Extraction of vocal features for health assessment and early diagnosis - Effects of measurement uncertainty on classification algorithms

Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Metrology (34^{th}cycle)

Effects of measurement uncertainty on classification algorithms, as applied to vocal features for health assessment and early diagnosis

By

Alessio Atzori

******

Supervisor(s):
Prof. Alessio Carullo, Department of Electronics and Telecommunications, Politecnico di Torino

Doctoral Examination Committee:
Prof. Eric J. Hunter , Referee, Michigan State University
Prof. Sten Olov Ternstrom, Referee, Kungliga Tekniska Högskolan, Stoccolma
Prof. ....., University of...
Prof. ....., University of...
Prof. ....., University of...

Politecnico di Torino
2022

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Alessio Atzori
2022

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

I would like to dedicate this thesis to my loving parents

Abstract

The present manuscript deals with the analysis of vocal features applied to health assessment and early diagnosis of patients with the Parkinson's Disease and other pathological voices. In particular, the whole measuring chain has been characterized in order to identify the main uncertainty contributions of voice features and study their effect on classification algorithms. The main features analysed in this work are the pseudo-period and amplitude stability metrics, such as Jitter and shimmer, of sustained vowels recorded with a microphone in air. Additionally, the Cepstral Peak Prominence Smoothed, the Root Mean Square values and the Harmonics to Noise Ratio sequences, extracted from the sustained vowels, have been analysed in terms of descriptive statistics. To evaluate the uncertainty contributions of the stability metrics, an analytical evaluation has been carried out highlighting that the main uncertainty contributions for the period stability metrics are the time-base resolution of the Analog to Digital Converter, while for the amplitude stability metrics the Integral NonLinearity and the Gain Error were identified as important contributions along the amplitude resolution. The analytical evaluation of the features uncertainty showed to be very challenging for some stability metrics so a Monte Carlo uncertainty propagation has been carried out. The Monte Carlo propagation has been performed on the stability metrics to evaluate the effect of time and amplitude resolution on the bias and dispersion of the metrics under analysis, highlighting an important bias contribution on some stability metrics. Moreover a study on the effect of background noise has showed that the extraction algorithm represents the main uncertainty contribution for the evaluated stability metrics. In order to evaluate the uncertainty contribution of each component of the measuring chain, a vowel *re-synthesis* method has been proposed to produce artificial vowels with known pseudo-periods and amplitudes sequences. This method is based on the sampling of the original distributions of pseudo-periods and amplitudes sequences, which are used to produce reference sequences that are statistically comparable

to the original ones. Using this method a characterization of the measuring chain, composed by a microphone in air, a portable audio digital recorder and an extraction algorithm, has been carried out to evaluate the main uncertainty contributions to the voice features. This has been made possible thanks to the use of a Head and Torso Simulator, which replaces the original subject in the measuring chain and allows to perform repeatable and reproducible measurements. The analysis of the feature uncertainty highlighted that the vocal features are affected by a bias contribution and a dispersion contribution. The bias contribution can be different depending on the subject and on the length of the measuring chain, while the dispersion showed to be almost constant. This study highlighted that the extraction algorithm is the measuring chain component that mainly affects the evaluation of the voice features because the contribution of the whole measuring chain is comparable to the extraction contribution alone. The uncertainty evaluations performed in this work have been used to train binary weighted logistic regression models using a number of features in a range between 2 and 6. To train the models various strategies have been used to take advantage of the uncertainty evaluations of the measuring chain. In particular the effect of bias removal and the effect of evaluating the mixed terms of the uncertainty have been tested. The uncertainty analysis of the predicted probabilities was used to produce confidence intervals around the probabilities and thus the definition of a third class of non-classified. Thanks to this definition new classification metrics have been proposed to evaluate the classification performances and in particular the Realistic Accuracy has been defined excluding from the accuracy evaluations the non-classified subjects. These evaluations leaded to a maximum realistic accuracy of 100 % for the training and 77 % for the validation of the classification of Parkinson subject with respect to an healthy control group. For the Parkinson vs. Pathologic classification the training realistic accuracy reached values up to 96 % and 92 % for the validated one.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

A_{RMS}	Root mean square value of the amplitude
apq	Amplitude Perturbation Quotient
f_o	Fundamental frequency
GE	Gain Error
HNR	Harmonics to Noise Ratio
jit	Local jitter
jit_{abs}	Absolute jitter
rap	Relative Average Perturbation
shi	Local shimmer
shi_{dB}	Absolute shimmer
SPL	Sound Pressure Level
$V/UV\%$	Voiced-Unvoiced ratio
vAm	Coefficient of Amplitude variation
vf_o	Coefficient of Fundamental frequency variation
ACO	Acoustic contribution
ACQ	Acquisition contribution

ADC Analog to Digital Converter

CFS Cross Fade Smoothing

CPPS Cepstral Peak Prominence Smoothed

CS Common model and features Selection

DAC Digital to Analog Converter

DAQ Digital AcQuisition device

DC Direct Current

EXT Extraction algorithm contribution

FFT Fast Fourier Transform

GLM Generalized Linear Model

GUM Guide to the evaluation of Uncertainty in Measurements

HE Healthy

HY Hoehn-Yahr rating scale

INL Integral NonLinearity

KS Kolmogorov-Smirnov

LR Logistic Regression

LSB Least Significant Bit

MAS Moving average smoothing

MCMC Markov chain Monte Carlo

MH Metropolis Hastings

NSR Noise to Signal Ratio

PA Pathologic non-Parkinsonian

PD Parkinson's Disease

PM Perturbative method

ppq Pitch Period Perturbation Quotient

PS Proposed model and features Selection

RMS Root Mean Square

SNR Signal to Noise Ratio

SR Sampling Rate

UPDRS Unified Parkinson's Disease Rating Scale

WOE Weight of Evidence

Chapter 1

Introduction

1.1 General introduction

The main objective of this work consists in implementing an artificial intelligence that produces clinical predictions, which has the following characteristics:

- safety: the training and the predictions of an artificial intelligence should be safe for the patient
- repeatability: repeated training and predictions should give always the same outcome if the experimental conditions do not change.
- trustablity: the prediction should be given in terms of confidence and risk in order to leave the final decision to the patient or to the clinician
- traceability: an incorrect prediction should be back-traced to the source which caused the error
- accountability: the responsibility of the prediction should never lie with the patient or the clinician. The responsibility of an incorrect prediction should always lie with the artificial intelligence
- adaptablity: the artificial intelligence should be able to choose the best set of features which maximises the prediction accuracy in various experimental set-ups.

Keeping in mind such criteria, a series of analogies with "natural intelligences" will be presented in the following sections

1.2 A conceptual analogy with a "natural" intelligence

The following scenario is considered: a person wants to take safe-driving lessons. One of the tasks asked to the driver to pass the final exam is to demonstrate his ability to brake the car in a given space as shown in Fig. 1.1.

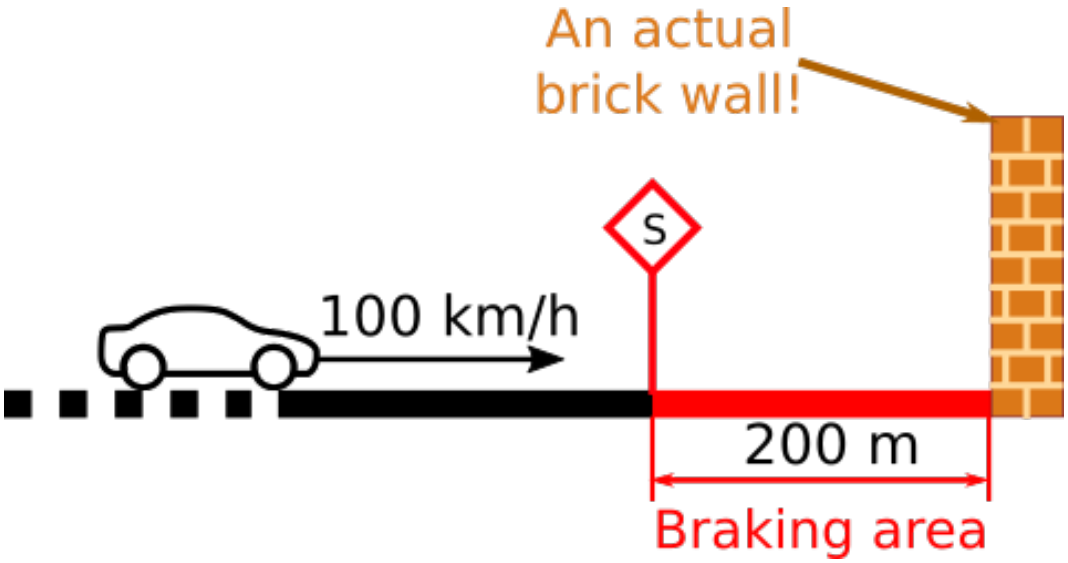


Fig. 1.1 An unsafe safe-driving task

As shown in Fig. 1.1, the task consists in accelerating the car until it reaches a speed of 100 km/h. Once the car reaches the stop sign S the driver have to try to stop it before hitting a brick wall. This is a practical example of a bad designed training task, because if the car reaches the wall with an high speed the car crashes, the driver dies and the driving teacher gets arrested. Such training experiment does not fulfill the safety characteristics described earlier and no one with a little common sense would agree to try this experiment.

1.3 Safety: how to stop a car?

In order to make this experiment safe the brick wall have to be removed. A possible solution is depicted in Fig. 1.2

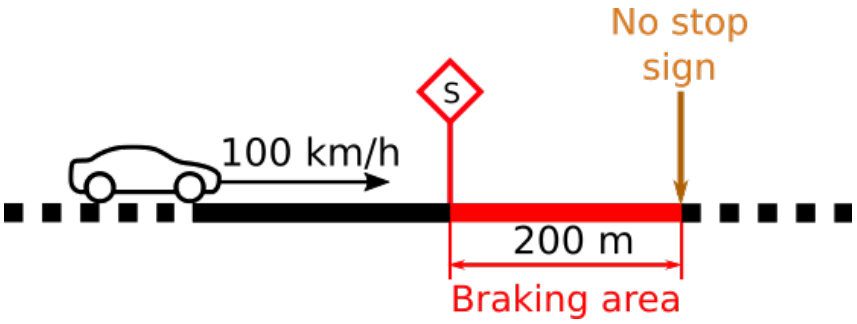


Fig. 1.2 A safe safe-driving task

In the example scenario depicted in Fig 1.2, the drivers are asked to stop after 200 m beyond the stop sign without having a stop sign at the end of the breaking area or a mark on the road. In this scenario each driver will stop the car after a distance which will be evaluated by its own perception of distance, so everyone will stop the car in different positions. Such a training experiment does not fulfill the repeatability characteristics defined in the list above, because it is impossible to evaluate if the drivers are actually able to stop the car before hitting the wall.

1.4 Repeatability: how to train to stop a car?

So how to make such training experiment repeatable? A possible solution is depicted in Fig. 1.3

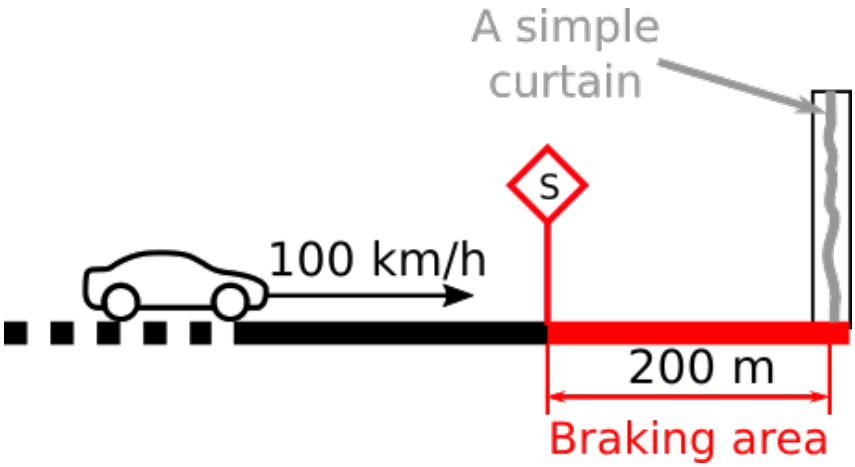


Fig. 1.3 A safe and repeatable safe-driving task

Replacing the brick wall with a soft curtain can make the training experiment repeatable because if the car hits the curtain nothing happens. With this setup a safe series of training experiments is possible and also an evaluation of the driver ability of braking the car is easy to implement.

1.5 Trustability: how to be confident of the braking

One way to evaluate such ability is to measure the distance between the car and the curtain rest position after the car has completely stopped. Repeating the experiment for a reasonable number of times, the stopping distances can be collected to build a statistical description of the experiment as shown in Fig. 1.4.

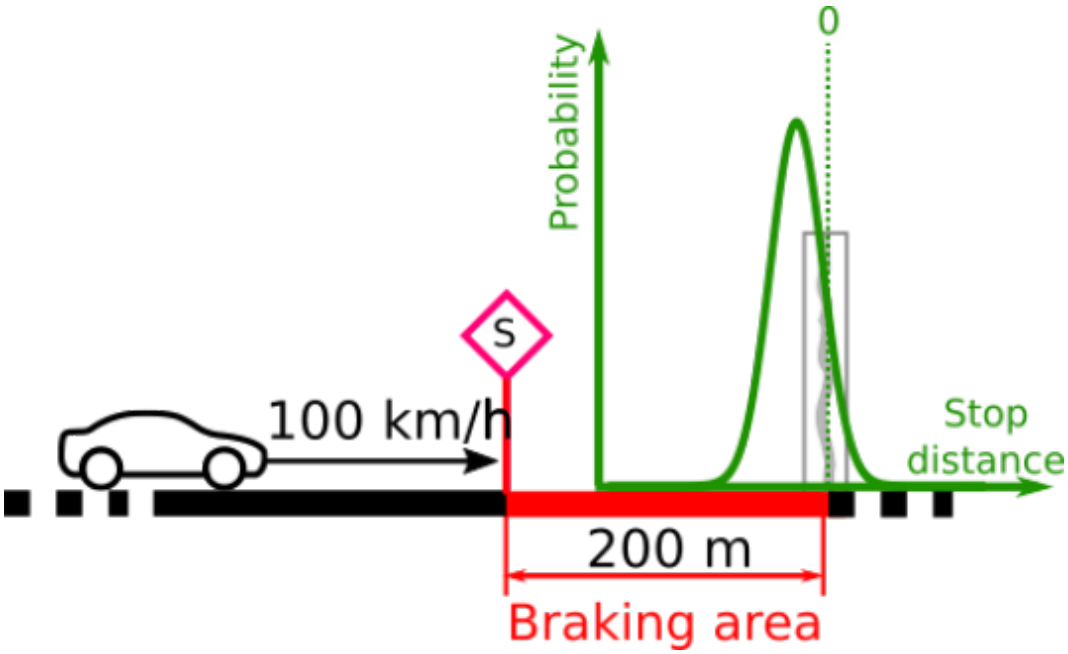


Fig. 1.4 A statistical evaluation of a driver braking ability

With this setup, the driver could decide if he is willing to perform the final exam, basing his choice on the cumulative probability of hitting the obstacle. After thousand of trials, the definition of a confidence-risk framework can be implemented in order to decide whether or not to perform the final exam. So, if the driver wants to have a 99 % confidence of not hitting the obstacle he is taking an 1 % risk of hitting it. If such a risk is low enough for the driver, he could decide to take a chance and try the final exam.

1.6 Traceability: who to blame when something goes wrong?

The statistical evaluation, depicted in Fig. 1.4 is useful to determine the cumulative probability of hitting the wall but it does not allow to distinguish the human contribution from the machine contribution. If something goes wrong it is impossible to establish the responsibility of a possible accident. Let's exemplify the human error contribution as simply caused by his reaction time using two sensors as shown in Fig. 1.5

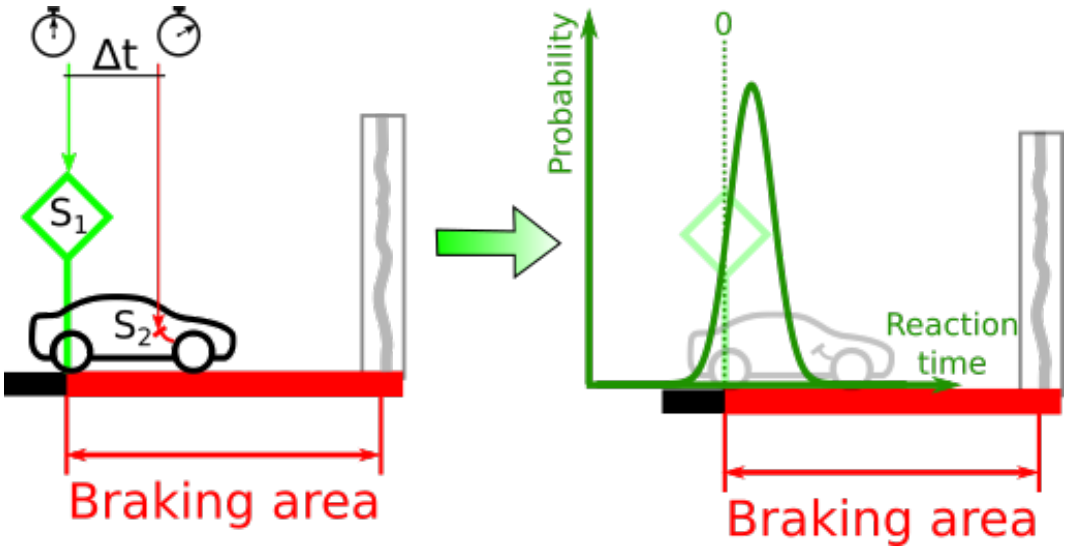


Fig. 1.5 Measuring method to evaluate the reaction time of a driver using two sensors producing a statistical evaluation

In this example, a photocell sensor S_1 starts a timer that is stopped when the driver pushes the sensor S_2 mounted on the brake pedal. According to this measurement, it is possible to evaluate statistically the reaction time Δt of the driver respect to a visual signal indicating the start of the braking area.

1.7 Accountability: how to discriminate the human from the machine

If the driving instructor wants to be sure that the car mechanics does not affects the result of the brake, an automatic braking system could be mounted to reduce the reaction time as close as zero. Such a breaking system could be a simple electro-mechanical actuator mounted on the brake pedal to simulate the driver foot push on the pedal. The actuator can be connected via wireless to the S_1 sensor to be activated as soon as the car crosses the sensor. Using this setup, the driving instructor can evaluate the probability of hitting the wall using a repeatable and trusted brake actuator. if the probability of hitting the wall is close to 0 %, then the responsibility of an accident lies just on the driver reaction time Δt and not on the car mechanics.

1.8 Adaptability: how to train a "natural intelligence"

Before the driver decides if he wants to take the risk of hitting the obstacle, he needs to evaluate his ability to take the correct decision. In this particular case the decision is clear: you have to brake the car! How to do it is not only a matter of car mechanics and dynamics physics, but also the driver perception of the experiment is highly involved in the decision [1]. In this example some of our senses are involved and in particular the sight and the sense of touch. The sight is a fundamental sense to evaluate the residual distance between the car and the obstacle, while the sense of touch is what gives to the driver the perception of car acceleration through the push of the seat on his back and the tension of the arm muscles flexing on the wheel. Also the hearing could give some informations on the car speed and spatial localisation but, respect to the sight and the sense of touch, such sense is less considered in such experience. All this informations coming from our senses have a natural importance hierarchy depending on the task we are performing. For this example a reasonable hierarchy could be:

1. sight (perception of residual distance)
2. touch (perception of acceleration)
3. hearing (perception of speed and body balance)
4. smell (no useful perception)
5. taste (no useful perception)

This hierarchy depends on the confidence one has in each of the senses, which is given by the reliability of the metric perception of the task that the subject is carrying out. If the braking area in Fig. 1.1 is extended to 500 m the sense of sight becomes less relevant in the execution of the task because it would take just a glance from time to time to evaluate the residual distance. In this alternative scenario the hierarchy could be:

1. touch (perception of acceleration)
2. sight (perception of residual distance)
3. hearing (perception of speed and body balance)

4. smell (no useful perception)

5. taste (no useful perception)

In fact the execution of such a task also depends on more complex perception models regarding the external environment, like the proprioception (perception of body positioning in space) and objective permanence (the ability of predicting the temporal evolution of a body in space, even when the body cannot be sensed). In this case the driver is allowed to get distracted and look at something other than the remaining distance and trust his own proprioception and his objective permanence to evaluate the residual distance between the car and the obstacle [2] [3].

1.9 Main topics of this thesis

The paradigms described in the previous sections were used to design an artificial intelligence capable of discriminating patients with the Parkinson's Disease (PD) , from an HEalthy control group (HE) and a PAthologic non-Parkinsonian (PA) set of patients.

1.9.1 The Parkinson's Disease

The Parkinson's Disease is a degenerative disorder which affects the central nervous system. The Parkinson's disease affects the production of dopamine in an area of the brain called *substantia nigra*. The symptoms include tremor, bradykinesia (slowness of the movements), limb rigidity and can also have an effect on voice production. To assess the clinical status of a PD patient some international rating scales have been developed. The most used is the UPDRS (Unified Parkinson's Disease Rating Scale) [4] developed in 1980s and updated in 2019. The UPDRS scale is based on an empirical evaluation of some aspects of the daily living of the PD patients and on a motor skills examination performed by a neurologist. In particular, the UPDRS scale is composed of four parts:

- Part I: non-motor experiences of daily living (13 questions);

- Part II: motor experiences of daily living (13 questions);

- Part III: motor examination (18 questions);
- Part IV: motor complications (6 questions).

For each question it is necessary to assign a score from 0 to 4. The total cumulative score is in a range between 0 (no disability) and 199 (total disability) [4, 5]. Another common scale used for PD assessment is the Hoehn-Yahr (HY) scale [5], which is used to evaluate the disease status and its progression over time. It has five grades ordered according to the severity of the disease:

- 0: no signs of disease;
- 1: symptoms on one side only (unilateral)
- 2: symptoms on both sides but no impairment of balance
- 3: balance impairment, mild to moderate disease, physically independent
- 4: severe disability, but still able to walk or stand unassisted
- 5: needing a wheelchair or bedridden unless assisted

1.9.2 Effects of the Parkinson's Disease on voice production

The Parkinson's Disease affects also the phonatory system and, in fact, the UPDRS score includes qualitative speech evaluations in parts 2 and 3. The Parkinson's Disease, which affects the motor system, makes difficult also the voice production. In particular the most common symptoms that can occur in voice production are:

- Hypophonia: loss of tonality and modulation; in same cases patients show total loss of the voice (aphonia)
- Monotone voice
- Stuttering: progressive acceleration of words and uncontrolled repetitions
- Dysarthria: difficulty in speaking, incorrect pronunciation of sentences

Several studies have been carried out on voice emission of PD patients [6] [7] [8] [9] [10] [11], which are based on the analysis of vocal material using machine learning techniques. In the majority of such studies the most recurring voice features are the vowel stability metrics (i.e. jitter and shimmer) and other spectral measurements such as the Harmonics to noise ratio. Another important voice quality measurement is the Cepstral Peak Prominence Smoothed (CPPS), which evaluates the *harmonicity* of a subject voice. Such a measurement has been used in several studies [12] [13] [14] to try to relate the CPPS evaluation to the Parkinsons' Disease status and its evolution over time. The CPPS is also adopted as a generic voice quality indicator to assess the health status for different voice pathologies [15] as well as to evaluate the general quality of Normophonic Subjects [16] (the healthy control group HE studied in this work).

1.9.3 Voice features and their uncertainties

Based on the existing literature, the above-mentioned voice features were analyzed in this work. In particular, such features are extracted from audio recordings of sustained vowels /a/ acquired with a microphone in air.
Regarding the uncertainty of the considered features, some studies have been carried out [17] [18] using parametric mechanical models of the human phonatory apparatus. According to author's knowledge, at the time of the writing of this manuscript, no studies have been found in the literature about the analytical evaluation of the voice stability metrics uncertainty. Moreover, a Monte Carlo uncertainty propagation of such voice features has not been carried out yet, so no studies are present in the scientific literature about this topic.
In order to fill this gap, this work tries to evaluate the uncertainty of voice features using different approaches. An analytical evaluation of some simple stability metrics, such as jitter and shimmer, was carried out, as presented in Chapter 3. Moreover, in the same chapter, a Monte Carlo uncertainty propagation of the stability metrics used in this work will be presented.
For this work, one of the main objective is to separate the human contribution from the machine contribution to the voice features uncertainty using a metrological approach, as will be showed in Chapter 4 and Chapter 5. In Chapter 4, a novel vowel synthesis method is proposed to evaluate the uncertainty of the voice features used to train classification algorithms for the recognition of the Parkinson's Disease. A

similar synthesis method [19] was implemented to produce synthetic voice samples with target arbitrary jitter and shimmer, where a perceptual assessment of the quality of these synthetic voices was performed by trained listeners. Using the proposed synthesis method, the effects of the measuring chain perturbations on voice features uncertainty were studied, as reported in Chapter 5. In order to evaluate if such perturbations may affect the training and the predictions of a binary classification algorithm, the original and synthetic vowels were used to train and validate logistic regression models, as will be shown in Chapter 6.

Chapter 2

Materials and methods

2.1 Acquisition Devices

For the present work, the voices of the involved subjects were recorded using a cheek condenser microphone in air (MIPRO MU 55-HR) connected to a portable audio recorder (Edirol Roland R-09HR). The sampling rate of the audio recorder was set to 44100 Sa/s and the bit resolution to 16 bit. The input gain of the audio recorder was set in order to have an input level around -6 dB_{fs}, as indicated by the recorder display. The microphone capsule was placed in front of the mouth opening at a distance in a range between 2 cm and 4 cm. The subjects' voices were also recorded using a contact piezoelectric microphone placed on the neck, as shown in Fig. 2.1:

Fig. 2.1 Contact microphone and microphone in air positioning

Both contact and in-air microphone were connected to the recorder through a split Y cable. The contact microphone was used to conduct other voice experiments

which will not be presented in this manuscript. The reason for this choice can be understood if the architecture of the whole characterization experiment is considered, as will be showed in Sec. 5. For the work presented in this manuscript, the author wanted to discriminate the human contribution from the machine contribution to the vocal features uncertainty substituting the subject under test with an artificial human simulator. This simulator is a Brüel & Kjær Head And Torso Simulator (HATS) which has a loudspeaker placed inside tits mouth, as shown in Fig. 5.21 (b). Unfortunately, the HATS is not equipped with a vocal folds simulator, therefore placing the contact microphone on the simulator's neck is meaningless. Without being able to replicate the input stimulus using the simulator, the characterization of the whole measuring chain was not possible and the recordings acquired with contact microphone were discarded for this work.
Both microphones need to be powered by a low DC voltage applied to the terminals of the microphones. Such voltage is called plug-in power and commonly is set between 3 V and 5 V depending on the power supply of the recorder. The electric audio signal coming from microphones is DC decoupled by series capacitors mounted on the recorder circuitry. Commonly the output impedance of a plug-in powered microphone is 2.2 kΩ so this value was taken as a reference in some of the experiments performed for this work .

2.2 Recordings

The audio recordings were collected at "Città della Salute, Torino". Three classes of voices were recorded:

- Parkinson patients (PD): number of subjects N=44 (29 Male, 15 Female) , mean age 67, standard deviation 13
- Healthy subjects (HE): N=58 (29 M, 29 F), mean age 31, standard deviation 15
- Pathological non-Parkinsonian patients (PA): N=61 (25 M, 36 F), mean age 52, standard deviation 16

The PD group was diagnosed by two Neurologists which assigned a clinical evaluation to each patient through UPDRS and Hoehn yahr evaluation scales. The

voice health status for the remaining classes was certified by endoscopic analysis.

The three subsets are very unbalanced regarding the age, so a subject selection was performed in order to balance age and gender. In particular, the 10 youngest PD subjects and the 10 oldest HE subjects were selected in order to balance the ages. For the PA dataset, the subjects was selected in order to balance the ages with the PD and HE dataset. The dataset used in this work is composed by 10 subjects from each class:

- Parkinson patients (PD): N=10 (5 M, 5 F), mean age 52, standard deviation 6
- Healthy subjects (HE): N=10 (5 M, 5 F), mean age 51, standard deviation 7
- Pathological non-Parkinsonian patients (PA): N=10 (5 M, 5 F), mean age 52, standard deviation 6

In the plot in Fig. 2.2, the mean ages and standard deviation of each class are showed:

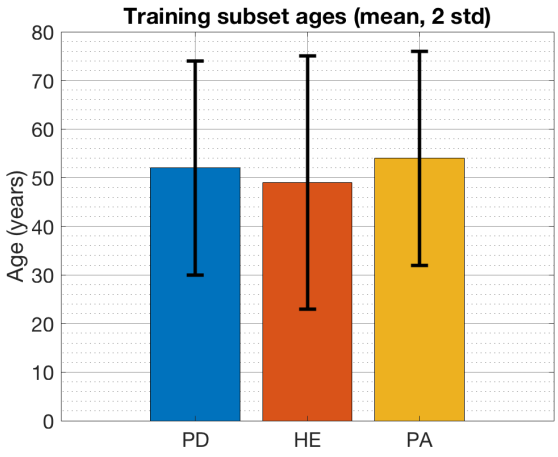


Fig. 2.2 Mean ages of the balanced dataset

The recordings took place in a public, non-acoustic clinic room with closed doors and windows. During the recordings, the people inside the ambulatory room were asked to stay as quiet as possible. Some background noise was found in the audio recordings and, to evaluate it, the Signal to Noise ratio (SNR) of the recordings was measured. For the three classes, the average SNR was found to be as high as 30 dB.

2.3 Tasks

The tasks asked to the three groups consisted in the repetition of three /a/ phonemes at a confortable pitch, level and duration. The three repetitions take the total number of audio recordings analyzed in this manuscript to 90. The subjects were asked also to read a phonetically balanced text in Italian. In addition, a minute of free speech on a topic of subject's choice was recorded. For this work, all attention was focused on the vowel repetitions. This was done in order to develop a method to evaluate the voice features uncertainties through the generation of artificial vowels and use the collected informations to train weighted classification models.

2.4 Extracted features

As already stated, for this work just the sustained vowels have been considered. The sustained vowels are pseudo-periodic signals with variable periods, amplitudes and spectral characteristics. As shown in Fig. 2.3 pseudo-periods and amplitude markers can be set over a vowel signal to identify their time evolution.

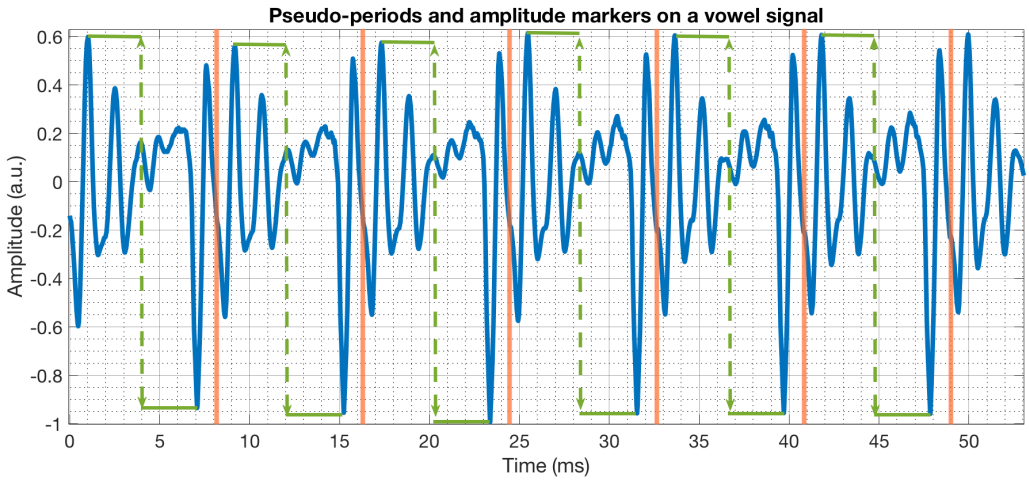


Fig. 2.3 An example of a sustained vowel. The red markers represent the pseudo-periods starting and ending times and the green ones represent the peak-to-peak amplitudes.

2.4.1 Algorithm for pseudo-period and amplitude extraction

To evaluate the pseudo-period lengths and amplitudes, the autocorrelation method was implemented. Such a method consists in multiplying a signal frame, containing at least two pseudo-periods, with a delayed version of the same signal using the Eq. 2.1 [20]:

$$A_c(lag) = \sum_{i=0}^{N} s(i) \cdot s(i+lag) \tag{2.1}$$

where $s(i)$ is the signal frame, $s(i+lag)$ is the signal frame delayed by lag and N is the number of samples of the current frame. Collecting autocorrelation values A_c at different lags an example plot as in Fig. 2.4 can be obtained.

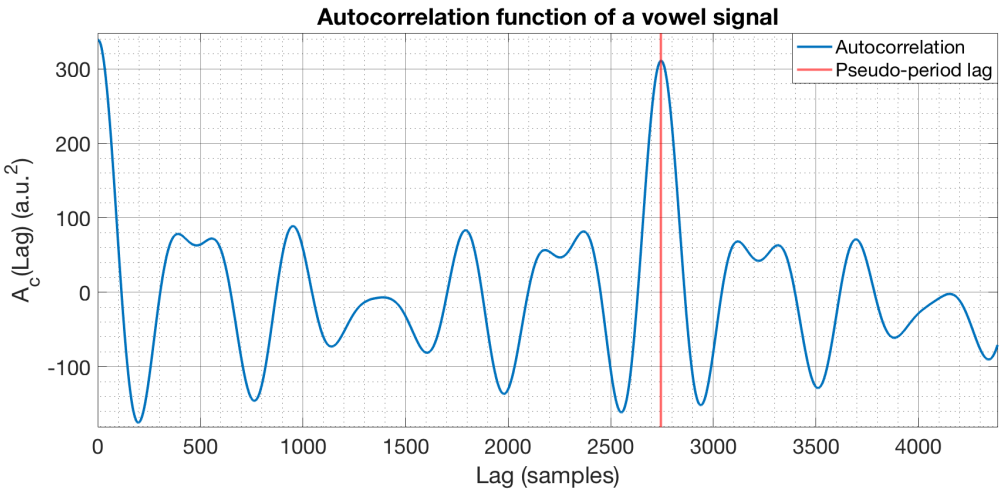


Fig. 2.4 Autocorrelation function of a vowel signal

The plot in Fig. 2.4 represent an autocorrelation curve of a vowel signal acquired with a sampling rate of 44100 Sa/s and then linearly oversampled by a factor of 8. In this way the oversampled signal wave will have a sample rate of 352800 Sa/s. The autocorrelation function assumes the maximum value at lag 0 and such value corresponds to the power of the signal frame ($s(t)^2$). The lag of the first maximum highlighted in red is called the period lag and it has two important properties:

- The *lag* itself identifies the length of the pseudo-period expressed in samples
- The autocorrelation value at the period lag can be used to estimate the harmonicity of the signal

For an ideal periodic signal, multiplying a frame vector to itself has the same result of multiplying a vector to a version of itself delayed by a period T. In such condition the peak of autocorrelation at lag 0 is equal to the peak at lag T. For a vowel signal the period peak is mostly of the time lower than the power peak. To evaluate the harmonicity H the following ratio [20] is calculated:

$$H = \frac{A_c(T)/A_c(0)}{[1 - A_c(T)]/A_c(0)} \tag{2.2}$$

where H is the harmonicity, $A_c(T)$ is the autocorrelation at *lag* T and $A_c(0)$ is the autocorrelation at *lag* 0 (the power of the signal frame). According to Eq. 2.2, the period peak $A_c(T)$ is normalized to the peak $A_c(0)$ in order to refer the peak to the power of the frame. For an ideal periodic signal $A_c(T) = A_c(0)$ and then H tends to infinite, while for an inharmonic signal ($A_c(T) << A_c(0)$), H tends to 0. To evaluate the Harmonics to Noise Ratio parameter (*HNR*), a logarithmic scale is applied:

$$HNR = 10 \cdot Log_{10}(H) \text{ (dB)} \tag{2.3}$$

Sometimes, especially for rising envelope transients and in general with noisy voices, the lag peak could be higher than the power peak. In such conditions, H can assume negative values so the the logarithm can not be calculated. To fix this problem, a simple moving average between the normalized autocorrelations $Ac(T)/Ac(0)$ of consecutive frames is sufficient. The algorithm used for this work is a synchronous autocorrelation method that performs the pseudo-periods extraction and amplitude evaluations of the signals. The main algorithm characteristic is the evaluation of the *HNR* [20] in order to discriminate valid (Harmonic), from invalid (unHarmonic) signal frames. Such a discrimination is not critical for healthy voices where the *HNR* parameter is mostly above 0 dB. Unhealthy voices as in the PD and PA subsets can lead to negative *HNR* evaluations, which corresponds to signals where the Harmonic component has an energy which is lower than the noise level. In such conditions, the extraction of pseudo-periods is a very difficult task so the frames evaluated with an $HNR < 0$ are labeled as invalid and they are not processed for feature extraction purposes. An additional condition is set to label the frames valid or not and it is a frequency jump condition. If the current frequency evaluation is more than half octave higher or lower than the previous frequency then the frame is labeled as invalid. Such condition helps to avoid octave jumps that occurs when processing unsteady and harsh voices. When an octave jump event occurs, the previous frequency is

stored for subsequent comparisons, but the evaluated frequency is discarded. The extraction algorithm used in this work can be summarized by the flow-chart in Fig. 2.5

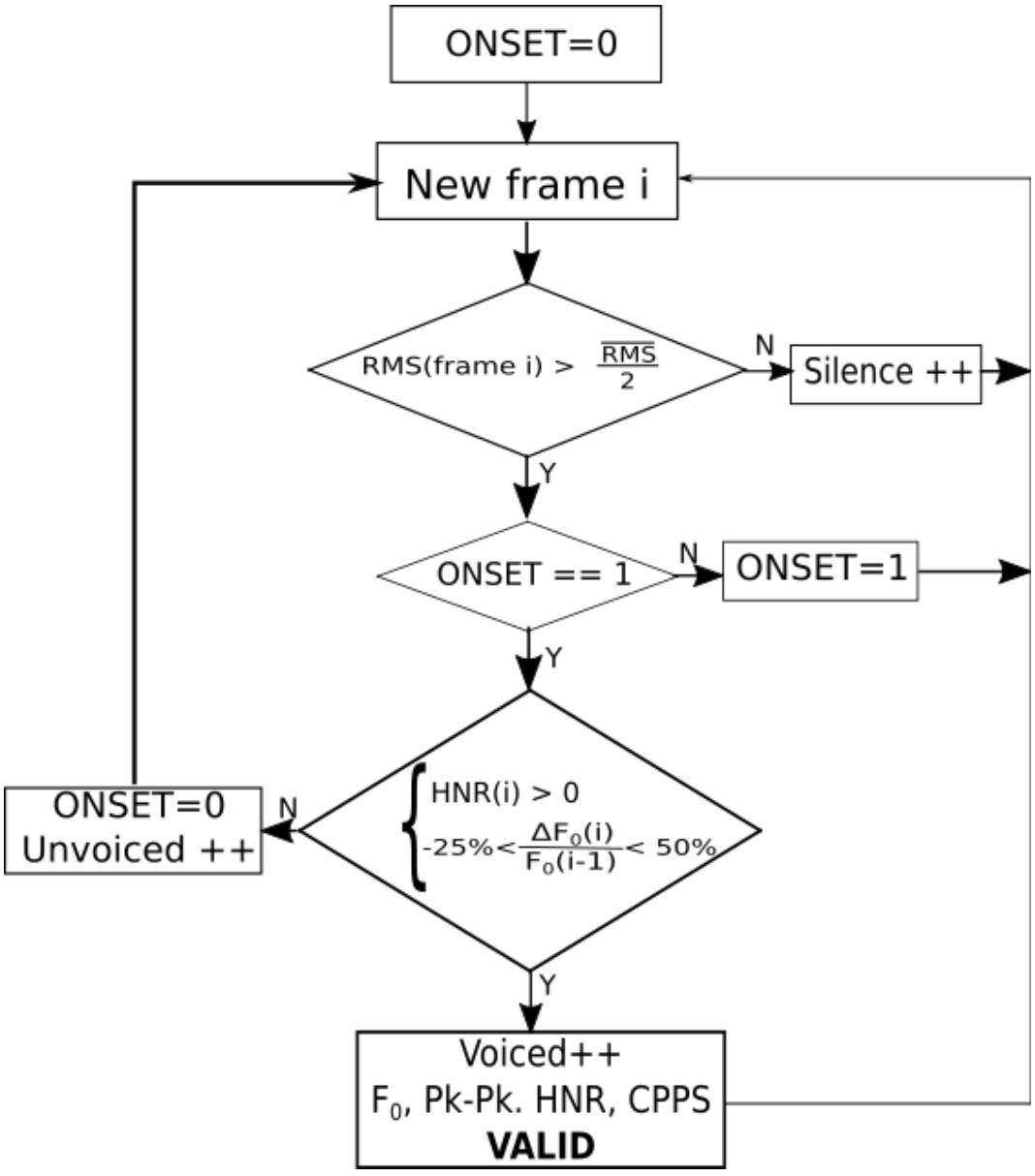


Fig. 2.5 Feature extraction algorithm

As shown in the flow-chart, the ONSET variable is initially set to 0 and a new frame is acquired from the vowel signal. If the Root Mean Square (RMS) of the current frame *i* is less than the mean RMS of the entire signal divided by 2 then the

Silence variable is incremented and a new frame is acquired. If the RMS condition is met, the ONSET variable is evaluated and if it is equal to 0 a new frame is acquired. If the ONSET variable is equal to 1, the *HNR* and the relative frequency jump between consecutive frames $\frac{\Delta f_o(i)}{f_o(i-1)}$ are evaluated. If the *HNR* is greater than 0 dB and the relative frequency jump is less than an half octave, the frame is labeled as valid, the Voiced variable is incremented and the features of the frame are extracted. If the *HNR* and frequency jump conditions are not met then the ONSET is set to 0, the Unvoiced variable is incremented and a new frame is collected. For this work, the vowel signal was normalized to the absolute peak of the entire signal so the amplitude is expressed in arbitrary units (a.u.). The length of the frame is fixed and is equal to the maximum evaluable pseudo-period ($\approx$ 12 ms), while the frame time shift is equal to the previous evaluated pseudo-period.

2.4.2 Period and amplitude stability metrics

For clarity, this work focus mostly on two measurements: local jitter (jit) and local shimmer (shi)

$$jit = \frac{N}{N-1} \cdot \frac{\sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\sum_{i=1}^{N} T_i} \cdot 100\ (\%) \tag{2.4}$$

$$shi = \frac{N}{N-1} \cdot \frac{\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\sum_{i=1}^{N} A_i} \cdot 100\ (\%) \tag{2.5}$$

where T_i are the vowel estimated pseudo-periods and A_i the corresponding peak-to-peak amplitudes. Other stability metrics have been considered for this work. In this manuscript a unique identificative number is associated to such metrics:

1. jit: Local jitter (%) - Eq. 2.4
2. jit_{abs}: Absolute jitter (μs)
3. rap: Relative Average Perturbation (%)
4. ppq: Pitch Period Perturbation Quotient (%)
5. vf_o: Coefficient of Fundamental frequency variation (%)
6. shi: Local shimmer (%) - Eq. 2.5

7. shi_{dB} : Absolute shimmer (dB)
8. apq: Amplitude Perturbation Quotient (%)
9. vAm: Coefficient of Amplitude variation (%)

The definition of these parameters is provided in Appendix A

2.4.3 Other period and amplitude metrics

From the pseudo-period extraction algorithm depicted in Fig. 2.5, additional parameters are extracted from the vowel signals:

- Harmonics to noise ratio HNR: described by Eq. 2.3 (dB)
- Fundamental frequency f_o: defined as $f_o = 1/T$ (Hz)
- Root mean square value of the amplitude: $A_{RMS} = \sqrt{(\sum_{n=1}^{N} s_n^2)/N}$ (a.u.)

In particular, the A_{RMS} values are calculated from the signals that were normalized to 1. The peak normalization of the signals causes a loss of information about the original signal RMS. The recordings took place in a very busy public clinic room where the patients were visited by the Neurologist before the recordings. An absolute calibration of the cheek microphone was tested with an adapted acoustic reference calibrator to fit the size of the microphone capsule. The microphone calibration would have made it possible to obtain information about the absolute Sound Pressure Level (SPL) . However, the evaluation of the SPL also depends on the distance of the microphone, which in this work could not be fixed accurately due to several factors such as the time required to complete the task and the patients facial morphology. The area around the mouth shows an important variation in sound intensity, as shown by several studies . Vocal emission patterns have a great variation even in artificial simulators of human voice, as will be shown in Sec. 5.5 (Fig. 5.22). Furthermore it has not been possible to set the same gain for each of the recordings, which was set in order to have an average level of -6 dB_{fs} for each of the recordings. For this reason the author preferred to not perform absolute SPL comparisons between subjects, so the peak normalization was carried out on vowel signals to simulate a real world application where the repeatability of the signal acquisition procedure cannot

be guaranteed, as in the case of unsupervised acquisition or when using different microphones such as the internal microphone of a smartphone.
The collected sequences of HNR, f_o and A_{RMS} have the same length of the sequences of the extracted pseudo-periods and amplitudes. The sequences of extracted parameters were transformed in statistical distribution which can be described with statistical metrics in order to reduce the size of collected data and to achieve a more representative identification of each vowel. In particular the following descriptive statistics are calculated:

- Mean value
- Median value
- Mode value
- Range
- Standard deviation
- 5° percentile
- 95° percentile
- Skewness
- Kurtosis

For this manuscript, the localization metrics such as mean, median and mode have not been considered for the f_o parameters because such values depends on the subject will and ability to produce high or low pitched vowels. In a similar way, the localization parameters of the descriptive statistics of the A_{RMS} evaluations may be altered by the signal normalization described earlier. For this reason such metrics have been considered as non informative and potentially as a source of errors for the classification algorithms. The last extracted feature is the Voiced-Unvoiced ratio $V/UV\%$, defined as the ratio between the number of harmonic ($HNR > 0$) and inharmonic frames ($HNR < 0$).

2.4.4 Cepstral Peak Prominence Smoothed (CPPS)

The harmonicity of the signal can be evaluated through cepstral measurement. The word "Cepstrum" comes from the anagram of spectrum and it is a transformation of the signal spectrum. In particular each frame is multiplied with a Hanning window to minimize the effects of non coherent sampling. A Fourier transform is performed to the windowed frame and then transformed with a log_{10} function and an another Fourier transform is taken on such frame. The module of such transform is called "Cepstrum".

$$C_p = 20 \cdot Log_{10}|\mathscr{F}\{20 \cdot Log_{10}(|\mathscr{F}\{s(t)\}|) \text{ (dB)} \tag{2.6}$$

where C_p is the cepstrum vectror, $\mathscr{F}$ is the Fourier transform of the variable and $s(t)$ is the signal time series. The cepstrum can be plotted as a function of the variable called "quefrency" (anagram of frequency) to obtain a curve similar to the last row of the example in Fig. 2.7

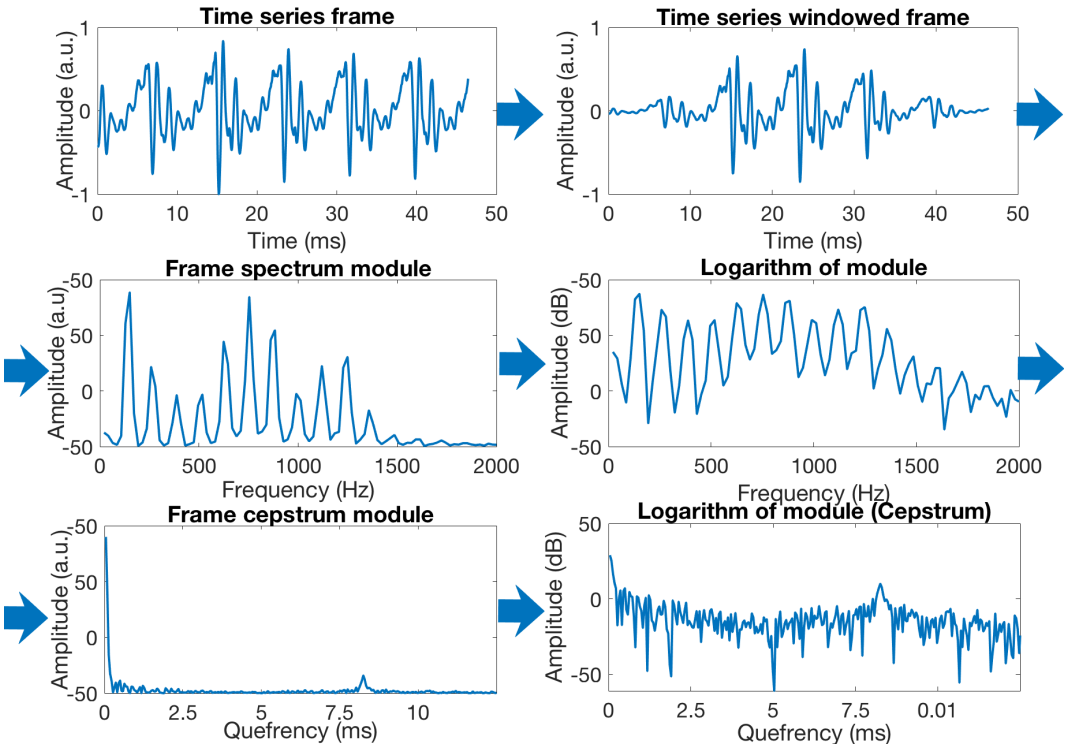


Fig. 2.6 Cepstrum extraction sequence

As shown in Fig. 2.6, the time series frame is windowed with a Hanning function and the module of the Fourier transform is calculated. The logarithm of the module spectrum is then transformed with a Fourier transform to obtain the curve in the bottom left. To better evaluate such a transformation, the cepstrum is converted in a logarithmic scale. The role of the second Fourier transform on the spectrum array is to highlight the periodicity in the spectrum. The spectrum of a sustained vowel has an harmonic appearance as can be noticed in Fig. 2.7, so the harmonic peaks are positioned on integer multiples of the fundamental frequency f_o. The most important feature of a cepstrum is the cepstral peak as shown in Fig 2.7

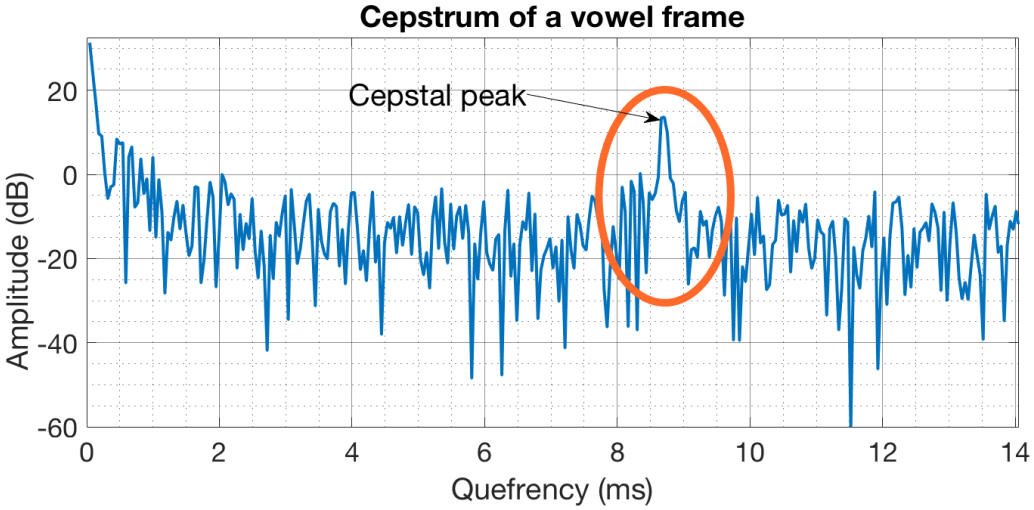


Fig. 2.7 Cepstrum

Such a peak is positioned at a quefrency equal to the estimated pseudo-period length and its prominence, evaluated from the noise floor, is a measure of signal harmonicity. In fact the capability of the Fourier transform of finding periodicity in signals can be adapted to find periodicity in the frequency domain. An harmonic spectrum with a low noise floor is seen as a periodic wave with a certain offset and amplitude from the successive Fourier transform performed to obtain the cepstrum. Therefore the more the amplitude of the spectrum is (distance between harmonics peaks and noise floor) the higher the cepstral peak will be. As shown in Fig 2.7, the cepstrum curve can be very noisy and estimating the prominence of the cespstral peak can be a challenging task, especially when the audio recordings are performed in a non-treated room [21]. To fix this problem, a two dimensional smoothing of collected cepstrum frames is performed [15] [16] [12] [22]. The cepstra sequence is

first smoothed in the time domain with a moving average filter with a frame size of 7 samples [23] [24]and then is smoothed in the quefrency domain using the same moving average filter. Once the collected cepstra were smoothed, for each cepstrum the regression line of the noise floor is estimated from 1 ms to the end of the cepstrum. The value of 1 ms was chosen to exclude the frequencies above 1 kHz where very small harmonic energy is present in vowels signals. The height of the projection of the smoothed cepstral peak on the tendency line is the evaluated Cepstral peak prominence smoothed (CPPS) for the current frame, as shown in Fig 2.8

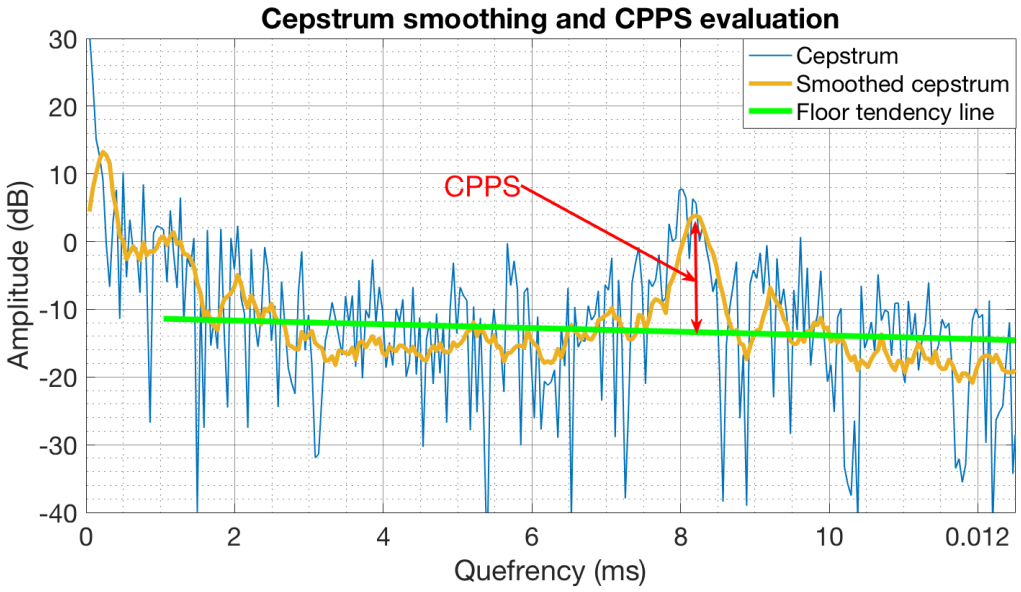


Fig. 2.8 Cepstral peak prominence smoothed

Chapter 3

Uncertainty evaluation of the extracted parameters

The main goals of this chapter are the identification and the evaluation of the contributions that affect the uncertainty of the extracted parameters. The first part of this chapter refers to the uncertainty analysis of pseudo-period and amplitude stability parameters defined in chapter 2.4.2, paying particular attention to the quantization error. The second part analyses the error contribution of the extraction algorithm by means of a re-synthesis method that provides reference signals, which allow the error due to the extraction algorithms to be evaluated.

3.1 Uncertainty evaluation of period and amplitude stability metrics

The period and amplitude stability metrics defined in Sec. 2.4.2 are evaluated starting from measured sequences of pseudo-periods and amplitudes. The uncertainty of these measured sequences affect the stability-metrics uncertainty. The first step consists in identifying the main uncertainty contributions of pseudo-period and amplitude measurements, which are:

- time-base tolerance
- time-base drift

- time resolution
- gain error
- integral nonlinearity
- amplitude resolution
- noise
- extraction algorithm

The contributions described in the list above can be depicted in the diagram in Fig. 3.1

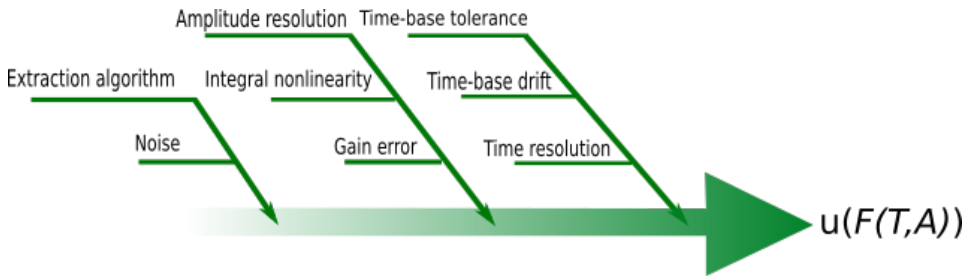


Fig. 3.1 Main uncertainty contributions for pseudo-periods and amplitudes evaluations

The figure 3.1 summarizes the uncertainty contributions that have been taken into account, where u(*F(T,A)*) represents the standard uncertainty of the generic feature f, which depends on pseudo-period (T) and amplitude (A). The noise affects more the extraction algorithm than the feature evaluation itself as will be showed in the next sections.

