Exploring Pathological Speech Quality Assessment with ASR-Powered Wav2Vec2 in Data-Scarce Context

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

Exploring Pathological Speech Quality Assessment with ASR-Powered Wav2Vec2 in Data-Scarce Context

Tuan Nguyen[1], **Corinne Fredouille**[1], **Alain Ghio**[2], **Mathieu Balaguer**[3]
Virginie Woisard[3,4,5]
[1]LIA, Avignon Université, Avignon, France
[2]Aix-Marseille Univ, LPL, CNRS, Aix-en-Provence, France
[3]IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
[4]IUC Toulouse, CHU Toulouse, Service ORL de l'Hôpital Larrey, Toulouse, France
[5]Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France
{manh-tuan.nguyen, corinne.fredouille}@univ-avignon.fr
alain.ghio@univ-amu.fr, mathieu.balaguer@irit.fr, woisard.v@chu-toulouse.fr

Abstract

Automatic speech quality assessment has raised more attention as an alternative or support to traditional perceptual clinical evaluation. However, most research so far only gains good results on simple tasks such as binary classification, largely due to data scarcity. To deal with this challenge, current works tend to segment patients' audio files into many samples to augment the datasets. Nevertheless, this approach has limitations, as it indirectly relates overall audio scores to individual segments. This paper introduces a novel approach where the system learns at the audio level instead of segments despite data scarcity. This paper proposes to use the pre-trained Wav2Vec2 architecture for both SSL, and ASR as feature extractor in speech assessment. Carried out on the HNC dataset, our ASR-driven approach established a new baseline compared with other approaches, obtaining average $MSE = 0.73$ and $MSE = 1.15$ for the prediction of intelligibility and severity scores respectively, using only 95 training samples. It shows that the ASR based Wav2Vec2 model brings the best results and may indicate a strong correlation between ASR and speech quality assessment. We also measure its ability on variable segment durations and speech content, exploring factors influencing its decision.

Keywords: Speech intelligibility, Speech severity, Pathological speech, Automatic speech quality assessment, Self-supervised learning, Automatic speech processing

1. Introduction

In the past two decades, speech disorder has gained more and more attention from the computer science community. Not only in helping the patient in daily life task such as speech synthesis, automatic speech recognition,...(Hu et al., 2023) but also in supporting experts in evaluating the subject's condition (Castillo Guerra and Lovey, 2003). For the assessment of the patient's speech quality, a panel of experts will conduct evaluations based on perceptual information. The drawback of perceptual methods is that they are really costly, time-consuming and may not be always consistent. Therefore, automatic evaluation system based on perceptual information of experts as ground truth has gained attention from the research community. Automatic speech quality assessment can provide consistent performance compared to human experts. Hence, this system can encompass a wide range of tasks, from classification (distinguishing between speech disorders and normal speech) to rating aspects such as intelligibility or severity score.

Building upon the previous discussion, it is worth noting that the fundamental nature of classification has led to numerous research with promising results. Conversely, due to the complexity and lack of uniformity in rating scales, regression tasks have often received less attention. One of the challenges associated with regression is the scarcity of datasets with score rating compared to classification. Nonetheless, this exacerbates the challenge of developing automated assessment systems, particularly in the context of today's speech technologies, which are primarily data-driven and require a substantial amount of data for processing and generating accurate assessments.

This paper focuses more deeply on the domain of regression task, with the goal of exploring assessment scoring using automated systems. Current research often enriches data by segmenting patients' audio files into multiple samples (excluding non-speech segments) in an effort to expand datasets in order to leverage the use of deep learning. However, this approach has inherent limitations, as it indirectly associates overall audio scores with individual (smaller) segments, potentially missing crucial context. Responding to these limitations, we propose an approach in which the system is designed to learn and provide assessment scores on the whole audio rather than focusing on segments.

In the context of this paper, two pre-trained Wav2Vec2 models are proposed, each originating from distinct scenarios: one pre-trained via self-supervised learning (SSL) and the other through

fine-tuning for automatic speech recognition (ASR). These pre-trained models will be further fine-tuned to serve as feature extractors for the speech assessment task. Indeed, we assess the performance of these two architectures for the task of prediction of both intelligibility and severity scores on a speech corpus produced by patients suffering from Head and Neck Cancers. We compare this performance on existing approaches available in the literature, applied on the same corpus.
In addition, based on the best pre-trained Wav2Vec2 model we propose, further experiments and analyses were carried out to investigate how the quality of audio files, in terms of duration and therefore content, can impact model performance.

2. Corpus

The experiments are conducted on two different French speech corpora, C2SI (Woisard et al., 2021) and SpeeCOmco (Balaguer et al., 2023; Quintas et al., 2023), recorded in the context of Head and Neck Cancers (HNC). Additionally, a third speech corpus related to the Parkinson's disease, named Aix Hospital Neurology (AHN) corpus (Ghio et al., 2012), is also used to further analyze the generalization capabilities of the system. Lastly, Common Voice dataset (Ardila et al., 2020) is also used with the purpose of creating pre-trained Wav2Vec2 model through ASR tasks. They will be presented individually in the subsequent sections.

