Interpretable Assessment of Speech Intelligibility using Deep Learning: A Case Study on Speech Disorders due to Head and Neck Cancers

Sondes Abderrazek[1], **Corinne Fredouille**[1], **Alain Ghio**[2], **Muriel Lalain**[2]
Christine Meunier[2], **Mathieu Balaguer**[3], **Virginie Woisard**[3,4,5]
[1]LIA, Avignon Université, 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
sondes.abderazek@gmail.com,corinne.fredouille@univ-avignon.fr

Abstract

This paper sheds light on a relatively unexplored area which is deep learning interpretability for speech disorder assessment and characterization. Building upon a state-of-the-art methodology for the explainability and interpretability of hidden representation inside a deep-learning speech model, we provide a deeper understanding and interpretation of the final intelligibility assessment of patients experiencing speech disorders due to Head and Neck Cancers (HNC). Promising results have been obtained regarding the prediction of speech intelligibility and severity of HNC patients while giving relevant interpretations of the final assessment both at the phonemes and phonetic feature levels. The potential of this approach becomes evident as clinicians can acquire more valuable insights for speech therapy. Indeed, this can help identify the specific linguistic units that affect intelligibility from an acoustic point of view and enable the development of tailored rehabilitation protocols to improve the patient's ability to communicate effectively, and thus, the patient's quality of life.

Keywords: Speech Disorders, Speech Intelligibility assessment, Deep Learning, Interpretability and explainability, Head and Neck Cancers, Clinical phonetics

1. Introduction

Speech intelligibility is an essential component of effective communication. It refers to the extent to which a speaker's intended message can be understood by a listener. This fundamental component of communication can be hampered as a consequence of speech disorders, leading to a deteriorated quality of life for affected individuals (Kent, 1992). In the context of Head and Neck Cancers (HNC), speech can be notably impacted by the presence of tumors within the speech production system. Nonetheless, the primary cause of speech impairment in HNC typically arises from the treatments administered to manage tumors, such as surgery, radiotherapy, chemotherapy, or a combination thereof (Meyer et al., 2004). In such cases, the evaluation of speech quality is crucial to assess the communication deficit of patients and develop targeted treatment plans. Traditionally perceived as the gold standard in clinical practice for evaluating speech disorders, perceptual measures were often criticized for their inherent subjectivity and lack of reproducibility (Revis, 2004; Pommée et al., 2021). In recent years, the application of deep learning (DL) for the automatic assessment of speech disorders has emerged as a promising tool to complete and enhance perceptual measures (Bin et al., 2019; Quintas et al., 2020; Gupta et al., 2021). These tools have shown their ability to yield more reliable and specific measurements providing clinicians with access to new information contained in the speech signal.

Despite the advances in this direction, only a few studies addressed this subject from a DL interpretability point of view. In this context, we can find a research work that was conducted with a focus on dysarthric speech by Tu Ming et al. (Tu et al., 2017). The authors trained a model to predict the severity of dysarthric speech from the input signal. On the other hand, they took steps to make the model interpretable by incorporating a specific bottleneck layer. They used transfer learning to learn both clinically-interpretable labels perceived by speech-language pathologists (SLPs) (e.g. vocal quality and articulatory precision) and the final severity score. The result is a model that not only improved the accuracy of dysarthria assessment but also provided justifications for its predictions by exhibiting high correlations with the interpretable bottleneck features. An extension of this work was recently proposed by (Xu et al., 2023). Instead of relying on perceptual labels provided by SLPs, the authors of this work trained the interpretable layer to learn four acoustic features that characterize different aspects of dysarthria (articulatory precision, consonant-vowel transition precision, hypernasality,

LREC-COLING 2024, pages 9170–9179

and vocal quality). Authors extracted these acoustic features from the speech samples they have in possession. They also applied SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) as an explanation tool to further analyze the contribution of each acoustic feature in the interpretable layer to the final prediction. Very close to Tu Ming et al., authors in (Korzekwa et al., 2019) proposed a DL model for the detection and reconstruction of dysarthric speech. Their model not only provides interpretable characteristics of dysarthria but also tries to reconstruct healthy speech.

Although these works address one major requirement of DL in a clinical application which is DL interpretability, their methodology based on the incorporation of a bottleneck layer raises the need for a large dataset of speech pathology. This need arises because they trained their DL models from scratch using a dataset of dysarthric speech. For Tu Ming et al. (Tu et al., 2017), this data requirement is even more important since they need extra labels, in addition to the severity score, for the training of the bottleneck layer. In addition, if we consider that these intermediate labels could be subjective since they are provided by humans (SLPs), this leads to the incorporation of a subjectivity characteristic in the interpretability of the final score. Recently, the authors in (Abderrazek et al., 2023) shed light on these issues and proposed an alternative methodology for an interpretable speech assessment using DL. One of the authors' main contributions was addressing the issue of data requirements. Indeed, collecting a significant amount of data, especially for pathological speech, can be a difficult and expensive task, making this factor a crucial aspect to take into account. The authors' starting point was a DL-based model trained on healthy speech that encodes the characteristics of "normal" reference. Later on, they proposed an explainability framework and brought to light an interpretable dimension that emerges automatically within this DL-based model. This dimension serves later to interpret the final assessment of patients. As a result, interpretability can be achieved without the need for additional labels or data, and without introducing any possible subjective factors.

