Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

BIOINGEGNERIA

Ciclo XXVIII

Settore Concorsuale di afferenza: 09/G2 - BIOINGEGNERIA

Settore Scientifico Disciplinare: ING-INF/06 - BIOINGEGNERIA ELETTRONICA E INFORMATICA

DEVELOPMENT OF MARKERLESS SYSTEMS FOR AUTOMATIC ANALYSIS OF MOVEMENTS AND FACIAL EXPRESSIONS: APPLICATIONS IN NEUROPHYSIOLOGY

Presentata da: Ing. Andrea Bandini

Coordinatore Dottorato
Prof.ssa Elisa Magosso

Relatore
Prof. ssa Claudia Manfredi

Correlatore
Prof. Andrea Corvi

Controrelatori
Prof. Ugo Della Croce
Dr. Hugo Jair Escalante

Esame finale anno 2016

“Everyone must leave something in the room or left behind when he dies, my grandfather said. A child or a book or a painting or a house or a wall built or a pair of shoes made. Or a garden planted. Something your hand touched some way so your soul has somewhere to go when you die, and when people look at that tree or that flower you planted, you're there. It doesn't matter what you do, he said, so long as you change something from the way it was before you touched it into something that's like you after you take your hands away. The difference between the man who just cuts lawns and a real gardener is in the touching, he said. The lawn-cutter might just as well not have been there at all; the gardener will be there a lifetime.”

- Ray Bradbury, Fahrenheit 451

Contents

Abstract

This PhD project is focused on the development of markerless methods for studying facial expressions and movements in neurology, focusing on Parkinson's disease (PD) and disorders of consciousness (DOC).

Parkinson's disease is a neurodegenerative illness that affects around 2% of the world population over 65 years old. Impairments of voice and speech are among the main signs of the disease and affect high percentages of Parkinsonian patients. These patients may present alterations related to all speech dimensions (i.e., the four subsystems of speech production): respiration, phonation, articulation and prosody. This set of impairments is known as hypokinetic dysarthria, because of the reduced range of movements involved in speech production. This reduction can be visible also in other facial muscles, leading to a hypomimia typical of PD patients. Both impairments lead PD patients to have difficulties in social relationships, with a tendency to isolate themselves.

Despite the high percentage of patients that suffer from dysarthria and hypomimia, only a few of them undergo speech therapy and rehabilitation processes with the aim to improve the dynamic of articulatory and facial movements. The main reason is the lack of low cost methodologies that could be implemented at home in order to perform therapeutic exercises during the daily activities.

Disorders of consciousness after coma in severe brain injury are Vegetative State (VS), characterized by the absence of self-awareness and awareness of the environment, and Minimally Conscious State (MCS), in which certain behaviors, although functionally inconsistent, are sufficiently reproducible to be distinguished from simple reflex responses. The interest in these conditions has gradually spread in the medical and scientific community in relation to the relevant bioethical theme of end of life decision-making processes.

The differential diagnosis between VS and MCS can be extremely hard and prone to a high rate of misdiagnosis (around 40%). This differential diagnosis is currently based on neuro-behavioral scales, instrumental examinations and functional neuroimaging techniques, although the latter techniques are limited to the research field due to the high costs. A key role to plan therapy and rehabilitation treatments for DOC patients is played by the first diagnosis after the end of coma. In fact, MCS patients who undergo a well-defined rehabilitation treatment are more prone to a better consciousness recovery than VS patients.

Concerning PD patients the aim is the development of a contactless system that could be implemented to study symptoms related to speech and facial movements/expressions. The method proposed here, based on acoustical analysis and video processing techniques, would allow tracking the disease progression and providing support to patients during speech therapy also at home. In case of DOC patients the project was focused on the assessment of reflex and cognitive responses to standardized stimuli. This would allow objectifying the perceptual analysis performed by clinicians, developing a

markerless system for monitoring DOC patients and assessing reactions related to the perception of pain in patients that are unable to communicate.

To this end, the following topics are covered in this dissertation: speech processing for biomedical applications, video-based analysis of articulatory movements, facial expression recognition and heart rate estimation from videos.

Summary

The PhD project presents new and effective methods for the analysis of movements, facial expressions and voice in two neurological disorders: Parkinson's disease and disorders of consciousness, i.e. vegetative state and minimally conscious state.

