0.01928141526 | 0.01825185886 | 0.01875508536 | 0.01799868099 | 0.02265250758 |
| speaking_rate | 0.00949531612 | 0.008706728431 | 0.009149666883 | 0.009288570987 | 0.009673628871 | 0.008421046499 | 0.007932631021 | 0.008018199961 | **0.00788175189** | 0.009372725227 | 0.01003070456 | 0.009355750841 |
| articulation_rate | 0.01460807176 | 0.012276978 | 0.01199722544 | 0.01293632749 | 0.01212447962 | **0.01100245681** | 0.01102734852 | 0.01178566118 | 0.01273003023 | 0.01180807437 | 0.01229814885 | 0.0123860972 |
| longpfreq | 0.006085115024 | 0.005659941734 | 0.005149693695 | 0.005310365257 | **0.004731439031** | 0.005081345847 | 0.005215093074 | 0.005338007252 | 0.00520273901 | 0.004949189748 | 0.0055988443 | 0.005908655898 |
| average_syllables_in_words | 0.0405414943 | 0.03934107353 | 0.03889618546 | 0.04012435788 | 0.03491881224 | 0.03203875843 | **0.02692308332** | 0.03189995042 | 0.03070087745 | 0.03460050468 | 0.04644216529 | 0.04072065185 |
| wordsyll2 | 0.02971014255 | 0.02958088167 | 0.02909421088 | 0.02779293592 | 0.02733638405 | 0.02679802803 | **0.02158009075** | 0.02380347881 | 0.02561642531 | 0.02566025422 | 0.03042140132 | 0.02845798602 |
| repetition_freq | 0.02574293891 | 0.02627080771 | 0.02602259791 | 0.02620179685 | 0.02627396395 | **0.02551717195** | 0.02575151809 | 0.02705670945 | 0.02671655863 | 0.02590812334 | 0.02628673194 | 0.02628793326 |

Table 11: Results (MSE) for fluency features on wave2vec2.0 for non-native read speech corpus (L2 Arctic)

| **Pronunciation** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| StressDistanceSyllMean | 0.01085782733 | 0.01098628002 | 0.01108407769 | 0.01070625137 | **0.01044733789** | 0.01092516488 | 0.01057964372 | 0.01062937802 | 0.01096083754 | 0.01096456235 | 0.01082172927 | 0.01086185439 |
| StressDistanceMean | 0.01445041269 | 0.01437040954 | 0.01474399634 | 0.01455219575 | 0.01444723178 | 0.01438453577 | **0.01380486132** | 0.01465175479 | 0.01415742156 | 0.01446328933 | 0.0146644549 | 0.01456196221 |
| vowelPercentage | 0.006582178908 | 0.006036077598 | 0.005798679384 | 0.005019396605 | 0.004906005752 | 0.005742765319 | 0.005376293419 | 0.005847743266 | **0.004815154536** | 0.005328370126 | 0.006170894301 | 0.006101583732 |
| consonantPercentage | 0.01077557184 | 0.009301757647 | 0.009350718316 | 0.01172386174 | 0.00765887745 | **0.006690978593** | 0.008678450289 | 0.008014754067 | 0.00881095758 | 0.008519494114 | 0.009191255784 | 0.009435561027 |
| vowelDurationSD | 0.004968164154 | 0.005062242745 | 0.004466884003 | 0.004497627784 | 0.004240898993 | **0.004156724155** | 0.004320067699 | 0.004319318428 | 0.004305789569 | 0.004623777541 | 0.004952676432 | 0.00479170822 |
| consonantDurationSD | 0.008936564392 | 0.008598281277 | 0.008308606219 | 0.008175608065 | 0.008195857606 | **0.007876085461** | 0.008254148679 | 0.008291035084 | 0.008223764793 | 0.008239658054 | 0.009342117536 | 0.008895579444 |
| syllableDurationSD | 0.01806393282 | 0.01700667742 | 0.0167047632 | 0.01594318009 | 0.01699094383 | 0.01574212378 | **0.01560875624** | 0.01609173706 | 0.01600048516 | 0.01620207649 | 0.01767428638 | 0.01906834934 |
| vowelSDNorm | 0.007290093544 | 0.007268939415 | 0.007572882747 | 0.007195812177 | 0.007125734843 | 0.007204871667 | 0.007109662866 | 0.007452009186 | **0.0071058745** | 0.007387514968 | 0.00728065025 | 0.007137284653 |
| consonantSDNorm | 0.01111690383 | 0.01012459115 | 0.01039347629 | 0.01005310843 | 0.01032912051 | 0.01007985877 | 0.01034796039 | 0.01050947246 | **0.01000415971** | 0.01125236326 | 0.01140357022 | 0.01113252911 |
| syllableSDNorm | 0.01593478455 | 0.016013385 | 0.01563592597 | 0.01565452532 | 0.01731780043 | 0.01532025746 | **0.01483609854** | 0.0163971165 | 0.01779906823 | 0.0157345589 | 0.01658006541 | 0.01767856745 |
| vowelPVINorm | 0.006011102844 | 0.006212722731 | 0.006066495108 | 0.006022844786 | 0.0059080979 | 0.005946239177 | **0.00579974942** | 0.005853436331 | 0.005861750356 | 0.005970207649 | 0.006059811336 | 0.005955741749 |
| consonantPVINorm | **0.009807191844** | 0.009824664954 | 0.01023335258 | 0.01036442773 | 0.01015554178 | 0.01186972196 | 0.009981847411 | 0.01027775468 | 0.01083973459 | 0.011682993 | 0.01066331915 | 0.01051632762 |
| syllablePVINorm | 0.01513100113 | 0.01469861179 | 0.01460090845 | 0.01424850078 | 0.01509495692 | 0.01560843725 | **0.01386009991** | 0.01513577809 | 0.01471565868 | 0.01435015978 | 0.01558518122 | 0.01490329087 |

