Exploring ASR-Based Wav2Vec2 for Automated Speech Disorder Assessment: Insights and Analysis

Tuan Nguyen, Corinne Fredouille, Alain Ghio, Mathieu Balaguer, Virginie Woisard

EXPLORING ASR-BASED WAV2VEC2 FOR AUTOMATED SPEECH DISORDER ASSESSMENT: INSIGHTS AND ANALYSIS

Tuan Nguyen[1], *Corinne Fredouille*[1], *Alain Ghio*[2], *Mathieu Balaguer*[3], *Virginie Woisard*[3,4]

[1]Avignon Université, [2]Aix-Marseille Université, [3]Université de Toulouse, [4]IUC Toulouse

ABSTRACT

With the rise of SSL and ASR technologies, the Wav2Vec2 ASR-based model has been fine-tuned for automated speech disorder quality assessment tasks, yielding impressive results and setting a new baseline for Head and Neck Cancer speech contexts. This demonstrates that the ASR dimension from Wav2Vec2 closely aligns with assessment dimensions. Despite its effectiveness, this system remains a black box with no clear interpretation of the connection between the model ASR dimension and clinical assessments. This paper presents the first analysis of this baseline model for speech quality assessment, focusing on intelligibility and severity tasks. We conduct a layer-wise analysis to identify key layers and compare different SSL and ASR Wav2Vec2 models based on pre-trained data. Additionally, post-hoc XAI methods, including Canonical Correlation Analysis (CCA) and visualization techniques, are used to track model evolution and visualize embeddings for enhanced interpretability.

Index Terms— Speech quality assessment, Interpretability, Pathological speech, ASR, SSL

1. INTRODUCTION

In the 21st century, people have a variety of communication choices, which make life easier. Nonetheless, verbal communication continues to play an irreplaceable role in the culture of humanity. Because only through verbal communication can we, as humans, fully comprehend and express all the intricate aspects of a subject, including emotions, and more. The lack of ability to communicate using speech often referred to as a speech disorder, represents a significant loss and necessitates the need for treatment. Speech disorders can be caused by various reasons such as Parkinson's disease, throat cancer, stroke, etc [1, 2]. This leads to various treatment methods specific to different disease stages and causes. Moreover, each patient may have different responses or adaptations to the same treatment method. Regular assessment of speech has to be conducted after a certain period of time to ensure the effectiveness of the treatment method as well as to monitor the patient's condition [3]. However, this assessment process demands substantial resources and expertise. Therefore, efforts to develop an automated speech quality assessment architecture as an alternative or support to this process have been increasing in recent years [4, 5]. Some automatic systems have shown robust performance and stability by learning from expert decisions [6, 7].

In 2024, Nguyen et al. [8] introduced a system that leverages the Automatic Speech Recognition (ASR) based Wav2Vec2 model [9], known for its strong capability in learning speech representations. This approach compared self-supervised learning (SSL) and the ASR dimension for speech quality assessment. It is shown that the fine-tuning of SSL models, using the ASR dimension, achieves the best results for the downstream task [8]. Despite its good performance, this assessment system is not perfect, and the actual behavior of the model is not well understood, which could pose significant problems, especially in the medical domain. Therefore, it is important to clarify and understand the decision-making process to ensure trustworthiness for humans, making these systems applicable in real-life scenarios.

In this paper, we present the first analysis of the ASR-based Wav2Vec2 model for speech disorder assessment, focusing on the prediction task of both intelligibility and severity scores. Our study begins with a layer-wise analysis of the model performance to identify which layers are most dedicated to these tasks, providing insights crucial for future research in the community. This analysis involves freezing and fine-tuning parts of the model up to selected layers to better understand how training should be approached in future. Furthermore, we conducted this layer-wise analysis across different versions of Wav2Vec2 models, based on the amount of pre-trained data. By comparing their performance in the speech quality assessment tasks, we aim to establish the initial connection between pre-trained SSL data and its impact on speech quality assessment, a question that has yet been addressed in prior work [8]. This exploration promises benefits not only for the speech disorder community but also for the SSL speech community, offering insights into the effects of data quantity and characteristics on model performance. In the parallel, we use a post-hoc eXplainable AI (XAI) method to gain more insights. Specifically, we utilize Canonical Correlation Analysis (CCA) to track how the model evolves across layers. Finally, we visualize the embedding information to enhance interpretability.

