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[345]

[1]*LIA, Avignon University, France,*
[2]*Aix-Marseille Univ, LPL, CNRS, France*
[3]*IRIT, Toulouse University, CNRS, France*
[4]*IUC Toulouse, CHU Toulouse, Larrey Hospital, France*
[5]*LNPL, UR 4156, Toulouse University, France*

PRESENTOR:

(Manh) Tuan NGUYEN
PhD Candidate

LREC COLING 2024

Table of Contents

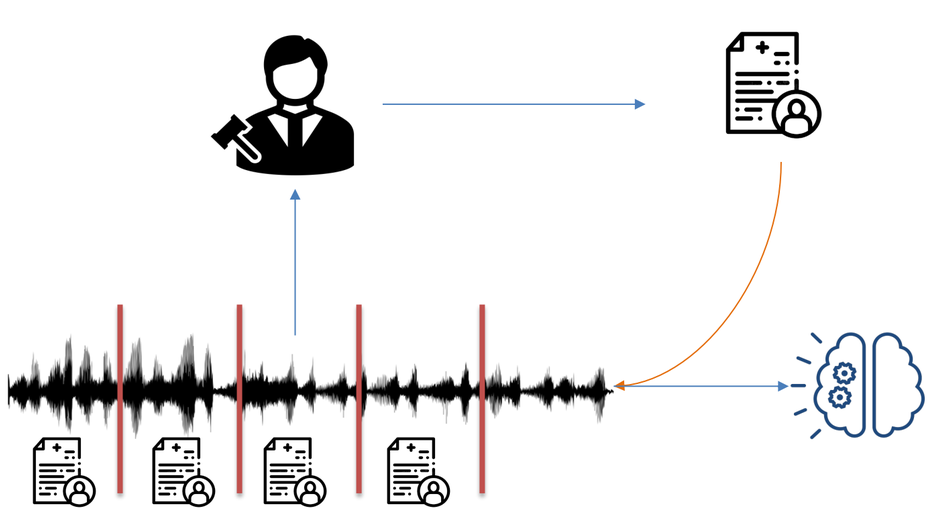
Context

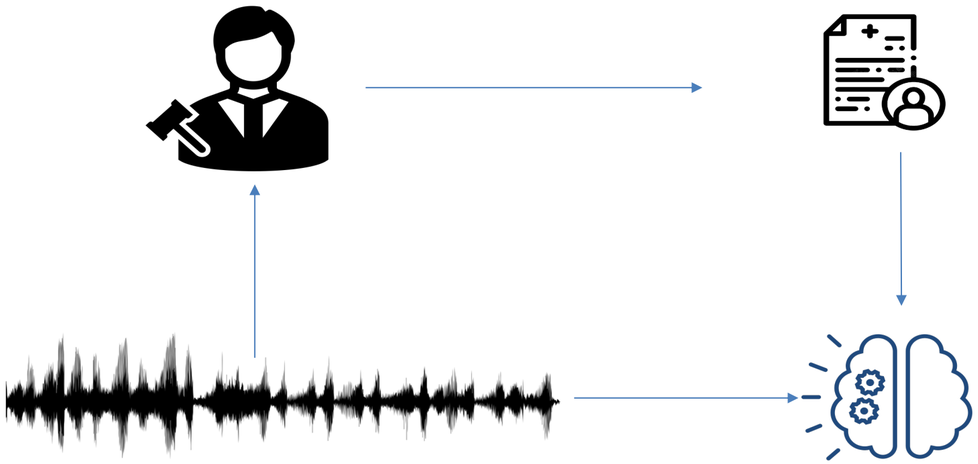
- Speech and voice disorders
- Evaluating patients' speech is crucial
- Traditional approach relies on subjective judgments
- Automatic speech quality assessment is gaining attention as an alternative or supplement to traditional clinical evaluation.