Table 12: Results (MSE) for pronunciation features on wave2vec2.0 for non-native read speech corpus (L2 Arctic)

| **Vocab** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total adjectives | 0.05510421979 | 0.05573282852 | 0.05543830605 | 0.05820005871 | 0.05378823381 | 0.05328800051 | 0.05165909671 | **0.05101566702** | 0.05514225681 | 0.05501988631 | 0.05822923791 | 0.05865896362 |
| Total adverbs | 0.03235472971 | 0.03103184496 | **0.03050799792** | 0.03141402481 | 0.03414797832 | 0.03062025662 | 0.03362480616 | 0.03124593999 | 0.03223135553 | 0.03249981824 | 0.03078615433 | 0.03073787498 |
| Total nouns | 0.03152909146 | 0.03221436911 | 0.03405263043 | 0.03293163967 | 0.03294914251 | 0.02952474792 | **0.02835238334** | 0.02928062923 | 0.02950680222 | 0.03049105247 | 0.03218764593 | 0.03326532311 |
| Total verbs | 0.03806369809 | 0.03730369597 | 0.04160129284 | 0.03769314576 | 0.03674953763 | 0.04077345616 | **0.03550568766** | 0.03743405782 | 0.04529775482 | 0.04024061532 | 0.03756253321 | 0.03765887944 |
| Total pronoun | 0.0195913671 | 0.01954425637 | 0.01955983322 | 0.01959607485 | 0.01956247168 | **0.01944526418** | 0.01969910536 | 0.01944636298 | 0.01950609594 | 0.01949140182 | 0.01948452475 | 0.01949405056 |
| Total conjunction | 0.04257824983 | 0.04249800371 | 0.04201713096 | 0.04132180095 | 0.04235767698 | 0.04219216089 | 0.04165918617 | 0.04138203177 | **0.04112746427** | 0.04187836853 | 0.04245932476 | 0.04278077157 |
| Total determiners | 0.01977967616 | 0.01961318789 | 0.01991113196 | 0.01972002446 | 0.01972667769 | 0.01958241395 | 0.01964575983 | **0.0195345228** | 0.01957967543 | 0.01961906192 | 0.01960968918 | 0.01977558393 |
| Unique Word count | 0.01632767888 | 0.01526919263 | 0.01442965304 | 0.0142056617 | 0.01618741767 | 0.01649467498 | **0.01319559305** | 0.01332996183 | 0.01482716115 | 0.01525202137 | 0.01619794752 | 0.01812350579 |
| Word Complexity | 0.02357972535 | 0.02507438627 | 0.02657050341 | 0.02360681218 | 0.02386869462 | **0.02275726452** | 0.02455551883 | 0.02325487976 | 0.02408732581 | 0.02473021066 | 0.02585488513 | 0.02602586518 |