In this paper, we suggest building upon the work proposed by the authors in (Abderrazek et al., 2023) in order to have an end-to-end solution for an interpretable assessment of speech intelligibility for HNC patients. The rest of this paper is organized as follows. In section 2, we provide an overview of the methodology proposed by the authors, which serves as the foundation of this study. In section 3, we briefly describe the two main datasets dedicated to speech disorders due to HNC that we use in this work. Then, section 4 is devoted to the description of the experimental setup, including the data preprocessing, the architecture dedicated to the score assessment, and the training details. Subsequently, results and several analyses are reported in section 5. We follow up with a case study in section 6 to demonstrate the potential of this research as an end-to-end solution for the objective assessment and interpretation of speech disorders in a clinical setting. Finally, section 7 concludes this work and gives some perspectives.

2. Methodology Overview

In the **first step** (Abderrazek et al., 2020), authors proposed a DL-based model (Convolutional Neural Network - CNN) dedicated to the task of French phoneme classification. The model was trained exclusively on healthy speech in order to address the issue of limited data availability in speech pathology while also meeting the relatively high data requirement of deep learning applications. As shown in

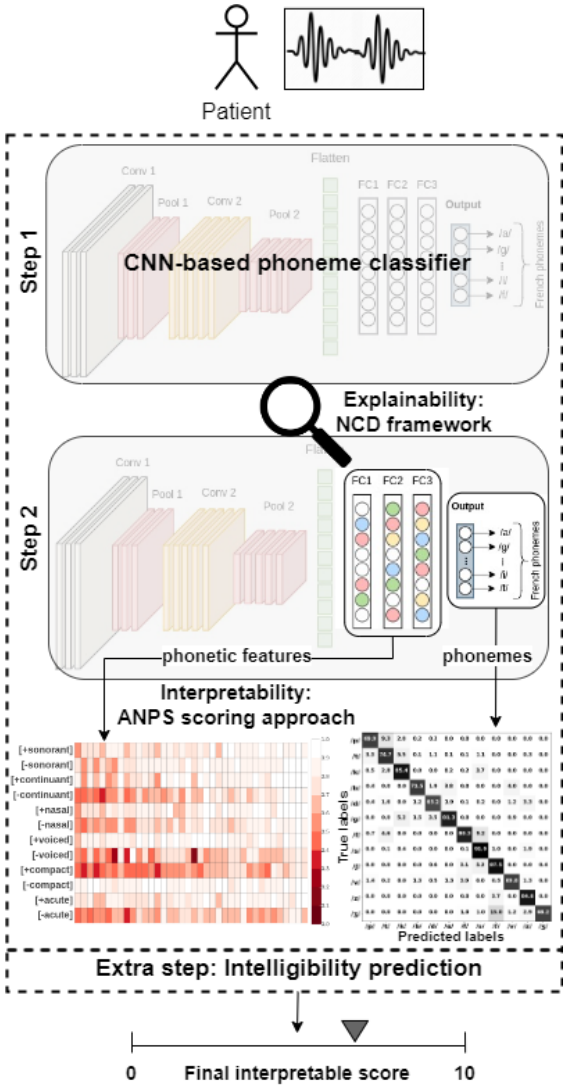


Figure 1: Overview of the methodology

figure 1, this methodological choice allowed the authors to have deep representations of French phonemes (hidden layers of the CNN) in addition to the phoneme dimension (output layer of the CNN). In the **second step** (Abderrazek et al., 2022), the authors' goal was to investigate the capacity of the CNN-based phoneme classifier to yield relevant knowledge related to the characteristics of speech pathology. For this sake, the authors proposed the framework, **Neuro-based Concept Detector (NCD)**, a general analytic framework for the explainability of hidden neurons/layers of a DL-based model performing a classification task. By applying NCD for the CNN explainability, the authors brought to light an interpretable dimension of great relevance in the clinical phonetics context which is phonetic features. Subsequently, they proposed a scoring approach, **Artificial Neuron-based Phonological Similarity (ANPS)**, to retrieve fine-grained interpretations of the speech impairment based on the emergent dimension of phonetic features. This scoring approach is associated with heatmaps to facilitate the visualization and understanding of interpretable information by clinical experts. In this paper, we suggest adding an extra step to this proposed methodology in order to make it an end-to-end solution dedicated to an interpretable assessment of speech intelligibility in the context of speech disorders due to HNC. The details of this extra step implementation are provided in the next sections.

3. Data

This work relies on two main corpora, involving both recordings of disordered speech due to HNC and their perceptual measures.

3.1. C2SI corpus

The C2SI corpus (Carcinologic Speech Severity Index) (Woisard et al., 2021), is a corpus including the recordings of 87 HNC patients and 40 healthy control speakers. It was proposed in the C2SI study with the goal to assess how the treatment for upper aerodigestive tract cancers (i.e. pharynx and oral cavity) affects speech production using both perceptual and automated speech processing techniques. To qualify for inclusion, patients need to have successfully completed their therapeutic protocol for a minimum of six months, achieved clinical remission, and exhibited no speech disorders of alternative causes.