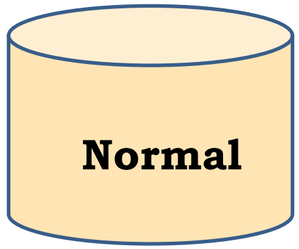
Challenge

- Limited data availability restricts research success to basic tasks like binary classification.
- Segmenting audio files in current methods to augment datasets indirectly links overall scores with individual segments, presenting limitations.

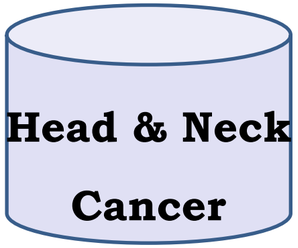
This work propose a regression system learns at the audio level despite data scarcity, leveraging the pre-trained Wav2Vec2 architecture for both SSL and ASR as feature extractors in speech assessment.







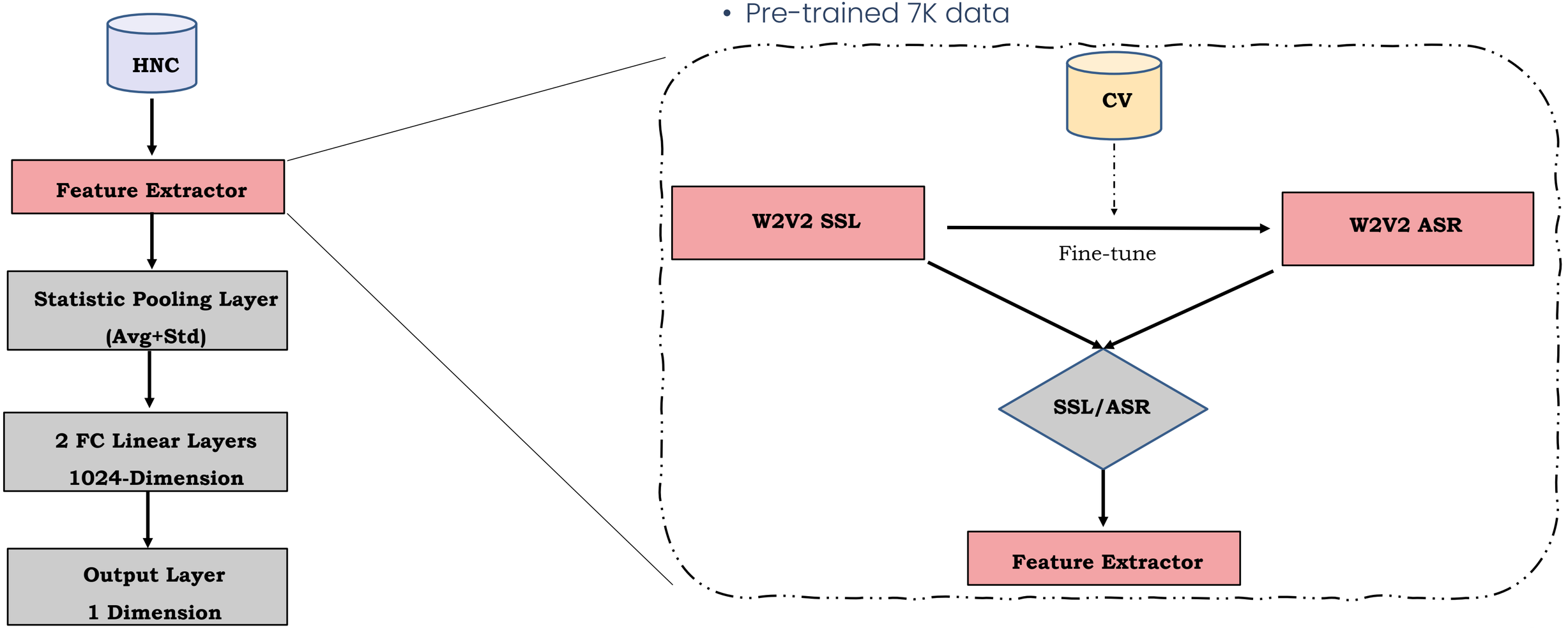
- CommonVoice corpus
  - Specifically used for fine-tuning Wav2Vec2 for the ASR task
  - Version 6.1