This research work was funded by the **LIAvignon AI Chair**

2. CORPUS

This paper utilized four different corpora. Different variants of Wav2vec2-based ASR were trained using the Common Voice corpus [10]. The BREF [11] corpus was used to develop a phoneme recognition system intended for subsequent layer-wise analysis. Other analysis experiments of the paper is based on two additional French speech corpora: C2SI [12] and SpeeCOmco [13, 14], recorded within the context of Head and Neck Cancers (HNC).

2.1. Common Voice

First introduced in 2019 by Mozilla, Common Voice responded to the problem of training data scarcity for speech technology, which was unavailable for most languages or otherwise prohibitively expensive at that time. It is a multilingual, open-sourced corpus designed specifically for ASR. Data collection is conducted through crowd-sourcing, where participants are asked to record their speech by reading sentences displayed on the screen via the project application or website. In the context of this paper, the French corpus (version 6.1) is used to align with the work of [8].

2.2. BREF

The BREF corpus, introduced in 1991, is a comparable data for French, similar to other major corpora in different languages such as TIMIT [15]. Specifically designed for assessing automatic speech recognition systems and studying phonological variations, this paper uses the BREF-120 corpus, featuring 120 speakers primarily from the Paris region. These participants were given a short reading test, which contains sentences collected from LeMonde newspaper. In total, 115 hours of read-speech data from 65 females and 47 males were collected.

2.3. C2SI

C2SI is a French corpus recorded from 2015 to 2017 as part of the Carcinologic Speech Severity Index (C2SI) INCa project, comprises speech recordings from healthy controls (HC) and patients diagnosed with Head and Neck Cancer. Recorded tasks include sustaining /a/ vowels, describing pictures, and reading text passages or pseudowords, facilitating analysis of speech distortion at multiple levels, including phonation, continuous speech production, and prosody-specific aspects.

This paper relies on different sets of recordings from 106 speakers - 82 patients and 24 HC - to conduct experiments. The first set is based on passage reading task. The first paragraph of *La Chèvre de monsieur Seguin*, a tale by Alphonse Daudet, was read by participants. A group of six experts then listened to these audio recordings and provided individual perceptual evaluations on speech intelligibility and severity. The evaluation is on a scale from 0 to 10, where a score of 0 represents severe speech disorder or unintelligible speech, and a score of 10 represents normal or highly intelligible speech. Another set of audio recordings used in this paper was based on the sustained vowel task. These recordings contain the production of 3 sustained vowels /a/. They could provide information on lower dimensions of speech such as voice level, stability, harmonics contents, etc.

2.4. SpeeCOmco

Comprising 27 patients suffering from Head and Neck Cancer, Speech and Communication in Oncology (SpeeCOmco) is an additional corpus for C2SI. Similar to the C2SI passage reading set, participants provided audio recordings of reading *La Chèvre de monsieur Seguin*, which were evaluated by the same panel of experts using the same metrics as in C2SI. In this study, SpeeCOmco is used to test extended speech quality assessment models, following the approach proposed by [8].

3. BASELINE SYSTEM

In the study by [8], the authors introduce an architecture using Wav2Vec2-based ASR as the initial component or feature extractor for speech quality assessment tasks. Subsequently, these features pass through intermediate layers, which include a statistical pooling layer (mean and standard deviation) and two linear layers of size 1024. Finally, a basic linear layer with a dimension of 1 is employed to generate output scores for intelligibility or severity. The model's performance is measured based on the Mean Squared Error (MSE) between the predicted scores and the ground truth. All layers, including Wav2Vec2, are updated during training to align them with the downstream task space.

They compared the *Wav2Vec2-3K-Large* and *Wav2Vec2-7K-Large* models, both fine-tuned on ASR tasks using the CommonVoice dataset. These models were pre-trained on self-supervised tasks using approximately 3000 hours and 7700 hours of healthy speech, respectively. The authors observed that starting with Wav2Vec2-based ASR outperformed Wav2Vec2-based SSL in fine-tuning for speech quality assessment, achieving superior results in both intelligibility and severity in the context of HNC patients. Another finding is also interesting since the 3K model performs better compared with 7K model despite less pre-trained SSL data. However, it is not totally clear what could caused this difference.