Recorded tasks: The patient recordings were generated through various tasks, each designed for a specific type of analysis. This study primarily concentrates on the subset of C2SI recordings issued respectively from the reading task (C2SI-LEC), the image description task (C2SI-DES), the prosodic tasks (C2SI-SYN, C2SI-FOC, and C2SI-MOD), and the sentence verification task (C2SI-SVT). It is worth mentioning that not all C2SI speakers were recorded for all the tasks. The exact included number of speakers will be given later in table 1. The recordings were therefore analyzed by a jury composed of six clinicians whose expertise area is speech disorder evaluation.

Perceptual measures: Among the different perceptual measures that were conducted, we outline the most significant ones to the current study which are intelligibility and severity. The instructions given to the experts included the following definitions (Balaguer et al., 2019). Intelligibility is defined as *"the comprehensibility of the message sent by the signal"*, while severity is defined as *"the degree of the overall deterioration of the audible signal"*. Both measures are assessed on a scale from 0 to 10, where 0 corresponds to the strongest alteration/unintelligible speech, and 10 corresponds to the absence of alteration/perfectly intelligible speech. In the rest of this paper, we exclusively use the intelligibility and severity measures that were perceptually assessed on the picture description task. We assign a single overall severity score and a single overall intelligibility score to the set of recordings produced by each speaker regardless of the real task of the recording in question. This choice is explained by the fact that the task of image description leads to less predictable linguistic content compared to the reading task (Lalain et al., 2020), therefore, to a more valuable perceptual assessment by the experts.

Metadata: In addition to the recordings and perceptual measures, C2SI corpus includes some clinical information about the patients such as the treatment type (surgery, radiotherapy, chemotherapy), cancer region, values of T and N criteria from UICC Tumor/Node/Metastasis (TNM) classification, etc.

3.2. SpeeCOmco

Proposed by (Balaguer, 2021), SpeeCOmco is a corpus including 27 patients treated for cancer of the oral cavity or oropharynx. Similarly to C2SI, the patients in SpeeCOmco corpus recorded different tasks and were subject to several perceptual assessments. In this work, we only focus on the recordings issued from the reading task (the same clinical text used in C2SI). Regarding the perceptual measures, we use the intelligibility and severity measures of SpeeCOmco patients that were assessed on the recordings of the semi-directed interview. These assessments were conducted using not only the same instructions and rating scales as those used in C2SI corpus but also the same experts, which ensures greater comparability in the perceptual measures of the two corpora.

4. Experimental Setup

In this section, we outline the implementation of the extra step which serves our final goal: predicting an interpretable intelligibility score for HNC patients.

4.1. Input Data Preparation

In this section, we present the process of data preparation for the score prediction task, which we illustrate in figure 2. Basically, we take the speech productions from every speaker and apply data preprocessing to make it compatible with the input of the CNN-based phoneme classifier. The outcome of this data preprocessing stage is a set of acoustic feature matrices at the frame level, as described in (Abderrazek et al., 2020), that we refer to as CNN input samples. Next, we consider these CNN input samples by blocks of 100 consecutive samples which reflect almost one second of speech produced by a particular speaker. Each of these blocks is then fed to the trained CNN. The choice of one-second segments leading to blocks of 100 consecutive samples has been driven by the necessity of sufficient data for the intelligibility score prediction process, regarding the speech disorder corpora available in our context. Indeed, we cannot consider the set of overall speech recordings available per patient, but smaller speech segments to augment the processed data. As illustrated in figure 2, we select the set of 985 interpretable neurons across the different fully-connected layers of the CNN. These interpretable neurons have been identified as phonetic feature detectors in (Abderrazek et al., 2022). Now, as aforementioned, we fed the blocks of 100 input samples to the CNN. We retrieve the activations of the selected set of interpretable neurons and concatenate them into embedding vectors with a dimension of 985. Here, a single embedding vector matches a single input sample. That is to say, for one block of 100 input samples reflecting almost one second of speech, we obtain 100 embedding vectors. We refer to these resulting embedding vectors as **phonetic feature embeddings** as they represent the input speech signal in terms of phonetic features. A block of 100 phonetic feature embeddings is considered later as one input sample to the next model responsible for predicting the final score.

4.2. Score Prediction process

Building on the previous input preparation, we use the blocks of 100 phonetic feature embeddings, generated for each speaker, as input to a Shallow Neural Network (SNN) that aims to predict a final score. As detailed in figure 3, the SNN generates a score prediction for each block of 100 vectors. In other words, for each speaker, we will have an assessment of his/her speech production for almost every second. It is worth noting that it is possible to obtain an overall score for an utterance or a speaker. For instance, to get an utterance-level score, we average the scores generated for each second of the utterance. Similarly, a global score for a given speaker results from averaging all the scores generated for each second across all utterances produced by that speaker. As regards the SNN architecture, it simply consists of an average pooling layer followed by a fully-connected layer, then a final output layer. Considering the phonetic feature embeddings that we generated as detailed in figure 2, this pooling layer takes 100 vectors each composed of 985 activation values, and converts them to a 985-dimensional vector. This transformation can be considered as passing from a frame-level representation to a segment-level representation (one-second segment). This latter is then fed to one fully connected layer with a ReLU activation function. The number of neurons within this layer is a hyper-parameter that we tune and fix later based on the task in question. Finally, the output layer corresponds to the final score (i.e. the assessment of the one-second input segment). To ensure that the predicted score is between 0 and 10, we use the bounded activation function *sigmoid* which maps any input to a value between 0 and 1, and then we scale the output of this function to map it to the range [0, 10]. This requirement stems from the fact that the perceptual measures in our possession, which serve as the ground-truth scores for training the regression model, are evaluated within a range of 0 to 10 as detailed in section 3.