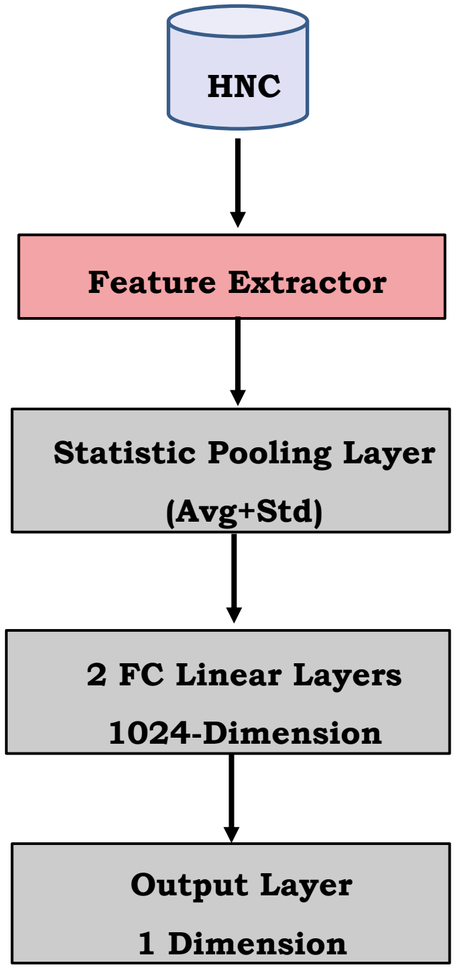


- C2SI corpus
  - 105 speakers (84 patients & 21 controls)
  - Reading of *La Chèvre de monsieur Seguin*
  - Measures intelligibility/severity (0-10)
- SpeCOmco corpus
  - Additional corpus for C2SI
  - 27 patients

- AHN corpus
  - 15 patients
  - Reading of La Chèvre de monsieur Seguin and Le Cordonnier
  - Measures intelligibility/severity (0-4)

- LeBenchMark large model
- Pre-trained 3K data
- Pre-trained 7K data





- Epochs: 20
- Batch size: 1
- Loss function and Evaluation metric: MSE
- 10-fold cross-validation using C2SI corpus
- Testing on SpeCOmco

System's performance

Model	Intelligibility MSE	Severity MSE
3K-SSL	1.65 ± *0.43*	2.10 ± *0.83*
7K-SSL	1.84 ± *0.49*	1.83 ± *0.71*
3K-ASR	0.73 ± *0.18*	1.15 ± *0.14*
7K-ASR	0.98 ± *0.26*	1.15 ± *0.16*
Baseline 1 ECAPA-TDNN	*1.75*	*1.91*
Baseline 2 CNN based	*2.97*	*3.05*

- Compared with existing baselines using the same SpeeCOmCo dataset:
  - Shallow Neural Network-based system: MSE reduction of 58% for intelligibility assessment and 41% for severity assessment.
  - CNN-based system: MSE reduction of 75% for intelligibility assessment and 62% for severity assessment.

Our proposed system consistently outperforms these baselines, demonstrating its effectiveness in speech quality assessment.