4. ANALYTICAL APPROACHES

This paper undertakes an analysis of the current baseline of automatic speech quality assessment, as described in the previous section. Firstly, we add more models using different pre-trained starting points to the original models proposed by Nguyen et al. [8]. Secondly, we extract frame-level features using passage reading , which are then use in Canoni-

cal Correlation Analysis (CCA) [16] framework to gain more insights about the models. This analysis is conducted layer-wise. In parallel, different layer embeddings are employed to train for the final task in order to identify the best layer. Finally, based on the results of CCA and layer-wise training, visualization using scatter plots is performed at the phoneme-level for read speech and at the frame-level for sustained vowels.

4.1. Additional models for comparative analysis

To observe the impact of the amount of data used for pre-training SSL, we compared our five different models (comprising two Wav2Vec2-based SSL and three Wav2Vec2-based ASR).

LeBenchmark recently published another pre-trained model, *Wav2Vec2-14K-Large*, which was trained on an additional 7000 hours of data from the 7K-model [17]. We fine-tuned this new 14K model, along with the previous *Wav2Vec2-1K-Large* pre-trained SSL model, using the Common Voice dataset for the ASR downstream task. This process was carried out using the end-to-end approach provided by SpeechBrain [18], which is identical to that used by the current baseline. For a comprehensive comparison, we additionally fine-tuned the 7K model using the Common Voice 6.1 dataset, aligning with other models, instead of following the approach in [8] where the available 7K model ASR from SpeechBrain, fine-tuned for Common Voice 14, was used.

Moving forward, we will employ the following labels: **1K-ASR**, **3K-ASR**, **7K-ASR**, and **14K-ASR** to represent ASR-based models, and **1K-SSL**, **3K-SSL**, **7K-SSL**, and **14K-SSL** for SSL-based models.

4.2. Layer-wise training

For a more robust investigation of the impact of each layer from the pre-trained Wav2Vec2-based ASR model, we trained this model in layer-wise manner for speech disorder assessment task. To investigate which layers in the ASR model provide relevant information for downstream tasks, we freeze all layers of the ASR-based Wav2Vec2 model and extract representations layer-wise for downstream training. In parallel, we conducted partial fine-tuning experiments, we conducted partial fine-tuning experiments, progressing layer by layer from one layer, to two layers, and so forth up to all layers.This approach aimed to explore potential reductions in computational costs, preservation of critical information,etc. These findings could provide valuable insights into feature analysis for the target task.

4.3. Canonical correlation analysis

Canonical Correlation Analysis (CCA) is a statistical method that measure the relationship between continuous-valued vectors by maximizing their linear projections' correlations. In the context of neural networks, CCA is well-known for evaluating the similarity of representations either between different models or within a single model. Its ability to remain invariant to linear transformations makes it particularly useful for this purpose. Consequently, CCA is commonly utilized to investigate the characteristics of deep learning models [19, 20, 21].

Since then, multiple variants of CCA have been introduced, but notable ones include Singular Vector CCA (SVCCA) [22] and Projection-Weighted CCA (PWCCA) [23]. Both SVCCA and PWCCA are designed to address the issue where not all dimensions (neurons) of a neural network layer may be utilized or active during the training task. SVCCA employs singular value decomposition (SVD) to remove low variance neurons that primarily introduce noise. On the other hand, PWCCA calculates a weighted mean of the correlation per neuron, assigning higher weights to directions that contribute more to the input. Both variants have demonstrated increased robustness compared to the original method. Given that SVCCA requires determining a threshold for the number of dimensions to be used, we decided to use PWCCA variant instead. For simplicity, we will refer to this variant as CCA from this point forward.

CCA is utilized to evaluate the similarity between layer representations of different Wav2Vec2 feature extractors, as described in Section 4.1, and their counterparts : (i) from corresponding layer of pre-trained ASR models (**CCA-ASR**), (ii) pre-trained SSL model (**CCA-SSL**) and (iii) from the acoustic information of phoneme recognition model (**CCA-phoneme**). For this analysis, we exclusively used models with equivalent pre-trained SSL data as the feature extractor variants (1K, 3K, 7K, and 14K).