Table 13: Results (MSE) for text features on wave2vec2.0 for non-native read speech corpus (L2 Arctic)

| Audio | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| total_duration | 0.01132428695 | **0.003762919523** | 0.007238992203 | 0.005671930936 | 0.005220973059 | 0.007031149497 | 0.007505373695 | 0.007055257454 | 0.007160907835 | 0.007566914968 | 0.01002525633 | 0.01267732305 |
| stdev_energy | 0.01531330077 | 0.01533933898 | 0.01649358362 | 0.01449291551 | 0.01402955807 | 0.01536183221 | 0.01397247866 | 0.01796714901 | 0.01486330595 | 0.01347379434 | **0.01117071068** | 0.01251225659 |
| mean_pitch | 0.01631877178 | 0.01786401 | 0.01384232861 | 0.01427588393 | 0.01120626363 | 0.01479962637 | 0.01793363005 | 0.01136428547 | 0.008506216093 | 0.003964858149 | **0.00336777112** | 0.003792742136 |
| voiced_to_unvoiced_ratio | 0.002840575782 | 0.002697371658 | 0.002844036818 | 0.002683747683 | 0.002754297038 | 0.002925922295 | 0.002919951457 | 0.002638006071 | 0.002201766288 | 0.001792214941 | **0.001610400597** | 0.00221409082 |
| zero_crossing_rate | 0.01198278218 | 0.013095671 | 0.01180463814 | 0.0120515795 | 0.01135187312 | 0.01377190836 | 0.01748469921 | 0.01238682109 | 0.007931431331 | 0.00732755264 | **0.00490843802** | 0.005564741017 |
| energy_entropy | 0.01426082695 | 0.01395716063 | **0.01138648094** | 0.01175751104 | 0.01330351103 | 0.01249528354 | 0.01563216641 | 0.01412510058 | 0.01682605393 | 0.01212163778 | 0.01392596306 | 0.01502481167 |
| spectral_centroid | 0.000002846928862 | 0.000002708621608 | 0.000002917653246 | 0.000002864165578 | 0.000004297931043 | **0.000002537697294** | 0.000004529206696 | 0.000002645481248 | 0.000002706540701 | 0.000003055551443 | 0.000002719192597 | 0.000004625777187 |
| localJitter | 0.009555001024 | 0.009915405168 | 0.008387510554 | 0.009048920991 | 0.008969662653 | 0.01148685269 | 0.01161788982 | 0.0077060057 | 0.007508060421 | 0.007772406808 | 0.007588351459 | **0.007486868943** |
| localShimmer | 0.007763135906 | 0.007365633079 | 0.00610032398 | 0.006087114938 | 0.006471823685 | 0.006133956514 | 0.007093345793 | 0.00578951809 | 0.006633051749 | 0.005442358034 | 0.006894810167 | **0.004808022677** |

Table 14: Results (MSE) for audio features on Mockingjay for non-native read speech corpus (L2 Arctic)