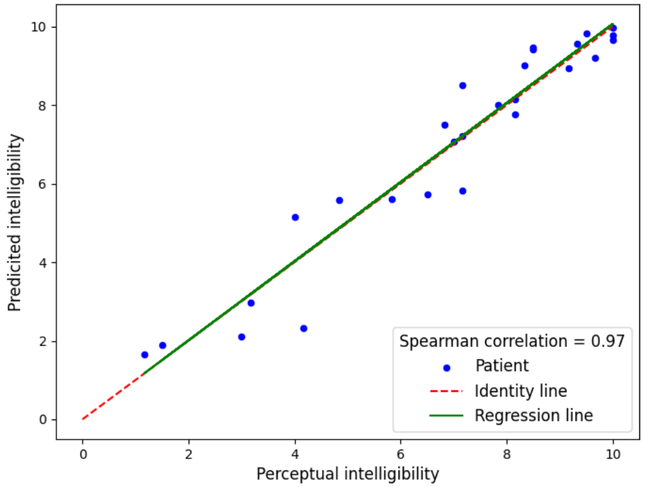
System's performance

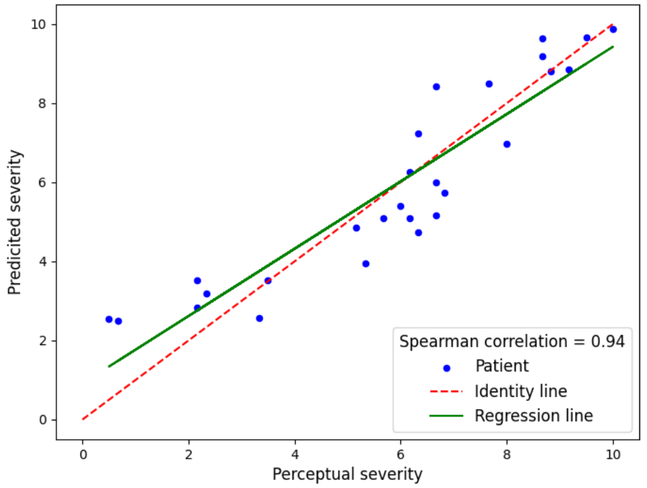
Model	Intelligibility MSE	Severity MSE
3K-SSL	1.65 ± *0.43*	2.10 ± *0.83*
7K-SSL	1.84 ± *0.49*	1.83 ± *0.71*
3K-ASR	**0.73 ± *0.18***	**1.15 ± *0.14***
7K-ASR	0.98 ± *0.26*	1.15 ± *0.16*
Baseline 1 ECAPA-TDNN	*1.75*	*1.91*
Baseline 2 CNN based	*2.97*	*3.05*

- Among different feature extractors:
  - 3K-ASR obtained the best result
  - Pre-trained based ASR outperforms pre-trained based SSL
  - 3K model in general performs better than 7K

System's performance

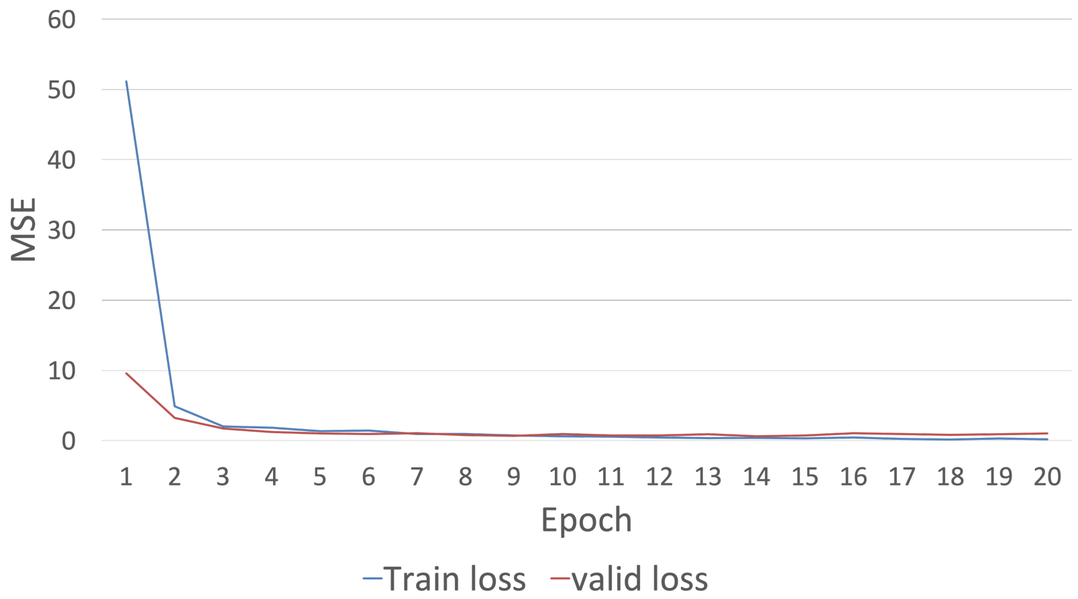
Model	Intelligibility MSE	Severity MSE
3K-SSL	1.65 ± *0.43*	2.10 ± *0.83*
7K-SSL	1.84 ± *0.49*	1.83 ± *0.71*
3K-ASR	**0.73 ± *0.18***	**1.15 ± *0.14***
7K-ASR	0.98 ± *0.26*	1.15 ± *0.16*
Baseline 1 ECAPA-TDNN	*1.75*	*1.91*
Baseline 2 CNN based	*2.97*	*3.05*





