Kanagawa University of Human Services
Graduate School of Health Innovation

【Doctoral Dissertation】

"Severity Classification of COVID-19 Patients Using Voice Biomarker with DTW Indices and Mahalanobis Distance Thresholding: A Feasibility Study "

Doctoral Program
Scheduled Completion in 2024-March

Student ID Number: 62145004
Name: Teruhisa Watase

Research Supervisor: Prof. Shinichi Tokuno
Assistant Supervisor: Prof. Hiroto Narinatsu
Associate Prof. Ryo Watanabe

To Mika

Table of Contents

Table of Tables

Table of Figures

Table of Appendices

List of Original Articles

This thesis is based on the following article, which will be referred to in the text by their Roman numerals:

I. Teruhisa Watase, Yasuhiro Omiya, Shinichi Tokuno
Severity classification using Dynamic Time Warping-Based voice biomarkers for patients with COVID-19: Feasibility Cross-sectional Study
JMIR Biomedical Engineering 2023, 8: e50924, **doi:10.2196/50924**

I Introduction

I-1 Background

The coronavirus disease 2019 (COVID-19) originated in Wuhan, China, in December 2019 and escalated into a global pandemic. By December 2022, approximately 650 million individuals had been infected by this disease, resulting in the tragic loss of more than 6.64 million lives. Although new infections appeared to have abated in the spring of 2023, the past explosion of infections strained medical care systems in several countries. In coping with this pressure, these countries have changed their responses toward infected patients based on the severity of their illness. In Japan, as illustrated in Table 1, displaying the Ministry of Health, Labour and Welfare guidelines on the severity of COVID-19 [1], responses were categorized into four levels of severity, spanning from mild to serious illness. Mild illness is characterized by “an oxygen saturation of 96% or higher, or the absence of respiratory symptoms or coughing without experiencing shortness of breath (SoB) with no sign of pneumonia in either scenario.” Conversely, moderate illness I is categorized by “an oxygen saturation greater than 93% but less than 96%, or the presence of shortness of breath or pneumonia in the clinical

condition." Moderate illness II is described as "having an oxygen saturation of 93% or lower, or necessitating oxygen administration." The target population for this study was individuals recovering at home or in recuperation facilities. Therefore, they were theoretically patients with mild illness. Nonetheless, due to worsening conditions or shortcomings of medical services, this population included patients with moderate illness I who should have been treated in a hospital. Therefore, accurately classifying these two adjacent severity categories (mild illness and moderate illness I) is essential in determining appropriate measures, such as early hospitalization, by detecting worsening conditions in patients with mild illness.

Table 1. Definition of the Severity of Covid-19 Infections

Severity	**Oxygen Saturation**	**Clinical Conditions**
Mild illness	SpO2 ≧ 96%	No respiratory symptoms or coughing without shortness of breath, but no evidence of pneumonia in either case
Moderate illness I	93% < SpO2 < 96%	Shortness of breath and pneumonia are visible
Moderate illness II	SpO2 ≦ 93%	Requires oxygen administration
Serious illness	---	Admitted to ICU or requires a ventilator

Oxygen saturation (SpO2) measurements using a pulse oximeter were crucial for assessing the severity

of illness. Daily measurements of SpO2 and body temperature, along with the assessment of physical conditions, were essential for monitoring disease progression from mild illness to moderate illness I over approximately 1 week or more during the recuperation period. However, the explosive increase of patients with COVID-19 made it challenging to distribute pulse oximeters to all patients with mild illness. In particular, it was more problematic for young patients who had to recuperate at their homes rather than in healthcare facilities. This unexpected shortage of pulse oximeters has motivated us to devise alternative and cost-effective ways to monitor for worsening medical conditions in persons exhibiting mild illness. Notably, all pulse oximeters supplied by Kanagawa prefecture were an identical model.

I-2. Voice Biomarkers

Previous research on Parkinson's disease, Alzheimer's disease, depression, and other psychiatric disorders such as stress [2-12] has shown that voice biomarkers can be leveraged. Specifically, they can be leveraged to noninvasively and cost-effectively identify the presence or absence of diseases, classify symptoms, and monitor conditions. Voice biomarkers could also be an alternative method to detect changes in disease severity from mild illness to moderate illness I in COVID-19. COVID-19 is a respiratory disease that has been reported to cause acoustic oscillations in the voice due to inflammation of the pharynx in the vocal tract, vocal cords, or both, and a lower expiratory volume

due to pneumonia [13]. Moreover, significant differences in jitter (fluctuation of the fundamental frequency on the time axis), shimmer (fluctuation of the amplitude on the power axis), and harmonic-to-noise ratio (HNR) were reported between healthy participants and those with COVID-19 [13-16]. Some reports state that COVID-19 can be detected from acoustic data obtained from a patient's cough [17].

I-3. Dynamic Time Warping Distance

Dynamic time warping (DTW) is a practical algorithm for measuring the similarity between two patterns. The DTW distance, a computational result obtained by the DTW algorithm using two waveform features, progressively approaches zero as the features become more similar. In contrast, it increases as the features become less similar (Figure 1).

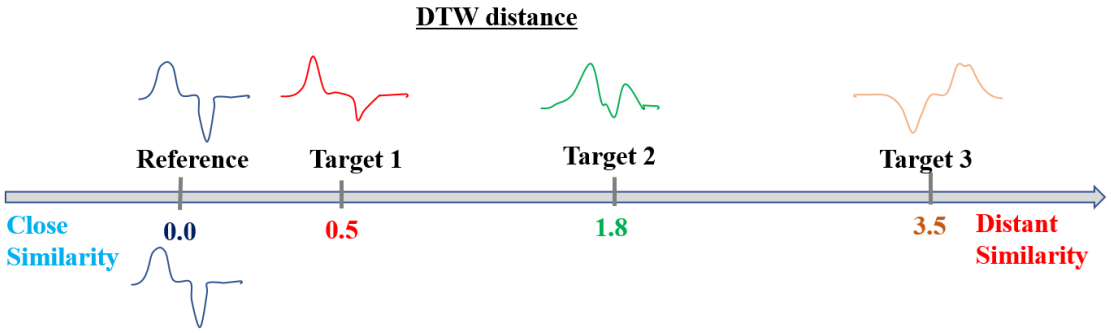


Figure 1. Illustration of the DTW distance, with three target example waves against a reference wave.

This metric has been widely used since the 1980s in motion recognition, speech recognition, and time-series data analysis [18-26]. For example, DTW could differentiate between healthy people and people with walking disabilities with high accuracy by processing differences in the gait patterns acquired by accelerometer sensors on smartphones [18]. The effectiveness of an automated scoring system applied in conjunction with the DTW algorithm for evaluating the progress of speech audiometric rehabilitation was similar to that of conventional manual scoring methods [23]. It has also been reported that complementing the mel-frequency cepstral coefficient (MFCC) algorithm with the DTW algorithm improved voice recognition performance. The DTW algorithm has been introduced as a feature-matching technique for voice recognition [25]. The feature-matching implementation of the DTW algorithm (i.e., the scoring method for 2D feature similarity) may function effectively for the desired classification of long vowel samples.

I-4. Mahalanobis Distance

In the process of some promising DTW distance-based indices, we added another algorithm of Mahalanobis Distance (MD) with the concept of *Normal Space* (NS) and *thresholding* to detect any latent implausible observation data that may exist in arbitrary observation data to be classified in this study. We expect this algorithm to automatically set an "out of specification" flag against either NS of two groups when some arbitrary observation data is calculated to classify. Implausible observation

data, such as outliers, measurement error data, or data significantly deviating from the normal observation data, have higher MD Scores (MDS) than those within NS. Mahalanobis-Taguchi System (MTS), including some modified and expanding MD versions, is a well-established technique to detect implausible observations that deviate from the normative data along the statistically analyzed space and well-setting thresholding. Many MTS applications have been reported in aerospace, medical diagnostics, manufacturing, structural damage detection, environmental risk management, and public administration. [27-33]. According to previous research, Ramlie et al. [34] said, "In MTS, to classify any two or more samples among the sample groups, MD values for each sample are calculated based on their common feature datasets." The MD values computed are viewed as points in the high dimensional space, and they represent the distances of the corresponding samples from each in a univariate scale; Ramlie et al. [34] also stated, "an accurate classification result depends on a threshold value or a cut-off MD value that can effectively separate the two classes. Obtaining a reliable threshold value is very crucial". Their study compared four thresholding methods commonly used in MTS methodology (probabilistic thresholding method [35], Type-I and Type-II errors method [36], ROC curve method [37], and control chart method via Box-Cox transformation [38]. They employed multiple data sets and concluded that none of the four thresholding methods outperformed one of the others in most of the datasets. The effective use of the four thresholding methods to produce

promising classification performances is dataset-dependent. Given that this study used the AUC of the ROC curve method to validate the binary classification model and algorithm, our MD threshold setting was also validated by the ROC curve method.

I-5. Study Objectives

This feasibility study aimed to determine if DTW-based voice biomarkers could effectively and significantly differentiate between mild and moderate illness I for COVID-19 through binary classification.

II Methods

II-1. Study Design

We conducted a cross-sectional study using the voice samples of patients with COVID-19.

II-2. Ethics Approval

The study was conducted per the Declaration of Helsinki. Further, the Ethics Committee of Kanagawa University of Human Services approved the study (SHI 3-001, dated May 27, 2021, and SHI 26, dated November 25, 2021). Informed consent was obtained from all participants involved in the study.

II-3 Participants

This study recruited participants through a brochure distributed exclusively to patients with COVID-

19 aged 20 years and above. In particular, those who tested positive for SARS-CoV-2 in PCR testing recuperated at designated facilities or homes in Kanagawa prefecture, Japan, between June 2021 and March 2022. Patients who consented to the study's objectives were requested to register for participation using the QR code on the brochure through their smartphones. The participants were asked to provide data on their daily vital signs, including voice recordings during recuperation. Nonetheless, they could withdraw from the study (opt-out) anytime. A ¥1000 (US $6.68) Amazon gift card was given to participants as compensation. This study only included patients with mild illness or moderate illness I. Thus, for those who did not require hospitalization, patients with moderate illness II were not included. The participants were categorized into mild and moderate I groups, as illustrated in Table 1.

II-4 Data Collection

Those who agreed to participate were asked to install a smartphone app and enter their basic information, vital signs data (temperature and SpO2), symptom scores, and voice recordings on the first day of recuperation. From the second day onward, the participants were required to enter their vital signs data, symptom scores, and voice recordings daily until the last day of recuperation. Voice data were stored with text data indicating the symptoms and vital signs on a highly secure dedicated server. The voice recordings were in the WAV format with a sampling rate of 48 kHz and a bit depth

of 16 bits using the three long vowels /a/, /e/, and /u/. Participants were asked to explain their reasoning if they wished to withdraw from the study. Table 2 shows the timing and data entry items of the participants. The /a/, /e/, and /u/ sounds are in many worldwide languages. Therefore, by obtaining samples using long vowels in these sounds, the research results are expected to be useful across international and regional boundaries of human races.

Table 2. Timing and items for the input data and information from participants

Timing and items	Description
Baseline	
Orientation	Explain the purpose of the study and obtain participant consent
Basic information	Sex, Age, Symptom onset date, Diagnosis confirmation date, Treatment start date
On a daily basis during recuperation	
Vitals sign data	Body temperature, Blood oxygen saturation
Questionnaire	Change in symptoms, Symptomatic or not, Respiratory distress, Taste/olfactory disorder, Cough/sputum, Chest pain, Runny nose/nasal congestion, Sore throat, Nausea/vomiting, Diarrhea, Appetite, Fatigue, Headache, Joint pain, Rash, Red eyes
Voice recording	Three sustained vowels: /a/, /e/, /u/
Dropout during recuperation	
Dropout	Confirm the reason for dropout from the research

II-5. Waveform Samples Cutout and Standardization (Paper I)

II-5-(1) Waveform Samples Cutout and Standardization for DTW Distance

In calculating the DTW distance, a 10-cycle waveform sample was extracted for each date from the participants' long-vowel recordings in the WAV format using Audacity (version 3.1.3; Audacity Team) at a sampling rate of 48 kHz (Figure 2). Then, standardization was achieved along the power axis within the range of –1 to +1 as the maximum amplitude, and the time axis involved 1000 data points, multiplied by 1/48,000 seconds, considering the length of a 10-cycle waveform. R (version 4.4.2; R Core Team) with the tuneR package (version 1.4.0) was used to read and standardize the WAV data.

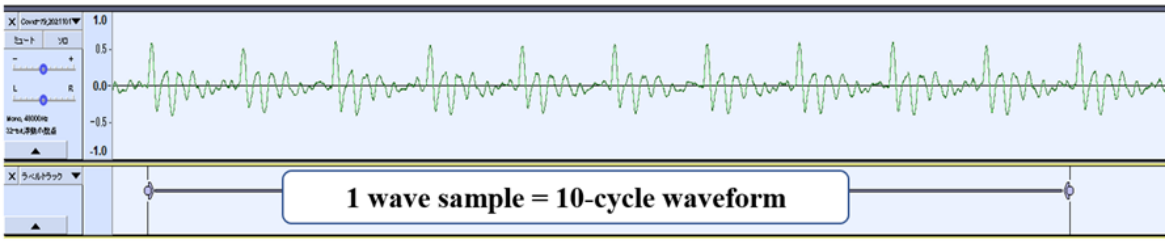


Figure 2. Screenshot showing 10-cycle waveform data extracted for each date from each patient's voice recording of vowels.

II-5-(2) Calculation of the DTW Distance for two Groups

After standardizing the power and time of the 110 waveform samples, the DTW distance was calculated for each sample paired with those of the remaining 109 samples. The obtained DTW distances were divided into 2 categories based on 2 kinds of labels for the 109 waveform samples.

Therefore, each sample was assigned two variables for DTW distance. The mild group had 61 or 60 DTW distances; the moderate I group had 48 or 49 DTW distances. The average and variance of the DTW distances were calculated for each group. For the mild group, the average index (i.e., the mild-group filtering [MiF] average) and variance index (i.e., MiF variance) of the DTW distance was obtained. In contrast, for the moderate I group, the average index (i.e., moderate-group filtering [MoF] average) and variance index (MoF variance) of the DTW distance were obtained. These four indices were obtained from a single waveform sample. The indices for the three vowels, /a/, /e/, and /u/, were prepared and are represented in Table 3. Thus, 12 indices were used in the subsequent analyses. The two groups' average values and variances of the DTW distances were statistically investigated to determine whether they exhibited significant values for the 2-group classification scheme.

Table 3. Twelve indices were generated from 3 vowels and 4 indices.

Indices	**Vowels**		
	/a/	**/e/**	**/u/**
MiF_average	**/a/**-MiF_average	**/e/**-MiF_average	**/u/**-MiF_average
MoF_average	**/a/**-MoF_average	**/e/**-MoF_average	**/u/**-MoF_average
MiF_variance	**/a/**-MiF_variance	**/e/**-MiF_variance	**/u/**-MiF_variance
MoF_variance	**/a/**-MoF_variance	**/e/**-MoF_variance	**/u/**-MoF_variance

MiF: mild-group filtering, MoF: moderate-group filtering

II-5-(3) Waveform Samples Cutout for Jitter, Shimmer, and HNR

Another Jitter, Shimmer, and HNR dataset was prepared for the comparison model. The identical 10-cycle waveforms from the same 110 patients, dates, and three vowels as used for DTW-distance were employed to extract these three acoustic indices of jitter, shimmer, and HNR, except for standardization to time and power axis because standardization may result in a loss of acoustic features of them.

II-6 Data Analysis

II-6-(1) Linear Discriminant Analysis Considering the Average and Variance Indices of the DTW Distance (Paper I)

The Mann-Whitney U test was used to determine whether any statistical significance existed between the mild and moderate I groups. This test was performed on 12 indices that measured the average and variance of the DTW distance for the 3 vowels /a/, /e/, and /u/. A significance level of 1% was established, with the null hypothesis of no statistical significance between the two groups. Box plots and linear discriminant analysis (LDA) were used to determine the indicators of the three vowels most effective for determining statistical significance between the two groups. The confusion matrix obtained from the LDA results was displayed with a specific index for the true positive rate (TPR), true negative rate (TNR), and balanced accuracy (BA). The boxplot function was calculated and plotted using the R ggplot package (version 3.4.0), and the LDA function was calculated using the R

MASS package (version 7.3).

II-6-(2) Generalized Linear Model with the Promising Indices of DTW Distance (Paper I)

The significant indices from the four categories of DTW distance were used to distinguish between severity levels (mild or moderate I). These indices were then used as explanatory variables to create generalized linear models (GLMs) for each vowel. A five-fold cross-validation method with 110 waveform samples was used to train the model for each vowel, which was subsequently used to predict the severity classification. R was used for GLM modeling and label prediction. The pROC package (version 1.18.0) for R was used to obtain the receiver operating characteristic (ROC) curve and calculate the area under the curve (AUC). In contrast, confusion matrices were generated using the Caret package (version 6.0) for R.

II-6-(3) Generalized Linear Model with Mahalanobis Distance

II-6-(3)-A. Normal Space and Thresholding of Mahalanobis Distance

When collecting observation data from the real world, but not limited to arbitrary observation data, it is commonly believed that the data include some amount of implausible observation, even if it is sparse. Some datasets for this study needed to exclude implausible observations using the MD method. Here is the procedure for that.

1. First, for the 110 patient samples, the MDSs were obtained from the two DTW-based variance indices, MDS1 for the MiF-Variance and MDS2 for the MoF-Variance, respectively.

2. The mean and variance-covariance for the mild and moderate I groups were calculated, and the MDS was calculated as the distance from the norm. For this calculation, the “Mahalanobis” function contained in the "stats" Package (ver. 3.6.2) of R was applied.

3. Each sample was assigned MDS coordinates (MDS1 and MDS2), allowing a review of the distribution on a two-dimensional plane. The center coordinates (Mu1 and Mu2), and the level envelope of the NS in an ellipse is displayed alongside each group’s coordinates. If the level value is set to 90%. 90% of the data points can be found within an ellipse.