4.4. Phoneme encoder

As observed in previous studies [24, 25, 26], phoneme information strongly influences the assessment of severity and intelligibility in speech disorders. To compare the information present in the system's feature extractor, we fine-tuned a phoneme recognition model using Wav2Vec2 **7K-SSL** in an end-to-end approach with the Connectionist Temporal Classification (CTC) loss function. The model was trained on the BREF corpus, with an 80-10-10% split between the training, validation, and test sets. It achieved its best performance, with a Phone Error Rate of 3.4%, on the test set. This model encodes meaningful phoneme representations, supported by prior research [27, 28]. Consequently, the output of the **last layer of this Wav2Vec2 model (layer 24)** was employed to analyze the phoneme information of the systems using *CCA-phoneme* as described in 4.3.

4.5. t-SNE visualization

Building upon the results obtained from the above analysis methods, we will further examine and **visualize the representations of last layer of feature extractor Wav2Vec2 -**

the 24th layer. To do that, we applied t–Stochastic Neighbourhood Embedding (t-SNE) method [29] to reduce the information from layer 24 to 2-dimensional plane. The focus here is to observe the system behavior, in terms of speech representation, at the phoneme level for read speech and at the frame level for the sustained vowel production task. For the phoneme level, the approach involves averaging all frames of each phoneme utterance using mean and standard deviation to generate a representative vector for the corresponding utterance.

5. INSIGHTS

All experiments were conducted for both intelligibility and severity assessment targets. Due to page limitations, readers should expect similar behavior across both tasks if only a single task is reported without specification. Additionally, all visualization clustering techniques were applied to the last layer of the feature extractor Wav2Vec2 (layer 24).

For all subsections 5.1, 5.2 and 5.3, the C2SI reading corpus was used to train the system and compare it with the baseline system. This corpus was also utilized to extract embeddings of the system and calculate CCA in subsections 5.4, 5.5, 5.6. In subsection 5.7, we used the C2SI corpus sustained vowel audios to extract embeddings and visualize them. The SpeCOmco corpus was used to report performance in subsections 5.1 and 5.3, as well as to visualize phoneme information in subsection 5.6 with t-SNE.

5.1. Relationship between pre-trained SSL data and speech quality assessment

Following the training and evaluation process outlined in [8], Table 1 illustrates the performance of additional models detailed in Section 4.1 using 10-fold validation, compared with the models from the same study. The results are reported on SpeCOmco corpus using MSE as described in Section 3

Comparing the feature extractors based on SSL, it's evident that the two additional models, 1K-SSL and 14K-SSL, have poorer performance compared to 3K-SSL and 7K-SSL. It's clear that having thousands of data points less significantly impacts the performance of 1K-SSL on both tasks. Despite the additional 7000 hours of data, 14K-SSL fails to perform anywhere close to the levels achieved by 3K-SSL and 7K-SSL. This result may be due to the additional 7000 hours of data in the 14K-SSL model, which includes Niger-Mali French [17], leading to a broader range of acoustic information captured by the model. This broader scope may not align well with the C2SI corpus, which primarily includes French mainland speakers, making it more challenging for tasks related to comprehensibility or intelligibility.

However, for the severity task, which focuses more on acoustic or low-level speech information, all SSL models have shown similar performance. This is logical since SSL models are known for their capability to capture speech representations well, especially acoustic information, resulting in similar performance among them.

Table 1: MSE results (mean ± std) for severity and intelligibility prediction tasks with different pre-trained models

	Intelligibility MSE	**Severity MSE**
Feature Extractor Based on Pre-trained SSL		
3K-SSL [8]	1.65 ± 0.43	2.1 ± 0.83
7K-SSL [8]	1.84 ± 0.49	1.83 ± 0.71
1K-SSL	3.65 ± 1.44	2.30 ± 0.53
14K-SSL	3.25 ± 1.4	2.23 ± 0.89
Feature Extractor Based on Pre-trained ASR		
3K-ASR baseline [8]	$\mathbf{0.73 \pm 0.18}$	$\mathbf{1.15 \pm 0.14}$
7K-ASR baseline [8]	$\boldsymbol{0.98 \pm 0.26}$	$\boldsymbol{1.15 \pm 0.16}$
1K-ASR	0.9 ± 0.17	1.33 ± 0.21
7K-ASR	$\boldsymbol{1.1 \pm 0.23}$	$\boldsymbol{1.76 \pm 0.50}$
14K-ASR	0.86 ± 0.19	1.28 ± 0.15