Background

Parkinson's disease

Alterations of voluntary and involuntary movements of the human being are clinical features common to many neurodegenerative diseases. Considering the constant increase in life expectancy of the world population and the high rate of incidence of neurological disorders, the automation of the diagnostic routines and of the rehabilitation processes is highly desirable. One of the most significant cases is Parkinson's disease (PD), a neurodegenerative illness that affects around 2% of the world population over 65 years old. Idiopathic Parkinson's disease involves the loss of neurons in the zona compacta of the *substantia nigra* of the midbrain and other pigmented nuclei. This loss of neurons leads to an insufficient synthesis and production of dopamine, an essential neurotransmitter for the control of body movements.

Impairments of voice and speech are among the main signs of the disease. They affect about 70% of PD patients. PD patients may present alterations related to all speech dimensions (i.e., the four subsystems of speech production): respiration, phonation, articulation and prosody. Therefore, they might exhibit low intensity and fading during speech, vocal tremor, monotone pitch, hesitations in starting to speak, difficulties in pronunciation and articulation, speech acceleration and loss of timing control and nasalization. These alterations result in a reduced intelligibility of these patients that may cause a tendency to a social isolation. The severity of speech and voice impairments increases with the disease progression.

This set of speech and voice impairments in PD patients is usually known as "dysarthria". This term refers to a group of speech disorders caused by an alteration of the muscular control of the pneumo-phono-articulatory organs due to damage in the basic motor processes involved in speech production. In most of PD patients, dysarthria is usually "hypokinetic", because of the reduced range of movements involved in speech production. In fact, hypokinetic dysarthria is characterized by reduced peak velocities and displacements of the articulators (i.e., jaw, lips and tongue) during speech movements.

The other motor sign considered in this PhD project is facial bradykinesia consisting in the reduction and slowness of facial movements. Facial bradykinesia may affect both the upper and the lower face. PD patients exhibit a blink rate reduction and reduced facial expressions (both spontaneous and posed), as well as impaired voluntary orofacial movements (as reported above for hypokinetic dysarthria). Face is an important means for conveying information about the emotional state of a

person. Facial bradykinesia is reflected in a reduction or loss of spontaneous facial movements and emotional facial expressions called hypomimia. The impairment of spontaneous facial expressions is consistent with the neuroanatomic evidence of a damage to the extrapyramidal motor system. Moreover, PD patients may experience difficulties in performing voluntary facial expressions and facial movements.

The main signs of hypomimia in PD are: wider palpebral distances (that in conjunction with the reduced blink rate gives the impression of a staring expression), flattened nasolabial folds and unintentional lips separation (mouth opening). PD face is often a "masked" or "poker face", typical of people without any interest in the surrounding environment. External observers and in particular practitioners, may have difficulties in decoding the emotional state behind the "mask". This inaccurate impression could lead to problems with social relationships and interactions in PD patients with higher masking.

One of the most altered facial expressions in Parkinsonian patients is smile. When PD patients experience spontaneous happiness, their smiles are perceived as fake and insensitive, because of the lack of cheeks raising and due to a loss of tone of the orbicularis oculi muscles. This reduction in smiling could also be related to depression. Today a big debate in the research community is focused on the nature of hypomimia in PD patients. In fact, many works support the hypothesis that the loss of facial expressions could be due to an impairment in facial emotion recognition, with main deficits in the decoding of disgust, fear and sadness.

Disorders of consciousness

Disorders of consciousness (DOC) after coma in severe brain injury are Vegetative State (VS), characterized by the absence of self-awareness and awareness of the environment, and Minimally Conscious State (MCS), in which certain behaviors, although functionally inconsistent, are sufficiently reproducible to be distinguished from simple reflex responses. The interest in these conditions has gradually spread in the medical and scientific community in relation to the relevant bioethical theme of end of life decision-making processes.

Consciousness is a multifactorial concept composed by two main elements: self-awareness and awareness of the surrounding environment ("awareness" for simplicity) and "wakefulness". In neurology, coma (also defined as "unarousable unresponsiveness state") is a deep unconsciousness state, characterized by the absence of both aspects of consciousness (awareness and wakefulness), and caused by the inactivation of those neuronal circuits responsible of maintaining a waking state. These circuits are in the area of the brainstem called reticular activating system (RAS). In most cases, coma is a transient condition that can last for a few weeks, ending when patients open the eyes, recovering the wakefulness. Although the re-opening of the eyes is sufficient to establish the end of this first acute phase, it is not enough for determining what the clinical outcome of patients is. In fact, it is necessary to understand whether patients recover the awareness and at what level. According to these criteria,

the main post-comatose outcomes can be: vegetative state, minimally conscious state, locked-in syndrome, brain death and recovery.