2.1. C2SI

C2SI corpus is a set of healthy control (HC) and patients that have been diagnosed with oral cavity or oropharyngeal cancer originating from different tumor locations. To maintain the stability of their speech impairment throughout the study corpus, patients were required to complete their treatment plan at least 6 months before enrollment and to be in clinical remission. Both the control and patient groups were instructed to record their speech in variety of tasks including sustaining vowels, picture description, delivering spontaneous speech, pseudo-word or passage reading.
In this study, our focus is on the passage reading task. The participants were asked to read the first paragraph of *La Chèvre de monsieur Seguin*, a short story by Alphonse Daudet. The audio files were then evaluated on different perceptual criteria by a panel of six expert clinicians. Within the context of the paper, intelligibility and severity metrics are taken into account. Each participant is given by a score from 0 to 10 reflecting their level of intelligibility or severity where a score of 0 indicates severe speech disorder or unintelligible speech, while a score of 10 is normal or related to highly intelligible speech. The final decision (either the score of intelligibility or severity) is considered to be the average of the six scores given by experts. A set of 105 speakers (84 patients and 21 controls) is used as train and valid set for our automatic system.

2.2. SpeeCOmco

Speech and communication in oncology (SpeeCOmco) is an additional corpus dedicated to HNC similar. It is composed of 27 patients, varying from really severe patients to quite normal ones. As in the C2SI corpus, these participants were also recorded while performing the same reading task (the first paragraph of *La Chèvre de monsieur Seguin*) among other speech production tasks. Their speech productions were evaluated by the same panel of expert clinicians using the same metrics. Therefore, SpeeCOmco is used as a test set in this paper, which has never been exposed to the model during the training phase.

2.3. Parkinson's disease - AHN

15 patients suffering from the Parkinson's disease are involved in this study. They are part of the larger Aix Hospital Neurology (AHN) corpus, which has 990 dysarthric patients and 160 healthy controls. Most of patients suffer from Parkinson's disease (601) or Parkinsonian syndromes (98). All participants of the AHN corpus were recorded using EVA workstation (Teston et al., 1999) on multiple speech tasks such as sustained vowels, text reading with several speed instructions, spontaneous speech and more.
The 15 patients were chosen because they provided diverse reading contexts, useful for further analysis in this paper. Indeed, as detailed in (Ghio et al., 2012), they had performed a double task of reading, the first one on the same text reported before (*La Chèvre de monsieur Seguin*) and the second one on an additional French text called *Le Cordonnier*. This set of 15 patients, exhibiting two different samples of read speech production, aims at assessing the generalization ability of the proposed system and at providing additional analyses considering varying linguistic content. A panel of eleven expert clinicians listened to these recordings and provided speech quality scores considering severity and intelligibility measurement. For this corpus, they used a 4-point scale, where 0 represents healthy or intelligible speech respectively, while the opposite represents the most severe condition or the lowest level of intelligibility. The mean score for severity is approximately 0.56, and for intelligibility is 0.3.

2.4. Common Voice

Common Voice is a multilingual open-source dataset created by Mozilla which is primarily designed for training ASR systems. It comprises a huge amount of transcribed speech data acquired through the crowdsourcing of reading text. The French corpus within Common Voice version 6.1 is a set of 375K utterances or approximately 475 hours of audio from 10K French speakers. This corpus is challenging for ASR system development since it contains different accents, noise and other variation factors. Because of these reasons, Common Voice is an excellent choice for developing a robust ASR model which is capable of covering multiple aspects of speech signal. Furthermore, since the audio data in Common Voice is collected from reading activities, it aligns with the same domain of audio that this paper uses for speech assessment.

3. Proposed Approach

Deep learning has recently seen significant developments and achieved impressive results across different domains. However, when working in the context of pathological speech, one of the most significant challenge is the limitation of data samples, which is crucial for deep learning. To address this limitation, a potential solution when it comes to data up-sampling is processing speech at the level of individual audio segments. A single audio recording can be segmented into several smaller segments, where each segment is assigned with the same score as the average score provided by the experts to the entire audio. Subsequently, the model is trained using these segment samples to generate prediction. The final decision is determined by averaging the predictions for samples from the same audio. Additionally, various data augmentation techniques, like speed and tempo distortion, may be applied to the segments. This method has been applied in certain works (Quintas et al., 2020; Vaysse et al., 2021) and has shown promising results.

In contrast to those promising results, we consider this approach to have a few concerns. In the first place, presuming the segment (local) score to be identical with overall (global) audio score could lead the system to behave differently than actual expert assessment. Indeed, each local segment should be assigned a score linked to its context. For instance, a patient might encounter difficulty in pronouncing some particular speech segments like məˈsjʌ ˈsəˈgɛ (ˈɱistər səgɛ) but not navɛ ʒamɛ y (ˈnɛvər hæd). Consequently, local scores should be adjusted accordingly, depending on segments and their degrees of production difficulty for the patient.