- High correlation between predictions and targets
- Correlation ranging from 0.94 to 0.97 (p-value < 0.01)
- Correlation with severity assessment is slightly lower than intelligibility assessment.
- The model overestimates severity for severe patients and underestimates for mild patients.

Generlization ability



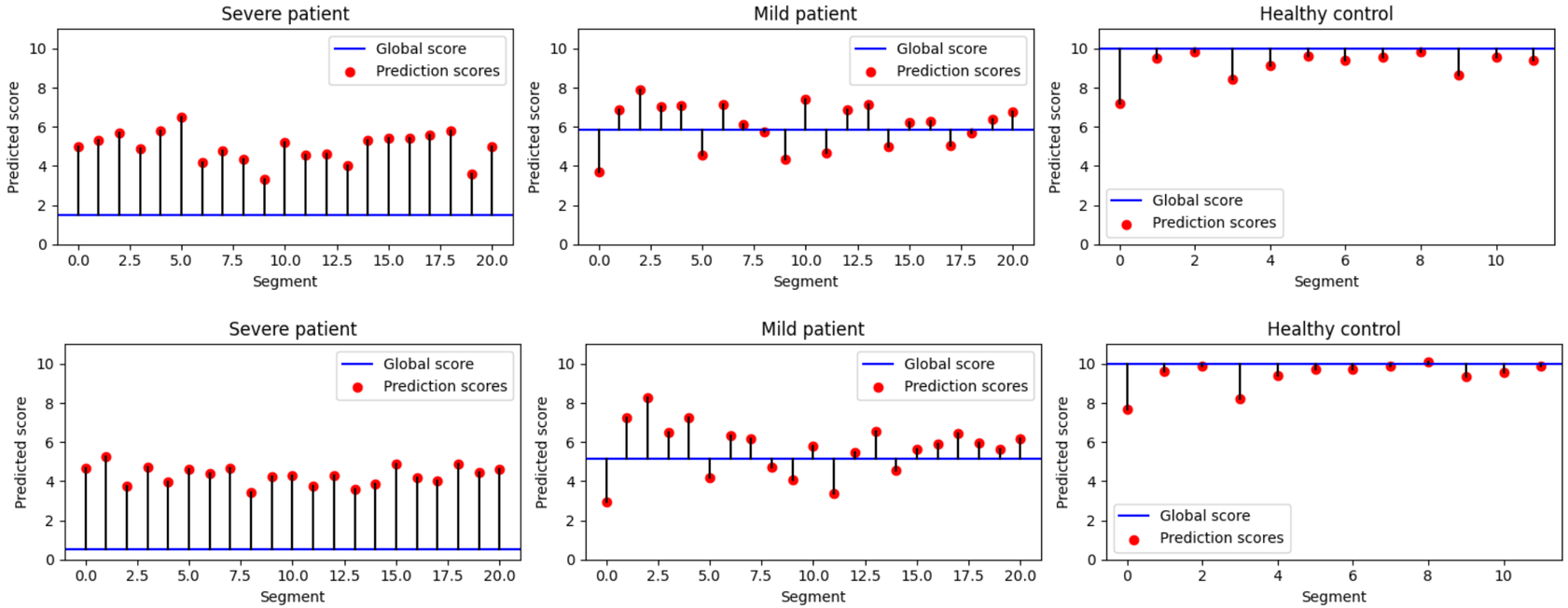
- Model continue to perform well and stable through out 10-fold validation
- Valid and train loss curve both decrease
- Cross-domain testing on AHN corpus: MSE=0.22 (intelligibility), MSE=0.37 (severity)