4.3. Datasets and Training details

The proposed SNN model is trained to map an input of 100 phonetic feature embeddings at the frame level to a particular score of interest. To this end, we use the datasets C2SI-SVT, C2SI-FOC, C2SI-MOD, and C2SI-SYN as input for the training process. A collection of one-second segments is issued from the different speakers' productions in these datasets (i.e. patients and healthy control speakers) and then, as described in section 4.1, prepared to be an input to the regression model. We further use the dataset C2SI-LEC as a validation set to monitor the training and tune the experimental settings. As regards the test, we use SpeeCOmco dataset to evaluate the resulting model. Still in table 1, we report some details about the input samples to the regression task for training, validation, and testing. Some details about the target score distributions (i.e. intelligibility and severity), within the training and validation sets, are summarized in the same table. The Mean Squared Error (MSE) is taken as a loss function for the score regression task.

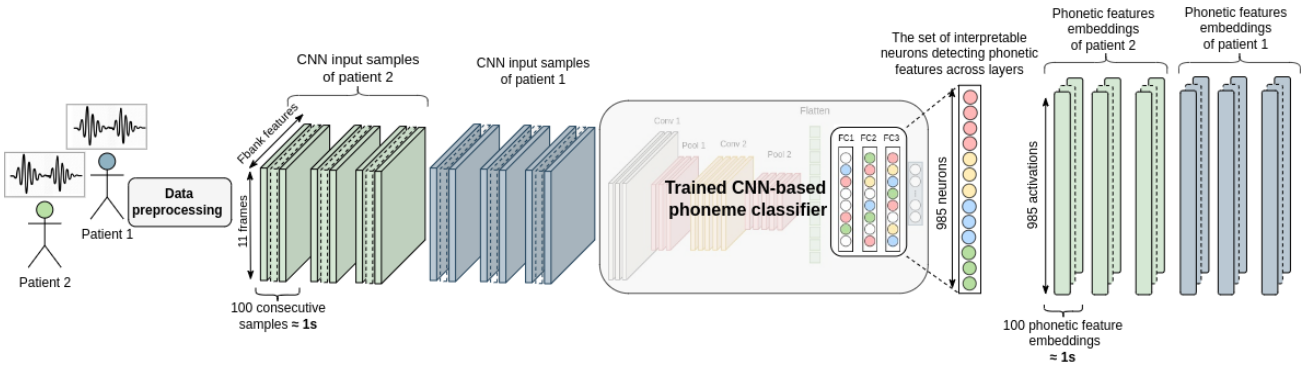


Figure 2: Preparation of the input for score prediction: Phonetic feature embeddings

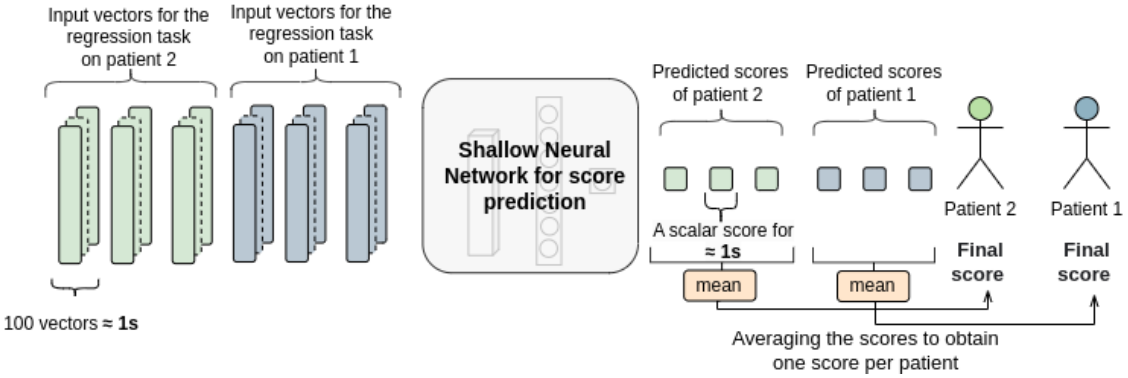


Figure 3: The process of score prediction

5. Results

In this section, we report and discuss the results of different regression experiments. The outcomes are depicted in Table 2, showcasing two distinct losses: the Mean Absolute Error (MAE) and the MSE. These losses provide insights into various aspects of the errors present in the model predictions. In the table columns, we specify the target scores for which the model was trained to make predictions, along with the number of neurons used in the hidden fully connected layer of the SNN. Therefore, for each of these configurations, we provide the different loss values obtained on both the validation set (C2SI-LEC corpus) and the test set (SpeeCOmco corpus). We can observe that the best model for severity prediction is the one with 64 hidden neurons. Regarding the intelligibility prediction, the best model is the one with 256 hidden neurons. **All the analyses and comparisons below are based on these two best models.**