Furthermore, by repeating the same scores to numerous samples could also introduce overfitting issue. Segment samples may lack of higher-level speech information, such as prosodic prominence, rhythmic group coherence, and temporal dimensions, which are more represented at the entire audio level (depending on the length of segments compared with the entire audio file). Choosing the right segment duration is also a challenge. On the other hand, data augmentation techniques, which usually include speech signal modification such as the addition of noise or changes in speech rate, could lead to the loss of important information or an excess of unrealistic data.

In order for the automatic assessment system to behave as closely as possible to that of the experts, we propose to train the related model at the audio level without data augmentation. By doing this, different important information from speech to non-speech (extra-linguistic features) could be preserved without alteration. Pre-training technique is proposed here to handle the data scarcity.

3.1. Wav2Vec2 model

Wav2Vec2 (Baevski et al., 2020) was originally introduced by Facebook as a pre-trained model for SSL task. The model was learnt on vast amounts of unlabeled audio data, allowing it to extract meaningful representations from audio signals without the need for supervision. The pre-trained Wav2Vec2 model has demonstrated its effectiveness across various applications due to its ability to learn different dimensions of speech signals (Pasad et al., 2021). It can be fine-tuned on small amounts of label data to excel in tasks like speech recognition or speech classification (Tirronen et al., 2023a).

Wav2Vec2 consists of a feature extractor, transformer encoder and quantization block. The feature extractor, powered by a convolutional network, processes the raw audio into a latent representation. Then the transformer encoder captures contextual information and generates continuous embedding from latent space. Finally, the quantization block quantizes these continuous embeddings, creating an efficient representation for further processing. In this study, we plan to integrate LeBenchmark Wav2Vec2 large model, as introduced by (Evain et al., 2021), which has been pre-trained on French corpus to transfer its knowledge to our assessment system of speech quality by fine-tuning. This could help the system overcome the data scarcity while providing consistent and reliable results.

In the context of feature extraction, as observed in prior research (Mohamed et al., 2022; Violeta et al., 2022; Hu et al., 2023; Cho et al., 2023) it is evident that Wav2Vec2 pre-trained through self-supervision effectively captures relevant information for the assessment task. However, when applied to more complicated assessment task,

Wav2Vec2 pre-trained using SSL encounters difficulties in reaching a convergence point (Tirronen et al., 2023b). On the other hand, one of the well-established methods of speech intelligibility assessment relies on the comparison between the manual transcription of speech signal by expert with the "ground-truth" annotation of the linguistic content produced by speakers. In this way, different studies (Christensen et al., 2012; Van Nuffelen et al., 2009) have shown the potential correlation between intelligibility scores and word error rates computed from the outputs of an ASR system. With this in mind, we propose the use of Wav2Vec2 fine-tuned for ASR task as a pre-trained model for the assessment system, in comparison to the original Wav2Vec2 pre-trained through self-supervision.

3.1.1. Wav2Vec2 pre-trained through SSL

LeBenchmark introduces to community several pre-trained Wav2Vec2 models via SSL for French language. This paper compares the 2 distinct models: *Wav2Vec2-3K-Large* [1] and *Wav2Vec2-7K-Large* [2]. The difference between both models lies on the volume of unlabeled data the model was learnt on. One model was trained on approximately 3000 hours of healthy speech while the second one was trained on more extensive dataset consisting of 7700 hours of such data. From now, these two regression models will be referred as **3K-SSL** and **7K-SSL** for ease of reference.

3.1.2. Wav2Vec2 pre-trained through ASR

Taking the two pre-trained models described in section 3.1.1, we fine-tuned them on an ASR downstream task with the Common Voice French dataset. The 7K model, which serves as an ASR baseline, has been provided by SpeechBrain (Ravanelli et al., 2021), an open-source toolkit dedicated to automatic speech processing. This model achieved a Word Error Rate (WER) of 9.96% on Common Voice corpus. For the 3K model, we adopted an end-to-end fine-tuning approach with Connectionist Temporal Classification (CTC) loss function. After 50 epochs, our ASR model achieved its best WER performance at 13.57%. Finally, we extracted the Wav2Vec2 block from both systems to be placed in the feature extraction in section 3.2. Moving forward, we will use the labels **3K-ASR** and **7K-ASR** to denote these two models.

[1] https://huggingface.co/LeBenchmark/Wav2Vec2-FR-3K-large

[2] https://huggingface.co/LeBenchmark/Wav2Vec2-FR-7K-large

3.2. Speech Assessment Architecture

In this section, we look into the system architecture, which combines Wav2Vec2 with additional layers to create an end-to-end solution optimized for regression tasks. The model architecture can be broken down into three key components:

1. **Feature Extractor**: In the initial stage, Wav2Vec2, as detailed in section 3.1, takes on the role of processing raw audio data to derive meaningful representation with dimension of 1024. Wav2Vec2 will not remain static but will undergo fine-tuning to transfer the knowledge it acquired from its pre-trained task. The output of Wav2Vec2 from the last layer (layer 24) will be passed to the next intermediate layers.