VS is a syndrome characterized by the lack of awareness, and the lack of intentional and communicative responses. VS is also called Unresponsive Wakeful Syndrome (UWS), since VS patients maintain a sleep-wake rhythm, they can smile, crying, grimacing and emit sounds, but without certain reasons. MCS, instead, is characterized by a partial recover of the awareness. There may be some cognitive-mediated behaviors, inconsistent in functional terms, but maintained long enough to be distinguished from simple reflex responses. These patients may experience smiles or other responses after the administration of external (audio-visual, emotional and linguistic stimuli), suggesting their ability to interact with the surrounding environment. MCS is often considered as an intermediate state between VS and the presence of consciousness.

The differential diagnosis between VS and MCS can be extremely hard and prone to a high rate of misdiagnosis (around 40%). This differential diagnosis is currently based on neuro-behavioral scales (Coma Recovery Scale Revised - CRS-R, Glasgow Coma Scale - GCS, etc.), instrumental examinations (EEG, evoked potentials) and functional neuroimaging techniques, although the latter techniques are limited to the research field due to the high costs. A key role to plan therapy and rehabilitation treatments for DOC patients is played by the first diagnosis after the end of coma. In fact, MCS patients who undergo a well-defined rehabilitation treatment are more prone to a better consciousness recovery than VS patients. Despite the number patients with severe brain injury and low consciousness levels is negligible if compared to other neurological diseases (between 56 and 70 cases per million), the impact on the economy can be very important.

Aims

In case of PD patients, the aim was to extract acoustical and kinematic parameters related to speech disorders and facial expression impairments. This allows:

- Objectifying the perceptual analysis performed by clinicians;
- Developing a markerless system to monitor symptoms related to hypokinetic dysarthria and hypomimia;
- Developing a markerless system for speech therapy that can be used in a domestic environment.

Thus, by using acoustical analysis and video processing techniques the aim was to develop a contactless system that could be implemented to track the disease progression (in particular for studying symptoms related to speech and facial movements/expressions) and for aiding patients during the speech therapy process also at home. In fact, despite speech disorders concern a quite large population of PD patients, speech therapy is applied to a small part of them.

In case of DOC patients the project was focused on the assessment of reflex and cognitive responses to standardized stimuli administered according to the CRS-R protocol. This allows:

- Objectifying the perceptual analysis performed by clinicians;

- Developing a markerless system for monitoring DOC patients;
- Assessing reactions related to the perception of pain in patients unable to communicate.

The proposed system merges two of the methods developed in this PhD projects: the analysis of facial expressions (also used for studying hypomimia in PD patients) and the contactless estimation of the heart rate (HR). This system could be used as an aid to the monitoring of these patients, even when there are no caregivers or family members near them.

Both case studies (PD and DOC patients) share the study of facial movements and facial expressions by means of video based techniques.

Methods

After a careful literature review of the main topics of the project (speech disorder, facial expressions impairments in PD patients and awareness evaluation in DOC patients), the following collaborations were established:

- Unit of Neurology - Hospital "Nuovo San Giovanni di Dio" (Firenze) and *Associazione Italiana Parkinsoniani* (Firenze), for the recruiting of PD patients and of healthy control (HC) subjects;
- "Villa delle Terme" rehabilitation center (Impruneta, Firenze) and "Don Carlo Gnocchi" Foundation IRCSS (Scandicci, Firenze) for the recruiting of DOC patients.