4. We set the NS level at 90% for both groups and observed how much the two NSs overlapped on the graph. We also ensured that most MDS points outside the NSs could be taken as *implausible observations.* If not, we adjusted the NS level by 5% up or down to reach an optimal level. We expected the NS to include as many normal samples as possible at this level while excluding most implausible observations. Finally, each group was given its own MD thresholding (MDt)

corresponding to the given NS level. The dataset of 110 patient samples was reconstructed by excluding MDSs with values more than the MDt.

5. This study analyzed two GLM models: one that used MDS1 and MDS2 in an optimized dataset and another that utilized DTW-based variance indices. To determine the accuracy and performance of both models, we used the confusion matrix and AUC of the ROC curve. If the new model shows improvement in these areas, we can conclude that the optimization of NSs was successful. We defined the most optimized normal space as the *silver standard of normal spaces (SSNS)* for this study, which provides the *standard parameters* of the center of coordination and the variance-covariance information to classify arbitrary datasets.

II-6-(3)-B. MDS-based Generalized Liner Model to Classify Arbitrary Observation Data

The procedure for handling arbitrary observation data using the DTW Distance and SSNS is as follows.

1. Using the center coordinates, variance and covariance information of SSNS, generate MDS1 and MDS2 from the variance indices of an arbitrary dataset. We plotted them on a 2D coordinate system, including the center points and an ellipse envelope displaying the NSs of arbitrary observation to analyze the distribution of all MDS points, including implausible observations.

2. To obtain the most optimized MD thresholds (MDt) for the mild group (K1) and the moderate I group (K2), we shifted the level of the NS of the arbitrary dataset from 100% to 65% in 5% increments. We then determined the optimized combination of K1 and K2 by examining a line graph of the MDt (K1 and K2) to the NS level. The dataset of arbitrary observation was reconstructed for the severity classification by excluding the MDSs with values more significant than the MDt (K1 and K2).

3. The MDS-based GML model was generated based on the probability-of-fit and cut-off values obtained from SSNS and then used to classify the arbitrary observations data, which were within the designated level of NSs.

4. We evaluated the results' balanced accuracy, sensitivity, specificity, and AUC using confusion matrices and ROC curves. We aimed to determine whether MDt could effectively filter out the implausible observations of arbitrary observation data, resulting in reasonable NSs finally used for a binary classification.

5. Finally, we analyzed the MDS information beyond MDt individually to determine if it was due to patient bias, model limitations, or unlikely observations with no reason.

II-6-(4) Generalized Linear Model using Jitter, Shimmer, and HNR Indices

The significant difference between the two was tested using the Mann-Whitney U test. The significance level was set to 1%, with the null hypothesis of no significance between the two groups. The distribution of severity labels for each factor is also displayed as boxplots per factor and vowel. A GLM model with three variables–jitter, shimmer, and HNR–was created using 5-fold cross-validation and evaluated using the ROC curve/AUC and confusion matrix. The R packages used to extract jitter, shimmer, and HNR indices from the waveform samples were the Sonicscrewdriver package (ver. 0.0.4) and the Soundgen package (ver. 2.5.2).

III Results

III-1 Participants

In June 2021 and March 2022, approximately 540,000 patients with COVID-19 were recorded in Kanagawa prefecture. After requesting approximately 10,000 individuals to participate, our study recruited 295 participants in the same period. Of these, 291 were eligible to participate because participants who did not meet the inclusion criteria, such as minors, those with invalid data registration, or those who withdrew midway through evaluation, were excluded.

A total of 74 participants who reported no symptoms of coughing, throat pain, chest pain, or SoB

during recuperation were assigned to the mild group. Of the 217 participants who reported any symptoms, 68 with symptoms of SoB were posted to the moderate I group. Of the 149 participants who reported symptoms other than SoB, 6 with SpO2 values of less than 96% were posted to the moderate I group, and 143 participants who reported SpO2 values of no less than 96% were posted to the mild group. The 291 participants were classified into 2 groups: 217 as mild and 74 as moderate I. Figure 3 shows a flowchart of study participation.

The primary periods of infection in Japan were during the Delta period, from July to December 2021, and the Omicron period, from January to June 2022. According to previous reports [39, 40], COVID-19 exhibits varying levels of infectivity, severity, and symptoms, depending on the type of mutant strain present. We identified the time of infection in Japan and carefully matched the 291 study participants who had already been labeled into 2 groups. Table 4 shows the attribution matrix for the participants by the time of infection, sex, and severity. Table 5 shows the attribution matrix for the same sample based on the time of infection, sex, and age group. Finally, 110 participants (61 with mild illness and 49 with moderate illness I) were included in the study.

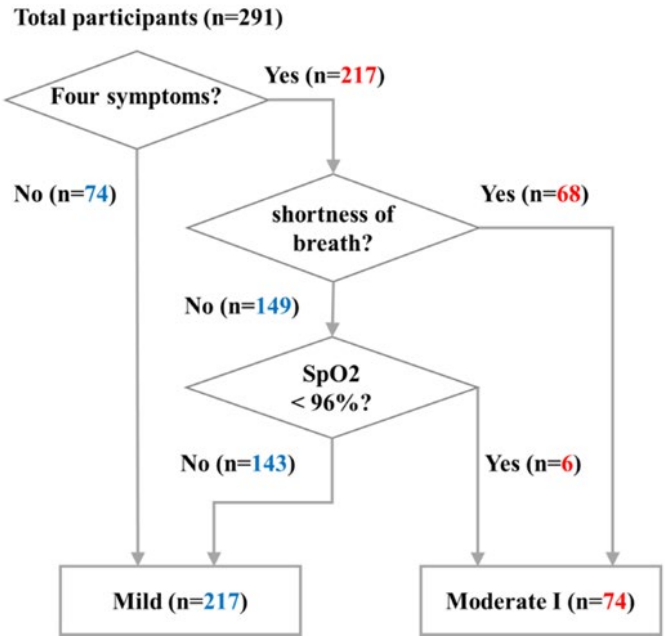


Figure 3. Flowchart showing classification of 291 participants into the mild illness and moderate illness I groups. "Four symptoms" refers to the 4 major symptoms: coughing, throat pain, chest pain, and shortness of breath. SpO2: oxygen saturation.

Table 4. Participants' attribution matrix by the time of infection, sex and severity(n=110)

Severity	Delta period [a] (n=46)	Omicron period [b](n=64)
Mild illness (n=61)		
Male	14	16
Female	13	18
Total	27	34
Moderate illness I (n=49)		
Male	10	16
Female	9	14
Total	19	30

a Delta period from July 2021 to December 2021.

b Omicron period from January 2022 to June 2022.

Table 5. Participants' attribution matrix by the time of infection, sex, and age group (n=110).

Age group (years)	Delta period [a] (n=46)	Omicron period [b] (n=64)
20-29		
Male	9	10
Female	16	15
Total	25	25
30-3		
Male	4	9
Female	1	9
Total 9	5	18
40-49		
Male	7	6
Female	2	5
Total	9	11
50-59		
Male	3	4
Female	3	3
Total	6	7
60-69		0
Male	0	2
Female	0	0
Total	0	2
≥ 70		
Male	1	1
Female	0	0
Total	1	1

a Delta period from July 2021 to December 2021.

b Omicron period from January 2022 to June 2022.

III-2 Linear Discriminant Analysis with DTW Distance Indices (Paper I)

III-2-(1) Distribution of the DTW Distance Indices

Table 6 displays the two groups' Mann–Whitney U test results based on three vowels and four indicators. Of the 12 indices, 6 were found to be significant; they included /u/-MiF average, /a/-MiF variance, /e/-MiF variance, /a/-MoF variance, /e/-MoF variance, and /u/-MoF variance. The only significant index among the average indices was /u/-MiF average, while /u/-MiF variance was the only insignificant index among the variance indices. This indicates that the variance indices were more significant overall. Figures 4 and 5 illustrate the distributions of MiF average and MoF average, as well as MiF variance and MoF variance, for the two groups.

Table 6. Results of Mann-Whitney's U-test for mild illness and moderate illness I group classification.

Indices	**Vowels**, *P* values		
	/a/	**/e/**	**/u/**
MiF [a] _average	.03	.71	<.001
MoF [b] _average	.60	.43	.03
MiF_variance	<.001	<.001	.45
MoF_variance	<.001	<.001	<.001

a MIF: mild-group filtering

b MoF: moderate-group filtering

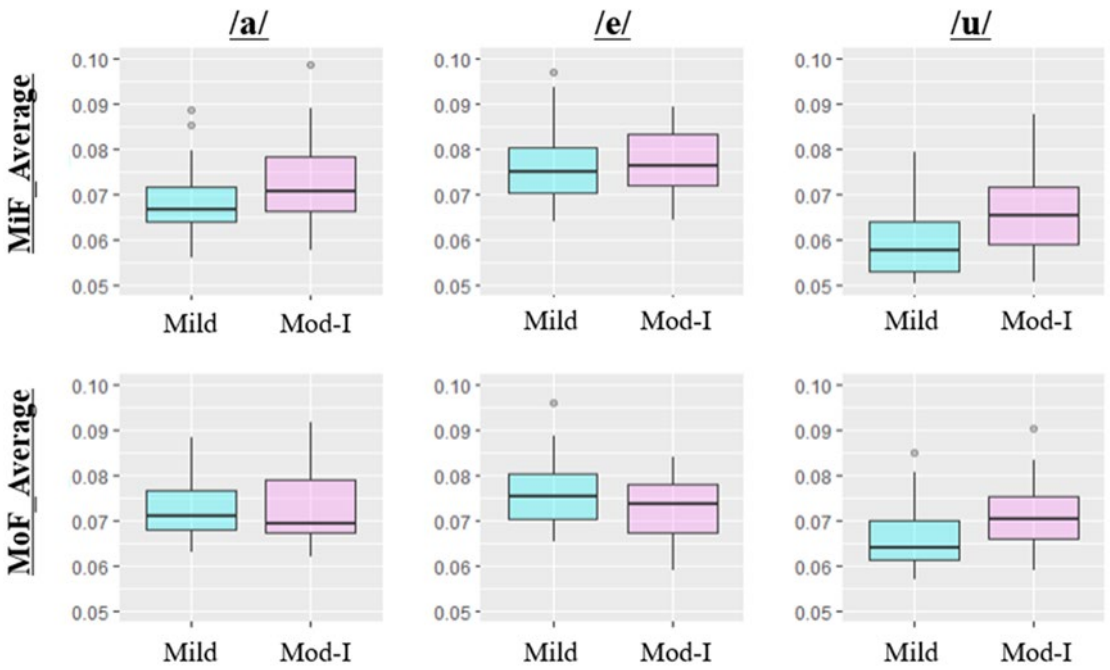


Figure 4. Distribution in the 2 groups of the average index for the dynamic time warping distance (MiF average and MoF average). MiF: mild-group filtering; MoF: moderate-group filtering.

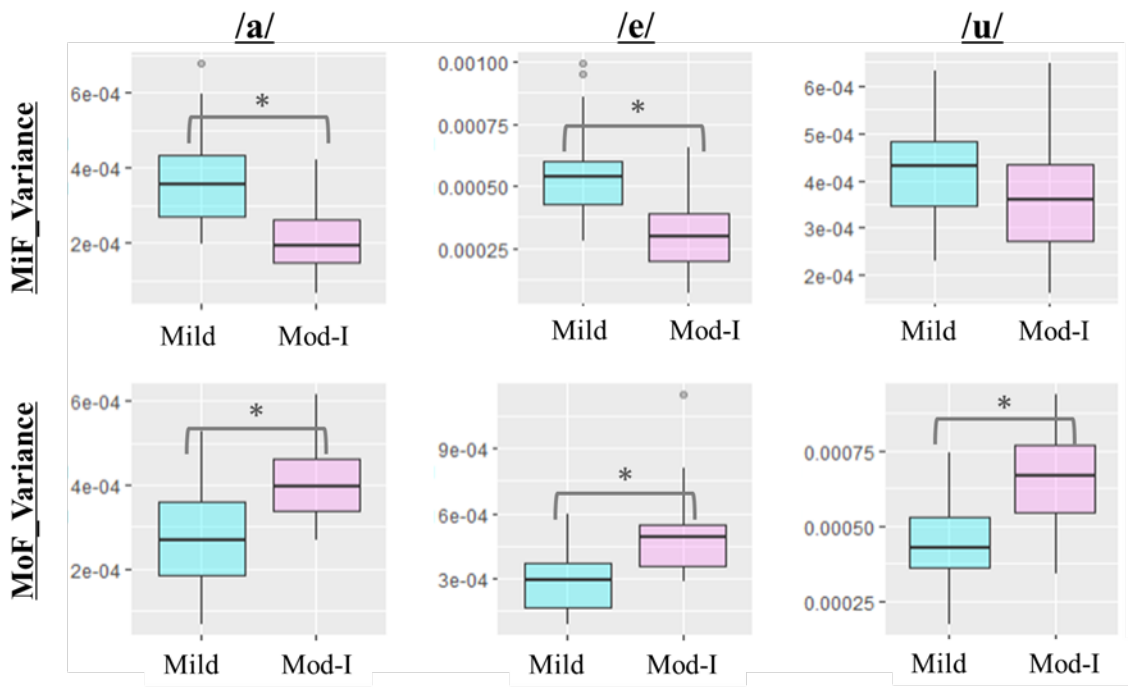
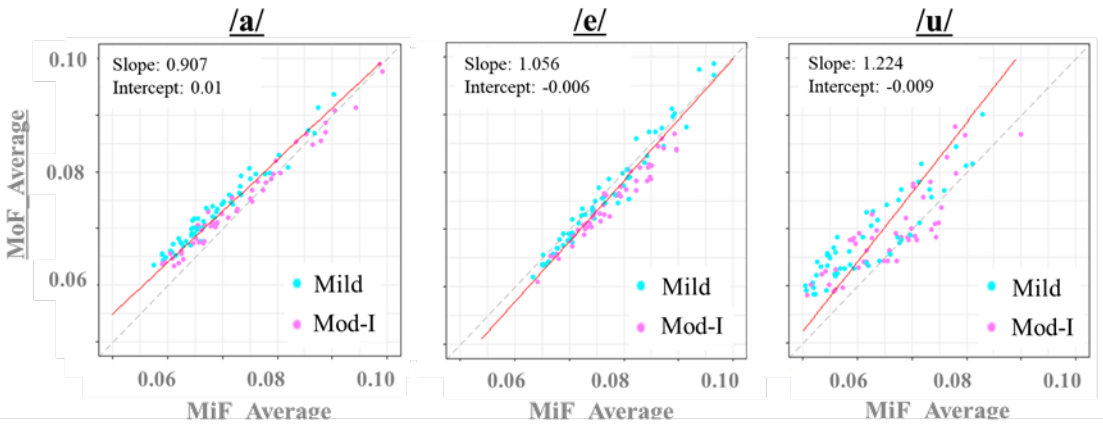


Figure 5. Distribution in the 2 groups of the variance indices for the dynamic time warping distance (MiF variance and MoF variance) *P<.001. MiF: mild-group filtering; MoF: moderate-group filtering.

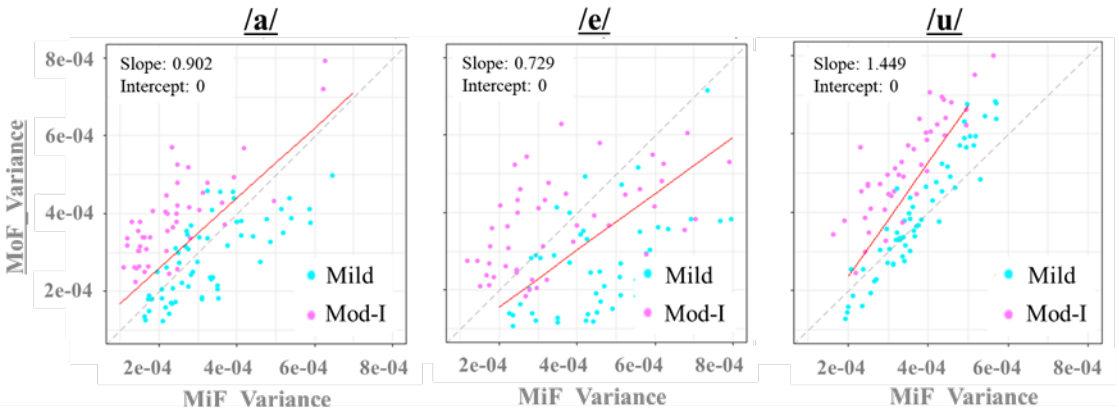
III-2-(2). Results for Linear Discriminant Analysis

Figure 6 shows a scatter plot of the average indices, and Figure 7 shows a scatter plot of the variance indices of the DTW distance, along with the confusion matrix, TPR, TNR, and BA values of the LDA. The straight line represents the discriminant line obtained using LDA. The variance indices of the DTW distance provided superior results for the classification indicators of the confusion matrix, including TPR, TNR, and BA, compared to the average indices. The ease of classification can be visually verified by observing the plots achieved via LDA.



	Mild labeled	Mod-I labeled	Mild labeled	Mod-I labeled	Mild labeled	Mod-I labeled
Mild predicted	54	12	48	17	44	23
Mod-I predicted	7	37	13	32	17	26
TPR	75.5		65.3		53.1	
TNR	88.5		78.7		72.1	
BA	82.0		72.0		62.6	

Figure 6. Linear discriminant analysis results and confusion matrix of the MiF average and MoF average indices. BA: balanced accuracy; MiF: mild-group filtering; MoF: moderate-group filtering; TPR: true positive rate; TNR: true negative rate.