2. **Intermediate Layers**: This part includes a pooling layer followed by 2 linear layers.

   (a) *Pooling layer*: Reduce the temporal dimension from Wav2Vec2 to ensure the consistent shape among data samples using *Statistic Pooling*, which has second-order calculations of standard deviation. Other papers (Okabe et al., 2018; Boureau et al., 2010) have shown that, by considering both mean and standard deviation, the system can effectively capture not only the information present throughout the whole sequence but also the fluctuations in the data.

   (b) *Linear layers*: 2 linear layers with dimension of 1024. They are responsible for learning and identifying complicated patterns from the feature extraction stage into a meaningful information connected to final decision.

3. **Output Layers**: A simple linear layer with dimension of 1 corresponding to predictive score. The prediction is evaluated using MSE metric.

3.3. 10-fold validation

A 10-fold validation technique was applied to the training phase due to the data scarcity. At each fold, 90% of data or approximately 95 speakers were involved in the training process and 10 were set aside as the validation set. The system was trained for 20 epochs using a small batch size of 1. This training was carried out on a single NVIDIA Tesla A100 with 40GB of VRAM. This choice of batch size aims to introduce the randomness into the training process (Keskar et al., 2017). This randomness was intended to enhance the generalization of model and to prevent the risk of overfitting. Also due to the size of Wav2Vec2 with more than 300M trainable parameters, using small batch size

helps reduce the computational resource demands. SpeeCOmco corpus was used as test set similarly for every fold. The entire process is implemented using SpeechBrain[3].

4. Results and Discussion

4.1. Baseline performance

To evaluate the effectiveness of our proposal, we compare our best system with existing works, for which the same dataset, SpeeCOmCo, was involved for both training and testing, permitting comparison. The first baseline is established by an automatic system using a Shallow Neural Network based on speaker embedding extraction (*x-vectors* or ECAPA-TDNN) (Quintas et al., 2023). As reported by the authors, the automatic prediction system provides a best MSE result of 1.75 (RMSE of 1.32 in the paper) for speech intelligibility assessment, and 1.91 (RMSE of 1.38) for speech severity, using reading passage task. The second baseline relies on a Convolution Neural Network (CNN) based system, trained for a typical French phone classification task to provide a healthy speech representation, and coupled with a Shallow Neural Network for score prediction (Abderrazek, 2023; Abderrazek et al., 2024). This baseline system, which aims at providing interpretable phonetic knowledge with the prediction score, reaches a best MSE result of 2.97 for speech intelligibility, and 3.05 for speech severity.

Remarkably, most of the systems we proposed consistently outperform these existing baseline systems. With our best model, we achieved between **58% to 75% MSE reduction** compared with the two baselines reported above for intelligibility assessment and between **40% to 62% to MSE reduction** for severity assessment within the context of SpeeCOmco corpus. This outstanding performance puts the system at the forefront of the field, highlighting its ability in speech quality assessment.

4.2. Comparison of feature extractors

The performance of the different regression models, described in sections 3.1.1 and 3.1.2, was measured on SpeeCOmco corpus. The results presented in table 1 indicate highly promising performance in tasks related to speech intelligibility and speech severity assessment. Our proposed architectures achieved outstanding results without requiring data augmentation. Specifically, we obtained an average best MSE at 0.73 for the intelligibility prediction task and 1.15 for the severity prediction task. Among the four different pre-trained Wav2Vec2 models, it is interesting to note that 3K-SSL model brings the worst performance in term of severity assessment. Meanwhile, the 7K-SSL model performs the poorest in the intelligibility prediction task. It is not entirely clear why these models exhibit distinct behaviors, but our hypothesis is related to approximately 4,700 hours of differing data. This additional data, sourced from the European Parliament event, may include French non-native speakers who exhibit distinct articulation, accent, or atypical speech patterns. These differences could cause the slightly lower performance of the 7K-SSL model in intelligibility assessment. In contrast, they are essential for severity assessment, help the model is more likely to have captured richer speech representations, including vocal bursts, prosody, and other acoustic cues. Further investigation is required to confirm this hypothesis.

While comparing feature extractor based on pre-trained SSL with pre-trained ASR, it is surprising that pre-trained ASR extractor outperformed the pre-trained SSL one. Not only with better average MSE, pre-trained ASR extractor also shows a more consistent performance with a significantly smaller standard deviation. The system itself is less sensitive to data variability and results in the learning of more robust features. This finding highlights a concrete connection between the ASR task and speech assessment, potentially shedding new light on future research directions. This implies the potential for considering the Wav2Vec2 ASR component as a feature extractor. ASR, being a more specialized task compared to SSL, may offer a more straightforward interpretability, making it a more practical choice for feature extraction.

Additionally, it is worth to mention that 3K-ASR model achieves better results compared with 7K-ASR model on both assessment tasks even the 7K-ASR model yields a much better general WER on Common Voice corpus (9.96%). This once again emphasizes the argument regarding the bias toward healthy speech in these models.