Two audio/video acquisition protocols for PD and DOC patients were defined. PD patients were evaluated during the repetition of sentences, syllables and during the displaying of acted and imitated facial expressions (neutral, anger, disgust, happy and sad). DOC patients were video-recorded during the CRS-R protocol, were standardized stimuli of different nature (acoustic, visual, tactile, etc.) are administered for assessing patients' reactions. During the PhD project the following subjects were enrolled:

- 27 PD patients (18 male, 9 female), age: 71.6 ± 8.2 years, disease duration 8.5 ± 5.0 years, Hoehn & Yahr scale: 2.1 ± 0.4, UPDRS motor score: 16.6 ± 10.4. Audio recordings during the speech task were performed on 25 patients (16 male, 9 female). Video recordings (color and depth streams) during the speech task were performed on 16 patients (10 male, 6 female), while video acquisitions (color stream only) of facial expressions were performed on 18 patients (14 male, 4 female).
- 32 HC subjects (18 male, 14 female), age: 67.7 ± 7.8 years. Audio recordings during the speech task were performed on 23 subjects (11 male, 12 female). Video recordings (color and depth streams) during the speech task were performed on 19 patients (13 male, 6 female), while video acquisitions (color stream only) of facial expressions were performed on 17 subjects (6 male, 11 female).
- 13 DOC patients (8 male, 5 female), age: 54.9 ± 15.8 years, months after the onset: 35.3 ±27.9, 11 VS patients and 2 MCS patients, CRS-R score 6.1 ± 1.4, etiology: 8 post-anoxic, 5 traumatic brain injury. Because of the poor quality of some acquisitions, only nine (7 VS and 2 MCS) of the 13

patients were considered for the analysis of facial features and vital signs during the administration of external stimuli.

According to the aims described above, during the three years of doctoral studies the following issues were addressed: acoustical analysis in PD patients, video analysis of articulatory movements PD patients, video analysis of facial expressions in PD patients and the video analysis of facial expressions and heart rate in DOC patients.

Acoustical analysis in PD patients

20 PD patients disease were recruited at the Department of Neurology of the Hospital "San Giovanni di Dio", Firenze, Italy (14 male, 6 female; age: 72.2 ± 8.6 years; disease duration: 8.3 ± 4.6 years; Hoehn and Yahr scale: 2.1 ± 0.4; UPDRS part III: 15.7 ± 9.9; UPDRS speech item: 0-1). Moreover, a group of 19 HC subjects were tested (9 male, 10 female; age: 68.1 ± 8.4 years). Each subject was asked to repeat a standardized Italian voiced sentence ("*Il bambino ama le aiuole della mamma*") at least 10 times as spontaneously as possible at comfortable loudness without hastening the speech, therefore properly separating the consecutive sentences.

The aim of this part of the work was the development and the implementation of an automatic voiced-unvoiced (AVU) segmentation algorithm for studying dysprosody in PD patients.

The AVU algorithm splits the whole signal in short frames of the same length whose energy is evaluated and stored in an "energy vector". The Otsu's method applied to the energy histogram allows finding two thresholds for the separation between two classes (voiced and unvoiced frames).

On each sentence, the following parameters were extracted: sentence duration ($T_{sentence}$ in seconds) as the time interval between the beginning of a sentence and that of the next one; inter-sentence duration (T_{inter} in seconds) as the time interval between the end of a sentence and the beginning of the next one; pause duration (T_{pause} in seconds) as the sum of "breaks" (short pauses) inside a sentence; duty cycle (D%) as the percent of voiced time with respect to the sentence duration; net speech rate (NSR in syllables/s) defined as the number of syllables of the sentence, divided by the effective speech time ($T_{sentence} - T_{inter} - T_{pause}$).

Video analysis of articulatory movements PD patients

A markerless system for tracking lips movements during speech was developed by using a 3D structured light sensor (Primesense Carmine 1.09) and a face tracking algorithm. This system was firstly tested on 2 HC subjects, in order to compare the performances of tracking articulatory movements in 3D with respect to an optoelectronic marker-based method. For the automatic identification of the facial features, the *Intraface* tracking algorithm was used. Lips were modeled as a set of 18 points: 12 on the outer border and 6 on the inner border.

Each subject was asked to read and pronounce a corpus composed by 50 meaningful sentences, 100 meaningful words and 30 repetitions of the syllable /pa/.

The comparison with the reference trajectories was obtained computing the root-mean-square error (RMSE) in mm. This comparison was also performed on the following articulatory parameters: lips width, lips opening and lip protrusion. Afterwards, for each syllable repetition the following kinematic parameters (for both systems) were computed: the maximum velocity (V_{open}) and acceleration (A_{open}) during the opening phase, the maximum velocity (V_{close}) and acceleration (A_{close}) during the closing phase. For each syllable repetition the Pearson's correlation coefficient between trajectories, velocities and accelerations extracted with both systems was computed. Correlation values close to 1 indicate that the trends of displacement, speed and acceleration calculated with the proposed method are very similar to the ground truth.