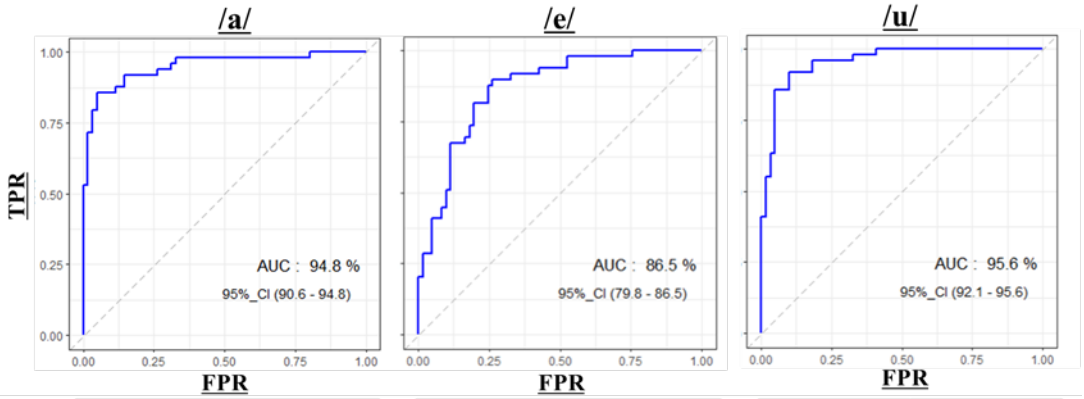


	Mild labeled	Mod-I labeled	Mild labeled	Mod-I labeled	Mild labeled	Mod-I labeled
Mild predicted	53	5	49	11	57	8
Mod-I predicted	8	44	12	38	4	41
TPR	89.8		77.6		83.7	
TNR	86.9		80.3		93.4	
BA	88.4		79.0		88.6	

Figure 7. Linear discriminant analysis results and confusion matrix of the MiF-variance and MoF-variance indices. BA: balanced accuracy; MiF: mild-group filtering; MoF: moderate-group filtering; TPR: true positive rate; TNR: true negative rate.

III-3. Generalized Linear Model with Promising Indices of the DTW Distance (Paper I)

We used the variance indices of the DTW distance as predictors of the GLM model because they achieved better classification performances than the average indices. Figure 8 shows the ROC and AUC values of the GLM model for each vowel with the confusion matrix, including the TPR, TNR, and BA data. The models of all three vowels exhibited high performance regarding the AUC and mid-to-high accuracy for TPR, TNR, and BA.



	Mild labeled	Mod-I labeled	Mild labeled	Mod-I labeled	Mild labeled	Mod-I labeled
Mild predicted	53	7	43	5	55	7
Mod-I predicted	8	42	18	44	6	42
TPR	85.7		89.8		85.7	
TNR	86.9		70.5		90.2	
BA	86.3		80.2		88.0	

Figure 8. ROC/AUC results of the generalized linear model with confusion matrices for the mild-group filtering variance and moderate-group filtering variance indices. AUC: area under the curve; BA: balanced accuracy; FPR: false positive rate; TPR: true positive rate; TNR: true negative rate; ROC: receiver operating characteristics curve.

III-4. Generalized Linear Model with Mahalanobis Distance

III-4-(1) Silver Standard of Normal Spaces (SSNSs) and Thresholding of Mahalanobis Distance (MDt)

Using the MD method, we used the /u/ variance indices of DTW distance as a representative example for generating SSNS and MDt. It showed the highest classification accuracy in the GLM (88.0%, Figure 8). Our goal is to observe the overlap between the SSNSs of the two groups and the number of out-of-SSNS observations generated to the given SSNS level (e.g., 85%), as shown in Figure 9.

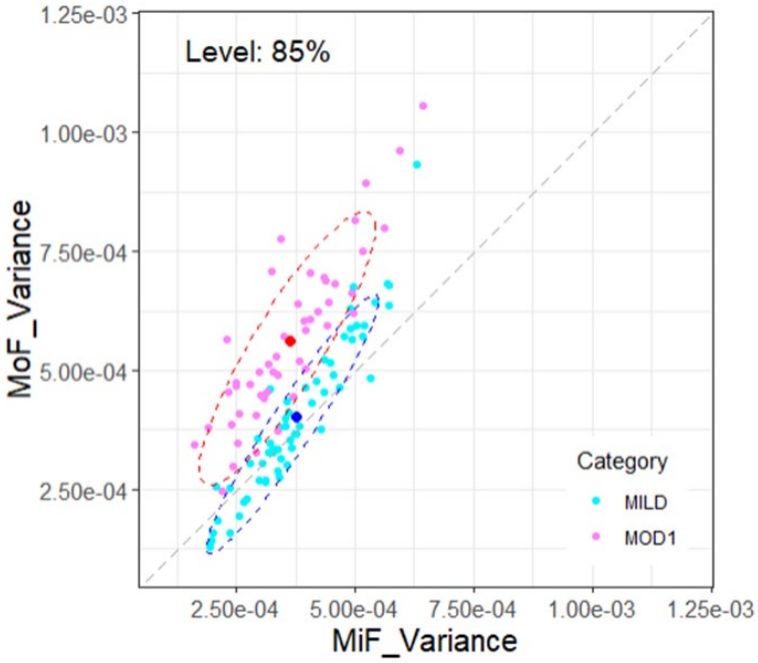


Figure 9. Distribution of the variance indices of DTW distance with 85 % of SSNS level.

We utilized the MD method to transform the variance indices of DTW distance into MD Scores (MDS). Then, we applied the same level of ellipse equation range (85%) to the distribution state of 100% of MDSs for comparison as is Figure 10. The figure shows that the distribution state of 100 % MDSs for each group and the positioning of each SSNS.

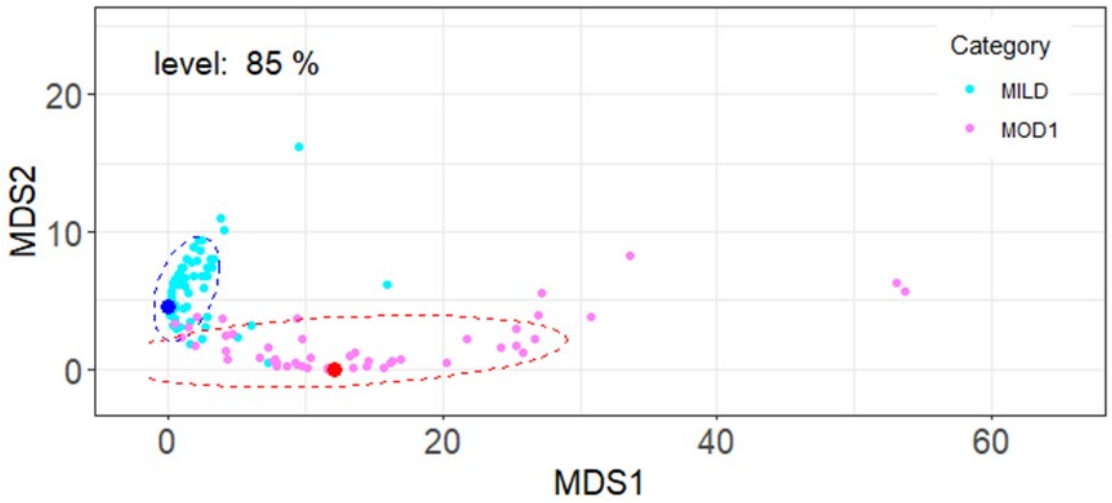


Figure 10. Distribution of MD Scores (MDS1 and MDS2) with 85% of SSNS level.

An example GLM model was generated corresponding to SSNS (85% level), and the model performance of binary classification is shown in Figure 11 as ROC curve and AUC, and the model accuracy is tabulated in Table 7 just for reference. For this model, out-of-SSNS MDSs were excluded by MDt parameters.

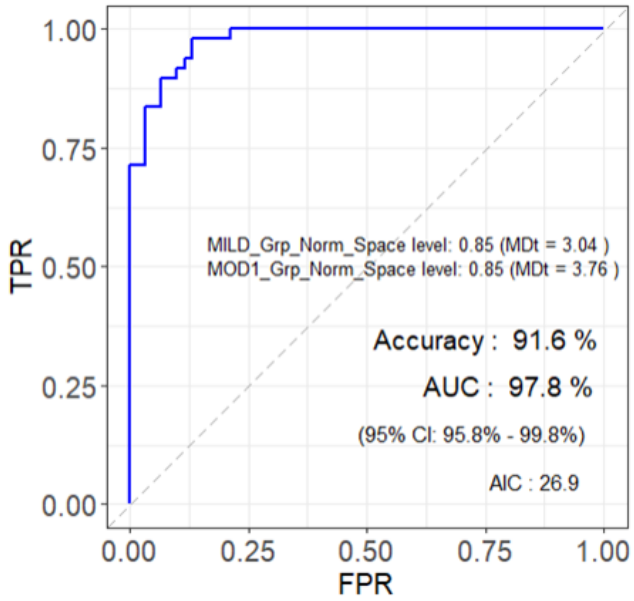


Figure 11. A representative example of a ROC curve with two SSNS levels of 85%.

Table 7. A representative example of the model accuracy with two SSNS levels of 85%.

Model Accuracy	**(%)**
Sensitivity	89.8
Specificity	93.4
Balanced Accuracy	91.6

Next, determined each group's optimal combination of levels. This involves selecting the appropriate level, removing implausible observations outside the SSNS at the given level, and maintaining the number of samples within the given levels as many as possible. We calculated two MD thresholds (k1 for the mild group and k2 for the moderate I group), the balanced accuracy, sensitivity, specificity, and AUC for each of the given levels of SSNS, which was shifting from 70% to 100% in 5% increments and plotted, as shown in Figure 12. Our analysis concluded that around 98% contributed to the best balance of the four model accuracy indices of AUC, sensitivity, specificity, and balanced

accuracy for SSNS. This level was used to finalize the SSNSs using MDt for the mild group (k1) at 9.49 and MDt for the moderate I group (k2) at 8.30.

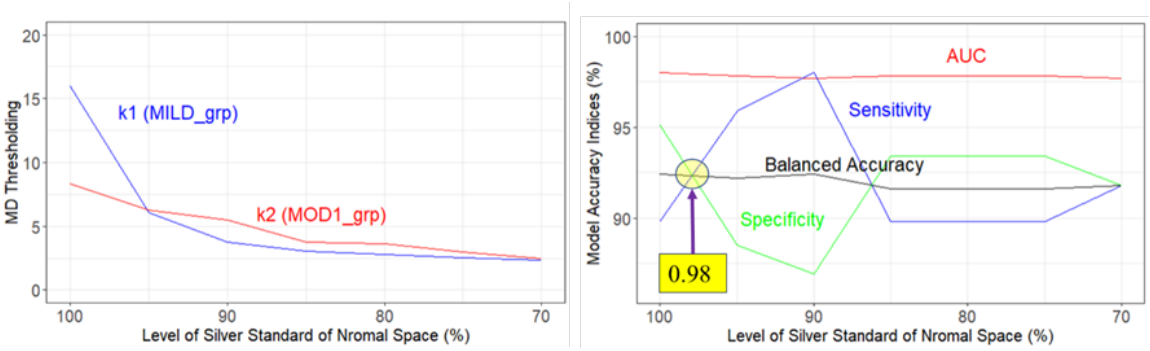


Figure 12. Transient of MDt (k1, k2) and model accuracy indices for each level for SSNS.

The SSNS for this study for the /u/ vowel was described as follows:

1. The ROC Curve (Figure 13) has an AUC of 98.0% and a balanced accuracy of 92.4%.
2. The model accuracy is shown in Table 8 with a sensitivity of 89.8%, specificity of 95.1%, and balanced accuracy of 92.4%.
3. Figure 14 shows the distribution state of 109 MDSs fitting the SSNS.
4. The only MDS excluded from the SSNS was one MDS with ID #02F618 and a mild label, as shown in Figure 15.

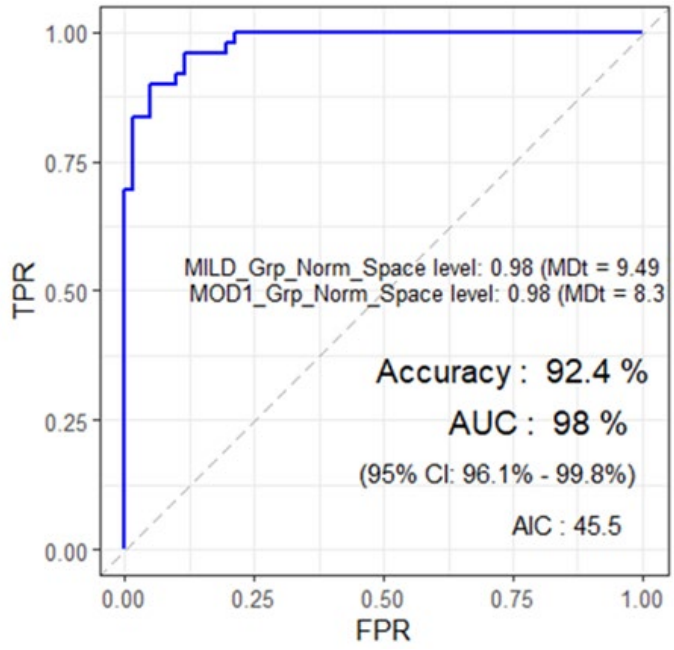


Figure 13. ROC curve of the GLM model with the SSNS

Table 8. The model accuracy with the SSNS.

Model Accuracy	**(%)**
Sensitivity	89.8
Specificity	95.1
Balanced Accuracy	92.4

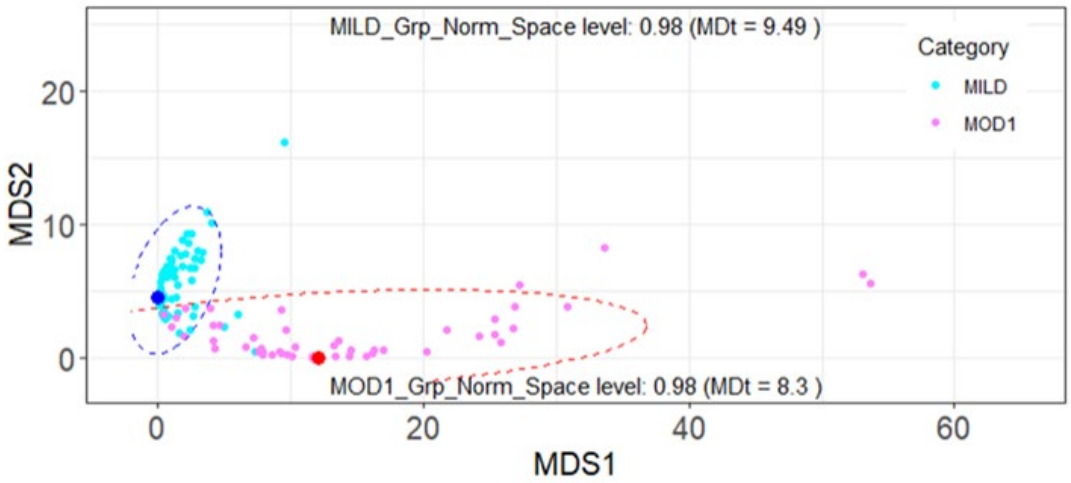


Figure 14. The state of fitting MDSs for each group at 98% of the NS level.

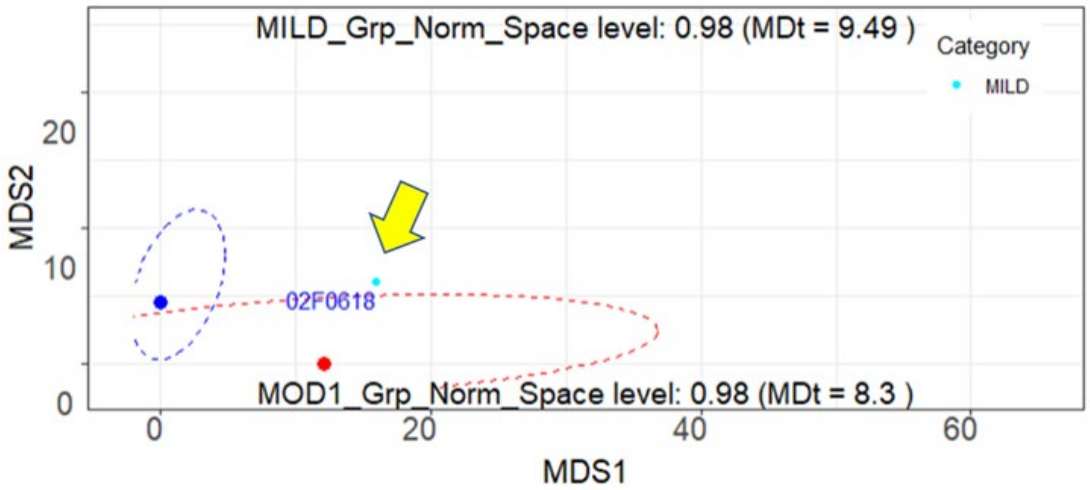


Figure 15. Identifying the out-of-SSNS MDSs

III-4-(2) MDS-based Generalized Linear Model to Classify Arbitrary Observation Data

Figure 16 presents the distribution of DTW_distance variance indices and the 85% of NS for the mild and moderate I groups in arbitrary observations and indicates the location of the SSNS center coordinates in this graph. The arbitrary observation dataset means all observation data other than the 110 samples used to SSNS, with a sample of 172 (mild: 94, moderate I: 78) for /u/. This graph was prepared as a comparison of the state of the distribution when DTW-variance indices were converted to MD scores.

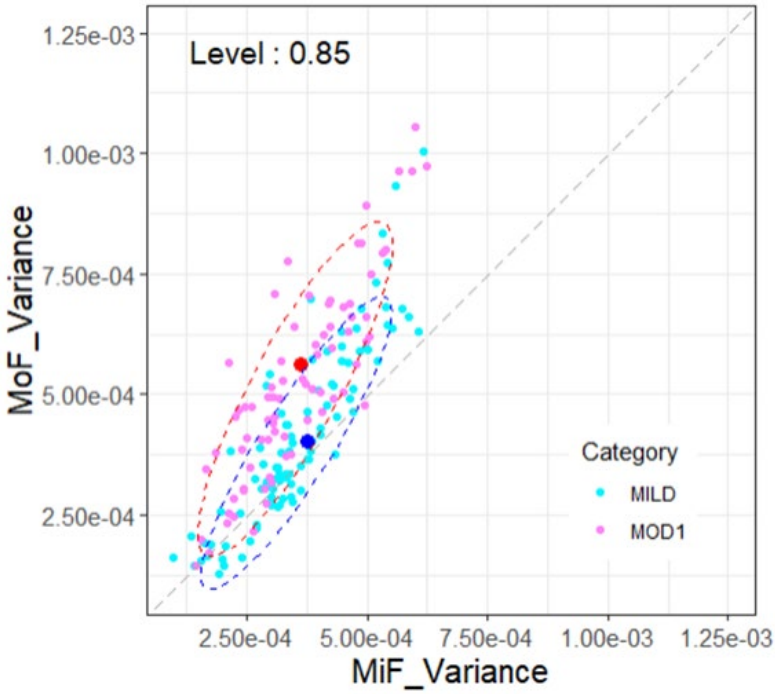


Figure 16. Distribution of DTW variance indices of arbitrary observation for each group. 100% of MDSs (172) are plotted. Two 85% ellipse equation ranges are reference only.