3.1.1 Time-base tolerance, aging and resolution

The contribution of time-base tolerance affects the sampling phase of the analog-digital conversion of the Analog to Digital Converter (ADC) and thus all the time measurements. To evaluate the effect of time base tolerance and aging the time measurements were modeled using the formula:

$$T_i = \frac{n_i}{sr_i} \tag{3.1}$$

where n_i is the number of samples that corresponds to the period Ti and that is assumed as exact, while sr_i is the sampling rate affected by tolerance and aging:

$$sr_i = sr_n \cdot (1 \pm \varepsilon \pm k \cdot \Delta t) \tag{3.2}$$

where sr_n is the nominal sampling rate, ε is a relative tolerance, k is a relative time drift expressed in s^{-1} and Δt is the duration of the acquisition interval. Combining the equations 3.1 and 3.1 the number of counted samples can be expressed as:

$$n_i = T_i \cdot sr_i = T_i \cdot [sr_n \cdot (1 \pm \varepsilon \pm k \cdot \Delta t)] = n_{i0} \cdot (1 \pm \varepsilon \pm k \cdot \Delta t) \tag{3.3}$$

where $n_{i0} = T_i \cdot sr_n$ is the expected number of samples. Substituting n_i in Eq. 3.1:

$$T_i = T_{i0} \pm T_{i0} \cdot (\varepsilon + k \cdot \Delta t) \tag{3.4}$$

where $k \cdot \Delta t$ is the linear time drift over the time interval Δt and T_i is the duration of the i-th period. An evaluation of the effect of aging on time measurements can be done considering a vowel with a duration in the order of 10 s. Considering a worst case scenario of a commercial Room Temperature Crystal Oscillator (RTXO), with an aging time drift $k = 10^{-6}$ month$^{-1} \approx 3.8 \cdot 10^{-13}$ s^{-1}, the term $k \cdot \Delta t$ is in the order of 10^{-12}, which corresponds to 10^{-6} ppm, so the aging uncertainty contribution can be considered as negligible.
A tolerance ε of the timing crystal frequency affects negligibly the evaluation of periods duration. As an example, a pessimistic clock tolerance in the order of 100 ppm, on a typical clock frequency of 11.28 MHz, leads to a frequency error of 1128 Hz. Assuming that there is no drift caused by the clock divisor circuit, the standard audio sampling frequency of 44100 Sa/s will have an error around 4.41 Sa/s. Regarding the period evaluation error, an example of the effect of a positive clock error ε around 100 ppm can be evaluated on a 10 ms period measurement as:

$$T_i = T_{i0} \cdot (1 + \varepsilon) \approx 10.001 \text{ ms} \tag{3.5}$$

As shown in this particular example, the timing crystal tolerance can produce an important effect on time measurements. However, the tolerance ε on the timing crystal frequency on period stability evaluations has no effect due to the relative nature of such measurement. As an example the effect of tolerance on jitter measurements can be evaluated by the following equation:

$$jit^* = \frac{N}{N-1} \cdot \frac{\sum_{i=1}^{N-1} |T_{i0} \cdot (1 \pm \varepsilon) - T_{(i+1)0} \cdot (1 \pm \varepsilon)|}{\sum_{i=1}^{N} T_{i0} \cdot (1 \pm \varepsilon)} \cdot 100 = \frac{N}{N-1} \cdot \frac{\sum_{i=1}^{N-1} |T_{i0} - T_{(i+1)0}|}{\sum_{i=1}^{N} T_{i0}} \cdot 100 = jit \quad (3.6)$$

As shown in Eq. 3.6 the terms $(1 \pm \varepsilon)$ cancel out in the jitter equation so no effect is expected due to the timing crystal frequency tolerance.
Another important contribution on time measurements is caused by the finite resolution of the ADC time-base. Because of the lack of synchronicity between the input signal and the time-base, the start and ending instants of each evaluated pseudo-period have an absolute uncertainty of ± 1 sample corresponding to a sampling period of ± 1 T_s. The period evaluation, which is the difference between the ending and starting instants will have an absolute uncertainty of ± 2 samples. If this uncertainty is supposed to be caused by a random perturbation with an uniform distribution extending from -2 T_s to +2 T_s the period uncertainty is evaluated as $4 \cdot T_s / \sqrt{12} = 2 \cdot T_s / \sqrt{3}$ As an example, with a sampling rate $sr = 44100$ Sa/s the sampling period is $T_s = 22.6$ μs, so $u(T)$ is in the order of 26 μs. Considering the fundamental frequency of a vowel between 80 Hz and 400 Hz, the relative uncertainty is in a range between 0.2 % and 1 %. These values are larger than the tolerance (10^{-4}) and the aging (10^{-12}) uncertainty contributions. In conclusion, the time-base resolution of time measurements is the main contribution to the pseudo-period uncertainty.

3.1.2 Analytical uncertainty propgation of jitter and shimmer

In this section, the analytical evaluation of features uncertainty is carried out for some of the period and amplitude stability measurements as recommended by the Guide to the evaluation of uncertainty in meaurements (GUM) [25]. In particular, the uncertainty propagation of jitter and shimmer is carried out according to the following general equation:

$$u(F) = \sqrt{\sum_{j=1}^{N} \sum_{i=1}^{N} \left(\frac{\partial F}{\partial X_i} \cdot \frac{\partial F}{\partial X_j} \right) \cdot \sigma_{X_{ij}}} \quad (3.7)$$

where F is a generic feature, $X_{i,j}$ are the input variables and $\sigma_{X_{ij}}$ is the variance-covariance matrix. If the correlation between input variables is negligible, a simpli-

fied version of the equation can be used:

$$u(F) = \sqrt{\sum_{i=1}^{N} (\frac{\partial F}{\partial X_i} \cdot \sigma_{X_i})^2} \tag{3.8}$$

The feature jitter is a function of measured pseudo-periods T_i, according to equation 2.4, and then the corresponding uncertainty can be obtained as:

$$u(shi)=u(A)\cdot\sqrt{\frac{1}{(\Sigma A_i)^2}\cdot\{N\cdot shi^2+2\cdot(\frac{100\cdot N}{N-1})^2\cdot[(N-1)-\sum_{i=1}^{N-2} sign(A_i-A_{i+1})\cdot sign(A_{i+1}-A_{i+2})]\}} \tag{3.9}$$

$$u(jit) = u(T) \cdot C(T,N) \tag{3.10}$$

where $u(T)$ is the period evaluation uncertainty and $C(T,N)$ is the sensitivity coefficient of *jit* with respect to T, which depends on the period T and the number N of processed periods. With reference to the sensitivity coefficient $C(T,N)$, particular attention is paid towards the term:

$$SUM_{sgn} = \sum_{i=1}^{N-2} sign(T_i - T_{i+1}) \cdot sign(T_{i+1} - T_{i+2}) \tag{3.11}$$

Such a term depends on the temporal evolution of consecutive periods and its statistical evaluation was performed on the balanced dataset described in Sec. 2.2 to evaluate its order of magnitude. In particular, it was evaluated as a fraction of the number of consecutive pseudo-periods couples (N-2). This choice is justified considering the extremal case of a strictly rising or falling periods sequence, where the product of the *sign* functions is 1 so that $SUM_{sgn} = (N-2)$.

$$F_N = SUM_{sgn}/(N-2) \tag{3.12}$$

as shown in Fig. 3.2 F_N can vary from -0.5 and 0.3 with a median value of -0.2.

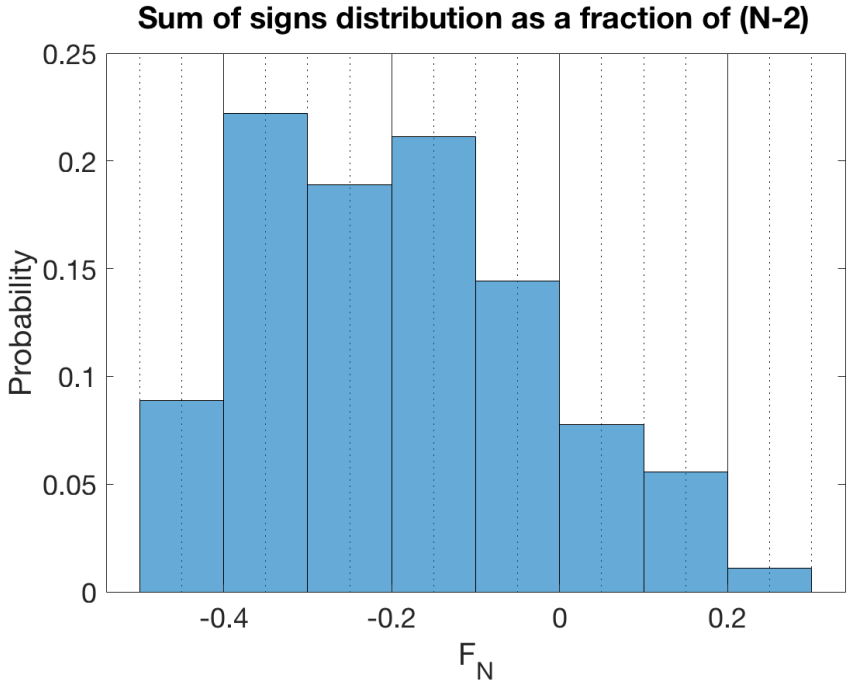


Fig. 3.2 Distribution of the term SUM_{sgn} normalized respect to (N-2)

Substituting the evaluated median value of F_N and assuming also a mean period duration $\overline{T} = \sum T_i/N$, the term $(\sum T_i)^2$ becomes $(N \cdot \overline{T})^2$ so the Eq. 3.9 becomes a function of $N, \overline{T}$, jit and F_N:

$$u(jit) = u(T) \cdot \sqrt{\frac{1}{(N \cdot \overline{T})^2} \cdot \{N \cdot jit^2 + 2 \cdot (\frac{100 \cdot N}{N-1})^2 \cdot [(N-1) - (N-2) \cdot F_N]\}} \quad (3.13)$$

Assuming a vowel with a fundamental frequency of 100 Hz and a variable duration from 1 to 10 s, a jitter in the range from 0.1 & to 5 % and $F_N = -0.2$, the sensitivity coefficient depends just on jit and N ($C(T,N) \rightarrow C(jit,N)$) and the representation of such term can be depicted as in the heatmap in Fig. 3.3

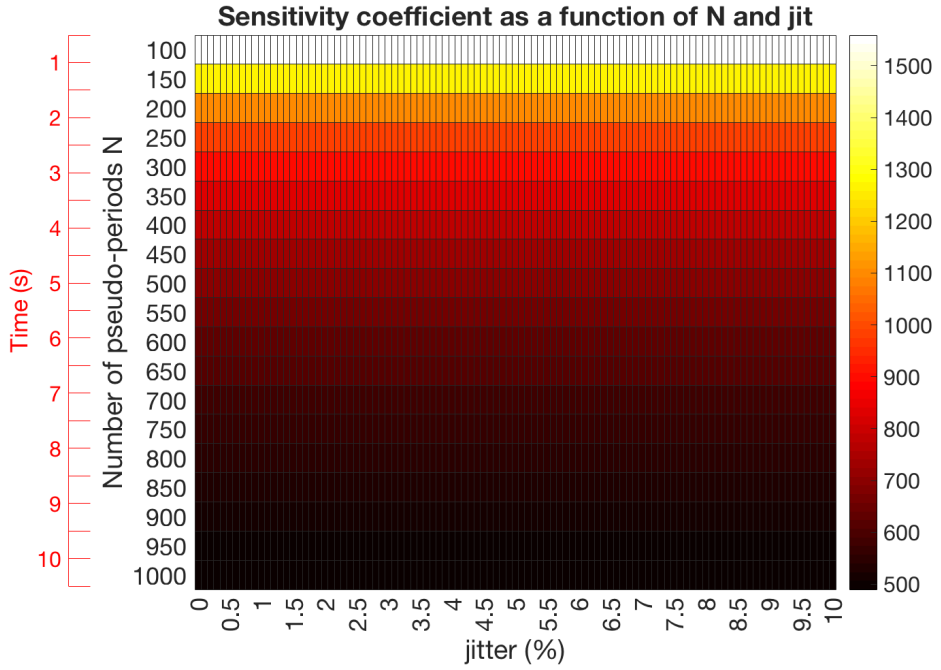


Fig. 3.3 Sensitivity coefficient of the analytical evaluation of jitter uncertainty.

As shown in Fig. 3.3, the sensitivity factor is lower for larger amounts of collected pseudo-periods and a weakest dependence on the evaluated jitter can be noticed.

To evaluate the order of magnitude of jitter uncertainty, the period measurement uncertainty $u(T)$, evaluated in the previous section is considered. In this example, with a sampling rate $sr = 44100$ Sa/s, $u(T)$ is in the order of $2.6 \cdot 10^{-5}$ s. If a 100 Hz vowel with a 5 s duration and a jitter of 5 % is considered, the sensitivity coefficient is around 700 (%/s). Multiplying such a sensitivity coefficient by $u(T)$, the uncertainty of jitter measurements is $u(jit) = 0.018$ %. For the same vowel example, substituting the minimum and maximum values of the evaluated F_N (-0.5 and 0.3), the respective propagated uncertainties are 0.02 % and 0.014 % . If such values are compared to the scale of jitter measurements extracted from the dataset defined in Sec. 2.2, which ranges from 0.12 % to 5.35 %, their relative uncertainties spans from 0.3 % to 16 %. For shimmer measurements (*shi*), thanks to the fact that Eq. 2.5 has the same mathematical structure of *jit* (Eq. 2.4), the same considerations can be made substituting the periods measurements T_i with peak-to-peak amplitude measurements A_i. To evaluate the influence of F_N on amplitude sequences, the same evaluation on our dataset was performed, obtaining the result that is summarized in Fig. 3.4.

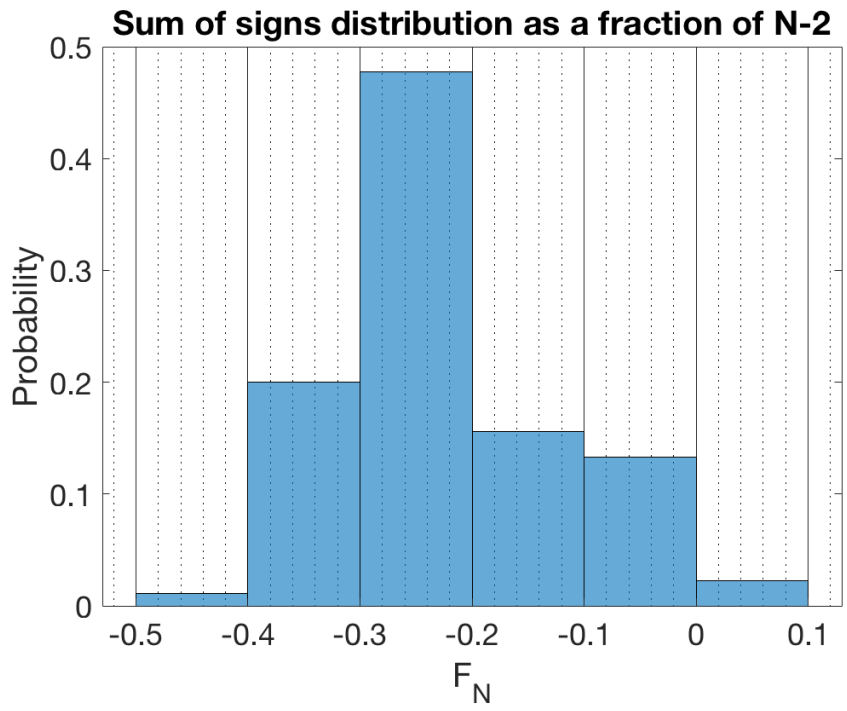


Fig. 3.4 F_N distribution for shimmer

As shown in Fig. 3.4, the F_N term has a median value of -0.23 with a more prominent peak relative to the period duration evaluations. In order to evaluate the sensitivity coefficients for shimmer measurements, a mean constant value A=1 a.u. (half of the full scale range) was set for the amplitude measurements. In this way, the term $(\sum A_i)^2$ becomes $(N \cdot \overline{A})^2$. In this way is possible to evaluate the sensitivity coefficient $C(shi,N)$ and the representation of such term can be depicted as in the heatmap in Fig. 3.5

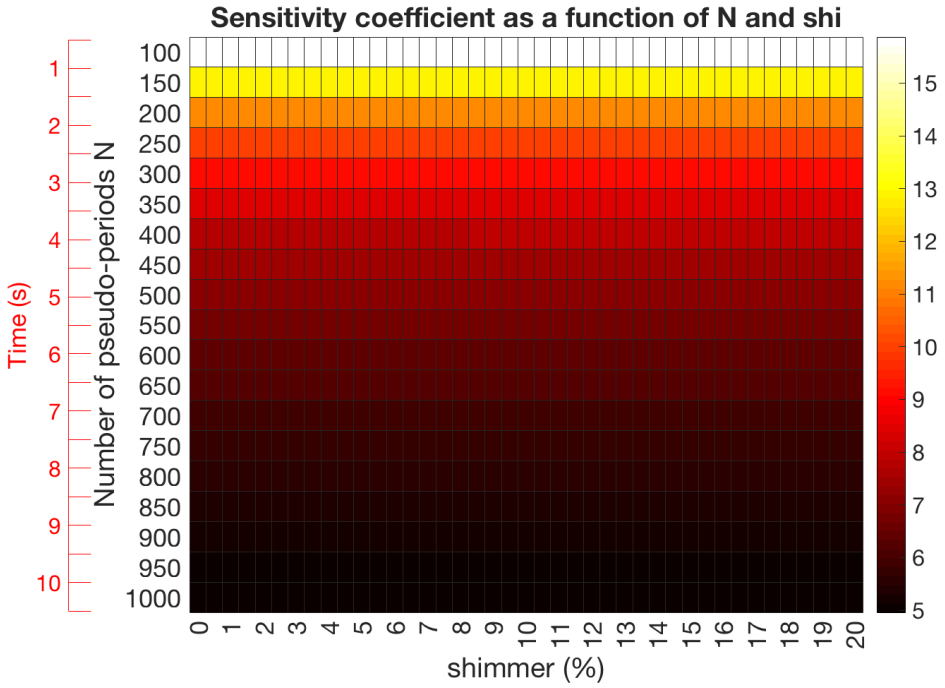


Fig. 3.5 Sensitivity coefficient heatmap for shimmer

The heatmap in Fig. 3.5 highlights that the sensitivity coefficient has a small dependence on N and is almost insensitive with respect to the parameter *shi*, though the sensitivity coefficient values are lower if compared to the jitter sensitivity coefficient. For *shi* measurements, an uncertainty contribution is the amplitude resolution, which affects the evaluation of peak-to-peak amplitudes. As showed for the period quantisation contribution, a uniform random perturbation of ± 1 LSB can affect the amplitude evaluations. Another amplitude uncertainty contribution is the Integral Nonlinearity (INL), which for a medium quality ADC is stated by the manufacturer as high as ± 2 LSB on the instantaneous amplitude measurement. The third contribution is the gain error, which for a medium level ADC is stated as the 0.05 % of the full-scale. To express each uncertainty contributions in LSB, an ADC with 16 bit of resolution is considered. A peak normalization is applied to the input signals in order to measure a maximum absolute amplitude of 1 a.u. If a signal with an average peak-to-peak amplitude of 1 a.u. (i.e. 0.5 a.u. peak amplitude) is considered, the main amplitude uncertainty are:

- Quantisation: $LSB_Q = \pm 1$ LSB

- Integral Nonlinearity: $LSB_{INL} = \pm 2$ LSB
- Gain Error: $LSB_{GE} = GE(\%) \cdot S_{in} \cdot 2^{N_b - 1} = 0.0005 \cdot 0.5 \cdot 2^{15} = \pm 8$ LSB

where LSB_Q, LSB_{INL} and LSB_{GE} are respectively the quantisation, INL and GE contributions expressed in LSB. $GE(\%)$ is the gain error and S_{in} is the signal amplitude expressed in a.u.. These contribution are supposed to have an uniform distribution extended between the LSB limits listed above, so the quadratic sum of each contribution is performed using the following equation:

$$u_{LSB}(A) = \sqrt{2 \cdot (\frac{LSB_Q}{\sqrt{12}})^2 + 2 \cdot (\frac{LSB_{INL}}{\sqrt{12}})^2 + 2 \cdot (\frac{LSB_{GE}}{\sqrt{12}})^2} \approx 2.5 \text{ LSB} \qquad (3.14)$$

the quadratic terms are multiplied by 2 to consider that the peak-to-peak amplitude measurement is the difference between two instantaneous signal codes. Considering this specific example, the uncertainty of amplitude measurements is $u(A) = u_{LSB}(A)/2^{16} \approx 3.8 \cdot 10^{-5}$ a.u.. As already done in the previous case, the shimmer uncertainty can be evaluated considering a vowel with a fundamental frequency of 100 Hz and 5 seconds duration. For a 10 % *shi*, the corresponding sensitivity coefficient is close to 7 (%/a.u.). Multiplying this term by the amplitude uncertainty leads to a shimmer uncertainty $u(shi) = 0.0004$ %. Such a value is negligible if compared to the scale of shimmer measurement extracted from the dataset defined in Sec. 2.2, which ranges from 1.14 % to 18.5 %. Such values lead to a relative uncertainty ranging from 0.002 % to 0.03 %. This example is to be considered as a typical application, even though ADC with lower performances ratings exist. If we consider an ADC with $GE(\%) = 0.5$ % then $u_{LSB}(A) \approx 24$ LSB, $u_(shi) \approx 0.002\%$ and the range of relative uncertainties on shimmer measurements spans between 0.01 % and 0.035 %.

The same analytical procedure can be adopted for the uncertainty evaluations of the other stability metrics defined in Appendix A (rap, ppq, apq). Due to their nested mathematical architecture, it proved to be very challenging to propagate through the analytical method. Therefore, a numeric Monte Carlo Error propagation was carried out to solve this problem.

3.2 Monte Carlo uncertainty propagation

To evaluate the effect of perturbations on period and amplitude stability measurements, a Monte Carlo error propagation was performed. Particular attention has been paid to the uncertainty contribution caused by the quantisation of time and amplitude measurements. The method is based on the perturbation of extracted periods and amplitudes with random vectors with specific statistical distributions. In the previous section it was highlighted that the main uncertainty contribution for time stability measurements is the quantisation of the time-base. For this reason, to evaluate the effects of quantisation noise on period estimations, a random uniform variable between -1 and 1 was added to each pseudo period evaluation expressed as counted samples. The uncertainty of amplitude measurements is affected by the quantisation of the amplitude scale and other effect such as the Integral Non Linearity (INL) and the Gain Error, as shown in the previous section. To evaluate the effects of such perturbations on amplitude estimations, a random uniform variable between -2.5 LSB and 2.5 LSB was added to each pseudo amplitude evaluation. Each perturbed period and amplitude sequence was transformed in the stability metrics described in Appendix 1A. This procedure was repeated 10^6 times to produce an array of perturbed values for each stability metrics and the statistical distributions of such arrays were extracted. As an example, the statistical distribution of the perturbed *shi* obtained from two vowels uttered by the same subject is depicted in Fig. 3.6.

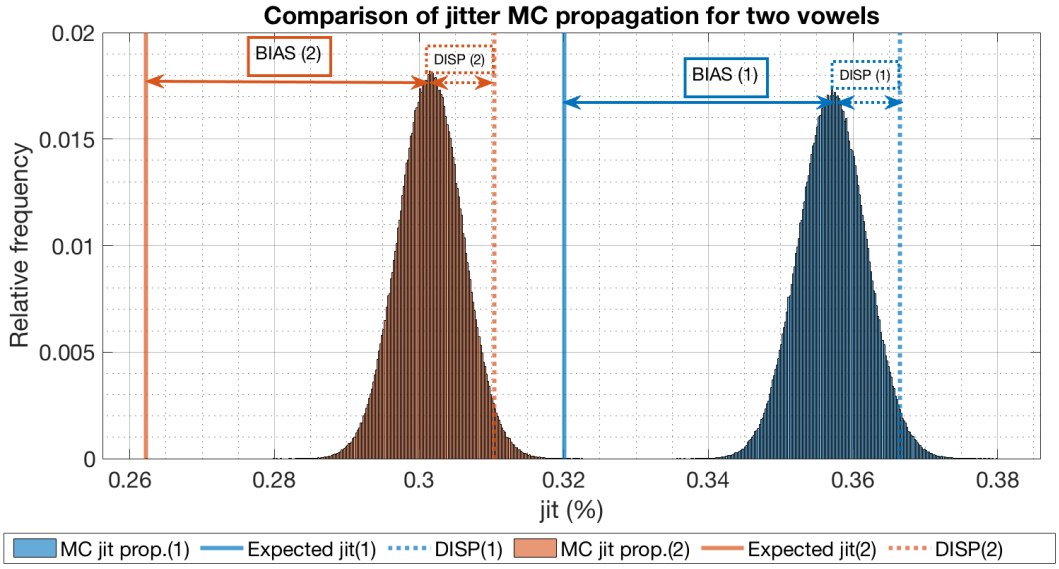


Fig. 3.6 Monte Carlo uncertainty propagation of jitter for two vowels emitted by the same subject

In Fig. 3.6 the blue and orange distributions are respectively relative to the first and second extracted vowels. The vertical solid lines represent the expected jitter values obtained from the un-perturbed pseudo-period sequences and the dotted lines represent the dispersion of the perturbed distributions expressed by the equation:

$$DISP = c(95\ \%) \cdot std(F_{MC}) \tag{3.15}$$

where $std(F_{MC}(i))$ is the standard deviation of the Monte Carlo simulation distributions, $c(95\%)$ is a coverage factor calculated as the t-student inverse with a confidence limit of 95%. As shown in Fig. 3.6, BIAS (1) and BIAS (2) were evaluated as the distance between expected value and the mean value of the perturbed distributions:

$$BIAS = \overline{F_{MC}} - F_{exp} \tag{3.16}$$

where $\overline{F_{MC}}$ is the feature mean value calculated from the Monte Carlo simulation and F_{exp} is the feature expected value. For the parameters where the absolute value operator is present in the definition (jit, jit_{abs}, rap, ppq, shi, shi_{abs}, apq), the Monte Carlo evaluations showed a significative bias respect to the expected (unperturbed) value. Such an effect is caused by the presence of the absolute value in the features formulae, which is responsible for a strictly positive accumulation of consecutive

differences. This assumption suggests that a bias is always present in the evaluation when the input variables are perturbed. A further analysis on the available dataset showed that this bias is not constant neither among different subjects nor among different vowels uttered by the same subject. For this reason, such a bias cannot be considered a systematic effect and then it has to be combined to the other uncertainty contributions. To take into account the measurement bias, the following relation was used to evaluate the measurement uncertainty with a 95% confidence level:

$$\overline{U_{F(bias)}(class)} = \frac{\sum_{i=1}^{N_s} \sqrt{BIAS(i)^2 + DISP(i)^2}}{N_s} \tag{3.17}$$

where N_s is the number of subjects of each class. In a similar way, a Mean dispersion parameter was defined as:

$$\overline{DISP_F(class)} = \frac{\sum_{i=1}^{N_s} DISP(i)}{N_s} \tag{3.18}$$

Such metrics were evaluated in order to separate the bias contribution from the dispersion contribution and to compare their order of magnitude.

3.2.1 Oversampling effect

The stability metrics were extracted from recordings with a sampling rate of 44100 Sa/s and a bit resolution of 16 bit. To simulate the effect of different sampling rates on the feature uncertainties, a linear oversampling of the audio signal was applied using the Matlab function *resample* [26]. The *resample* function performs a linear interpolation of the original signal and applies a FIR antialiasing filter to compensate the high frequency effects caused by the interpolation process. the The Monte Carlo error propagation was performed for four different oversampling factors:

- 1 (44100 Sa/s)
- 2 (88200 Sa/s)
- 4 (176400 Sa/s)
- 8 (352800 Sa/s)

The table 3.1 summarises the mean uncertainties, calculated using Eq. 3.17 and 3.18, for each stability metrics and each clinical class.

Table 3.1 Oversampling effect on stability metrics dispersions for different oversampling factors

$\overline{DISP_F(class)}$ **- oversampling=1 (bit resolution=16bit)**									
Class	*jit* (%)	jit_{abs} μs	rap (%)	ppq (%)	vf_o (%)	SHI (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$2.4{\cdot}10^{-2}$	1.5	$1.6{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$	$1.8{\cdot}10^{-2}$	$5.7{\cdot}10^{-4}$	$5.4{\cdot}10^{-5}$	$3.7{\cdot}10^{-4}$	$2.4{\cdot}10^{-2}$
HE	$1.5{\cdot}10^{-2}$	$9.1{\cdot}10^{-1}$	$1.0{\cdot}10^{-2}$	$9.6{\cdot}10^{-3}$	$1.3{\cdot}10^{-2}$	$4.0{\cdot}10^{-4}$	$3.7{\cdot}10^{-5}$	$2.5{\cdot}10^{-4}$	$1.5{\cdot}10^{-2}$
PA	$2.9{\cdot}10^{-2}$	1.4	$1.9{\cdot}10^{-2}$	$1.8{\cdot}10^{-2}$	$2.1{\cdot}10^{-2}$	$6.3{\cdot}10^{-4}$	$5.8{\cdot}10^{-5}$	$3.9{\cdot}10^{-4}$	$2.9{\cdot}10^{-2}$
$\overline{DISP_F(class)}$ **- oversampling=2 (bit resolution=16bit)**									
Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_o (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$1.3{\cdot}10^{-2}$	$7.9{\cdot}10^{-1}$	$8.6{\cdot}10^{-3}$	$8.0{\cdot}10^{-3}$	$8.9{\cdot}10^{-3}$	$5.7{\cdot}10^{-4}$	$5.4{\cdot}10^{-5}$	$3.7{\cdot}10^{-4}$	$3.8{\cdot}10^{-4}$
HE	$8.2{\cdot}10^{-3}$	$4.8{\cdot}10^{-1}$	$5.7{\cdot}10^{-3}$	$5.2{\cdot}10^{-3}$	$6.4{\cdot}10^{-3}$	$4.0{\cdot}10^{-4}$	$3.7{\cdot}10^{-5}$	$2.5{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$
PA	$1.6{\cdot}10^{-2}$	$7.9{\cdot}10^{-1}$	$1.0{\cdot}10^{-2}$	$9.6{\cdot}10^{-3}$	$1.1{\cdot}10^{-2}$	$6.3{\cdot}10^{-4}$	$5.8{\cdot}10^{-5}$	$3.9{\cdot}10^{-4}$	$4.0{\cdot}10^{-4}$
$\overline{DISP_F(class)}$ **- oversampling=4 (bit resolution=16bit)**									
Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_o (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$6.6{\cdot}10^{-3}$	$4.1{\cdot}10^{-1}$	$4.5{\cdot}10^{-3}$	$4.1{\cdot}10^{-3}$	$4.4{\cdot}10^{-3}$	$5.7{\cdot}10^{-4}$	$5.4{\cdot}10^{-5}$	$3.7{\cdot}10^{-4}$	$3.8{\cdot}10^{-4}$
HE	$4.4{\cdot}10^{-3}$	$2.6{\cdot}10^{-1}$	$3.0{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$3.2{\cdot}10^{-3}$	$4.0{\cdot}10^{-4}$	$3.7{\cdot}10^{-5}$	$2.5{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$
PA	$8.2{\cdot}10^{-3}$	$4.1{\cdot}10^{-1}$	$5.5{\cdot}10^{-3}$	$5.1{\cdot}10^{-3}$	$5.4{\cdot}10^{-3}$	$6.3{\cdot}10^{-4}$	$5.8{\cdot}10^{-5}$	$3.9{\cdot}10^{-4}$	$4.0{\cdot}10^{-4}$
$\overline{DISP_F(class)}$ **- oversampling=8 (bit resolution=16bit)**									
Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_o (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$3.3{\cdot}10^{-3}$	$2.1{\cdot}10^{-1}$	$2.3{\cdot}10^{-3}$	$2.1{\cdot}10^{-3}$	$2.2{\cdot}10^{-3}$	$5.7{\cdot}10^{-4}$	$5.4{\cdot}10^{-5}$	$3.7{\cdot}10^{-4}$	$3.8{\cdot}10^{-4}$
HE	$2.3{\cdot}10^{-3}$	$1.3{\cdot}10^{-1}$	$1.6{\cdot}10^{-3}$	$1.5{\cdot}10^{-3}$	$1.6{\cdot}10^{-3}$	$4.0{\cdot}10^{-4}$	$3.7{\cdot}10^{-5}$	$2.5{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$
PA	$4.2{\cdot}10^{-3}$	$2.1{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$2.7{\cdot}10^{-3}$	$6.3{\cdot}10^{-4}$	$5.8{\cdot}10^{-5}$	$3.9{\cdot}10^{-4}$	$4.0{\cdot}10^{-4}$

As shown in Table 3.1, the amplitude related measurements are not affected by the oversampling factor, as expected. The period related measurements dispersions seem to consistently decrease as the oversampling factor increase. In Fig. 3.7 an example of Mean absolute uncertainties (Eq. 3.17) and dispersions (Eq. 3.18) of jitter and shimmer measurements is shown.

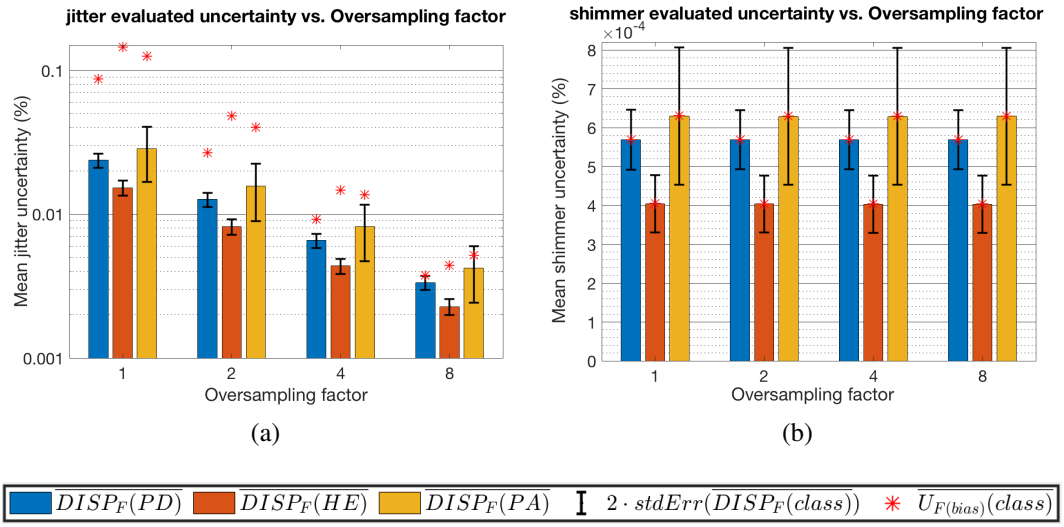


Fig. 3.7 Oversampling effect on jitter (a) and shimmer (b) evaluations

The bars in the plots of Fig. 3.7 represent the dispersion $\overline{DISP_F}$ of jitter and shimmer for the three classes (PD in blue, HE in red, PA in yellow). The red asterisks represent the extended uncertainties $\overline{U_{jit(bias)}}$ evaluated using Eq. 3.17. As shown in the plot, the distance between $\overline{U_{jit(bias)}}$ and the $\overline{DISP_{jit}}$ decreases as the oversampling rises, highlighting that the *BIAS* contribution on Eq. 3.17 have less influence on time stability metrics for higher oversampling factors. To evaluate the effect of the bias on jitter and shimmer measurements, the Tab. 3.2 reports the differences between the average expected jitter and shimmer and the average of the evaluated ones obtained as the mean values of the Monte Carlo simulation ($\overline{F_{MC}}$).

Table 3.2 Oversampling effect on Expected jitter and shimmer compared to the Evaluated measurements

oversampling=1 (bit resolution=16bit)				
Class	Exp. *jit* (%)	Eval. *jit* (%)	Exp. *shi* (%)	Eval. *shi* (%)
PD	0.62	0.71	5.75	5.75
HE	0.32	0.47	2.8	2.8
PA	0.44	0.56	4.19	4.19
oversampling=2 (bit resolution=16bit)				
Class	Exp. *jit* (%)	Eval. *jit* (%)	Exp. *shi* (%)	Eval. *shi* (%)
PD	0.64	0.66	5.79	5.79
HE	0.31	0.36	2.8	2.8
PA	0.44	0.48	4.18	4.18
oversampling=4 (bit resolution=16bit)				
Class	Exp. *jit* (%)	Eval. *jit* (%)	Exp. *shi* (%)	Eval. *shi* (%)
PD	0.63	0.64	5.83	5.83
HE	0.32	0.34	2.81	2.81
PA	0.44	0.45	4.18	4.18
oversampling=8 (bit resolution=16bit)				
Class	Exp. *jit* (%)	Eval. *jit* (%)	Exp. *shi* (%)	Eval. *shi* (%)
PD	0.64	0.64	5.82	5.82
HE	0.32	0.33	2.81	2.81
PA	0.44	0.44	4.18	4.18

As shown in Tab. 3.2, regarding shimmer measurements, no differences can be noted between expected and evaluated shimmer. For a sampling rate of 44100 Sa/s, a common value for consumer level recording devices, the difference between the expected and evaluated *jit* is not negligible and it can produce relative uncertainties up to 46 % for the HE class. As reported in Tab. 3.2, this difference is reduced for higher oversampling factors. From this considerations it is evident that sampling the signal at higher sampling rates or post process it with a linear oversampling can noticeably reduce the uncertainties on stability metrics. Such a practice can be very computationally expensive and not efficient if implemented in embedded and portable devices. Another way to reduce the stability metrics uncertainty may be the implementation of a compensation method to correct such effects for lower

oversampling factors (or lower sampling rates). As an example, in Fig. 3.8 the evaluated jitter, respect to the expected values using two different oversampling factor (1, 8), is presented.

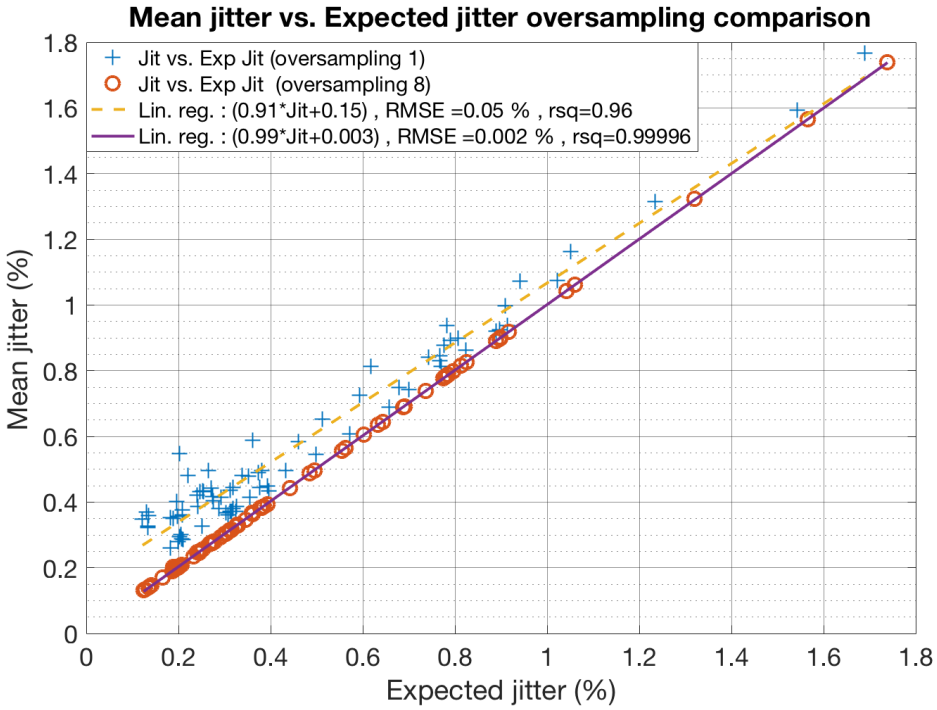


Fig. 3.8 Comparison between two oversampling factors for Expected versus evaluated jitter. The dashed and solid lines represent the linear regressions of the experimental points respectively for an oversampling factor of 1 and 8

Fig. 3.8 highlights a clear difference between different oversampling factors. The linear regression of The data extracted using an oversampling factor of 8 is clearly more linear than the data extracted without the oversampling of the signal. A significative difference between the bias of the compared oversampling factors can be noticed in the regression without oversampling. Moreover, the slope of the regression line without oversampling is significatively lower than 1 and this fact suggests that without any correction the jitter evaluations will be overestimated, above all for low jitter values. To correct this issue, the inverse of the linear regression with no oversampling can be calculated. For this particular example the correction equation would be:

$$jit = \frac{jit^* - o}{m} = \frac{jit^* - 0.15}{0.91} \tag{3.19}$$

where jit is the corrected jitter, jit^* is the evaluated one, m and o are respectively the slope and offset of the linear regression. Such an operation could correct the effect of low sampling rates at low jitter values, but the residual contribution that is related to the identification of the coefficients m and o of by the linear regression have to be considered in the uncertainty budget.

The same considerations can be made for each measurement where the absolute value is present (jit, jit_{abs}, rap, ppq).

3.2.2 Amplitude resolution effect

To evaluate the effect of amplitude resolution on the stability metrics, the signals were bit-reduced using the Matlab function *uencode* [27] which is capable of vertically resampling a signal to a given number of bits. The average uncertainties using Eqs.3.17 and 3.18 for each feature are summarized in Tab. 3.3. For this characterisation the oversampling factor was set to 8.

Table 3.3 Bit resolution effect on stability metrics dispersions for different number of bits

$\overline{DISP_F(class)}$ **- resolution=10 bit (oversampling=8)**									
Class	jit (%)	jit_{abs} (µs)	rap (%)	ppq (%)	vf_o (%)	shi (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$3.8{\cdot}10^{-3}$	$2.3{\cdot}10^{-1}$	$2.5{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$2.2{\cdot}10^{-3}$	$5.5{\cdot}10^{-2}$	$5.3{\cdot}10^{-3}$	$2.5{\cdot}10^{-2}$	$2.5{\cdot}10^{-2}$
HE	$4.4{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$9.1{\cdot}10^{-2}$	$8.4{\cdot}10^{-3}$	$2.1{\cdot}10^{-2}$	$1.7{\cdot}10^{-2}$
PA	$5.2{\cdot}10^{-3}$	$2.7{\cdot}10^{-1}$	$3.4{\cdot}10^{-3}$	$3.1{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$8.3{\cdot}10^{-2}$	$7.6{\cdot}10^{-3}$	$2.8{\cdot}10^{-2}$	$2.6{\cdot}10^{-2}$
$\overline{DISP_F(class)}$ **- resolution=12 bit (oversampling=8)**									
Class	jit (%)	jit_{abs} (µs)	rap (%)	ppq (%)	vf_o (%)	shi (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$3.8{\cdot}10^{-3}$	$2.3{\cdot}10^{-1}$	$2.5{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$9.8{\cdot}10^{-3}$	$9.4{\cdot}10^{-4}$	$5.9{\cdot}10^{-3}$	$6.1{\cdot}10^{-3}$
HE	$4.4{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$9.2{\cdot}10^{-3}$	$8.5{\cdot}10^{-4}$	$4.1{\cdot}10^{-3}$	$4.3{\cdot}10^{-3}$
PA	$5.2{\cdot}10^{-3}$	$2.7{\cdot}10^{-1}$	$3.4{\cdot}10^{-3}$	$3.2{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$1.2{\cdot}10^{-2}$	$1.1{\cdot}10^{-3}$	$6.3{\cdot}10^{-3}$	$6.5{\cdot}10^{-3}$
$\overline{DISP_F(class)}$ **- resolution=16 bit (oversampling=8)**									
Class	jit (%)	jit_{abs} (µs)	rap (%)	ppq (%)	vf_o (%)	shi (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$3.8{\cdot}10^{-3}$	$2.3{\cdot}10^{-1}$	$2.5{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$5.7{\cdot}10^{-4}$	$5.4{\cdot}10^{-5}$	$3.7{\cdot}10^{-4}$	$3.8{\cdot}10^{-4}$
HE	$4.4{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$4.0{\cdot}10^{-4}$	$3.7{\cdot}10^{-5}$	$2.5{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$
PA	$5.2{\cdot}10^{-3}$	$2.6{\cdot}10^{-1}$	$3.4{\cdot}10^{-3}$	$3.2{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$6.3{\cdot}10^{-4}$	$5.8{\cdot}10^{-5}$	$3.9{\cdot}10^{-4}$	$4.0{\cdot}10^{-4}$

The jit and shi evaluated uncertainties presented in Tab 3.3 are depicted in Fig. 3.9:

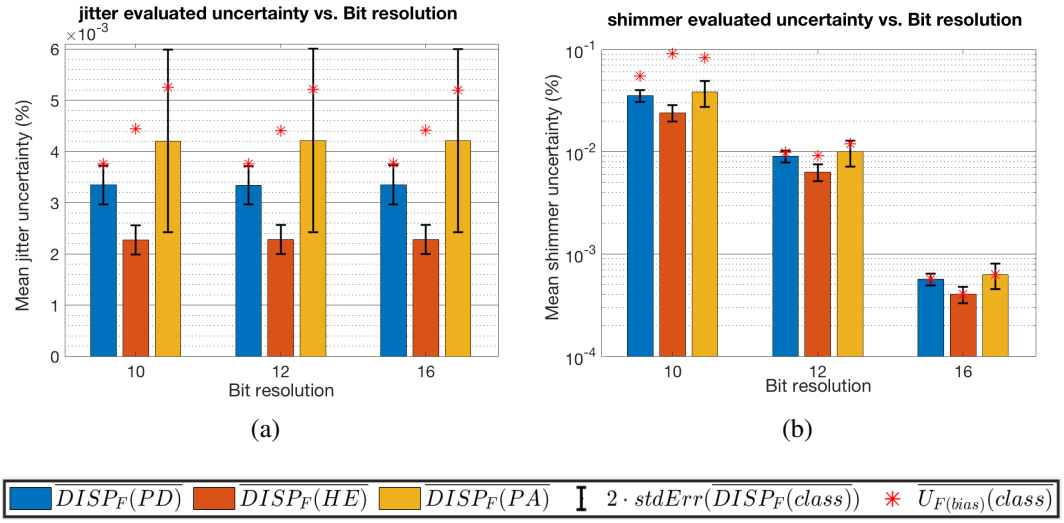


Fig. 3.9 Bit resolution effect on jitter (a) and shimmer (b) evaluations

As shown in Fig. 3.9, the amplitude resolution does not affect the jitter evaluations while it has an evident influence on shimmer evaluations. As an example, the correction method proposed for jitter evaluations (Fig. 3.8) can be applied to the shimmer evaluations as shown in Fig. 3.10.

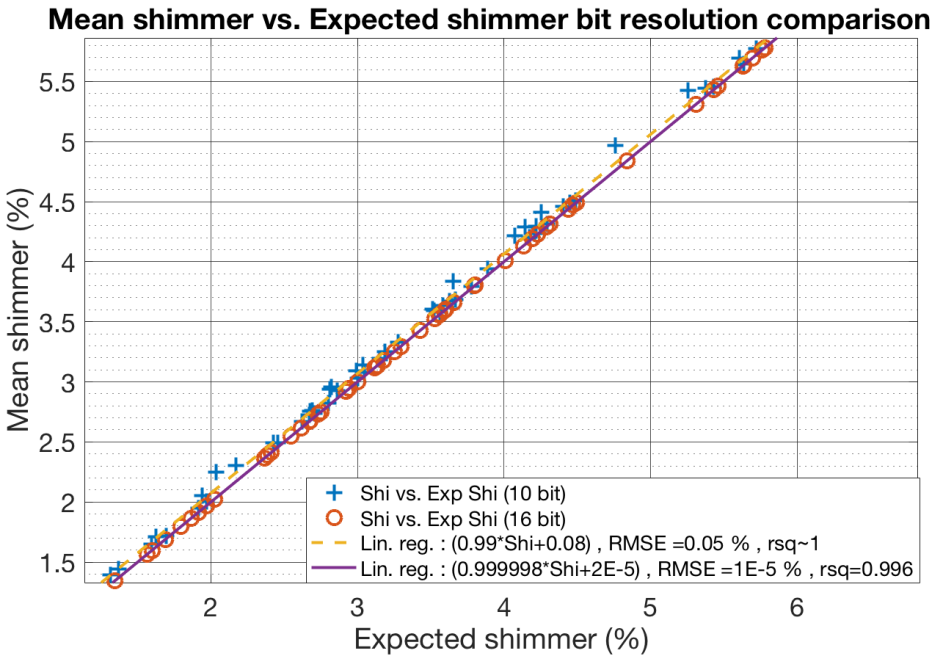


Fig. 3.10 Comparison between two amplitude resolutions for expected versus evaluated shimmer. The dashed and solid lines represent the linear regressions of the experimental points respectively for an amplitude resolution of 10 bit and 16 bit

As shown in Fig. 3.10, the shimmer correction is fairly unnecessary due to the enhanced linearity of the evaluated regression ($m \approx 1$, $o \approx 0$).

3.2.3 Background noise effect

To evaluate the effect of noise on the extraction of voice features, a Gaussian white noise was added to the original acquired audio signals using the Matlab function *wgn* [28]. To give a target Noise to signal ratio (NSR) to the signals under test, the A_{RMS} of the original signal was evaluated to establish the noise level to add to the original signal using the following equation:

$$A_{RMS}(noise) = A_{RMS}(signal) \cdot 10^{\frac{NSR}{20}} \tag{3.20}$$

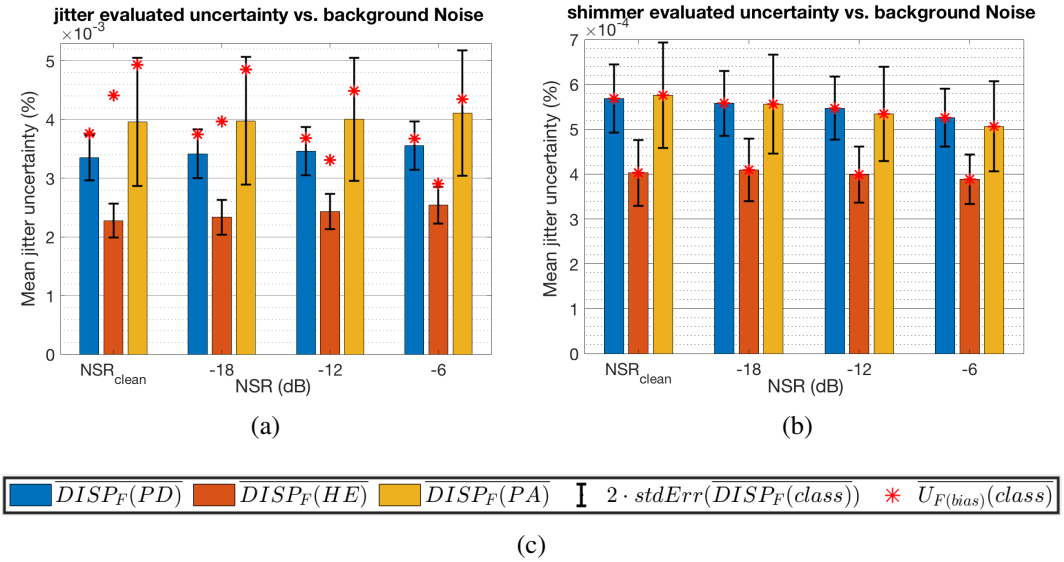
It is necessary to specify that for "clean signal" it is intended that the background noises levels (environmental and electronics noises), of the signals under test, should be lower than any noise level analysed in this experiment. As already stated in Sec.2.2, the original background SNR was estimated as high as 30 dB (NSR<-30

dB), so a NSR of -18 dB was chosen as minimum target NSR for the added noise. For the same reason the other test NSR were chosen as -12 dB and -6 dB. In table 3.4 the mean uncertainties, calculated using Eqs.3.17 and 3.18, for each stability metrics and each clinical class are presented.

Table 3.4 Noise effect on stability metrics dispersions for different NSR

$\overline{DISP_F(class)}$ **- NSR=NSRclean**									
Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_o (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$3.8{\cdot}10^{-3}$	$2.3{\cdot}10^{-1}$	$2.5{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$2.2{\cdot}10^{-3}$	$5.5{\cdot}10^{-2}$	$5.3{\cdot}10^{-3}$	$2.5{\cdot}10^{-2}$	$2.5{\cdot}10^{-2}$
HE	$4.4{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$9.1{\cdot}10^{-2}$	$8.4{\cdot}10^{-3}$	$2.1{\cdot}10^{-2}$	$1.7{\cdot}10^{-2}$
PA	$5.2{\cdot}10^{-3}$	$2.7{\cdot}10^{-1}$	$3.4{\cdot}10^{-3}$	$3.1{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$8.3{\cdot}10^{-2}$	$7.6{\cdot}10^{-3}$	$2.8{\cdot}10^{-2}$	$2.6{\cdot}10^{-2}$
$\overline{DISP_F(class)}$ **- NSR=-18 dB**									
Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_o (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$3.8{\cdot}10^{-3}$	$2.3{\cdot}10^{-1}$	$2.5{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$9.8{\cdot}10^{-3}$	$9.4{\cdot}10^{-4}$	$5.9{\cdot}10^{-3}$	$6.1{\cdot}10^{-3}$
HE	$4.4{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$9.2{\cdot}10^{-3}$	$8.5{\cdot}10^{-4}$	$4.1{\cdot}10^{-3}$	$4.3{\cdot}10^{-3}$
PA	$5.2{\cdot}10^{-3}$	$2.7{\cdot}10^{-1}$	$3.4{\cdot}10^{-3}$	$3.2{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$1.2{\cdot}10^{-2}$	$1.1{\cdot}10^{-3}$	$6.3{\cdot}10^{-3}$	$6.5{\cdot}10^{-3}$
$\overline{DISP_F(class)}$ **- NSR=-12 dB**									
Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_o (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$3.8{\cdot}10^{-3}$	$2.3{\cdot}10^{-1}$	$2.5{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$5.7{\cdot}10^{-4}$	$5.4{\cdot}10^{-5}$	$3.7{\cdot}10^{-4}$	$3.8{\cdot}10^{-4}$
HE	$4.4{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$4.0{\cdot}10^{-4}$	$3.7{\cdot}10^{-5}$	$2.5{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$
PA	$5.2{\cdot}10^{-3}$	$2.6{\cdot}10^{-1}$	$3.4{\cdot}10^{-3}$	$3.2{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$6.3{\cdot}10^{-4}$	$5.8{\cdot}10^{-5}$	$3.9{\cdot}10^{-4}$	$4.0{\cdot}10^{-4}$
$\overline{DISP_F(class)}$ **- NSR=-6 dB**									
Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_o (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	$3.8{\cdot}10^{-3}$	$2.3{\cdot}10^{-1}$	$2.5{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$2.3{\cdot}10^{-3}$	$5.7{\cdot}10^{-4}$	$5.4{\cdot}10^{-5}$	$3.7{\cdot}10^{-4}$	$3.8{\cdot}10^{-4}$
HE	$4.4{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$1.7{\cdot}10^{-3}$	$4.0{\cdot}10^{-4}$	$3.7{\cdot}10^{-5}$	$2.5{\cdot}10^{-4}$	$2.7{\cdot}10^{-4}$
PA	$5.2{\cdot}10^{-3}$	$2.6{\cdot}10^{-1}$	$3.4{\cdot}10^{-3}$	$3.2{\cdot}10^{-3}$	$2.8{\cdot}10^{-3}$	$6.3{\cdot}10^{-4}$	$5.8{\cdot}10^{-5}$	$3.9{\cdot}10^{-4}$	$4.0{\cdot}10^{-4}$

The results of this evaluation are depicted in Fig. 3.11, where the mean uncertainty for jitter and shimmer measurements is reported.



(a) (b)

(c)

Fig. 3.11 Noise effect on jitter (a) and shimmer (b) evaluations

As shown in Fig. 3.11, a negligible effect can be noticed in the jitter and shimmer evaluations. This is due to fact that the noise, in this case, is added to the signal and then the sequences of pseudo-periods and amplitudes are extracted. Perturbing these sequences with MC generations has no effect on the dispersions and biases of the evaluated metrics. To evaluate the effect of noise on the extraction of stability metrics, the contribution of the extraction algorithm to the total uncertainty has to be evaluated.

3.2.4 Extraction algorithm effect

To evaluate the contribution of the extraction algorithm to the total uncertainty, the mean absolute error was evaluated taking as reference measurements the features extracted from the signal with the lowest noise level:

$$F_{exp}(-18\text{dB}) = F_{exp}(-12\text{dB}) = F_{exp}(-6\text{dB}) = \mathbf{F_{exp}}(\mathbf{NSR_{clean}}) \tag{3.21}$$

In this way, the original signal without any added noise is taken as a "golden standard" for this work, substituting $\mathbf{F_{exp}}(\mathbf{NSR_{clean}})$ in Eq 3.17. Improving the recording technique and using better recording devices and algorithms could produce a more reliable reference signal, however the term "golden standard" used in this

section is intended as relative to this specific work. In Fig. 3.12, an example of the dispersions and $\overline{U_{jit}(bias)}$ for jitter and shimmer evaluation is presented. As in the previous cases the height of the bar represent the dispersions $\overline{DISP_F(class)}$ (Eq. 3.18) and the red asterisks represent the extended uncertainty $\overline{U_{F(bias)}(class)}$ (Eq. 3.17).