| **Fluency** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| filled_pause_rate | 0.000001473276996 | 0.000005419860981 | 0.000001650465328 | 0.000008519413545 | 0.000003878072607 | 0.00001318502424 | **0.00000008692807352** | 0.00001343803189 | 0.00007752478416 | 0.00000156222934 | 0.00000356914358 | 0.00005134366334 |
| general_silence | 0.01307881676 | 0.01167528975 | 0.01133589357 | 0.01092097086 | 0.01077687113 | 0.01092365747 | 0.01254632489 | 0.01162321885 | 0.01090970089 | **0.00954652106** | 0.0110049385 | 0.01233310859 |
| mean_silence | 0.009038383626 | 0.008801689111 | 0.00912950239 | 0.009512880408 | 0.01012747188 | 0.01160653425 | 0.01098129752 | 0.008231421116 | 0.007322015957 | 0.00804670488 | **0.0072769857** | 0.007886753264 |
| silence_abs_deviation | 0.008988055153 | 0.007935119196 | 0.009433308074 | 0.009916727547 | 0.008995566402 | 0.009183798126 | 0.009436651227 | 0.008632040293 | **0.007513496565** | 0.008630425092 | 0.008548152859 | 0.009036964244 |
| SilenceRate1 | 0.01136621292 | 0.0108004546 | 0.01101192452 | 0.009303176583 | 0.009495197318 | 0.01058382713 | 0.00993428986 | 0.009442205917 | 0.009309233593 | 0.00890018279 | **0.008661307729** | 0.009029231575 |
| SilenceRate2 | 0.02076327719 | 0.01990882985 | 0.01952633015 | 0.01843555537 | 0.02110262895 | 0.01852757613 | 0.01941679605 | 0.01836522243 | 0.0172794501 | 0.01710937857 | **0.0166202417** | 0.01818893651 |
| speaking_rate | 0.01057500348 | 0.01003365801 | 0.009616637147 | 0.009627113545 | 0.009733180997 | 0.01042946507 | 0.01385399793 | 0.01038978347 | 0.01006254321 | 0.01157778177 | 0.0101515701 | **0.009589242748** |
| articulation_rate | 0.01566377477 | 0.01457197335 | 0.01536387335 | 0.01410481012 | 0.01377338301 | 0.01519435582 | 0.01773659911 | 0.01638551305 | 0.01530340419 | 0.01374970851 | **0.01359579013** | 0.01501365605 |
| longpfreq | 0.007634478426 | 0.006733958909 | 0.009045465163 | 0.006315528741 | 0.005608732161 | 0.00631258586 | 0.006255921864 | 0.005071465668 | 0.004950666884 | 0.004762317562 | **0.004740796861** | 0.004900868777 |
| average_syllables_in_words | 0.04781052576 | 0.05374880648 | **0.04171374348** | 0.04345499262 | 0.04466745407 | 0.04727407451 | 0.04362590181 | 0.04413596202 | 0.0430084352 | 0.04317014443 | 0.04380146089 | 0.04577109174 |
| wordsyll2 | 0.03708758571 | 0.03631960768 | 0.03530161093 | **0.03477106527** | 0.03723229604 | 0.03800885004 | 0.03685256715 | 0.03744263533 | 0.03709040197 | 0.03557814656 | 0.03657478211 | 0.03585507681 |
| repetition_freq | 0.02641030373 | 0.02625820395 | 0.02641122354 | 0.02624297484 | **0.02617873088** | 0.02633850854 | 0.02649016891 | 0.02648242403 | 0.02648564996 | 0.02651871265 | 0.02619171225 | 0.02630454237 |

Table 15: Results (MSE) for fluency features on Mockingjay for non-native read speech corpus (L2 Arctic)