We utilized the MD method to transform the variance indices of DTW distance into two MD Scores (MDS1 for MiF_Variance and MDS2 for MoF_Variance) by applying the center coordinates and variance/covariance information of the SSNS to the arbitrary dataset. The distribution state of MDS is shown in Figure 17.

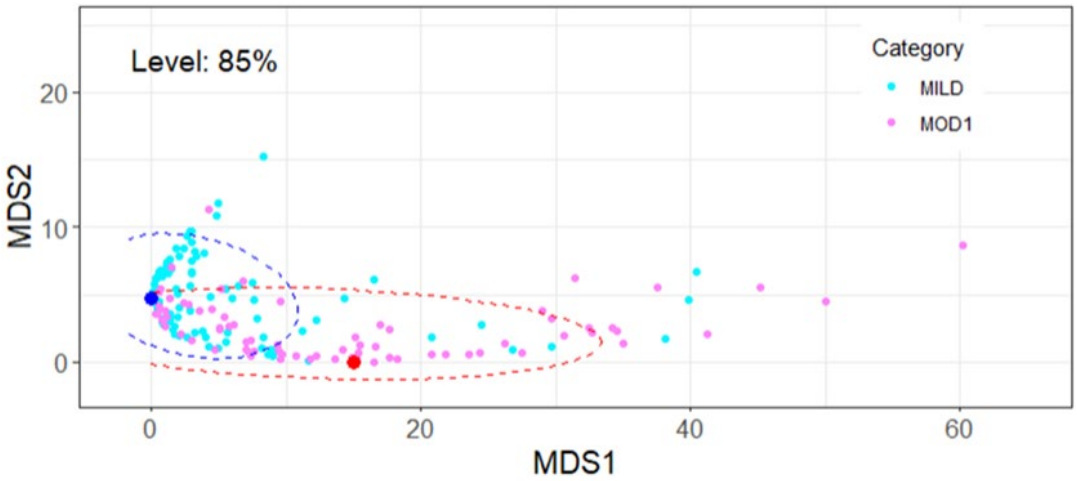


Figure 17. Distribution of MD Scores (MDS1 and MDS2) with 85% of NS level of arbitrary observations.

Figure 18 shows transitions in the MD threshold (MDt), AUC, balanced accuracy, sensitivity, and specificity for each group when the NS level of arbitrary observations was changed from 100% to 65% by 5%. Based on the insights obtained from Figure 18, the mild group's NS level was 85% (MDt of 8.64), and the NS level of the moderate I group was set at 75% (MDt of 3.95).

1. The rate of change in K1 showed a gentle slope when the NS level of the mild group was in the range of 80 to 90% (Figure 18, left). In addition, the specificity, the detection rate of the mild group, also showed a gentle slope of around 90% as the NS level is between 85 and 75% (Figure 18, right). Therefore, the greater NS level for the mild group between these two ranges is 85%.

2. The slope of the K2 rate changed to gentle after 90% of the NS level, allowing us to judge that implausible observation data were well filtered out (Figure 18, left). In addition, the sensitivity, which is the detection rate of the moderate I group, continuously increased as the NS level moved from 85 to 65% (Figure 18, right). Therefore, the NS level of the moderate I group needs to be 75% or less if you want 80% or more of the sensitivity of the model.

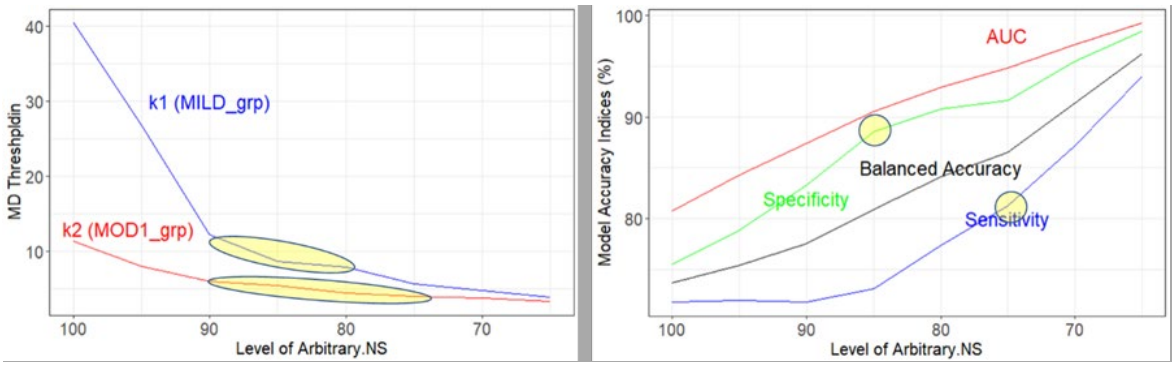


Figure 18. Transient of MDt (K1, K2) and model accuracy indices for each NS level of arbitrary observation datasets.

The ROC curve generated by the MDSs was extracted from the arbitrary observation dataset (Figure 19). Table 9 shows the confusion matrix results. Figure 20 shows the distribution of the 147 data fitted to the NS. Figure 21 maps implausible observations with wrong predictions against the true labels (25 observations, 14.5% of 172 whole arbitrary samples). The model's classification performance for the MDSs fitted to the NS was 93.9% of AUC, indicating high performance, and the balanced accuracy of the model was 85.0% (sensitivity 81.3%, specificity (88.6%), which is considered a reasonable

result.

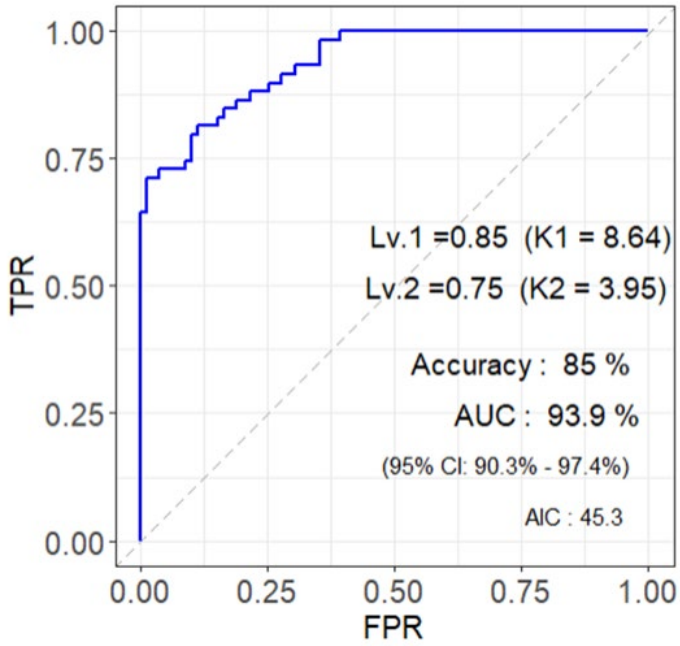


Figure 19. ROC curve of MDS fitted to the NS of the arbitrary observations.

Table 9. The model accuracy of arbitrary observation dataset with the gold NS method.

Model Accuracy	**(%)**
Sensitivity	81.4
Specificity	88.6
Bal.Acc	85.0

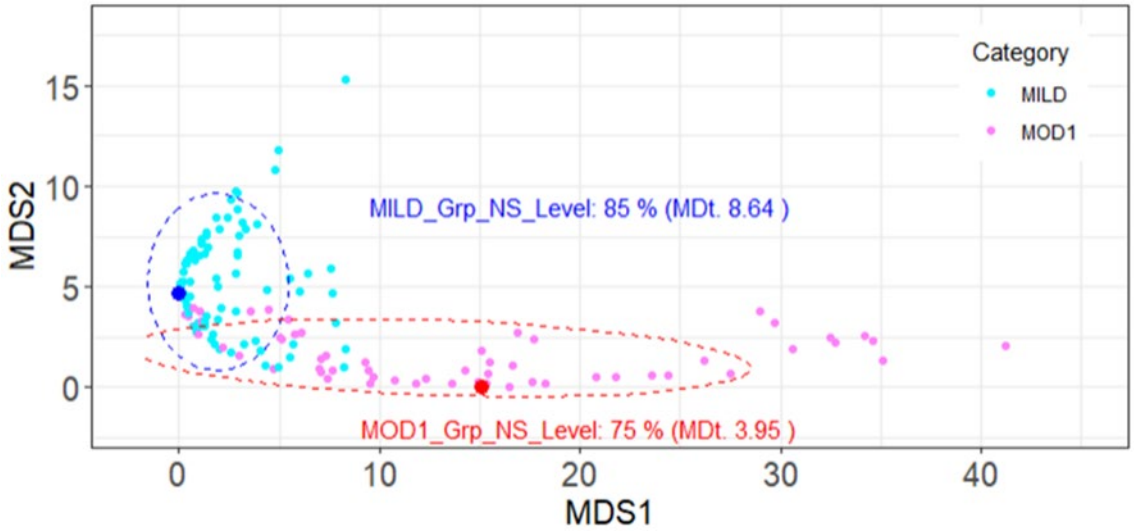


Figure 20. The distribution state of the MDSs fitting to the NS of each group.

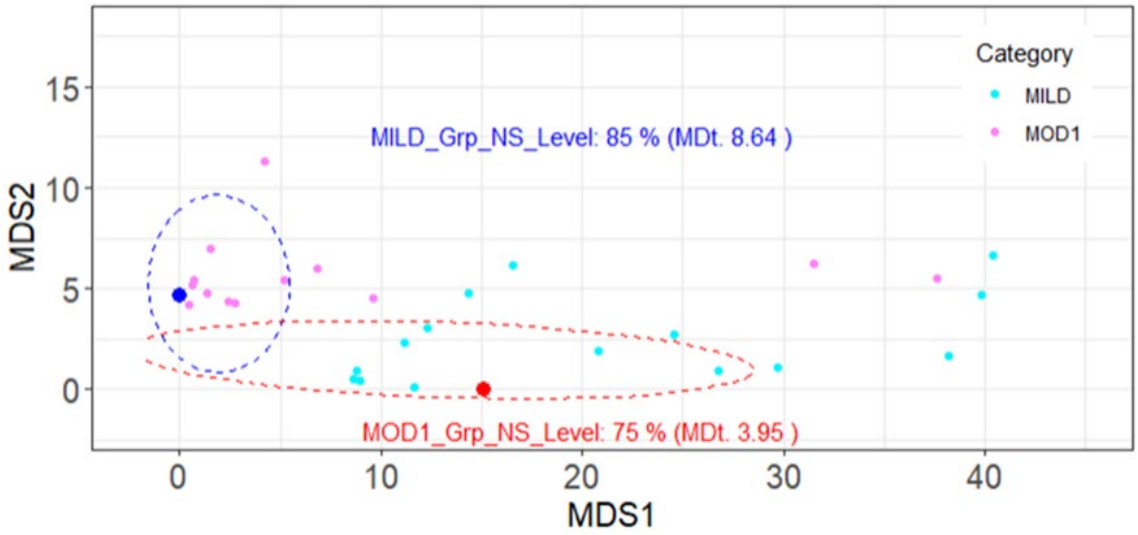


Figure 21. Mapping of implausible observations with wrong predictions against the true labels.

The table in Appendix 1 displayed incorrect moderate I predictions for true mild labels. It included 14 samples, and the associated distribution mapping is shown in Figure 22. The table in Appendix 2 displayed incorrect mild predictions for the true moderate I label. It included 11 samples, and the

associated distribution mapping is shown in Figure 23. Our best insights obtained from the related tables and figures were described in the “Possibility or Assumption” column of tables in Appendix 1 and 2. We can summarize why the label mismatching happened in 25 cases as follows:

(1) Patient bias (11 cases)

(2) Current model limitation (12 cases)

(3) Unknown - yet implausible observation (2 cases)

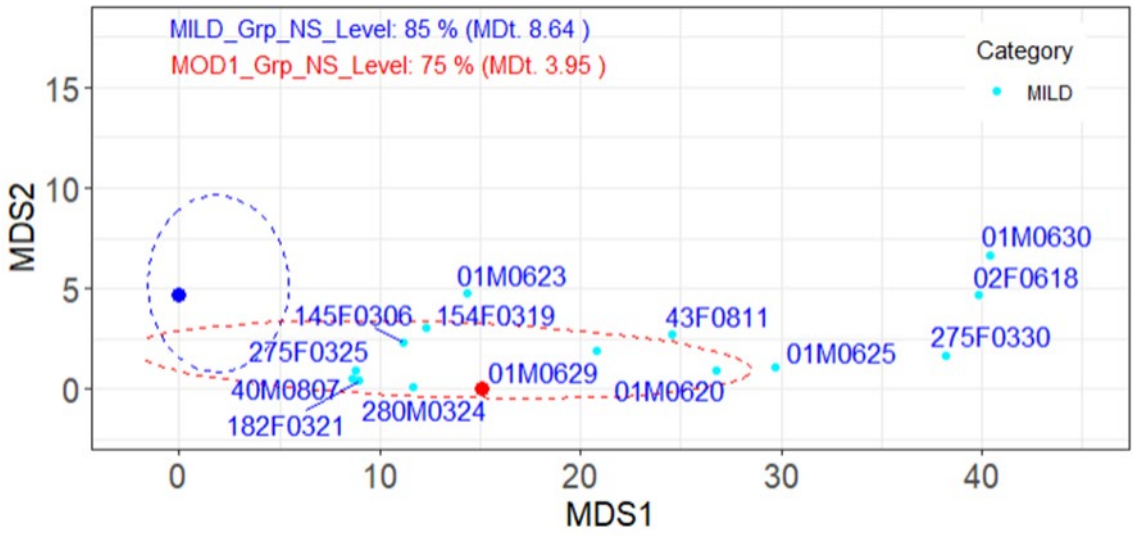


Figure 22. Mapping of the wrong 14 predictions as moderate I against true mild labels.

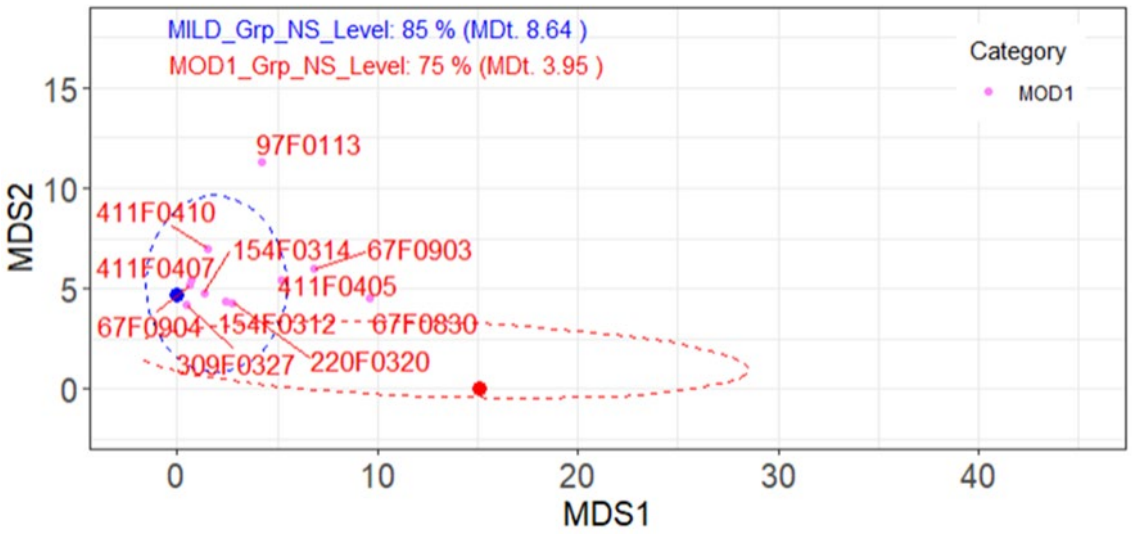


Figure 23. Mapping of the wrong 11 predictions as mild against true moderate I labels

III-3 Generalized Linear Model with Jitter, Shimmer, and HNR Indices

Table 10 presents the results of the Mann–Whitney U test for each of the three vowels and each factor for the two groups' classification. No index has a significance level of 1%. Figure 24 presents the distribution of the two groups for the jitter, shimmer, and HNR indices.

Table 10. Results of Mann-Whitney's U-test for Results of Mann-Whitney's U-test for mild illness and moderate illness I group classification, using Jitter, Shimmer and HNR

Indices	**Vowels,** *P* values		
	/a/	**/e/**	**/u/**
Jitter	.18	.76	.39
Shimmer	.07	.72	.55
HNR	.67	.99	.60

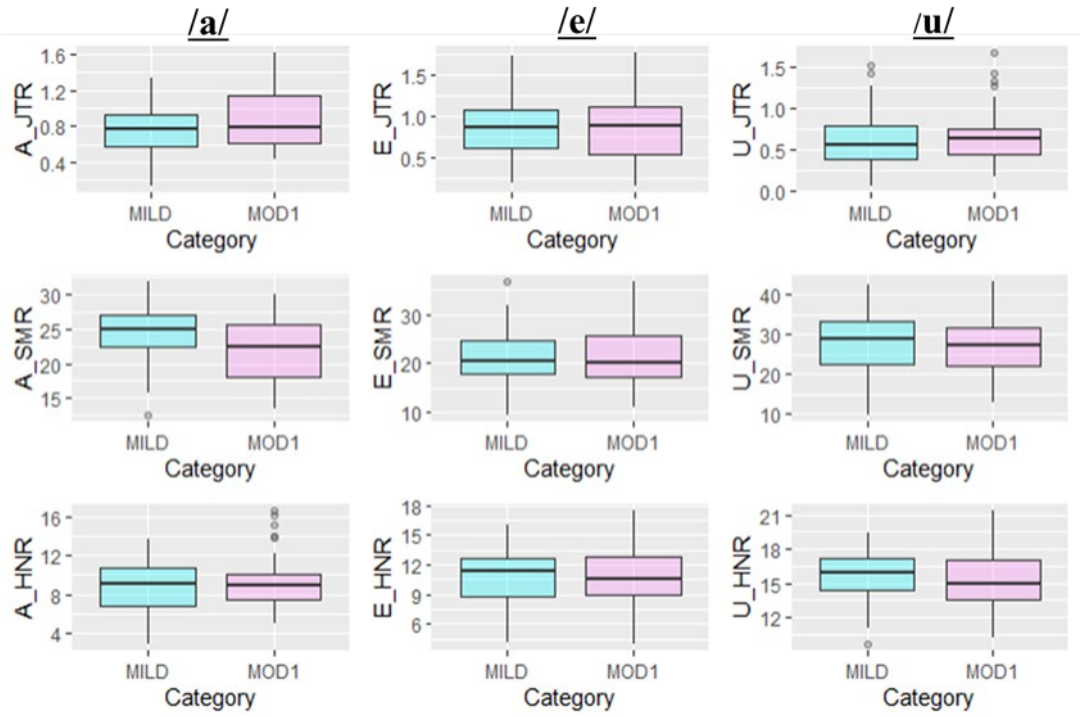


Figure 24. The boxplot of the two groups shows no significant difference in the data distribution. MILD: mild group, MOD1: moderate I group.