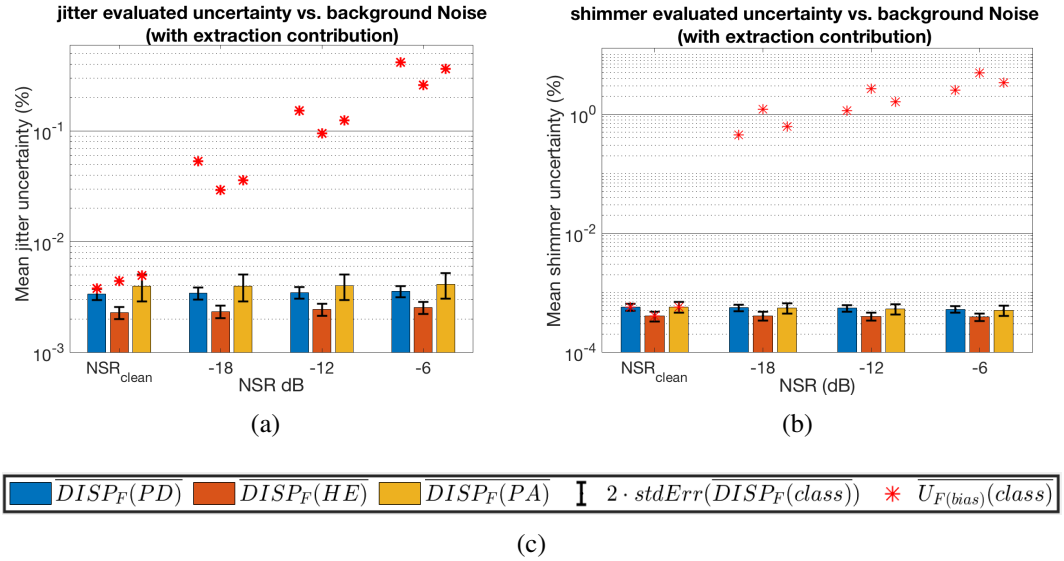


Fig. 3.12 Noise effect on jitter (a) and shimmer (b) evaluations considering the features extracted from the clean signals as golden standards

The plots in Fig. 3.12 shows clearly that the signals with the highest noise level (-6 dB) highlight higher uncertainties in the evaluation of jitter and shimmer. In particular, the contribution of the uncertainty associated to the bias looks predominant respect with the dispersions.

To get a global evaluation of the uncertainty including the contribution of the extraction algorithm, the Eq. 3.17 is calculated using as F_{exp} the one relative to the best measurement conditions in order to refer each measurement to a "golden standard", as already done for the noise contribution (Eq.3.21). In particular for the oversampling measurements:

$$F_{exp}(ovrsmp = 1) = F_{exp}(ovrsmp = 2) = F_{exp}(ovrsmp = 4) = \mathbf{F_{exp}(ovrsmp = 8)} \tag{3.22}$$

and for the bit reduction measurements:

$$F_{exp}(bit = 10) = F_{exp}(bit = 12) = \mathbf{F_{exp}(bit = 16)} \tag{3.23}$$

To get an overview of the extraction algorithm contribution, the Tab. 3.5 shows the error of each evaluated feature averaged along the three clinical classes. The golden standard for each evaluation is highlighted.

Table 3.5 Extraction algorithm contribution for different oversampling factors, bit resolutions and NSRs respect to golden standard measurements (highlighted in golden color)

Features:	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_o (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
Oversampling effect on $\overline{\mathbf{U_F(bias)}}$ respect to the golden standard (factor=8)									
8	$4.3{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$2.2{\cdot}10^{-3}$	$5.2{\cdot}10^{-4}$	$4.9{\cdot}10^{-5}$	$3.3{\cdot}10^{-4}$	$3.4{\cdot}10^{-4}$
4	$1.5{\cdot}10^{-2}$	$8.3{\cdot}10^{-1}$	$1.0{\cdot}10^{-2}$	$9.3{\cdot}10^{-3}$	$8.5{\cdot}10^{-3}$	$1.5{\cdot}10^{-2}$	$1.6{\cdot}10^{-3}$	$7.9{\cdot}10^{-3}$	$1.7{\cdot}10^{-2}$
2	$4.5{\cdot}10^{-2}$	2.5	$2.9{\cdot}10^{-2}$	$2.8{\cdot}10^{-2}$	$2.2{\cdot}10^{-2}$	$3.7{\cdot}10^{-2}$	$4.2{\cdot}10^{-3}$	$2.0{\cdot}10^{-2}$	$3.6{\cdot}10^{-2}$
1	$1.1{\cdot}10^{-1}$	6.4	$7.1{\cdot}10^{-2}$	$7.8{\cdot}10^{-2}$	$4.6{\cdot}10^{-2}$	$6.0{\cdot}10^{-2}$	$6.1{\cdot}10^{-3}$	$3.1{\cdot}10^{-2}$	$4.2{\cdot}10^{-2}$
Bit resolution effect on $\overline{\mathbf{U_F(bias)}}$ respect to the golden standard (bit=16)									
16	$4.3{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$2.2{\cdot}10^{-3}$	$5.2{\cdot}10^{-4}$	$4.9{\cdot}10^{-5}$	$3.3{\cdot}10^{-4}$	$3.4{\cdot}10^{-4}$
12	$6.6{\cdot}10^{-3}$	$3.7{\cdot}10^{-1}$	$4.4{\cdot}10^{-3}$	$3.9{\cdot}10^{-3}$	$5.4{\cdot}10^{-3}$	$1.7{\cdot}10^{-2}$	$1.7{\cdot}10^{-3}$	$1.3{\cdot}10^{-2}$	$1.5{\cdot}10^{-2}$
10	$8.8{\cdot}10^{-3}$	$4.7{\cdot}10^{-1}$	$5.9{\cdot}10^{-3}$	$5.5{\cdot}10^{-3}$	$8.1{\cdot}10^{-3}$	$1.7{\cdot}10^{-1}$	$1.6{\cdot}10^{-2}$	$1.3{\cdot}10^{-1}$	$6.4{\cdot}10^{-2}$
Noise effect on $\overline{\mathbf{U_F(bias)}}$ respect to the golden standard (NSR=NSRclean)									
NSRclean	$4.4{\cdot}10^{-3}$	$2.4{\cdot}10^{-1}$	$2.8{\cdot}10^{-3}$	$2.6{\cdot}10^{-3}$	$2.2{\cdot}10^{-3}$	$5.2{\cdot}10^{-4}$	$4.8{\cdot}10^{-5}$	$3.3{\cdot}10^{-4}$	$3.4{\cdot}10^{-4}$
NSR=-18 dB	$3.9{\cdot}10^{-2}$	2.1	$2.6{\cdot}10^{-2}$	$2.5{\cdot}10^{-2}$	$2.3{\cdot}10^{-2}$	$7.6{\cdot}10^{-1}$	$6.8{\cdot}10^{-2}$	$4.6{\cdot}10^{-1}$	1.6
NSR=-12 dB	$1.2{\cdot}10^{-1}$	6.6	$7.6{\cdot}10^{-2}$	$8.0{\cdot}10^{-2}$	$5.6{\cdot}10^{-2}$	1.8	$1.6{\cdot}10^{-1}$	1.1	3.2
NSR=-6 dB	$3.5{\cdot}10^{-1}$	$1.9{\cdot}10^{1}$	$2.1{\cdot}10^{-1}$	$2.2{\cdot}10^{-1}$	$1.3{\cdot}10^{-1}$	3.6	$3.1{\cdot}10^{-1}$	2.3	5.9

The data on Tab. 3.5 shows that the sampling rate (oversampling) and the vertical resolution can have a relevant effect on the evaluation of pseudo-periods and amplitudes stability metrics performed by the extraction algorithm. Such contribution does not affect the evaluation dispersion but just the bias contribution as seen in Fig. 3.12

3.3 Cross-talk contribution

The data used in the uncertainty evaluations of this work was collected using a portable audio recorder which captured the signals from two microphones: a contact throat microphone and a microphone in air. All the evaluations carried out in this manuscript have been performed on the signals from the microphone in air. It is possible that the signal coming from the contact microphone affected the signal from

the microphone in air because of the cross-talk effect. In multi-channel electronic acquisition devices, the cross-talk is an effect due to the internal circuitry of the device under test. To evaluate the possibility that, recording two signals at the same time the cross-talk can affect the acquisition, an experimental measurement was carried out. This evaluation was carried out as a side measurement of a larger experiment on the cross-talk effects on Digital AcQuisition devices (DAQ) , which resulted in the publication of a paper on the IEEE Transactions on Instrumentation and Measurement Journal [29]. The device used for this characterization is the ZOOM H2N, an handheld recording microphone which is capable of capturing the signal from up to four microphones in different configurations. In a common application the cross-talk effect for this device is not an issue since the embedded microphones capture an acoustic field which already has a natural signal "overlap" in the acoustic domain. For this reason a sound coming from the left is detected also in the right channel and the recorded signal will be always higher than any cross-talk voltage generated at the electronics level. For applications where the LINE IN input is used some issues can arise if different kind of devices are connected. The LINE IN input can accept line level signals (output impedance 600 Ω, Output voltage max. 1V) as well as electret microphone signals (output impedance 2.2 kΩ, average voltage 10 mV). To power the electret microphones , a pull-up resistor is connected to the right and left channels of the LINE IN input giving a constant voltage around 1.5 V. This kind of input configuration is commonly called plug-in. In this setup, a line level disturbance was applied to the Left channel to evaluate the effect of a large signal disturbance on Right channel. To simulate the presence of an electret microphone, a 2.2 kΩ resistor was connected between the Right channel and ground. Such a measurement was performed implementing the experimental architecture thath is shown in Fig. 3.13.

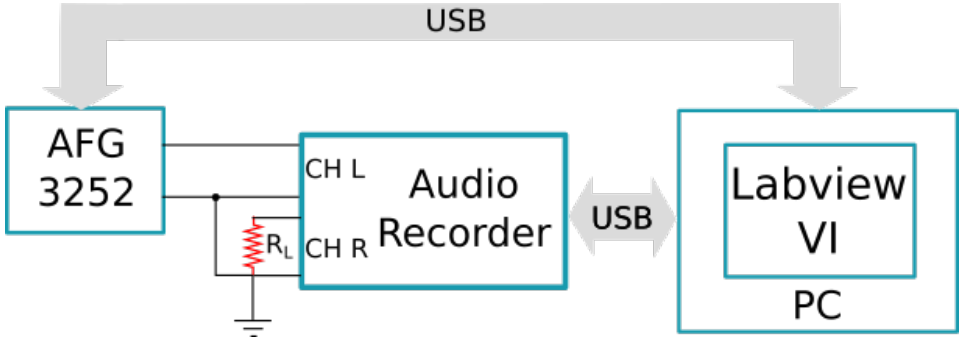


Fig. 3.13 Experimental setup of the cross-talk evaluation of an audio device

As shown in Fig. 3.13, an arbitrary function generator (Tektronix AFG-3252), which is controlled via USB interface by a LabVIEW Virtual Instrument (VI) running on a PC, is connected to the input Left channel of the audio recording system and configured to produce sinusoidal voltage signals at different frequencies. The voltage signal on the Left channel is set to have a peak-to-peak amplitude which produces a RMS full scale level of -6 dB_{fs} as indicated by the audio device display. The input channel R is connected to ground through the 2.2 kΩ resistor R_L.

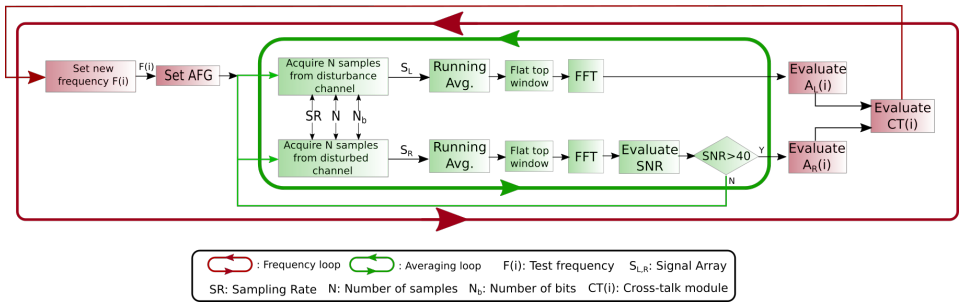


Fig. 3.14 Schematic of the LabVIEW script implemented to evaluate the cross-talk of the audio device

As shown in Fig. 3.14, the VI acquires signal samples from the two active channels and process them in order to estimate the disturbance source at the L channel and the error term at the R channel. The Sampling Rate SR of each channel is set to 44100 Sa/s. The collection of signals is performed by the LabVIEW Express VI *Acquire sound.vi* and the collected samples are processed with a flat-top window and then converted in the frequency domain through a Fast Fourier Transform (FFT) algorithm using the LabVIEW module *Amplitude and Phase Spectrum.vi*. In order to obtain a reliable measurement of the rms value, repeated frames are averaged in the time domain until a minimum Signal to Noise Ratio (SNR) of 40 dB is obtained. The SNR is estimated as the ratio between the rms at the tested frequency by the rms of all the other spectral components. The peak-to-peak amplitude values of the disturbance signal A_L and the error term A_R are evaluated as the FFT modules at the index corresponding to the tested frequency. The ratio between such rms values is calculated to estimate the cross-talk according to the equation:

$$CT = 20 \cdot log_{10}(\frac{A_R}{A_L}) \tag{3.24}$$

The evaluation of $CT(i)$ was carried out at different frequencies from 100 Hz to 22050 Hz with 100 Hz step. The evaluation of $CT(i)$ as a function of frequency produced the plot shown in Fig. 3.15

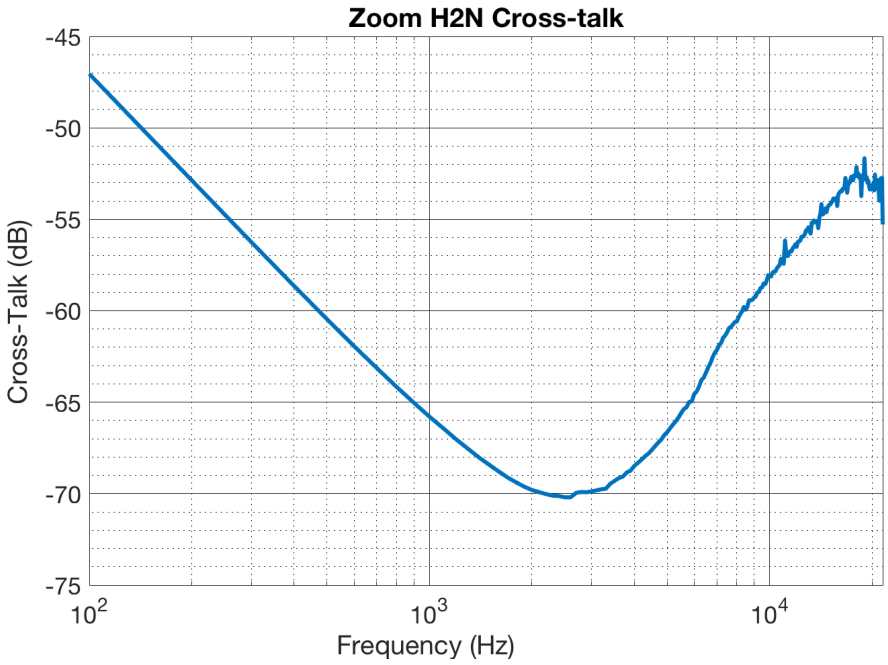


Fig. 3.15 Cross-talk evaluation of the audio device as a function of disturbance frequency

As shown in Fig. 3.15, the cross-talk can reach values as high as -50 dB at low frequencies while it is lower for higher frequencies. The average noise background level of the audio recordings was evaluated measuring an average SNR higher than 30 dB. The audio files were peak-normalised such that the background noise level is as low as -30 dB_{fs}. The maximum evaluated cross-talk ($\approx$-47 dB_{fs}) is lower than the background noise level, so no or negligible effect is expected when using both channel of the device under test.

3.4 Final considerations on the evaluation of the uncertainty of stability metrics

The presented section highlights the main uncertainty contributions in the extraction of vocal features commonly used in speech analysis and diagnosis.

3.4.1 Time and Amplitude contribution

The effects of time-base tolerance and aging on vocal features uncertainties have been investigated. Due to the relative architecture of jitter, a frequency drift over time does not affect the evaluation, thanks to the fact that the measured events (vowels), lasts lot less than a typical frequency drift that can become relevant after hours of operation. An extreme case scenario of a 100 ppm frequency tolerance on timing crystal has been analysed and the results showed that the period measurements are not significantly affected.

3.4.2 Analytical Error propagation:

A GUM oriented analysis on jitter uncertainty evaluation has been performed in order to evaluate the order of magnitude of the propagated uncertainty. This analysis highlighted the difficulties in the uncertainty propagation of period and amplitude stability features. In particular some terms such as SUM_{sgn} (Eq. 3.11) can not be analytically reduced so a statistical evaluation of such term using our dataset became necessary to evaluate the jitter and shimmer errors. To have a hint on the order of magnitude on *jit* and *shi* measurement, a common acquisition device with a sampling rate of 44100 Sa/s and an amplitude resolution of 16 bit has been considered. For the *jit* uncertainty, the main contribution was identified as the time-base quantisation which leads to a relative uncertainty between 0.3 % and 16 %. This consideration highlights that the sampling rate of the ADC plays an important role on the evaluation of *jit*. For the *shi* uncertainty, considering a medium quality ADC, relative uncertainties were estimated between. 0.004 % 0.07 %. This consideration suggests that the *shi* measure is less sensitive to the ADC amplitude characteristics (quantisation, INL, GE), as will be highlighted in the next conclusions.

3.4.3 Monte Carlo Error propagation of the quantization contribution

The Monte Carlo uncertainty propagation has been performed in order to evaluate the uncertainty contributions of the stability metrics, which can be challenging to propagate analytically. The uncertainty evaluation of the Monte Carlo propagation showed to be comparable to the analytical propagation. As an example, the quanti-

sation uncertainty contribution on jit, evaluated through the analytical method, is $u(jit)$ =0.018 %, which is comparable to the one obtained through the Monte Carlo simulation with oversampling=1 (0.015 %<$\overline{DISP_{jit}}$<0.029 %) as shown on Tab. 3.1 and Fig. 3.7a.
The Monte Carlo simulation showed an evident bias error in the evaluation of period stability parameters where the absolute value operator is present in their definition. Such a bias is not constant for every subjects and even for different vowels produced by the same subject. To compensate such a bias, the Monte Carlo simulations performed in this work seems to be the best option thanks to the fact that, this algorithm, showed good evaluation performances even using a relatively small number of trials.
A characterisation of the uncertainty contributions relative to a "golden standard" of the measurement was carried out. In particular, an example of jitter evaluation with a low oversampling factor (1) was characterised respect to the evaluation obtained with the maximum oversampling factor (8) through a linear regression of the experimental data. In a research application, where the execution time of the extraction algorithms is not a critical issue, the best practice should be to maximize the sampling parameters (sampling rate and bit resolution) in order to minimize the measurement uncertainty. For this work, the sampling parameters that minimize the stability metrics uncertainties were identified as:

- Sampling Rate $44100 \cdot 8 = 352800$ Sa/s
- Bit resolution 16 bit

In a more consumer oriented application, the computational power (and so the algorithm execution time) budget have to be balanced with the cost of the device. In this scenario, the sampling parameters could be lowered at the expense of a rising of measurement uncertainties. To take into account this uncertainties a correction method can be applied to the extracted features, as shown in Fig. 3.8, in order to develop a more computational efficient device using low cost components.
The effect of additive noise has been investigated and, as expected, it has negligible or no effect on the estimation of extracted features if the contribution of the extraction algorithm is considered as negligible. To evaluate how the extraction algorithm affects the stability metrics, an evaluation of measurement error respect to a golden standard (jitter and shimmer extracted from the clean signal) was performed. Such an evaluation showed enormous bias errors caused by the extraction algorithm which has proved to be very sensitive to the background noise level.

3.4.4 Cross-talk error contribution

As shown in the plot of Fig. 3.15, the cross-talk error contribution can affect the acquired signals depending on the evaluated frequency. The maximum cross-talk was evaluated at 100 Hz and is close to -47 dB_{fs} and the minimum value is reached around 2.5 kHz. The human voice, when a sustained vowel is emitted, show most of its spectral energy in a band which ranges from 100 Hz and 1 kHz. In such a range the evaluated cross-talk error is lower than the environmental background noise level of the recordings, which was estimated as low as -30 dB_{fs}. For this reason the cross-talk contribution have a negligible effect on the total measurement uncertainty.

Chapter 4

Evaluation of the measuring chain contributions to the features uncertainty

This chapter analyses the uncertainty contributions on the feature evaluation of the entire measuring chain. Differently from what has been described in the previous chapter, several issues arise when the uncertainty contribution of each component of the measuring chain has to be evaluated.

4.1 Uncertainty evaluation strategy

The single contributions to the total features, uncertainties coming from the measuring chain and from environmental conditions such as background noise, have to be evaluated in order to determine their influence in the measurement. Some aspects of the measuring chain can be evaluated a priori as showed in the previous sections (3), where the error evaluation of uncertainty contributions due to time and amplitude quantisation have been obtained. This was possible because the statistical nature of the disturbance is known, so a Monte Carlo simulation is advisable to evaluate the quantization effect. The main focus and challenge of this work is to evaluate the whole measuring chain error contribution and to achieve this result the architecture, that is summarized in Fig. 4.1, is proposed:

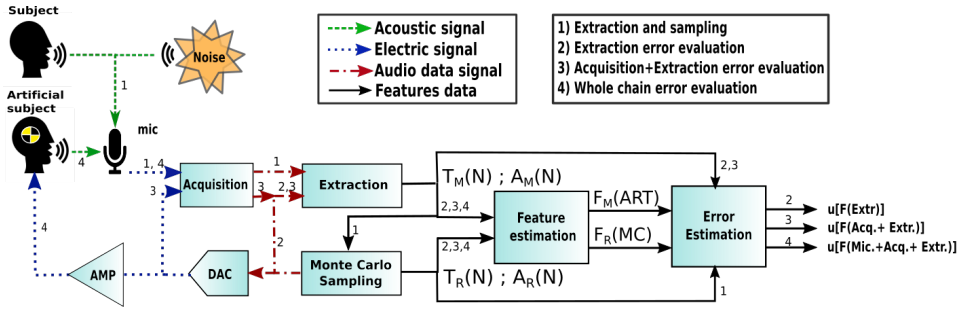


Fig. 4.1 Architecture of the proposed method for the measuring chain error evaluation

The proposed error evaluation method depicted in in Fig. 4.1 is based on four evaluation steps:

1. **Extraction and sampling**: The subject vowel is recorded with a microphone including the contribution of the background noise (green dashed lines). The microphone transforms the acoustic signal into an electrical signal (blue dotted lines). The acquisition module represent the sampling device used to record the subject voice, thus obtaining an audio data signal (red dashed lines). The audio signal is processed by an extraction algorithm that produces measured time sequences of pseudo-periods $T_M(N)$ and amplitudes $A_M(N)$ that can be combined to obtain the metrics described in Chapter 3. The extracted pseudo-periods and amplitudes are then fed to a Monte Carlo sampling algorithm, which produces reference time sequences of pseudo-periods $T_R(N)$ and amplitudes $A_R(N)$ that can be combined to obtain reference features $F_R(MC)$

2. **Extraction uncertainty evaluation**: An artificial signal is synthesised using as reference the sequence of generated $T_R(N)$ and $A_R(N)$. The artificial reference signal is fed back to the extraction algorithm and new sequences of pseudo-periods $T_M(N)$ and amplitudes $A_M(N)$ are extracted and combined to obtain measured artificial features $F_M(ART)$. $F_R(MC)$ and $F_M(ART)$ are compared to obtain an evaluation of the extraction uncertainty u[F(Extr.)]

3. **Acquisition+Extraction error evaluation**: The artificial test signal is converted into the electrical domain by means of a Digital to Analog Converter (DAC) and fed back to the acquisition device through the use of a cable. The same procedure applied to evaluate the extraction error is adopted to evaluate the Extraction and Acquisition contribution u[F(Acq.+Extr)].

4. **Whole chain error evaluation**: The artificial signal, after being transformed in the electrical domain, is amplified and then fed to a loudspeaker mounted inside a torso simulator as will be showed in the next sections. In this way the artificial signal is converted in the acoustic domain and recorded by a microphone connected to the acquisition device. Using the same procedures of the previous steps, an evaluation of the whole chain error contribution u[F(Mic.+Acq.+Extr.)] is obtained.

The architecture depicted in Fig. 4.1 is a four step feedback error evaluation algorithm, that produces a test signal which is capable of "jumping" between different dimensional domains in order to test each part of the measuring chain. The comparison between known and measured features occurs from the second iteration onwards because a dimensional domain transformation is needed to perform a further extraction. To achieve the domain transformation, a signal resampling method was proposed for this work. In the first iteration of the evaluation model in Fig. 4.1 a sequence of pseudo-periods lengths and amplitudes is produced. Such sequences can be visualised as "markers" on the time evolution of the vowel signal as shown in Fig. 2.3. The resampling method was performed using the Matlab functions *interp1* [30] and *linspace* [31] and an amplitude re-normalization using simple multiplications and divisions according to the code reported below:

```
/pseudo-code/
for each period i:
new_period(i)= linspace(1,'original length','new length')
* 'new amplitude'/'original amplitude');
period(i)=interp1('original_range', 'original_period' , new_period(i)) ;
test_signal=vertcat(test_signal,period(i))
end
```

The algorithm described above resamples the audio signal of each period to the desired new length and then re-normalize its amplitude to the desired one. The resampled periods are then concatenated using the Matlab function *vertcat* [32] to produce a test signal as showed in Fig. 4.2.

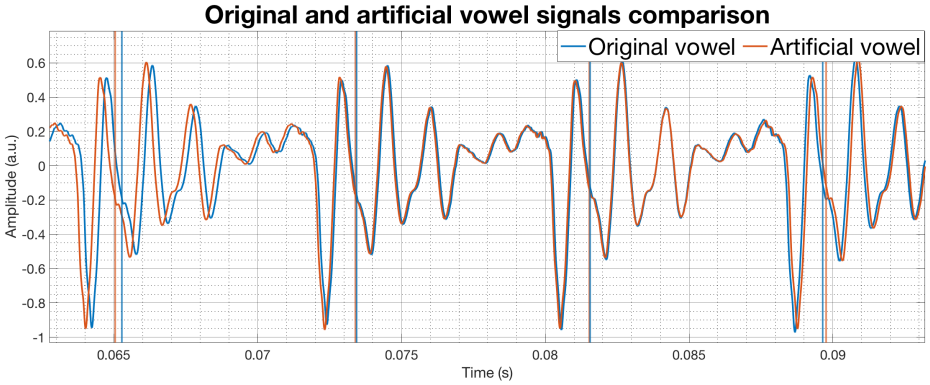


Fig. 4.2 An example of an original vowel signal in blue and an artificial one generated with the proposed resampling method in orange. The vertical blue and orange bars represent the periods start and ending points respectively of the original and artificial signals.

As shown in Fig. 4.2, the artificial vowel signal slightly differ from the original one in terms of amplitude and periods duration. The signal produced by the resampling algorithm has the same spectral characteristics as the original signal but a perturbed sequence of pseudo-periods and amplitudes. The artificial signal concatenation have to deal with invalid frames as described in chapter 2. For dysphonic voices, inharmonic frames could be produced during the phonation task. Such frames are dropped during the feature extraction so also the artificial signal should have the same inharmonic frames dropped. To guarantee the coherence between the original and artificial signal, the invalid frames are concatenated to the artificial signal without any period or amplitude transformation. Differently from the Monte Carlo evaluation of the quantization uncertainty showed in Sec. 3.2, the perturbation of the input measurements is not known. The uncertainty contribution of the measuring chain could be evaluated if the input signal were perfectly repeatable. In such ideal conditions, the uncertainty could be evaluated as the dispersion of repeated measurements using the same input stimulus and a statistical characterisation of the perturbation could also be possible. Unfortunately, the features extracted from vocal signals are not repeatable neither among the same clinical class nor among the same subject repetitions. Moreover the statistical distribution of extracted periods and amplitude showed to be different for every subject and task in our test dataset. From these considerations it was concluded that, to obtain a test signal with the same statistical characteristics as the original signal, a sampling of the pseudo-period and amplitude sequences were necessary, as will be described in the next section.

4.2 Monte Carlo Sampling

To produce a test signal which is statistically comprable to the original signal, a study on the statistical nature of pseudo-period and amplitudes of vowel signal was carried out on our dataset. As an example, the evolution of pseudo-periods duration of three vowels uttered by the same subject (PD) is shown in Fig. 4.3

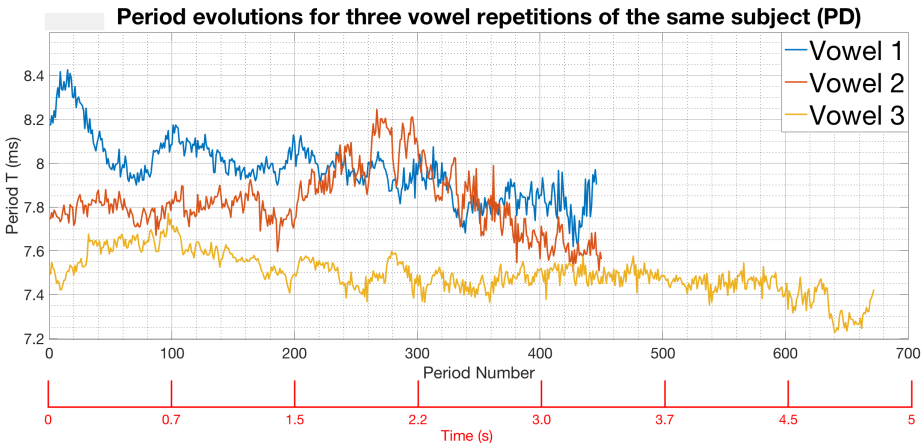
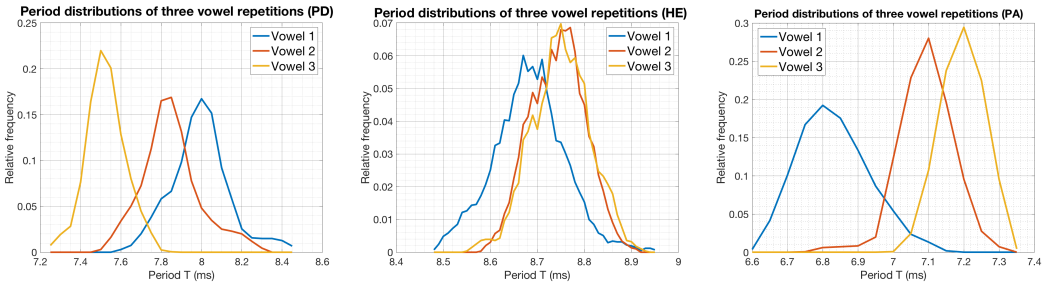


Fig. 4.3 An example of the period evolution of three vowel repetitions from the same subject. The red time scale is not linear due to the variability of the period evaluations and thus is an approximate scale

The evolution of pseudo-periods duration clearly resemble a random walk with different length and dispersion for vowels produced by the same subject. A statistical analysis of the pseudo-periods and amplitude sequences highlighted the absence of a statistical model which can adapt to each subject or to each clinical class. Moreover, even if the same subject emits consecutive vowels, the extracted periods and amplitudes showed very different statistical distributions. As an example, the statistical distribution of three repetitions of the vowel /a/ is reported in Fig. 4.4 for the three clinical classes.



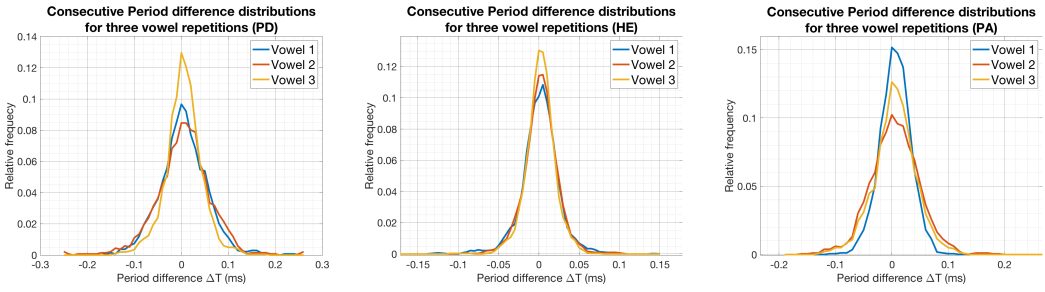
(a) (b) (c)

Fig. 4.4 Period duration distribution of three vowel repetitions from the same PD subject (a), HE subject (b) and PA subject (c)

As shown in Fig. 4.4, the period duration distributions can be very different in terms of positioning, dispersion and shape. The period duration distribution may depend on multiple factors such as vocal education, age, gender and health status. A recurrent evaluation on the metrics described in the previous section is the difference of consecutive periods which represent an important measurement to evaluate the voice frequency stability. A study on our dataset has been carried out to evaluate that statistical distributions of consecutive period differences as expressed by the following equation.

$$\Delta T_i = T_i - T_{i-1} \tag{4.1}$$

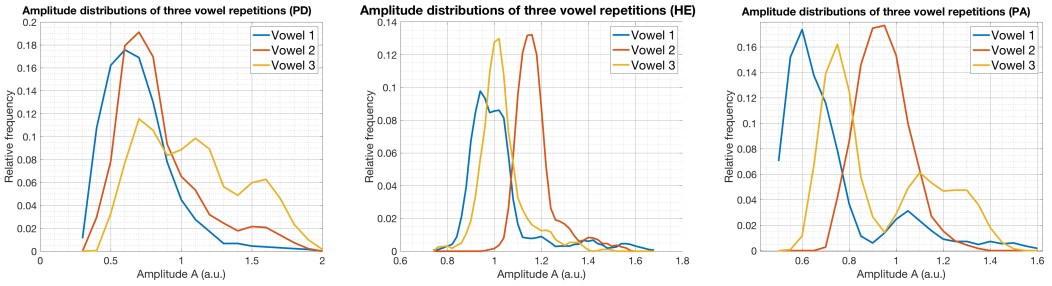
The statistical distributions of ΔT_i can be obtained from the sequence and plotted as shown in Fig. 4.5 i



(a) (b) (c)

Fig. 4.5 Consecutive period difference distributions of three vowel repetitions from the same PD subject (a), HE subject (b) and PA subject (c)

As shown in Fig. 4.5, the distributions of consecutive periods differences are less influenced by the vocal performances and exhibit a enhanced repeatability respect to the case of periods distribution. Such distributions follow a bell shaped curve which is zero centred and the repeatability of dispersion is affected by the health status.
For the pseudo-period peak-to-peak amplitude evaluation, the same analysis has been carried out on three vowel repetitions of three subjects from each clinical class as shown in Fig. 4.6



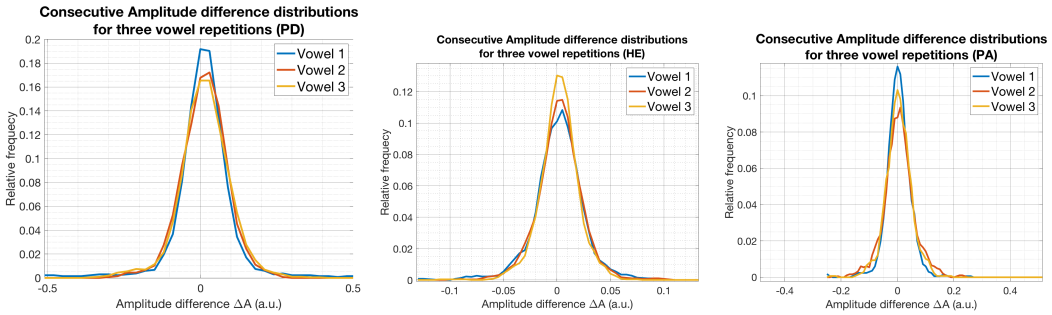
(a) (b) (c)

Fig. 4.6 Amplitude distributions of three vowel repetitions from the same PD subject (a), HE subject (b) and PA subject (c)

As shown in the examples in Fig. 4.6, the distributions of amplitudes highlight a very high variability in terms of positioning, dispersion and shape respect to the case of periods distributions. The statistical study on the consecutive amplitude differences has been carried out, as already done for the pseudo-periods case, defining the following equation:

$$\Delta A_i = A_i - A_{i-1} \tag{4.2}$$

An example of the statistical distributions of consecutive amplitude differences is shown in Fig. 4.7 for three vowel repetitions of three subjects from each clinical class.



(a) (b) (c)

Fig. 4.7 Consecutive amplitude difference distribution of three vowel repetitions from the same PD subject (a), HE subject (b) and PA subject (c)

The "bell" shape of the distributions, which is independent to the subject, may suggest that the consecutive period differences are dominated mainly by physical phenomena. Such phenomena depends on the ability of the phonatory system to jump from a period to the next which is limited by mechanical restrictions such as elasticity, thickness, density and length of the vocal folds. Such an important aspect allows to model the time sequence of periods as a perturbative random walk based on the extracted sequences of periods and amplitudes as shown in Fig. 4.8

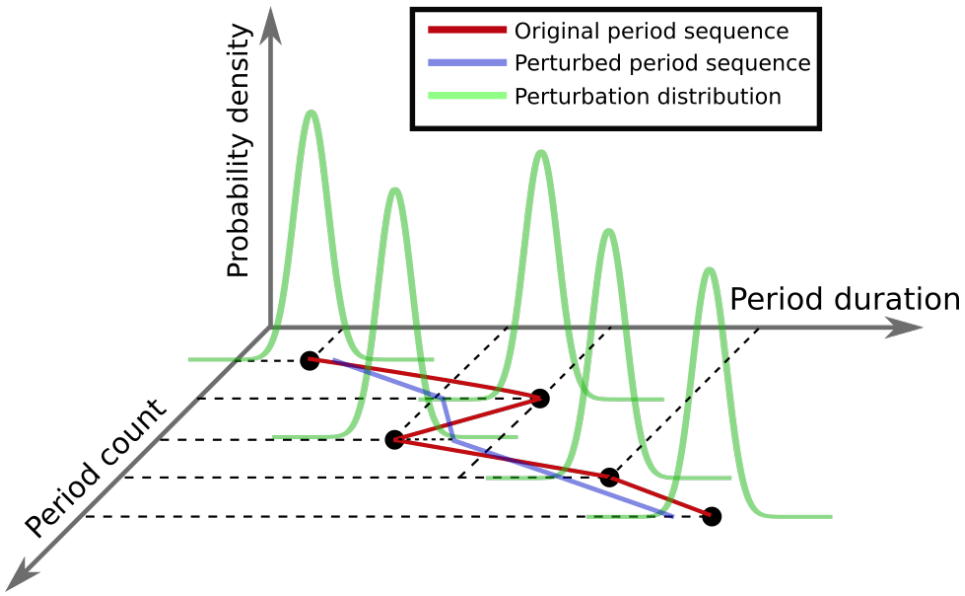


Fig. 4.8 Example of a period random walk perturbed by a random jump extracted from the consecutive difference distributions

The plot in Fig. 4.8 represent a temporal evolution of an extracted pseudo-period sequence (in red) and a perturbed sequence (in blue). Each pseudo-period perturbation is randomly extracted from the consecutive difference distribution represented by the green bell curves. The method described in Fig. 4.8 can be formalized using the following equations:

$$T_i^p = T_i + T_p \;\; ; \;\; T_p = f^{-1}_{ecdf(\Delta T)}(R_T) \tag{4.3}$$

$$A_i^p = A_i + A_p \;\; ; \;\; A_p = f^{-1}_{ecdf(\Delta A)}(R_A) \tag{4.4}$$

$$R_T, R_A \in \Re \in [R_{T,A}(min), R_{T,A}(max)] \in [0,1]$$

where T_i^p is the perturbed period, T_i is the original unperturbed period. $f^{-1}_{ecdf(\Delta T)}(R_T)$ is the inverse of the empirical cumulative distribution function which has as argument a random variable R_T with a uniform probability density function. The choice of such a variable, which ranges from zero to one, is limited between two values: $R_T(min)$ and $R_T(max)$. Such limits are defined by the equations:

$$R_T(min) = f_{ecdf(\Delta T)}[max(\Delta T_{min}, T_{i-1}^p - T_i + \Delta T_{min})]$$
$$R_T(max) = f_{ecdf(\Delta T)}[min(\Delta T_{max}, T_{i-1}^p - T_i + \Delta T_{max})]$$

where ΔT_{min} and ΔT_{max} are respectively the minimum and maximum consecutive period difference. For the perturbed amplitudes the same considerations have to be made. The definition of the terms $R_T(min)$ and $R_T(max)$ limits the choice of the uniform random number between two values that may differ from 0 and 1. This assures that the generated random jump will not exceed the maximum original jump between two consecutive periods or amplitudes. For clarity sake, in this manuscript I will refer to this method as Perturbative Method (**PM**).

Another way to produce test signals with comparable statistical distributions is the Markov chain Monte Carlo (**MCMC**) method, which is based on the free evolution of a random walk:

$$T_i^p = T_{i-1}^p + T_p \;\; ; \;\; T_p = f^{-1}_{ecdf(\Delta T)}(R_T) \tag{4.5}$$

$$A_i^p = A_{i-1}^p + A_p \;\; ; \;\; A_p = f^{-1}_{ecdf(\Delta A)}(R_A) \tag{4.6}$$

$$R_T, R_A \in \Re \subset [0,1]$$

where T_{i-1}^{p} is the previous generated random period. For this work the Metropolis-Hastings (**MH**) [33] algorithm has been used to accept or reject the proposed period and amplitude perturbations, as will be showed on the next section. Once the Monte Carlo generation has produced at least as many samples as the original sequence, the statistical distribution of the generated random walk have to be compared to the target distributions of the extracted pseudo-periods and amplitudes. This is achieved using the Kolmogorov-Smirnov two-sample test (**KS**) [34], as will be showed in the next section.

4.3 Perturbative method and Markov chain Monte Carlo generation algorithm

In order to formalize the methods described in the previous section, the Monte Carlo generation algorithm has been implemented in Matlab. The proposed algorithm is here reported as a pseudo-code. In some points of the algorithm the word "OR" and "AND" are reported in capital letters to highlight the logic valence of the algorithm step.

1. Using the Matlab function *histcounts* [35] target discrete statistical distributions of periods D_T and amplitudes D_A are produced using the Freedman-Draconis method [36] to choose the best number of bins.

2. The difference between consecutive samples is calculated to obtain distributions $D_{\Delta T}$ and amplitudes $D_{\Delta A}$, the relative empirical cumulative curves are estimated using the Matlab function *ecdf* [37] and such curves are then smoothed using the function *smooth* [38]

3. The exit condition is based on the two samples Kolmogorov-Smirnov test (using the Matlab function *kstest2* [39] with a confidence level of 99 %). The exit condition is met if the target distribution of periods is compatible with the generated one AND if the target distribution of amplitudes is compatible with the generated one.

4. Until the exit condition is not satisfied OR the generated sequence is shorter than the original do:

4.1. a period and an amplitude jump is generated and added to the the original one using Eq. 4.3 and 4.4 if adopting **PM** or added to the previous step if using Eq. 4.5 and 4.6 if adopting **MCMC**

4.2. generate proposal period count as: min(1,Dt(i)/Dt(i-1)) (same for amplitudes)

4.3. if an uniform random number between 0 and 1 is less than proposal, the sample is accepted, otherwise is rejected (Metropolis-Hastings). The acceptance condition have to be met for both period AND amplitude proposal.

4.4. -for MCMC: if the length of the generated array is equal to the original one shift the array to the left removing the first sample and concatenating the proposal (burn-in removal)
-for PM: if the length of the generated array is equal to the original one substitute the first sample with a new proposal, at the next iteration substitute the next sample and so on

4.5. refresh the generated periods and amplitudes distributions and the exit condition defined in step 3.

As an example the plots of distributions and sequences generated with the two methods are presented in Fig. 4.9 and Fig. 4.10

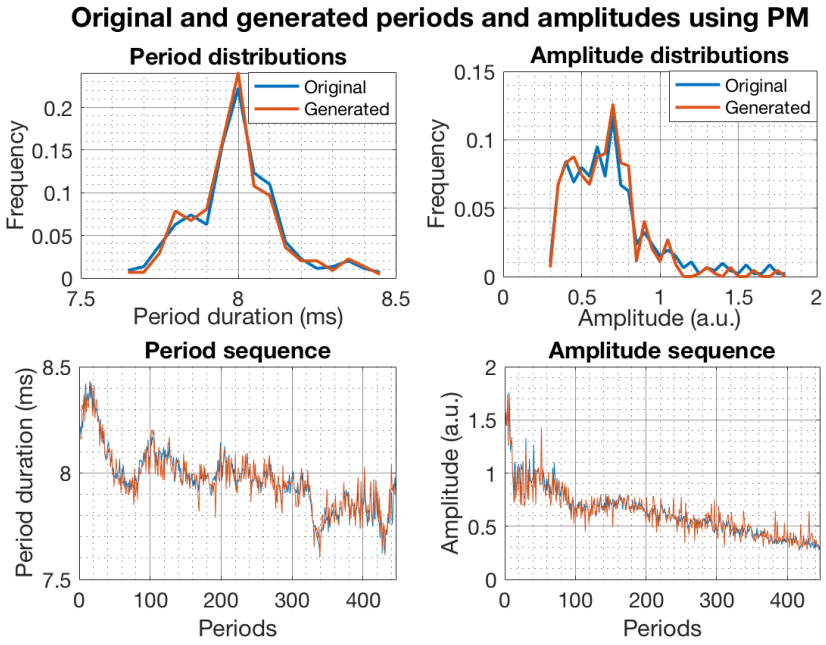


Fig. 4.9 Example of an original and generated periods and amplitudes distributions (top) and time evolutions (bottom) using the Perturbative method

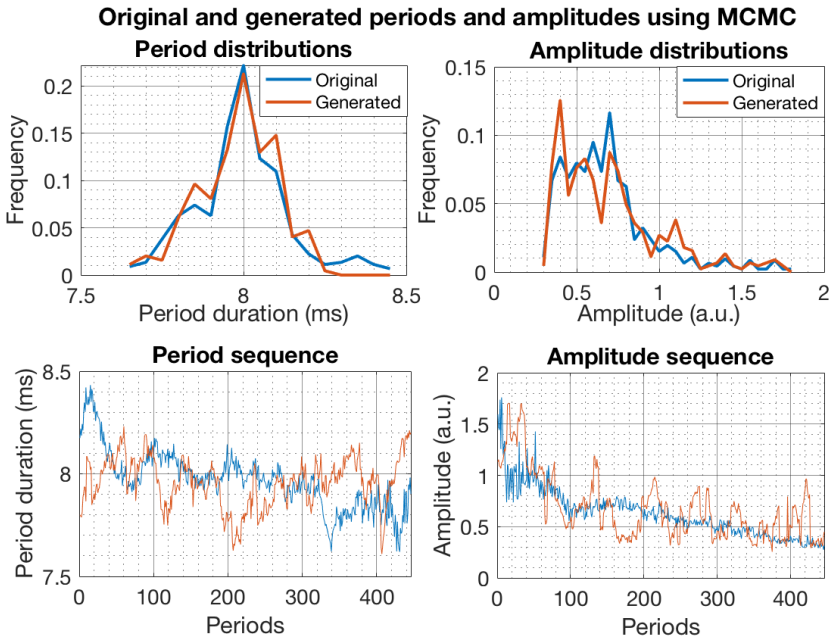


Fig. 4.10 Example of an original and generated periods and amplitudes distributions (top) and time evolutions (bottom) using the Markov Chain Monte Carlo Method

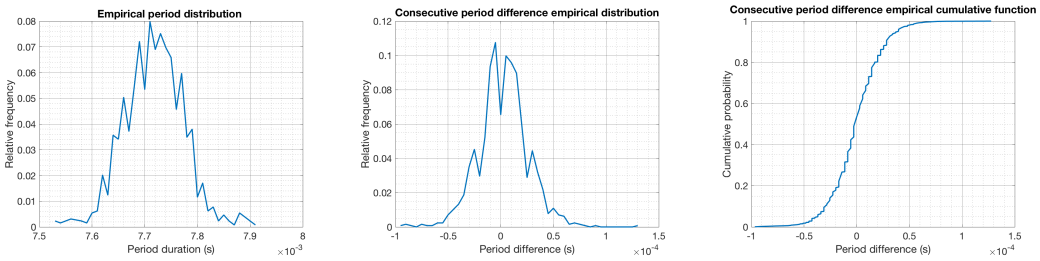
Between the two methods, the **PM** seems to be the most reliable and efficient in terms of similarity to the original distribution and sequences and in terms of acceptance conditions. The main issue with the **MCMC** method is that the random generation is less dependent to the original sequence and the statistical compatibility obtained with the **KS** test could be not reached at the end of the generation. The first generated samples are most always accepted by the **MH** algorithm because the correspondent distributions are still "empty" and unrepresentative. The first set of samples of a **MCMC** generation is called burn-in period and commonly is removed from the sequence in order to give the **MCMC** algorithm the time to reach its target distribution. To solve this problem, the generated sequence is shifted to the left (smaller periods counts) once the generation has reached the same length of the original sequence, removing the first sample and concatenating the next at the end of the generation array.
In Sec. 4.5, a comparison between the perceptual, spectral and temporal characteristics of the artificial vowels generated with the PM and MCMC methods will be carried out to evaluate their performances.

4.4 Considerations about the proposed algorithms

The algorithms used in this work has some critical issues that can produce unrepresentative samples during the Monte Carlo generation. The main source of error in the generation of samples is the quantization of target and proposed distributions and their relative cumulative functions.

4.4.1 Target distribution discretization

To transform a pseudo-period time sequence in a statistical distribution, a discretization of the input time series must be performed. Several methods are available to choose a suitable input variable discretization, such as the Friedman-Draconis method used in this work. As an example, in Fig. 4.11, the empirical period distribution (a), the consecutive difference distribution (b) and cumulative difference distribution (c) for a PA subject are shown for a pseudo-periods evaluation:



(a) (b) (c)

Fig. 4.11 An example of poor quantisation of the period distribution (a), Consecutive difference period distribution (b) and the empirical cumulative distribution function (c) of a PA subject

As can be noted in Fig. 4.11 (a), the empirical target period distribution can be very different from a continuous theoretical distribution and it can produce a unrepresentative statistical model for some of the evaluated subjects.
The best case scenario would be to find an analytical representation of the distributions in order to overcome the issues rising from the discretization of input variables. As already highlighted in the previous sections, the empirical statistical distributions of pseudo-periods and amplitudes are far from being repeatable even if these distributions are relative to vowels emitted from the same subject. Regarding the consecutive difference distributions, bell shaped curves with an enhanced repeatability have been found. In order to evaluate the statistical characteristics of the consecutive difference distributions, a study on the available dataset has been performed. In particular, the mean excess kurtosis and skewness parameters have been evaluated for each clinical class:

Table 4.1 Skewness and Excess Kurtosis of consecutive difference distributions of periods and amplitudes: mean values (standard errors)

Class	**Skewness** $D_{\Delta T}$	**Exc. Kurtosis** $D_{\Delta T}$	**Skewness** $D_{\Delta A}$	**Exc. Kurtosis** $D_{\Delta A}$
PD	-0.2 (0.1)	9 (3)	0.09 (0.09)	4 (1)
HE	0.13 (0.09)	10 (4)	-0.009 (0.07)	12 (5)
PA	-0.2 (1.4)	2.9 (0.1)	-0.03 (0.06)	1.9 (0.4)

The skewness parameter evaluates the symmetry of a distribution and for a perfectly symmetric distribution, such as a Gaussian, the skewness value is 0. The Kurtosis parameter evaluates the “closeness” to a normal distribution which has

a theoretical kurtosis value of 3. In Tab. 4.1 the mean values of excess kurtosis, calculated as the kurtosis minus 3, are reported. The consecutive amplitudes difference distributions seems to be more symmetric than the period distributions and no prevalence of left-right asymmetry can be noticed considering the mean values and the standard error of the skewness values. The excess kurtosis values instead, show a significative distance from 0 (the theoretical value for a normal distribution) and the prevalence of positive values, which indicates that the distributions are “leptokurtic” and therefore they have "fatter tails", that tend to 0 slower, if compared to a normal distribution. Such considerations exclude the possibility of fitting the distribution curves of consecutive differences with a Gaussian function in order to produce Monte Carlo samples that come from a continuous distribution. This practical limit has an important relevance in the Monte Carlo generation algorithm. The empirical period difference distribution and its relative cumulative function suffer from the choice of the number of bins, even though this choice is delegated to a general criterion such as the Friedman-Draconis method. The target distributions are used as reference for the MH algorithm to accept or reject a proposal sample. In fact, the statistical nature of the MH algorithm allow to have less representative target distributions at the expense of the computational power necessary to produce the proposal samples. Regarding the consecutive period difference empirical cumulative function, a coarse quantisation, as shown in Fig. 4.12, can lead to repeated proposal samples which can produce an unrepresentative proposal distribution:

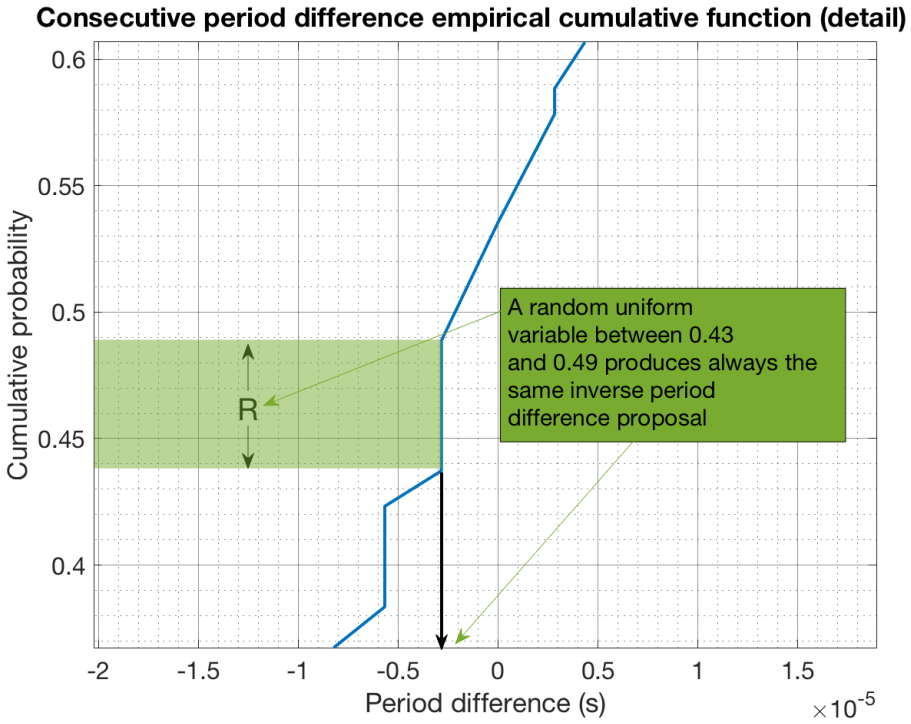


Fig. 4.12 A detailed view of the quantisation effect on the determination of the empirical cumulative function of Fig. 4.11 (c)

To avoid the issues due to the quantization of target distributions, a smoothing on the empirical cumulative function is performed using the Matlab function *smooth*, which is a simple moving average filter. Moreover, a linear interpolation between consecutive values of the empirical cumulative function is performed in order to produce proposal samples with an arbitrary resolution. In Fig. 4.13 a detailed view of the smoothing and interpolating method is shown.