Looking at the ASR-based models, we observe a similar performance among all ASR models. Since the ASR models have been fine-tuned with the Common Voice dataset, this may lead to some forgetting of information from the original SSL model, effectively pulling all SSL models towards better alignment with the task. This suggest that the ASR dimension is closer and more important for speech disorder assessment. A notable observation is that after ASR fine-tuning, the 7K model exhibits poorer performance compared to the others. This could be attributed to more than half of the 7K pre-trained data (4000 hours) leaning towards spontaneous speech, which might be emphasized in ASR task.

5.2. Relationship between ASR performance and speech quality assessment

Comparing our new 7K-ASR model with the 7K-ASR baseline used in [8], we observe a decline in performance across both tasks, particularly in the severity task where both performance and stability are significantly worse. This decline can be attributed to the amount of ASR data used to obtain the ASR model. As indicated in [8], the pre-trained 7K-ASR model served as a baseline provided SpeechBrain, trained on CommonVoice 14, which includes approximately 400 more hours of ASR data compared to the version used in our study.

On the other hand, the *7K-ASR baseline* model yielded better ASR performance, with a Word Error Rate (WER) of 10.24%, while our version only achieved a WER of 13.45%. Similar behavior is observed among the 1K, 3K, and 14K ASR models. The 3K model achieved the lowest WER, whereas the 1K model had the highest. Interestingly, this ASR performance pattern aligns with the performance of

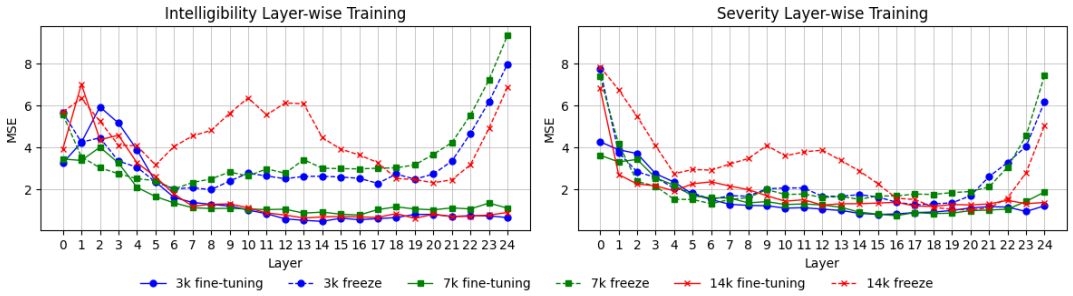


Fig. 1: Performance comparison of freeze and fine-tuned layer-wise feature extractor training on speech quality assessment tasks

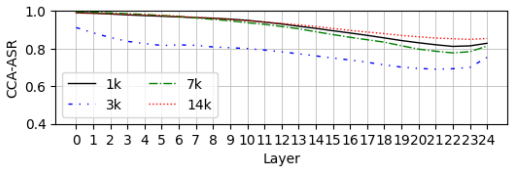
speech quality assessment, with the 3K model outperforming the others.

The new 7K-ASR model, despite having a better WER than the 1K and 14K models on the same version of Common-Voice corpus (13.45% WER compared to 16.64% and 15.52% WER, respectively), performed worse in the assessment tasks. This should be attributed to the amount of pre-trained SSL data, which is more lean towards spontaneous speech, as explained in section 5.1.

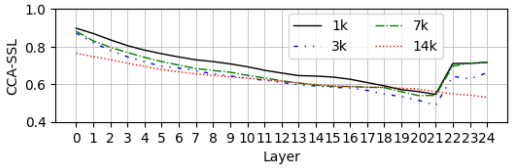
5.3. Layer-wise training analysis

Figure 1 illustrates the results of layer-wise training, both when freezing and fine-tuning the feature extractor model as described in Section 4.2, across different assessment tasks.