This markerless technique was then used to analyze movements of the lower lip during a syllable repetition task, both in PD patients and HC subjects. The aim was that of testing the reliability of a cheap and markerless approach for assessing signs of hypokinetic dysarthria (in particular, alterations of peak velocities and accelerations).

14 PD patients were recruited at the Unit of Neurology of the Florence Health Authority ("San Giovanni di Dio" Hospital, Firenze, Italy), and at the *Associazione Italiana Parkinsoniani (AIP) – Sezione di Firenze*¸ Firenze, Italy (9 male, 5 female; age: 71.6 ± 7.0 years; Hoehn and Yahr scale: 2.0 ± 0.3; UPDRS part III: 16.0 ± 12.0; UPDRS speech item: 0-1). A group of 14 HC subjects with no history of neurological disease was tested (8 male, 6 female; age: 69.0 ± 7.4 years).

The speech task consisted in the repetition of the syllable /pa/ for at least 25 times within a single breath, in a comfortable steady pace.

As described above the peak values of velocity and acceleration of the lower lip during the speech task were computed. The maximum value is relative to the opening phase ($v_{opening}$ and $a_{opening}$) while the minimum value refers to the closing phase ($v_{closing}$ and $a_{closing}$). Moreover, for each repetition the following parameters were computed: normalized range of opening ($\Delta Opening_{norm}$ as difference between the maximum and the minimum opening values divided by its mean value) and the normalized maximum opening value ($MaxOpening_{norm}$ as the maximum opening value within a repetition, divided by the width of the lips).

Video analysis of facial expressions in PD patients

PD patients often experience serious difficulties in displaying both voluntary and spontaneous facial expressions. Thus, an objective evaluation of this sign is essential for the assessment of hypomimia (as an aid to the facial expressions item of UPDRS part III) and for rehabilitation, in particular for speech therapy.

17 PD patients were recruited at the Department of Neurology of the Hospital "San Giovanni di Dio", Firenze, Italy (13 male, 4 female; age: 71.9 ± 9.2 years, disease duration: 8.2 ± 5.0 years; Hoehn and Yahr scale: 2.1 ± 0.4; UPDRS part III 17.5 ± 10.3). A group of 17 HC subjects was tested as control group (6 male and 11 female; age: 68.8 ± 7.5 years).

Each subject was asked to perform the following tasks: displaying a neutral expression for at least 10 seconds; displaying basic expressions (happiness, anger, disgust and sadness) upon request of the clinician; displaying basic expressions (happiness, anger, disgust and sadness) by imitating emotive faces shown on a screen.

The subjects' face was recorded using the Microsoft Kinect for Windows sensor (using only the color stream). On the recorded videos, two tests were performed:

1. The analysis of facial features with respect to the neutral baseline, in order to find the most discriminative features between HC subjects and PD patients and giving an objective quantification of the facial hypomimia in PD;
2. The automatic facial expressions recognition, in order to study how much far the PD expressions are from the standard expressions.

Twenty geometric features of the face were extracted from the facial landmarks provided by the *Intraface* tracker: angles and raising of the eyebrows, distances of the eyes and of the mouth.

During the first test (analysis of facial features with respect to the neutral baseline), a baseline of the neutral state was built, considering an average facial template from the neutral videos. After the set-up of the neutral baseline, the analysis was performed on the expressive videos (both acted and imitated expressions). For each subject and for each expressive video the 20 facial features were extracted and compared with the neutral baseline of each subject, computing the Euclidean distance between expressive frames and neutral template. This distance provides global information about the displacement of facial features from the neutral expression, during the displaying of different facial expressions.

During the second test (automatic facial expressions recognition) the analysis of facial expressions was performed using an automatic facial expression classifier trained on different databases of posed and spontaneous expressions. However, most of the available facial expressions databases are composed by videos and images taken from healthy subjects. Starting from this consideration, the aim was to measure the intensity of facial expressions that PD patients are able to reach, with respect to HC subjects, that are assumed to express "standard" expressions. A multiclass Support Vector Machine (SVM) with radial basis function kernel was trained on the Extended Cohn-Kanade database (CK+) and the Radboud Faces database (RaFD) for a total of 3696 frames (divided into the 5 expressions considered in this experiment: neutral, anger, disgust, happiness and sadness). The accuracy of the trained classifier was computed through a 10-fold cross validation on the training set, and it was equal to 88%. Once the classifier has been trained, we used as test set the database composed by HC subjects and PD patients enrolled for the project.