Looking across the tasks, both types of feature extractor perform slightly better and more stable with intelligibility assessment than severity assessment.

	Intelligibility MSE	Severity MSE
3K-SSL	1.65 ± 0.43	2.1 ± 0.83
7K-SSL	1.84 ± 0.49	1.83 ± 0.71
3K-ASR	**0.73** $\mathbf{\pm 0.18}$	**1.15** $\mathbf{\pm 0.14}$
7K-ASR	0.98 ± 0.26	1.15 ± 0.16

Table 1: MSE Results on Severity and Intelligibility prediction tasks at the Audio Level according to different pre-trained models (mean and standard deviation considering the 10-fold validation)

[3]https://speechbrain.github.io

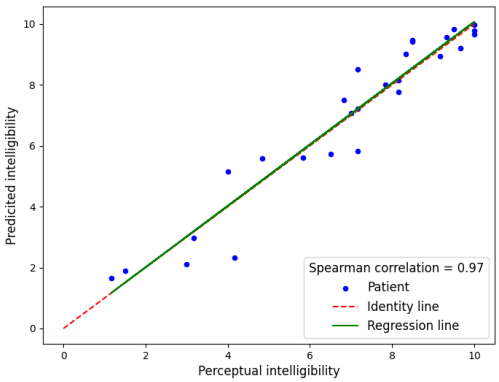


Figure 1: Scatter plot of intelligibility prediction

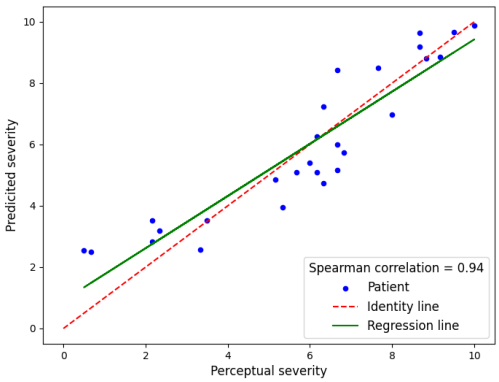


Figure 2: Scatter plot of severity prediction

However, with ASR based feature extractor, the difference between two tasks narrows, making it the winner.
In the following sections, our analysis focuses on the 3K-ASR model, as it obtained the best results and the highest level of stability. Due to the stability of 3K-ASR model with just 0.18 and 0.14 of standard deviation in both assessment tasks, without losing generality, we can analyze any random fold within the 10-fold model cross-validation for discussion purpose. In the following sections, we are discussing the first fold where the model has an MSE of 0.54 for intelligibility and 1.05 for severity in following sections.

4.3. Prediction evaluation

A more detail comparison between the prediction scores provided by the 3K-ASR model with the perceptual scores given by the experts will be present in the following sections.

4.3.1. Intelligibility prediction

Figure 1 illustrates the scatter plot of intelligibility scores predicted by the automatic system versus the perceptual scores given by the experts. This visualization provides insight into the relationship between predicted and target scores.
On the plot, the red dashed line represents the identity line, where $X = Y$, indicating the perfect match between predicted and target scores. The green solid line indicates the regression line which shows the best linear relationship between perceptual scores and predictions as determined through regression analysis.
The plot clearly shows a remarkably high correlation between the predictions and targets, visualized by the fact that regression and identity lines almost overlap. This high correlation level is confirmed by Spearman's correlation coefficient of 0.97. Across all 10-fold validation, the correlation level ranges from 0.94-0.97, with p-values always less than 0.01, highlighting a statistical significance.

4.3.2. Severity prediction

Similar to section 4.3.1, figure 2 depicts the predicted severity scores versus the perceptual scores. The level of correlation with the severity assessment task is slightly worse than with the intelligibility assessment, as seen in the figure. The regression line gradually diverges from the identity line, where the lower segment of the regression line tilts upward, and the final segment does the opposite. It suggests an overestimate for severe patients, as indicated by the lower segment of the plot, and an underestimate for mild patients observed at the point of intersection of the two lines.

4.4. Generalization and Overfitting

4.4.1. Learning curves

One of the main concerns regarding performance is the possibility of overfitting, particularly when considering the 95 training samples available for each fold and a large model size. However, using 10-fold cross-validation with a fixed test set that the system has never seen before, the model continues to perform well on the test data with high stability. Furthermore, when examining the loss curve for the intelligibility task, as shown in figure 3, it clearly indicates that there is no overfitting problem. At the convergence point, the difference between the training and validation loss is minimal. Both loss curves steadily decrease during training until they reach an optimal point. The same behavior can be observed for the severity performance. Therefore, it appears that overfitting is not a significant issue, which further strengthens the model's generalization capabilities.