Figure 25 presents the predictive ability of the GLM model for each vowel in terms of ROC curves and AUCs, with Table 11 of the confusion matrix with sensitivity, specificity, and balanced accuracy. GLM using one-dimensional information of Jitter, Shimmer, and HNR as predictors did not work well in classifying mild and moderate I groups.

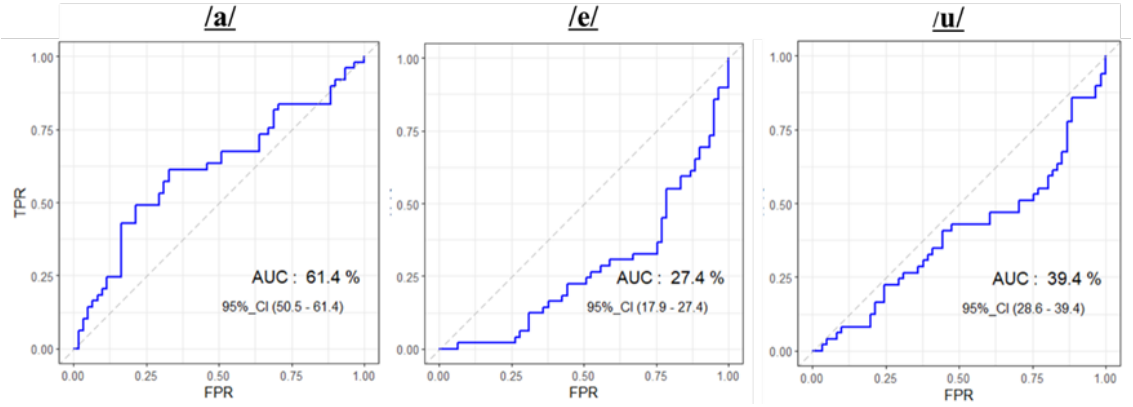


Figure 25. The result of the GLM model's ROC/AUC using Jitter, Shimmer, and HNR indices.

Table 11. The model accuracy of GLM with Jitter, Shimmer, and HNR indices

Model accuracy	**Vowel, (%)**		
	/a/	**/e/**	**/u/**
Sensitivity	53.1	30.6	24.5
Specificity	67.2	44.3	65.6
Balanced Accuracy	60.2	37.5	45.0

IV. Discussion

IV-1. Principal Results

This feasibility study demonstrated that the DTW distance-based voice biomarkers generated by the GLM had a balanced accuracy ranging from 80.2% to 88.0% and a high model performance indicated by the AUC ranging from 86.5% to 96.5% for 3 vowels of 110 samples of patients with COVID-19 to classify mild and moderate illness I. Furthermore, the two-story classification model, an improved version supported by MD thresholding to deal with arbitrary observation that may include a few implausible observations, achieved a reasonable model accuracy of 85.0% and high performance of

93.9% for arbitrary observation data of /u/ vowel.

IV-2. Comparison with Prior Works

IV-2-(1). Key 1D Features of Acoustic Parameters

Sondhi et al. [15] and Pah et al. [41] found that classifying subjects with and without COVID-19 was possible using jitter, shimmer, and HNR indices. However, they did not demonstrate classification using models incorporating these parameters. Pah et al. [41] stated, “The statistical analysis and SVM classification indicated that the voice features of sustained phoneme corresponding to vocal tract modulation (Mel Frequency Cepstral Coefficient (MFCC), Formants, Vocal-tract Length, and Intensity-SD) could potentially be adopted as a COVID-19 biomarker compared to the features of vocal fold vibration (jitter, shimmer, pitch, HNR, and NHR)” [41]. This suggests that a simple model using only jitter, shimmer, and HNR cannot differentiate between participants with and without disease. Therefore, machine learning and deep learning approaches are essential for performing pathophysiological analyses such as classification, presence or absence determination, and monitoring using key 1D features of acoustic parameters, such as MFCC and formants (Paper I). Several scholars have already performed such approaches with significant results [41, 42]. Omiya et al. [42] developed a voice marker using a machine-learning algorithm model with LightGBM and five-fold cross-validation in one instance. This resulted in an accuracy rate of over 88% and AUC in

distinguishing between “asymptomatic or mild (symptom)” and “moderate I (symptom)” by using three types of sustained vowels - /a/, /e/, and /u/. The speech recordings generated 13,998 types of 1-dimensional features. Nonetheless, only the features selected to null importance were those related to MFCC (Mel-frequency Cepstral-Coefficients), auditory spectrum, and magnitude spectrum, which included five types for /a/, 8 for /e/, and 16 for /u/. Thorough statistical analysis for feature selection was made to avoid overfitting the model result [42].

IV-2-(2) Two-Dimensional Feature Matching of DTW Algorithms

Conversely, the DTW score of 2D-feature matching was a simple and significant way compared with the ML approach using key 1-dimensional acoustic parameters. In particular, the DTW variance indices significantly differed from the average indices (Paper I). The difference between the two categories discriminated by the DTW variance became more pronounced in the distribution map in Figure 20, where the parameters were converted to MD scores and showed the distribution trend of the two Normal Spaces. The reason for this is considered as follows. Moderate illness I group spreads over a wide range because we assumed the patients have various kinds of combinations of swelling location and airflow intensity. Conversely, the distribution pattern of the mild illness group is relatively condensed around the center of gravity because they have less swelling or inflammation, and maybe airflow intensity is considered stable. Our discovery of “DTW variance indices” was “a novelty point”

in this study to make a binary classification simple and significant. However, the advanced model using 2D-features has not yet been developed to classify coronavirus strains or to infer mechanisms such as swelling or lowered airflow using voice changes to date. In addressing this issue, machine learning–based voice analysis systems that focus on learning 1D parameters, such as spectral and prosodic speech features [41-48], may bring quick results. To the best of our knowledge, the paper I is the first published attempt to use DTW to classify a disease severity from the patient vowels.

IV-2-(3). Effectiveness of MD method

In not a few studies, including this study, where sample size cannot increase easily, noise data such as outliers, patient bias, and measurement errors would affect the accuracy or performance of the classification model. Since the MD score corresponds to the distance from the center of gravity of each group, by appropriately tuning the threshold of MDS, it is possible “objectively” to identify data belonging to the Normal Space (NS) or not. Taking this concept, the MD method is a very practical pre-processing method to filter out the noise data from the given real-world datasets, which most likely include some amount of noise [27]. By using this method of filtering out some outliers, we were able to judge the model's accuracy much better. Furthermore, by applying the mathematical codes (variance/covariance matrix and center of gravity), which were extracted from the SSNS (Silver

Standard Normal Space) for the model predictors, we had the model generalized to handle any arbitrary datasets.

IV-2-(4) Meta-Information for the MDS with Wrong Label and Implausible Observations

Data with MDS greater than the MDt were considered implausible observations and set apart from standard classification. The supplementary information attached to the excluded MDS revealed the details of the patient's conditions, which might cause implausible observation data. For example, (1) a one-day moderate I label suddenly appeared in the middle of a week-long mild condition, (2) a series of actual mild labels with fever were predicted moderate I, and (3) a mild label appeared on the following day of a series of moderate I was predicted moderate I, among others. The data suggest that about half of the cases included patient bias due to subjective feelings toward symptoms and a disposition to endure. This study observed 11 cases of wrong label predictions and two instances of outliers with unknown origin, which could be attributed to patient bias. If we accept a 7.7% (13 cases out of 172) inclusion rate for implausible observation, the initial NS level for eliminating implausible observations from arbitrary observation datasets should be reasonably set at approximately 90–95%. However, if the sample size is small, we may need to squeeze the NS level down to 75%, as we experienced in this study, where 12 wrong labeling of 172 (7.0%) were observed. It was considered a

model limitation because no relation to implausible observation was found in the patient's incidental status. 93% of AUC seems to be near the upper limit of the model performance of this method. Suppose the MDt was set higher in the classification of arbitrary observation datasets. In that case, some implausible observations might remain in the MDS predictors that generate the GLM, resulting in poor model performance or classification accuracy. Conversely, if the MDt were set to minor, the number of objects that could be classified would be minimal, resulting in poor model versatility for classification with arbitrary observations. Therefore, finding the appropriate MDt for a given arbitrary observation dataset is essential in this two-story classification.

IV-2-(5). Advantages of the Standardization of Waveform Samples

By standardizing the time and power axes of the waveform samples in the DTW algorithm before computing, the fundamental frequency (F0) and volume were consequently transformed as parts of the elements forming a 10-cycle waveform in the unit envelope. Confounding factors derived from fundamental frequencies that vary by sex and age can be avoided as much as possible in advance. Thus, allowing the DTW distance's direct evaluation of the classification results [13,42,49,50]. This study examined 110 wavelet samples to determine the coefficient of variation (CV) of F0 estimates before and after standardization. We aimed to analyze the variety of F0 distribution based on sex and age groups. The results indicated that the CV of the wavelets without standardization ranged from

10.31% to 11.4% for the 3 long vowels /a/, /e/, and /u/, which is considered a significant confounding factor. However, after standardization, the CV ranged from 0.81% to 2.4%, significantly reducing and effectively minimizing confounding factors. Appendix 3 provides more details and comparison figures (Paper I).

IV-2-(6) Why Select a 10-cycle Waveform as a Unit Sample?

Our previous pilot study using the DTW algorithm to differentiate between participants' voices with and without COVID-19 was investigated using different waveforms with 1, 3, 5, 10, 20, 30, and 50 cycles. Among these waveform cycles, 10 and 20 cycles resulted in reasonable discrimination between the subjects with and without COVID-19 for all 3 participants tested. (A summary of the test results is shown in Appendix 4.) We selected a 10-cycle waveform rather than a 20-cycle waveform to reduce the total computing cost incurred when using the DTW algorithm as much as possible (see the section "Computing Cost") (Paper I).

IV-2-(7) Robustness to Noise

While recent smartphones have improved recording features that can aid in voice analysis [51], it should be noted that the environmental noises present during the recording process were not completely controllable because the voice samples used in this study were self-recorded by the

participants with their smartphones. As Qi [52] reported, “DTW was evaluated using both synthetic and natural voices, and significant reductions in noise were achieved.” Given that the DTW distance is considered to be robust to noise, it may be more practical when used for classification purposes with voices obtained from the real world compared to acoustic parameters such as jitter, shimmer, and HNR, which are considered to be sensitive to environmental noise [8,30] (Paper I).

IV-2-(8) Sample Size Consideration

There are known statistical correlations between significance level, power, effect size, and sample size [53]. We tested the validity of the sample size used for 110 participants (61 in the mild group and 49 in the moderate I group) in this study. We calculated an effect size of 0.666 with the significance level of 1% used in this study and a power of 0.8, which generally meets requirements from a statistical point of view [54]. The pwr package (version 1.3.0) of R was used for the calculation. A Cohen d score of 0.5 is regarded as a medium effect size, and 0.8 is considered a large effect size [55]. Therefore, we believe the sample size used in this study was appropriate because the effect size of 0.666 is between medium and large. (These validation processes are disclosed in Multimedia Appendix 5.) (Paper I)

IV-2-(9) Future Expectations

DTW distance–based voice biomarkers may effectively supplement pulse oximeters as an objective indicator when distinguishing moderate illness I from mild illness among patients during recuperation. Even if pulse oximeters are scarce, this biomarker can be accessed through a patient's smartphone. Suppose persons with disease recuperating at home can detect a worsening of symptoms to moderate illness I based on changes in their voice. In that case, they can determine whether they should seek medical care. Additionally, this system may allow health care providers to use voice biomarkers in addition to body temperature and pulse oximeter readings as objective and quantitative indicators to diagnose worsening symptoms and expedite inpatient treatment appropriately. Fischer et al. [56] said that "implementing vocal biomarkers in a digital solution for remote patient monitoring of frequently reported symptoms of COVID-19 is highly interesting." Utilizing voice biomarkers via smartphones is a convenient method of remotely gathering diagnostic information, which can help to reduce patient bias towards physicians. This approach is non-invasive and less time-consuming than responding to a daily questionnaire, making it less burdensome for patients (Paper I). In addition, it is highly expected that voice biomarkers will be implemented in some future medical settings, such as online medical treatment. If advanced models can go beyond the current skill of a binary classification on the disease severity and develop the skill of several pathology estimations, such as swelling condition, airflow intensity or phlegm status and so on, the voice biomarkers would bring clinical significance and value

as digital health devices. Although cost estimation is not possible yet today to develop such advanced skills for voice biomarkers, it is necessary to continuously verify various correlations with images and clinical data capturing upper respiratory tract and lung conditions, as well as with large amounts of voice recording data.

Suppose this algorithm were to be implemented in society as a smart phone application. It is first necessary to prepare a demonstration model of a system that processes the recorded voice algorithm in the cloud and simultaneously delivers the results to the patient and the doctor in charge. Such a PoC (proof-of-concept) approach would require the cooperation of demonstration system facilities owned by Kanagawa Prefecture research institutes or private entities. In this PoC, the communication speed between the smartphone and the cloud and the quality of voice recordings would also need to be validated. Moreover, although the algorithm will initially be built based on SSNSs (Silver Standard Normal Spaces) related to the regions, including overseas countries and time of period, if there is room to build GSNSs (Gold Standard Normal Spaces) from multiple SSNSs, the algorithm can be significantly simplified, and the computational cost can be reduced.

Since the /a/, /e/, and /u/ vowels employed in this study are found in many languages spoken worldwide,

the research results can be deployed usefully across international and regional boundaries of human races.

IV-3. Limitations

IV-3-(1) Computing Cost

In this study, our approach involved standardizing waveform samples, comprehensively computing the DTW distance for each sample, and subsequently using the resulting indices to determine the severity of illness using the GLM. However, this approach is computationally expensive, making it unsuitable for integration into standalone smartphone applications. Nevertheless, it can be used as a cloud application on a network. Despite some limitations, significant advancements in network transmission speed and information technology suggest that it may soon be practically applicable. Another potential solution for reducing computation time is to make the gold normal spaces of DTW distance-based indices fit for data diversity by reading more samples.

IV-3-(2) Patient Bias

Although the definition of moderate illness I used in this study is based on pulse oximeter readings (SpO2 93–96%) or subjective reports from patients of their clinical condition (i.e., SoB, as shown in Table 1), errors during labeling of the voice samples due to patient bias may have occurred as different individuals may have varying methods of describing the sensation of SoB. For example, the patient

with ID.01M reported SoB on day 5 only during 13 days of recuperation. During the period, no SpO2 lowering (< 96) was reported, but some fever (> 37.5) was reported from days 4 to 8. The prediction failed on days 3, 6, 8, 12, and 13, which were predicted as moderate illness I to the actual label of mild illness. Figure 22 shows that five points of 01M-series MDSs are far from the NS of mild groups and inside the NS of the moderate I group. Due to these factors, the patient's condition may have worsened to moderate I until day 8, when the fever was observed. Further, the days 12 and 13 may be due to aftereffects. Unfortunately, this issue is challenging to overcome, considering the use of patient-reported data. However, including objective indicators such as voice biomarkers during the diagnosis-making process may allow for more objective labeling of data in the future (Paper I).

IV-3-(3) Current Model Limitation

We refer to the patient ID# 411F as an example illustrating the model's limitations. During its recuperation, patient ID#411F reported shortness of breath all six days. SpO2 reading lowered (<96) on days 1 and 6, with a fever (>37.5) observed on day 1. Considering all the information, there is no doubt that patient ID#411F was in moderate illness status from day 1 to 6. However, the model predicted three data as “Mild.” Figure 23 displays all of them within Mild-NS. Our best assumption of why such prediction errors happen is (1) model training data for moderate illness I cases were not good enough to make the most coverage of moderate illness I conditions, and (2) it was not done to

effectively reduce intersectional area between two NSs with the limited predictors of MDS based on DTW variance. This feasibility study was also a challenge to see how far a single DTW index actuated practically as a single predictor in a GLM model. By adding other parameters (e.g., body temperature) as some other predictors for the current GLM model, it may be improved in future trials. As of today, in this study, we recognize the current model limitation.

IV-3-(4) Correlation with Pulse Oximeters

The number of moderate I illnesses judged by the pulse oximeter was just 6 of 291 participants (2%), which is not a large enough sample size to correlate the pulse oximeter and voice biomarker significantly. Therefore, a new opportunity for voice recording with simultaneous pulse oximeter readings is needed.

V. Conclusions

Medical treatments for COVID-19 vary depending on the severity of the illness. Patients with mild illness may need only to recuperate at home or a designated facility, whereas patients with moderate illness I may need to be hospitalized. This study tested the DTW distance–based voice biomarker to distinguish between mild and moderate illness I. A balanced accuracy ranging from 80.2% to 88% was achieved, and the model performance indicated by the AUC ranged from 86.5% to 96.5% for the

vowels /a/, /e/, and /u/ (Paper I). In addition, the improved model equipped with the MD method was applied to the arbitrary observation dataset of the /u/ vowel, which was a different dataset from 110 samples, for classification. The achievement was the model performance of 93.9% and balanced accuracy of 85.0%, considered reasonable for real-world datasets. This voice biomarker system is generalized for mild and moderate I illness severity classification. However, not all severity data were included in this study, so Moderate illness II and Serious illnesses are out of scope for the severity classification. In case of an unexpected shortage of pulse oximeters, it can be used as an alternative and cost-effective method for monitoring worsening medical conditions in patients with mild illness, recuperating at home or a medical facility. To the best of our knowledge, Paper I is the first published attempt to use DTW to classify a disease severity from the patient vowels.