For the severity prediction task, the best model achieves an MAE of 1.25 and an MSE of 2.55, as average errors on the C2SI-LEC dataset. As regards the best regression model predicting the intelligibility score, an MAE of 1.21 and an MSE of 2.97 are achieved on the same data. It is worth mentioning that these best models demonstrate remarkable performance on the SpeeCOmco corpus as well (test set). We can see from table 2 that the best regression model for severity prediction achieves an MAE equal to 1.4 and an MSE equal to 2.97 on the SpeeCOmco dataset. As regards intelligibility prediction, the best model achieves even better results with an MAE of 1.32 and an MSE of 2.97 on the same data. Despite having relatively few examples to learn from (25K one-second segments, see table 1), the models are able to accurately predict scores for another set of HNC patients, that were never seen in the training and validation phases. Importantly, this sheds light on the ability of the resulting models to generalize well to a completely different set of patients and confirms that they are not subject to overfitting on the C2SI patients.

To complete our analysis, we plot the scatter plots of the mean predicted severity (resp. intelligibility) vs. the true perceptual severity (resp. intelligibility) of C2SI-LEC and SpeeCOmco speakers. We organize the analysis based on the target task.

5.1. Analysis of severity prediction

The scatter plot of the mean predicted severity vs. the true perceptual severity of C2SI-LEC speakers is depicted in figure 4a. We highlight the best-fit line between the mean predicted scores and the

	Training	Validation	Testing
Dataset	C2SI-SYN & C2SI-MOD & C2SI-SVT & C2SI-FOC	C2SI-LEC	SpeeCOmco
#speakers	105	114	27
#input samples (1s segments)	25637	3542	867
Perceptual Intelligibility (mean±std)	7.9±2.5	7.9±2.5	6.7±2.6
Perceptual Severity (mean±std)	6.5±2.6	6.5±2.6	5.7±2.6

Given that a single perceptual intelligibility or severity rating was assigned to an entire recording, we replicate this assignment to every one-second segment comprising the entirety of that recording.
The mean and standard deviation values are calculated on one-second segments.

Table 1: Datasets for the training, validation, and testing of the Shallow Neural Network

		Task					
		Severity Prediction			**Intelligibility Prediction**		
		#Neurons			**#Neurons**		
		64	**128**	**256**	**64**	**128**	**256**
C2SI-LEC	**MAE**	**1.25**	1.28	1.26	2.13	1.3	**1.21**
	MSE	**2.55**	2.74	2.62	10.73	3.36	**2.97**
SpeeCOmco	**MAE**	**1.4**	1.44	1.4	3.29	1.45	**1.32**
	MSE	**2.97**	3.22	3.05	17.58	3.57	**2.97**

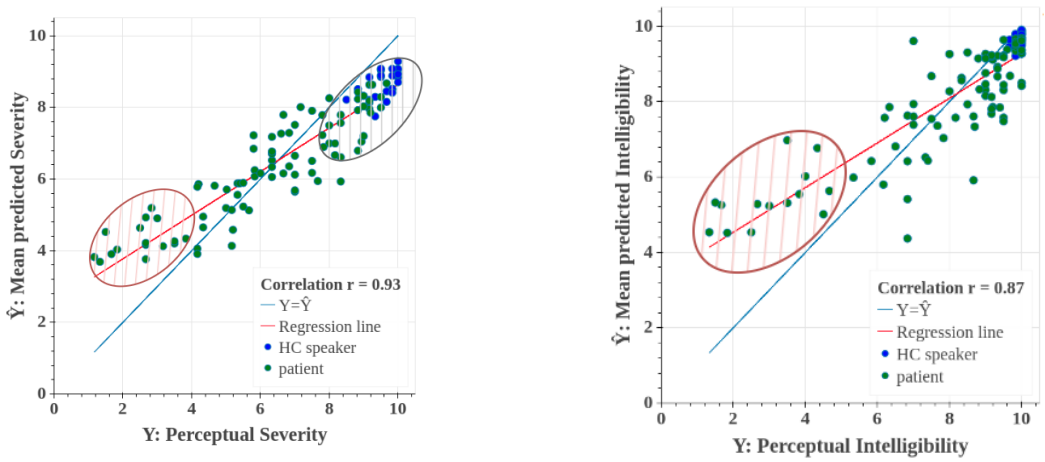
Table 2: Results of regression on phonetic feature embeddings according to the final task and the number of neurons in the SNN hidden layer

perceptual score, in addition to the line $Y = \hat{Y}$ to visualize any possible pattern in the errors. Healthy control speakers and patients are distinguished with blue and green colors, respectively.