| **Pronunciation** | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| StressDistanceSyllMean | 0.01104542613 | 0.01109904291 | 0.01132480895 | 0.01087026261 | 0.01099540125 | **0.01082515577** | 0.01099589881 | 0.01100831576 | 0.01102862626 | 0.01115663156 | 0.01083237715 | 0.01120610025 |
| StressDistanceMean | 0.01508909442 | 0.01540528173 | 0.01499919857 | 0.01580319327 | **0.01475674539** | 0.01507330306 | 0.01513319981 | 0.01513321 | 0.01508071928 | 0.01489617973 | 0.01515311937 | 0.01584713657 |
| vowelPercentage | 0.007385191339 | 0.007231984856 | 0.00714794423 | 0.006742874537 | 0.009473037671 | 0.007246873266 | 0.007359432344 | 0.006418323083 | 0.006181449067 | 0.005328299893 | **0.005268095936** | 0.005667836572 |
| consonantPercentage | 0.0116318778 | 0.01213936938 | 0.0109606066 | 0.01166343992 | 0.01047406819 | 0.01305005633 | 0.01413893024 | 0.01115384658 | 0.008864534174 | **0.008217352218** | 0.008636426617 | 0.01102221575 |
| vowelDurationSD | 0.005576002021 | 0.005648573514 | 0.005741291856 | 0.005688154593 | 0.005604476273 | 0.005606742182 | 0.005670487645 | 0.005817736235 | 0.005528790984 | 0.005221120835 | **0.004916082459** | 0.005037117742 |
| consonantDurationSD | 0.009748503712 | 0.009874701213 | 0.009595141269 | 0.009227518199 | 0.009424982592 | 0.009572179407 | 0.009823950086 | 0.009433136911 | 0.009011294895 | 0.008617355423 | **0.00828695312** | 0.008858853208 |
| syllableDurationSD | 0.02132959553 | 0.02184669811 | 0.0226484616 | 0.01974255433 | 0.02013796378 | 0.02047414778 | 0.02192453187 | 0.02083005298 | 0.02065268392 | 0.0189069598 | **0.01776821627** | 0.01812854238 |
| vowelSDNorm | 0.007599164413 | 0.007843354775 | 0.008135076697 | 0.007658485044 | 0.007601849988 | 0.007556601614 | 0.00763172515 | 0.007809241658 | 0.007583258576 | **0.007363074176** | 0.007511993508 | 0.007514772523 |
| consonantSDNorm | 0.01214739807 | 0.0120310838 | 0.01166347967 | 0.01236910295 | 0.01162965308 | 0.01164295335 | 0.01195547213 | 0.01212879421 | 0.01196931571 | **0.01097453874** | 0.01135074416 | 0.01105961095 |
| syllableSDNorm | 0.01804111848 | 0.01637122238 | 0.01882830924 | 0.01927251511 | **0.01621654996** | 0.0166021328 | 0.01686714405 | 0.0163813078 | 0.01657909281 | 0.01710862262 | 0.01746782886 | 0.01690434325 |
| vowelPVINorm | 0.006215734333 | 0.006449337514 | 0.007165993656 | 0.006114570327 | 0.006628356752 | 0.006099774583 | 0.006222376875 | 0.006118055118 | 0.006215811826 | 0.006102456314 | **0.006066908159** | 0.006612337197 |
| consonantPVINorm | 0.0116556712 | 0.01048152507 | 0.01176013518 | 0.01188734875 | 0.01042246125 | 0.01100122118 | 0.01046942329 | 0.01050718765 | 0.01059977782 | 0.01176892219 | **0.01010445988** | 0.0102762624 |
| syllablePVINorm | 0.01475997474 | 0.01477713179 | 0.01559573645 | **0.01460650209** | 0.01563738917 | 0.01487725143 | 0.01486211243 | 0.01462639155 | 0.01491031254 | 0.01470857668 | 0.01541867249 | 0.014618768 |

Table 16: Results (MSE) for pronunciation features on Mockingjay for non-native read speech corpus (L2 Arctic)

| **Vocabulary** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total adjectives | 0.06318404862 | 0.0619247938 | 0.06079716269 | 0.06014132744 | **0.05987823091** | 0.06053751686 | 0.06247047087 | 0.06301413201 | 0.06361835457 | 0.06178066896 | 0.06281169399 | 0.06382546757 |
| Total adverbs | 0.03117171396 | 0.03095706324 | 0.03091227592 | 0.03186724566 | 0.03075210676 | 0.03163696857 | 0.03176016577 | **0.03020573131** | 0.0308514338 | 0.03167641853 | 0.03124024759 | 0.03138128394 |
| Total nouns | 0.03479309525 | 0.04124192357 | 0.03509293774 | **0.03460912932** | 0.03484343636 | 0.03937797862 | 0.03621976189 | 0.03965186853 | 0.03527331659 | 0.03480436596 | 0.03614236555 | 0.03960483099 |
| Total verbs | **0.03815241929** | 0.03843911825 | 0.04098537793 | 0.03909517882 | 0.04191215109 | 0.04264500828 | 0.03921201085 | 0.03855232573 | 0.0391547778 | 0.03854201282 | 0.04359669904 | 0.03981245254 |
| Total pronoun | 0.01971188699 | 0.01985237799 | 0.01996660004 | 0.01965335094 | 0.01971121361 | 0.01976920775 | 0.01964397728 | **0.0196416389** | 0.01967384263 | 0.01966713911 | 0.01971822402 | 0.01979745081 |
| Total conjunction | 0.04272169798 | **0.04139294662** | 0.04453893254 | 0.04275823407 | 0.04250406794 | 0.04203422813 | 0.04204185542 | 0.04190527645 | 0.04221825341 | 0.04224755 | 0.0422208323 | 0.04247788975 |
| Total determiners | 0.01973337315 | 0.01982656053 | 0.01975389895 | 0.01970792552 | 0.0198162249 | 0.01972706823 | 0.01980878641 | 0.01978879803 | 0.01979176333 | **0.01953582744** | 0.01963054884 | 0.01960785383 |
| Unique Word count | 0.01989839225 | **0.01728327798** | 0.01783175585 | 0.02423742574 | 0.01881541414 | 0.01975514188 | 0.02222050837 | 0.0226465011 | 0.02065547339 | 0.02429926023 | 0.02616072579 | 0.02887803155 |
| Word Complexity | 0.02514394797 | 0.0261101335 | 0.02470550073 | 0.02725880817 | 0.02493061146 | 0.02470394819 | 0.02668094082 | 0.02487191537 | 0.02470169555 | 0.02530130615 | **0.02433745363** | 0.02536928852 |