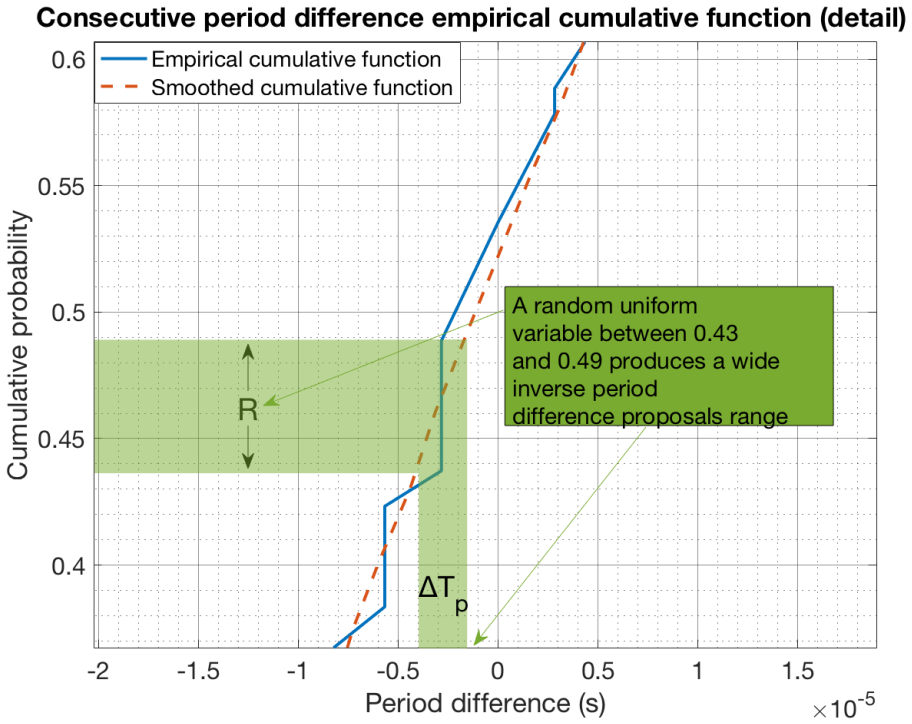


Fig. 4.13 A detailed view of the quantisation effect on the determination of the empirical cumulative function of Fig. 4.11 (c) and the effect of curve smoothing

4.4.2 Periods and amplitudes correlation

The acceptance criteria of the MH algorithm and the KS test are performed on period and amplitude proposal distributions.The acceptance criterion is evaluated using an AND condition to join the results of compatibility tests of proposed periods and amplitudes. This means that the proposed periods and amplitudes are tested independently from each other. This operation is possible only if the evaluated periods and amplitudes are negligibly correlated. In fact, if the periods and amplitude are correlated, the proposal of a new amplitude is dependent on the proposed period and therefore, the joint test using an AND condition is meaningless. To evaluate the feasibility of the acceptance tests, a study on the Pearson correlation coefficient ($\rho = \frac{\sigma_{TA}}{\sigma_T \cdot \sigma A}$) on the available dataset has been performed. As an example, in Fig. 4.14 the plot of coupled pseudo-periods and amplitudes is shown for a PA subject

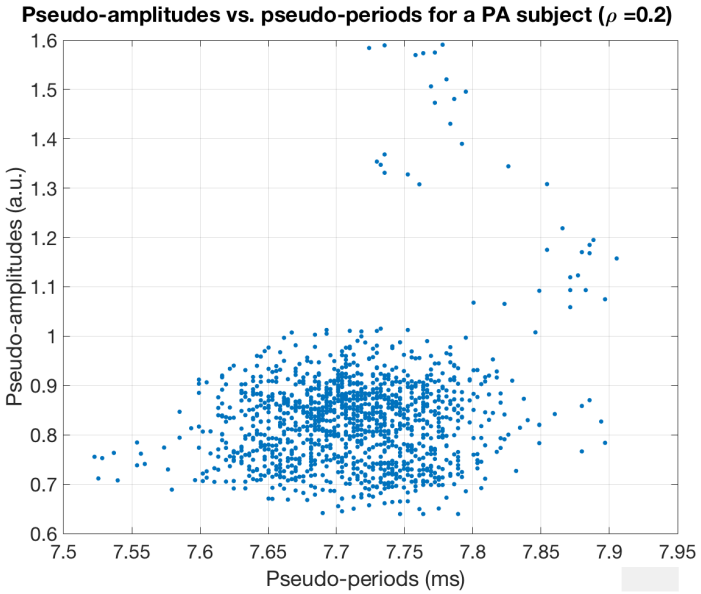


Fig. 4.14 An example of a scatter plot of periods and amplitudes extracted from a vowel emitted by a PA subject. The amplitude scale is normalised respect to a full-scale range of ± 1 a.u. so the peak-to-peak amplitude is in a range between 0 and 2 a.u.

The results of such a study showed a negligible correlation for the three clinical classes. In particular the mean correlation coefficient is -0.02 for the PD subset, -0.05 for the HE subset and 0.03 for the PA dataset. Each correlation coefficient has been evaluated with a p-value<0.05. The low correlation between periods and amplitudes justifies the choice of the independent logic tests used to accept new samples through the MH algorithm and to accept the final proposal distribution using the KS test.

4.5 Time, spectral and cepstral characteristics of the artificial vowels

The domain transformation performed by the resampling algorithm is capable of producing artificial signals with the same statistical characteristics of the original in terms of period and amplitude sequences. From a perceptual point of view, the original and the artificial signal can be evaluated downloading the audio files using the QR codes from Figs. 4.15, 4.16 and 4.17 (Clicking on the QR code opens the web link).

(a) PD (OR)

(b) HE (OR)

(c) PA (OR)

Fig. 4.15 Web link to download audio examples of an original vowel for the three clinical classes

(a) PD (PM)

(b) HE (PM)

(c) PA (PM)

Fig. 4.16 Web link to download audio examples of an artificial vowel, re-synthesized with the PM method, for the three clinical classes

(a) PD (MCMC)

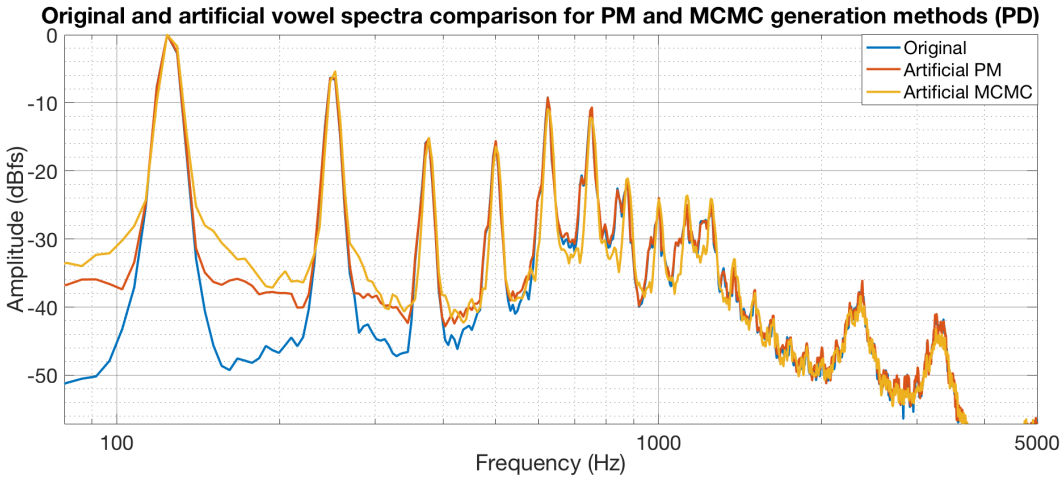
(b) HE (MCMC)

(c) PA (MCMC)

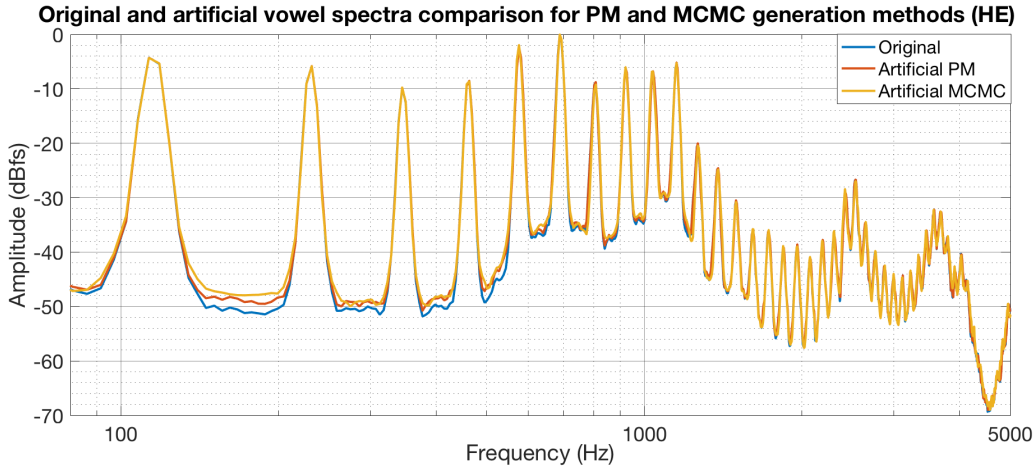
Fig. 4.17 Web link to download audio examples of an artificial vowel, re-synthesized with the MCMC method, for the three clinical classes

As can be heard from the audio files, the PM method sounds clearly better than the MCMC method. This is caused by the fact that the MCMC method may produce periods and amplitudes sequences that are not "natural" for a human voice, even though the generated sequences are statistically comparable to the original ones. For this work some spectral measurement have been implemented to enlarge the set of available features, so the artificial test signals should have the same spectral characteristics of the original ones.

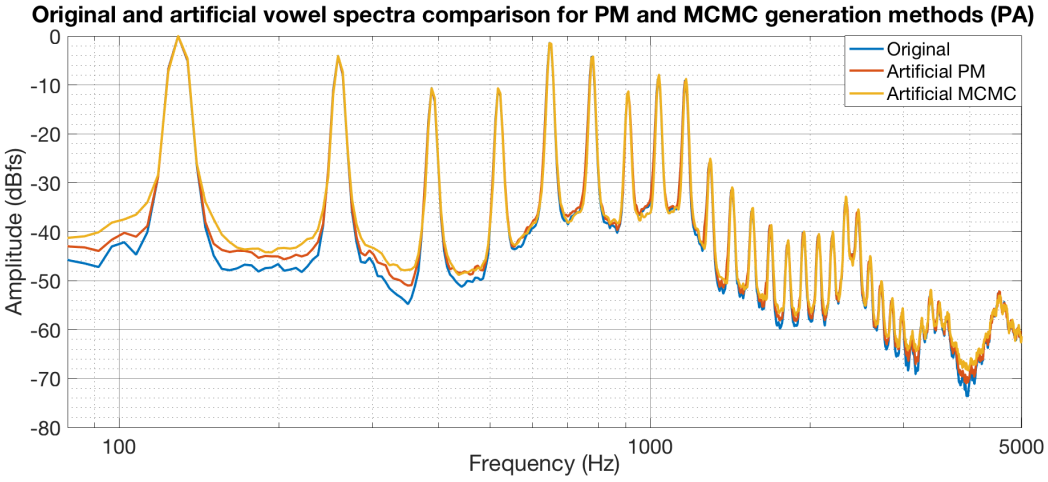
As an example, the peak-normalized spectra (with a resolution of 2.7 Hz), for the three clinical classes using the PM and MCMC methods are presented in Fig. 4.18



(a)



(b)



(c)

Fig. 4.18 An example of spectra comparison between the original vowel, an artificial one generated with PM and an artificial vowel generated with MCMC for a PD subject (a), a HE subject (b) and a PA subject (c)

As shown in Fig. 4.18, the spectra of the artificial vowel are very similar except for the noise floor in the frequency range from 0 to 1 kHz, which seems more pronounced for subjects with high period and amplitude instability (PD and PA). The lowest noise floor in this frequency range is produced by the PM algorithm. In such a range the period and perturbations applied by the resampling algorithm to the original signal could produce some spectral artefacts which are caused by:

- mismatched joints between consecutive periods
- offset errors caused by the amplitude re-normalization
- excessive (yet possible) periods or amplitude perturbations respect the original signal

The joint mismatch between consecutive periods, as shown in Fig. 4.19, has showed to be an important issue with the proposed concatenation method. Such impulses are relatively rare but their presence in the signal may cause instantaneous wide-band rising of the spectrum, as appear in Fig. 4.20.

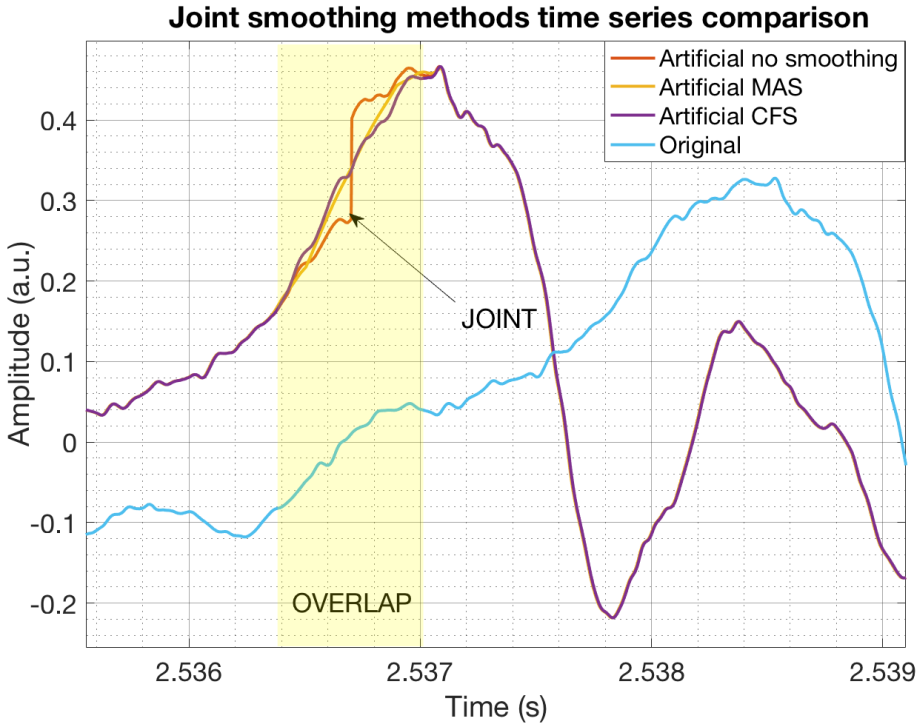


Fig. 4.19 Example of a bad joint between consecutive resampled periods and the methods used to smooth-out the discontinuity

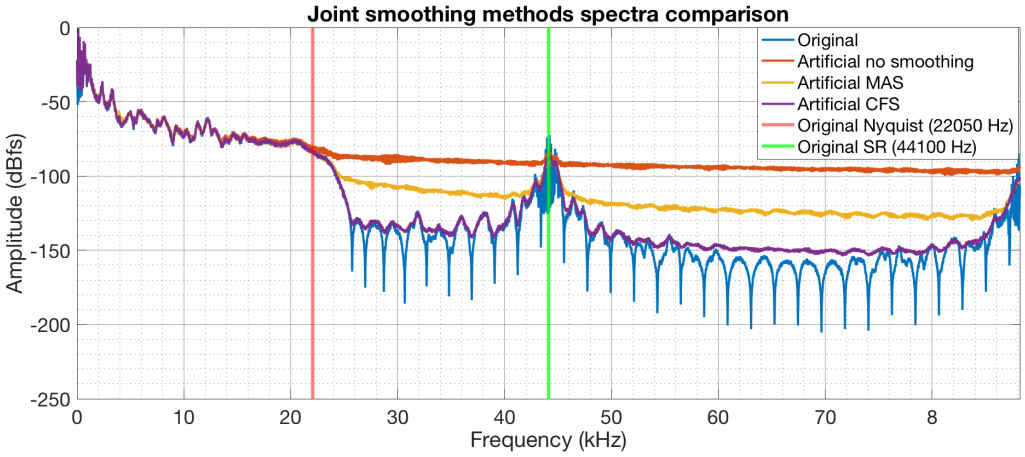


Fig. 4.20 Example of the effect of joint discontinuity in the frequency domain and how the smoothing methods acts on the relative spectra

This problem has been limited filtering the samples around the joint point using a moving average smoothing (MAS) algorithm as shown in Fig. 4.19. Another method that has been tested to solve the problem of mismatched joints is the cross-fade smoothing (CFS). Such a method is similar to the overlap-add technique, where a windowed frame is added to the next with an overlap which is less than the frame length. For this work the overlap length has been chosen as the mean period divided by 20. In this way, the overlap region for the current period wave is the last twentieth of a period and the twentieth before the first sample of the next period that is being concatenated. In this way, for a vowel with an average frequency of 100 Hz the overlap length is 0.5 ms and for a vowel of 400 Hz the overlap length is 0.125 ms. These practical limits have also been chosen in order to not over smoothing the high frequency content of the signal and therefore, for a 100 Hz vowel, the frequencies interested in the smoothing process are higher than 2 kHz. For this work, a Hanning window has been used to scale the overlap region. As an example, a comparison of the two methods for a PD vowel originally sampled at 44100 Sa/s and then oversampled by a factor of 8 is reported in Fig. 4.19 and 4.20 The choice of the joint smoothing method can affect the time series signal as well as the spectrum. In particular the spectra seem influenced at very high frequencies, higher than the original Nyquist frequency. Such a consideration may suggest that the choice of the joint smoothing method should not affect the signal, because the effects of such a choice fall in a frequency range that is well above the voice spectrum band. In practice the presence of mismatched joints can heavily influence the extraction of

pseudo-periods which is performed through a synchronous autocorrelation algorithm, as shown in Sec. 2.4.1.

The spectral issues already showed in the previous paragraph suggest that the re-synthesis algorithm could alter also the cepstrum and, consequently, the CPPS measurements. The rising of the noise floor at low frequencies, which present in the PD example spectrum in Fig. 4.18 (a), can affect the ratio between harmonic energy and inharmonic energy of the spectrum. As an example, the effect of different generation methods on the statistical distribution of extracted CPPS from the original signal and from the artificial ones is presented in Fig. 4.21

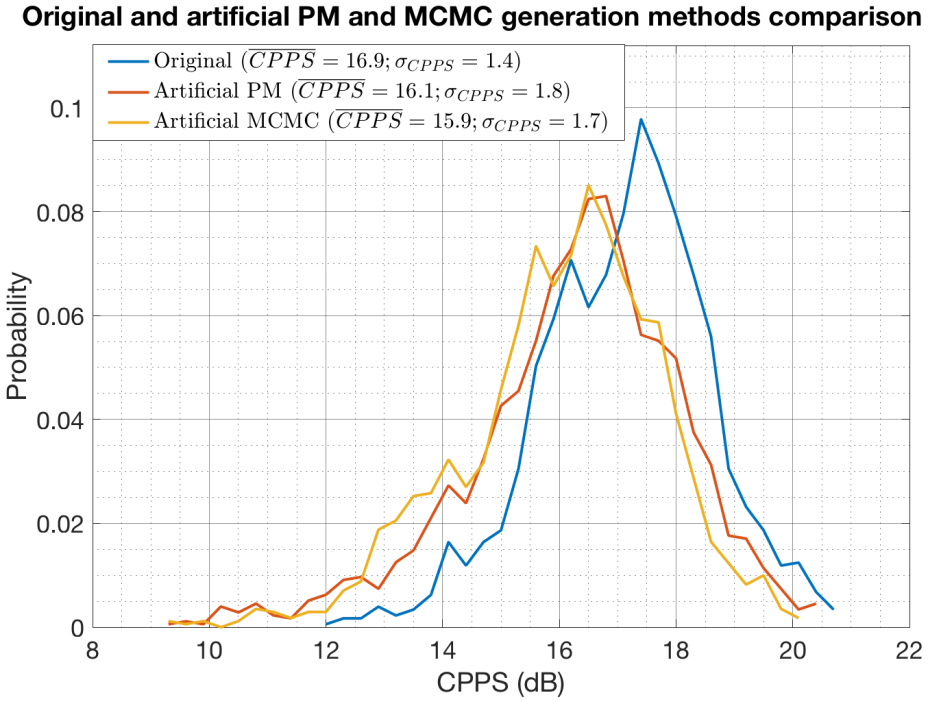


Fig. 4.21 Example of the effect of different generation methods on CPPS distributions

As shown in Fig 4.21, the PM and MCMC method showed similar performances regarding the statistical distribution of extracted CPPS. A slight shift toward smaller CPPS values can be noticed in the plot in Fig. 4.21. Such a bias will be extensively analysed in the next section. An example of the effect of different smoothing methods on a CPPS distribution is shown in Fig. 4.22:

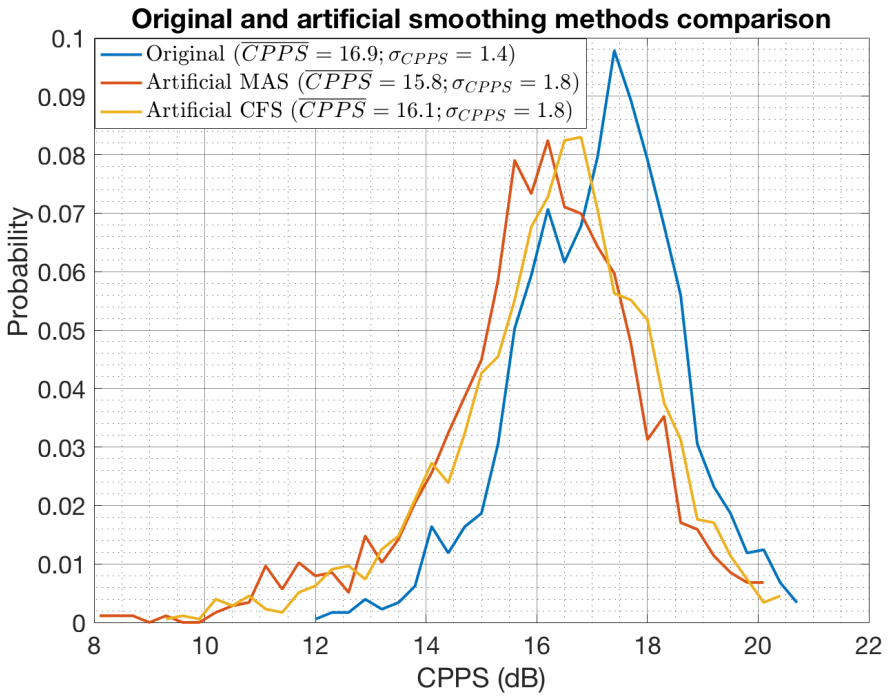


Fig. 4.22 Example of the effect of smoothing methods on the CPPS distributions

As shown in Fig. 4.22, a little shift toward lower CPPS values is evident for each smoothing methods. The best smoothing method seems the CFS because of the reduced distance between the mean values despite a rising in the standard deviation. This behaviour has been noted in each considered subject and clinical class.

Chapter 5

Evaluation of the uncertainty contributions of the whole measuring chain

This chapter concerns the evaluation of the uncertainty contributions of the various measuring chain components. In particular three principal contributions can be identified

1. Extraction algorithm contribution (EXT)
2. Acquisition contribution (ACQ)
3. Acoustic contribution (ACO)

The uncertainty contributions were evaluated adding one contribution at a time in order to evaluate the effects of an increasingly long measuring chain.

1. Extraction algorithm contribution (EXT)
2. Acquisition + Extraction contribution (ACQ+EXT)
3. Acoustic + Acquisition + Extraction contribution (whole chain - (ACQ+EXT+ACO))

5.1 Effects of the extraction algorithm on stability and CPPS metrics (perturbative method)

In order to evaluate the uncertainty associated to the extraction algorithm a separation between the different uncertainty contributions must be performed. As shown in Fig. 5.1, a recursive error evaluation architecture was implemented to evaluate the uncertainty contribution of each measuring chain component. The first part of the evaluation focuses on the period and amplitude stability metrics described in section 2.4.2

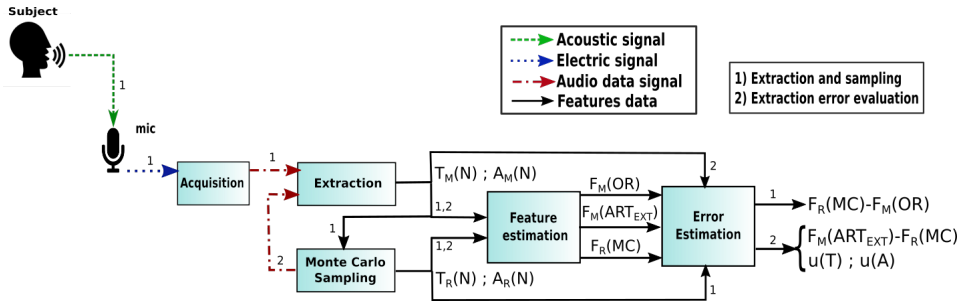


Fig. 5.1 Architecture of the extraction algorithm contribution evaluation method

The scheme in Fig. 5.1, performs various uncertainty evaluations at each of the path indicated by the numbers above the arrows:

- Path 1: the voice of the subject is recorded by the acquisition system and the audio data signal is processed in order to extract sequences of pseudo-periods and amplitudes $T_M(N)$ and $A_M(N)$. At the same time, the extracted periods and amplitudes are used as statistical reference to produce random **R**eference sequences of period and amplitude $T_R(N)$ and $A_R(N)$. The extracted and generated pseudo-periods and amplitudes are then used to calculate a set of **M**easured features $F_M(OR)$ and reference features $F_R(MC)$, where OR stands for original and MC stands for Monte Carlo.

- Path 2: the sequence of generated pseudo-periods and amplitudes are used to produce an artificial test signal using the resampling method described in Sec. 4.1. The new sequences of pseudo-periods and amplitudes $T_M(N)$ and $A_M(N)$ are compared to the reference sequences to obtain an average evaluation of periods and amplitude uncertainties $u(T)$ and $u(A)$. At the same time $T_M(N)$

and $A_M(N)$ are used to produce a set of measured features $F_M(ART)$, where ART stands for artificial. Such features are compared to the reference ones $F_R(MC)$ to obtain an estimation of the extraction uncertainty. The artificial features are also compared to the original ones to produce an mean overall estimation of the measurement error.

For the first step (path 1, described in Sec. 4.1) of the proposed uncertainty evaluation technique, a feature generation method was developed in order to evaluate the capability of producing artificial features which are compatible to the original ones. To clarify such a concept, a parallel analogy with voltage measurements can be made as shown in Fig. 5.2

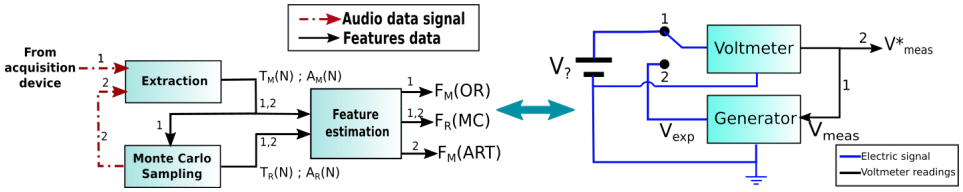


Fig. 5.2 An electrical measurement analogy with the proposed evaluation method

The scheme in Fig. 5.2 depicts the proposed method on the left, which is compared to an electrical measurement on the right. In the electrical measurement described above, suppose we have previous knowledge of a physical phenomenon that should produce a voltage of 1 V. Suppose that we have just an uncalibrated voltmeter which estimates the voltage at $V_{meas} = 0.97$ V (step 1). If a more accurate voltmeter is not available, the uncertainty of the measurement can not be evaluated. If a more accurate voltage generator is available instead, the voltmeter under test can be calibrated respect to a trusted reference voltage source. If we have no previous knowledge of the physical phenomenon and the expected voltage, the calibration should start from the value stated by the untrusted voltmeter $V_{meas} = V_{exp}$ and measure again a voltage V^*_{meas} with the voltmeter to evaluate the bias as the difference $V_{exp} - V^*_{meas}$. Sometimes is not possible to set the reference voltage to the same value of the measured one because the voltages can be set only in a quantized fashion so a little difference between $V_{exp} - V_{meas}$ is expected. In practice the choice of a reference voltage to calibrate the voltmeter is not critical unless the chosen voltage is very different from the one being evaluated. For clarity sake the Table 5.1 summarizes the role of the three performed evaluations:

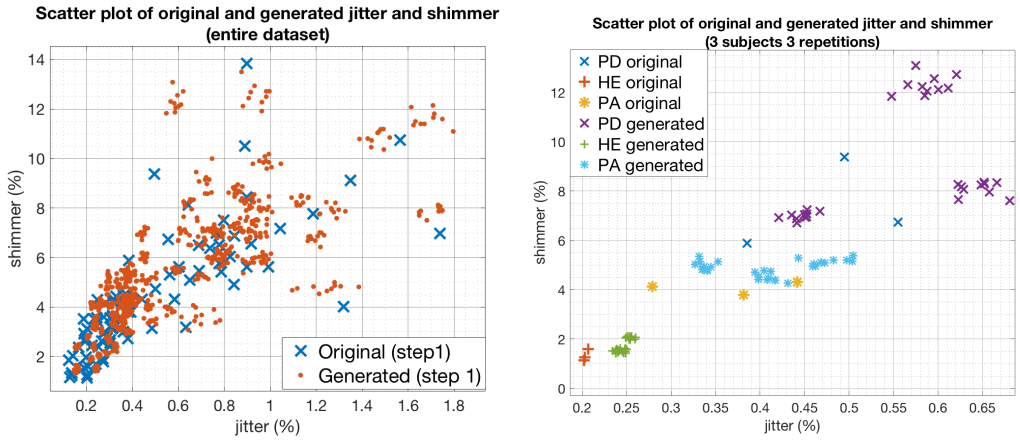
Table 5.1 A conceptual analogy between the proposed method and an electrical measurement

VOICE FEATURES	VOLTAGES
MC-OR: is an evaluation of how much close the reference feature is to the one being evaluated	$V_{exp} - V_{meas}$: is an evaluation of how much close the reference voltage is to the one being evaluated
ART-MC: is an evaluation of the measurement error referred to a **trusted "feature generator"**	$V^*_{meas} - V_{exp}$: is an evaluation of measurement error referred to a **trusted voltage generator**
ART-OR: is an evaluation of the measurement error when a reference feature is not available	$V^*_{meas} - V_{meas}$: is an evaluation of how much close the trusted voltage is to the untrusted one after the calibration

The green row on Tab. 5.1 highlights the importance of this evaluation in the uncertainty estimation method proposed for this work.
In order to minimise the uncertainties due to quantization, all the evaluations were carried out using the “Golden Standard” parameters described in Sec. 3.2.4. In particular the oversampling factor was set to 8 and the bit resolution to 16. For each subject and repetition of the task, 10 artificial vowels were synthesised to produce a dataset that includes 900 entries (30 subject x 3 repetitions x 10 artificial repetitions), that were combined to the original 90 repetitions of the task to get a total of 990 entries. For each repetition, the seed of the Monte Carlo random generation is initialized to the incremental count of the file being processed.

5.1.1 Evaluation of the generation method effects on stability metrics (path 1)

The first step (path 1 described in Sec. 4.1 of the uncertainty evaluation estimates the capability of the Monte Carlo sampling algorithm to produce sets of features that are comparable with the extracted ones, in a similar way of the electrical example of Tab. 5.1, where an expected V_{exp} is set to characterize the voltmeter. The first important consideration to make is strictly bounded to the concept of trust in measurements. The statistical compatibility between the sequences of extracted pseudo-periods and amplitudes and the sequences of the generated ones, as shown in the algorithm in Sec. 4.3 (step 3), is a necessary but not sufficient condition to obtain sets of features compatible to the original ones. As an example, a scatter plot of measured and generated jitter and shimmer is shown in Fig. 5.3



(a) (b)

Fig. 5.3 Scatter plot of 90 original and 900 generated jitter and shimmer (a) and a detailed example of 90 generated vowels from three repetitions of three subjects (PD, HE, PA) (b).

As shown In Fig. 5.3, the generated features (($F_R(MC)$ - red dots) show an overall bias with respect to the original ones ($F_M(OR)$ - blue Xs). A detailed view of original and generated jitter and shimmer values, are shown in Fig. 5.3 (b), which were extracted from three repetitions of a vowel from three subjects. The bias and the dispersion of the generated data seem to depend on the jitter and shimmer values. To investigate the bias and dispersions of the generated features, the mean bias using Eq. 5.1 and mean dispersion using Eq. 5.2 for each clinical class was evaluated for the comparisons between Original ($F_M(OR)$) and Generated ($F_R(MC)$) stability metrics. The mean bias and dispersion for each class were evaluated on the dataset defined in Sec. 2.2 (10 subjects x 3 repetitions = 30 vowels) using the following equations:

$$\overline{BIAS_{MC-OR}(class)} = \frac{\sum_{j=1}^{30}\left(\frac{\sum_{i=1}^{10} F_R^{ij}(MC) - F_M^j(OR)}{10}\right)}{30} \tag{5.1}$$

$$\overline{DISP_{MC}(class)} = \frac{\sum_{j=1}^{30}\sqrt{\frac{\sum_{i=1}^{10}(F_R^{ij}(MC) - \overline{F_R^j(MC)})^2}{10}}}{30} \tag{5.2}$$

where $F_R^{ij}(MC)$ is the array of 10 generated features for the j$-th$ vowel, $F_M^j(OR)$ is the array of extracted features and $\overline{F_R^j(MC)} = \sum_{i=1}^{10} F_R^{ij}(MC)/10$ is the array of features averaged every 10 generations for the j$-th$ vowel. It is easy to demonstrate

that the dispersion parameter can also be evaluated as the mean of the standard deviation of the distances $F_R^{ij}(MC) - F_M^j(OR)$. The data presented in Tab. 5.2 summarizes the results of the bias evaluation of the generated stability metrics respect to the original ones.

Table 5.2 Generated features bias - $\overline{BIAS_{MC-OR}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.065	3.9	0.041	0.056	-0.054	1.1	0.1	0.68	-0.27
HE	0.048	2.9	0.031	0.041	-0.04	0.63	0.057	0.38	-0.39
PA	0.055	2.9	0.034	0.049	-0.042	0.75	0.069	0.53	-0.025

As shown in Tab. 5.2, the features extracted from generated period and amplitude sequences show both positive and negative biases with respect to the original values. The highest bias, if compared to the expected values, appears to affect more the shimmer evaluation of the PD class. The dispersion of generated points have also been evaluated using Eq. 5.2 as reported in Tab. 5.3:

Table 5.3 Generated features dispersions - $\overline{DISP_{MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.02	1.3	0.01	0.01	0.02	0.22	0.02	0.13	0.15
HE	0.01	0.7	0.01	0.01	0.01	0.08	0.007	0.05	0.08
PA	0.02	1.1	0.02	0.01	0.02	0.15	0.01	0.09	0.14

Once again the highest dispersion affects the shimmer evaluation of the PD class. In order to compare the dispersion of generated features to the natural dispersion of the subjects, the Eq. 5.2 has been modified as follow:

$$\overline{DISP_{OR}(class)} = \frac{\sum_{j=1}^{10} \sqrt{\frac{\sum_{i=1}^{3}(F_M^{ij}(OR) - \overline{F_M^j(OR)})^2}{3}}}{10} \tag{5.3}$$

where $F_M^{ij}(OR)$ is the array of original features for the j$-th$ subject and the i$-th$ vowel repetition and $\overline{F_M^j(OR)}$ is the averaged array along the three repetitions. The evaluated dispersions of the task repetitions are reported in Tab. 5.4:

Table 5.4 Measured Original dispersions - $\overline{DISP_{OR}(class)}$

Class	*jit* (%)	jit_{abs} (µs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.19	11	0.12	0.09	0.29	1.30	0.11	0.95	6.2
HE	0.09	5.7	0.06	0.05	0.27	0.74	0.06	0.48	3.8
PA	0.33	15	0.18	0.22	0.46	1.3	0.11	0.9	4.6

The original mean dispersion values of each class is almost always higher than the dispersion of generated features. As an example, in Fig. 5.4 the plots of bias and dispersion of the generation uncertainty contribution are compared to the original dispersion for jitter and shimmer measurements.

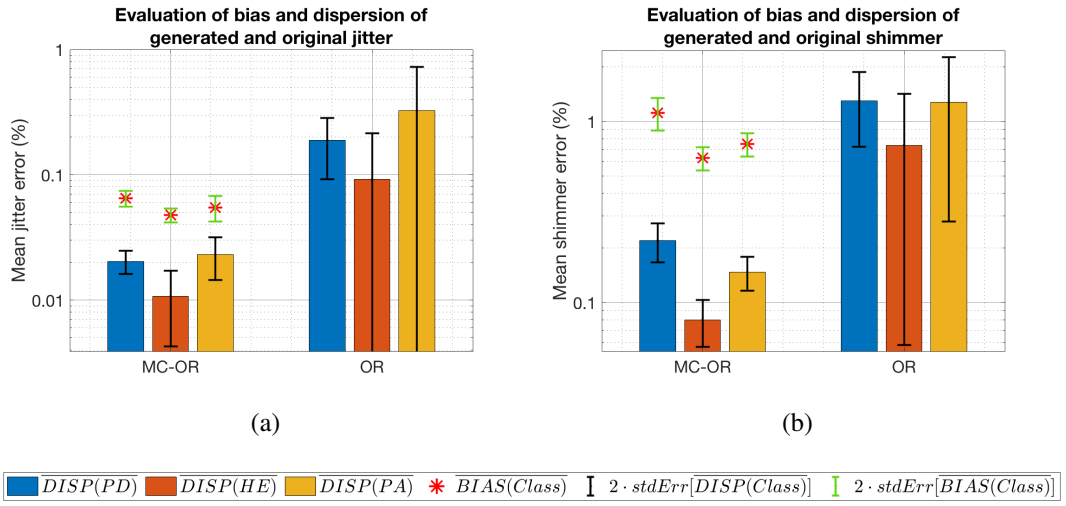


Fig. 5.4 Evaluation of generation mean bias and dispersion of jitter (a) and shimmer (b) for the three clinical classes.

As shown in Fig. 5.4, the dispersion parameters of the generated jitter and shimmer is lower than the intra-subject dispersion. This considerations will be discussed later in the conclusion of this chapter.

5.1.2 Evaluation of the extraction contribution to the period and amplitude uncertainty (path 2)

In order to evaluate the period and amplitude extraction uncertainty, the raw data from the period and amplitude extraction algorithm was analysed as shown in Fig.

2.3. To evaluate such an uncertainty, the root mean squared difference between coupled samples was evaluated:

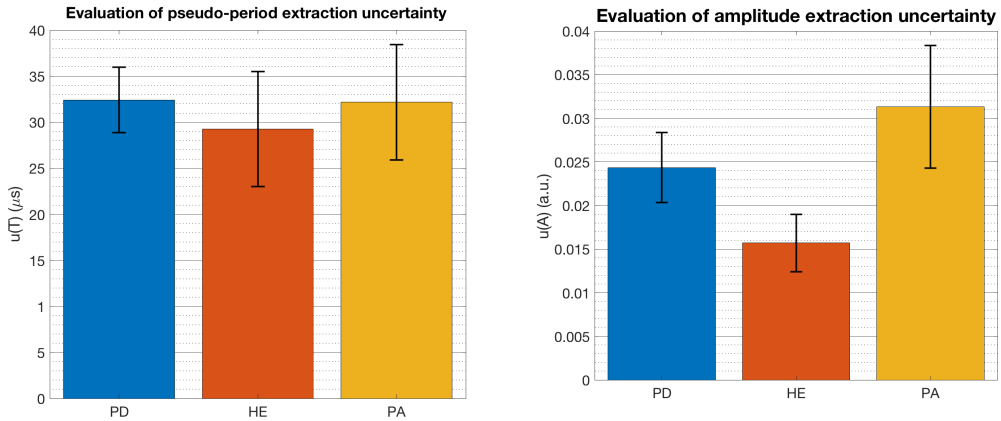
$$u(T) = \sqrt{\frac{\sum_{i=1}^{N}(T_{MC}^{i} - T_{EXT}^{i})^2}{N}} \tag{5.4}$$

A similar equation was used to evaluate the amplitude extraction error. The results of this analysis are reported in Tab. 5.5

Table 5.5 Pseudo-periods and amplitudes mean extraction error - $u(T)$, $u(A)$

Class	u(T) (μs)	u(A) (a.u.)
PD	32	0.02
HE	29	0.01
PA	32	0.03

The results of this analysis highlight the uncertainty contribution of the extraction algorithm on evaluating sequences of periods and amplitudes. The averaged values of period extraction uncertainties does not seem to depend on the clinical class, while for the amplitude, the HE class shows a lower value respect to the other classes. In Fig. 5.5 the plots of pseudo-period and amplitude uncertainty is presented.



(a) (b)

Fig. 5.5 Mean error evaluation for the three clinical classes for pseudo-periods (a) and amplitudes (b) measurements.

The evaluated period uncertainty has the same order of magnitude of the sampling period of the recordings used to perform this evaluation (1/(44.1 kSa/s) $\approx$ 22 μs) so the contribution of the extraction algorithm is similar to the quantisation contribution evaluated in Sec. 3.1.2 ($\approx$ 26 μs). Regarding the amplitude uncertainty evaluation, the recordings were normalized to have a maximum absolute peak of 1 so that the full scale range spans from -1 to 1. The evaluated uncertainty is considerably larger than the amplitude uncertainty evaluated in Sec. 3.1.2 ($\approx 3.8 \cdot 10^{-5}$ a.u.).

5.1.3 Evaluation of the extraction uncertainty contributions of stability metrics (path 2)

To evaluate the uncertainty contribution of the extraction algorithm, the artificial signal is processed by the extraction algorithm to obtain the measured features $F_M(ART)$. Using the dataset defined in Sec. 2.2, an example of a scatter plot of generated and measured jitter and shimmer is presented in Fig. 5.6 .

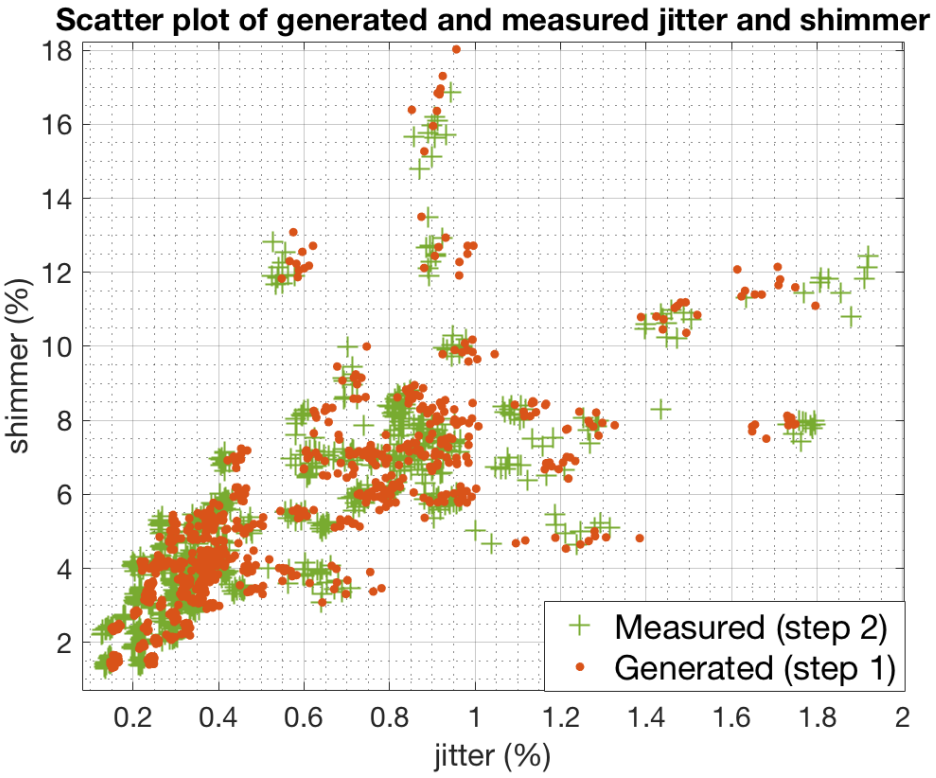


Fig. 5.6 Scatter plot of generated (MC) and measured (ART) jitter and shimmer

The distance between generated and extracted data seems less evident in the scatter plot of generated $F_R(MC)$ and measured $F_M(MC)$ jitter and shimmer as can be noted in Fig. 5.6. To investigate the errors of measured and generated features, the mean distance for each clinical class was evaluated for the comparisons between Generated (MC) and Artificial (ART) features. For the $ART - MC$ bias evaluation, 900 generated entries were compared to the matching 900 extracted entries. using the following equation:

$$\overline{BIAS_{ART_{EXT}-MC}(class)} = \frac{\sum_{i=1}^{300} F_M^i(ART_{EXT}) - F_R^i(MC)}{300} \tag{5.5}$$

To evaluate the dispersions of the extracted artificial features the following equation was used:

$$\overline{DISP_{ART_{EXT}}(class)} = \frac{\sum_{j=1}^{30} \sqrt{\frac{\sum_{i=1}^{10}(F_M^{ij}(ART_{EXT}) - \overline{F_M^j(ART_{EXT})})^2}{10}}}{30} \tag{5.6}$$

The data presented in Tab. 5.6 and Tab. 5.7 summarizes the results of the evaluation of Eq. 5.5 and Eq. 5.6.

Table 5.6 Measured artificial bias of the extraction contribution - $\overline{BIAS_{ART_{EXT}-MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	-0.02	-1.4	-0.02	-0.02	0.02	-0.07	-0.006	-0.032	0.051
HE	-0.02	-1.3	-0.02	-0.01	0.02	-0.02	-0.003	-0.007	0.260
PA	-0.01	-0.8	-0.01	-0.01	0.02	-0.07	-0.006	-0.037	-0.002

Table 5.7 Measured artificial dispersions of the extraction contribution - $\overline{DISP_{ART_{EXT}}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.02	0.9	0.010	0.011	0.02	0.22	0.020	0.13	0.16
HE	0.01	0.7	0.007	0.006	0.02	0.08	0.008	0.05	0.09
PA	0.02	1.1	0.014	0.015	0.03	0.18	0.016	0.10	0.13

Comparing the data presented in Tab. 5.7 and Tab. 5.3, it can be noted that the bias change between the two step evaluation, while the dispersions have very similar values. As an example, in Fig. 5.7 the plots of bias and dispersion of the extraction contribution are compared to the generation dispersion for jitter and shimmer measurements.

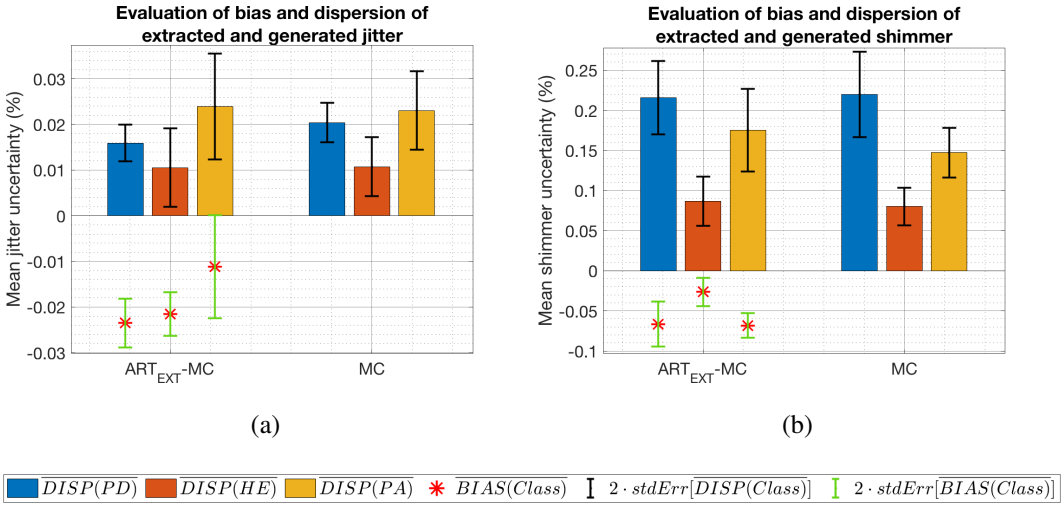


Fig. 5.7 Evaluation of extraction mean bias and dispersion of jitter (a) and shimmer (b) for the three clinical classes.

As shown in Fig. 5.7, the dispersion of the extracted jitter and shimmer is comparable to the dispersion of the generated ones. Respect to the example in Fig. 5.4 a negative bias is present for jitter and shimmer evaluations. In the previous section the uncertainty of the extracted amplitudes was estimated around 0.02 a.u, which is noticeably larger than the analytical estimation of $\approx 3.8 \cdot 10^{-5}$ a.u.. Such consideration can explain the larger dispersion of the extracted shimmer values ($\approx 10^{-1}$ %) respect to the quantisation contribution evaluated in Sec. 3.1.2 ($\approx 3 \cdot 10^{-4}$ %).

5.1.4 Evaluation of the extraction contribution to CPPS features uncertainty

The weak point of the proposed re-synthesis method is the alteration of the spectral and cepstral characteristics of the original vowels as shown in Sec. 4.5. The proposed re-synthesis method is not capable of producing sequences of known and trusted CPPSs, therefore any alteration of such values have to be intended as a contribution to the measurement uncertainty. Without a trusted CPPS synthesis method it is impossible to attribute the difference between original and artificial CPPS values to the re-synthesis or to the extraction method. For this reason the bias on CPPS metrics is evaluating comparing the 30 original (OR) extracted features with the CPPS

metrics extracted from the 300 artificial signals (ART) as shown in the equation:

$$\overline{BIAS_{ART_{EXT}-OR}(class)} = \frac{\sum_{i=j}^{30}(\frac{\sum_{i=1}^{10} F_M^{i,j}(ART_{EXT}) - F_M^j(OR)}{10})}{30} \tag{5.7}$$

In Tab. 5.8 and Tab. 5.9, the bias and dispersion evaluation are reported for the CPPS metrics defined in Sec. 2.4.4, using Eq. 5.7 for the bias evaluation and Eq. 5.6 for the dispersion evaluation.

Table 5.8 Measured artificial bias of CPPS metrics - $\overline{BIAS_{ART_{EXT}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	-0.51	-0.53	-0.42	0.09	0.30	-0.68	-0.35	0.06	-0.17
HE	-0.24	-0.24	-0.30	0.02	0.33	-0.27	-0.22	-0.05	0.34
PA	-0.39	-0.37	-0.31	-0.01	0.21	-0.39	-0.39	-0.02	0.05

Table 5.9 Measured artificial dispersion of CPPS metrics - $\overline{DISP_{ART_{EXT}}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.06	0.07	0.54	0.06	0.79	0.18	0.11	0.13	0.37
HE	0.04	0.05	0.42	0.03	0.62	0.10	0.08	0.09	0.27
PA	0.06	0.07	0.51	0.05	0.64	0.14	0.14	0.12	0.25

As reported in Tab. 5.8 and 5.9, the bias and dispersion seem negligible if compared to the scale of CPPS evaluations, which ranges from 0 to 26 dB. To have a comparison with the original dispersion of CPPS metrics, the data in Tab. 5.10 summarise the evaluated intra-subject dispersions.

Table 5.10 Measured original dispersion of CPPS metrics - $\overline{DISP_{OR_{EXT}}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.4	0.4	0.6	0.2	1.8	0.7	0.4	0.3	0.9
HE	0.7	0.7	0.9	0.1	1.1	0.8	0.5	0.3	0.7
PA	1.2	1.2	1.4	0.2	1.5	1.5	1.1	0.3	0.4

As can be noted in Tab. 5.10, the original intra-subject dispersions $\overline{DISP_{OR_{EXT}}(class)}$ are higher than the artificial dispersions $\overline{DISP_{ART_{EXT}}(class)}$. An example of bias and dispersion of extracted and original mean CPPS is reported in Fig. 5.8

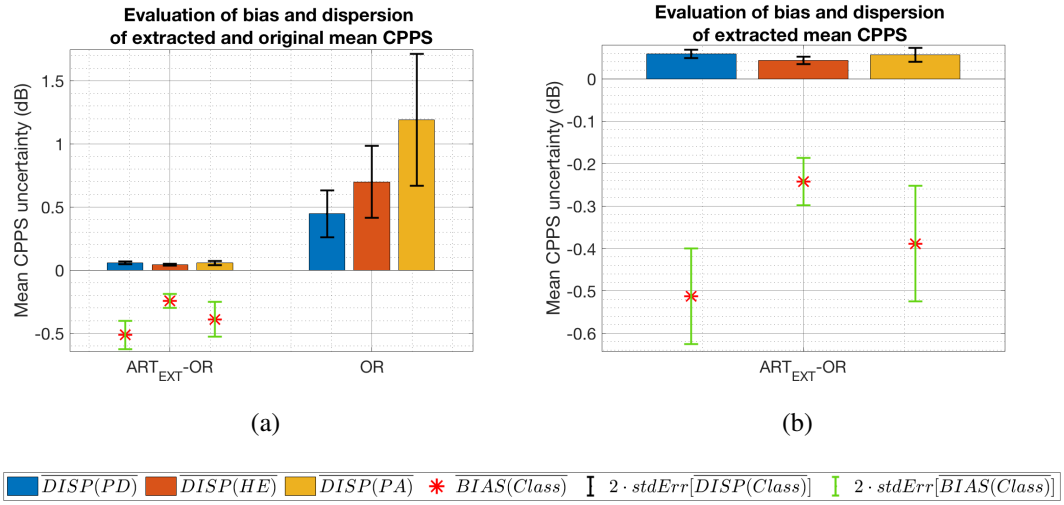


Fig. 5.8 Mean bias and dispersion evaluations of artificial and original Mean CPPS (a). A detailed view is shown in (b)

5.2 Evaluation of the MCMC generation method

The same evaluations performed for the PM generation method were performed using artificial signals generated with the MCMC method. In particular, the sequence of the tables presented in this section is:

- Evaluation of the generation method effects on stability metrics (Bias: Tab. 5.11; Dispersion: Tab. 5.12)
- Evaluation of period and amplitude uncertainty (Tab. 5.13)
- Evaluation of the extraction uncertainty contribution on stability metrics (Bias: Tabs. 5.14; Dispersion: Tab. 5.15)
- Evaluation of the extraction contribution to CPPS features uncertainty (Bias: Tab. 5.16; Dispersion: Tab. 5.17)

The considerations on the presented data will be extensively covered in the next section.

Table 5.11 Generated features bias (MCMC) - $\overline{BIAS_{MC-OR}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	-0.17	-9.5	-0.12	-0.05	-0.05	-0.88	-0.06	0.68	-0.71
HE	-0.13	-8.9	-0.09	-0.03	-0.10	-0.47	-0.04	0.56	-0.57
PA	-0.25	-11.0	-0.18	-0.11	-0.03	-1.00	-0.08	0.77	-0.77

Table 5.12 Generated features dispersion (MCMC) - $\overline{DISP_{MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.02	1.20	0.01	0.02	0.19	0.23	0.02	0.32	2.80
HE	0.01	0.82	0.01	0.01	0.15	0.13	0.01	0.19	2.70
PA	0.02	1.10	0.01	0.02	0.21	0.20	0.02	0.25	2.60

Table 5.13 Pseudo-periods and amplitudes mean extraction uncertainty (MCMC) - $u(T)$, $u(A)$

Class	u(T) (μs)	u(A) (a.u.)
PD	49	0.05
HE	53	0.04
PA	58	0.05

Table 5.14 Measured artificial bias of the extraction contribution (MCMC) - $\overline{BIAS_{ART-MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.220	12	0.140	0.090	0.093	0.180	0.013	-0.006	-0.330
HE	0.160	11	0.094	0.062	0.073	0.097	0.008	0.002	-0.064
PA	0.360	17	0.220	0.200	0.180	0.250	0.021	0.032	-0.150

Table 5.15 Measured artificial dispersions of the extraction contribution (MCMC) - $\overline{DISP_{ART}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.03	1.60	0.02	0.02	0.19	0.26	0.02	0.32	2.70
HE	0.02	1.60	0.02	0.01	0.13	0.13	0.01	0.18	2.70
PA	0.04	1.80	0.02	0.03	0.21	0.24	0.02	0.26	2.60

Table 5.16 Measured artificial bias of CPPS metrics (MCMC) - $\overline{BIAS_{ART-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	-0.7	-0.6	-0.6	0.15	0.5	-1.0	-0.4	0.02	-0.2
HE	-0.4	-0.4	-0.4	0.08	0.8	-0.5	-0.3	-0.03	0.2
PA	-0.5	-0.4	-0.5	0.08	0.8	-0.7	-0.4	-0.13	0.2

Table 5.17 Measured artificial dispersions of CPPS metrics (MCMC) - $\overline{DISP_{ART}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.07	0.09	0.6	0.07	0.8	0.2	0.12	0.1	0.3
HE	0.05	0.06	0.4	0.04	0.7	0.1	0.09	0.1	0.3
PA	0.08	0.10	0.6	0.06	0.7	0.2	0.13	0.1	0.4

5.3 Final considerations on the extraction uncertainty evaluation

From the data presented in the previous sections, some interesting considerations can be made. For clarity reasons such conclusive remarks will be faced separately to better understand each evaluation step.