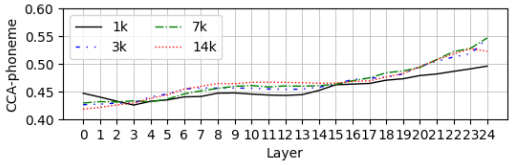
There is a similarity in performance between freeze and fine-tuning for the initial layers, indicating that the information in these layers remains relatively stable throughout fine-tuning and could be frozen to faster the process. Looking at the intelligibility task, there is a notable performance difference between freeze and fine-tuning from higher layers (starting from layer 8). In contrast, for the severity task, the representations of ASR models at intermediate layers exhibit similar performance with fine-tuning, with some layers achieving identical performance at layer 22. This suggests that ASR models encapsulate acoustic or low-level information for severity assessment across intermediate layers, requiring only minor adjustments in fine-tuning to achieve convergence. On the other hand, for intelligibility, the necessary information seems less clear with ASR models, requires fine-tuning to achieve optimal performance. Nevertheless, the overall trend for both fine-tuning and freezing indicates that intermediate layers (layer 8 onwards) contain relevant information for speech quality assessment. However, the 14K-ASR model exhibits a different behavior compared to the others from layer 9 to layer 16. This unusual behavior in the mid-layers suggests that the additional data from the African accent provide more different speech dimension which is not observed in the other models. Further analysis is required which could provide more insight into the impact



(a) CCA-ASR



(b) CCA-SSL



(c) CCA-phoneme

Fig. 2: CCA similarity between fine-tuned feature extractors with pre-trained ASR Wav2Vec2, SSL Wav2vec2 models and phoneme encoder

of data on SSL models and to understand why the ASR performance of the 14K model is inferior to that of the 3K or 7K models, as indicated in [17] despite having more pre-trained data.

5.4. Impact of fine-tuning on the ASR model

As observed in previous sections, the 3K-ASR baseline continues to demonstrate the best performance among all models. Therefore, starting from this section onward, all analyses are conducted using the **3K-ASR baseline** model.

The CCA-ASR analysis (Fig. 2a) revealed a consistent and relatively high similarity between the ASR pre-trained model and the feature extractor of the automatic assessment

system, gradually decreasing from layer 0 to layer 24 but consistently remained at a minimum level of around 0.7. Given this observation, it could be inferred that freezing the upper layers during training is viable, as their similarity is exceptionally high. This finding provides additional support for the conclusion drawn in [8] that pre-trained ASR serves as a strong initialization for both severity and intelligibility assessment.

5.5. The SSL representation within model

Looking at CCA-SSL (Fig. 2b), we can clearly observe a continuance decrease in similarity across layers. This aligns with CCA-ASR as well, where similarity primarily relates to ASR. However, in the last 3 layers, the CCA score sharply increases to nearly 0.8. This is intriguing as [20] demonstrated that SSL follows an encoder-decoder style, suggesting that these final layers closely resemble the input as if they are reconstructing the input signal. Additionally, a notable point is that the 14K model exhibits a different behavior compared to the others, with the CCA score consistently decreasing linearly.

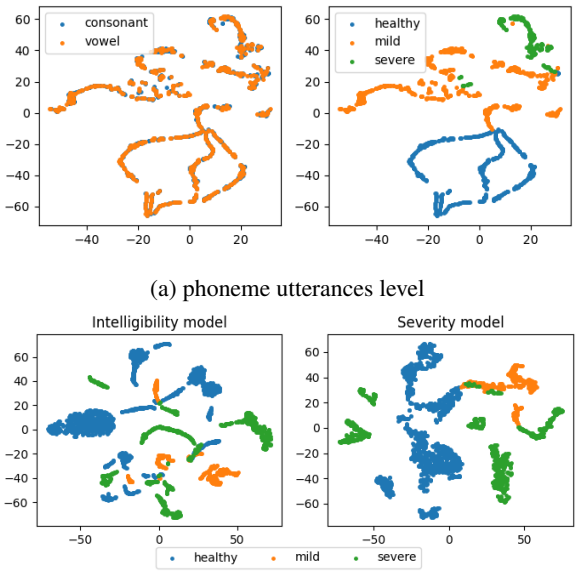
5.6. Phonetic information in feature extractor

When comparing the phoneme information encoded within the feature extractor with the last layer of the phoneme recognition system, we observed a relatively low similarity score of approximately 0.6 across layers (Fig. 2c).