Table 17: Results (MSE) for text features on Mockingjay for non-native read speech corpus (L2 Arctic)

| Audio_reg | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| total_duration | 0.004848900229 | 0.003863915258 | 0.004287122943 | **0.003783825339** | 0.004344520906 | 0.004427644402 | 0.003811237183 | 0.005109543088 | 0.004882294667 | 0.005763219731 | 0.009990349682 | 0.00750392505 |
| stdev_energy | 0.01730283908 | **0.01681536258** | 0.01796875486 | 0.01966112687 | 0.0205387239 | 0.02083439084 | 0.02079950793 | 0.02027682245 | 0.01946188919 | 0.01971326701 | 0.02184660291 | 0.01947816552 |
| mean_pitch | 0.01006024988 | **0.00658550357** | 0.008036226643 | 0.00811293786 | 0.008285121024 | 0.008348004637 | 0.01001664103 | 0.01012975076 | 0.009512276668 | 0.009188184194 | 0.01351123816 | 0.0119256139 |
| voiced_to_unvoiced_ratio | 0.006834375191 | 0.007376795598 | 0.005870861794 | 0.005814311666 | 0.00700179955 | 0.006451495571 | 0.007265555558 | **0.005396942508** | 0.006471980731 | 0.007256321426 | 0.008058287228 | 0.00715184344 |
| zero_crossing_rate | 0.01357175479 | 0.0140099222 | **0.013133127** | 0.0136683096 | 0.01394808329 | 0.01504515634 | 0.01491236375 | 0.01575922416 | 0.01368447107 | 0.01354429307 | 0.01973247031 | 0.01492608689 |
| energy_entropy | 0.01286934208 | 0.01425085582 | 0.01343616879 | 0.01308097535 | 0.01387268033 | **0.01239496646** | 0.01511617563 | 0.01359821082 | 0.01616405567 | 0.01966480175 | 0.015758514 | 0.01739258232 |
| spectral_centroid | **0.004403675623** | 0.004427087809 | 0.004423378077 | 0.004408777549 | 0.004420320531 | 0.004409092315 | 0.004418122369 | 0.004420023754 | 0.004420583009 | 0.004421105874 | 0.004423533266 | 0.004407367195 |
| localJitter | **0.01373034504** | 0.01404961373 | 0.01389421042 | 0.01494240126 | 0.01455549524 | 0.01497129501 | 0.01477446334 | 0.01573872289 | 0.0148343247 | 0.01415938999 | 0.01562866786 | 0.01418735225 |
| localShimmer | 0.01205946784 | 0.009315373126 | 0.01028803862 | 0.009537074303 | **0.00927095069** | 0.0105189144 | 0.01024419213 | 0.01027528666 | 0.009515543941 | 0.009711281336 | 0.01138879615 | 0.01044949956 |

Table 18: Results (MSE) for audio features on wave2vec2.0 for native spontaneous speech corpus (Mozilla Common Voice)