5.3.1 Generation method evaluation (path 1)

The reliability of the proposed method is based on the trust on the Monte Carlo sampling algorithm, which produces sets of features that are known with high accuracy. This is due to the fact that the generated periods and amplitudes are **numbers** with a floating point 64 bit double-precision (IEEE 754 format).
As shown in Sec. 3.1.2, the absolute uncertainty of pseudo-period measurements of a vowel sampled at 44100 Sa/s is around 26 μs. In the same section, the absolute uncertainty on amplitude measurements, for a signal sampled with a resolution of 16 bit, were estimated approximately as $3.8 \cdot 10^{-5}$ a.u.. Regarding the pseudo-period extraction uncertainty evaluated in Sec. 5.1.2, the measured uncertainty (29-32 μs) is very close to the one analytically estimated. The evaluated amplitude uncertainty in Sec. 5.1.2 show larger values (0.01-0.03 a.u.) if compared to the one analytically evaluated.
A generated period is not prone to the limits of the sampling process, since the resolution of any generated period or amplitude depends on the precision format of

the number representation. In a double precision number, the first 12 bits are used to represent the sign and the exponent of the number expressed in scientific notation. The remaining 52 bits are used to represent the number which is multiplied by the exponent as:

$$Double = (-1)^{sign} \cdot (1 + \sum_{i=1}^{52} b_{52-i} \cdot 2^{-i}) * 2^{exp-1023} \tag{5.8}$$

where *Double* is the floating-point number, *sign* is the first bit, *exp* is the exponent represented by 11 bits, and $(1 + \sum_{i=1}^{52} b_{52-i} \cdot 2^{-i})$ is called *mantissa* and it is a number with 52 bit precision. The numbers in this format are represented with an uncertainty that depends on the number value. In the Matlab environment the function "eps" returns the floating-point relative accuracy of a given number. As an example the representation of pseudo-periods values from 2.5 ms (0.0025 s) to 10 ms (0.01 ms) have a relative accuracy that is depicted in Fig. 5.9

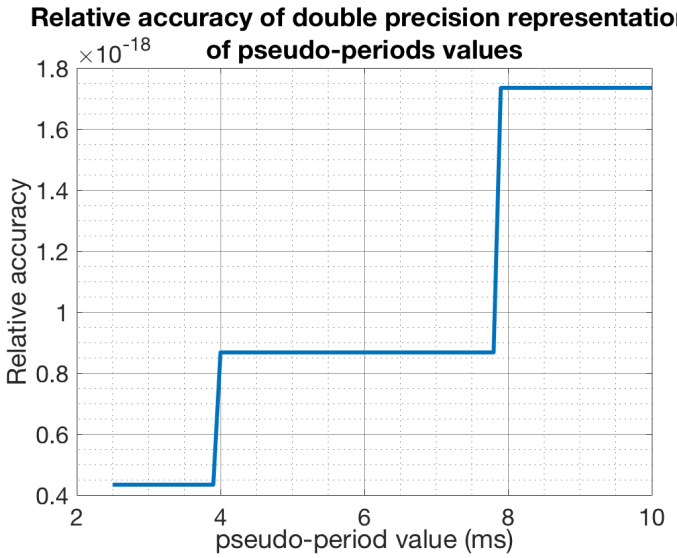


Fig. 5.9 Relative accuracy of pseudo-periods values represented in a double precision format

As shown In Fig. 5.9, the relative accuracy changes as the represented pseudo-period varies. As an example, a 0.01 s period length (100 Hz), would be represented with a relative accuracy of $\approx 1.7 \cdot 10^{-18}$, thus giving an absolute period uncertainty of $\approx 1.7 \cdot 10^{-20}$ s, which is much less than the evaluated one. Finally if the Tabs. 5.3 and 5.4 and the plots of Fig. 5.4 are considered, one should notice that the dispersion of the generated features are always lower than the intra-subject dispersion. Such a

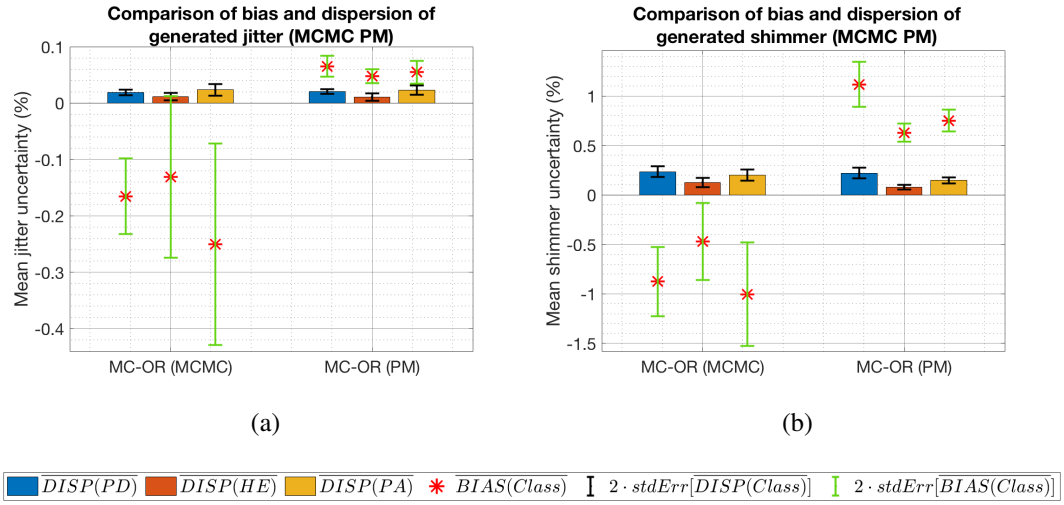
consideration highlights that the generation method produces artificial values that are statistically closer to the single original vowel, used as a reference, instead of being scattered within the subject natural dispersion of task repetitions.

5.3.2 Extraction algorithm uncertainty evaluation (path 2)

Looking at the Tabs. 5.6, 5.7 and plots in Fig. 5.7, biases and dispersions can be noticed in each feature evaluation. In particular the dispersion of the extracted artificial features $F_M(ART)$ are statistically comparable to the dispersion of the generated reference features $F_R(MC)$. This consideration may suggest that the extraction algorithm only adds some bias contribution to the uncertainty (as can be noted in Fig. 5.7 (b)), while leaving unchanged the dispersion of the reference features. The evaluation of the pseudo-periods and amplitudes extraction uncertainty highlighted a comparable contribution respect with the uncertainty evaluations in sections 3.1.2 and 3.2. In particular, the extraction period uncertainty estimation is very close to the quantisation period, therefore if the evaluation of period uncertainty is substituted in Eq. 3.13, a jitter uncertainty around 0,02 % is expected. Such expected uncertainty is very close to the values in the first column of Tab. 5.7. The evaluation of the extraction bias allows to correct such an effect if a sufficient number of artificial vowels is produced, as will be showed in the next chapter.
Regarding the CPPS estimation uncertainties, the negligible entity of bias and dispersions may suggest that the re-synthesis method proposed for this work is quite "transparent" for the cepstral characteristics of the signal. This fact could allow to consider any alteration of the CPPS metrics as caused by the experimental set-up contributions and not to the re-synthesis method itself.

5.3.3 Comparison between PM and MCMC generation methods

The uncertainty assessment methods carried out using the MCMC method produced different evaluations of each feature uncertainty. In particular the MCMC generation method seems to produce similar dispersions to the PM method, while the bias evaluation highlighted a larger and less repeatable estimation between the subjects respect to the PM method as shown in Fig. 5.10



(a) (b)

Fig. 5.10 Mean bias and dispersions generation uncertainty comparison of jitter (a) and shimmer (b) for the three clinical classes.

In order to evaluate if the MCMC method produces artificial vowels whose stability metrics are correctly measured by the extraction algorithm, the extraction contribution was evaluated as summarised in Tabs. 5.14 and 5.15. The MCMC method shows a worst behaviour as concerning the biases which are larger and more dispersed between the subjects of each class. Again the evaluated dispersions seem compatible between the generation methods even though the ones relative to the PM method seem smaller. Such considerations can be noticed in Fig. 5.11

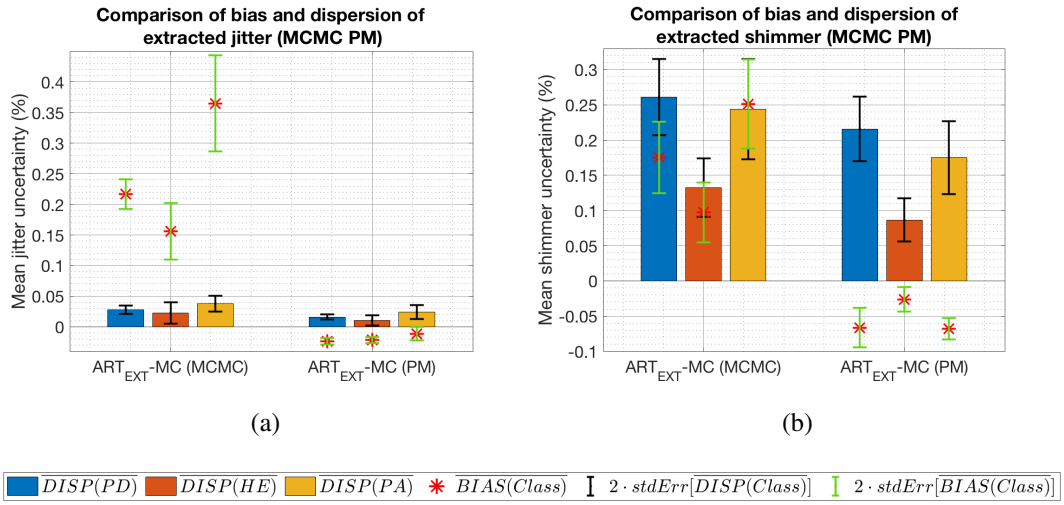


Fig. 5.11 Mean bias and dispersions extraction uncertainty comparison of jitter (a) and shimmer (b) for the three clinical classes.

For the mean CPPS evaluation the two analysed methods seem to be equivalent, even though the MCMC method produces less dispersed biases, as shown in Fig. 5.12.

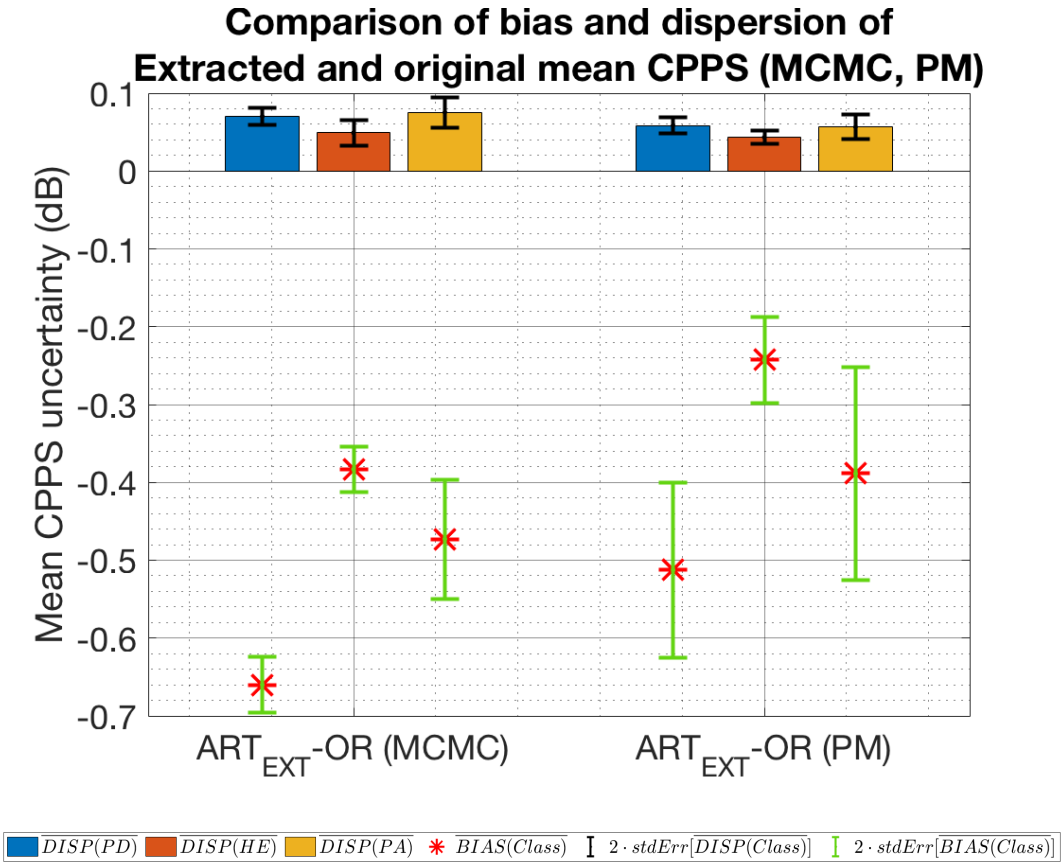


Fig. 5.12 Mean CPPS uncertainty comparison between generation methods

From the consideration made in this conclusive remarks the best generation method seems the perturbative method. For this reason, in the next section, just the PM was used to produce the artificial vowels.

5.4 Effects of the acquisition device on stability and CPPS metrics (PM)

The second part of the measuring chain uncertainty evaluation is focused on the acquisition device used to record the original voice samples. To evaluate the uncertainty contribution of the acquisition device, the test signals were converted into an electrical signal using a Digital to Analog Converter as shown in Fig. 5.13.

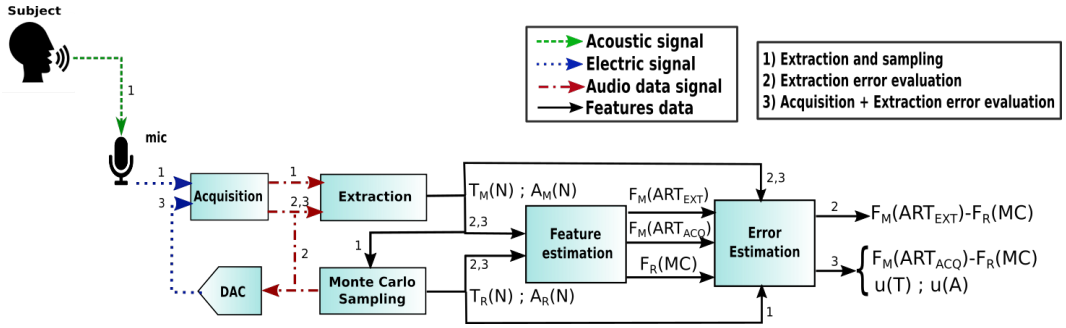


Fig. 5.13 Architecture of the evaluation method for the acquisition contribution

As shown in Fig. 5.14, the transformation of the vowel signals from the digital domain to the electrical domain was carried out using an audio interface device (MOTU Audio Express), which is connected with a USB cable to a computer. The output of the audio interface has a DC output impedance of about 100 Ohm and the acquisition device expects a plug-in power impedance (in the range of 1 kΩ to 3 kΩ). To simulate the presence of such an impedance, a 2.2 kΩ resistor was connected between the input terminal of the left channel and ground. For the same reason, the right channel was connected to ground with a 2.2 kΩ resistor.

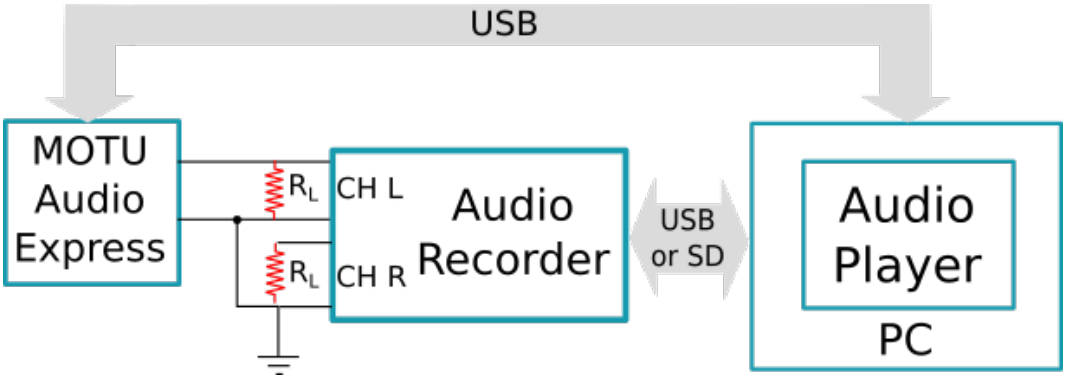


Fig. 5.14 Schematic of the acquisition contribution evaluation

In the scheme in Fig. 5.13, respect to the architecture depicted in Fig. 5.1, the measured artificial features can be compared between two domains:

- Digital domain Features $F_M(ART_{EXT})$
- Electrical domain Features $F_M(ART_{ACQ})$

For this evaluation, the MC-OR comparison is exactly like the comparison performed for the extraction uncertainty evaluation, so the bias and dispersion are the same of Tabs. 5.2, 5.3 and 5.4. To avoid excessive effort in recording 900 separate files, a long file was built concatenating the original vowels followed by ten repetitions of the relative artificial vowel. At the same time, a marker file that contains the starting and ending time of each vowel is created for later use. The file, which is almost 2 hours long, is played by a wave editor (e.g. Audacity) and recorded by the acquisition device. Once the recording was completed, the recorded file is extracted from the acquisition device memory and processed to produce separate audio files. In particular, a cross-correlation is performed between the digital and the acquired file to determine the gross delay between the two waves. In order to align each couple of vowel signals over time, a finer cross-correlation between the signal of the digital and acquired vowel is performed. In particular the digital vowel is trimmed between the time markers previously saved so the signal of the acquired vowel can be aligned using the cross-correlation.

5.4.1 Effects of the non-idealty of the chain

When there is an electrical connection between the output of the audio device and the input of the acquisition device some issues regarding the voltage scales of the two devices may arise. In fact, even tough we expect an unitary transformation when using a cable, the acquired signal is not identical to the expected digital signal. This fact are caused by some critical issues that can affect the measuring chain as well as the uncertainty evaluation chain (in this section the audio output device as shown in Fig. 5.13). In particular, such effect may be caused by:

- Gain error
- Offset error
- Sampling rate perturbations
- DAC-ADC Sampling rate mismatch
- Frequency response
- Chain non-linearity

As an example, in ideal conditions, the coincidence between the minimum and maximum values of the digital and the acquired signals is expected. In practice, even after a normalization, the minimum and maximum values could be different and the amplitude transfer function may not be unitary as shown in Fig. 5.15.

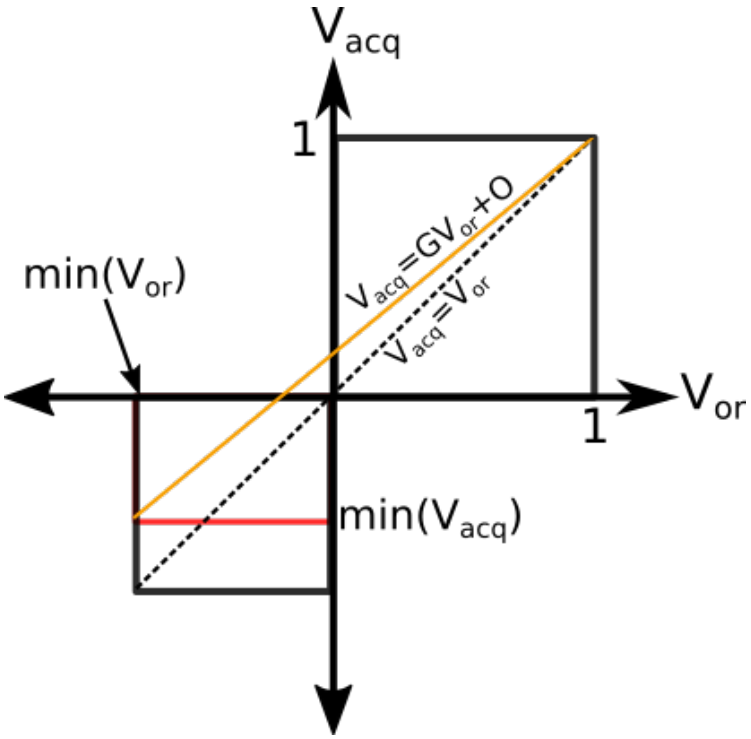


Fig. 5.15 An example of gain and offset error between an original and an acquired signal both normalised to 1 a.u.

The plot in Fig. 5.15 shows an example case where the original V_{or} and the acquired V_{acq} signals were normalized to 1 a.u. so the maximum value for both is 1 a.u.. In this example the effect of gain and offset error produces an overestimation of the minimum value of the acquired signal. In a research oriented application, where a trustable arbitrary function generator is available, the gain and offset errors can be attributed to the acquisition device and thus a characterization of this device is possible. If using a consumer level DAC as a reference device, such as the audio device used in this evaluation, the non-ideality of this chain cannot be characterized unless a trusted digital acquisition system is used to charecterize the DAC. Evaluating the gain and offset of such transfer function could help to correct the amplitude

values removing the offset O and dividing for the gain G as defined in the equations:

$$G = \frac{max(V_{acq}) - min(V_{acq})}{max(V_{or}) - min(V_{or})} \tag{5.9}$$

$$O = max(V_{acq}) - G \cdot max(V_{or}) \tag{5.10}$$

$$V^*_{acq} = \frac{V_{acq} - O}{G} \tag{5.11}$$

where V_{or} is the original digital amplitude, V_{acq} is the acquired digital amplitude and V^*_{acq} is the corrected acquired amplitude. In order to try to evaluate these issues, without using professional level instruments, a characterization of the whole feedback chain (Audio interface + Acquisition device) was carried out. To evaluate the gain and offset error of the device under test, a study on the artificial vowels dataset with 900 entries was performed in order to produce mean estimations of G and O for the three clinical classes:

Table 5.18 Mean Gain and Offset errors and their relative standard errors of the acquisition device

Class	$\overline{G}$	stdERR($\overline{G}$)	$\overline{O}$	stdERR($\overline{O}$)
PD	0.9980	0.0006	-0.0047	0.0006
HE	0.9978	0.0003	-0.0022	0.0003
PA	0.9992	0.0003	-0.0010	0.0003

As shown in Tab. 5.19, the Gain $\overline{G}$ are close to 1, even though a relative error of about 0.2 % was evaluated. The offsets $\overline{O}$ are close to 0 and their relative errors around 0.3 %.

In a practical application, such kind of method can lead to an inaccurate evaluations of the gain and offset errors because of the phase and spectral distortions introduced by the acquisition device or the audio interface as shown in Fig. 5.16:

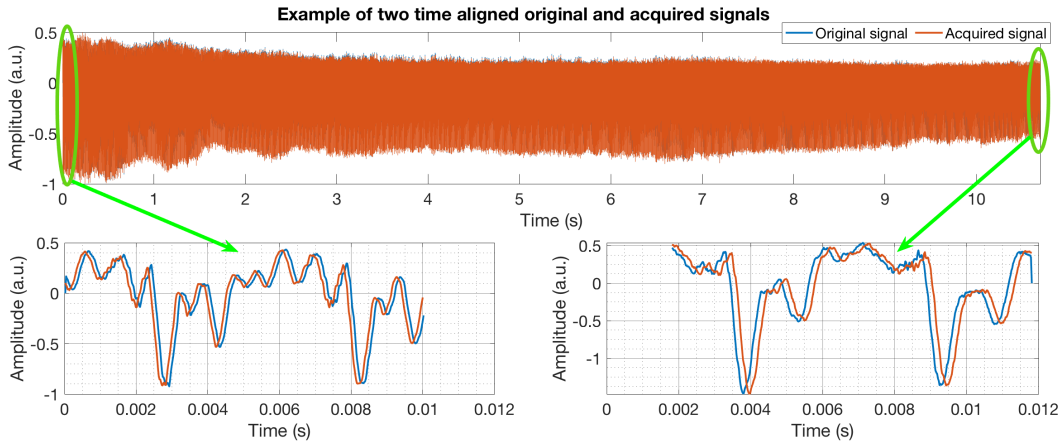


Fig. 5.16 An example of an original (in blue) and the acquired (in red) signals time aligned with the cross-correlation method

As shown in Fig. 5.16, after the time alignment carried out through the application of the cross-correlation method, the phase and the amplitude of the original and acquired signal may vary at different time instants. This may be partly due to a slight difference between the sampling rate of the DAC on the output audio interface and the sampling rate of the ADC on the input acquisition device (asynchronous sampling). Moreover the frequency response of the acquisition device (or the audio interface) may be not "flat" and some differences in the shape of the signal waves is expected between the original an the acquired signal.

To take into account such a misalignment, a regression method could be used to statistically evaluate gain and offset errors. As shown in Fig. 5.17 the phase and spectral difference between the time aligned signals can produce non unitary responses and the XY plot of the two signals highlights a variable phase alignment.

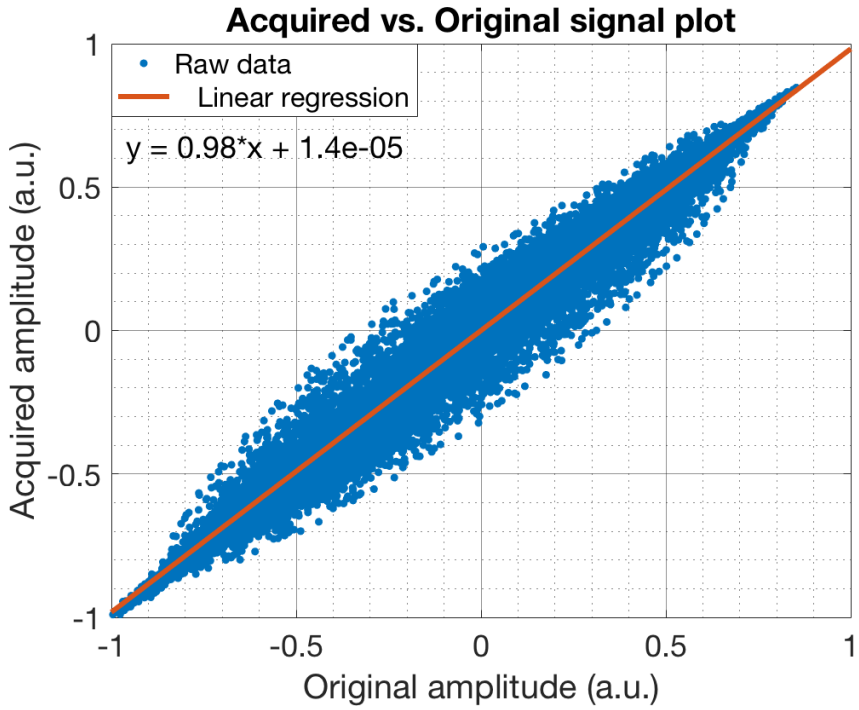


Fig. 5.17 An example of a scatter plot of an acquired vs. original signal

Such a technique can produce very different gain and offset error estimations, respect to the previously proposed method, as summarised in Tab. 5.19.

Table 5.19 Mean Gain and Offset errors and their relative standard errors of the acquisition device

Class	$\overline{G}$	stdERR($\overline{G}$)	$\overline{O}$	stdERR($\overline{O}$)
PD	0.9855	$6 \cdot 10^{-4}$	$1 \cdot 10^{-5}$	$2 \cdot 10^{-8}$
HE	0.9313	$3 \cdot 10^{-3}$	$1 \cdot 10^{-5}$	$5 \cdot 10^{-8}$
PA	0.9578	$2 \cdot 10^{-3}$	$1 \cdot 10^{-5}$	$4 \cdot 10^{-8}$

As shown in Tab. 5.19, the values of $\overline{G}$ lead to a relative gain error in a range between 2 % and 7 %. The evaluated offsets $\overline{O}$ are smaller than the ones evaluated with the previously proposed method and they produce a relative offset error of about 0.001 %.

This considerations may suggest that a correction method could be implemented to reduce the features uncertainties. For this work, such a correction did not take

place to evaluate the effects that these contributions have on the evaluation of voice features, as will be showed in the next section.

5.4.2 Evaluation of the acquisition contribution to period and amplitude uncertainty

As already done in the previous section, the evaluation of extracted pseudo-periods and amplitudes uncertainty was carried out on the signals recorded with the acquisition device. The data presented in Tab. 5.20 summarizes the results of this evaluation.

Table 5.20 Pseudo-periods and amplitudes mean extraction uncertainty of the (ACQ+EXT) contribution - $u(T)$, $u(A)$

Class	u(T) (μs)	u(A) (a.u.)
PD	38	0.04
HE	34	0.04
PA	38	0.04

As shown in Tab. 5.20 the period and amplitude uncertainty has increased slightly respect to the case of the extraction contribution ($u_{EXT}(T) \approx 31$ μs, $u_{EXT}(A) \approx 0.02$ a.u.).

5.4.3 Evaluation of the acquisition uncertainty contribution

To evaluate the acquisition uncertainty contribution, the bias were evaluated using Eq. 5.5. The results of this evaluation are summarised in Tab. 5.21 and Tab. 5.22

Table 5.21 Measured artificial bias of ACQ+EXT contribution - $\overline{BIAS_{ART_{ACQ}-MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	-0.02	-1.1	-0.02	-0.02	0.03	-0.3	-0.03	-0.13	0.03
HE	-0.02	-1.2	-0.02	-0.01	0.02	-0.1	-0.01	-0.04	0.23
PA	-0.01	-0.7	-0.01	-0.01	0.02	-0.1	-0.01	-0.08	0.03

Table 5.22 Measured artificial dispersion of ACQ+EXT contribution - $\overline{DISP_{ART_{ACQ}}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.01	0.9	0.009	0.010	0.02	0.19	0.017	0.13	0.15
HE	0.01	0.8	0.008	0.007	0.01	0.08	0.007	0.05	0.08
PA	0.02	1.1	0.014	0.015	0.02	0.15	0.014	0.09	0.13

To evaluate the results summarized in Tabs. 5.21 and 5.22, the bias and dispersion of the ACQ+EXT contribution are compared to the contributions that are due to the extraction (EXT) algorithm only, obtaining the outcomes summarized in Fig. 5.18.

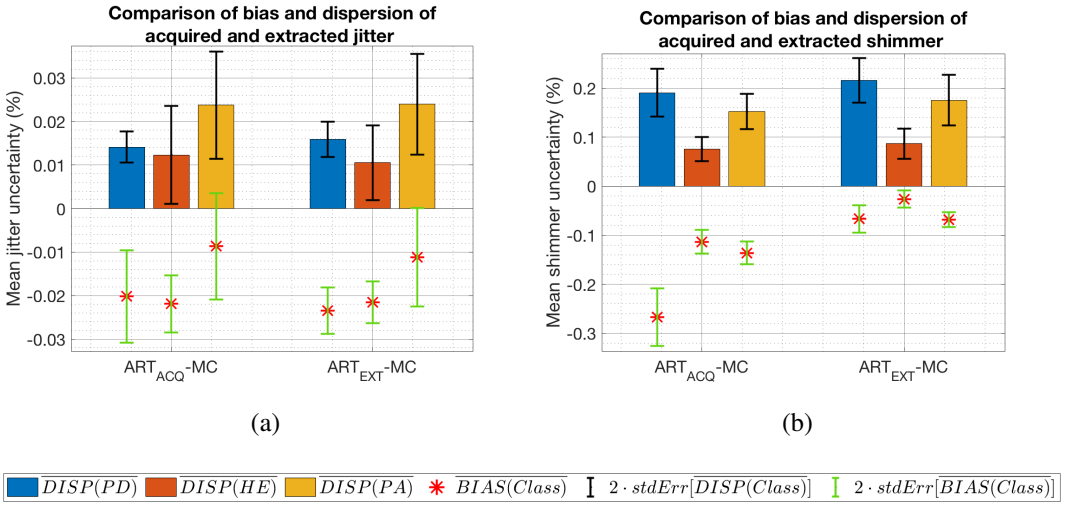


Fig. 5.18 (ACQ+EXT) contribution of mean bias and dispersion evaluations of jitter (a) and shimmer (b) for the three clinical classes.

As shown in Fig. 5.18, the $ART_{ACQ} - MC$ dispersions are compatible with those of the $ART_{ACQ} - MC$ evaluation. The biases for the jitter evaluation are compatible with the contribution comparison $ART_{ACQ} - MC$, while for the shimmer evaluation higher negative biases are clearly visible. Such effect is due to the gain and offset error as described in Sec. 5.4.1. The evaluation of the dispersion of the (ACQ+EXT) contributions are very similar to the dispersion of the EXT contribution only, so a negligible effect is expected on stability metrics. The evaluation of the extraction bias allows to correct such an effect if a sufficient number of artificial vowels is produced, as will be showed in the next chapter.

5.4.4 Evaluation of the acquisition contribution to CPPS features uncertainty

As already done for the extraction contribution, the effect of the acquisition device on CPPS features uncertainty was analysed. The results of this analysis are presented in Tabs. 5.23 and 5.24:

Table 5.23 Measured artificial bias of CPPS metrics for the (ACQ+EXT) contribution - $\overline{BIAS_{ART_{ACQ}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	-0.6	-0.6	-0.5	0.093	0.3	-0.7	-0.4	0.06	-0.19
HE	-0.2	-0.1	-0.2	-0.004	0.2	-0.2	-0.3	-0.13	0.38
PA	-0.5	-0.6	-0.6	0.013	0.3	-0.5	-0.5	0.06	-0.04

Table 5.24 Measured artificial dispersions of CPPS metrics for the (ACQ+EXT) contribution - $\overline{DISP_{ART_{ACQ}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.06	0.08	0.6	0.06	0.7	0.2	0.10	0.13	0.4
HE	0.04	0.05	0.4	0.03	0.6	0.1	0.07	0.09	0.3
PA	0.06	0.08	0.5	0.05	0.6	0.1	0.13	0.11	0.2

To evaluate the effect of the acquisition device on the mean CPPS measurements, a comparison between the EXT and the ACQ+EXT contributions is depicted in Fig. 5.19.

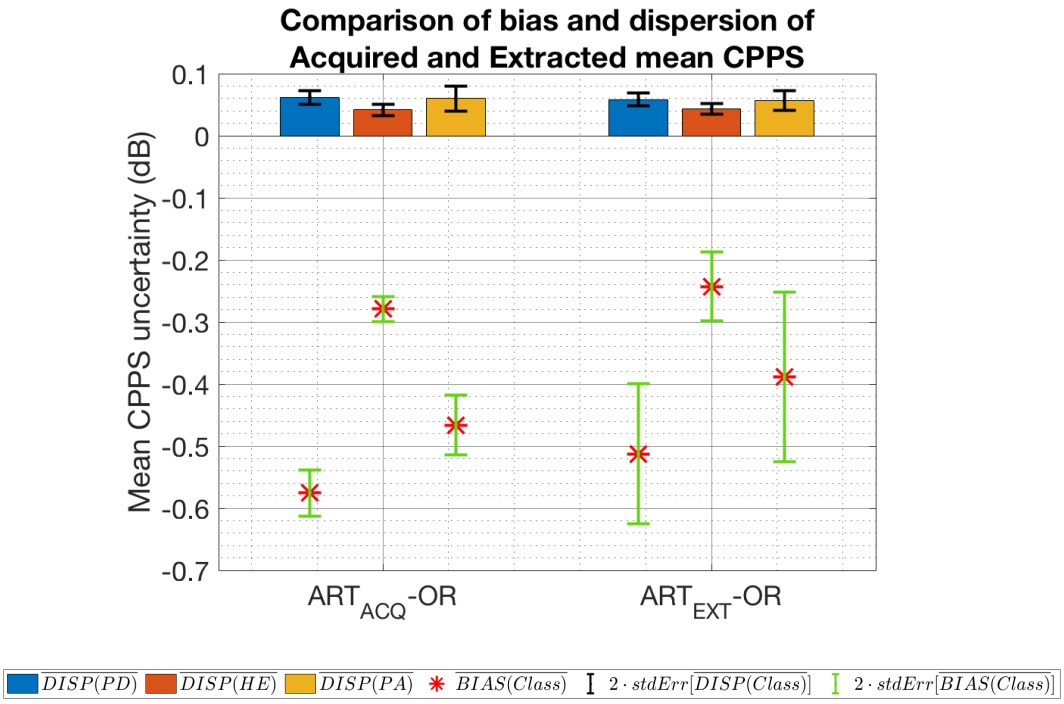


Fig. 5.19 Mean CPPS uncertainty comparison between the ACQ+EXT and EXT contribution

As shown in Fig. 5.19, the dispersion of the mean CPPS are comparable with the dispersion of the EXT contribution. The evaluated biases are comparable with the EXT contribution and their repeatability (the size of the green bars) is increased respect with the EXT contribution. For this reason the ACQ contribution to mean CPPS uncertainties can be considered as negligible.

5.5 Evaluation of the whole chain contribution on stability and CPPS metrics

The third part of the measuring chain uncertainty evaluation is focused on the microphone used to capture the acoustic waves. In Fig. 5.20, the architecture of the evaluation method is presented.

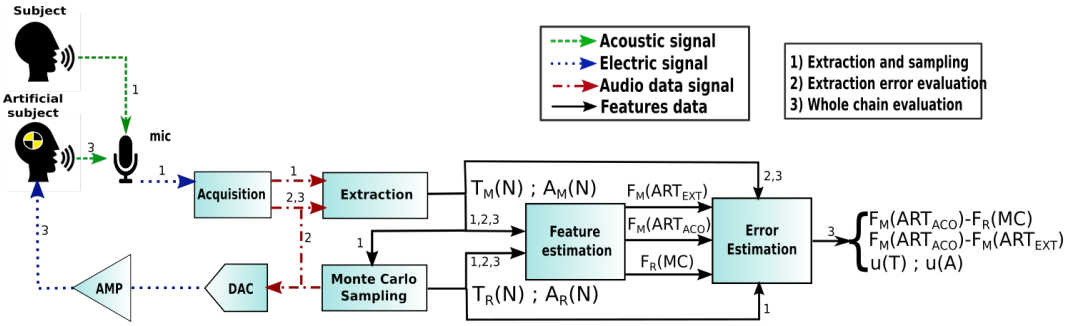
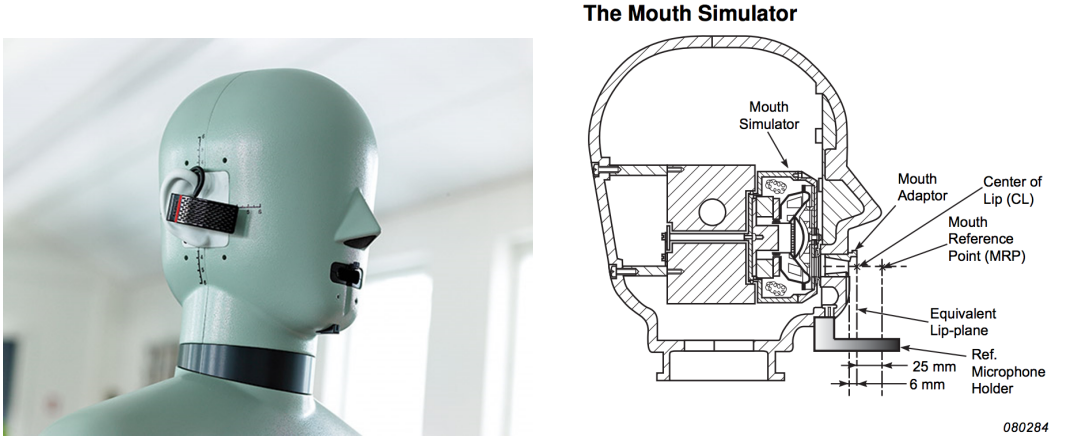


Fig. 5.20 Architecture of the whole chain uncertainty contributions evaluation

To evaluate the uncertainty contribution of the whole chain, the test signals were converted into an electrical signal by using an audio interface device (MOTU Audio Express) connected with a USB cable to a computer. The output of the audio device was connected to a power amplifier (Alpine MRP-F200) powered with a 12 V battery. The output of the audio amplifier goes to the input of a torso simulator (Brüel & Kjær HATS Head and Torso Simulator type 5128) as shown in Fig. 5.21

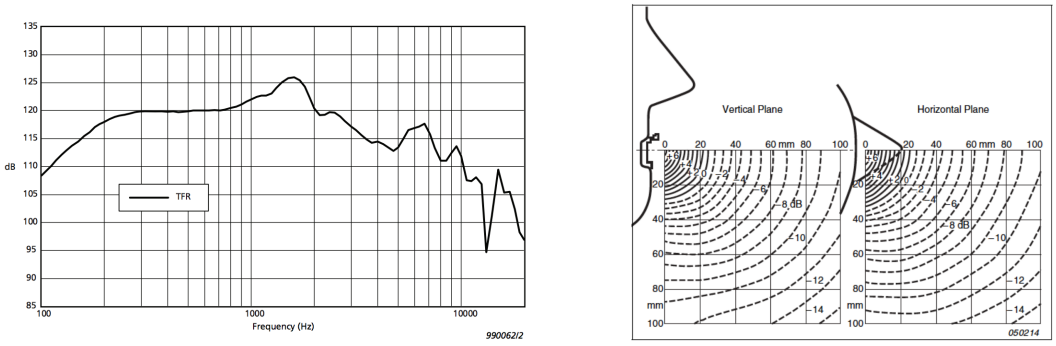


(a) (b)

Fig. 5.21 The head piece of the HATS with a bluetooth earset (a) and a schematic diagram of the mouth simulator (b)

The torso simulator represents the final link in the uncertainty evaluation chain and the first of the measuring chain. Such a torso simulator takes the place of the subject, who cannot produce the input stimulus with an acceptable repeatability. The torso simulator, instead, produces acoustic waves that are highly repeatable and reproducible in terms of extracted features. All the specifications of the HATS are

known and extensively reported in the specifications manual. As an example, the Talk Frequency Response (TFR) at 25 mm from the mouth opening is reported in Fig. 5.22 (a) along the sound pressure level distribution around the mouth (b).



(a) (b)

Fig. 5.22 HATS in-axis frequency response (a) and sound pressure level distribution around the mouth simulator (b)

The TFR evaluated by the manufacturer at a distance 25 mm has a fairly "flat" response in the frequency range of a sustained vowel. The microphone distance of 25 mm used in this evaluation has the same order of magnitude of the distance of the microphone in air adopted in the recordings of the vowels used in this work (between 20 mm and 40 mm).

The HATS was placed inside the anechoic room of the Department of Energy of the Polytechnic of Turin and the same microphone in air used in the collection of vowel samples, a cheek headset microphone (**CM**, MIPRO MU 55-HR), was placed on the HATS head. Moreover, a reference microphone (**RM**, NTI Audio M4261) was placed at 1 meter from the HATS mouth. Another measurement was performed using a smartphone (**SP**, Iphone X) placed in the standard video-call position. The cheek microphone were tested in 4 different positions around the HATS mouth. To sum up, a total of 6 evaluations was carried out to evaluate the whole chain uncertainty contributions:

1. **CM** Position 1 (golden standard): 20 mm from the mouth opening, in axis with its center (Fig. 5.23)
2. **CM** Position 2: 20 mm from the mouth opening, 20 mm above the mouth plane (Fig. 5.26)

3. **CM** Position 3 20 mm from the mouth opening, 20 mm under the mouth plane (Fig. 5.27)
4. **CM** Position 4: 40 mm from the mouth opening, in axis with its right corner (Fig. 5.28)
5. **RM** Reference microphone: 1 m from the mouth opening (5.29)
6. **SP** IphoneX: 50 cm from the mouth opening (5.30)

With this setup, the evaluation of the measurement uncertainty includes the acoustic domain and it makes possible to study the possible perturbations of features caused by microphone placement and type.
The input gain of the acquisition device, connected to the cheek microphone, was set to produce a measured full-scale level of -6 dB_{fs} indicated by the acquisition device display. In a similar way, the input gain of the audio interface, connected to the reference microphone, was set to measure a full-scale level of -6 dB_{fs}, as indicated by the audio interface display. As regarding the smartphone recordings, they were carried out using the iOS application "Voice Memos" and the smartphone was placed 50 cm from the HATS face. The position of the smartphone was set in order to place the HATS face in the center of the frame captured by the smartphone front video-camera. In this way the standard position of a video-call was simulated. All the evaluations were carried out setting an appropriate output gain on the uncertainty evaluation chain (DAC + AMP in Fig. 5.20). The output gain of the uncertainty evaluation chain was set in order to produce a Sound Pressure Level (SPL) around 70 dB [40] measured by a calibrated Sound Level Meter placed at a distance of 1 m from the mouth opening.

5.5.1 Microphone position 1 (golden standard)

The first evaluation tries to replicate the experimental conditions of the original acquisition. The cheek microphone is placed at 20 mm from the mouth opening in axis with its center, aligned with the tip of the nose. This microphone position is approximately the same used in each of the original recordings, for this reason this evaluation is considered as the golden standard respect to the other 3 positions. In Fig. 5.23 the positioning of the cheek microphone on the HATS head is shown.

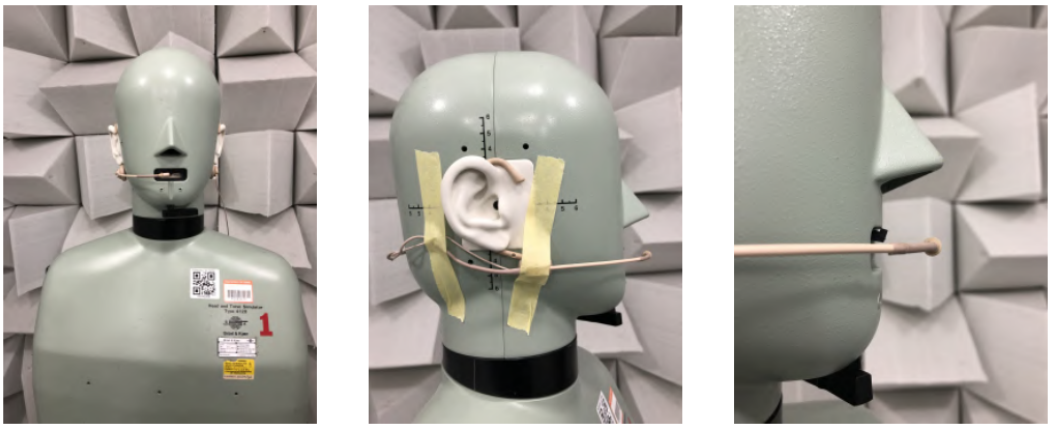


Fig. 5.23 Microphone position 1

As already done in the previous sections, bias and dispersion of stability metrics extracted from the 900 artificial signals are presented in Tabs. 5.25 and 5.26. The term ACO refers to the contribution evaluated in the acoustic domain, which coincides with the whole chain contribution.

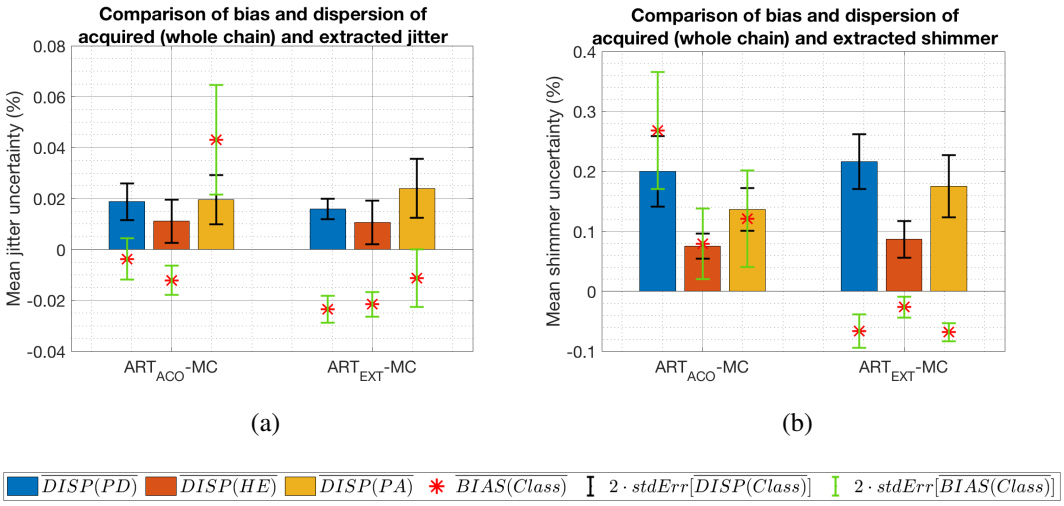
Table 5.25 Measured artificial bias of the whole chain contribution (CM position 1) - $\overline{BIAS_{ART_{ACO}-MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	-0.004	-0.3	-0.004	-0.011	0.01	0.27	0.017	0.35	2.9
HE	-0.012	-0.6	-0.011	-0.009	0.02	0.08	0.005	0.09	1.6
PA	0.043	1.7	0.022	0.019	0.05	0.12	0.008	0.10	1.2

Table 5.26 Measured artificial dispersions of the whole chain contribution (CM position 1) - $\overline{DISP_{ART_{ACO}}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.01	1.20	0.012	0.012	0.02	0.20	0.018	0.14	0.18
HE	0.01	0.71	0.007	0.007	0.02	0.07	0.007	0.05	0.15
PA	0.02	0.93	0.011	0.014	0.02	0.14	0.012	0.09	0.14

As can be noted in Fig. 5.24, the biases of the whole chain are higher and more dispersed than the biases of the extraction contribution (short chain). The dispersion is clearly comparable to the extraction contribution dispersion.



(a) (b)

Fig. 5.24 Mean bias and dispersion evaluations of jitter (a) and shimmer (b) for the three clinical classes of the whole chain contribution (CM position 1)

Evaluation of the whole chain contribution to period and amplitude uncertainty

As already done in the previous sections, the evaluation of extracted pseudo-periods and amplitudes uncertainty was carried out on the acoustic waves emitted by the HATS and recorded with the acquisition device. The data presented in Tab. 5.27 summarizes the results of such evaluation.

Table 5.27 Pseudo-periods and amplitudes mean uncertainty of the whole chain contribution (CM position 1) - u(T), u(A)

Class	u(T) (μs)	u(A) (a.u.)
PD	43	0.1
HE	37	0.1
PA	41	0.1

As shown in Tab. 5.27, the period uncertainty is slightly larger than the one evaluated for the extraction and the acquisition contribution. The amplitude uncertainty for the whole chain contribution instead, is an order of magnitude larger than the extraction contribution. This is caused by the offset and gain error as shown in Sec. 5.4.1.

Evaluation of the whole chain uncertainty contribution to CPPS features

The effect of the whole chain on CPPS metrics was investigated as already done in the previous sections. The results of this evaluation are reported in Tabs. 5.28 and 5.29:

Table 5.28 Measured artificial bias of CPPS metrics of the whole chain contribution (CM position 1) - $\overline{BIAS_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	-0.6	-0.6	-0.6	0.05	0.2	-0.7	-0.5	0.060	-0.12
HE	0.8	0.8	0.8	-0.02	0.3	0.7	0.7	-0.160	0.52
PA	0.2	0.2	0.2	-0.05	0.1	0.3	0.2	0.002	0.01

Table 5.29 Measured artificial dispersion of CPPS metrics of the whole chain contribution (CM position 1) - $\overline{DISP_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.07	0.08	0.5	0.06	0.7	0.2	0.12	0.13	0.4
HE	0.07	0.07	0.4	0.04	0.6	0.1	0.09	0.09	0.3
PA	0.08	0.10	0.5	0.05	0.6	0.1	0.12	0.11	0.2

As shown in Fig. 5.25 higher biases are noticeable in the HE and PA classes for mean CPPS measurements. The dispersion of the mean CPPS, instead is comparable with the dispersion of the extraction contribution.