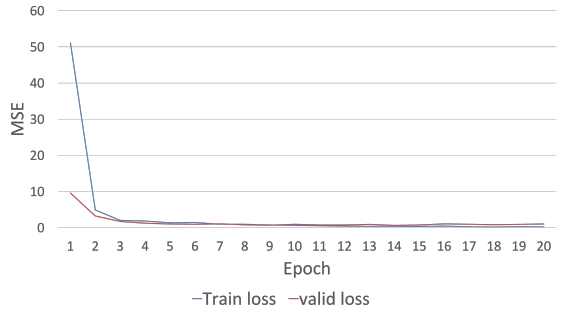


Figure 3: Train and validation loss (MSE) curves from a random fold

4.4.2. Cross-domain testing

Another factor that can reinforce the generalization is the application of Cross-domain testing technique. To do that, the model is evaluated using data from a different but related domain, specifically, the AHN dataset described in sec 2.3
Normally, patients suffering from Parkinson's disease experience different and relatively mild speech symptoms compared to patients with HNC. This could lead to consider the AHN dataset as a new domain to our system.
By evaluating the generalization ability of the 3K-ASR model using the audio *La Chèvre de monsieur Seguin*, we observed a consistent pattern in its predictions for intelligibility and severity assessment. By converting the scale to a range of 0-10 (which is the scale of the system), we can easily see that the predictions made by 3K-ASR are in line with the perceptual evaluations of experts. Despite the cross-domain nature, the model consistently shows good performance, achieving an *MSE=0.22* for intelligibility and an *MSE=0.37* for the severity task. One possible reason why the model achieved better performance with Parkinson's patients compared to SpeeCOmco corpus due to the fact that most of the patients are not severely affected, as indicated by the average score provided in section 2.3.
Based on the findings from the cross-domain corpus and our observations in section 4.4.1, we can reinforce our conclusion that overfitting is not a significant concern.

5. Content impact analysis

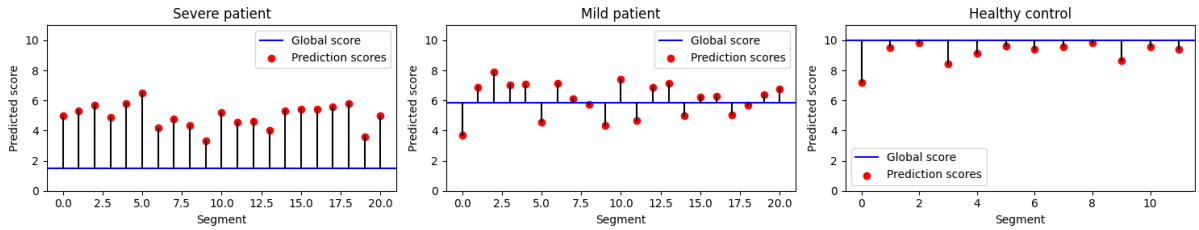
This section presents first observations concerning the influences of speech content on the behavior of the speech quality assessment model.

5.1. Limited content

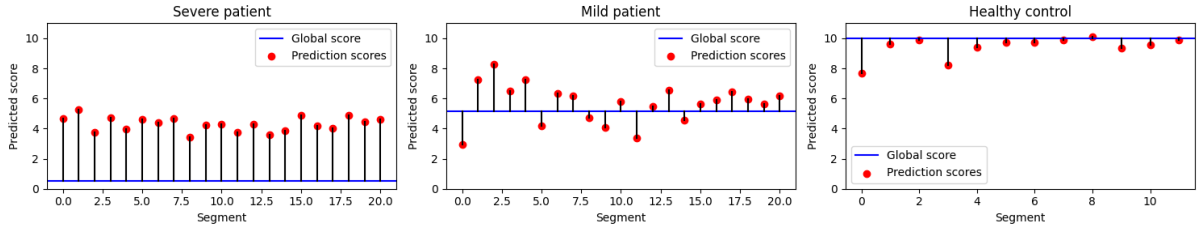
As argued in section 3, training the model on the entire audio file ensures that the model can encompass the entire content, context, and various aspects of speech, making the model behavior closer to the global assessment performed by experts. Furthermore, considering short segments of speech extracted from the main audio file, the scores may vary depending on the patient's condition. This section analyzes in detail how the model behaves with shorter test segments and examines their impact on the model performance.
Based on provided arguments, our hypothesis is that the model should provide distinct scores to different segments, depending on the content expressed in those segments. The average scores across segments are close or relatively close to the global score.
For healthy control group, as they do not experience any symptoms, their segment scores should be similar and close to the global score. This pattern also applies for severe patients, who has extreme difficulty in speaking, resulting in similar low segment scores. On the other hand, mild patients whose symptoms are not so significant should not experience extreme difficulty in pronouncing all types of phonemes. Consequently, it is expected that the segment scores for this group vary the most, with both high and low segment scores.
To validate this hypothesis, the model generated predictions at a two-second segment level. Three speakers were carefully selected to ensure that they well represent their respective groups for examination:

- **Severe group**: a patient with perceptual score of 1.5 for intelligibility and 0.5 for severity.
- **Mild group**: a patient with scores of 5.8 for intelligibility and 5.1 for severity.
- **Control group**: a healthy speaker with scores of 10 for both intelligibility and severity.