First, it is worth mentioning that the range of the mean predicted severity is [3.7; 9.3] which means that this score is reduced and does not cover the complete range of severity [0; 10]. As regards the regression line, we can see that a positive strong relationship exists between Y and $\hat{Y}$. This is confirmed by a high Pearson correlation, equals to 0.93, between the predicted and perceptual severity values. This may indicate that the model is able to capture some of the underlying patterns in the phonetic feature embeddings. However, it is important to note that a high correlation does not necessarily imply high accuracy or precision in the predictions. Even if the model is able to capture some of the overall trends in the data, it may still be making significant errors in individual predictions, which could lead to incorrect conclusions. To this end, we analyze the results based on the line $Y = \hat{Y}$. From this perspective, we can see that the regression model actually underestimates high severity scores (i.e. the upper right area hashed in grey) and overestimates low severity scores (i.e. the bottom left area hashed in red). Consequently, this may suggest that the model has a systematic bias in its predictions. Specifically, the model may be *"flattening"* the predicted scores towards the mean, rather than capturing the full range of variation in the target variable. In other words, this indicates that there is room for improvement in the regression model we proposed in order to tackle this specific behavior. Now moving to the model prediction analysis on the test set, we would like to discard any possibility of misleading conclusions from the previous analysis due to the fact that it was conducted on the validation set. Figure 5 depicts the scatter plot of the mean predicted severity vs. the true perceptual severity on SpeeCOmco patients. First, the scatter plot shows exactly the same trends as the one in figure 4 conducted on C2SI-LEC speakers, with a variation of the mean predicted severity in the range [3.4; 8.7]. The model bias previously observed towards underestimating high severity scores and overestimating low severity scores is still noticeable.

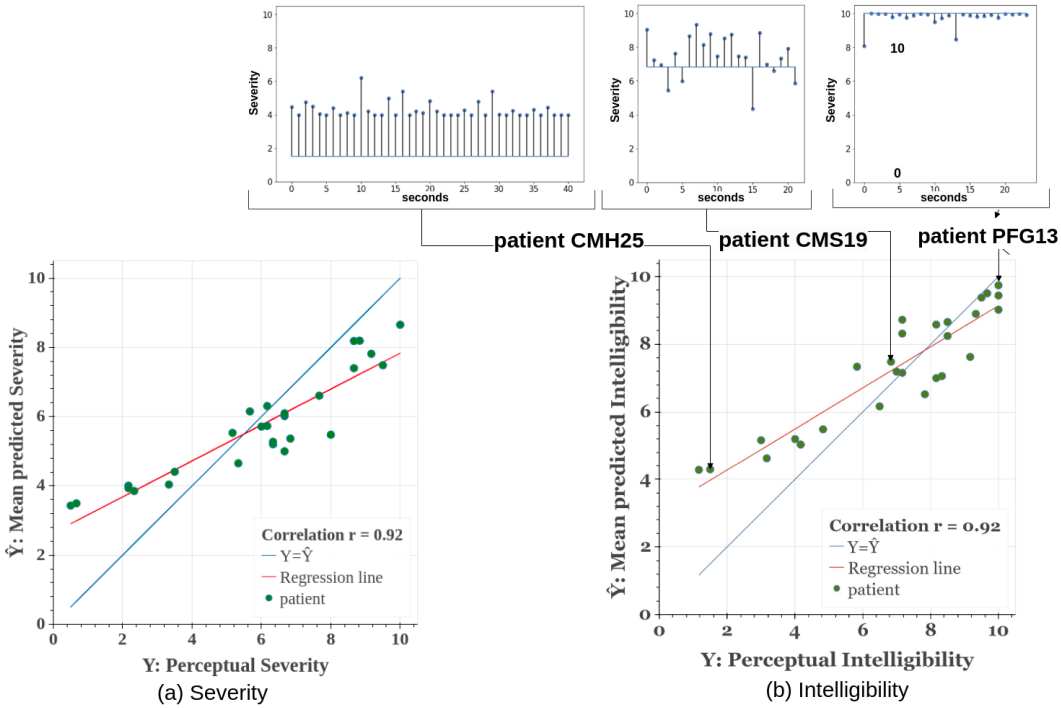
5.2. Analysis of intelligibility prediction

Similarly, figure 4b depicts the scatter plot of the mean predicted intelligibility vs. the true perceptual intelligibility of C2SI-LEC speakers. It is worth noting that the range of the mean predicted intelligibility is [4.4; 9.9]. This range is indeed slightly higher than the range of mean predicted severity. Moreover, the Pearson correlation between the mean predicted intelligibility and the true perceptual intelligibility is less than the correlation calculated on severity, but still very important (r=0.87). Additionally, we observe that the model exhibits a clear bias towards overestimating low scores in the prediction of speech intelligibility, consistent with the previously noted bias in the prediction of speech severity. However, unlike the bias observed in the



(a) Severity (b) Intelligibility

Figure 4: Scatter plot of the mean predicted measure vs. the true perceptual measure on C2SI-LEC speakers: (a) Severity, (b) Intelligibility.



(a) Severity (b) Intelligibility

Figure 5: Scatter plot of the mean predicted measure vs. the true perceptual measure on SpeeCOmco patients: (a) Severity, (b) Intelligibility

prediction of speech severity, the bias towards underestimating high scores is not readily apparent in the case of speech intelligibility prediction.