Video Analysis of DOC patients: facial movements and vital signs

The aim of this part of the work was the implementation of a markerless method for monitoring and studying DOC patients, in particular after the administration of standardized stimuli. The proposed

system merges two methods: the analysis of facial expressions and the contactless estimation of the heart rate.

The aim was to extract quantitative and objective information to help clinicians in diagnostic assessment through the analysis of possible reactions after standardized stimuli in DOC patients. Moreover, this system could be used as an aid to the monitoring of these patients, especially when there are no caregivers or family members near them.

Nine DOC patients were video recorded at the "Villa delle Terme" rehabilitation center, Firenze, Italy (7 male, 2 female; age: 53.4 ± 14.2 years; months after the brain injury: 31.7 ± 27.6 months; CRS-R score: 6.2 ± 1.5). For five patients, the etiology was post-anoxic/ischemic, four patients had a cerebral hemorrhage and one patient a traumatic brain injury. DOC patients were evaluated during the administration of the CRS-R. Although all the CRS-R items were video recorded, only the motor evaluation function was considered for this study. During this item a noxious stimulus was administered to hands or feet. Our aim was the study of facial expressions and vital signs after the administration of the noxious stimulation.

During the CRS-R administration, patients' faces were recorded with a webcam. For this work we merged the analysis of facial features with respect to the neutral baseline (already used for study hypomimia in PD patients) and the contactless video-based estimation of the heart rate.

The analysis of facial features comes from the consideration that facial expressions could bring important information about patient's reactions to external stimuli and pain as they are unable to communicate. In fact, most of the perceptual ratings performed by clinicians are based on the assessment of the facial mimicry. However, a standard facial expression recognition approach (i.e., training of a classifier on basic expressions and then testing on video frames, like the approach proposed for PD patients) is not well-suited for these patients. In fact, they do not exhibit standard expressions, even if sometimes they can display smiles and grimaces. In this case, a more "patient-fitted" approach that allows extracting relative information about the evolution of the studied features during tests is best suited. Thus, for each video recording, the following processing steps were performed:

- Baseline building on video frames without stimuli administration;
- Extraction of the 20 facial features (as explained above) on the whole video recording;
- Euclidean distance calculation between the facial features vector and the baseline built for the each subject. This distance provides a global information about the displacement of facial features from the neutral state (i.e., before the administration of the external stimulus).

The HR could provide important information to evaluate patient's reactions and the level of pain. However, a continuous and constant monitoring of HR could be expensive and uncomfortable for the presence of sensors attached to the patient's body. Thus, we estimated HR through the following steps:

- Face tracking and Regions of Interest (ROIs) detection. Two ROIs were automatically located with the *Intraface* tracking algorithm: one on the forehead and one in the central facial region (including cheeks and nose).
- For both ROIs the pre-processing was applied to the temporal trends of the mean values of the R, G and B channels: detrend based on a smoothness priors approach (smoothing parameter = 100, cutoff frequency = 0.66 Hz); mean value subtraction and normalization with respect to the standard deviation.
- Estimation of the three independent sources by means of Independent Component Analysis (ICA) based on the joint approximate diagonalization of the eigenmatrices algorithm (JADE). Afterwards, a smoothing with a 5-point moving average window and a bandpass filtering between 0.6 and 3.5 Hz with a 128-points Hamming window were applied to the three sources.
- Heart rate estimation through an autoregressive (AR) model with recursive least squares (RLS) estimation on the estimated sources (model order = 4; forgetting factor = 0.9889, in order to take into account measurements up to the previous 3s). At each time instant the first peak of the AR power spectral was detected. The performance of this algorithm were firstly tested on four HC subjects monitored with an ECG in order to test the accuracy during HR variations due to a mild exercise (pedaling on an indoor cycling machine).

Results and discussion

Acoustical analysis in PD patients

The results of this study show that PD patients exhibit an alteration of prosodic patterns of speech during a sentence repetition task. With respect to control subjects, PD patients have longer pauses between each sentence repetition (T_{inter}) and a lower percentage of "voiced time" during the entire repetition period (duty cycle D%), as reported in Tab. I. On the other hand, a decrease of duty cycle leads to an increase of NSR. Therefore PD patients tend to have a shorter time period occupied by speech than healthy controls, but at the expense of a longer recovery time, since no significant difference was found in $T_{sentence}$ (the time from the start of a sentence to the start of the next one).