| **Vocabulary_reg** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total adjectives | 0.0335636095 | 0.0355528849 | 0.03562800051 | 0.03371269054 | 0.035960629 | 0.03484367332 | 0.03336296315 | 0.03454088316 | 0.03774241945 | 0.03567314315 | 0.03737013864 | **0.03309152118** |
| Total adverbs | 0.03499237064 | 0.03490227723 | 0.03548898718 | 0.03591213537 | 0.03465381191 | 0.03488443067 | 0.03478685717 | 0.0349435478 | 0.03489453385 | 0.03488539124 | 0.03506322218 | **0.03446789372** |
| Total nouns | 0.0259992195 | 0.02575678079 | 0.02825899267 | **0.02286530921** | 0.02753577135 | 0.02649220177 | 0.02576713354 | 0.02951983639 | 0.02885208712 | 0.02460092158 | 0.02772971892 | 0.02685838834 |
| Total proper nouns | 0.001246035635 | 0.001240899417 | 0.00125103115 | **0.001239943418** | 0.001239440379 | 0.001239651024 | 0.001245050603 | 0.001241447194 | 0.001279094276 | 0.001251123873 | 0.001287215026 | 0.001280808881 |
| Total verbs | 0.02174248396 | 0.02238550591 | 0.02164992172 | 0.0214138707 | 0.02250983884 | 0.02167924295 | 0.02164178445 | 0.02179472569 | 0.02206596364 | 0.02152242078 | 0.0216113466 | **0.02106053148** |
| Total pronoun | 0.004966920314 | 0.004971733874 | 0.004971720508 | 0.004968270775 | 0.004969505559 | **0.004966416064** | 0.004993177728 | 0.004969222783 | 0.0049912574 | 0.004994949394 | 0.004968574727 | 0.005062977071 |
| Total conjunction | 0.01850148618 | 0.01841563206 | 0.01848532752 | 0.01843472013 | 0.01846578636 | 0.01846183381 | 0.01852863046 | 0.01848158558 | 0.01848480125 | **0.01836849423** | 0.01870337746 | 0.01845905571 |
| Total determiners | 0.001353849711 | 0.001253704095 | 0.001272171462 | **0.001249276101** | 0.001262623841 | 0.001303429942 | 0.001362165212 | 0.001254933033 | 0.001293262048 | 0.001445716578 | 0.001389417068 | 0.001452194368 |
| Unique Word count | 0.01845888857 | 0.02003891249 | 0.02003383619 | 0.01776118224 | 0.02201419969 | 0.01871178851 | 0.0180176312 | 0.02021435048 | 0.01904716594 | 0.01893145189 | 0.02137563037 | **0.01735364781** |
| Word Complexity | 0.01760485511 | 0.01667454042 | 0.01630606359 | 0.01949920931 | 0.01730683897 | 0.01996059248 | 0.01737334921 | 0.02020322766 | **0.01521792993** | 0.01757990552 | 0.01608085374 | 0.01588515035 |

Table 19: Results (MSE) for text features on wave2vec2.0 for native spontaneous speech corpus (Mozilla Common Voice)

| Audio | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| total_duration | 0.01969611883 | 0.009559035356 | **0.009083062274** | 0.01035598879 | 0.009994780165 | 0.01041355546 | 0.01413635177 | 0.01290650319 | 0.01739108976 | 0.0174465089 | 0.02315019196 | 0.03310960901 |
| stdev_energy | 0.02301282557 | 0.02356082611 | 0.02357562409 | 0.02312384893 | 0.02322197109 | 0.02277250205 | 0.02332409925 | 0.02315600874 | 0.02222899431 | **0.02034480139** | 0.02071422557 | 0.0225063406 |
| mean_pitch | 0.03195826549 | 0.03202319787 | 0.03432047275 | 0.03161315384 | 0.03527633403 | 0.03387834162 | 0.03807346721 | 0.0301615608 | 0.02604102958 | 0.01752459431 | 0.01540516296 | **0.01320302476** |
| voiced_to_unvoiced_ratio | 0.01132498675 | 0.01224456425 | 0.01150825352 | 0.01050957408 | 0.010532844 | 0.01125972421 | 0.01182617455 | 0.009867347706 | 0.008246490254 | 0.006430362119 | **0.005966189793** | 0.007566473671 |
| zero_crossing_rate | 0.02514514977 | 0.02344684766 | 0.02265917967 | 0.02391377785 | 0.02312772603 | 0.0256152895 | 0.02416192244 | 0.02181344027 | 0.0199218134 | 0.01959072123 | **0.01791414976** | 0.01906959729 |
| energy_entropy | 0.02260997071 | 0.01864840583 | 0.01924748623 | 0.02238052367 | 0.02471348779 | 0.02809989892 | 0.02233257516 | 0.02344543182 | 0.02040420553 | **0.01593418513** | 0.02460551242 | 0.02125200622 |
| spectral_centroid | 0.00441998788 | 0.004420464559 | 0.004415645452 | 0.00442057248 | 0.004414127048 | 0.004421131345 | 0.004420333751 | 0.004423290144 | 0.004420883487 | 0.004419840753 | **0.004398830114** | 0.004422171867 |
| localJitter | 0.01830855997 | 0.01807919782 | 0.01801409873 | 0.01824081742 | 0.01715655285 | 0.01833400694 | 0.01812983634 | 0.01789327138 | 0.01664063823 | 0.01590755723 | **0.0144273038** | 0.01582481371 |
| localShimmer | 0.01804036728 | 0.0166693069 | 0.01632145589 | 0.01672439724 | 0.01701765074 | 0.01663488258 | 0.01663078698 | 0.01690814134 | 0.0151106039 | 0.01350417914 | **0.01234748409** | 0.01306185737 |