On the other hand, figure 5 depicts the scatter plot of the mean predicted intelligibility vs. the true perceptual intelligibility of SpeeCOmco patients. First, the variation of the mean predicted intelligibility of SpeeCOmco patients is in the range [4.3; 9.7]. Obviously, the trends and observations described for the intelligibility predictions on C2SI-LEC speakers remain valid. In addition, we add examples of regression plots per second on three patients in order to have visibility on the model behavior at the one-second segment level. The selection of patients was performed while varying their perceptual intelligibility levels (i.e. *"PFG13"* for high perceptual intelligibility, *"CMS19"* for medium perceptual intelligibility, and *"CMH25"* for deteriorated perceptual intelligibility). The X-axis of these plots represents the seconds of the speech production,

which number depends on the time each patient takes to read the same text. The Y-axis represents the severity range. The horizontal blue line is the perceptual severity level of the patient in question, while the blue dots are the predicted severity scores at the one-second segment level. The vertical black lines are the residuals (i.e. the difference between a true and a predicted value) at each second. It is worth mentioning that the detailed predictions at the one-second segment level reveal that the model exhibits a high degree of confidence and consistently makes the same decision for all seconds of the patient *"PFG13"*. As regards the patient *"CMS19"*, we can see that the model predictions vary largely depending on the one-second segment in question. This behavior tends to be coherent. Indeed, with nearly "normal" speech (the case of patient *"PFG13"*), it would be expected to have a few altered one-second segments as observed. Conversely, with very severe impairment (the case of patient *"CMH25"*), it would be expected that almost all one-second segments would be altered as observed. Finally, more variation between one-second segments should be expected with moderate impairment (the case of patient *"CMH19"*), with some "normal" speech segments, and others more altered. Still, additional analyses are necessary to investigate the characteristics of the one-second segments.

6. An end-to-end application of the interpretable methodology

In this section, we consider the three patients *"PFG13"*, *"CMS19"*, and *"CMH25"* belonging to SpeeCOmco corpus to illustrate an end-to-end application of the global methodology for an interpretable assessment of speech intelligibility. These patients were rated by the experts 10, 6.8, and 1.5 in terms of intelligibility respectively. Regarding the automatic prediction of intelligibility scores, as shown in Figure 6, we obtain prediction scores of 9.7, 7.4, and 4.2 for the same three patients, respectively. Thanks to Step 2 of the interpretability methodology (Abderrazek et al., 2022), we can associate these different predicted intelligibility scores with a deeper analysis based on the altered phonetic features as depicted in the figure 6 heatmaps. Indeed, this figure reports the local ANPS scores per phonetic features for both consonants and vowels for all the patients of the SpeeCOmco dataset (heatmaps), sorted according to their perceptual intelligibility scores (from the most intelligible on the right to the least intelligible on the left). Proposed in (Abderrazek et al., 2023), this score assesses how well acoustic/articulatory characteristics related to phonetic feature t are produced by speaker s. The three patients *"PFG13"*, *"CMS19"*, and *"CMH25"* are specifically highlighted in the figure with their local ANPS scores surrounded. Comparing these three patients, we can clearly see different configurations of local ANPS scores, showing a consistent deterioration of score values compared to the predicted intelligibility scores of these patients. This association between the predicted intelligibility score and the heatmaps displaying ANPS scores should enable clinicians to directly associate a score with alterations in phonetic features at a specific time (t). Additionally, it should allow the comparison of different pairs of scores/heatmaps for the same patient in a longitudinal way to measure the efficiency of a rehabilitation program or of a specific prosthesis.

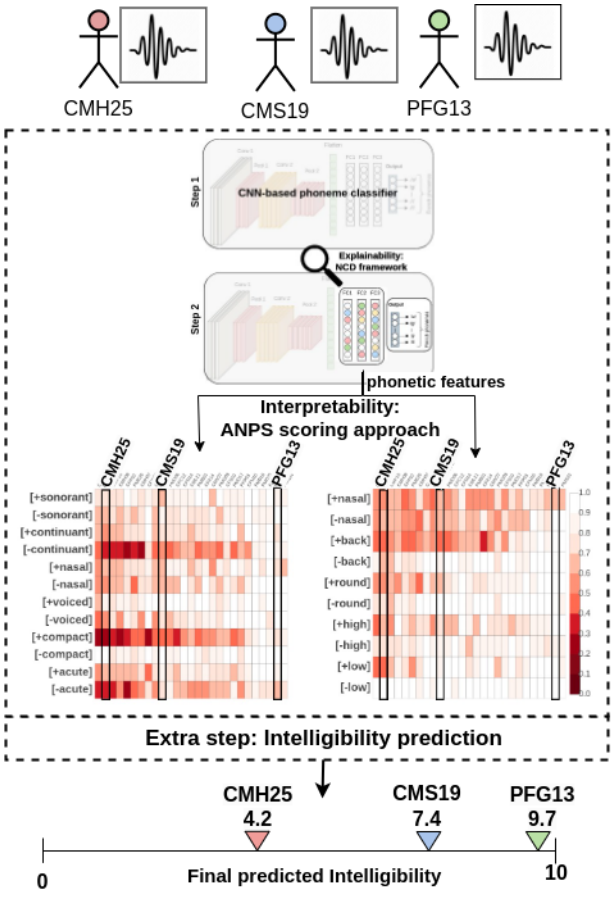


Figure 6: End-to-end application of the global methodology on three SpeeCOmco patients

7. Discussion

In this paper, we built upon an interpretability methodology that was proposed by (Abderrazek et al., 2023) in order to have an end-to-end solution for an interpretable assessment of speech intelligibility in the context of speech disorders. Promising results have been obtained regarding the prediction of speech intelligibility and severity of HNC patients while giving relevant interpretations of the

final assessment both at the phonemes and phonetic feature levels. By examining this speech in terms of production at these two granularity levels, clinicians can gather more useful information for speech therapy and develop tailored rehabilitation protocols. This is especially important because we are aware that phonemic alterations are the primary challenge in speech production for HNC patients, significantly affecting their communication skills and, consequently, their quality of life.
One interesting perspective of this work would be to confirm these results in clinical practice and more specifically within a longitudinal study. This type of study involves following a group of individuals with speech disorders for a period of time which allows clinicians to observe and track changes in speech production over time. In other words, if a longitudinal study can confirm the effectiveness of the proposed methodology and demonstrate consistent and meaningful interpretations, it can provide clinicians and researchers with a deeper understanding of how speech disorders impact individuals and inform better treatment approaches.