VI. Acknowledgments

I want to express my gratitude to Mindy Fang for introducing me to the pleasures of statistical analysis programming in R. I am also grateful to Dr. Yasuhiro Omiya for instructing me on how to utilize Audacity and Praat acoustic software, the ROC/AUC model performance validation method, and the five-fold cross-validation method for model creation, all of which were integral to our study. He also gave me a comprehensive overview of DTW and inspired me when I struggled with my research outcomes. Dr. Shinichi Tokuno also deserves my gratitude. Despite his busy schedule, he reviewed

and supervised my research, even at night, and provided me with helpful guidance and challenging assignments. Thanks to his support, I could structure my research, and I believe I have significantly improved. I cannot thank him enough. I am grateful for the unwavering support and understanding my wife, Mika Watase, has always provided me throughout my research endeavors.

VII. Conflicts of Interest

None declared.

VIII. Abbreviations

AUC: Area Under the Curve

Bal.Acc: Balanced Accuracy

DTW: Dynamic Time Warping

EPV10: Events Per Variable 10

GLM: Generalized Linear Model

HNR: Harmonic-to-Noise Ration

LDA: Linear Discriminant Analysis

MD: Mahalanobis Distance

MDS: Mahalanobis Distance Score

MDt: Mahalanobis Distance Thresholding

MILD: Mild illness group

MOD1: Moderate illness I group

MiF: mild-group Filtering

MoF: moderate group Filtering

MFCC: Mel-Frequency Cepstral Coefficient

NSs: Normal Spaces

ROC: Receiving Operator Curve

SpO2: Saturation of peripheral Oxygen – the saturation of oxygen in the peripheral blood

SoB: Shortness of Breath

SSNS: Silver Standard of Normal Space

TPR: True Positive Rate

TNR: True Negative Rate

References

1. Ministry of Health, L. a. W. (October 5, 2022). *Japan A Guide to Medical Care of COVID-19, Version 8.1 (Japanese)*. https://www.mhlw.go.jp/content/000936655.pdf .

2. Ding, H., Mandapati, A., Karjadi, C., Ang, T. F. A., Lu, S., Miao, X., . . . Lin, H. (2022). Association Between Acoustic Features and Neuropsychological Test Performance in the Framingham Heart Study: Observational Study. *J Med Internet Res*, *24*(12), e42886. https://doi.org/10.2196/42886

3. Hajjar, I., Okafor, M., Choi, J. D., Moore, E., Abrol, A., Calhoun, V. D., & Goldstein, F.

C. (2023). Development of digital voice biomarkers and associations with cognition, cerebrospinal biomarkers, and neural representation in early Alzheimer's disease. *Alzheimers Dement (Amst)*, *15*(1), e12393. https://doi.org/10.1002/dad2.12393

4. Ma, A., Lau, K. K., & Thyagarajan, D. (2020). Voice changes in Parkinson's disease: What are they telling us? *J Clin Neurosci*, *72*, 1-7. https://doi.org/10.1016/j.jocn.2019.12.029

5. Chintalapudi, N., Battineni, G., Hossain, M. A., & Amenta, F. (2022). Cascaded Deep Learning Frameworks in Contribution to the Detection of Parkinson's Disease. *Bioengineering (Basel)*, *9*(3). https://doi.org/10.3390/bioengineering9030116

6. Costantini, G., Cesarini, V., Di Leo, P., Amato, F., Suppa, A., Asci, F., . . . Saggio, G. (2023). Artificial Intelligence-Based Voice Assessment of Patients with Parkinson's Disease Off and On Treatment: Machine vs. Deep-Learning Comparison. *Sensors (Basel)*, *23*(4). https://doi.org/10.3390/s23042293

7. Tracy, J. M., Özkanca, Y., Atkins, D. C., & Hosseini Ghomi, R. (2020). Investigating voice as a biomarker: Deep phenotyping methods for early detection of Parkinson's disease. *J Biomed Inform*, *104*, 103362. https://doi.org/10.1016/j.jbi.2019.103362

8. Tougui, I., Jilbab, A., & Mhamdi, J. E. (2020). Analysis of Smartphone Recordings in Time, Frequency, and Cepstral Domains to Classify Parkinson's Disease. *Healthc Inform Res*, *26*(4), 274-283. https://doi.org/10.4258/hir.2020.26.4.274

9. Suppa, A., Costantini, G., Asci, F., Di Leo, P., Al-Wardat, M. S., Di Lazzaro, G., . . . Saggio, G. (2022). Voice in Parkinson's Disease: A Machine Learning Study. *Front Neurol*, *13*, 831428. https://doi.org/10.3389/fneur.2022.831428

10. Higuchi, M., Nakamura, M., Shinohara, S., Omiya, Y., Takano, T., Mitsuyoshi, S., & Tokuno, S. (2020). Effectiveness of a Voice-Based Mental Health Evaluation System for Mobile Devices: Prospective Study. *JMIR Form Res*, *4*(7), e16455. https://doi.org/10.2196/16455

11. Shinohara, S., Nakamura, M., Omiya, Y., Higuchi, M., Hagiwara, N., Mitsuyoshi, S., . . . Tokuno, S. (2021). Depressive Mood Assessment Method Based on Emotion Level Derived from Voice: Comparison of Voice Features of Individuals with Major Depressive Disorders and Healthy Controls. *Int J Environ Res Public Health*, *18*(10). https://doi.org/10.3390/ijerph18105435

12. Kappen, M., van der Donckt, J., Vanhollebeke, G., Allaert, J., Degraeve, V., Madhu, N., . . . Vanderhasselt, M. A. (2022). Acoustic speech features in social comparison: how stress impacts the way you sound. *Sci Rep*, *12*(1), 22022. https://doi.org/10.1038/s41598-022-26375-9

13. Asiaee, M., Vahedian-Azimi, A., Atashi, S. S., Keramatfar, A., & Nourbakhsh, M. (2022). Voice Quality Evaluation in Patients With COVID-19: An Acoustic Analysis. *J Voice*, *36*(6), 879.e813-879.e819. https://doi.org/10.1016/j.jvoice.2020.09.024

14. Kaur, S., Larsen, E., Harper, J., Purandare, B., Uluer, A., Hasdianda, M. A., . . . Jariwala, S. (2023). Development and validation of a respiratory-responsive vocal biomarker-based tool for generalizable detection of respiratory impairment: Independent case-control studies in multiple respiratory conditions including asthma, chronic obstructive pulmonary disease, and COVID-19. *J Med Internet Res*, *25*, e44410. https://doi.org/10.2196/44410

15. Sondhi, S., Salhan, A., Santoso, C. A., Doucoure, M., Dharmawan, D. M., Sureka, A., . . . Hedwig, R. (2021). Voice processing for COVID-19 scanning and prognostic indicator. *Heliyon*, *7*(10), e08134. https://doi.org/10.1016/j.heliyon.2021.e08134

16. Tohidast, S. A., Mansuri, B., Memarian, M., Ghobakhloo, A. H., & Scherer, R. C. (2021). Voice quality and vocal tract discomfort symptoms in patients With COVID-19. *J Voice*. https://doi.org/10.1016/j.jvoice.2021.09.039

17. Erdoğan, Y. E., & Narin, A. (2021). COVID-19 detection with traditional and deep features on cough acoustic signals. *Comput Biol Med*, *136*, 104765. https://doi.org/10.1016/j.compbiomed.2021.104765

18. Adhikary, S., & Ghosh, A. (2022). Dynamic time warping approach for optimized locomotor impairment detection using biomedical signal processing [Article]. *Biomedical Signal Processing and Control*, *72*, 12, Article 103321. https://doi.org/10.1016/j.bspc.2021.103321

19. Dolatabadi, E., Taati, B., & Mihailidis, A. (2016). Automated classification of pathological gait after stroke using ubiquitous sensing technology. *Annu Int Conf IEEE Eng Med Biol Soc*, *2016*, 6150-6153. https://doi.org/10.1109/EMBC.2016.7592132

20. Li, M., Tian, S., Sun, L., & Chen, X. (2019). Gait Analysis for Post-Stroke Hemiparetic Patient by Multi-Features Fusion Method. *Sensors*, *19*(7), 1737. https://doi.org/10.3390/s19071737

21. Renggli, D., Graf, C., Tachatos, N., Singh, N., Meboldt, M., Taylor, W. R., . . . Schmid Daners, M. (2020). Wearable inertial measurement units for assessing gait in real-world environments. *Front Physiol*, *11*, 90. https://doi.org/10.3389/fphys.2020.00090

22. Roth, N., Küderle, A., Ullrich, M., Gladow, T., Marxreiter, F., Klucken, J., . . . Kluge, F. (2021). Hidden Markov Model based stride segmentation on unsupervised free-living gait data in Parkinson's disease patients. *J Neuroeng Rehabil*, *18*(1), 93. https://doi.org/10.1186/s12984-021-00883-7

23. Venail, F., Legris, E., Vaerenberg, B., Puel, J. L., Govaerts, P. J., & Ceccato, J. C. (2016). Validation of the French-language version of the OTOSPEECH automated scoring software package for speech audiometry. *Eur Ann Otorhinolaryngol Head Neck Dis*, *133*(2), 101-106. https://doi.org/10.1016/j.anorl.2016.01.001

24. Barry, S. J., Dane, A. D., Morice, A. H., & Walmsley, A. D. (2006). The automatic recognition and counting of cough. *Cough*, *2*, 8. https://doi.org/10.1186/1745-9974-2-8

25. Muda, L., Begam, M., & Elamvazuthi, I. (2010). Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques.

26. Ng, G., & Andrysek, J. (2023). Classifying Changes in Amputee Gait following Physiotherapy Using Machine Learning and Continuous Inertial Sensor Signals. *Sensors (Basel)*, *23*(3). https://doi.org/10.3390/s23031412

27. Pang, J., Liu, D., Peng, Y., & Peng, X. (2023). Temporal dependence Mahalanobis distance for anomaly detection in multivariate spacecraft telemetry series. *ISA Trans*. https://doi.org/10.1016/j.isatra.2023.06.002

28. Guerrero-Gonzalez, J. M., Yeske, B., Kirk, G. R., Bell, M. J., Ferrazzano, P. A., & Alexander, A. L. (2022). Mahalanobis distance tractometry (MaD-Tract) - a framework for personalized white matter anomaly detection applied to TBI. *Neuroimage*, *260*, 119475. https://doi.org/10.1016/j.neuroimage.2022.119475

29. Ibáñez, D., Garcia, E., Soret, J., & Martos, J. (2023). Incipient wear detection of welding gun secondary circuit by virtual resistance sensor using Mahalanobis Distance. *Sensors (Basel)*, *23*(2). https://doi.org/10.3390/s23020894

30. Janeliukstis, R., & Mironovs, D. (2023). Wavelet-based output-only damage detection of composite structures. *Sensors (Basel)*, *23*(13). https://doi.org/10.3390/s23136121

31. Vode, F., Tehovnik, F., Kosec, G., & Steiner Petrovič, D. (2022). Classification of hot-rolled plates using the Mahalanobis Distance of NMIs in ti-stabilized austenitic stainless-steel produced by secondary metallurgy. *Materials (Basel)*, *15*(2). https://doi.org/10.3390/ma15020684

32. Rabby, Y. W., Li, Y., & Hilafu, H. (2023). An objective absence data sampling method for landslide susceptibility mapping. *Sci Rep*, *13*(1), 1740. https://doi.org/10.1038/s41598-023-28991-5

33. Jung, Y., Longo, C., & Tompa, E. (2022). Longitudinal assessment of labor market earnings among patients diagnosed with cancer in Canada. *JAMA Netw Open*, *5*(12), e2245717. https://doi.org/10.1001/jamanetworkopen.2022.45717

34. Ramlie, F., Muhamad, W. Z. A. W., Harudin, N., Abu, M. Y., Yahaya, H., Jamaludin, K.

R., & Abdul Talib, H. H. (2021). Classification performance of thresholding methods in the Mahalanobis–Taguchi System. *Applied Sciences*, *11*(9), 3906. https://doi.org/10.3390/app11093906

35. Huang, C.L., Hsu, T.S., Liu, C.M. (2010). Modeling a dynamic design system using the Mahalanobis Taguchi system—Two steps optimal based neural network. *J. Stat. Manag. Syst,* 13, 675–688. https://doi.org/10.1080/09720510.2010.10701495

36. Huang, C.L., Hsu, T.S., Liu, C.M. (2009). The Mahalanobis–Taguchi system—Neural network algorithm for data-mining in dynamic environments. Expert Syst. *Appl*, 36, 5475–5480. https://doi.org/10.1016/j.eswa.2008.06.120

37. El-Banna, M. (2017). Modified Mahalanobis Taguchi system for imbalance data classification. *Comput Intell Neurosci*, *2017*, 5874896. https://doi.org/10.1155/2017/5874896

38. Kumar, S.; Chow, T.W.S.; Pecht, M. Approach to fault identification for electronic products using Mahalanobis Distance. *IEEE Trans. Instrum. Meas.* 2010, 59, 2055–2064. https://doi.org/10.1109/TIM.2009.2032884

39. Ribeiro Xavier, C., Sachetto Oliveira, R., da Fonseca Vieira, V., Lobosco, M., & Weber Dos Santos, R. (2022). Characterisation of Omicron Variant during COVID-19 Pandemic and the Impact of Vaccination, Transmission Rate, Mortality, and Reinfection in South Africa, Germany, and Brazil. *BioTech (Basel)*, *11*(2). https://doi.org/10.3390/biotech11020012

40. Skarbinski, J., Wood, M. S., Chervo, T. C., Schapiro, J. M., Elkin, E. P., Valice, E., . . . Kushi, L. H. (2022). Risk of severe clinical outcomes among persons with SARS-CoV-2 infection with differing levels of vaccination during widespread Omicron (B.1.1.529) and Delta (B.1.617.2) variant circulation in Northern California: A retrospective cohort study. *Lancet Reg Health Am*, *12*, 100297. https://doi.org/10.1016/j.lana.2022.100297

41. Pah, N. D., Indrawati, V., & Kumar, D. K. (2022). Voice features of sustained phoneme as COVID-19 Biomarker. *IEEE Journal of Translational Engineering in Health and*

Medicine, *10*, 1-9. https://doi.org/10.1109/jtehm.2022.3208057

42. Omiya, Y., Mizuguchi, D., & Tokuno, S. (2023). Distinguish the severity of illness associated with novel coronavirus (COVID-19) infection via sustained vowel speech features. *Int J Environ Res Public Health*, *20*(4). https://doi.org/10.3390/ijerph20043415

43. Hu, H.-C., Chang, S.-Y., Wang, C.-H., Li, K.-J., Cho, H.-Y., Chen, Y.-T., . . . Lee, O. K.-S. (2021). Deep learning application for vocal fold disease prediction through voice recognition: Preliminary Development Study. *Journal of Medical Internet Research*, *23*(6), e25247. https://doi.org/10.2196/25247

44. Altshuler E, Tannir B, Jolicoeur G, et al. Digital cough monitoring - A potential predictive acoustic biomarker of clinical outcomes in hospitalized COVID-19 patients. J Biomed Inform. Feb 2023,138:104283. https://doi.org/10.1016/j.jbi.2023.104283

45. Costantini G, Dr VC, Robotti C, et al. Deep learning and machine learning-based voice analysis for the detection of COVID-19: A proposal and comparison of architectures. Knowl Based Syst. Oct11, 2022, 253: 109539. https://doi.org/10.1016/j.knosys.2022.109539

46. Khalilzad Z, Hasasneh A, Tadj C. Newborn Cry-Based Diagnostic System to Distinguish between Sepsis and Respiratory Distress Syndrome Using Combined Acoustic Features. Diagnostics (Basel). Nov 15, 2022, 12 (11) https://doi.org/10.3390/diagnostics12112802

47. Verde, L., De Pietro, G., Ghoneim, A., Alrashoud, M., Al-Mutib, K. N., & Sannino, G. (2021). Exploring the use of artificial intelligence techniques to detect the presence of coronavirus Covid-19 through speech and voice analysis. *IEEE Access*, *9*, 65750-65757. https://doi.org/10.1109/access.2021.3075571

48. Liu Z, Xu Y. Deep learning assessment of syllable affiliation of intervocalic consonants. J Acoust Soc Am. Feb 2023;153(2):848 https://doi.org/10.1121/10.0017117

49. Abitbol J, A. P., Abitbol B. (1999). Sex hormones and the female voice. *J Voice.*, *Sep 13(3)*. https://doi.org/doi: 10.1016/s0892-1997(99)80048-4. PMID: 10498059.