Tab. I: - Mean value, standard deviation and t-test result for the acoustical parameters extracted from the sentences repetition task

Parameters	PD patients		Healthy subjects		t-test
	Mean	SD	Mean	SD	
$T_{sentence}$ (s)	3.05	0.67	3.14	0.56	$t(39) = 0.42$, $p = .67$
T_{inter} (s)	0.76	0.33	0.57	0.23	$t(39) = 2.03$, $p = .049$
D%	73.77	7.37	79.77	6.52	$t(39) = 2.68$, $p = .011$
NSR (syll./s)	6.54	0.97	5.79	0.77	$t(39) = 2.68$, $p = .011$

Our results concerning the increase of NSR in PD patients are consistent with other findings where an increase of the number of syllables per second was found during the reading of a text. However, other studies that used the reading of a passage did not show any significant difference in this parameter.

Thus, it is likely that speech rate alterations could be more easily raised from the repetition of a sentence rather than by the reading of a text.

In conclusion, our results show that in PD patients with no or minimal speech problems (UPDRS item 18 score between 0 and 1) the speech rate and temporal alterations are the most noticeable features of dysprosody during a sentence repetition task.

Video analysis of articulatory movements PD patients

The proposed markerless system is able to track lips movements during speech with errors between 1 and 3 mm (against the marker-based optoelectronic reference) for most of the lips points. Considering the low image resolution and the lower cost of the depth sensor (if compared with the optoelectronic method) this is a very promising result. Kinematic parameters of the lower lip show a tendency to underestimate the module of the maximum and the minimum speed values (closing and opening phases) with differences around 20 mm/s. However, high correlation values between trajectories (0.96 ± 0.03), velocities (0.95 ± 0.05) and accelerations (0.88 ± 0.10), confirm that trends of these measures are very similar to the reference. This suggests that a bias is present in the estimation of the kinematic parameters.

This bias might be due to the distance from the face at which the device was located (about 0.8 m), or to the different framerate of the systems (30 Hz for the depth sensor, 100 Hz for the marker-based method). This distance was a trade-off between the need to move the sensor as close as possible to the subject's face and its characteristic (range of work: 0.4-1.5 m), without interfering with the field of view of the cameras of the optoelectronic system.

During the comparison between PD patients and HC subjects in the syllable repetition task, significant differences were found in $v_{opening}$ ($t(28) = 2.49$, $p = .019$), $v_{closing}$ ($t(28) = 2.32$, $p = .028$) and $a_{opening}$ ($t(28) = 2.13$, $p = .043$). All these values are lower in PD patients. In particular, the opening and closing velocities are reduced by more than 25 mm/s (94.94 ± 33.40 mm/s vs 64.45 ± 30.94 mm/s for $v_{opening}$, 87.85 ± 31.28 mm/s vs 61.54 ± 28.49 mm/s for $v_{closing}$). Lower values were found also for $a_{opening}$, although not significant. Concerning the opening parameters ($\Delta Opening_{norm}$, $MaxOpening_{norm}$) the normalized range of opening is lower in PD patients (0.65 ± 0.36 vs 0.46 ± 0.23) although not significant, while the normalized maximum value is comparable with values around 0.4 in both groups (i.e., the maximum opening is about the 40% of the mouth width). Our results confirm the previous findings where PD patients exhibited reduced peak velocities of lower lip, both for opening and closing phases. In addition, the peak values of acceleration are reduced in PD patients with significant differences only for the opening phase. Thus, our work supports most of the literature on the kinematic analysis of the articulators in PD patients which states that most of these patients exhibit a downscaling of the articulatory movements. The novel contribution of this work concerns the assessment of this downscaling by means of a fully markerless and low-cost method. This is the first attempt to study articulatory movements in dysarthric patients by means of video-based contactless

methods. This method could be implemented at home in order to help these patients in performing speech therapy and rehabilitation exercises, since a large percentage of them suffers from dysarthria.

Video analysis of facial expressions in PD patients

Results from the automatic facial expression recognition showed that HC subjects perform anger and disgust better than PD patients during the acted task (42.51% vs 29.32% for anger, 43.42% vs 35.00% for disgust).