While these findings show great promise, we believe that further improvements are necessary to enhance the reliability and generalizability of the models. As a perspective, we suggest using an attention mechanism that can potentially improve the performance of the regression model by enabling the focus on the most relevant frames and features for the prediction. This technique is very used in the speaker recognition field (Okabe et al., 2018), where it has been shown that some frames are more unique and important for discriminating speakers than others, for a given utterance. In speech intelligibility assessment of patients with speech disorders, some frames of speech may contain more critical information for understanding the intended message than others. A possible further analysis could be to examine the input features used by the regression model and their importance in predicting the target score. An application of the SHAP (Lundberg and Lee, 2017) framework can be used to explain the predicted score of the model for a specific input by attributing a contribution value to each element in the phonetic feature embedding. This provides insight into which features are driving the predictions of the model and how they are influencing the final intelligibility score.

Acknowledgments

This work was supported by the French National Research Agency under RUGBI project untitled "Looking for Relevant linguistic Units to improve the intelliGiBIlity measurement of speech production disorders" (Grant n°ANR-18-CE45-0008-04).

8. Bibliographical References

Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, and Virginie Woisard. 2020. Towards Interpreting Deep Learning Models to Understand Loss of Speech Intelligibility in Speech Disorders — Step 1: CNN Model-Based Phone Classification. In *Proc. Interspeech 2020*, pages 2522–2526, China.

Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, and Virginie Woisard. 2022. 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 - International Conference on Acoustics, Speech and Signal Processing*, pages 7387–7391, Singapore.

Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, and Virginie Woisard. 2023. Interpreting Deep Representations of Phonetic Features via Neuro-Based Concept Detector: Application to Speech Disorders Due to Head and Neck Cancer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:200–214.

Mathieu Balaguer. 2021. *Mesure de l'altération de la communication par analyses automatiques de la parole spontanée après traitement d'un cancer oral ou oropharyngé*. Thesis, Université Paul Sabatier - Toulouse III.

Mathieu Balaguer, Aline Boisguerin, Anaïs Galtier, Nadège Gaillard, Michèle Puech, and Virginie Woisard. 2019. Factors influencing intelligibility and severity of chronic speech disorders of patients treated for oral or oropharyngeal cancer. *Eur. Arch. Otorhinolaryngol.*, 276(6):1767–1774.

Li Bin, Matthew C Kelley, Daniel Aalto, and Benjamin V Tucker. 2019. Automatic speech intelligibility scoring of head and neck cancer patients with deep neural networks. In *ICPhS'19, Melbourne, Australia*.

Siddhant Gupta, Ankur T. Patil, Mirali Purohit, Mihir Parmar, Maitreya Patel, Hemant A. Patil, and Rodrigo Capobianco Guido. 2021. Residual neural network precisely quantifies dysarthria severity-level based on short-duration speech segments. *Neural Networks*, 139:105–117.

Raymond D Kent. 1992. *Intelligibility in speech disorders: Theory, measurement and management*, volume 1. John Benjamins Publishing.

Daniel Korzekwa, Roberto Barra-Chicote, Bozena Kostek, Thomas Drugman, and Mateusz Lajszczak. 2019. Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech. In *Proc. Interspeech 2019*, pages 3890–3894.

Muriel Lalain, Alain Ghio, Laurence Giusti, Danièle Robert, Corinne Fredouille, and Virginie Woisard. 2020. Design and development of a speech intelligibility test based on pseudowords in french: Why and how? *Journal of Speech, Language, and Hearing Research*.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA.

T. K. Meyer, J. C. Kuhn, B. H. Campbell, A. M. Marbella, K. B. Myers, and P. M. Layde. 2004. Speech intelligibility and quality of life in head and neck cancer survivors. *Laryngoscope*, 114(11):1977–1981.

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

Timothy Pommée, Mathieu Balaguer, Julie Mauclair, Julien Pinquier, and Virginie Woisard. 2021. Assessment of adult speech disorders: current situation and needs in French-speaking clinical practice. *Logopedics Phoniatrics Vocology*, pages 1–15.

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.

J. Revis. 2004. *L'analyse perceptive des dysphonies : approche phonétique de l'évaluation vocale*. Ph.D. thesis, Université de la Méditerranée.

Ming Tu, Visar Berisha, and Julie Liss. 2017. Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In *Proc. Interspeech 2017*, pages 1849–1853.

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(1).

Lingfeng Xu, Julie Liss, and Visar Berisha. 2023. Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA Express Letters*, 3(1):015201.