50. Berti, L. C., Spazzapan, E. A., Queiroz, M., Pereira, P. L., Fernandes-Svartman, F. R., Medeiros, B. R., . . . Finger, M. (2023). Fundamental frequency related parameters in Brazilians with COVID-19. *J Acoust Soc Am*, *153*(1), 576. https://doi.org/10.1121/10.0016848

51. Uloza, V., Ulozaitė-Stanienė, N., Petrauskas, T., & Kregždytė, R. (2021). Accuracy of acoustic voice quality index captured with a smartphone - measurements with added ambient Noise. *J Voice*. https://doi.org/10.1016/j.jvoice.2021.01.025

52. Qi, Y. (1992). Time normalization in voice analysis. *J Acoust Soc Am*, *92*(5), 2569-2576. https://doi.org/10.1121/1.404429

53. Serdar, C. C., Cihan, M., Yücel, D., & Serdar, M. A. (2021). Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. *Biochemia medica*, *31*(1), 27-53. https://doi.org/10.11613/bm.2021.010502

54. Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: a review of three human research domains. *Royal Society Open Science*, *4*(2), 160254. https://doi.org/10.1098/rsos.160254

55. Calin-Jageman, R. J. (2017). The new statistics for neuroscience majors: Thinking in Effect Sizes. *Journal of Undergraduate Neuroscience Education*, *16*(2), E21. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6057753/

56. Fischer, A., Elbeji, A., Aguayo, G., Fagherazzi, G. (2022). Recommendations for successful implementation of the use of vocal biomarkers for remote monitoring of COVID-19 and Long COVID in Clinical Practice and Research. *Interact J Med Res*, 11(2). https://DOI: 10.2196/40655

Appendix-1: (1/1)

The Wrong Predictions as MOD1 against Actual MILD Labels, with a total of 14 Samples

data.id	Subj.ID	Number of data	Patient Condition from day1	Detail information	Possibility/Assumption
1	01M	5	Day1 to 9 were MILD, except only day 6 of MOD1	Intermediately continued coughing and phlegm, and soar throat. Feaver observed during day 4 to day 8.	Found data points within or near MOD1_NS. Maybe it is a patient with preserving or bias.
2	02F	1	Only one MILD data on day 1	Trend is unknown due to one data	Found a data point far from MOD1 and MILD. It is implausible due to the model algorithm
3	40M	1	Continued MILD from day 1 to the end	Day 1 and 2 indicated marginal SpO2 readings of 96 or 95	Found a data point within MOD1_NS. Maybe it is a patient with preserving or bias.
4	43F	1	Continued MILD from day 1 to day 5, only day 6 shows MOD1	Day 2 indicated feaver, which is this data, but label was reported as MILD	Found a data point near MOD1_NS. Maybe patient bias.
5	145F	1	Day 1 and 2 were MOD1. From day 3 to the end were MILD	Day 5, this data, showed MILD without any symptoms	Found a data point within MOD1_NS. Maybe it is a model limitation.
6	154F	1	Day 1 and 3 were MOD1. From day 4 to the end were MILD	Day 8, this data, showed MILD with coughing and phlegm	Found a data point on the edge of MOD1_NS. Maybe it is a model limitation.
7	182F	1	Continued MILD from day 1 to the end	Day 5, this data, showed MILD without any symptoms	Found a data point within MOD1_NS. Maybe it is a model limitation.
8	275F	2	Continued MILD from day 1 to the end	Day 1 and 2 indicated coughing and phlegm.	Found one data (275F330) is far from MILD_NS, which is implausible. Found the other (275F0325) within MOD1_NS, that may indicate patient bias
9	280M	1	Only one MILD data on day 1	Trend is unknown due to one data. Marginal SpO2 reading of 96.	Found a data point within MOD1_NS. Maybe it is patient bias

Appendix-2: (1/1)

The Wrong Predictions as MILD against Actual MOD1 Labels, with a total of 11 Samples

data.id	Subj.ID	Number of data	Patient Condition from day1	Detail information	Possibility/Assumption
1	67F	3	Continued MOD1 from day 1 to the end	Continued coughing, phlegm, and sore throat. Only day 1 predicted as MOD1	Found data points within MILD_NS. Maybe it is due to model limitation
2	97F	1	Only one MOD1 data on day 1	Trend is unknown due to one data	Found data points near MILD_NS. Maybe it is due to model limitation
3	154F	2	MOD1 from day 1 to 3, Condition moves to MILD from day 4	Coughing and phlegm were observed from day1 to the end	Found data points within MILD_NS. Maybe it is due to model limitation
4	220F	1	MOD1 was on day 1 only. From day 2, patient conditions showed MILD to the end.l	Coughing and phlegm were observed on day1 only	Found data points within MILD_NS. Maybe it is patient bias.
5	309F	1	Continued MOD1 from day 1 to the end. Prediction from day 2 were correct	Continued coughing, phlegm, and sore throat.	Found data points within MILD_NS. Maybe the model prediction was correct due to patient bias
6	411F	3	Continued MOD1 from day 1 to the end	Continued coughing, phlegm, and sore throat.	Found data points within MILD_NS. Maybe the model limitation

Appendix 3: Additional Analysis for F0 Confounding (1/3)

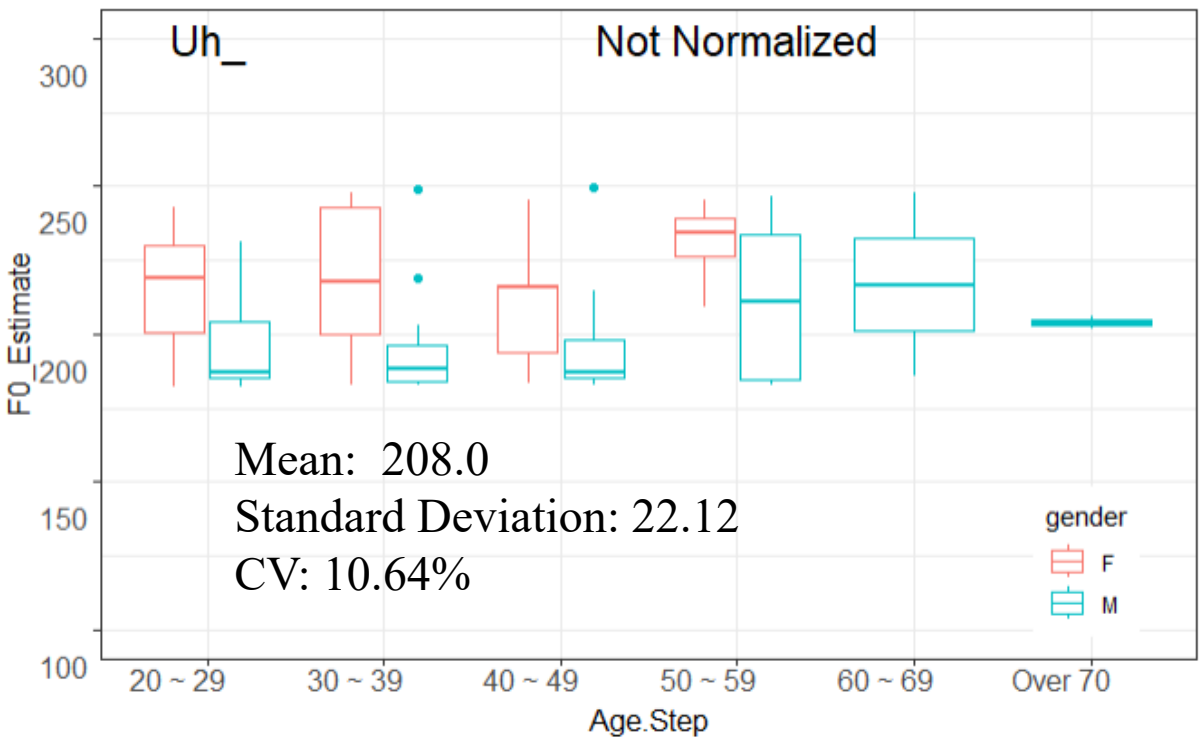
The fundamental frequency (F0) is known to vary with gender and age. However, when comparing wavelet features using the DTW algorithm, the difference in F0 is considered a confounding factor that makes disease severity misleading.

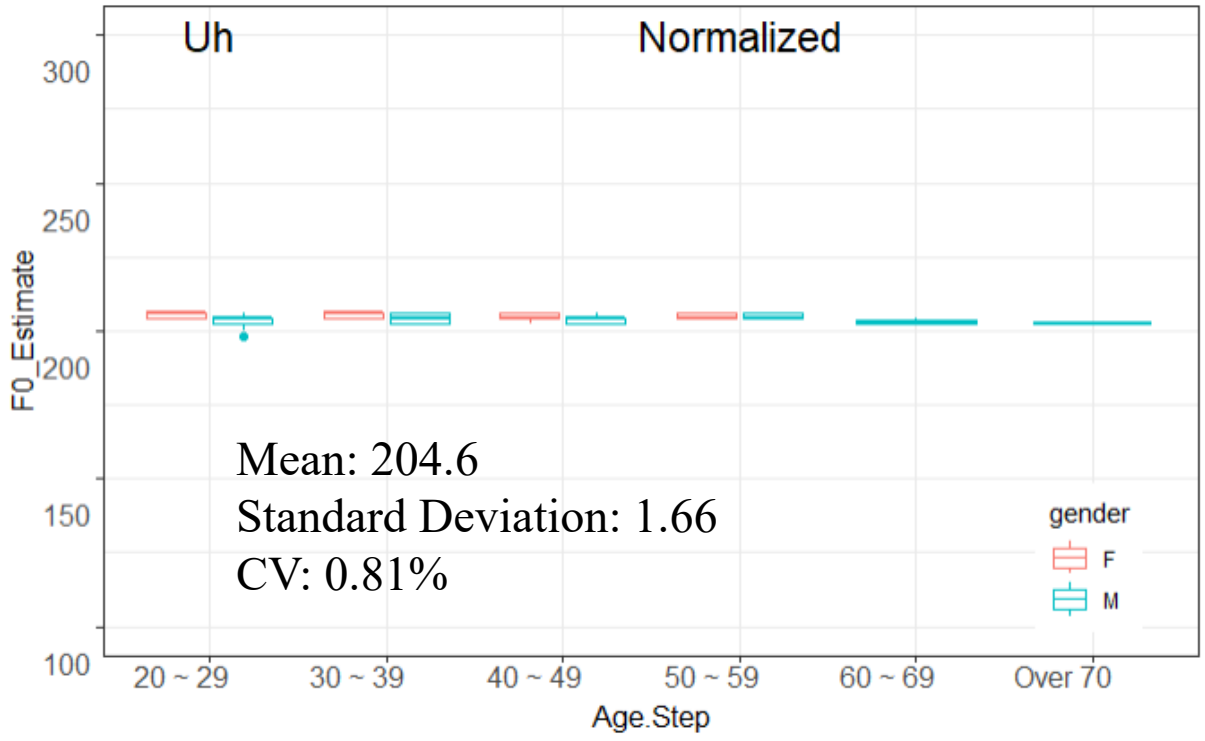
It is, therefore, necessary to standardize or normalize each wavelet to the time axis to remove a confounding factor before calculating the DTW distance.

The figure in the next slide indicates the variety of F0 distribution by sex and age groups. It also shows the effectiveness of the normalization on the wavelets that were used in this study.

Appendix 3: Additional Analysis for F0 Confounding (2/3)

110 Wavelets' F0 Distribution: Not Normalized (left) and Normalized (right)



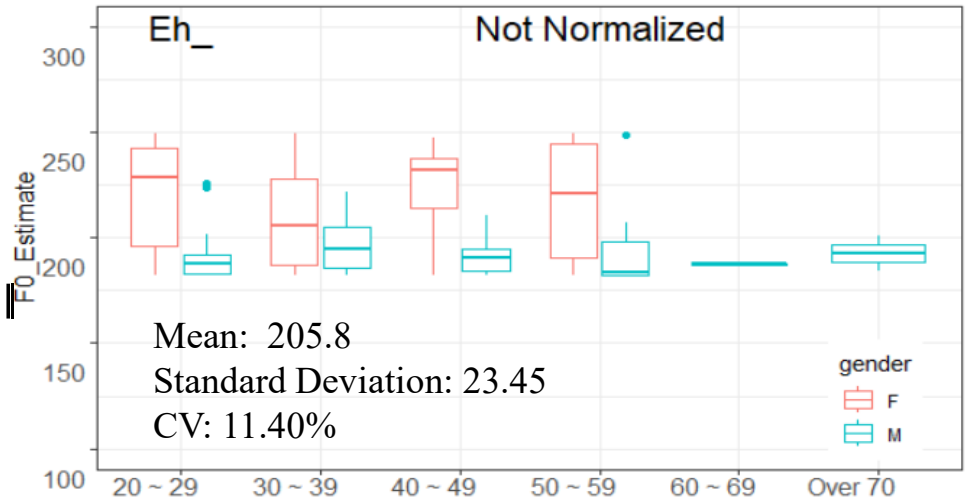


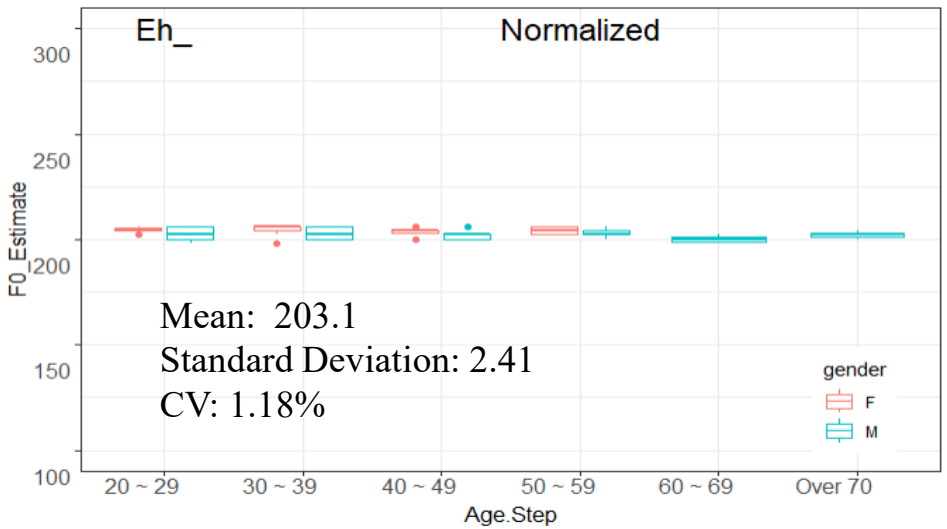
F0 was estimated using the Fast Fourier Transform function with additional sorting programming in R. 110 samples of 10-cycle wavelets were subjected to F0 estimation in either state, "not-normalized," or "normalize." The result of the Uh vowel is shown here with the Coefficient of Variance(CV), 10.64%(not normalized) to 0.81%(normalized). Ah and Eh are shown in the next slide.

Appendix 3: Additional Analysis for F0 Confounding (3/3)

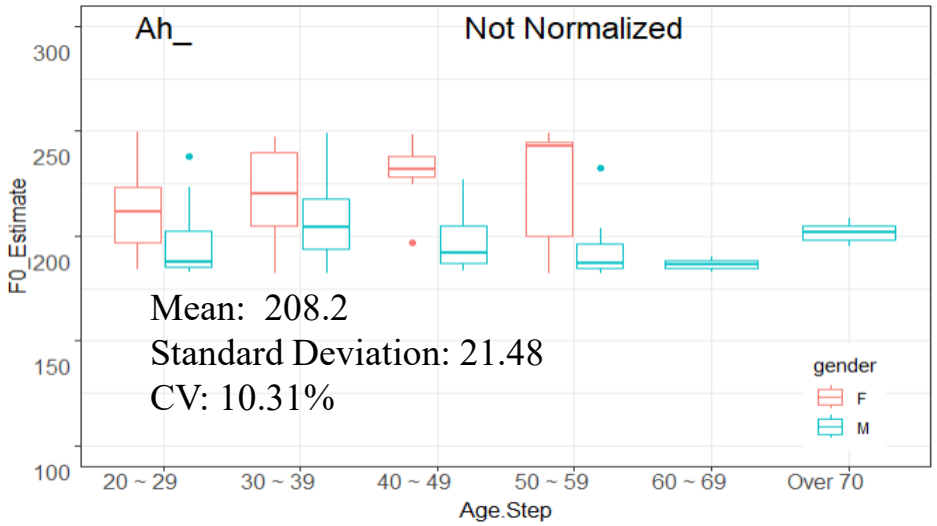
110 Wavelets' F0 Distribution: Not Normalized (left) and Normalized (right)