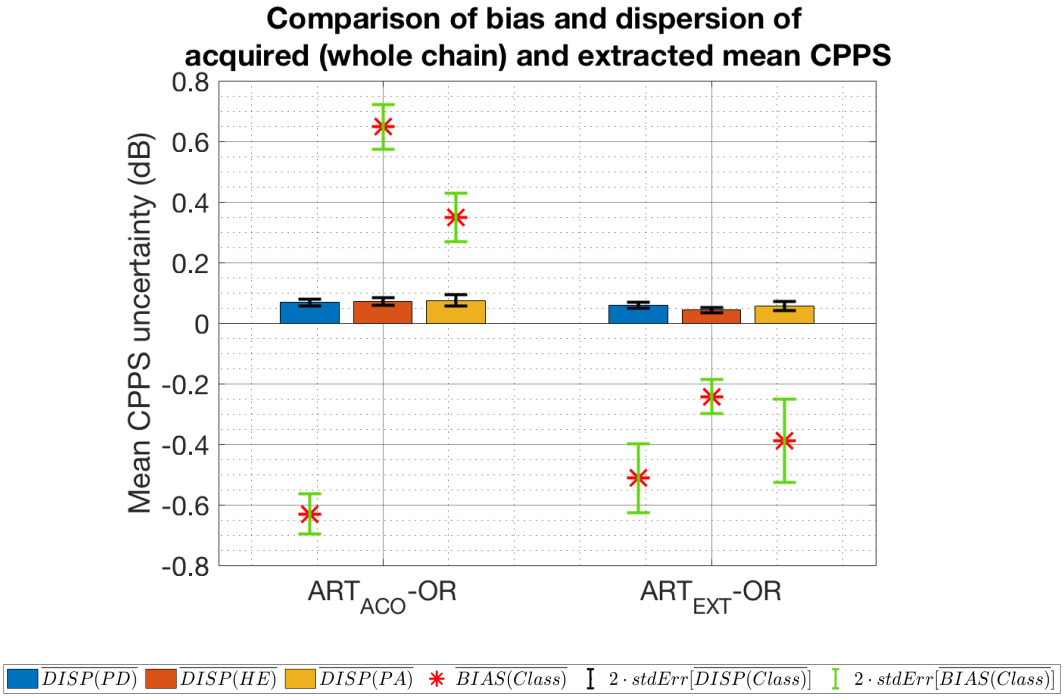


Fig. 5.25 Mean bias and dispersion evaluations of Mean CPPS of the whole chain contribution (CM position 1)

For the next microphone evaluations, the sequence of the presented data will be the same presented in this section:

- Evaluation of the whole chain uncertainty contribution on stability metrics (bias and dispersion)
- Evaluation of the whole chain uncertainty contribution to CPPS features (bias and dispersion)

5.5.2 Microphone position 2

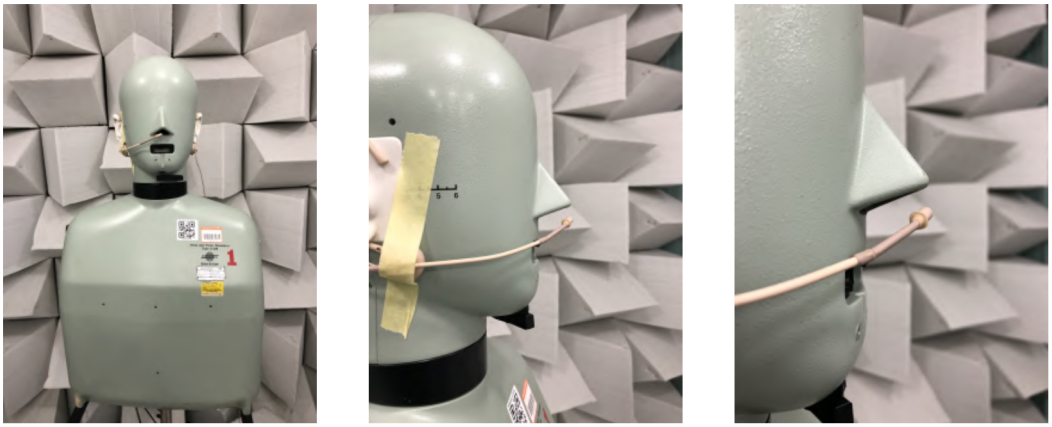


Fig. 5.26 Microphone position 2

Table 5.30 Measured artificial bias of the whole chain contribution (CM position 2) - $\overline{BIAS_{ART_{ACO}-MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	-0.002	-0.2	-0.003	-0.007	0.009	0.6	0.04	0.55	3.5
HE	-0.014	-0.7	-0.013	-0.010	0.022	0.1	0.01	0.14	1.8
PA	0.048	2.0	0.025	0.022	0.054	0.3	0.03	0.23	1.6

Table 5.31 Measured artificial dispersions of the whole chain contribution (CM position 2) - $\overline{DISP_{ART_{ACO}}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.02	1.1	0.011	0.011	0.02	0.21	0.019	0.14	0.19
HE	0.01	0.8	0.008	0.008	0.08	0.07	0.007	0.04	0.09
PA	0.02	0.9	0.012	0.013	0.02	0.14	0.012	0.09	0.14

Evaluation of the whole chain uncertainty contribution to CPPS features

Table 5.32 Measured artificial bias of CPPS metrics of the whole chain contribution (CM position 2) - $\overline{BIAS_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	-0.5	-0.5	-0.4	0.07	0.3	-0.6	-0.4	0.06	-0.09
HE	1.0	1.1	1.0	0.01	0.6	1.0	1.0	-0.09	0.44
PA	1.0	1.1	1.2	-0.02	0.2	0.9	0.9	-0.10	0.04

Table 5.33 Measured artificial dispersions of CPPS metrics of the whole chain contribution (CM position 2) - $\overline{DISP_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.07	0.08	0.5	0.06	0.7	0.2	0.1	0.1	0.5
HE	0.13	0.14	0.4	0.05	0.6	0.2	0.2	0.1	0.3
PA	0.32	0.33	0.6	0.08	0.7	0.3	0.3	0.1	0.3

5.5.3 Microphone position 3

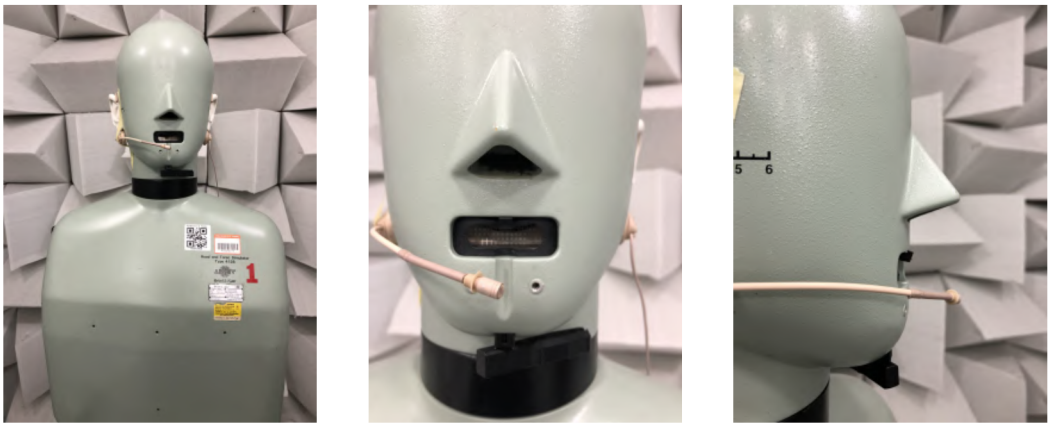


Fig. 5.27 Microphone position 3

Table 5.34 Measured artificial bias of the whole chain contribution (CM position 3) - $\overline{BIAS_{ART_{ACO}-MC}(class)}$

Class	*jit* (%)	jit_{abs} (µs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	-0.002	-0.2	-0.003	-0.009	0.02	0.35	0.023	0.41	3.2
HE	-0.011	-0.5	-0.011	-0.008	0.03	0.08	0.005	0.09	1.6
PA	0.041	1.6	0.020	0.018	0.05	0.11	0.006	0.01	1.3

Table 5.35 Measured artificial dispersion of the whole chain contribution (CM position 3) - $\overline{DISP_{ART_{ACO}}(class)}$

Class	*jit* (%)	jit_{abs} (µs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.02	1.3	0.014	0.014	0.03	0.22	0.021	0.15	0.64
HE	0.01	0.7	0.007	0.006	0.01	0.07	0.007	0.04	0.09
PA	0.02	0.9	0.012	0.014	0.02	0.14	0.013	0.09	0.14

Evaluation of the whole chain uncertainty contribution to CPPS features

Table 5.36 Measured artificial bias of CPPS metrics of the whole chain contribution (CM position 3) - $\overline{BIAS_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.06	3.4	0.04	0.05	-0.05	1.5	0.13	1.1	2.8
HE	0.03	2.1	0.02	0.03	-0.02	0.7	0.06	0.5	1.1
PA	0.11	4.9	0.06	0.08	0.04	0.8	0.07	0.6	0.9

Table 5.37 Measured artificial dispersion of CPPS metrics of the whole chain contribution (CM position 3) - $\overline{DISP_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.02	1.3	0.014	0.014	0.03	0.22	0.021	0.15	0.64
HE	0.01	0.7	0.007	0.006	0.02	0.07	0.007	0.04	0.09
PA	0.02	0.9	0.012	0.014	0.02	0.14	0.013	0.09	0.14

5.5.4 Microphone position 4

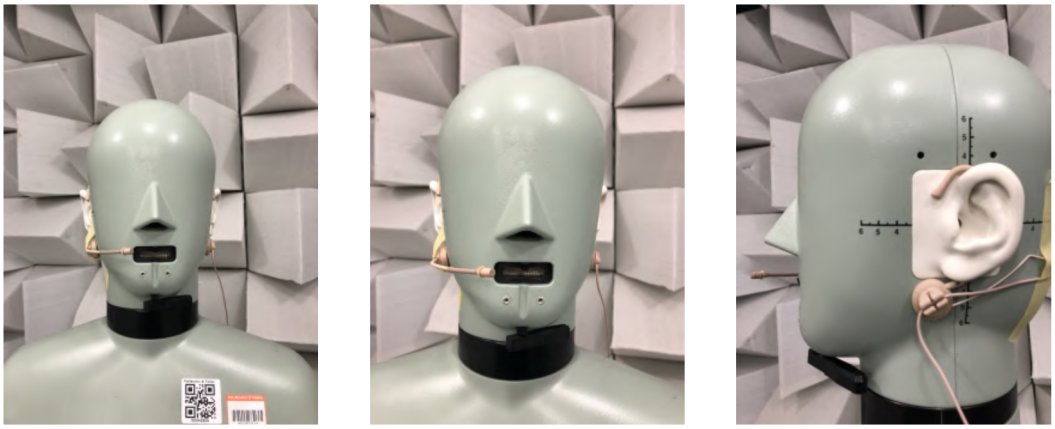


Fig. 5.28 Microphone position 4

Table 5.38 Measured artificial bias of the whole chain contribution (CM position 4) - $\overline{BIAS_{ART_{ACO}-MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.002	-0.03	0.00001	-0.007	0.01	0.5	0.036	0.5	3.4
HE	-0.011	-0.49	-0.01100	-0.008	0.03	0.1	0.008	0.1	1.7
PA	0.049	2.00	0.02600	0.024	0.05	0.2	0.015	0.2	1.5

Table 5.39 Measured artificial dispersion of the whole chain contribution (CM position 4) - $\overline{DISP_{ART_{ACO}}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.02	1.1	0.01	0.01	0.02	0.21	0.02	0.14	0.18
HE	0.01	0.6	0.01	0.01	0.02	0.08	0.01	0.05	0.09
PA	0.02	0.9	0.01	0.01	0.02	0.14	0.01	0.09	0.15

Evaluation of the whole chain uncertainty contribution to CPPS features

Table 5.40 Measured artificial bias of CPPS metrics of the whole chain contribution (CM position 4) - $\overline{BIAS_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	-0.5	-0.5	-0.5	0.06	0.24	-0.6	-0.3	0.08	-0.14
HE	0.8	0.8	0.8	0.02	0.65	0.8	0.8	-0.06	0.42
PA	0.7	0.7	0.77	-0.04	0.04	0.7	0.6	-0.07	0.04

Table 5.41 Measured artificial dispersion of CPPS metrics of the whole chain contribution (CM position 4) - $\overline{DISP_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.1	0.1	0.6	0.07	0.7	0.2	0.1	0.1	0.3
HE	0.1	0.1	0.4	0.04	0.7	0.1	0.1	0.1	0.2
PA	0.2	0.2	0.6	0.07	0.7	0.2	0.2	0.1	0.3

5.5.5 Reference microphone

Fig. 5.29 Acoustic uncertainty contribution evaluation for a reference microphone

Table 5.42 Measured artificial bias of the whole chain contribution (RM) - $\overline{BIAS_{ART_{ACO}-MC}(class)}$

Class	*jit* (%)	jit_{abs} (µs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.1	0.1	0.6	0.07	0.7	0.2	0.2	0.1	0.3
HE	0.1	0.1	0.4	0.04	0.7	0.1	0.1	0.1	0.3
PA	0.2	0.2	0.6	0.07	0.7	0.2	0.2	0.1	0.3

Table 5.43 Measured artificial dispersion of the whole chain contribution (RM) - $\overline{DISP_{ART_{ACO}}(class)}$

Class	*jit* (%)	jit_{abs} (µs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.02	1.2	0.01	0.01	0.02	0.2	0.02	0.1	0.2
HE	0.02	1.0	0.01	0.01	0.05	0.2	0.02	0.1	0.3
PA	0.02	1.1	0.01	0.02	0.03	0.2	0.02	0.1	0.1

Evaluation of the whole chain uncertainty contribution to CPPS features

Table 5.44 Measured artificial bias of CPPS metrics of the whole chain contribution (RM) - $\overline{BIAS_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	-0.3	-0.3	-0.2	0.07	0.3	-0.4	-0.2	0.08	-0.17
HE	0.5	0.5	0.4	0.07	0.9	0.3	0.6	-0.07	0.40
PA	0.8	0.8	0.9	-0.01	0.2	0.8	0.77	-0.08	0.02

Table 5.45 Measured artificial dispersion of CPPS metrics of the whole chain contribution (RM)- $\overline{DISP_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.1	0.1	0.5	0.1	0.8	0.2	0.2	0.1	0.4
HE	0.3	0.3	0.5	0.1	0.7	0.4	0.3	0.1	0.3
PA	0.2	0.2	0.6	0.1	0.7	0.3	0.2	0.1	0.3

5.5.6 Smartphone microphone

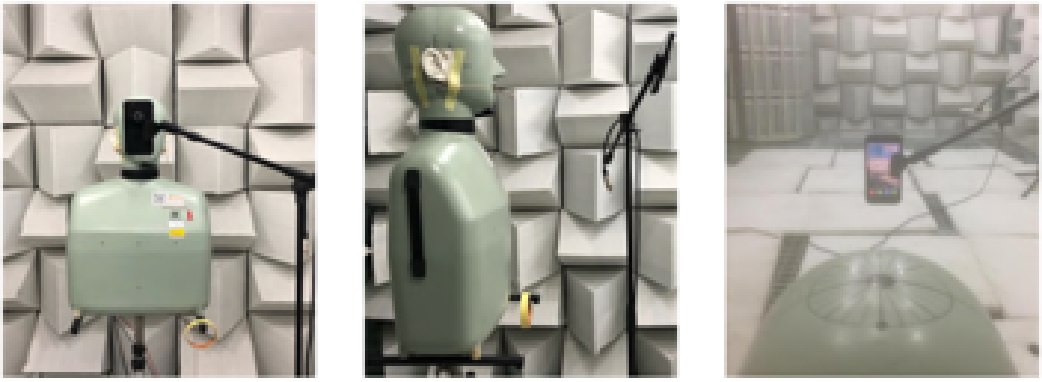


Fig. 5.30 Acoustic uncertainty contribution evaluation for smartphone microphone

Table 5.46 Measured artificial bias of the whole chain contribution (SP) - $\overline{BIAS_{ART_{ACO}-MC}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	-0.05	-2.9	-0.03	-0.03	-0.02	-0.1	-0.02	0.09	2.1
HE	-0.03	-1.9	-0.02	-0.02	0.03	-0.2	-0.02	-0.12	1.1
PA	-0.03	-1.7	-0.02	-0.02	0.07	-0.5	-0.05	-0.34	-1.2

Table 5.47 Measured artificial dispersion of the whole chain contribution (SP) - $\overline{DISP_{ART_{ACO}}(class)}$

Class	*jit* (%)	jit_{abs} (μs)	rap (%)	ppq (%)	vf_0 (%)	*shi* (%)	shi_{abs} (dB)	apq (%)	vAm (%)
PD	0.01	0.9	0.010	0.010	0.02	0.20	0.018	0.13	0.1
HE	0.01	0.9	0.009	0.008	0.03	0.08	0.008	0.05	0.2
PA	0.02	1.3	0.016	0.016	0.03	0.14	0.013	0.09	0.1

Evaluation of the whole chain uncertainty contribution to CPPS features

Table 5.48 Measured artificial bias of CPPS metrics of the whole chain contribution (SP) - $\overline{BIAS_{ART_{ACO}-OR}(class)}$

Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	-1.20	-1.20	-1.20	0.07	0.20	-1.30	-1.10	0.10	-0.19
HE	0.02	0.03	-0.04	-0.02	0.36	0.05	-0.04	-0.07	0.42
PA	-0.04	-0.01	0.04	-0.05	-0.02	-0.01	-0.16	-0.09	0.04

Table 5.49 Measured artificial dispersion of CPPS metrics of the whole chain contribution (SP) - $\overline{DISP_{ART_{ACO}-OR}(class)}$

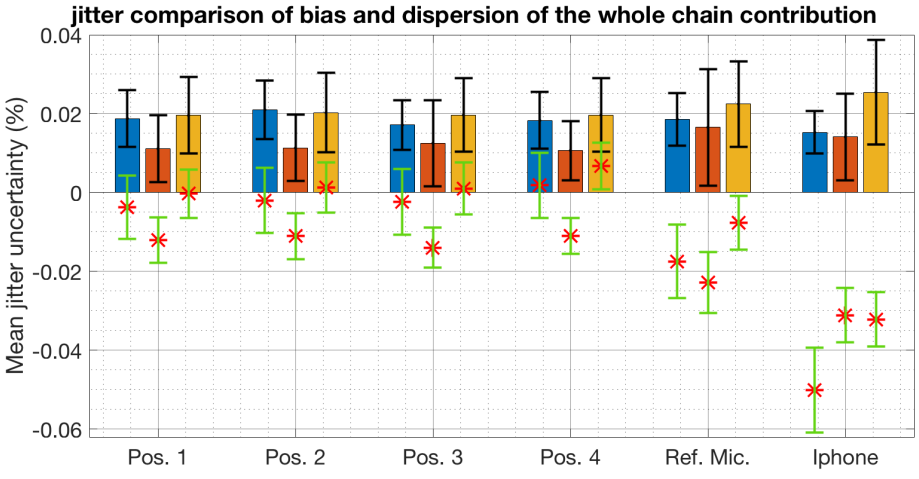
Class	Mean (dB)	Median (dB)	Mode (dB)	Std (dB)	Range (dB)	5°perc. (dB)	95°perc. (dB)	Skewness (%)	Kurtosis (%)
PD	0.11	0.12	0.5	0.06	0.7	0.2	0.14	0.1	0.4
HE	0.07	0.07	0.4	0.04	0.6	0.1	0.09	0.1	0.3
PA	0.11	0.12	0.5	0.05	0.6	0.2	0.15	0.1	0.2

5.6 Final considerations and comparisons on the whole chain contribution

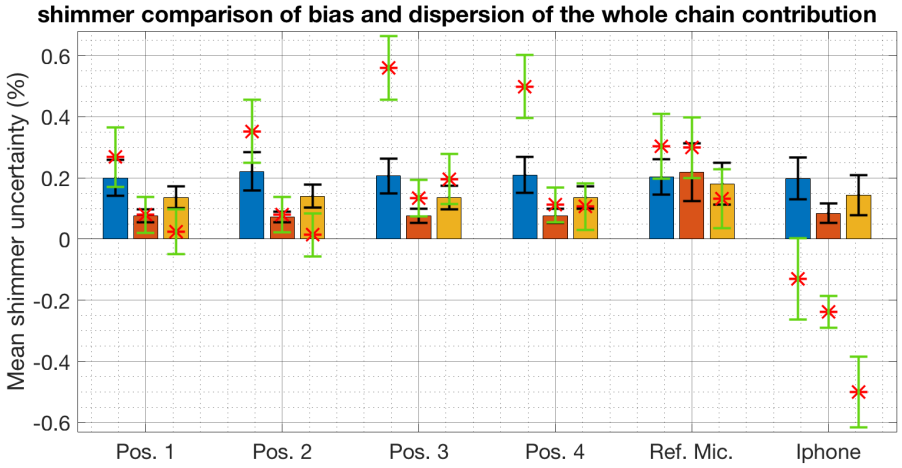
In this section a comparison between the different microphone positioning and types will be performed in order to determine if microphone perturbations can alter the evaluation of biases and dispersions. A comparison between the uncertainty evaluations for different measuring chain lengths will be also presented.

5.6.1 Microphone positioning and type comparison

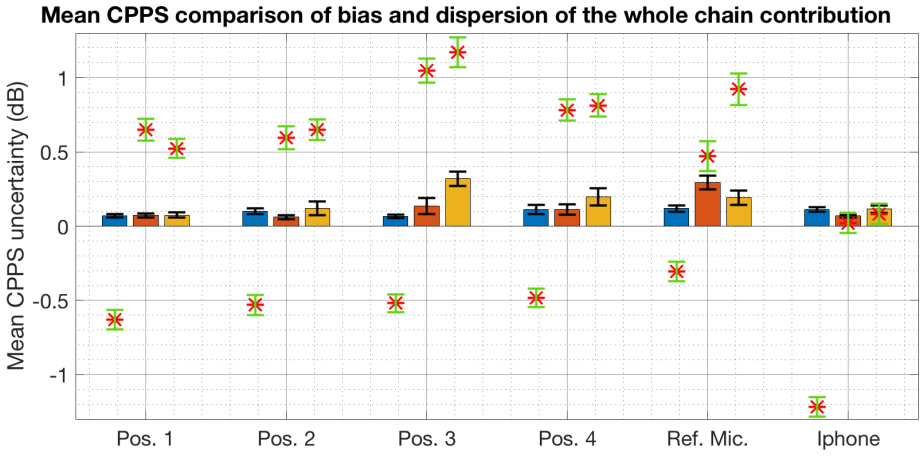
As showed in Tabs. from 5.25 to 5.46, the dispersion of the evaluated features seem to not depend on the microphone positioning or type. Such consideration is highlighted in the plot in Fig. 5.31, where the dispersion parameters for jitter, shimmer evaluations seem comparable across each microphone positioning and type. The biases of jitter evaluations are clearly comparable between the different microphone positions and also with the reference microphone. The jitter bias obtained in the evaluation of the Iphone microphone is the only one that is not comparable with the other microphone positions and the reference microphone. Regarding the shimmer evaluation some more variety on the biases is present between the comparisons. The mean CPPS evaluations showed more varied dispersion and biases which seem to depend on the clinical class. In particular the PD class shows negative biases while the HE and PA class highlights positive biases as shown in Fig. 5.31 (c).



(a)



(b)



(c)

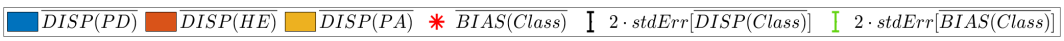
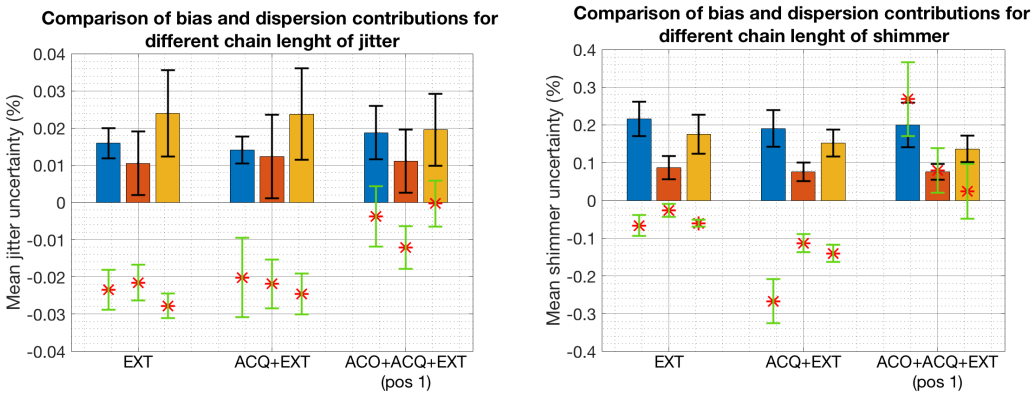


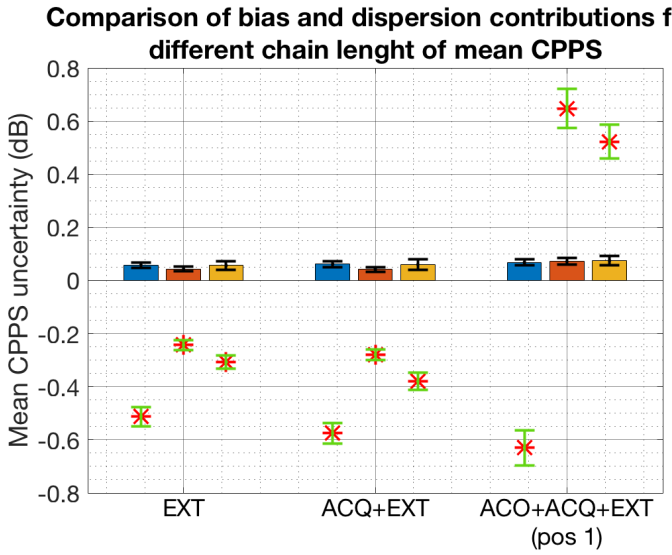
Fig. 5.31 A comparison between different microphone positioning and types for jitter (a), shimmer (b), and CPPS (c) evaluations

5.6.2 Effects of the measuring chain length on the features uncertainty

To evaluate the contributions of the measuring chain components to the total feature uncertainty, a comparison between the evaluations performed in sections 5.1, 5.4 and 5.5 is presented. As shown in Fig. 5.32, no difference is highlighted between the dispersion evaluations while some difference is noticeable in the bias evaluations. In particular the biases seem compatible for the Extraction (EXT) and the Acquisition+Extraction (ACQ+EXT) evaluation for jitter and mean CPPS measurements. The bias for shimmer and mean CPPS evaluations seem to change drastically as the measuring chain length increases.



(a) (b)



(c)

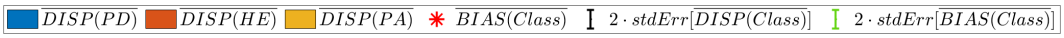


Fig. 5.32 A comparison between different chain lengths for jitter (a), shimmer (b), and CPPS (c) evaluations

Chapter 6

Machine learning algorithms

In this chapter the uncertainty evaluations performed in the previous sections will be used to implement a simple machine learning algorithm. Such an algorithm produces binary classifications that separate Parkinsonian subjects from Healthy and Pathological non-parkinsonian subjects.

6.1 The logistic regression

The logistic regression (LR) is a non-linear statistical model which is used to separate binary variables as in the case of pathological subject respect to an healthy control group. The logistic regression belongs to the class of generalized linear models (GLM) that uses the logistic function to model a binary dependent variable. In GLM regression analysis, this function is called *link function* because it transforms a linear combination into the desired target function. The logarithm of the odds (log-odds) for the positive class is a linear combination of independent variables X_i called predictors:

$$l = \log\frac{p}{1-p} = \Theta^T \cdot X \;\; ; \;\; \Theta^T \cdot X = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \cdots + \beta_N \cdot X_N \qquad (6.1)$$

where $p = p(y = 1|x)$ is the probability of belonging to the positive class and β_i are the regression coefficients. Inverting Eq. 6.1 the probability is obtained:

$$p = \frac{e^{\Theta^T \cdot X}}{1 + e^{\Theta^T \cdot X}} = \frac{1}{1 + e^{-(\Theta^T \cdot X)}} \qquad (6.2)$$

The probability evaluated with Eq. 6.2 can be visualized as in Fig. 6.1

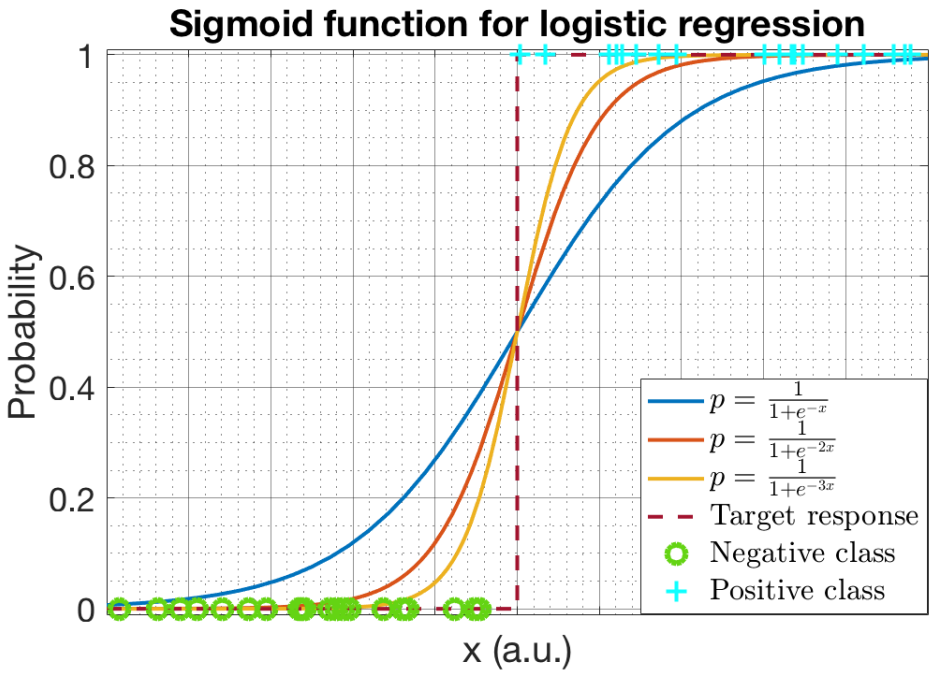


Fig. 6.1 An example of the sigmoid function used in a logistic regression analysis.

As shown in Fig. 6.1, an example of an ideal Target response (red dashed line) is compared to some probability sigmoids with different coefficients β_1. The higher the coefficient β_1, the steeper the sigmoid curve.
The aim of the logistic regression in binary classification problems is to reduce the distance between the regressed curves and an ideal step function between 0 and 1. To find the best combination of coefficients a minimisation problem must be solved

$$\Theta_r = \text{argmin}_\Theta J(\Theta) \tag{6.3}$$

where Θ_r are the regressed coefficients and J is an objective cost function defined as the negative log conditional likelihood:

$$J(\Theta) = -\log \prod_{i=1}^{N_S} p_\Theta(y^{(i)}|x^{(i)}) = -\sum_{i=1}^{N_S} \log p_\Theta(y^{(i)}|x^{(i)}) \tag{6.4}$$

where $p_\Theta(y^{(i)}|x^{(i)})$ is the conditional probability of having a $y^{(i)}$ response given an input $x^{(i)}$ and N_S is the number of training samples. To minimise this function several

methods have been implemented, in particular the least squared difference, gradient descent and the Newton method are commonly adopted by machine learning softwares. Such methods try to solve the partial derivatives of $J(\Theta)$ (Eq. 6.4) respect with the coefficients Θ to find the local minimum of the log-likelihood using deterministic or stochastic approaches. The search for the best regressed coefficients performed by the learning algorithm gives as results a set of best estimates of the coefficients. The learning algorithm can give also an estimate of the coefficient variances and covariances which can be used to evaluate the goodness of the regression model. The output of the logistic regression 6.2 is a continuous probability of belonging to a given positive class. To obtain a binary classification, the probability is compared to a fixed threshold commonly equal to 0.5 (50 %). If the probability is higher than the threshold then the data belongs to the positive class and, if lower, to the negative class. The trained model produces predictions which are compared to the true class label to obtain the following metrics:

- TP.: True Positive - Number of predictions correctly classified as belonging to the positive class
- TN.: True Negative - Number of predictions correctly classified as belonging to the negative class
- FP.: False Positive - Number of predictions incorrectly classified as belonging to the positive class
- FN.: False Negative - Number of predictions incorrectly classified as belonging to the negative class

Such quantities can be used to calculate some model metrics:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \cdot 100 \ \ (\%) \tag{6.5}$$

$$Sensitivity = \frac{TP}{TP+FN} \cdot 100 \ \ (\%) \tag{6.6}$$

$$Specificity = \frac{TN}{TN+FP} \cdot 100 \ \ (\%) \tag{6.7}$$

The *Accuracy* metrics is the most important parameter in classification models and it gives an evaluation of the fraction of correctly classified data. The *Sensitivity* metrics measures the fraction of data belonging to the positive class that is correctly

predicted by the classification model. The $Specificity$ metrics measures the fraction of data belonging to the negative class that is correctly predicted by the classification model.

6.1.1 Weighted logistic regression

To take advantage of the uncertainty evaluation techniques described in the previous sections, the evaluated uncertainties can be used as weights in the learning process. To weight the samples, a modified version of the objective cost function is used:

$$J_w(\Theta) = -\sum_{i}^{N_S} w_i \cdot \log p_\Theta(y^{(i)}|x^{(i)}) \tag{6.8}$$

where w_i are adimensional weights. Such formulation is often used in the training of surveyed data and especially in the case of an unbalanced dataset where different weight of evidence (WOE) [41] can be given to the input data. In this work the inverse of mean value of the relative uncertainties of the considered features was chosen as a weighting coefficient:

$$w_i = \frac{1}{(\sum_j^{N_F} \frac{U(F_j^i)}{F_j^i})/N_F} \tag{6.9}$$

where $\frac{U(F_j^i)}{F_j^i}$ is the relative expanded uncertainty of the j-th feature for the i-th sample and N_F is the number of features used in the training of the model.

6.2 A metrologic approach to the logistic regression

To evaluate the uncertainty of the prediction model, an analytical uncertainty propagation was carried out on the probability Eq. 6.2, as showed in Section 3.1.2:

$$p = \frac{e^{\Theta^T \cdot X}}{1 + e^{\Theta^T \cdot X}} = \frac{1}{1 + e^{-(\Theta^T \cdot X)}} \tag{6.10}$$

The partial derivatives with respect to features and coefficients of the probability equations are calculated to obtain the sensitivity coefficients:

$$\begin{aligned} \frac{\partial p_i}{\partial F_j} &= \beta_j \cdot p_i \cdot (1 - p_i) \\ \frac{\partial p_i}{\partial \beta_0} &= p_i \cdot (1 - p_i) \\ \frac{\partial p_i}{\partial \beta_j} &= F_j \cdot p_i \cdot (1 - p_i) \\ j &\in [1 \dots N_F] \end{aligned} \tag{6.11}$$

where N_F is the number of considered features. In this way, an analytical evaluation of the probability uncertainty can be obtained using the uncertainty propagation formula. One should note that the sensitivity coefficients are equal to zero for perfect predictions (p_i =0 %, p_i =100 %) an their effect increases as the prediction probabilities approach the value 50 %.

6.2.1 Correlation evaluation

The evaluation of correlation between couples of features F_j and couples of model coefficients β_j is a common practice in machine learning model training. To evaluate the effects of correlation between features and coefficients, an absolute conditional Pearson correlation matrix is calculated using the equation:

$$|r_c| = |r| \cdot S + \overline{S} \tag{6.12}$$

$$S = (p_{value} \leq 0.05) \tag{6.13}$$

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \tag{6.14}$$

where $|r|$ is the absolute Pearson correlation matrix given by the ratio between σ_{XY}, which is the covariance between feature X and Y and $\sigma_X \cdot \sigma_Y$, which is the product of the feature standard deviation. S is a logic binary condition that is 1 when the X-Y correlation is significative with a confidence level of 95 % and $\overline{S}$ is its logic complement. According to Eq. 6.12, the correlations estimated with a high significance level will have a $|r_c| < 1$ and those that are not significative or perfectly

correlated will have a $|r_c| = 1$. The conditional correlation matrix of Eq. 6.12 can be represented by a heatmap as shown in Fig 6.2.

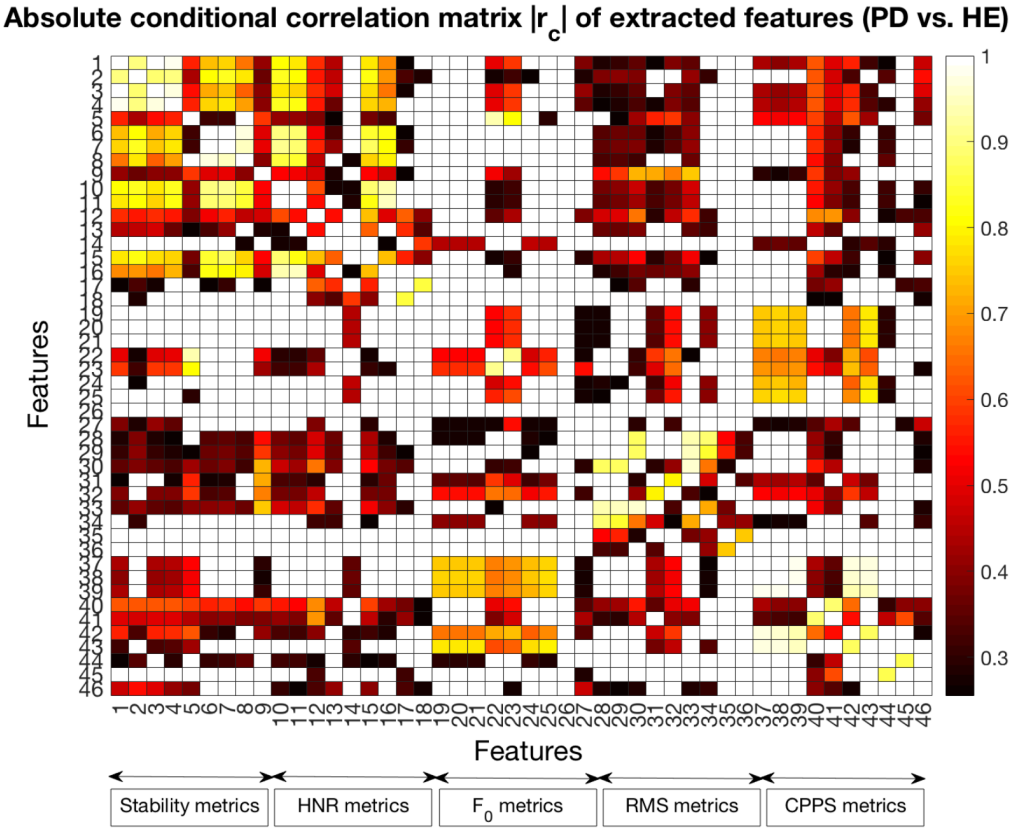


Fig. 6.2 An example of r_c correlation matrix. The cells in solid white represent the correlations that were evaluated with a low significance level (p-value>0.05).

As shown in the heatmap, the correlation coefficients can be very high and a considerable number of features couples show a correlation which has not been possible to evaluate with a high significance level (solid white cells).

6.2.2 First approach: negligible correlation

If the correlation between features is low enough, a simplified uncertainty evaluation can be performed using Eq. 3.8, here reported for clarity:

$$u(p) = \sqrt{\sum_{i=1}^{N} (\frac{\partial p}{\partial F_i} \cdot u(F_i))^2 + (\frac{\partial p}{\partial \beta_i} \cdot u(\beta_i))^2} \tag{6.15}$$

The uncertainties $u(F_i)$ of the features F_i can be extracted from the uncertainty evaluation techniques showed in the previous sections. The uncertainties $u(\beta_i)$ of the model coefficients β_i are given by the training process which produces an array of coefficients uncertainties. As an example, the Matlab function *fitglm* gives an estimation of the coefficients along with their respective standard errors which are evaluated with a certain p-value. The condition that allows to consider the mixed effects as negligible is a necessary but not sufficient condition which can be explained considering the correlation in Eq. 6.14. Considering a special case where $\sigma_X \approx \sigma_Y$

$$\sigma_{XY} = r \cdot \sigma_X^2 = r \cdot \sigma_Y^2 \tag{6.16}$$

The relation above states that the covariance term between X and Y is r times the variance term of X or Y. To have a significative difference between the covariance and the variance terms a r<0.1 ($r^2 = 0.01$) could be a feasible choice because it guarantees a difference of an order of magnitude between the two terms.

6.2.3 General approach: mixed-terms evaluation

As previously stated, the low correlation between features is a necessary condition to consider the mixed effects as negligible. From a general point of view this assumption is not true, because the covariance terms multiply combinations of sensitivity coefficients which can be very large and sometimes can produce even negative terms. To solve this issue, a general matrix formulation is used to calculate the uncertainty:

$$u(p) = \sqrt{J \cdot COV \cdot J^T} \tag{6.17}$$

where COV is the variance-covariance matrix and J is the Jacobian of the input features:

$$J_{ij}(F) = \frac{\partial p_i}{\partial F_j} \;;\; j \in [1 \ldots N_F] \;;\; i \in [1 \ldots N_S] \tag{6.18}$$

$$J_{ij}(\beta) = \frac{\partial p_i}{\partial \beta_j} \;;\; j \in [1 \ldots N_F + 1] \;;\; i \in [1 \ldots N_S] \tag{6.19}$$

where N_F is the number of considered features, N_S is the number of samples of the dataset, $J_{ij}(F)$ is the $N_S \times N_F$ Jacobian matrix of the features, $J_{ij}(\beta)$ is the $N_S \times (N_F + 1)$ Jacobian matrix of the coefficients. Due to the different dimensions

of the Jacobian and covariance matrices, the equation 6.17 can be rewritten as the square root of two terms:

$$u(p) = \sqrt{J_F COV_F J_F^T + J_\beta COV_\beta J_\beta^T} \tag{6.20}$$

6.3 Feature and model selection

For this work a brute force approach was used to select the combination of features to train the logistic regression models. To accept or reject a feature combination some common criteria have to be met:

- Correlation between each of the features lower than a certain threshold, evaluated with a high significance level (p-value<0,05)
- Model coefficients evaluated with high significance (p-value<0,05)

In order to find the best LR model that separates the classes, a model score metrics, using the equations described in equations 6.5 to 6.7, can be defined:

$$\text{Score} = Accuracy - |Sensitivity - Specificity| \tag{6.21}$$

Using the Eq. 6.21, the feature selection algorithm can choose the model which maximises the accuracy, balancing the Sensitivity and Specificity at the same time. In Fig. 6.3 the flow-chart of a Common feature and model selection algorithm is shown:

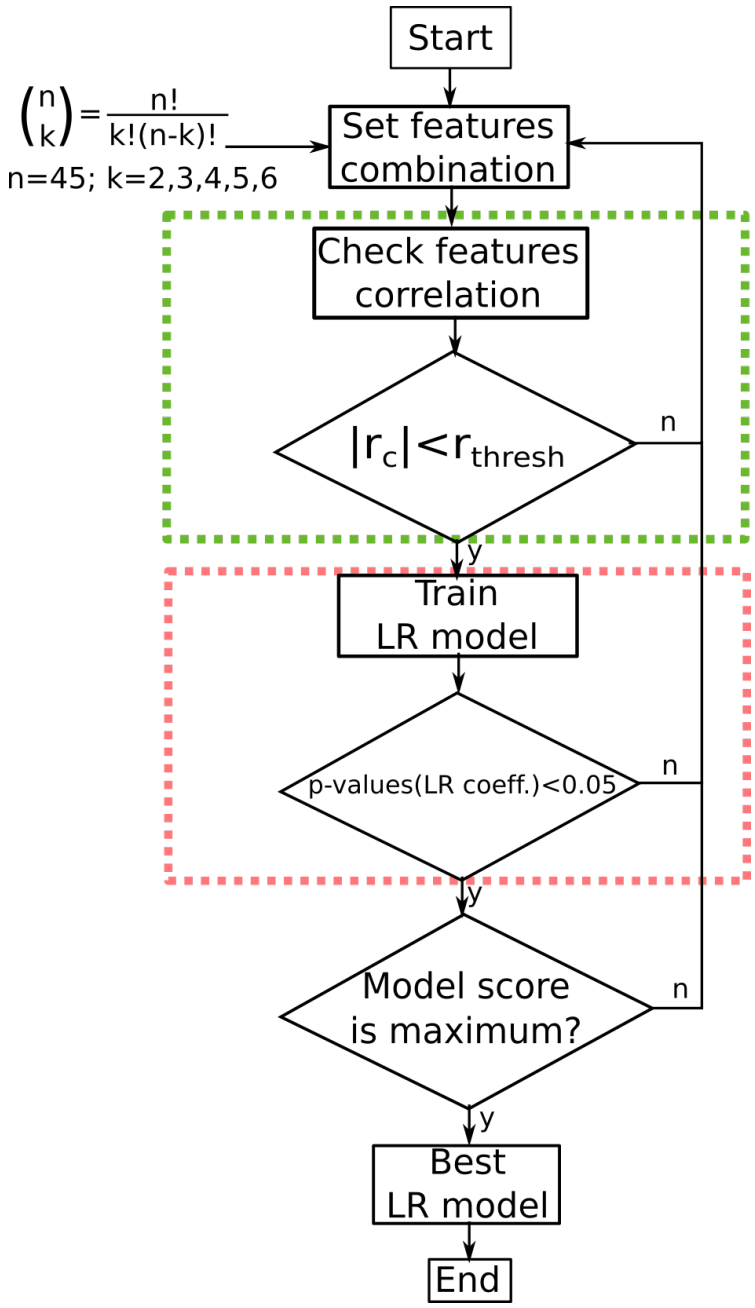


Fig. 6.3 Flow chart of a common feature and model selection algorithm.

As shown in the flow-chart in Fig. 6.3, a non-repeated combination of 45 features grouped in sets of 2,3,4,5 and 6 features is set using the Matlab function *nchoosek*.

The number of non-repeated features combinations is given by the equation:

$$\binom{n}{k} = \frac{n!}{k! \cdot (n-k)!} \tag{6.22}$$

where n is the total number of extracted features and k is the number of considered features. Considering n=45 extracted features (see the Appendix A for a detailed list) and k ranging from 2 to 6 features, the number of possible non-repeated combinations are:

- k=2: $\binom{n}{k} = 990$
- k=3: $\binom{n}{k} = 14190$
- k=4: $\binom{n}{k} = 148995$
- k=5: $\binom{n}{k} = 1221759$
- k=6: $\binom{n}{k} = 8145060$

As showed in the list, the number of combination to test can reach very high values so the feature selection algorithm becomes more computationally expensive as the number of considered features rises. The algorithm depicted in Fig. 6.3 has two principal components (highlighted by dashed boxes):

- Feature selection (green box)
- Model training (red box)

For each feature combination, the conditional correlation matrix $|r_c|$ is evaluated using Eq. 6.12. If the ij feature combination have a $|r_c|$ less than a certain threshold value the combination is valid and a Logistic Regression model is trained. If the LR coefficients are estimated with a high significance level the model Score is evaluated using Eq. 6.21. This evaluation goes on until every combination of features was tested. The trained model with the highest score is labeled as the best model. The algorithm depicted in Fig. 6.3 represent a common feature and model selection and to compare it to the proposed method it will be referred as common selection (**CS**) .

6.3.1 Proposed feature and model selection

The proposed feature and model selection (**PS**) uses the informations achieved in the previous section to produce weighted classification models. To take into account the features uncertainty evaluations, a modified version of the model training component (red dashed box), shown in Fig. 6.3 is presented in Fig. 6.4.

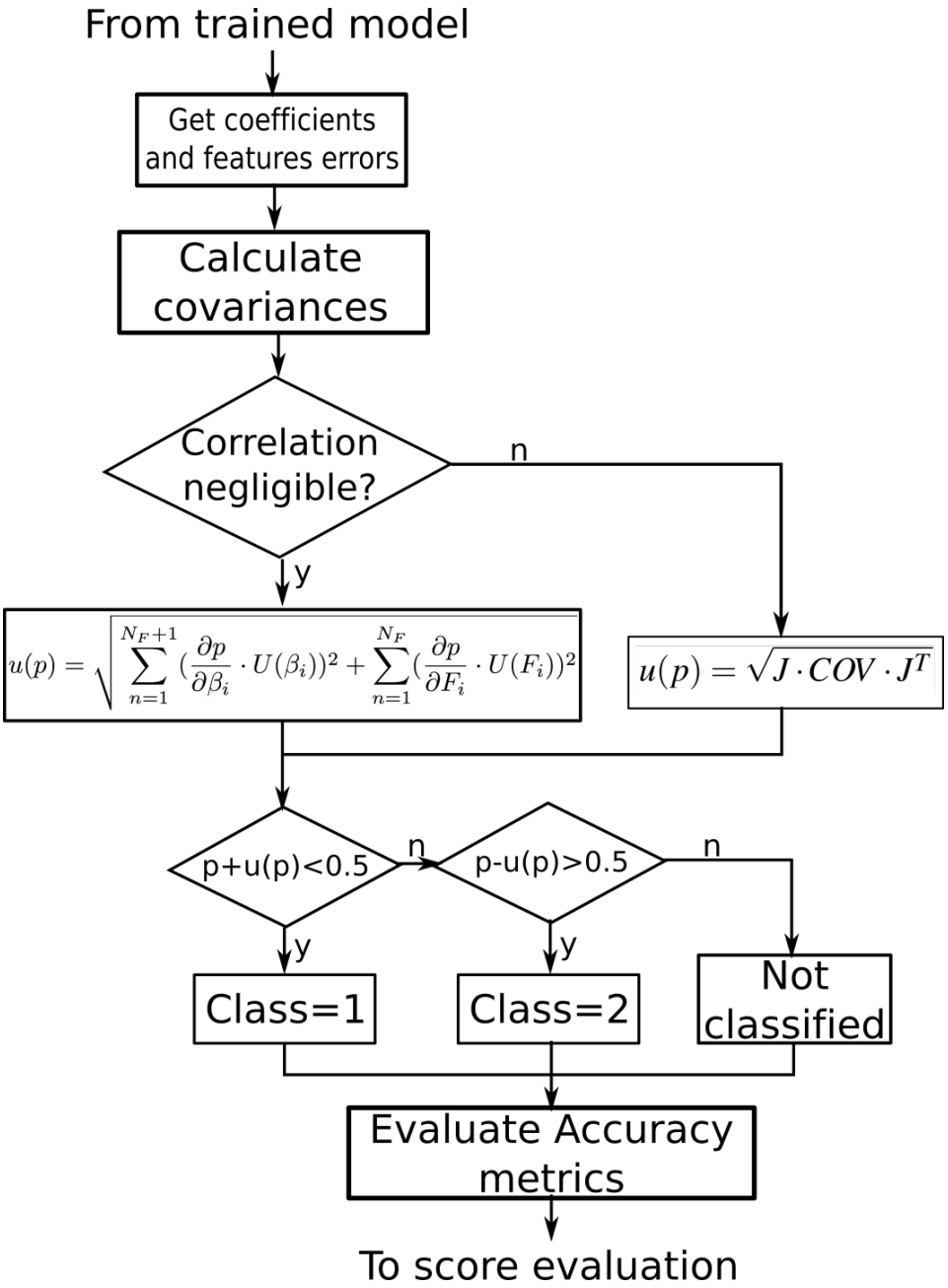


Fig. 6.4 Flow chart of the proposed model training algorithm.

As shown in the figure, an alternative model training algorithm is inserted after the correlation check module and before the score evaluation. Such module trains a LR model using the candidate set of features weighting each sample using Eq. 6.9. Moreover, the uncertainty of each predicted probability is evaluated to produce a confidence interval around the probability values. An example of two features

logistic regression model is presented in Fig. 6.5 to better understand the selection mechanism .

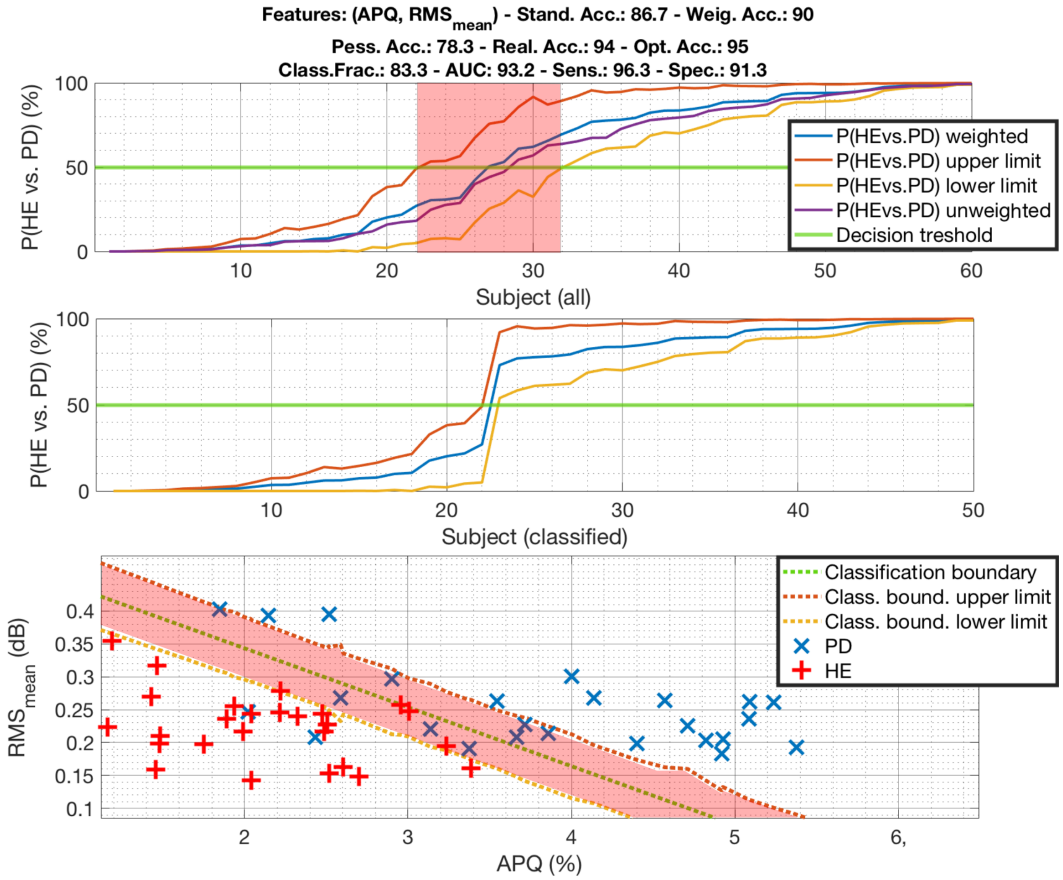


Fig. 6.5 Predicted probabilities of the PD vs. HE subset using two features. The highlighted red areas represent the subset of non-classified subjects. All the accuracy metrics are expressed as %.

As shown in Fig. 6.5, an example of two features (apq and $A_{RMS}(mean)$, see the Appendix A for details) trained model is presented. The plot on the top row represent the CS model (purple line, unweighted) and the PS model (blue line, weighted) predicted probabilities of belonging to the HE class or the PD class. The predicted probability values were sorted in ascending order to obtain the typical sigmoid curve. As shown in Fig. 6.5, the determination of a confidence interval around the predicted probabilities allow to define a third class of "non-classified" subjects (the red areas on the top and bottom plots). Such subjects have a predicted probability range which intersects the decisional threshold of 0.5. Removing the indeterminate predictions, the middle plot of Fig. 6.5 is obtained. The probability confidence intervals can be

projected on the features hyperplane, as shown in the bottom plot of Fig. 6.5. From a metrologic point of view, the subjects who fall in the red area are just indeterminate and nothing can be done about it except taking more measurements hoping that the confidence intervals will shrink after such an operation. Such a threshold comparison experiment is a very common practice in the metrologic evaluation of car speed performed by automatic road speedometers. Commonly the tolerance for the car speed evaluated by a speedometer is set around ± 5 km/h to take into account the accuracy of the driver in evaluating the car speed through the speedometer mounted in the car dashboard. So if the speed limit of a particular road is 80 km/h, any speed between 75 km/h and 85 km/h should be considered as indeterminate. How to deal with such an indetermination? In Italy the indetermination is solved using an Italian legal paradigm known as "*presunzione di innocenza*" which can be roughly translated to "*innocent until proven guilty*". So, in an indetermination situation, the government cannot prove beyond any reasonable doubt that the car was actually going over 80 km/h and the driver is automatically considered as innocent. Using a similar paradigm new classification metrics were defined:

- Accuracy of the proposed selection method (PS): the non-classified are classified (the probability confidence interval is not considered):

$$Accuracy_{PS} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \tag{6.23}$$

- Pessimistic Accuracy: the non-classified are False:

$$Accuracy_{p} = \frac{(TP_c+TN_c)}{TP+TN+FP+FN} \tag{6.24}$$

- Realistic Accuracy: the non-classified are not considered:

$$Accuracy_{r} = \frac{(TP_c+TN_c)}{(TP_c+TN_c+FP_c+FN_c)} \tag{6.25}$$

- Optimistic Accuracy: the non-classified are True:

$$Accuracy_{o} = \frac{(TP_c+TN_c)+(N-N_c)}{(TP+TN+FP+FN)} \tag{6.26}$$

- Fraction of classified: the fraction of classified subjects with respect to the total number of samples:

$$F_c = \frac{N_c}{N} \tag{6.27}$$

where the subscript c refers to the samples whose predictions have confidence intervals that do not intersect the decisional threshold, N is the total number of samples and N_c is the number of classified samples. According to the new metrics, an evaluation of worst case scenario (pessimistic accuracy) and best case scenario (optimistic accuracy) can be performed giving more information about the classification performances of the prediction models. A new score definition is proposed modifying Eq. 6.21 to maximise the performance of the selected model :

$$\text{Score}_p = Accuracy_p - |Sensitivity - Specificity| \tag{6.28}$$

where $Accuracy_p$ is the pessimistic accuracy. In this way the optimization process tries to maximise the pessimistic accuracy, which strongly depends on the Fraction of classified, while balancing sensitivity and specificity.

6.4 Training experiments

The production of artificial vowels tries to replace the ability of the subject to produce repeated tasks in a reproducible way. According to this procedure, it is possible to evaluate the effect of the measuring chain uncertainty contributions on the training of the classification algorithm. Some training experiments were carried out on the subjects dataset to evaluate the performance of the proposed method. The models were evaluated using the two approaches described in sections 6.2.2 and 6.2.3:

1. first approach: negligible correlation - $|r_c| < 0.1$ ($r^2 < 0.01$)
2. general approach: mixed-terms evaluation $|r_c| < 0.7$ ($r^2 \lesssim 0.5$)

A common training experiment (CS) as described in Sec. 6.3 was performed (setting $|r_c| < 0.7$) to compare the classification performances with the ones of the proposed method (PS). Each dispersion parameter was multiplied by a coverage factor chosen using the inverse T-student with a 95 % confidence limit (≈ 2). For each original vowel 10 artificial vowels were generated using the re-synthesis method described in

Sec. 4.1. A total of 900 artificial vowels were generated for the three clinical classes (PD, HE, PA). The Original and artificial data were combined in different ways to evaluate the differences between the models:

6.4.1 Using original data

Four combinations are possible when using the original dataset to train the models

1. **Training data**=Original data ($F_M(OR)$)
   Data uncertainty=squared sum of bias plus two times the standard error of the artificial data (Eq. 6.33)
   Prediction uncertainty: no mixed-terms evaluation (Eq. 6.15)

2. **Training data**=Original data ($F_M(OR)$)
   Data uncertainty=squared sum of bias plus two times the standard error of the artificial data (Eq. 6.33)
   Prediction uncertainty: with mixed-terms evaluation (Eq. 6.20)

3. **Training data**=Original data (bias removed, $F_M(OR)^*$, Eq. 6.30)
   Data uncertainty=two times the standard error of the artificial data (Eq. 6.32)
   Prediction uncertainty: no mixed-terms evaluation (Eq. 6.15)

4. **Training data**=Original data (bias removed, $F_M(OR)^*$, Eq. 6.30)
   Data uncertainty=two times the standard error of the artificial data (Eq. 6.32)
   Prediction uncertainty: with mixed-terms evaluation (Eq. 6.20)

The bias removal process is carried out considering the bias evaluations performed in the previous section, i.e. using Eq. 5.5 as in Section 5.1.3, where an average evaluation for each class was presented. For this training experiment, a different bias value for each i-th vowel was considered modifying Eq. 5.5 in order to obtain an evaluation of the bias of each i vowel using Eq. 6.29:

$$\overline{BIAS_{ART-MC}(i)} = \frac{\sum_{j=1}^{10} F_M^j[ART_{EXT}(i)] - F_R^j[MC(i)]}{10} \tag{6.29}$$

This bias evaluation can be used to correct the input data respect to a trusted source. The *BIAS* can be removed from the original data matrix $F_M(OR)$ using Eq. 6.30

$$F_M^i(OR)^* = F_M^i(OR) - \overline{BIAS_{ART-MC}(i)} \tag{6.30}$$

where $F_M^i(OR)^*$ is the original features matrix after the bias correction. The dispersion of the original data is evaluated as the standard error of the corresponding measured artificial data:

$$\overline{DISP_{ART}(i)} = \frac{\sqrt{\frac{\sum_{j=1}^{10}(F_M^j[ART_{EXT}(i)] - \overline{F_M^j[ART_{EXT}(i)]})^2}{10}}}{\sqrt{10}} \tag{6.31}$$

The uncertainty of the $i-th$ feature is evaluated as two times the standard error of the the artificial data when the bias is removed (using Eq. 6.30) from the original data $F_M^i(OR)$:

$$\overline{U_{OR}(i)} = 2 \cdot \overline{DISP_{ART}(i)} \tag{6.32}$$

If the bias is not removed, a different uncertainty parameter is calculated including the bias contribution:

$$\overline{U_{OR}(i)} = \sqrt{(\overline{BIAS_{ART-MC}(i)})^2 + \cdot(2 \cdot \overline{DISP_{ART}(i)})^2} \tag{6.33}$$

In this way, the original data, bias corrected ($F_M^i(OR)^*$) or not ($F_M^i(OR)$), is considered as an "expected mean value" surrounded by a cloud of 10 artificial values, which have a dispersion equal to the one evaluated in the previous sections. The standard error, which is the standard deviation divided by the square root of the number of samples, is considered as the uncertainty of an averaging process which actually never took place in this experiment. This happens because of the impossibility of asking to a subject to produce identical tasks in order to separate the human contribution from the machine contribution.

6.4.2 Using artificial data

The proposed method can work also as a data boosting procedure to increase the size of the training dataset. To take advantage of such a method, four different combinations of input data have to be defined:

1. **Training data**=Artificial data ($F_M(ART)$)
   Data uncertainty=squared sum of bias plus two times the standard deviation of the artificial data (Eq. 6.38)
   Prediction uncertainty: no mixed-terms evaluation (Eq. 6.15)

2. **Training data**=Artificial data ($F_M(ART)$)
   Data uncertainty=squared sum of bias plus two times the standard deviation of the artificial data (Eq. 6.38)
   Prediction uncertainty: with mixed-terms evaluation (Eq. 6.20)

3. **Training data**=Artificial data (bias removed, $F_M(ART)^*$, Eq. 6.35)
   Data uncertainty=standard deviation of the artificial data (Eq. 6.36)
   Prediction uncertainty: no mixed-terms evaluation (Eq. 6.15)

4. **Training data**=Artificial data (bias removed, $F_M(ART)^*$, Eq. 6.35)
   Data uncertainty=standard deviation of the artificial data (Eq. 6.36)
   Prediction uncertainty: with mixed-terms evaluation (Eq. 6.20)

According to these choices, each of the data of the artificial clouds have a certain probability of belonging to its distribution, which is equal to the standard deviation multiplied by a coverage factor. The bias removal process is different in this case because of the generation uncertainties seen in Sec. 5.1.1. To consider the artificial data as representative of the original vowel from which they were generated, their values should be scattered around the original data. To achieve this, the generation bias, evaluated in Sec. 5.1.1 and the extraction bias, evaluated in Sec. 5.1.3, have to be removed from the artificial data. The equations 6.29, 6.30 and 6.31 are then modified as follows:

$$\overline{BIAS_{ART-MC-OR}(i)} = \frac{\sum_{j=1}^{10} F_M^j[ART_{EXT}(i)] - F_R^j[MC(i)])}{10} + \frac{\sum_{j=1}^{10} F_M^j[MC(i)] - F_M^i(OR))}{10} \tag{6.34}$$

$$F_M^i(ART)^* = F_M^i(ART) - \overline{BIAS_{ART-MC-OR}(i)} \tag{6.35}$$

$$\overline{DISP_{ART}(i)} = \sqrt{\frac{\sum_{j=1}^{10}(F_M^k[ART_{EXT}(i)] - \overline{F_M^j[ART_{EXT}(i)]})^2}{10}} \tag{6.36}$$

The uncertainty of the i-th feature is evaluated as two times the standard deviation of the the artificial data when the bias is removed (using Eq. 6.34) from the artificial data $F_M^i(ART)$:

$$\overline{U_{ART}(i)} = 2 \cdot \overline{DISP_{ART}(i)} \tag{6.37}$$

If the bias is not corrected, a different equation is used to include the bias contribution :

$$\overline{U_{ART}(i)} = \sqrt{(\overline{BIAS_{ART-MC-OR}(i)})^2 + \cdot(2 \cdot \overline{DISP_{ART}(i)})^2} \tag{6.38}$$

6.5 Training experiments results on extracted features (EXT contribution)

The model metrics in this section are relative to the models trained with the features and their uncertainties extracted from the "short" measuring chain (using the architecture described in Sec. 5.1. The uncertainty evaluations of the extraction contribution were used to train the weighted logistic regressions and to evaluate the probability confidence intervals. In particular, the evaluated metrics in this section are:

- **CS features**: selected features of the common feature and model selection algorithm defined in Sec. 6.3 (green column)
- **PS features**: selected features of the proposed feature and model selection algorithm defined in Sec. 6.3.1 (light blue column)
- **CS Accuracy**: prediction accuracy (Eq. 6.5) of the common feature and model selection algorithm defined in Sec. 6.3 (green column)
- **PS Accuracy**: prediction accuracy (Eq. 6.23) of the proposed feature and model selection algorithm defined in Sec. 6.3.1, (light blue column)
- **AUC**: area under curve, a metric which measures the ability of the model to distinguish between classes.
- **Sens.**: sensitivity, as defined in Eq. 6.6
- **Spec.**: specificity, as defined in Eq. 6.7
- **Fraction of classified**: as defined in Eq. 6.27
- **Pess. Acc.**: pessimistic accuracy, as defined in Eq. 6.24
- **Real. Acc.**: realistic accuracy, as defined in Eq. 6.25
- **Opt. Acc.**: optimistic accuracy, as defined in Eq. 6.26

- **Score$_p$**: proposed model score, as defined in Eq. 6.28

The columns after the PS Acc report the model metrics of the PS algorithm. For graphical reasons, the selected features are identified with a number corresponding to the features reported in the Appendix A. The rows highlighted in red represent a failure of the feature and model selection algorithm in finding a set of features and/or a model that meet the conditions described in sections 6.3 and 6.3.1.