Table 20: Results (MSE) for audio features on Mockingjay for native spontaneous speech corpus (Mozilla Common Voice)

| **Vocabulary** | 1st layer | 2nd layer | 3rd layer | 4rth layer | 5th layer | 6th layer | 7th layer | 8th layer | 9th layer | 10th layer | 11th layer | 12th layer |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Total adjectives | **0.03241210244** | 0.03482287048 | 0.03529522366 | 0.03642654975 | 0.03582160011 | 0.03406383728 | 0.03420949581 | 0.03422122207 | 0.04153839558 | 0.03681910709 | 0.03511459393 | 0.03497248174 |
| Total adverbs | 0.03484643463 | 0.03475470728 | 0.03587847323 | 0.03477450888 | 0.03488850375 | 0.03495373018 | 0.03477768324 | 0.03487565293 | 0.03494904503 | **0.034743899** | 0.03476430007 | 0.0357219648 |
| Total nouns | 0.0260823038 | 0.02506542004 | 0.0257070428 | 0.02820721339 | 0.02674152729 | 0.02730540183 | 0.02673217732 | 0.02770532215 | 0.02663394189 | **0.02477734541** | 0.030839408 | 0.02581936672 |
| Total proper nouns | 0.00124667931 | 0.001262687502 | 0.001248209042 | 0.001258744441 | 0.001256570356 | 0.001245394034 | 0.00124379553 | 0.001278992117 | 0.001281759997 | **0.001244656718** | 0.001248538351 | 0.001296415104 |
| Total verbs | 0.02235103763 | **0.0212275423** | 0.02146956801 | 0.02512176048 | 0.02319622885 | 0.02142171765 | 0.02319601713 | 0.02195797096 | 0.02216082649 | 0.02169235418 | 0.02133443839 | 0.02674580057 |
| Total pronoun | 0.004979578775 | 0.004974892093 | **0.004955408927** | 0.005024264471 | 0.004964500056 | 0.005003028063 | 0.004994292048 | 0.004982513006 | 0.004984716703 | 0.004986542166 | 0.004986964387 | 0.004979221869 |
| Total conjunction | 0.01847687076 | 0.01896668912 | 0.01868157543 | 0.01854897153 | 0.01851183634 | 0.0184303229 | 0.0185904104 | 0.01844198175 | 0.0185097485 | 0.0186029227 | **0.01838761507** | 0.01878459163 |
| Total determiners | 0.001272624752 | 0.001256711454 | 0.001250373971 | 0.00125551682 | 0.001327768201 | 0.001259583735 | **0.001240219521** | 0.001260014356 | 0.001349371417 | 0.001254533257 | 0.001293317415 | 0.001350737932 |
| Unique Word count | 0.01862358005 | 0.01894741878 | 0.01866685513 | 0.018967376 | 0.01915162541 | 0.01874813055 | **0.01848215867** | 0.02065582885 | 0.01985518488 | 0.02016355156 | 0.01870237352 | 0.0186963021 |
| Word Complexity | **0.0153121312** | 0.01548550613 | 0.01703385729 | 0.01954327202 | 0.01886643626 | 0.0173613293 | 0.01806525091 | 0.01726932331 | 0.01595747218 | 0.01565404409 | 0.01712179528 | 0.01802478775 |

Table 21: Results (MSE) for text features on Mockingjay for native spontaneous speech corpus (Mozilla Common Voice)