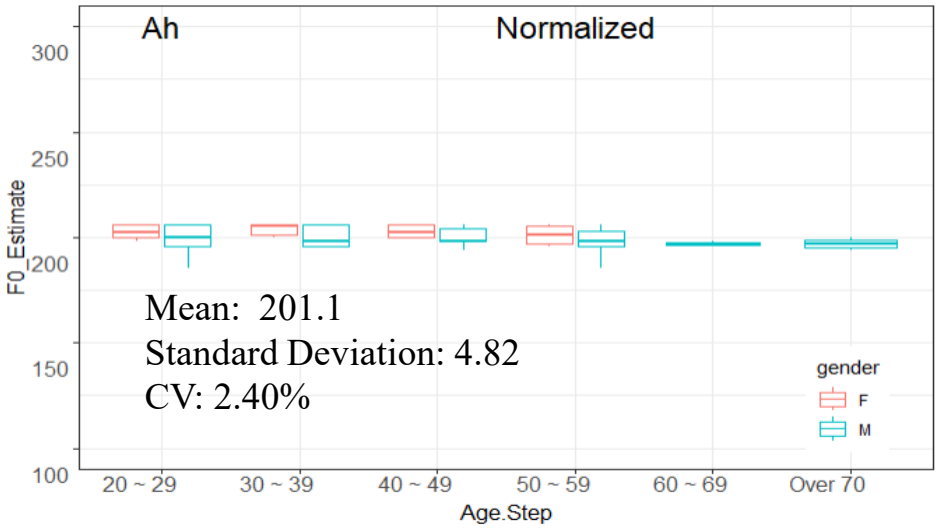
Eh =





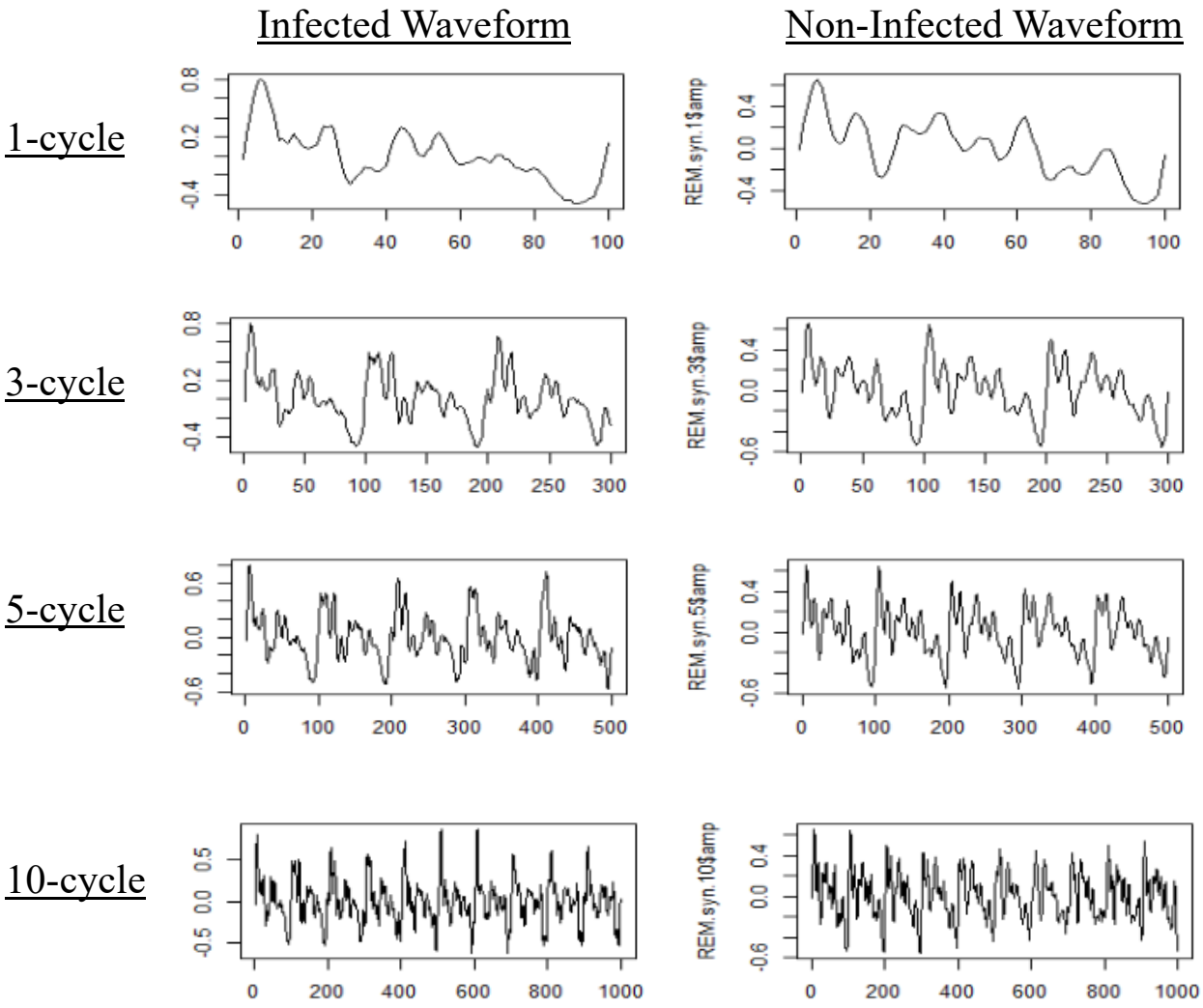
Ah =





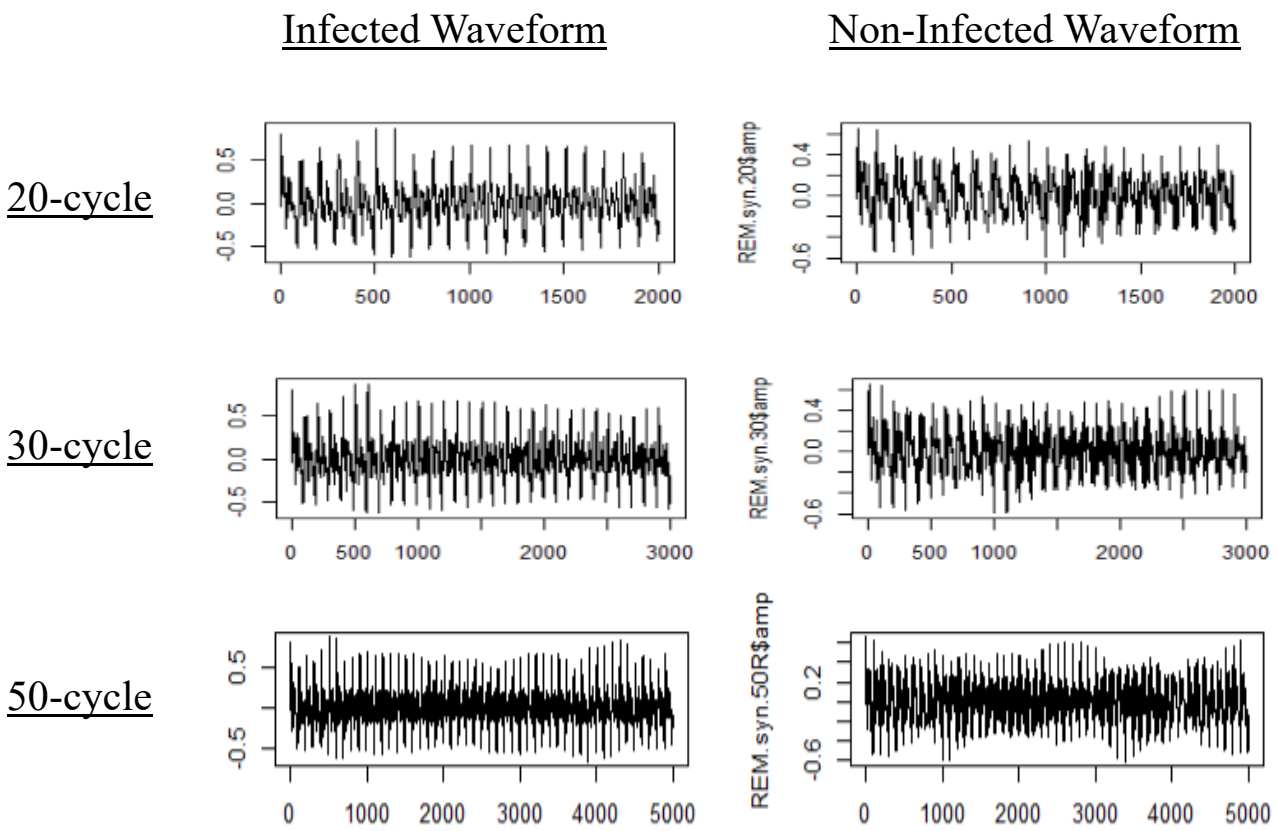
Appendix-4: Pilot Study for Adequate Wavelet Cycles (1/4)

Examples of 1-, 3-, 5, and 10-cycle waveforms from the same voice sample



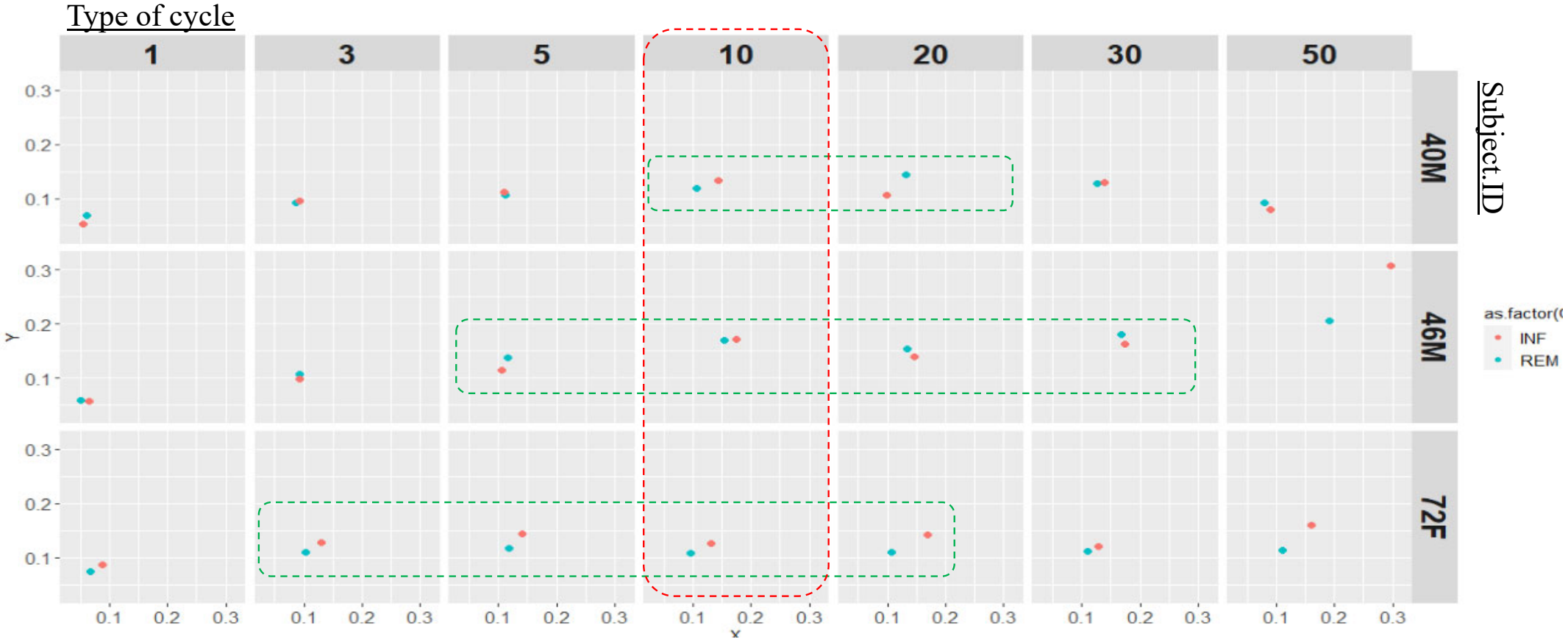
Appendix-4: Pilot Study for Adequate Wavelet Cycles (2/4)

Examples of 20-, 30-, 50-cycle waveforms from the same voice sample



Appendix-4: Pilot Study for Adequate Wavelet Cycles (3/4)

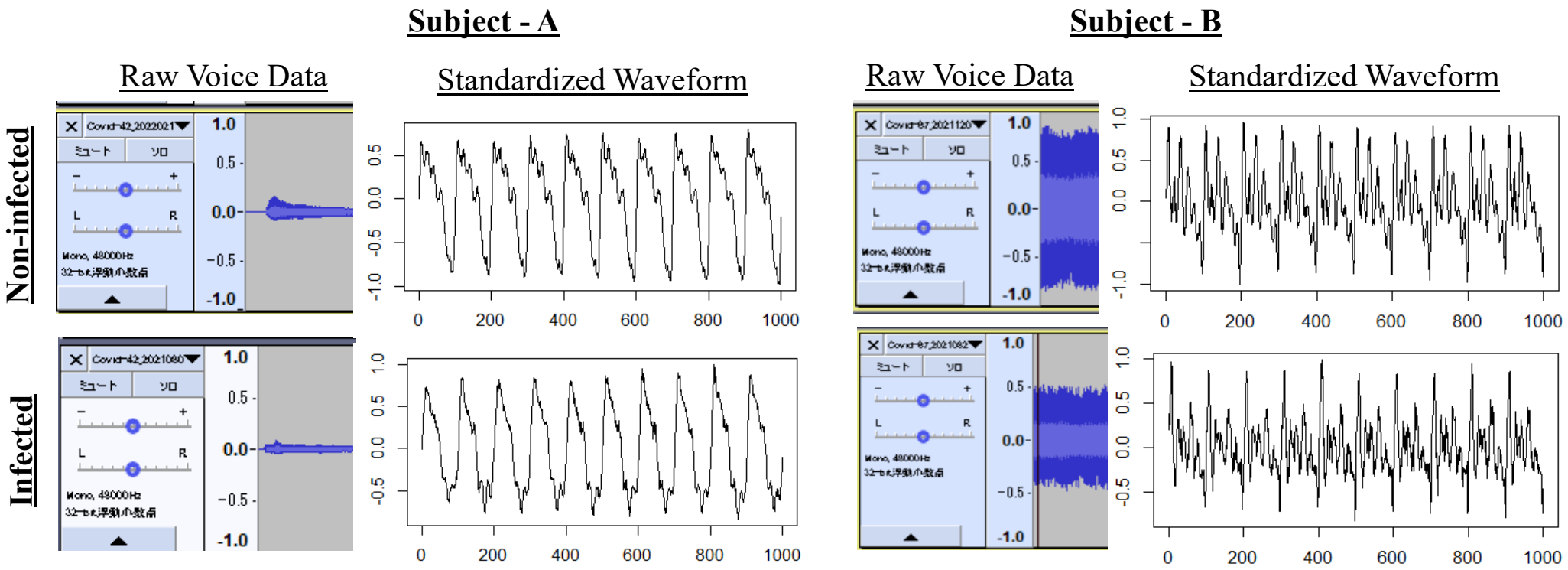
Example of Interpersonal Voice Classification using several-cycle waveforms



- The distance between the two points “infected” (INF, red dots) and “in remission” (REM, blue dots) seems to be adequate for discrimination in the range of 3 to 30 cycles in the three test cases (40M, 46M, 72F). Among them, waveforms with 10 or 20 cycles were commonly discriminable distances for all three subjects. In this study, the 10-cycle waveform was selected as the unit sample waveform rather than the 20-cycle due to computational cost.

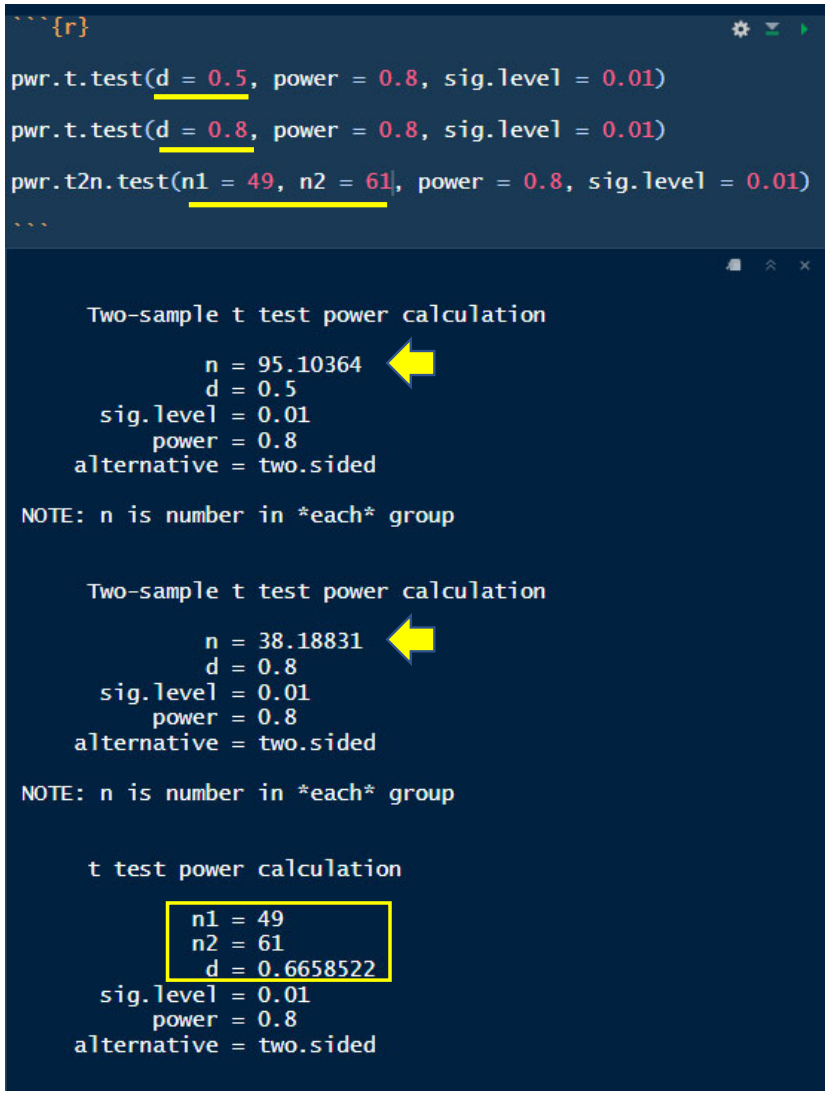
Appendix-4: Pilot Study for Adequate Wavelet Cycles (4/4)

Envelope Resulting from Standardization to Power and Time Axis



- All unit waveform samples should be fitted within the *envelope* resulting from standardization to power and time axis for better feature matching as using DTW algorithms.
- The Confounding factors of F0 and vocal volume among the participants were avoided within the envelope

Appendix-5: Sample Size Consideration (1/2)



```r
pwr.t.test(d = 0.5, power = 0.8, sig.level = 0.01)

pwr.t.test(d = 0.8, power = 0.8, sig.level = 0.01)

pwr.t2n.test(n1 = 49, n2 = 61, power = 0.8, sig.level = 0.01)
```

```
     Two-sample t test power calculation

              n = 95.10364
              d = 0.5
      sig.level = 0.01
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group


     Two-sample t test power calculation

              n = 38.18831
              d = 0.8
      sig.level = 0.01
          power = 0.8
    alternative = two.sided

NOTE: n is number in *each* group


     t test power calculation

             n1 = 49
             n2 = 61
              d = 0.6658522
      sig.level = 0.01
          power = 0.8
    alternative = two.sided
```

- At a significance level of 1% and a power of 0.8, each group sample size was calculated for the range of Cohen's effect size from 0.5 (medium effect) to 0.8 (large effect).
- The range was from 39 each (effect size: 0.8) to 96 each (effect size: 0.5). Each of our sample sizes (49 for MOD1 and 61 for MILD) is within this range.
- Then we calculated Cohen's effect size using different sample sizes of two groups and found it was about 0.666, which was considered effective to classify two groups as far as p-values are less than 1%.
- We concluded that the sample size we used for this feasibility study was validated.

Appendix-5: Sample Size Consideration (2/2)

p-value	**Ah**	**Eh**	**Uh**
MiF_average	**0.034**	**0.7115**	**7.66E-04**
MoF_average	**0.6029**	**0.4326**	**0.0316**
MiF_variance	**2.09E-05**	**8.28E-05**	**0.4468**
MoF_variance	**3.95E-05**	**5.81E-05**	**2.93E-06**

Cohen_d	**Ah**	**Eh**	**Uh**
MiF_average	**0.41**	**0.07**	**0.67**
MoF_average	**0.15**	**0.31**	**0.37**
MiF_variance	**0.74**	**0.78**	**0.14**
MoF_variance	**0.91**	**0.8**	**0.97**

- We calculated Cohen's effect size (bottom left) for each index using the P-values shown in Table 6 of the manuscript (top left)
- The effect sizes for the *variance* indices used for GLM classification were in the high range of 0.74 to 0.97. (except for 0.14 for the Uh-MiF-variance)