Therefore, it appears that overfitting is not a significant issue. The model also show the generalization ability through cross-domain data

Limited content

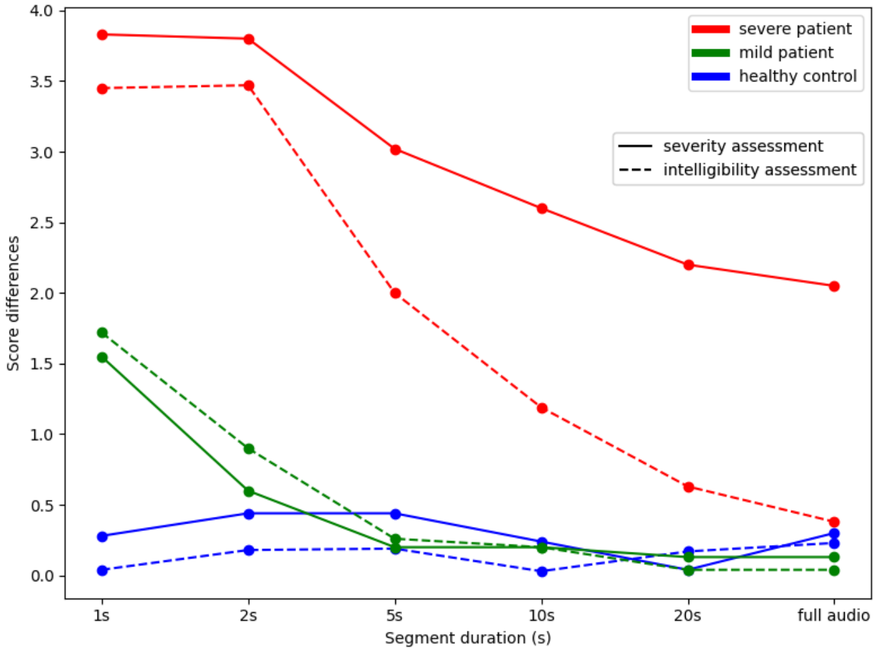
- **Severe group:** a patient with perceptual score of 1.5 for intelligibility and 0.5 for severity.
- **Mild group:** a patient with scores of 5.8 for intelligibility and 5.1 for severity.
- **Control group:** a healthy speaker with scores of 10 for both intelligibility and severity.

Limited content



- Results demonstrate consistent score generation for severe patients, with mild group scores varying more.
- Severe patient predictions tend to overestimate due to limited content information, indicating model struggle for accuracy.
- On the other hand, control group predictions show slight underestimation.

Limited content



- Longer durations enhance model performance by providing more content information.
- Duration does not affect healthy control group performance.
- Severe patients benefit from increased content, improving predictions, with mild patients showing similar behavior at a lower level.

Different content

- High alignment observed in predictions between texts *"La Chèvre de monsieur Seguin"* and *"Le Cordonnier"* within the AHN corpus.
- Despite differing phonetic contexts, the automatic system generates consistent predictions.
- Spearman's correlation analysis yields rates **of 0.96 for speech intelligibility** and **0.95 for severity assessment.**
- Indicating robust performance across different contents.
- Different contexts do not impact final decision, highlighting model's consistent performance.

- Proposed novel approach trains model on entire audio despite data scarcity
- achieving **58% MSE reduction (intelligibility)** and **41% MSE reduction (severity)** from current baseline
- ASR pre-trained context closely related to speech quality assessment.
- Duration of test segments impacts model decisions, particularly for severe patients.
- Changes in linguistic content between training and testing do not significantly affect model.

The End

Thank you for listening