In figure 4, the X-axis represents the index of segment along the temporal timeline while the Y-axis indicates model prediction. The blue line indicates the global score over time while red dots are the predicted scores at the two-second segment level. The vertical black lines represent the differences between the predicted score and the global score.
As expected, figure 4 demonstrates that the model consistently generated scores on different segments for severe patient and control group. With the mild group, scores vary more around the reference line. Nevertheless, when considering the severe patient, the predictions deviate significantly from reference line, exhibiting an overestimation of the scores by the model. This indicates that, with limited content information, although the model seems to recognize the consistency in the pathology between segments for severe group, it struggles to make accurate predictions and tends to give an overestimation assessment. In a more subtle way, an underestimation of predicted scores can



(a) Intelligibility prediction for two-second segment



(b) Severity prediction for two-second segment

Figure 4: Model behavior at the segment level

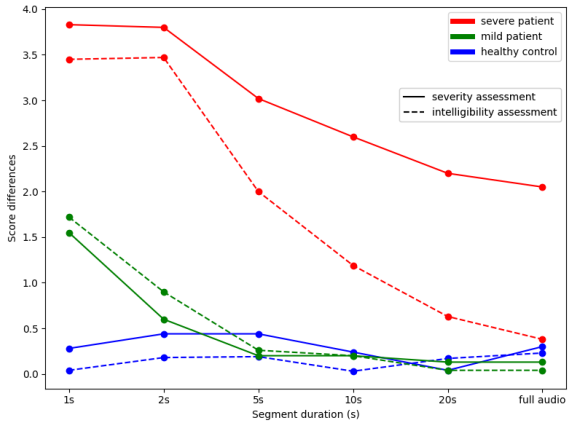


Figure 5: Absolute error variation across different segment durations

be observed in contrast for the control group.

To further investigate, we applied the same analysis method to segments of varying duration : one, two, five, ten and twenty seconds with same patients. The average scores of segments for each patient were compared will the full audio predictions by calculating the absolute error between the targets and these scores. The results are displayed in figure 5.

The solid line in the figure indicates the severity assessment task and the dashed line corresponds to the intelligibility assessment. Patients' speech disorders are represented by different colors: red represents severe patient, green represents mild patient and blue for healthy control. The graph demonstrates that as the duration increases, the model performance improves, resulting in a reduction in absolute error. It is logical since increasing duration means that the audio contains more content information and provides the model with more dimension to process. However, the segment duration does not seem to affect the healthy control group, as the healthy line remains consistent across various durations with good performance. On the other hand, for severe patient who experiences strong speech disorders, there is a close relationship with the amount of contents. The more contents it has, the better the predictions are. Same behavior can be observed with the mild group but at a lower level.

5.2. Different content

By comparing readings from both texts *La Chèvre de monsieur Seguin* and *Le Cordonnier*, within the AHN corpus, we also observed a high alignment in the predictions made by the automatic system. Despite the different phonetic contexts of these readings, the system generates high consistent predictions. Indeed, by applying the Spearman's correlation between the model decisions obtained with the two text readings, we obtained high correlation rates of 0.96 and 0.95 for speech intelligibility and severity assessment respectively, both with a p-value of less than 0.01 indicating a statistically significant correlation. The consistent performance of the model across different contexts indicates that different contents do not affect the final decision.

6. Conclusion

This paper proposes a novel approach dedicated to the assessment of speech quality to train model on the entire audio despite the data scarcity. To achieve this, we use a regression system with a Wav2Vec2 based model that serves as a feature

extractor. Through experimentation, we find that fine-tuning Wav2Vec2 on ASR yields better results compared to a typical pre-trained Wav2Vec2 SSL when it is fine-tuned for a final regression assessment task. Only using 95 training samples, we obtained the best result $MSE = 0.73$ for intelligibility prediction and $MSE = 1.15$ for severity prediction. Currently, the proposed system outperforms all previous competitors, achieving a significant **58% MSE reduction** for intelligibility assessment and a **41% MSE reduction** for severity assessment within the context of SpeeCOmco corpus, thereby setting a new performance baseline. From this, it can be concluded that the ASR pre-trained context is closely related to the speech quality assessment, involving both intelligibility and severity.
Moreover, additional analyses showed that the duration of test segments does impact the model decisions. This is particularly true for severe patients; the degree of impact decreases with the patient's speech impairment, as observed for healthy patients who seem less affected. Regarding now changes in linguistic content (between training and testing), the model does not seem to be significantly affected. Future work will take a closer look at segment content and, in particular, how certain phonetic contexts might influence the decision of predictive models.

7. Acknowledgements

The authors express their heartfelt gratitude to all anonymous reviewers for their insightful comments and suggestions. Additionally, we acknowledge the support of the **LIAvignon AI Chair**[4] for funding this research work.

8. Bibliographical References

Sondes Abderrazek. 2023. *Assessment of Speech Intelligibility using Deep Learning – Towards Enhanced Interpretability in Clinical Phonetics*. Theses, Université d'Avignon.

Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Mathieu Balaguer, and Virginie Woisard. 2024. Interpretable assessment of speech intelligibility using deep learning: A case study on speech disorders due to head and neck cancerst. In *2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, Torino, Italia.

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

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

Mathieu Balaguer, Julien Pinquier, Jérôme Farinas, and Virginie Woisard. 2023. Development of a holistic communication score (HoCoS) in patients treated for oral or oropharyngeal cancer: Preliminary validation. *International Journal of Language and Communication Disorders*, 58(1):39–51.

Y-Lan Boureau, Jean Ponce, and Yann LeCun. 2010. A theoretical analysis of feature pooling in visual recognition. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, pages 111–118, Haifa, Israel.

Eduardo Castillo Guerra and Denis F. Lovey. 2003. A modern approach to dysarthria classification. In *Proceedings of the 25th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE Cat. No.03CH37439)*, volume 3, pages 2257–2260 Vol.3, Cancun, Mexico.

Cheol Jun Cho, Peter Wu, Abdelrahman Mohamed, and Gopala K. Anumanchipalli. 2023. Evidence of vocal tract articulation in self-supervised learning of speech. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes Island, Greece.

Heidi Christensen, Stuart Cunningham, Charles Fox, Phil Green, and Thomas Hain. 2012. A comparative study of adaptive, automatic recognition of disordered speech. In *Proc. Interspeech 2012*, pages 1776–1779, Portland, OR, USA.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong, Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2021. LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from

[4] https://liavignon.fr

Speech. In *Proc. Interspeech 2021*, pages 1439–1443, Brno, Czech Republic.

Alain Ghio, Gilles Pouchoulin, Bernard Teston, Serge Pinto, Corinne Fredouille, Céline De Looze, Danièle Robert, François Viallet, and Antoine Giovanni. 2012. How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Communication*, 54(5):664–679. Advanced Voice Function Assessment.

Shujie Hu, Xurong Xie, Zengrui Jin, Mengzhe Geng, Yi Wang, Mingyu Cui, Jiajun Deng, Xunying Liu, and Helen Meng. 2023. Exploring self-supervised pre-trained asr models for dysarthric and elderly speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes Island, Greece.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. 2017. On large-batch training for deep learning: Generalization gap and sharp minima. In *International Conference on Learning Representations*.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Koji Okabe, Takafumi Koshinaka, and Koichi Shinoda. 2018. Attentive Statistics Pooling for Deep Speaker Embedding. In *Proc. Interspeech 2018*, pages 2252–2256, Hyderabad, India.

Abhimanyu Pasad, Ju-Chieh Chou, and Karen Livescu. 2021. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 914–921, Cartagena, Colombia.

Sebastião Quintas, Julie Mauclair, Virginie Woisard, and Julien Pinquier. 2020. Automatic Prediction of Speech Intelligibility Based on X-Vectors in the Context of Head and Neck Cancer. In *Proc. Interspeech 2020*, pages 4976–4980, Shangai, China.

Sebastião Quintas, Mathieu Balaguer, Julie Mauclair, Virginie Woisard, and Julien Pinquier. 2023. 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)*, pages 1–7, Taipei, Taiwan.

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. 2021. Speechbrain: A general-purpose speech toolkit.

Bernard Teston, Alain Ghio, and Benoît Galindo. 1999. A multisensor data acquisition and processing system for speech production investigation. In *International Congress of Phonetic Sciences (ICPhS)*, pages 2251–2254, San Francisco, United States. University of California.

Saska Tirronen, Farhad Javanmardi, Manila Kodali, Sudarsana Reddy Kadiri, and Paavo Alku. 2023a. Utilizing wav2vec in database-independent voice disorder detection. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, Rhodes Island, Greece.

Saska Tirronen, Sudarsana Reddy Kadiri, and Paavo Alku. 2023b. Hierarchical multi-class classification of voice disorders using self-supervised models and glottal features. *IEEE Open Journal of Signal Processing*, 4:80–88.

Gwen Van Nuffelen, Catherine Middag, Marc De Bodt, and Jean-Pierre Martens. 2009. Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language & Communication Disorders*, 44(5):716–730.

Robin Vaysse, Jérôme Farinas, Corine Astésano, and Régine André-Obrecht. 2021. Automatic Extraction of Speech Rhythm Descriptors for Speech Intelligibility Assessment in the Context of Head and Neck Cancers. In *Proc. Interspeech 2021*, pages 1912–1916, Brno, Czech Republic.

Lester Phillip Violeta, Wen Chin Huang, and Tomoki Toda. 2022. Investigating Self-supervised Pretraining Frameworks for Pathological Speech Recognition. In *Proc. Interspeech 2022*, pages 41–45, Incheon, Korea.

Virginie Woisard, Corine Astésano, Mathieu Balaguer, Jérome Farinas, Corinne Fredouille, Pascal Gaillard, Alain Ghio, Laurence Giusti, Imed Laaridh, Muriel Lalain, et al. 2021. C2si corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, 55:173–190.