The data points in Fig. 3a represent phoneme utterances, with each point labeled according to phoneme type (consonant or vowel) on the left, and speech quality (severe, mild, and healthy) on the right. The plot on the left indicates that feature extractors cannot distinguish between consonants and vowels, corroborating with CCA-phoneme. However, the plot on the right demonstrates that feature extractors can distinguish between patients based on their speech quality. This suggests that unlike ASR, the feature extractors may not clearly distinguish between phonemes (e.g., between vowels and consonants) but may instead capture lower-level phonetic information (nasal, labial, etc), consistent with findings from [25].

5.7. Voice production information in feature extractor

Figure 3b presents the visualization of sustained vowels at the frame level, with each point labeled according to the quality class of the respective patient. Since records related to the sustained vowels only contain a single vowel (in this particular case, vowel ”a”) pronounced continuously, it is typically used to measure voice quality [30]. Indeed, this type of audio, especially the stable part of the vowel, makes the measurement of voice characteristics such as jitter and breathiness easier. As observed in Fig. 3b, the three levels of speech quality (represented by the three patient groups - healthy, mild, and severe) are distinctly separated, with the severity



(a) phoneme utterances level

(b) vowel frame level

Fig. 3: 2D t-SNE visualization of the last Wav2Vec2 layer

task showing slightly clearer separation than intelligibility. This suggests a strong correlation between voice information and the model decision, which is logical considering that patients suffering from cancer often exhibit significant patterns of fatigue or discontinuous speech whereas healthy speakers do not. While testing the overall model performance using sustained vowel audio, the model MSE is notably high, with MSE values of 15.16 for the severity task and 17.11 for the intelligibility task. This is expected, as the model was trained primarily on read speech. Despite Wav2Vec2 ability to capture voice signal details, accurate scoring still requires speech-related dimensions like continuous speech, and phonetic variety. Combining this with Section 5.6, we can conclude that the model relies not only on speech dimensions such as articulation, resonance, and probably prosody (not studied here), but also needs to incorporate voice information to make final scoring decisions the most accurate.

6. CONCLUSION

This paper presents the first analysis on automatic speech quality assessment using ASR as a pre-trained starting point. We found that aligning the domain of pre-trained SSL data with downstream speech tasks (e.g., read speech with read speech) is more critical than the quantity of pre-trained data. Additionally, the experiments show a strong correlation between ASR performance and quality assessment, highlighting the impact of not only phonetic features but also low-dimensional voice signals. Finding of an unusual pattern in the 14K model’s layers suggests the need for further investigation into the effects of data quantity on model behavior.

7. REFERENCES

[1] Sonja C. Vernes, Jérôme Nicod, Fanny M. Elahi, Julie A. Coventry, Niamh Kenny, Anne-Marie Coupe, Louise E. Bird, Kay E. Davies, and Simon E. Fisher, "Functional genetic analysis of mutations implicated in a human speech and language disorder," *Human Molecular Genetics*, vol. 15, no. 21, pp. 3154–3167, 09 2006.

[2] Bastiaan R Bloem, Michael S Okun, and Christine Klein, "Parkinson's disease," *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[3] David G Pfister, Sharon Spencer, David M Brizel, Barbara Burtness, Paul M Busse, Jimmy J Caudell, Anthony J Cmelak, A Dimitrios Colevas, Frank Dunphy, David W Eisele, et al., "Head and neck cancers, version 1.2015," *Journal of the National Comprehensive Cancer Network*, vol. 13, no. 7, pp. 847–856, 2015.

[4] Mostafa Shahin, Usman Zafar, and Beena Ahmed, "The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 400–412, 2019.

[5] Duc Le, Keli Licata, Elizabeth Mercado, Carol Persad, and Emily Mower Provost, "Automatic analysis of speech quality for aphasia treatment," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4853–4857.

[6] Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, Mathieu Balaguer, and Virginie Woisard, "Interpretable assessment of speech intelligibility using deep learning: A case study on speech disorders due to head and neck cancers," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, May 2024, pp. 9170–9179, ELRA and ICCL.