6.5.1 Training experiments using original data

In this section, the results of the training experiments using the original data is presented. The data were pre-processed removing or not the bias using Eq. 6.30. The Data uncertainty were evaluated using equations 6.32 or 6.33. The probability uncertainty was evaluated using equations 6.15 or 6.20. The models were trained to find the best combination of features which maximises the score given by equations 6.21 and 6.28. The classification metrics in this section are not validated but an example of the model validations will be presented in Sec. 6.7. A summary of the results is shown in Tabs. 6.1 and 6.2.

PD vs. HE training results

In Tab. 6.1 the results of the training experiments for the PD vs. HE subset are presented. The numbers of the selected features refer to the equations and definitions summarised in the Appendix A.

Table 6.1 CS and PS accuracy metrics for the PD vs. HE subset.

N° Features	Mixed terms eval.	Bias removal	CS Features	PS Features	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	$Score_p$%
2	No	No	11 28	8 28	87	88	93	96	92	87	82	94	95	77
2	Yes	No	11 28	8 29	87	90	94	96	89	93	87	93	93	80
2	No	Yes	7 28	8 28	93	90	93	96	92	88	83	94	95	79
2	Yes	Yes	7 28	1 12	93	88	89	89	89	93	83	89	90	83
3	No	No	10 30 31	6 30 31	90	88	92	96	88	85	78	92	93	70
3	Yes	No	10 30 31	9 11 28	90	90	94	93	93	93	87	93	93	86
3	No	Yes	10 30 32	8 29 40	88	88	96	93	91	83	77	92	93	75
3	Yes	Yes	10 30 32	6 12 28	88	93	93	93	93	95	88	93	93	88
4	No	No	5 11 30 31	5 7 30 31	93	88	96	96	91	83	78	94	95	73
4	Yes	No	5 11 30 31	7 9 17 29	93	90	94	97	96	92	88	96	97	88
4	No	Yes	5 10 30 31	8 9 28 40	93	90	99	96	92	87	82	94	95	78
4	Yes	Yes	5 10 30 31	7 12 28 40	93	92	93	93	93	97	90	93	93	90
5	No	No	5 9 13 15 29	4 5 9 29 40	90	87	99	95	95	65	62	95	97	61
5	Yes	No	5 9 13 15 29	5 10 12 30 31	90	92	96	93	90	95	87	91	92	83
5	No	Yes	4 5 9 29 40	1 5 9 29 40	88	87	97	92	91	78	72	91	93	71
5	Yes	Yes	4 5 9 29 40	6 9 12 28 40	88	90	95	93	93	97	90	93	93	90
6	No	No	Not Found	Not Found										
6	Yes	No	Not Found	5 9 10 12 29 40		88	97	93	92	88	82	92	93	81
6	No	Yes	Not Found	Not Found										
6	Yes	Yes	Not Found	1 8 9 12 28 40		93	97	96	96	90	87	96	97	87

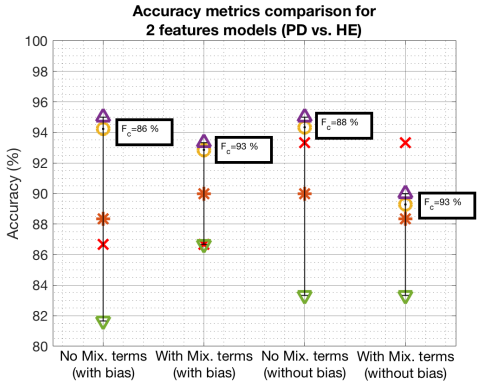
PD vs. PA training results

In Tab. 6.2 the results of the training experiments for the PD vs. PA subset are presented. The numbers of the selected features refer to the equations and definitions summarised in the Appendix A.

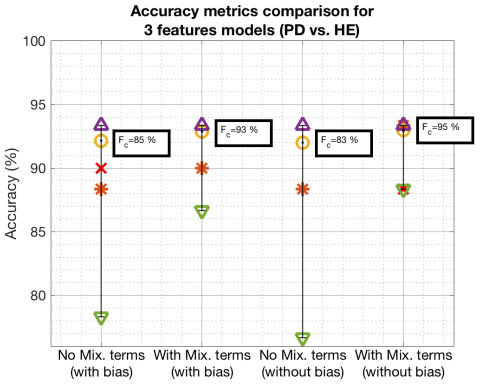
Table 6.2 CS and PS accuracy metrics for the PD vs. PA subset.

N° Features	Mixed terms eval.	Bias removal	CS Features	PS Features	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	$Score_p$%
2	No	No	32 37	9 28	77	67	69	75	59	77	52	67	75	36
2	Yes	No	31 37	5 43	77	70	72	76	71	88	65	74	77	60
2	No	Yes	31 37	32 37	77	75	90	89	81	58	50	86	92	42
2	Yes	Yes	31 37	31 38	77	77	82	83	70	93	72	77	78	59
3	No	No	31 34 43	32 34 43	85	75	95	100	91	47	45	96	98	36
3	Yes	No	31 34 43	1 5 38	85	70	73	76	75	88	67	75	78	66
3	No	Yes	31 34 43	32 34 43	80	78	94	94	92	48	45	93	97	44
3	Yes	Yes	31 34 43	14 23 38	80	78	82	78	78	90	70	78	80	70
4	No	No	Not Found	Not Found										
4	Yes	No	Not Found	5 7 12 38		72	75	74	75	85	63	75	78	62
4	No	Yes	Not Found	Not Found										
4	Yes	Yes	Not Found	22 31 34 43		78	84	82	77	90	72	80	82	66
5	No	No	Not Found	Not Found										
5	Yes	No	Not Found	5 9 12 32 38		77	88	88	79	83	70	84	87	61
5	No	Yes	Not Found	Not Found										
5	Yes	Yes	Not Found	5 8 16 41 42		67	84	78	73	82	62	76	80	56
6	No	No	Not Found	Not Found										
6	Yes	No	Not Found	Not Found										
6	No	Yes	Not Found	Not Found										
6	Yes	Yes	Not Found	2 5 12 16 41 42		72	78	71	74	72	52	72	80	49

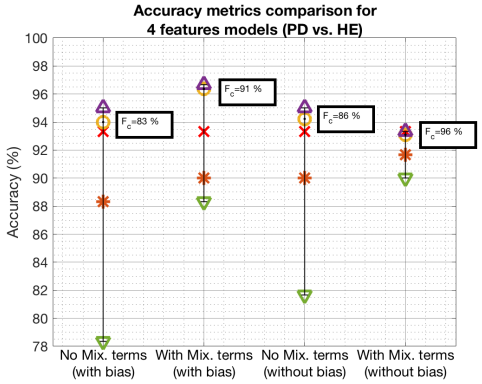
The results in Tabs. 6.1 and 6.2 are depicted in Figs. 6.6 and 6.7 where the CS, PS, realistic, pessimistic and optimistic Accuracy are presented.



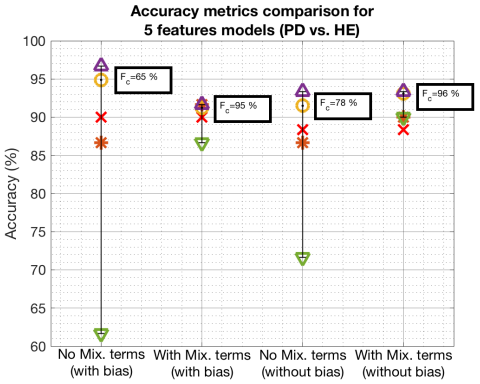
(a) 2 Features



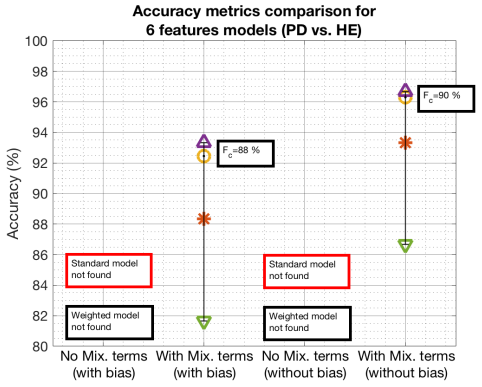
(b) 3 Features



(c) 4 Features



(d) 5 Features



(e) 6 Features

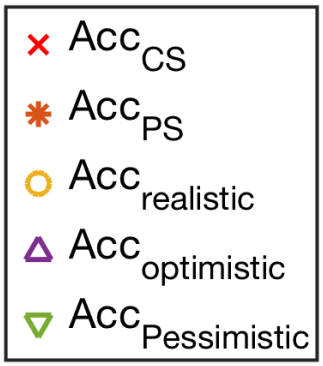
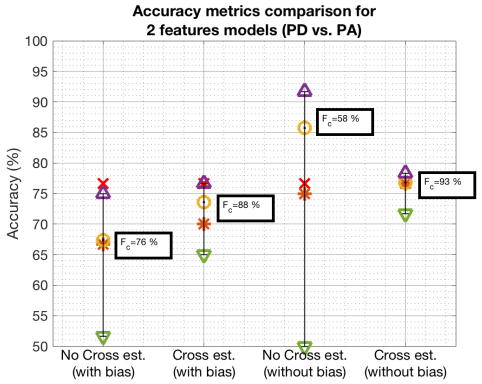
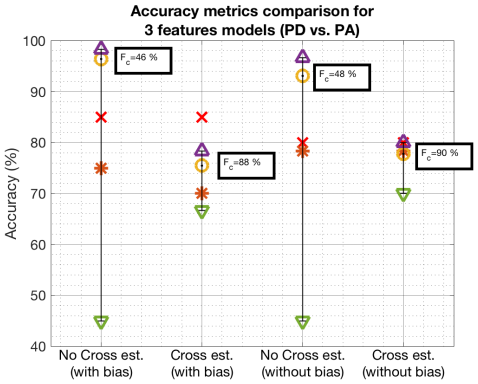


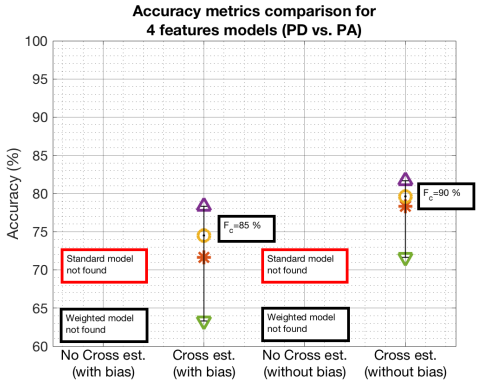
Fig. 6.6 Classification results for the PD vs. HE subset using original data.



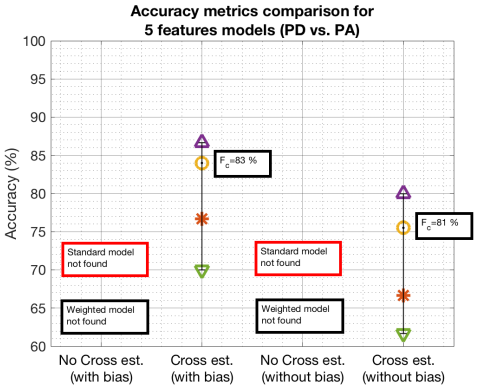
(a) 2 Features



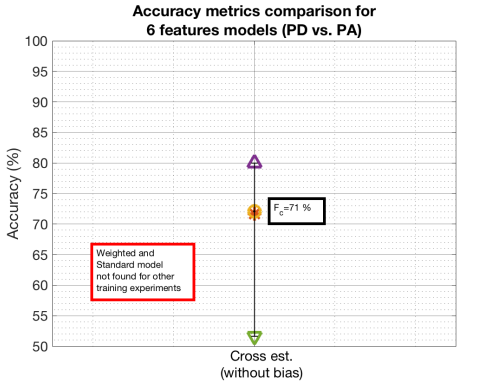
(b) 3 Features



(c) 4 Features



(d) 5 Features



(e) 6 Features

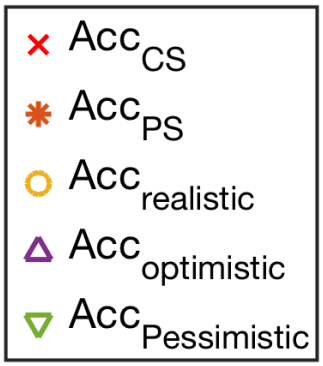


Fig. 6.7 Classification results for the PD vs. PA subset using original data.

6.5.2 Training experiments using artificial data (boosting technique)

The results of the training experiments using the artificial data are presented in this section. The data were pre-processed removing or not the bias using Eq. 6.35. The Data uncertainty were evaluated using equations 6.36 or 6.37. The probability uncertainty was evaluated using equations 6.15 or 6.20. The models were trained to find the best combination of features which maximises the score given by equations 6.21 and 6.28. The classification metrics in this section are not validated, but an example of the model validations will be presented in Sec. 6.7. A summary of the results is shown in Tabs. 6.3 and 6.4.

PD vs. HE training results

In Tab. 6.3 the results of the training experiments for the PD vs. HE subset are presented. The numbers of the selected features refer to the equations and definitions summarised in the Appendix A.

Table 6.3 CS and PS accuracy metrics for the PD vs. HE subset (boosted).

N° Features	Mixed terms eval.	Bias removal	CS Features	PS Features	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	$Score_p$%
2	No	No	10 29	2 31	89	81	85	85	80	91	75	83	85	70
2	Yes	No	10 29	8 34	89	90	94	95	87	96	88	91	91	80
2	No	Yes	10 34	6 34	92	86	92	97	88	90	84	93	93	75
2	Yes	Yes	10 34	10 29	92	90	91	89	93	93	85	91	92	81
3	No	No	11 28 43	11 27 34	91	91	95	96	94	78	74	95	96	73
3	Yes	No	11 28 43	11 34 43	91	92	97	93	92	95	89	93	93	88
3	No	Yes	11 22 34	8 28 36	93	88	96	96	92	89	84	94	95	80
3	Yes	Yes	11 22 34	11 34 37	93	93	95	93	93	97	90	93	93	90
4	No	No	11 23 34 37	7 22 27 34	93	92	94	96	96	79	76	96	97	76
4	Yes	No	11 23 34 37	7 12 34 43	93	93	94	93	93	97	91	93	93	90
4	No	Yes	14 16 30	8 9 12 29	97	91	96	96	96	86	82	96	97	82
4	Yes	Yes	14 16 30	11 34 37 44	97	93	96	93	93	98	92	93	93	91
5	No	No	9 10 32 34 42	7 27 31 34 42	95	92	97	96	95	80	77	96	97	76
5	Yes	No	9 10 32 34 42	7 12 28 31 43	95	93	96	93	93	98	92	93	93	91
5	No	Yes	4 9 34 39 40	11 27 30 31 42	99	95	95	97	93	97	92	95	95	88
5	Yes	Yes	4 9 34 39 40	4 9 28 39 40	99	99	100	100	100	92	92	100	100	91
6	No	No	8 9 23 34 37 40	8 9 23 34 38 40	98	98	100	100	100	86	85	100	100	85
6	Yes	No	8 9 23 34 37 40	1 5 8 30 31 37	98	95	99	97	96	96	93	96	97	92
6	No	Yes	9 11 14 32 34 38	4 9 12 34 39 40	100	99	100	100	100	92	91	100	100	91
6	Yes	Yes	9 11 14 32 34 38	4 9 34 39 40 44	100	100	100	100	100	95	95	100	100	95

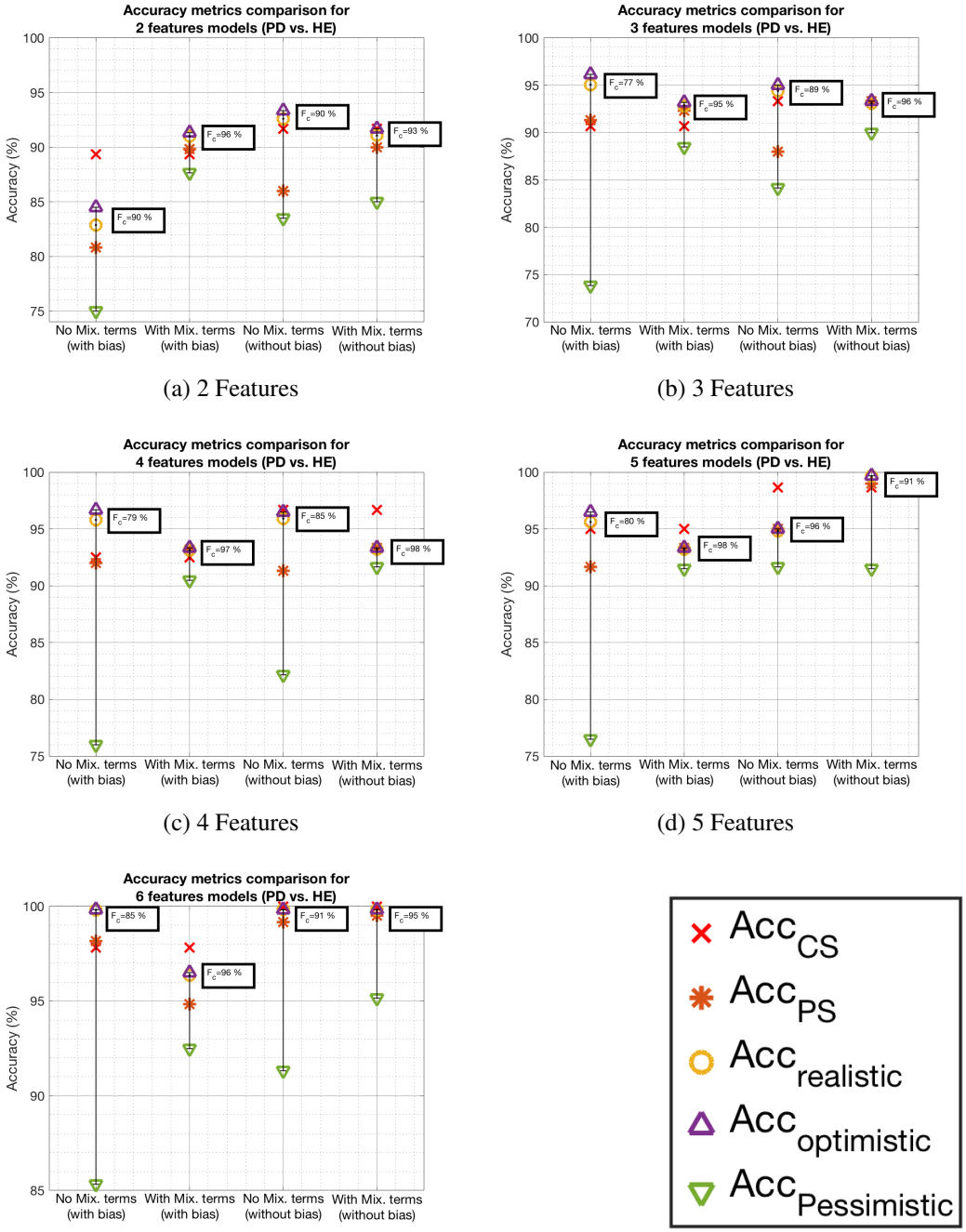
PD vs. PA training results

In Tab. 6.4 the results of the training experiments for the PD vs. PA subset are presented. The numbers of the selected features refer to the equations and definitions summarised in the Appendix A.

Table 6.4 CS and PS accuracy metrics for the PD vs. PA subset (boosted).

N° Features	Mixed terms eval.	Bias removal	CS Features	PS Features	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	$Score_p$%
2	No	No	32 43	9 28	76	66	69	70	64	84	57	67	73	50
2	Yes	No	32 43	8 32	76	72	73	70	75	89	65	72	75	60
2	No	Yes	32 37	8 31	77	65	68	66	68	81	55	67	73	53
2	Yes	Yes	32 37	31 43	77	77	81	86	72	95	75	79	80	62
3	No	No	6 34 43	32 33 38	84	80	98	96	96	60	57	96	98	57
3	Yes	No	6 34 43	8 34 38	84	84	92	89	84	91	79	86	88	75
3	No	Yes	15 34 38	15 34 37	83	80	87	86	83	75	63	84	88	60
3	Yes	Yes	15 34 38	15 34 37	83	80	86	85	82	92	77	84	85	74
4	No	No	8 32 33 42	8 30 32 38	90	89	100	100	100	72	72	100	100	71
4	Yes	No	8 32 33 42	8 32 33 37	90	89	98	93	91	91	84	92	93	82
4	No	Yes	10 29 32 43	10 32 34 38	87	85	97	96	95	73	70	95	97	70
4	Yes	Yes	10 29 32 43	15 30 32 42	0	85	95	93	92	88	82	92	93	81
5	No	No	8 23 32 33 38	5 8 32 33 38	90	90	100	99	100	72	71	99	100	71
5	Yes	No	8 23 32 33 38	5 8 32 33 37	90	90	98	93	92	92	85	93	93	84
5	No	Yes	10 14 29 32 43	7 23 32 34 38	90	89	97	95	95	73	69	95	96	68
5	Yes	Yes	10 14 29 32 43	10 30 32 38 41	90	85	98	93	92	90	83	93	93	83
6	No	No	5 8 32 33 34 38	5 8 30 32 34 38	92	91	100	100	97	74	72	98	99	69
6	Yes	No	5 8 32 33 34 38	8 12 30 32 34 42	92	91	96	95	94	89	84	95	95	84
6	No	Yes	16 30 32 37 41 45	7 23 30 32 34 38	93	86	99	97	95	73	71	96	97	69
6	Yes	Yes	16 30 32 37 41 45	10 12 30 32 34 37	93	88	94	93	93	92	85	93	93	85

The results in Tabs. 6.3 and 6.4 are depicted in Figs. 6.8 and 6.9 where the CS, PS, realistic, pessimistic and optimistic Accuracy are presented.



(a) 2 Features

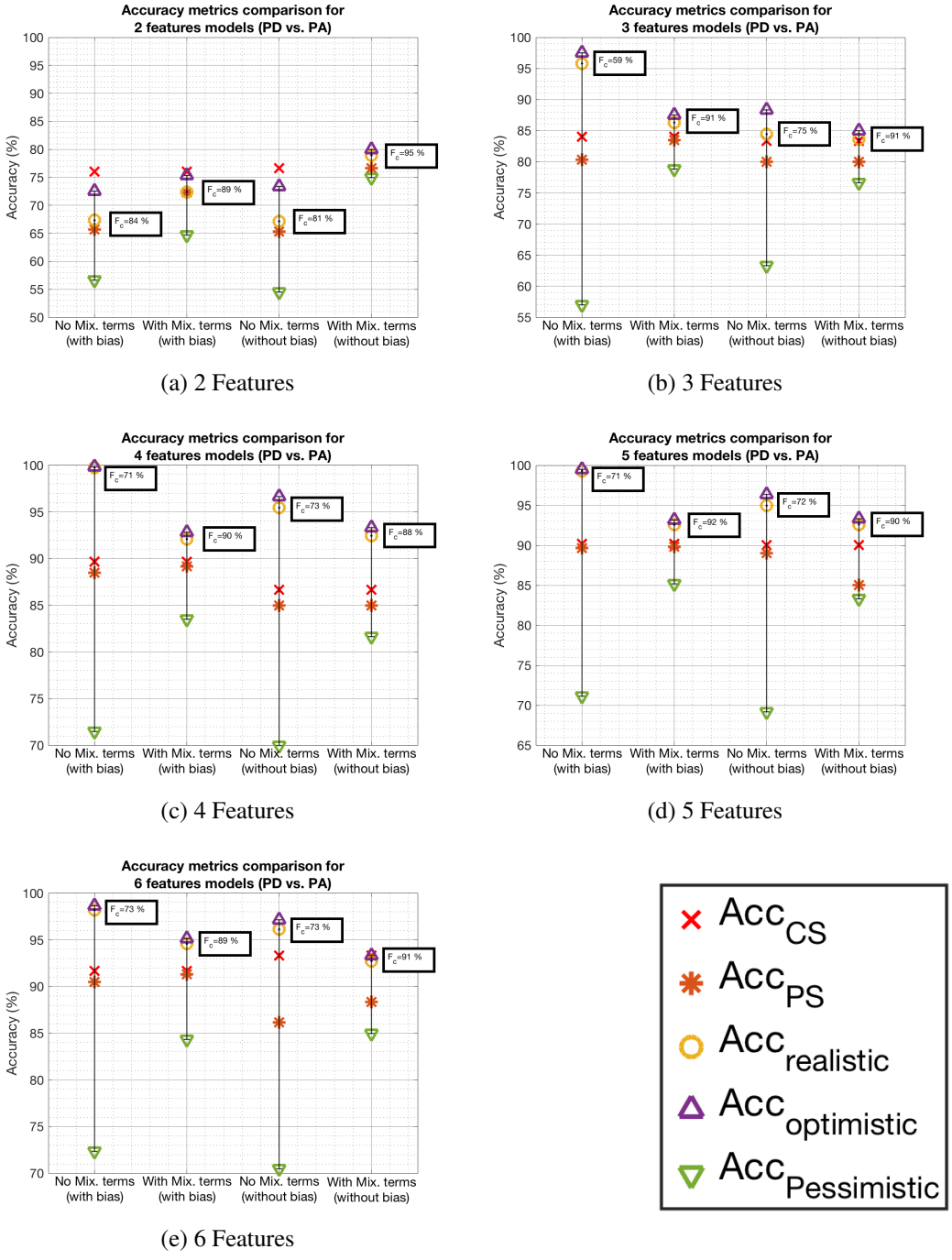
(b) 3 Features

(c) 4 Features

(d) 5 Features

(e) 6 Features

Fig. 6.8 Classification results for the PD vs. HE subset using artificial data.



(a) 2 Features

(b) 3 Features

(c) 4 Features

(d) 5 Features

(e) 6 Features

Fig. 6.9 Classification results for the PD vs. PA subset using artificial data.

6.6 Training experiments results on whole chain data (ACO+ACQ+EXT contribution)

PD vs. HE training results

In this section the results of the training experiments using the original data extracted from the audio recordings using the HATS (see Sec. 5.5) is presented. The data were pre-processed removing or not the bias using Eq. 6.35. The Data uncertainty were evaluated using equations 6.36 or 6.37. The probability uncertainty was evaluated using equations 6.15 or 6.20. The models were trained to find the best combination of features which maximises the score given by equations 6.21 and 6.28. The classification metrics in this section are not validated, but an example of the model validations will be presented in Sec. 6.7. A summary of the results is shown in Tabs. 6.5 and 6.6. In Tab. 6.5 the results of the training experiments for the PD vs. HE subset are presented. The numbers of the selected features refer to the equations and definitions summarised in the Appendix A.

Table 6.5 CS and PS accuracy metrics for the PD vs. HE subset (whole chain).

N° Features	Mixed terms eval.	Bias removal	CS Features	PS Features	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	Scorep %
2	No	No	11 28	10 28	87	92	92	92	92	83	77	92	93	76
2	Yes	No	11 28	10 29	87	90	91	90	93	93	85	91	92	82
2	No	Yes	8 33	10 28	87	92	92	92	92	83	77	92	93	76
2	Yes	Yes	8 33	10 29	87	90	91	90	93	93	85	91	92	82
3	No	No	10 30 31	10 22 31	90	80	88	89	88	85	75	88	90	74
3	Yes	No	10 30 31	5 7 29	90	92	93	93	93	95	88	93	93	88
3	No	Yes	10 30 31	10 22 31	90	80	88	89	88	85	75	88	90	74
3	Yes	Yes	10 30 31	10 30 31	90	90	93	90	89	95	85	89	90	84
4	No	No	5 11 30 31	10 22 30 31	93	92	96	96	92	85	80	94	95	76
4	Yes	No	5 11 30 31	4 5 31 42	93	90	95	93	93	95	88	93	93	88
4	No	Yes	5 10 30 31	10 22 30 31	90	92	96	96	92	85	80	94	95	76
4	Yes	Yes	5 10 30 31	5 11 30 31	90	88	96	93	93	93	87	93	93	86
5	No	No	5 9 13 15 29	4 5 12 31 42	90	90	100	100	100	63	63	100	100	63
5	Yes	No	5 9 13 15 29	6 12 30 31 40	90	92	94	93	93	95	88	93	93	88
5	No	Yes	4 5 9 31 39	4 5 9 31 37	90	92	99	96	96	77	73	96	97	73
5	Yes	Yes	4 5 9 31 39	5 10 12 30 31	90	88	95	93	93	95	88	93	93	88
6	No	No	Not found	Not found										
6	Yes	No	Not found	1 5 12 31 40 42		90	98	93	93	92	85	93	93	85
6	No	Yes	Not found	Not found										
6	Yes	Yes	Not found	5 6 9 12 29 41		90	93	93	93	93	87	93	93	86

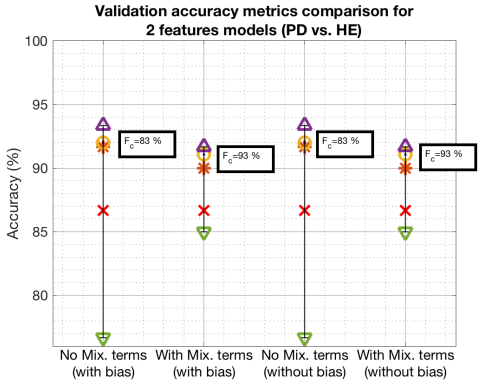
PD vs. PA training results

In Tab. 6.6 the results of the training experiments for the PD vs. PA subset are presented. The numbers of the selected features refer to the equations and definitions summarised in the Appendix A.

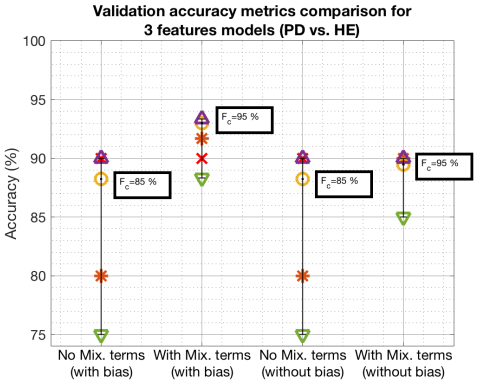
Table 6.6 CS and PS accuracy metrics for the PD vs. PA subset (whole chain).

N° Features	Mixed terms eval.	Bias removal	CS Features	PS Features	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	Scorep %
2	No	No	32 37	31 38	77	73	87	82	76	57	45	79	88	39
2	Yes	No	32 37	31 38	77	73	82	78	75	92	70	76	78	67
2	No	Yes	32 37	31 38	77	73	87	82	76	57	45	79	88	39
2	Yes	Yes	32 37	31 38	77	73	82	78	75	92	70	76	78	67
3	No	No	31 34 43	32 34 43	85	80	96	94	93	55	52	94	97	51
3	Yes	No	31 34 43	5 31 43	85	73	81	73	73	93	68	73	75	68
3	No	Yes	31 34 43	32 34 43	85	80	96	94	93	55	52	94	97	51
3	Yes	Yes	31 34 43	1 5 38	85	73	75	74	72	93	68	73	75	67
4	No	No	Not found	Not found										
4	Yes	No	Not found	12 23 38 40		73	74	73	73	87	63	73	77	63
4	No	Yes	Not found	Not found										
4	Yes	Yes	Not found	5 12 32 37		77	86	82	75	93	73	79	80	66
5	No	No	Not found	Not found										
5	Yes	No	Not found	5 9 12 32 38		78	85	84	78	87	70	81	83	64
5	No	Yes	Not found	Not found										
5	Yes	Yes	Not found	9 12 23 32 37		80	88	81	79	83	67	80	83	65
6	No	No	Not found	Not found										
6	Yes	No	Not found	Not found										
6	No	Yes	Not found	Not found										
6	Yes	Yes	Not found	9 12 14 23 32 37		80	94	86	89	67	58	88	92	56

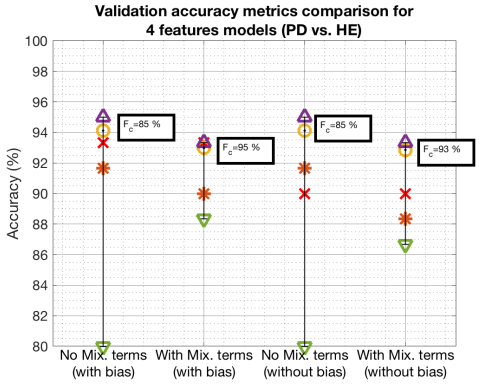
The results in Tabs. 6.5 and 6.6 are depicted in Figs. 6.10 and 6.11 where the CS, PS, realistic, pessimistic and optimistic Accuracy are presented.



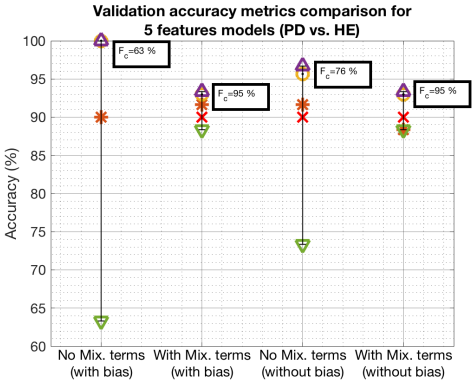
(a) 2 Features



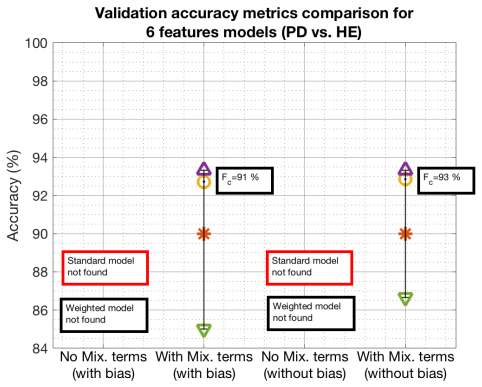
(b) 3 Features



(c) 4 Features



(d) 5 Features



(e) 6 Features

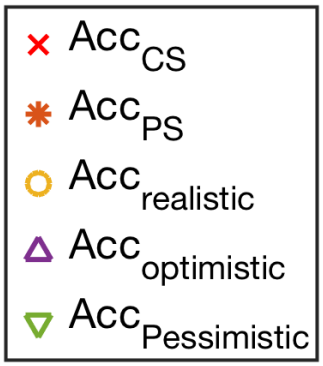
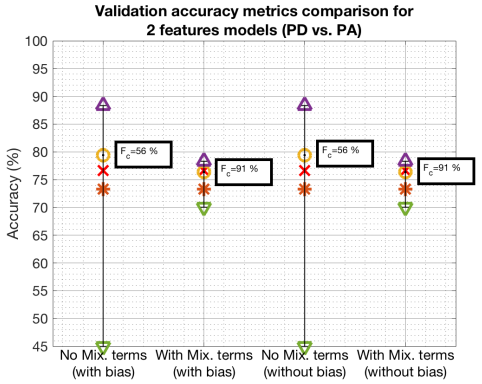
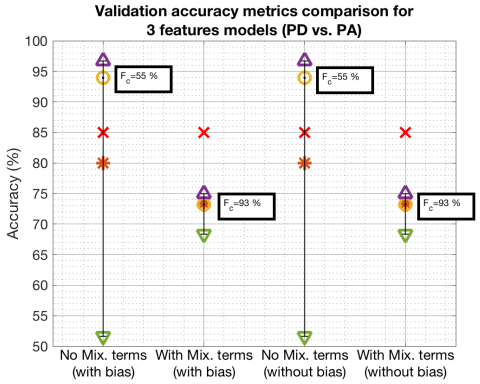


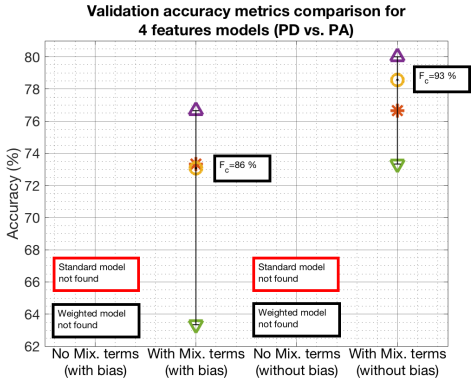
Fig. 6.10 Classification results for the PD vs. HE subset using original data extracted from the whole chain.



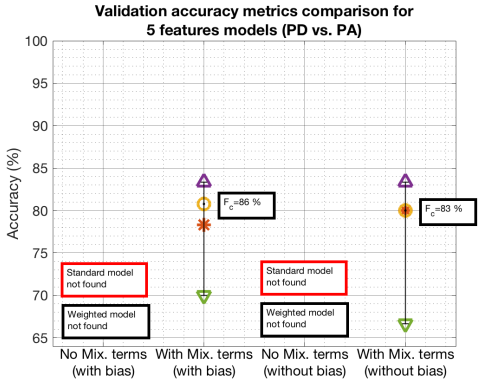
(a) 2 Features



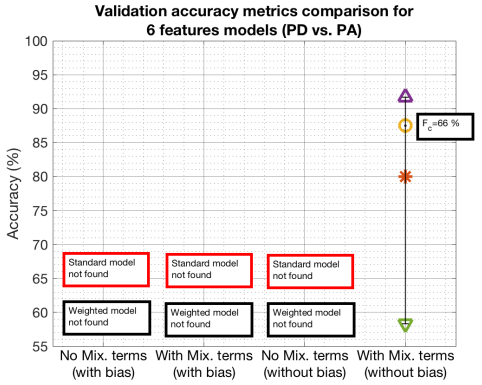
(b) 3 Features



(c) 4 Features



(d) 5 Features



(e) 6 Features

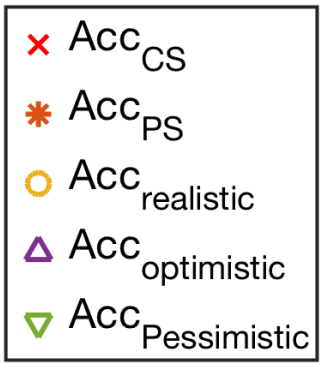


Fig. 6.11 Classification results for the PD vs. PA subset using original data extracted from the whole chain.

6.7 Classification models validation

In order to validate the models trained in the previous sections, a subset of subjects was extracted from the dataset. During the research activities, due to the difficulties in accessing public health facilities due to the Covid 19 emergency, the validation dataset could not be balanced in terms of subject age as in the case of the training subset. The validation subset includes:

- Parkinson patients (PD): N=10 (5 M, 5 F), mean age 59, standard deviation 7
- Healthy subjects (HE): N=10 (5 M, 5 F), mean age 36, standard deviation 9
- Pathological non-Parkinsonian patients (PA): N=10 (5 M, 5 F), mean age 46, standard deviation 12

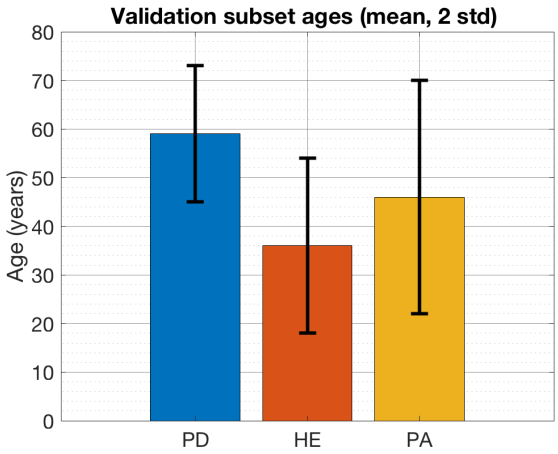


Fig. 6.12 Mean ages of the validation subset.

The audio files of the validation subset were processed in order to extract the features and evaluate the extraction uncertainty as shown in Sec. 5.1. The features and the relative uncertainties were used to predict the classification of PD, HE and PA subjects using the models trained in the previous sections. The tables 6.7 and 6.8 summarise the results of the classification metrics.

6.7.1 Validation of the models trained with the original data

In this section the models trained with the original data, as described in Sec. 6.5.1, were used to predict the health status of the PD vs. HE subset and the PD vs. PA subset.

PD vs. HE validation results

Table 6.7 CS and PS accuracy metrics for the PD vs. HE subset (validation).

N° Features	Mixed terms eval.	Bias removal	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	Scorep %
2	No	No	70	75	83	93	64	85	68	80	83	39
2	Yes	No	70	75	82	88	62	97	73	76	77	47
2	No	Yes	73	75	81	90	65	90	72	80	82	47
2	Yes	Yes	73	72	83	71	84	88	68	77	80	56
3	No	No	75	75	83	93	68	82	67	82	85	42
3	Yes	No	75	75	75	78	68	95	70	74	75	60
3	No	Yes	68	70	76	89	52	80	58	73	78	22
3	Yes	Yes	68	70	78	81	68	87	65	75	78	52
4	No	No	65	65	62	86	50	87	60	69	73	24
4	Yes	No	65	73	83	87	64	92	70	76	78	47
4	No	Yes	63	68	79	92	46	83	58	70	75	12
4	Yes	Yes	63	70	78	82	63	92	67	73	75	47
5	No	No	65	62	63	71	53	67	42	63	75	23
5	Yes	No	65	63	52	76	44	90	55	61	65	23
5	No	Yes	63	67	70	79	57	85	58	69	73	36
5	Yes	Yes	63	75	78	79	64	90	65	72	75	50
6	No	No	Not found	Not found								
6	Yes	No	Not found	63	56	80	45	87	57	65	70	22
6	No	Yes	Not found	Not found								
6	Yes	Yes	Not found	78	84	82	72	88	68	77	80	58

PD vs. PA validation results

Table 6.8 CS and PS accuracy metrics for the PD vs. PA subset (validation).

N° Features	Mixed terms eval.	Bias removal	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	Scorep %
2	No	No	49	63	90	96	70	58	51	88	93	25
2	Yes	No	49	60	74	33	96	91	58	63	67	-5
2	No	Yes	54	70	99	100	82	65	60	92	95	42
2	Yes	Yes	54	72	85	83	59	98	70	71	72	47
3	No	No	54	86	100	100	100	54	54	100	100	54
3	Yes	No	54	60	71	36	92	89	58	65	68	2
3	No	Yes	49	84	99	100	93	58	56	97	98	49
3	Yes	Yes	49	70	80	58	92	88	67	76	79	33
4	No	No	Not found	Not found								
4	Yes	No	49	70	70	48	100	81	60	74	79	7
4	No	Yes	Not found	Not found								
4	Yes	Yes	51	82	90	86	83	91	77	85	86	75
5	No	No	Not found	Not found								
5	Yes	No	53	86	98	96	95	77	74	95	96	73
5	No	Yes	Not found	Not found								
5	Yes	Yes	54	72	76	52	91	77	56	73	79	17
6	No	No	Not found	Not found								
6	Yes	No	Not found	Not found								
6	No	Yes	Not found	Not found								
6	Yes	Yes	Not found	65	64	47	95	63	46	72	82	-2

6.7.2 Boosting method validation

The models trained with the boosting method described in Sec. 6.5.2 were used to predict the subject's class using data extracted from the validation subset. The original data were boosted using the methods described in the previous sections.

PD vs. HE validation results (boosting method)

Table 6.9 CS and PS accuracy metrics for the PD vs. HE subset (validation, boosted).

N° Features	Mixed terms eval.	Bias removal	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	Scorep %
2	No	No	74	71	85	75	72	80	59	74	79	56
2	Yes	No	74	77	83	90	63	95	74	78	79	47
2	No	Yes	75	77	82	90	62	92	71	77	79	43
2	Yes	Yes	75	75	71	78	69	97	72	74	75	63
3	No	No	74	73	80	85	69	83	65	78	82	48
3	Yes	No	74	74	84	79	72	95	72	76	77	65
3	No	Yes	68	76	81	86	66	87	67	77	80	47
3	Yes	Yes	68	75	78	79	69	98	73	75	75	64
4	No	No	77	74	77	89	65	83	65	79	82	41
4	Yes	No	77	79	87	87	74	92	75	81	83	61
4	No	Yes	70	77	83	84	77	85	69	81	84	62
4	Yes	Yes	70	75	79	78	69	97	72	74	75	63
5	No	No	71	75	90	93	73	71	60	83	88	40
5	Yes	No	71	69	85	62	81	95	68	71	73	48
5	No	Yes	68	72	77	77	72	85	63	75	78	58
5	Yes	Yes	68	68	76	75	63	91	63	70	72	52
6	No	No	68	68	80	82	56	87	61	71	75	35
6	Yes	No	68	64	78	64	67	94	61	65	67	58
6	No	Yes	45	72	76	78	69	93	69	74	76	59
6	Yes	Yes	45	70	80	76	64	99	70	70	71	58

PD vs. PA validation results (boosting method)

Table 6.10 CS and PS accuracy metrics for the PD vs. PA subset (validation, boosted).

N° Features	Mixed terms eval.	Bias removal	CS Acc %	PS Acc %	AUC %	Sens. %	Spec. %	Fraction of classified %	Pess. Acc. %	Real. Acc. %	Opt. Acc %	Scorep %
2	No	No	56	67	86	93	42	81	58	72	77	7
2	Yes	No	56	58	67	79	40	88	52	59	64	13
2	No	Yes	58	52	55	66	35	76	40	53	65	10
2	Yes	Yes	58	77	88	83	76	88	70	80	82	64
3	No	No	59	83	96	100	86	71	68	95	96	53
3	Yes	No	59	87	96	97	91	89	84	94	95	78
3	No	Yes	53	84	90	90	83	67	58	87	91	51
3	Yes	Yes	53	84	88	85	81	84	70	83	86	66
4	No	No	71	90	98	100	89	72	69	96	97	58
4	Yes	No	71	89	98	100	81	92	84	92	93	66
4	No	Yes	58	89	96	100	90	77	74	95	96	64
4	Yes	Yes	58	86	96	100	76	86	77	90	91	53
5	No	No	76	86	97	100	80	74	69	93	95	49
5	Yes	No	76	84	97	100	72	91	80	88	89	52
5	No	Yes	66	87	99	100	95	76	74	98	98	69
5	Yes	Yes	66	88	96	100	84	93	86	92	93	70
6	No	No	67	89	96	100	79	70	64	92	95	43
6	Yes	No	67	87	95	100	74	89	80	89	90	54
6	No	Yes	47	91	99	100	95	79	78	98	98	73
6	Yes	Yes	47	89	94	93	88	96	88	91	91	82

6.8 Results discussion and comparisons

This section summarizes the obtained data in order to highlight the main results of the training experiments. In particular the effect of Bias removal, mixed terms evaluation, data boosting will be analysed. Moreover a comparison between the classification performances of the "short" measuring chain (considering just the feature extraction contribution) and the performances of the whole measuring chain will be discussed. Lastly, a discussion on the data presented for the validation experiments will be carried out in order to evaluate the prediction capability of the proposed classification methods.

6.8.1 Effects of Bias removal

As shown in Tabs. 6.1, 6.2 and in the plots in Fig. 6.6, 6.7, the accuracy metrics of the proposed method does not seem to improve after the Bias removal. As shown in the plot in Fig. 6.13 a comparison between the accuracy metrics obtained without the mixed terms evaluation is presented. The trained models, using a variable number of features between 2 and 5, are compared removing or not the bias as shown in Eq. 6.30. From the plot in Fig. 6.13, no improvement is noticeable in terms of realistic, pessimistic and optimistic accuracy when the Bias is removed except for the 5 features model, where a relevant increase in the pessimistic accuracy is present.

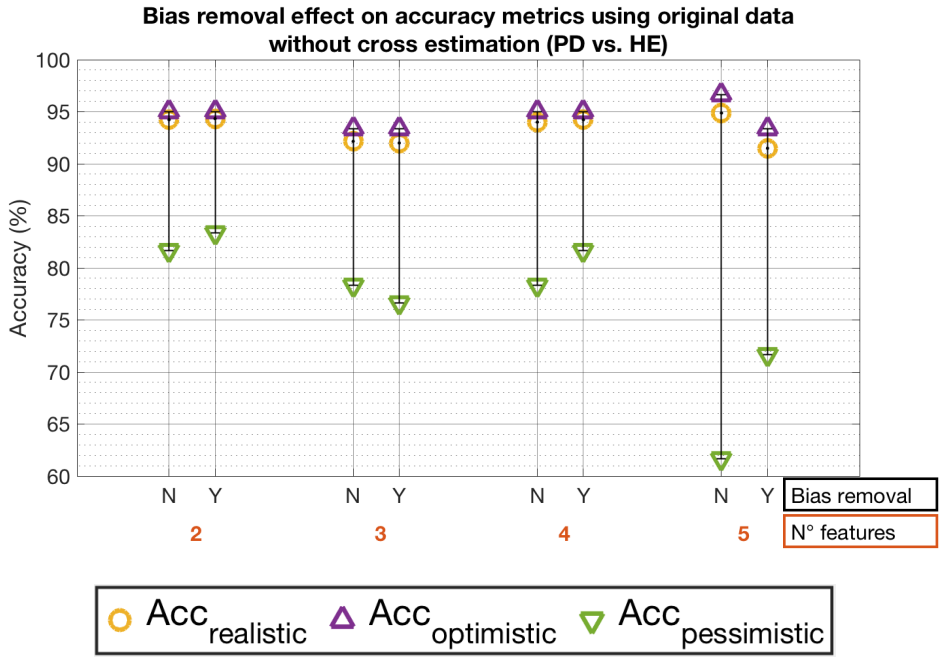


Fig. 6.13 Effect of Bias removal on the classification metrics.

6.8.2 Effects of mixed terms evaluation

As shown in Tabs. 6.1, 6.2 and in the plots in Fig. 6.6, 6.7, the accuracy metrics of the proposed method are highly affected by the mixed terms evaluation of features and model coefficient uncertainties. As in the previous case, the models trained with or without the mixed terms evaluation, using the bias removal process, are compared for a number of features from 2 to 5. As shown in Fig. 6.14, the mixed terms evaluation of uncertainties leads to a reduction of the distance between pessimistic and optimistic accuracy, thus giving more robust classification models. This is due to the reduction of the confidence interval around each probability value, which implies a reduced number of non-classified subjects and thus a higher pessimistic accuracy and fraction of classified. One should note that increasing the number of features makes impossible to find a combination of features that meet the correlation condition ($|r|<0.1$) that was set to consider the mixed terms of the uncertainty as negligible as shown in Sec. 6.2.2. In fact, looking at Tabs. 6.1 and 6.2 the number of "Not found" model increases as the number of used features rises. This issue is partially solved evaluating the mixed terms of the uncertainty with a higher correlation ($|r|<0.7$). The mixed terms on Eq. 6.20 (the synthetic formulation of Eq. 3.7) may be negative, so

the contribution of the quadratic terms of the uncertainty propagation equation is limited, thus the confidence interval size of the predicted probabilities is reduced. This fact leads to an increased fraction of classified and, consequently, a higher pessimistic accuracy.

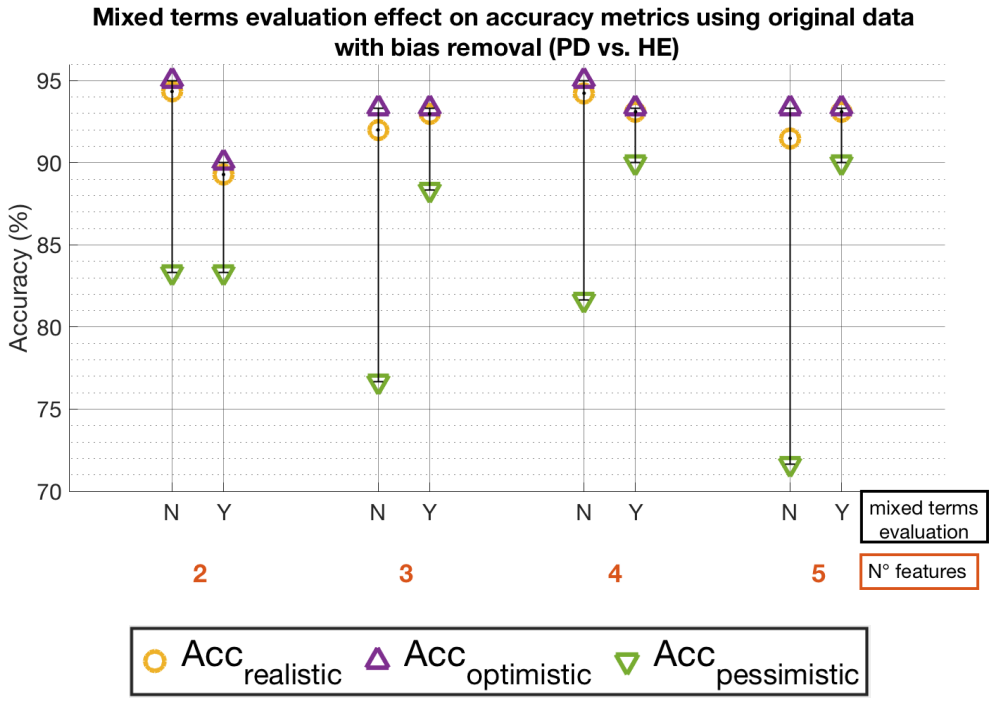


Fig. 6.14 Effect of mixed terms evaluation on the classification metrics.

Another way to improve the classification metrics consists in increasing the size of the dataset, as showed in the next section.

6.8.3 Boosting technique using the artificial data

As shown in the previous section, the best training practice for the classification model seems to be to evaluate the mixed-terms of the uncertainties and, even though a negligible effect is noticed, removing the bias from the training data. For this reason the plot in Fig. 6.15 shows only the non-validated accuracy metrics of the models trained with the artificial data that were treated removing the bias and evaluating the uncertainty mixed terms.

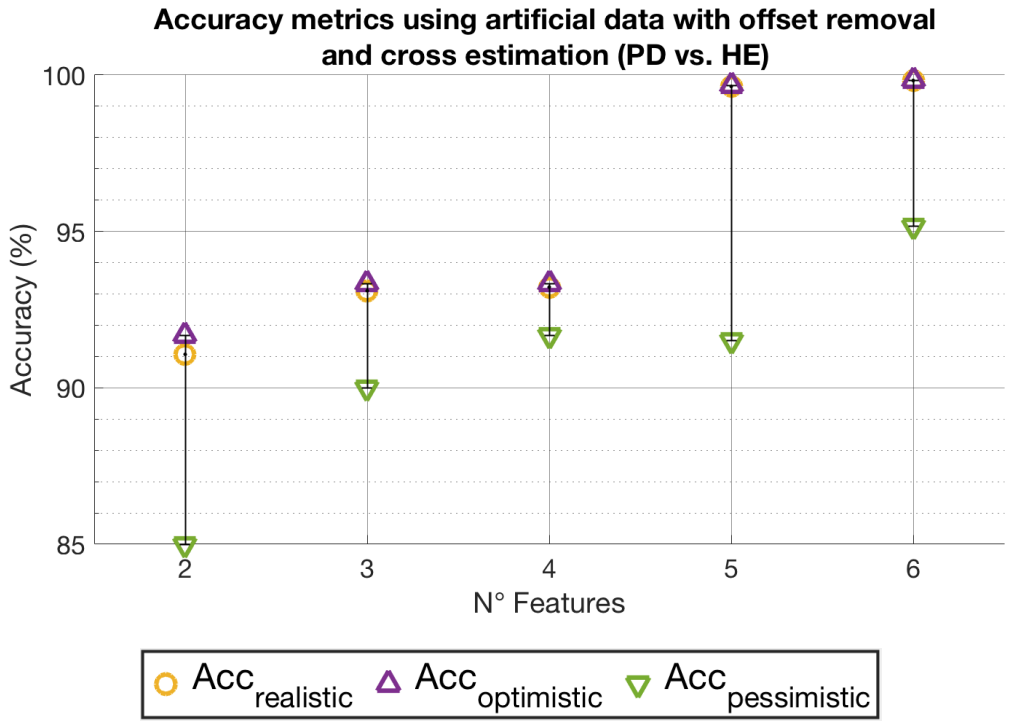


Fig. 6.15 Accuracy metrics comparison between models with different number of features.

Looking at Tabs 6.3 and 6.4, accuracy values of 100 % are present and such fact is often a sign of perfect separation which causes a large gap between the classes in the feature hyperplane. In such cases the metrologic frequentist approach does not allow to consider the gap area as a "dead zone" where no data is present and never will be. The frequentist approach forces us to consider the decisional boundary as a bundle of curves with very large coefficients uncertainties. In order to avoid this issue the perfect separation models was discarded for the boosting method searching for models with a score as high as the one being discarded. As an example, let's consider the last row of Tab 6.3, which reports the accuracy metrics for the PD vs. HE classification using 6 features with bias removal and mixed terms evaluation. Such a model reports a 100 % weighted, realistic and optimistic accuracy. Such model has the following coefficients:

Table 6.11 Coefficients and uncertainties of a 6 features model (last row of Tab. 6.3).

	β_0	β_1	β_2	β_3	β_4	β_5	β_6
Coefficients Estimates	786	-477	-7	-885	-27	162	24
Standard Errors	79	48	0.7	90	3	17	4
Coefficients relative Errors	10 %	10 %	10 %	10 %	10 %	11 %	15 %

The data presented in Tab. 6.11 show very large coefficients are present. As shown in Fig. 6.1, for high absolute values of the coefficient, the slope of the central part of the sigmoid rises tending to a step function. This fact may reduce the contribution of the features in sum of Eq. 6.1, so even a small feature variation may produce very large exponents in Eq. 6.2. The effect of such a perfect separation on prediction accuracy can be noticed in the validation of the trained models.

6.8.4 Measuring chain length

In this section, the non-validated Accuracy metrics of the short chain (Extraction contribution, EXT) and the long chain (Acoustic contribution, ACO) are compared. In Fig. 6.16 a comparison between the accuracy metrics obtained from the datasets extracted from the short chain (**S**) and the long chain (**L**) is presented. The comparison is performed using the models trained with the bias removal and the mixed terms evaluation for a number of features ranging between 2 and 6.

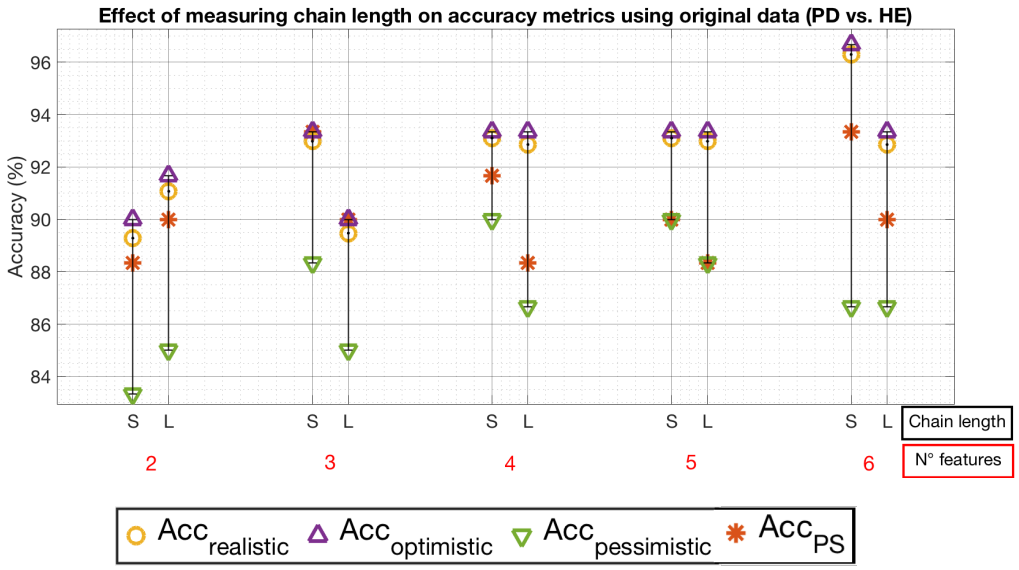


Fig. 6.16 Accuracy metrics comparison between models with different number of features.

As shown in Fig. 6.16 , the accuracy metrics obtained from the short and long measuring chain are comparable. However the accuracy metrics of the long chain shows reduced performances in terms of realistic, pessimistic and optimistic accuracy except for the model with 2 features. The models from the short and the long chain have selected different sets of features as summarised in Tab. 6.12.

Table 6.12 Selected features comparison between the models trained with the short and long measuring chain.

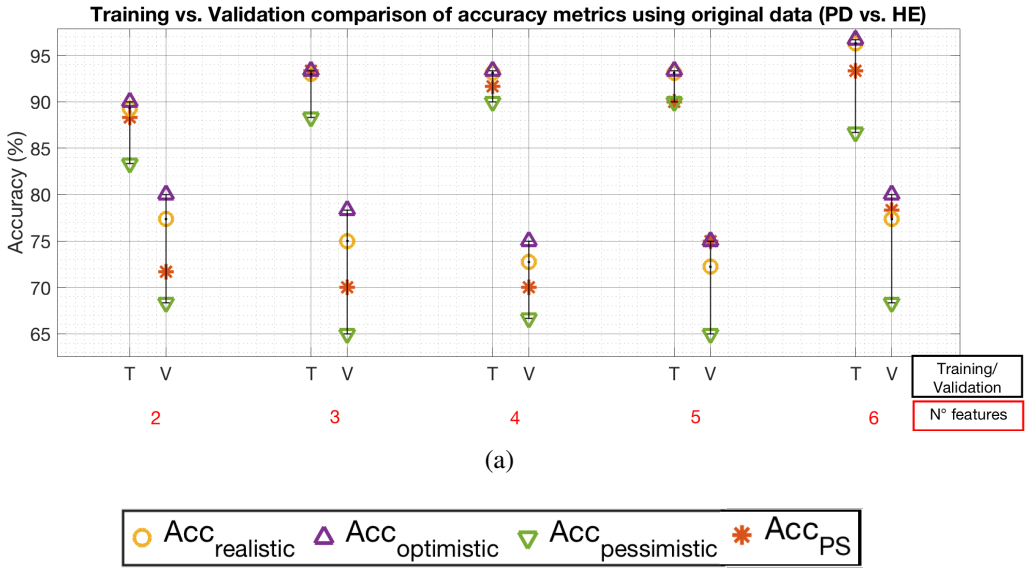
N° Features	**Short chain features**	**Long chain features**
2	jit, HNR_{mode}	HNR_{mean}, $A_{RMS}(med)$
3	shi, HNR_{mode}, $A_{RMS}(mean)$	HNR_{mean}, $A_{RMS}(mode)$, $A_{RMS}(std)$
4	shi_{abs}, HNR_{mode}, $A_{RMS}(mean)$, $CPPS_{std}$	vf_o, HNR_{med}, $A_{RMS}(mode)$, $A_{RMS}(std)$
5	shi, vAm, HNR_{mode}, $A_{RMS}(mean)$, $CPPS_{std}$	vf_o, HNR_{mean}, HNR_{mode}, $A_{RMS}(mode)$, $A_{RMS}(std)$
6	jit, apq, vAm, HNR_{mode}, $A_{RMS}(mean)$, $CPPS_{std}$	vf_o, shi, vAm, HNR_{mode}, $A_{RMS}(med)$, $CPPS_{range}$

As shown in Tab. 6.12, the selected features are almost always different for the short and long chain models. However, the corresponding selected features for the two chains belong to the same feature families (stability, HNR , f_o, A_{RMS}, CPPS). This consideration may suggest that the training of the models may change if a perturbation of the measuring chain is introduced, but the "informativeness" of some features remains unchanged. This behaviour may be conceptually compared to the

behaviour of a natural intelligence as described in Sec. 1.8. When the environmental set-up of the training experiment is perturbed, the selected features used to evaluate a decision may change to better adapt to the new conditions.

6.8.5 Models validation

In this section a comparison between the accuracy metrics of the unvalidated model and the validation predictions will be discussed. In Fig. 6.17, a comparison of the accuracy metrics of the unvalidated models and the accuracy of the predictions performed on the validation set is shown. The selected data were processed in order to remove the bias and the models were trained considering the mixed terms evaluations. The comparison was carried out using the original data of the PD and HE subset as features input.



(a)

Fig. 6.17 A comparison between the accuracy metrics of the trained models and the accuracy metrics of the predictions of the validation subset.

As shown in Fig. 6.17, the accuracy metrics of the predictions performed on the validation subset are considerably lower than the accuracy metrics of the unvalidated models. As an example, the realistic accuracies of the validation predictions (yellow circles) are in a range between 72 % and 77 %, while for the training accuracy such a range is between 89 % and 96 %. This effect can be caused by the fact that the validation dataset is unbalanced with respect to the ages of the subjects that compose

it and is also unbalanced with respect to the average age of the training dataset. The same analysis was carried out using the models trained with the boosting method. The selected data were processed in order to remove the bias and the models were trained considering the mixed terms evaluations. In Fig. 6.18 an example of the accuracy metrics of **T**rained and **V**alidated predictions is depicted.

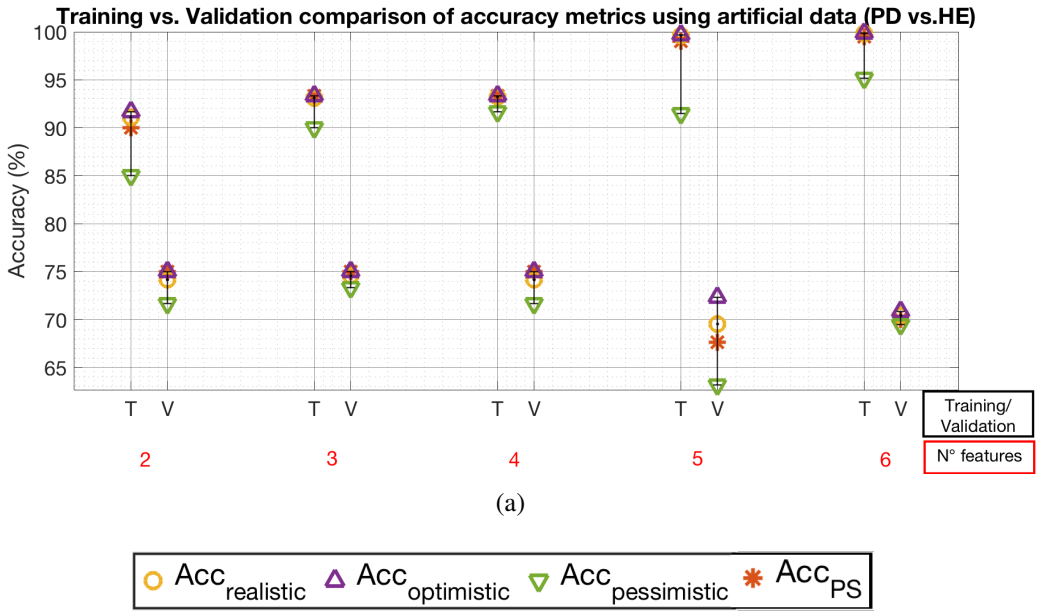


Fig. 6.18 A comparison between the accuracy metrics of the trained models and the accuracy metrics of the predictions of the validation subset (data boosting, PD vs. HE).

As shown in Fig. 6.18, the accuracy metrics of the predictions on the validation subset are noticeably lower than the training accuracy metrics. As an example, the realistic accuracies of the validation predictions (yellow circles) are in a range between 75 % and 78 %, while for the training accuracy such a range is between 91 % and 100 % Considering the PD vs. PA classification results reported in Tab. 6.7.2 higher accuracy were obtained from the prediction of the validation dataset, as shown in Fig. 6.19.

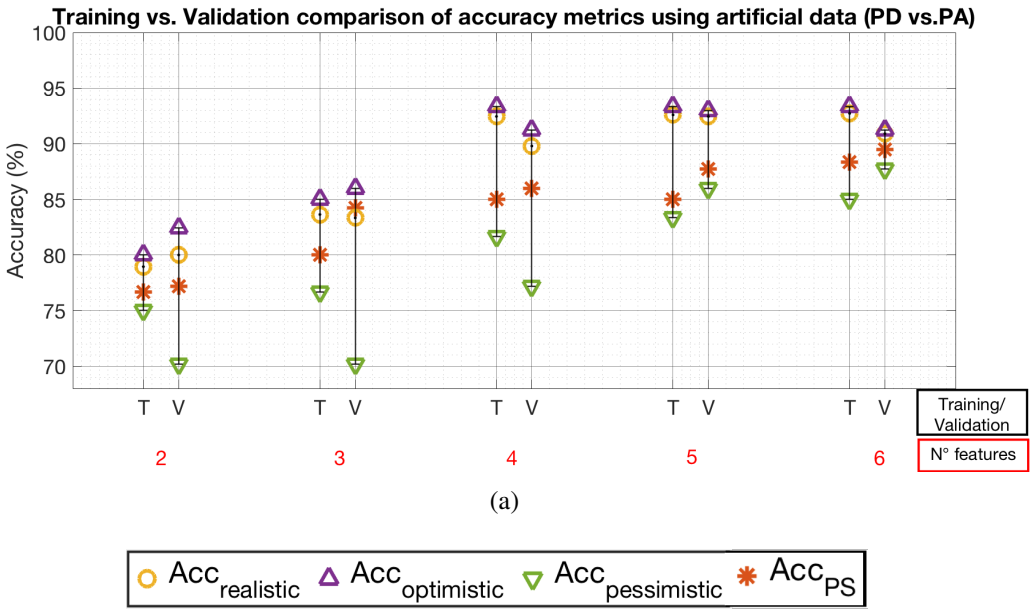


Fig. 6.19 A comparison between the accuracy metrics of the trained models and the accuracy metrics of the predictions of the validation subset (data boosting, PD vs. PA).

From the plot of Fig. 6.19, is clearly visible that the accuracy metrics of the training and the validation accuracy are comparable. This may be caused by the fact that the ages of the PD vs. PA subset (PD=59, PA=46) are more balanced than the ages of the PD vs. HE subset (PD=59, PA=36), therefore the trained models are more representative for the subjects whose ages are closer to the ages of the training dataset ($\approx$ 52 years).

These considerations on the accuracy metrics of the validation predictions, highlight the importance of having balanced datasets to train and validate the classifications models. Due to the difficulties in accessing in public health structures during the Covid-19 sanitary emergency, the collection of more data to train and validate the classification models was not possible.

Chapter 7

Conclusions

In this section, the final conclusions of the work described in this manuscript will be presented. For clarity reasons the chapters from 3 to 6 will be discussed separately in order to draw the main conclusions for each topic of this manuscript.

7.1 Chapter 3

In this chapter, the conclusions for the evaluations carried out in Chapter 3 will be presented.

7.1.1 Stability metrics: the time-base aging negligibly affects the pseudo-period evaluations

The effect of the aging of the timing crystal on period measurement was analysed in Sec. 3.1.1. A worst case scenario of a commercial Room Temperature Crystal Oscillator, with an aging time drift $k = 10^{-6}$ month^{-1}, was evaluated using Eq. 3.4. Considering a vowel with a duration of 10 s, the relative uncertainty contribution to period measurements was estimated as 10^{-6} ppm, therefore it can be considered negligible.

7.1.2 Stability metrics: the time-base tolerance does not affects the stability metrics

The effect of a tolerance on the timing crystal frequency was faced in Sec. 3.1.1. A pessimistic clock tolerance of 100 ppm was considered to evaluate its effect on period measurement and the stability metrics. Although the tolerance can considerably affect the period evaluation, it has no effect on some stability metrics. In particular the stability metrics equations which have the periods T_i at the denominator are not affected by the tolerance because the terms $(1+\varepsilon)$ cancel out in the metrics equation as exemplified in Eq. 3.6. The timing tolerance affects the jit_{abs} evaluations (see Eq. A.2) and, for a 100 ppm clock tolerance, the relative uncertainty is 0.01 %.

7.1.3 Stability metrics: the time-base resolution is the main uncertainty contribution for period metrics

As stated in Sec. 3.6, the resolution of the ADC time base has an important effect on the evaluation of period durations. The absolute uncertainty for a period measurement was considered as caused by a random perturbation with an uniform distribution, therefore it was estimated as $2 \cdot T_s/\sqrt{3}$, where T_s is the sampling period. For a vowel sampled at 44.1 kSa/s, the uncertainty was estimated as $u(T) \approx 26$ μs, therefore the relative contribution of the time-base resolution was estimated in a range between 0.2 % and 1 %. These values are higher than the aging and tolerance contributions so the time-base resolution is the main uncertainty contribution to pseudo-period evaluations and thus to the stability metrics. As shown in Sec. 3.2.1, an oversampling of the signal can reduce the effect of the time-base resolution on the period stability metrics.

7.1.4 Stability metrics: the amplitude resolution is NOT the main uncertainty contribution for amplitude metrics

An analysis of the amplitude uncertainty contribution was carried out in Sec. 3.1.2, where three contributions were identified: Quantisation, Integral Nonlinearity (INL) and Gain Error (GE). Considering these contributions as caused by a random perturbation with an uniform distribution, for a signal sampled with a 16 bit amplitude

resolution, the amplitude uncertainty was estimated as $u_{LSB}(A) \approx 2.5$ LSB. This evaluated uncertainty was used as a reference perturbation for the Monte Carlo propagation presented in Sec . 3.2. Considering a full scale range of ± 1 a.u., the amplitude uncertainty was evaluated as $u(A) \approx 3.8 \cdot 10^{-5}$ a.u., therefore it may be considered as negligible. However, other uncertainty contributions can affect the amplitude metrics and they are caused by the difficulties in setting an adequate input gain for each of the subjects. In a real life application, in example when the recordings are collected with a smartphone, the microphone gain settings cannot be guaranteed. An approximated gain setting could be performed in order to give the same headroom for each of the recordings ($A_{RMS}(mean)$ =-6 $\mathrm{dB_{fs}}$ for this work). However the setting of each recordings gain was performed by hand by the author taking as a reference the A_{RMS} level indicated by the portable recorder display. For this reason the author cannot guarantee that each recording was recorded with the same $A_{RMS}(mean)$ level, therefore a signal normalization was applied to each of the recordings in order to compare them.

7.1.5 Stability metrics: the analytical uncertainty propagation is easy to obtain for some simple metrics

A GUM oriented analytical propagation of jitter and shimmer metrics uncertainty was presented in Sec. 3.1.2. The analytical propagation formula for jitter measurements was obtained as in Eq. 3.9. Such a formulation depends on pseudo-periods T_i and the number of evaluated periods N. The jitter uncertainty equation was represented as the product of the uncertainty $u(T)$ and a sensitivity coefficient $C(T,N)$. To have an evaluation of the order of magnitude of the jitter uncertainty, a parametrization of the sensitivity coefficient was carried out to obtain Eq. 3.13 so that the sensitivity coefficient becomes a function of jit, N and a fractional term F_N (Eq. 3.12) To evaluate the term F_N, a statistical evaluation on the available dataset was carried out, as shown in Fig. 3.2. Replacing the median value of this statistical analysis, an evaluation of the sensitivity coefficient was carried out, as shown in the heatmap in Fig. 3.3. Considering a common case scenario of a vowel with a duration of 5 s and a fundamental frequency of 100 Hz, sampled at 44.1 kSa/s. , the absolute standard uncertainty of the jitter was estimated as $u(jit) = 0.018$ %. Considering the range of jitter measurement of the available dataset, its relative uncertainty was estimated in a range from 0.3 % to 16 %.

The same procedure was applied to the shimmer uncertainty evaluation, substituting the pseudo-periods T_i with the amplitudes A_i. A statistical analysis of the term F_N was carried out and the sensitivity coefficient was parametrized as in the jitter case. Considering the amplitude uncertainty contribution described in the previous conclusions, the shimmer uncertainty was evaluated for a vowel with a 5 s duration, a fundamental frequency of 100 Hz, sampled with an amplitude resolution of 16 bit, and its value was estimated as $u(shi) = 0.0004$ %. Considering the range of shimmer measurement of the available dataset, its relative uncertainty was estimated in a range from 0.02 % to 0.6 %.
As stated in Sec. 3.1.2, the analytical propagation of some stability metrics can be a very challenging task, especially for those equations with a nested mathematical structure (i.e. ppq in Eq. A.4 and apq in Eq. A.8). For these reasons a Monte Carlo uncertainty propagation was carried out, as described in the next section.

7.1.6 Stability metrics: the Monte Carlo uncertainty propagation highlighted a bias in some metric evaluations

A Monte Carlo uncertainty propagation on the stability metrics was carried out in Sec. 3.2. The sequences of pseudo-periods and amplitudes, extracted from the recordings of the available dataset, were perturbed considering the uncertainty contributions described in Sec. 3.1.2, $\pm 1\ T_s$ for the periods perturbation and ± 2.5 LSB for the amplitudes. This propagation highlighted a bias in each stability metrics where the absolute value operator is present in the equation, i.e. for each stability metric with the exception of vf_o (Eq. A.5) and vAm (Eq. A.9). This happens because the absolute value operator guarantees a strictly positive accumulation of the perturbation contribution causing a bias in the evaluation of the considered metric.

7.1.7 Stability metrics: the higher the sampling rate is the lower is the uncertainty of period metrics

To evaluate the effect of the time-base resolution on stability metrics, a linear oversampling of the vowel signals was carried out to simulate different sampling rates. Starting from the original sampling rate of 44.1 kSa/s, four oversampling factors were tested: 1, 2, 4 and 8, where the oversampling factor 1 corresponds to

the original sampling rate. The results presented in Tab. 3.1 highlight a reduced uncertainty on period stability metrics for higher oversampling factors, therefore for higher sampling rates. In particular the dispersion and bias contribution of the period metrics is lower for higher oversampling factors. To limit the effects of low sampling rates on jitter evaluations, a compensation method has been proposed comparing the expected (unperturbed) jitter values and the extracted ones using two oversampling factors: 1 and 8. As depicted in Fig. **??**, the linear regression of extracted vs. expected jitter values shows a linear relation for the oversampling factor 8. Keeping the original sampling rate of 44.1 kSa/s (oversampling factor 1), the extracted jitter values are more dispersed than in the case of oversampling 8. The linear regression obtained from this data highlighted a slope of 0.91 and an offset of 0.15 % with a larger RMSE. Using these informations, a compensation of the jitter uncertainty is possible removing the offset and dividing for the slope as exemplified in 3.19. As expected, the amplitude stability metrics are not affected by the oversampling factor.

7.1.8 Stability metrics: the higher the amplitude resolution is the lower is the uncertainty of amplitude metrics

To evaluate the effect of amplitude resolution on stability metrics, a bit reduction of the vowel signals was carried out to simulate different amplitude resolutions. Three bit resolutions were tested: 10 bit, 12 bit and 16 bit. The results presented in Tab. 3.3 highlight a reduced uncertainty on period stability metrics for higher amplitude resolutions. In particular the dispersion and bias contribution of the period metrics is lower for higher bit resolutions. As expected, the period stability metrics are not affected by the amplitude resolution. Differently from the jitter case, the linear regression, evaluated using the expected and extracted shimmer values, showed a very linear relation between the just mentioned data, so the compensation method is useless for shimmer measurements when using lower amplitude resolutions.

7.1.9 Stability metrics: the background noise negligibly affects the stability metrics if the extraction algorithm contribution is not considered

An analysis on the effect of background noise on stability measurements was carried out in Sec. 3.2.3. For this evaluation, a white gaussian noise was added to the original signals in order to obtain three target *NSRs*: -18 dB, -12 dB and -6 dB. If the sequences of extracted pseudo-periods and amplitudes are considered to be free of uncertainty, the background noise have a negligible effect on stability metrics as reported in Tab. 3.4. This is due to the trust the experimenter has on the capability of the extraction algorithm in evaluating given sequences of pseudo-periods and amplitudes.

7.1.10 Stability metrics: the extraction algorithm affects the stability metrics

If the extraction algorithm is considered as "not perfect", a definition of a golden standard measurement is useful to evaluate the effect of different sampling rates, amplitude resolutions and background noise levels. Considering the data obtained in the previous evaluation, the golden standard parameters were set as:

- oversampling factor: 8
- amplitude resolution: 16 bit
- noise $NSR < -30$ $\mathrm{dB_{fs}}$.

Comparing the stability metrics extracted using different configurations for oversampling, amplitude resolution and noise level with the one extracted using the golden standard, an important effect can be noticed in the stability metrics uncertainties, as reported in Tab. 3.5. In particular, comparing the worst case scenario with the golden standard, the background noise can produce a jitter uncertainty which covers three orders of magnitude, while for shimmer measurement, the uncertainty can cover a range of five orders of magnitude. The effect of the extraction algorithm contribution was extensively analysed in Sec. 5.1.3.

7.1.11 The cross-talk effect on voice features is negligible

The cross-talk effect on voice features was evaluated in Sec. 3.3. This evaluation was carried out on a consumer level portable recorder (very similar to the one used for the acquisition of the original vowels) and produced the result depicted in Fig. 3.15, where a maximum cross-talk of -47 dB was found at low frequency ($\approx$ 100 Hz) and the minimum of -70 dB around 2.5 kHz. Considering the Signal to Noise Ratio, that was statistically evaluated from the normalized audio recordings close to -30 dB_{fs}, the effect of cross-talk can be considered as negligible on the evaluated voice features using the ZOOM H2N portable recorder. Considering that the recorder used for the subject's voice collection is also a consumer device, it is quite safe to say that the cross-talk effect can be considered negligible for consumer level audio recorders.

7.2 Chapter 4

In this chapter, the conclusions for the evaluations carried out in Chapter 4 will be presented.

7.2.1 Everyone is unique, even with respect to themselves

To evaluate the uncertainty contributions of each component of the measuring chain, the architecture depicted in Fig. 4.1 was used. In order to achieve these evaluations, a novel vowel re-synthesis method has been proposed in Sec. 4.1. The artificial vowels should have comparable statistical characteristics of pseudo-periods and amplitudes with respect to the original ones, therefore a study on the statistical distributions of pseudo-periods and amplitudes was carried out, as described in Sec. 4.2. This study highlighted that the distributions of pseudo-periods and amplitudes may be very different in terms of positioning, dispersion and shape. Such a variability was found among different subjects and even between the vowel repetitions of the same subject, as shown in Figs. 4.4 and 4.6. For this reason an *apriori* distribution of the pseudo-periods and amplitudes can not be defined.

7.2.2 Everyone has approximately the same vocal apparatus

The distributions of the consecutive differences of pseudo-periods and amplitudes, showed zero-centred bell-shaped distributions, as shown in Figs. 4.5 and 4.7. These distributions are way more repeatable than the pseudo-periods and amplitude distributions. This characteristics may be associated more to the physical limits of the vocal apparatus than to the subject ability in producing a sustained vowel. A preliminary study on the *normality* of the consecutive difference distributions was carried out to evaluate if fitting these distributions with a Gaussian curve was possible. This study evaluated the mean skewness and excess kurtosis parameters of the consecutive differences distributions of pseudo-periods to have a measure of their *normality*. The skewness values reported in Tab. 4.1 are close to 0, therefore the consecutive difference distribution are highly symmetrical. The excess kurtosis values, instead, are very far from 0 so the distributions are *leptokurtic*. This statistical evaluation forced the author to adopt the Monte Carlo sampling method described in Sec. 4.3, which consider each sampled vowel as a "unique" phonation event.

7.2.3 The Monte Carlo Perturbative method is better than the Markov Chain Monte Carlo method

The artificial vowels were re-synthesised taking as a reference the sampled distributions of pseudo-periods and amplitudes and their correspondent consecutive difference distributions. Two generation methods have been proposed: the Perturbative method and the Markov Chain Monte Carlo method. To compare these methods, a perceptual evaluation can be performed by the reader, downloading the audio examples using the QR codes in Figs. 4.15, 4.16 and 4.17. As can be noted listening to the audio files, the vowels re-synthesised with the PM method sounds way more realistic than the ones re-synthesised with the MCMC method. This is due to the fact that the MCMC method, despite it produces comparable statistical distributions with respect to the original ones, is not "bounded" to the original periods and amplitudes time sequences, unlike the PM method. The performances of the two methods were evaluated also in terms of spectral and cepstral characteristics of the artificial vowels. As can be noticed in Fig. 4.18, the PM method seems to affect less the mean spectrum of the example vowel, if compared to the MCMC method. Moreover an example evaluation of the CPPS is depicted in Fig. 4.21, where negligible differences can

be noticed between the distributions relative to the two methods, even though both method seem to cause a shift of the distribution toward smaller values of CPPS. This effect was extensively evaluated in Chapter 5.

7.3 Chapter 5

In this chapter, the conclusions for the evaluations carried out in Chapter 5 will be presented.

7.3.1 The Monte Carlo generation algorithm is not perfect but it works

In Sec. 5.1.1, an evaluation of the performances of the PM Monte Carlo generation method was presented. As can be noticed in the example in Fig. 5.3, the artificial vowels present some bias and dispersion with respect to the original vowels and these contributions were evaluated as reported in Tab. 5.2 and Tab. 5.3. Regarding the dispersion of the artificial stability metrics, the evaluation of the intra-subject dispersion reported in Tab. 5.4 highlighted that the dispersion of the generated stability metrics is smaller than the intra-subject dispersion. Such a consideration highlights that the generation method produces artificial values that are statistically closer to the single original vowel, used as a reference, instead of being scattered within the subject natural dispersion of task repetitions.
The electrical analogy depicted in Fig. 5.2 has been used to explain the role of the Monte Carlo generation method in the proposed uncertainty evaluation technique. The Monte Carlo method is considered as a trusted feature generator, therefore if the reference values that produces have a bias with respect to the original value it is not a critical issue unless the generated values are too far from the original one. The generation bias reported in Tab. 5.2 is lower than intra-subject dispersions reported in Tab. 5.4, so the generated metrics fall "inside" the subject natural dispersion. Anyway, the generation bias, as well as all the other biases caused by the measuring chain components, is not a critical issue if a reasonable number of artificial vowel is produced because the bias can be removed from the features evaluation as it was showed in chapter 6.

7.3.2 The extraction algorithm is not perfect

An evaluation of the uncertainty contributions of the extraction algorithm to periods and amplitudes measurements was carried out in Sec. 5.1.2. The data presented in Tab. 5.5 represent the signals that were pre-processed adopting the golden standard parameters: oversampling=8, amplitude resolution=16 bit. The pseudo-period uncertainties reported in Tab. 5.5 are similar to the sampling period at 44.1 kSa/s ($T_s \approx 22\ \mu$s) and to the period uncertainty evaluated in Sec. 3.6 ($u(T) \approx 26\ \mu$s). These values are considerably higher than the values obtained with the analytical evaluation carried out in Sec. 3.6, where an uncertainty equation $u(T) = 2 \cdot T_s/\sqrt{3}$ has been defined. Using this equation and considering a $T_s = 22\ \mu\text{s}/8 \approx 2.8\ \mu$s the uncertainty should be $u(T) \approx 3.8\ \mu$s. This consideration may suggest that the quantisation is not the main period uncertainty contribution, as stated before, and the extraction algorithm adds some uncertainty to the pseudo-periods evaluations.
Regarding the amplitude uncertainty, the evaluations presented in Tab. 5.5 are noticeably higher than the values evaluated with the analytical method ($u(A) \approx 10^{-4}$a.u.). The same considerations made for the extraction contribution on pseudo-periods uncertainty can be sustained for the amplitude uncertainty. As stated for the period uncertainty evaluation, the amplitude contributions (resolution, INL, GE) are not the main uncertainty contributions, therefore some uncertainty is added by the extraction algorithm.
The bias and dispersion uncertainty contributions on the stability metrics were evaluated in Sec. 5.1.3. The data presented in Tab. 5.6, highlights a negative bias for every stability metrics except for vf_o and *vAm*, unlike the case of the generation bias, which presented positive biases for every metrics except for vf_o and *vAm*. Regarding the dispersion evaluations, summarised in Tab. 5.7, the obtained results are comparable to the generation dispersion evaluated in Sec. 5.1.1, therefore the extraction algorithm contribution to the data dispersion can be considered as negligible.
To evaluate the cepstral characteristics of the generation method, the CPPS metrics of the artificial vowels were evaluated by the extraction algorithm and compared to the metrics extracted from the original vowels. The bias evaluations reported in Tab. 5.8, highlighted both negative and positive biases for all the metrics. As reported by the comparison between the artificial and the original metrics, the dispersion of the artificial metrics is lower than the dispersion of the original ones. In conclusion,

the re-synthesis method alters just the sequences of periods and amplitudes of the original vowel leaving the cepstral characteristics unchanged.

7.3.3 The acquisition device negligibly affects the voice features

Using the architecture depicted in Fig. 5.13, the evaluation of the acquisition device contribution to the features uncertainty was carried out on a consumer level, portable audio recorder, connected as in the diagram of Fig. 5.14. The effects of non ideality of the chain were analysed in Sec. 5.4.1, where the offset and gain errors were estimated using two different evaluation methods. The results of this analysis highlighted the difficulties in evaluating the gain and offset errors because of the non linearity of the acquisition and evaluation chain (DAC+ADC), therefore the compensation of these uncertainty contributions has not been performed for this work.
The uncertainties reported in Tab. 5.20, represent the acquisition and extraction (ACQ+EXT) uncertainty contributions to pseudo-period and the amplitude measurements. The values in Tab. 5.20 are higher than the uncertainties of the extraction contribution, even though they are comparable. As expected, the non ideality of the DAC+ADC chain negligibly affects the period stability metrics, but it has an important effect on the amplitude metrics as summarised in Tab. 5.21 and 5.22.
The CPPS metrics, extracted from the artificial vowel acquired with the acquisition device, were evaluated as summarised in Tab. 5.23 ad Tab. 5.24. The results showed negligible differences between the ACQ+EXT and the EXT contributions to CPPS metrics uncertainty in terms of bias and dispersion, therefore the ACQ contribution can be considered as negligible.

7.3.4 The whole measuring chain affects the voice features

In order to evaluate the acoustic domain contributions to the features uncertainty, a human simulator (Head And Torso Simulator, HATS) was used to produce acoustic waves using as a reference the artificial vowels, as can be noticed in the architecture in Fig. 5.20. The substitution of the human subject with the simulator represent the final link to evaluate and separate the human contribution from the machine contribution to features uncertainty. This is possible thanks to the fact that using the HATS, the experimental conditions of repeatablity and reproducibility are met,

therefore repeated measurements of the artificial vowel features were carried out to evaluate the effect of perturbations in the acoustic domain. In particular four microphone positions of the Cheek Microphone (CM) were tested along an iPhoneX and a reference microphone. As summarised in the Tables from 5.25 to 5.49, the dispersion contribution of the whole chain (ACO+ACQ+EXT) is comparable to the dispersion of the previous evaluations (ACQ+EXT and EXT). An important effect can be noticed in the bias of the amplitude stability metrics. This consideration is confirmed by the data presented in Tab. 5.27, where the amplitude uncertainties $u(A)$ raised by an order of magnitude respect to the ACQ+EXT contribution, while the period uncertainty $u(T)$ has raised slightly.
The effect of the whole chain on the CPPS features was evaluated and negligible differences can be noticed between the dispersions and biases of the various experimental setups. This is true except for the data relative to the recordings with the SmartPhone (SP) where large biases were evaluated on the CPPS metrics as shown in Fig. 5.31.

7.4 Chapter 6

In this chapter, the conclusions for the evaluations carried out in Chapter 6 will be presented. The uncertainty evaluations presented in the previous chapters were used to train weighted logistic regression models and to define a confidence-risk framework for the predicted probabilities of these models.

7.4.1 Machine learning: a metrologic approach to the logistic regression

The definition of a metrologic framework to evaluate the effect of features and model uncertainty was described in Sec. 6.2. The analytical uncertainty evaluation of predicted probabilities was carried out considering two approaches:

- Negligible correlation (the mixed terms of the uncertainty propagation are considered as negligible with respect to the squared terms) Eq. 6.15
- The correlation is evaluated Eq. 6.17.

The uncertainty evaluations described in Chapter 5 were used to evaluate the features bias in order to be able to remove it from the original data. In particular, the bias contribution was removed or not in order to evaluate its effect on the classification metrics. The uncertainty of the features was used to define prior weights (Eq. 6.9) so that the set features evaluated with a low mean relative uncertainty are considered more by the proposed weighted logistic regression, whose cost function is defined in Eq. 6.8. The analytical propagation of the predicted probabilities has made possible a definition of a confidence interval around these values. Such a confidence interval may intersect or not the decisional threshold (0.5), therefore a third class of *non-classified* has been proposed to define new classification metrics:

- Pessimistic Accuracy
- Realistic Accuracy
- Optimistic Accuracy
- Fraction of classified

Using these metrics, training and validation experiments were carried out to evaluate the effect of different data processing techniques and different evaluation approaches.

7.4.2 Training experiments: removing the non-classified subjects improves the classification accuracy

The plots in Figs. 6.6 and 6.7, report the classification metrics, proposed in Sec. 6.3.1, of the models trained using different number of features and adopting the uncertainty evaluation strategy described in Sec. 6.2.2 and Sec 6.2.3. As can be noticed from the plots, the realistic accuracy $Acc_{realistic}$ is always higher than the accuracy Acc_{PS} of the proposed selection method (PS) and almost always higher than the accuracy Acc_{CS} of the common selection method (CS). This means that removing the non classified improves the classification accuracy ($Acc_{realistic}$>Acc_{PS}), because the non-classified predictions have a number of false predictions greater than the number of true predictions ($FP_{NC} + FN_{NC} > TP_{NC} + TN_{NC}$). The second consideration that can be done is that the proposed method showed almost always better classification metrics than the common feature and model selection, therefore the proposed method showed overall better performances than a common method.

7.4.3 Training experiments: removing the bias has a negligible effect on classification metrics

To evaluate the effect of bias on the classification prediction, the training of the models was performed without the mixed terms evaluation (approach 1, negligible correlation). These models were trained processing the input data in order to remove or not the bias evaluated in Sec. 5.1.3 and Sec. 5.1.4. As can be noted in the plot if Fig . 6.13, the bias removal process do not improve significatively the accuracy metrics of the models trained with the PD vs. HE subset if the mixed terms of the uncertainty are considered as negligible. The same considerations can be done looking at the plots in Fig. 6.6, where even for the accuracy metrics obtained with the mixed terms evaluation no significative difference can be noticed, even though slightly higher accuracy metrics are showed for 5 and 6 features models (Fig. 6.6 (d) (e)). The same consideration can be made for the PD vs. PA classification, as can be noted from the plots in Fig. 6.7. Regarding the accuracy metrics of the models trained with the artificial data slight improvements of the accuracy metrics can be noticed, especially for the PD vs. PA classification, as shown in Fig. 6.8.

7.4.4 Training experiments: evaluating the mixed terms improves the classification metrics

The effect of considering the mixed terms of the features uncertainty was evaluated by means of a comparison with the accuracy metrics of the models trained without the mixed terms evaluation, as shown in Sec. 6.8.2. As shown in Fig. 6.14, training the classification models using the uncertainty evaluation strategy described in Sec. 6.2.3 produces higher accuracy metrics than the approach described in 6.2.2. This happens because the mixed terms on Eq. 6.20 may be negative, so the contribution of the quadratic terms of the uncertainty propagation equation is limited, thus the confidence interval size of the predicted probabilities is reduced. This fact leads to an increased fraction of classified and, consequently, a higher pessimistic accuracy.

7.4.5 Training experiments: the artificial data can be used as a boosting technique

The recordings of the vowels used to perform the analysis presented in this manuscript were carried out between November 2018 and and October 2019. Unfortunately, the author and the research group realized soon that the collected data were unbalanced in terms of subjects ages. In spring 2020 a new data collection campaign should have took place but, due to the restrictions to non-medical workforce, such a campaign never took place and the restriction lasted until the time this manuscript is being written. In order to solve the problem with the scarcity of data, the idea of generating artificial vowels, to perform the analysis presented in this manuscript, came to the author's mind in April 2020. The features extracted from the artificial vowels were used to train weighted logistic regression models to obtain the classification metrics, reported in Sec. 6.5.2 and briefly discussed in Sec. 6.8.3, as depicted in Fig. 6.8 and 6.9. As already stated, the effect of bias removal and mixed term evaluation are visible on the plots and, in particular, the pessimistic accuracy seems to improve if the bias is removed and the mixed terms are evaluated. The realistic and the weighted accuracy reached values up to 100 %, as well as the accuracy of the common selection method. It is important to note that these values are the accuracies of the unvalidated models, as discussed in Sec. 6.8.5. An analysis of the 6 feature model for the PD vs. HE classification highlighted very high β_i coefficients with high relative uncertainties, as reported in Tab. 6.11. This means that the trained model may produce a large separation between the classes in the features hyperplane. For this reason, the models with 5 and 6 features can not be considered as reliable because of their high *epistemic* uncertainty [42], which is caused by the difficulties in defining a reasonably small confidence interval for the decision threshold curves in the features hyperspace

7.4.6 Training experiments: the length of the measuring chain affects the performance of the classification algorithms

An evaluation of the effect of the measuring chain length was carried out in Sec. 6.6 and briefly discussed in Sec. 6.8.4. Comparable stability metrics were found between the long measuring chain (ACO+ACQ+EXT) and the short one (EXT), as shown in Fig. 6.16. Despite having comparable performance, the trained models

have chose different sets of features, as reported in Tab. 6.12. This behaviour may be conceptually redirected to one of the characteristics an artificial intelligence should have: the adaptability. Such a characteristic was discussed in the Introduction of this manuscript in Sec. 1.8. In conclusion, when the environmental set-up of the training experiment is perturbed, the selected features used to evaluate a decision may change to better adapt to the new conditions.

7.4.7 Validation experiments: the classification metrics are lower if an unbalanced dataset is used

Due to the difficulties on accessing in public health structures due to the Covid-19 emergency, the validation dataset could not be balanced in terms of subjects age as in the case of the training subset. The validation dataset is composed by older PD subjects and younger HE and PA subjects with respect to the training dataset. This analysis was carried out in Sec. 6.7 and briefly discussed in Sec. 6.8.5. As can be noticed in Fig. 6.17 and Fig. 6.18, the validation of the trained models for the PD vs. HE subset highlighted reduced accuracy metrics for the models trained with the original data and for the data-boosted models. The models trained with the PD vs. PA subset, instead, are slightly more balanced in terms of subjects age and with respect to the mean ages of the training dataset ($\approx$ 52 years). This consideration highlights the importance of having balanced datasets when performing training and validation experiments.
Comparing the validated realistic accuracies of the PD vs. HE classification models (maximum 77 %) with the accuracies of the validated models found in the literature ($\approx$ 95 % for [7] and $\approx$ 81 % for [8]), the models proposed in this work present lower values of accuracy. Anyway the author's intent has never been to reach higher accuracies with respect to the existent literature, but to produce predictions that can be inserted in a confidence-risk framework. In conclusion: the author never wanted to be the best, just the more honest.

7.5 Final Conclusions: a conceptual link to the introduction of this manuscript

Safety

Uttering some vowels on a microphone can be considered a safe activity if it is done for a reasonable short amount of time. Substituting the human subject with an artificial one (HATS) allow the experimenters to perform thousand of repeated measurements of one vowel without involving the subject that has produced it.

Repeatability

Considering the conclusions stated in Sec. 7.3.4 and the ones stated in Sec. 7.4.6, the uncertainty contributions of the measuring chain were characterized, and their effects on the classification algorithm performance were evaluated. According to these evaluations, the whole measuring chain, which begins with the microphone and ends with a predicted probability, can be considered repeatable. This was possible tanks to the substitution of the original subject with the artificial subject.

Trustability

The definition of a confidence-risk framework, described in Sec. 6.3.1, allow to produce clinical predictions along with their respective confidence intervals. This feature allows the patient and his doctor to have an evaluation on the trust that can be given to each single prediction. Moreover, the definition of the accuracy metrics described in Sec. 6.3.1, allows to have an evaluation of the trust that can be given to the classification algorithm, defining *pseudo-confidence intervals* for the accuracy metrics. These intervals are defined in a range between the pessimistic and the optimistic accuracy, therefore they give an information on the distance between *worst case accuracy* and the *best case accuracy*.

Traceability

As already stated in the conclusions of Sec. 7.3.4 and Sec. 6.8.4, the entire measurement chain was made traceable, starting from the acoustic domain up to the

prediction domain. This was possible tanks to the substitution of the original subject with the artificial subject.

Accountability

Substituting the human subject with the artificial one removes the human dispersion contribution from the uncertainty budget, therefore the responsibility of each prediction always lies with the artificial intelligence and its artificial body (the measuring chain).

Adaptability

As exemplified in Sec. 1.8, the informativeness of the environmental features may be altered if the experimental conditions change. As already stated in the conclusions of Sec.6.8.4, when a perturbation of the measuring chain is introduced, the proposed model and features selection method choose different sets of features to better adapt the models to the uncertainty contributions of the features.

References

[1] Luis G. Hernández, Oscar Martinez Mozos, José M. Ferrández, and Javier M. Antelis. Eeg-based detection of braking intention under different car driving conditions. *Frontiers in Neuroinformatics*, 12, 2018.

[2] Jeferson Menegazzo and Aldo Von Wangenheim. Vehicular perception and proprioception based on inertial sensing: a systematic review. 10 2018.

[3] Heikki Summala, Dave Lamble, and Matti Laakso. Driving experience and perception of the lead car's braking when looking at in-car targets. *Accident Analysis and Prevention*, 30(4):401–407, 1998.

[4] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The unified parkinson's disease rating scale (updrs): Status and recommendations. *Movement Disorders*, 18(7):738–750, 2003.

[5] Christopher G. Goetz, Werner Poewe, Olivier Rascol, Cristina Sampaio, Glenn T. Stebbins, Carl Counsell, Nir Giladi, Robert G. Holloway, Charity G. Moore, Gregor K. Wenning, Melvin D. Yahr, and Lisa Seidl. Movement disorder society task force report on the hoehn and yahr staging scale: Status and recommendations the movement disorder society task force on rating scales for parkinson's disease. *Movement Disorders*, 19(9):1020–1028, 2004.

[6] Antonio Suppa, Giovanni Costantini, Francesco Asci, Pietro Di Leo, Mohammad Sami Al-Wardat, Giulia Di Lazzaro, Simona Scalise, Antonio Pisani, and Giovanni Saggio. Voice in parkinson's disease: A machine learning study. *Frontiers in Neurology*, 13, 2022.

[7] Far D. Shahbakhi, M. and Tahami. E. (2014) speech analysis for diagnosis of parkinson's disease using genetic algorithm and support vector machine.

[8] Lucijano Berus, Simon Klancnik, Miran Brezocnik, and Mirko Ficko. Classifying parkinson's disease based on acoustic measures using artificial neural networks. *Sensors*, 19(1), 2019.

[9] Rekha Viswanathan, Sridhar P. Arjunan, Adrian Bingham, Beth Jelfs, Peter Kempster, Sanjay Raghav, and Dinesh K. Kumar. Complexity measures of voice recordings as a discriminative tool for parkinson's disease. *Biosensors*, 10(1), 2020.

[10] Ömer Eskıdere, Ali Karatutlu, and Cevat Ünal. Detection of parkinson's disease from vocal features using random subspace classifier ensemble. In *2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)*, pages 1–4, 2015.

[11] Max A. Little *, Patrick E. McSharry, Eric J. Hunter, Jennifer Spielman, and Lorraine O. Ramig. Suitability of dysphonia measurements for telemonitoring of parkinson's disease. *IEEE Transactions on Biomedical Engineering*, 56(4):1015–1022, 2009.

[12] Michal Šimek and Jan Rusz. Validation of cepstral peak prominence in assessing early voice changes of parkinson's disease: Effect of speaking task and ambient noise. *The Journal of the Acoustical Society of America*, 150(6):4522–4533, 2021.

[13] Alessio Atzori, Alessio Carullo, Alberto Vallan, Viviana Cennamo, and Arianna Astolfi. Parkinson disease voice features for rehabilitation therapy and screening purposes. In *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pages 1–6, 2019.

[14] Shalini Narayana, Crystal Franklin, Elizabeth Peterson, Eric J. Hunter, Donald A. Robin, Angela Halpern, Jennifer Spielman, Peter T. Fox, and Lorraine O. Ramig. Immediate and long-term effects of speech treatment targets and intensive dosage on parkinson's disease dysphonia and the speech motor network: Randomized controlled trial. *Human Brain Mapping*, 43(7):2328–2347, 2022.

[15] Antonella Castellana, Alessio Carullo, Simone Corbellini, and Arianna Astolfi. Discriminating pathological voice from healthy voice using cepstral peak prominence smoothed distribution in sustained vowel. *IEEE Transactions on Instrumentation and Measurement*, 67(3):646–654, 2018.

[16] Christina Batthyany, Youri Maryn, Ilse Trauwaen, Els Caelenberghe, Joost van Dinther, Andrzej Zarowski, and Floris Wuyts. A case of specificity: How does the acoustic voice quality index perform in normophonic subjects? *Applied Sciences*, 9(12), 2019.

[17] Edson Cataldo, Rubens Sampaio, Jorge Lucero, and Christian Soize. Modeling random uncertainties in voice production using a parametric approach. *Mechanics Research Communications*, 35(7):454–459, 2008.

[18] E. Cataldo, C. Soize, and R. Sampaio. Uncertainty quantification of voice signal production mechanical model and experimental updating. *Mechanical Systems and Signal Processing*, 40(2):718–726, 2013.

[19] Y. Bennane, A. Kacha, J. Schoentgen, and F. Grenez. Synthesis of pathological voices and experiments on the effect of jitter and shimmer in voice quality perception. In *2017 5th International Conference on Electrical Engineering - Boumerdes (ICEE-B)*, pages 1–6, 2017.

[20] Paul Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17, 01 2000.

[21] Pasquale Bottalico, Juliana Codino, Lady Catherine Cantor-Cutiva, Katherine Marks, Charles J. Nudelman, Jean Skeffington, Rahul Shrivastav, Maria Cristina Jackson-Menaldi, Eric J. Hunter, and Adam D. Rubin. Reproducibility of voice parameters: The effect of room acoustics and microphones. *Journal of Voice*, 34(3):320–334, 2022/08/11 2020.

[22] Mark D. Skowronski, Rahul Shrivastav, and Eric J. Hunter. Cepstral peak sensitivity: A theoretic analysis and comparison of several implementations. *Journal of Voice*, 29(6):670–681, 2022/08/11 2015.

[23] Castellana Antonella, Carullo Alessio, Corbellini Simone, Astolfi Arianna, Spadola Massimo, and Colombini J. Cepstral peak prominence smoothed distribution as discriminator of vocal health in sustained vowel. 05 2017.

[24] James Hillenbrand and Robert Houde. Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of speech and hearing research*, 39:311–21, 04 1996.

[25] Joint Committee for Guides in Metrology (JCGM) (BIPM). Jcgm 100:2008 - evaluation of measurement data — guide to the expression of uncertainty in measurement, 2008.

[26] https://it.mathworks.com/help/signal/ref/resample.html. Matlab function *resample* reference.

[27] https://it.mathworks.com/help/signal/ref/uencode.html. Matlab function *uencode* reference.

[28] https://it.mathworks.com/help/comm/ref/wgn.html. Matlab function *wgn* reference.

[29] Alessio Atzori, Alessio Carullo, Simone Corbellini, and Alberto Vallan. Crosstalk effects in the uncertainty estimation of multiplexed data acquisition systems. *IEEE Transactions on Instrumentation and Measurement*, 70:1–11, 2021.

[30] https://it.mathworks.com/help/matlab/ref/interp1.html. Matlab function *interp1* reference.

[31] https://it.mathworks.com/help/matlab/ref/linspace.html. Matlab function *linspace* reference.

[32] https://it.mathworks.com/help/matlab/ref/vertcat.html. Matlab function *vertcat* reference.

[33] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 04 1970.

[34] Frank J. Massey. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, 46(253):68–78, 1951.

[35] https://it.mathworks.com/help/matlab/ref/histcounts.html. Matlab function *histcounts* reference.

[36] David Freedman and Persi Diaconis. On the histogram as a density estimator: L 2 theory. 1981.

[37] https://it.mathworks.com/help/stats/ecdf.html. Matlab function *ecdf* reference.

[38] https://it.mathworks.com/help/curvefit/smooth.html. Matlab function *smooth* reference.

[39] https://it.mathworks.com/help/stats/kstest2.html. Matlab function *kstest2* reference.

[40] Ergonomics — Assessment of speech communication, 2003.

[41] FP Agterberg, GF Bonham-Carter, Qiu-min Cheng, and DF Wright. Weights of evidence modeling and weighted logistic regression for mineral potential mapping. *Computers in geology*, 25:13–32, 1993.

[42] Eyke Hüllermeier and Willem Waegeman. Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. *Machine Learning*, 110(3):457–506, 2021.

Appendix A

Features equations

In this Appendix a list of the features used for this work is presented. The identificative numbers define the feature numbers used in the Tables of Chapter 6.

1. Local jitter:

$$jit = \frac{N}{N-1} \cdot \frac{\sum_{i=1}^{N-1} |T_i - T_{i+1}|}{\sum_{i=1}^{N} T_i} \cdot 100 \ \ (\%) \tag{A.1}$$

2. Absolute jitter:

$$jit_{abs} = \frac{10^6}{N-1} \cdot \sum_{i=1}^{N-1} |T_i - T_{i+1}| \ \ (\mu\text{s}) \tag{A.2}$$

3. Relative Average Perturbation (RAP3):

$$rap = \frac{N}{N-2} \cdot \frac{\sum_{i=1}^{N-2} |T_i - (\frac{1}{3} \sum_{r=i-1}^{r=i+1} T_r)|}{\sum_{i=1}^{N} T_i} \cdot 100 \ \ (\%) \tag{A.3}$$

4. Pitch Period Perturbation Quotient (PPQ5):

$$ppq = \frac{N}{N-4} \cdot \frac{\sum_{i=2}^{N-4} |T_i - (\frac{1}{5} \sum_{r=i-2}^{r=i+2} T_r)|}{\sum_{i=1}^{N} T_i} \cdot 100 \ \ (\%) \tag{A.4}$$

5. Coefficient of Fundamental frequency variation:

$$vf_o = \frac{\sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (f_i - \overline{f})^2}}{\overline{f}} \cdot 100 \ \ (\%) \ \ ; \ \ f_i = \frac{1}{T_i} \ \ ; \ \ \overline{f} = \frac{1}{N} \sum_{1}^{N} f_i \tag{A.5}$$

6. Local shimmer:

$$shi = \frac{N}{N-1} \cdot \frac{\sum_{i=1}^{N-1} |A_i - A_{i+1}|}{\sum_{i=1}^{N-1} A_i} \cdot 100 \ \ (\%) \tag{A.6}$$

7. Absolute shimmer:

$$shi_{abs} = \frac{1}{N-1} \cdot \sum_{i=1}^{N-1} |20 \cdot log_{10} \frac{A_{i+1}}{A_i}| \ \ (\text{dB}) \tag{A.7}$$

8. Amplitude Perturbation Quotient (APQ11):

$$apq = \frac{N}{N-10} \cdot \frac{\sum_{i=5}^{N-10} |T_i - (\frac{1}{11} \sum_{r=i-5}^{r=i+5} T_r)|}{\sum_{i=1}^{N} T_i} \cdot 100 \ \ (\%) \tag{A.8}$$

9. Coefficient of Amplitude variation:

$$vAm = \frac{\sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (A_i - \overline{A})^2}}{\overline{A}} \cdot 100 \ \ (\%) \ ; \ \overline{A} = \frac{1}{N} \sum_{1}^{N} A_i \tag{A.9}$$

Harmonics to noise ratio:

$$HNR = 10 \cdot Log_{10}(\frac{A_c(T)/A_c(0)}{[1 - A_c(T)]/A_c(0)}) \ \ (\text{dB}) \tag{A.10}$$

10. Mean HNR (dB)
11. Median HNR (dB)
12. Mode HNR (dB)
13. HNR Range (dB)
14. HNR Standard deviation (dB)
15. HNR 5° percentile (dB)
16. HNR 95° percentile (dB)
17. HNR Skewness (a.u.)
18. HNR Kurtosis (a.u.)

Fundamental Frequency:

$$f_o = 1/T \ \ (\text{Hz}) \tag{A.19}$$

19. Mean f_o (Hz)
20. Median f_o (Hz)
21. Mode f_o (Hz)
22. f_o Range (Hz)
23. f_o Standard deviation (Hz)
24. f_o 5° percentile (Hz)
25. f_o 95° percentile (Hz)
26. f_o Skewness (a.u.)
27. f_o Kurtosis (a.u.)

Amplitude Root Mean Square Value:

$$A_{RMS} = \sqrt{(\sum_{n=1}^{N} s_n^2)/N} \ \ (\text{a.u.}) \tag{A.28}$$

28. Mean A_{RMS} (a.u.)
29. Median A_{RMS} (a.u.)
30. Mode A_{RMS} (a.u.)
31. A_{RMS} Range (a.u.)
32. A_{RMS} Standard deviation (a.u.)
33. A_{RMS} 5° percentile (a.u.)
34. A_{RMS} 95° percentile (a.u.)
35. A_{RMS} Skewness (a.u.)
36. A_{RMS} Kurtosis (a.u.)

Cepstral Peak Prominence Smoothed (CPPS):
see Sec. 2.4.4 for the algorithm details.

37. Mean *CPPS* (dB)
38. Median *CPPS* (dB)
39. Mode *CPPS* (dB)
40. *CPPS* Range (dB)
41. *CPPS* Standard deviation (dB)
42. *CPPS* 5° percentile (dB)
43. *CPPS* 95° percentile (dB)
44. *CPPS* Skewness (a.u.)
45. *CPPS* Kurtosis (a.u.)
46. Voiced/unvoiced: $\frac{N_{periods}(HNR>0)}{N_{periods}(HNR\leq 0)} \cdot 100$ (%)