[7] Sebastião Quintas, Julie Mauclair, Virginie Woisard, and Julien Pinquier, "Automatic Prediction of Speech Intelligibility Based on X-Vectors in the Context of Head and Neck Cancer," in *21st INTERSPEECH (2020)*, Shangai (fully virtual conference), China, Oct. 2020, International Speech Communication Association (ISCA), pp. 4976–4980, ISCA.

[8] Tuan Nguyen, Corinne Fredouille, Alain Ghio, Mathieu Balaguer, and Virginie Woisard, "Exploring pathological speech quality assessment with ASR-powered Wav2Vec2 in data-scarce context," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italia, May 2024, pp. 6935–6944, ELRA and ICCL.

[9] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: a framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2020, NIPS '20, Curran Associates Inc.

[10] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France, May 2020, pp. 4218–4222, European Language Resources Association.

[11] Lori F. Larnel, Jean-Luc Gauvain, and Maxine Eskenazi, "Bref, a large vocabulary spoken corpus for french," in *2nd European Conference on Speech Communication and Technology (Eurospeech 1991)*, 1991, pp. 505–508.

[12] Virginie Woisard, Corine Astésano, Mathieu Balaguer, Jérome Farinas, Corinne Fredouille, Pascal Gaillard, Alain Ghio, Laurence Giusti, Imed Laaridh, Muriel Lalain, et al., "C2si corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers," *Language Resources and Evaluation*, vol. 55, no. 1, pp. 173–190, 2021.

[13] Mathieu Balaguer, Julien Pinquier, Jérôme Farinas, and Virginie Woisard, "Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer: Preliminary validation," *International Journal of Language and Communication Disorders*, vol. 58, no. 1, pp. 39–51, Jan. 2023.

[14] Sebastião Quintas, Mathieu Balaguer, Julie Mauclair, Virginie Woisard, and Julien Pinquier, "Can we use speaker embeddings on spontaneous speech obtained from medical conversations to predict intelligibility?," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Taipei, Taiwan, 2023, pp. 1–7.

[15] Victor Zue, Stephanie Seneff, and James Glass, "Speech database development at mit: Timit and beyond," *Speech communication*, vol. 9, no. 4, pp. 351–356, 1990.

[16] Harold Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics: methodology and distribution*, pp. 162–190. Springer, 1992.

[17] Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Estève, Mickael Rouvier, Jerôme Goulian, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier, "Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech," *Computer Speech & Language*, vol. 86, pp. 101622, 2024.

[18] Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[19] Ankita Pasad, Ju-Chieh Chou, and Karen Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 914–921.

[20] Ankita Pasad, Bowen Shi, and Karen Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[21] Elena Voita, Rico Sennrich, and Ivan Titov, "The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 4396–4406, Association for Computational Linguistics.

[22] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein, "Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability," in *Advances in Neural Information Processing Systems*. 2017, vol. 30, Curran Associates, Inc.

[23] Ari Morcos, Maithra Raghu, and Samy Bengio, "Insights on representational similarity in neural networks with canonical correlation," in *Advances in Neural Information Processing Systems*. 2018, vol. 31, Curran Associates, Inc.

[24] Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, and Virginie Woisard, "Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders step 2: Contribution of the emergence of phonetic traits," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7387–7391.

[25] Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, and Virginie Woisard, "Validation of the Neuro-Concept Detector framework for the characterization of speech disorders: A comparative study including Dysarthria and Dysphonia," in *Proc. Interspeech 2022*, 2022, pp. 3638–3642.

[26] Mostafa Shahin, Usman Zafar, and Beena Ahmed, "The automatic detection of speech disorders in children: Challenges, opportunities, and preliminary results," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 2, pp. 400–412, 2019.

[27] Elena Voita, Rico Sennrich, and Ivan Titov, "The bottom-up evolution of representations in the transformer: A study with machine translation and language modeling objectives," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, Nov. 2019, pp. 4396–4406, Association for Computational Linguistics.

[28] Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, Seattle, Washington, July 2022, pp. 83–91, Association for Computational Linguistics.

[29] Laurens Van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[30] Bruce R. Gerratt, Jody Kreiman, and Marc Garellek, "Comparing measures of voice quality from sustained phonation and continuous speech," *Journal of Speech, Language, and Hearing Research*, vol. 59, no. 5, pp. 994–1001, 2016.