During the imitation task there is a noticeable increase of anger and disgust only in HC subjects (from 42.51% to 60.98% for anger, from 43.42% to 76.89% for disgust), while PD patients show a small increase of the anger percentage (from 29.32% to 33.87%) and a decrease of the disgust (from 35.00% to 32.89%).

On average, HC subjects reported higher distances from the baseline than PD patients along the whole tasks (12.68 ± 5.05 for HC subjects vs 9.35 ± 3.85 for PD patients, $p < .00001$).

These results show that anger and disgust are the two expressions in which HC subjects show a higher increase in the target expression during the imitation task. This means that HC subjects benefit from the displaying of the target expression in order to perform the imitation of those facial expressions. This not true for PD patients, where the percentage of the target expressions (anger and disgust) remains stable or decreases from the acted task to the imitated task. These results confirm in an objective way the results reported in literature, that is the loss of facial expressions in PD patients could be due to an impairment in facial emotion recognition, with main deficits in the decoding of emotions with negative valence (like disgust, fear and sadness). During happiness expression PD patients and HC subjects show similar results, with low differences in the target expression between acted and imitated also for HC subjects.

Sadness is the expression with the worst results in terms of target expression percentage. In fact, after the experiment, the opinion of both groups was that it is very difficult to simulate sadness even when there is an image for the imitation. In this task both groups tend to assume different expressions, in particular anger and disgust.

In all the four acquisitions PD patients exhibited higher percentage of neutral with respect to HC subjects. This confirm the presence of a hypomimia that reduced the displaying of other facial expressions.

On average, HC subjects reported higher distances than PD patients along the whole tasks; this confirms that HC subjects show larger movements also during posed and imitated facial expressions. PD patients do not show any significant difference between acted and imitated expression in all the 4 tasks. This confirm what described above, namely PD patients do not improve their facial mimicry even when they are asked to imitate another expression. This result could confirm that facial hypomima is not only a motor disorder, but also the result of an impairment at the level of the emotion recognition processes, although further investigations are needed.

Video Analysis of DOC patients: facial movements and vital signs

A one-way repeated-measures analysis of variance (ANOVA) was performed to compare the effect of the noxious stimuli over 4 time instants (before the first stimulus - baseline - and after the first 3 stimuli) on the Euclidean distance of facial features from the neutral template and on the HR estimation. This analysis was performed on 7 patients. Results showed that the stimulus administration produced a significant increase in the Euclidean distance from the neutral template over the 4 time instants, $F(3,18) = 7.1994$, $p = .002$. No significant differences were found in HR values over the 4 time instants, $F(3,18) = 2.6155$, $p = .08$. Concerning the Euclidean distance: no significant differences exist between the baseline (mean: 10.99, SD: 6.97) and after the administration of the first stimulus (mean: 13.79, SD: 9.18), $t(6) = 1.73$, $p = .13$; there was a significant difference between the baseline and the interval after the second stimulus (mean: 19.28, SD: 10.09), $t(6) = 3.05$, $p = .02$, and between the baseline and the interval after the third stimulus (mean: 24.49, SD: 13.48), $t(6) = 4.21$, $p = .006$. Concerning HR, no significant differences were found during the comparison between the baseline and the three intervals after the stimuli.

No significant correlations were found between distance and CRS, age and months after the onset and between HR, CRS, age and months after the onset.

Thus, DOC patients showed a significant increase of the Euclidean distance of facial features from the neutral template during the course of the experiment (in particular after the second stimulus). This is consistent with an increase of facial mimicry and facial movements that might indicate a patient's reaction. This distance has lower values during the baseline time interval (i.e. before the administration of the first stimulus), reflecting a lower activity in facial mimicry with a facial expression similar to the neutral template. However, this is not true for the HR, where the mean value remains stable during the first 3 stimuli (around 80 bpm). Moreover, the variations along the administered stimuli are very small (around 5 bpm on average) and thus comparable with the accuracy of this estimation method. Thus, it is not possible to draw any conclusion on HR variations, although it is reasonable assume that only small changes occurred in DOC patients, at least considering the HR.

This is due to the small number of patients considered for this study (only 9 patients). These patients had different etiology, different age and different post-comatose outcome (7 VS patients and 2 MCS patients). More accurate results could be obtained from larger samples, differentiating between VS and MCS patients. The present work is a pilot study and a first attempt to make a contactless automated monitoring of DOC patients, that could be easily extended to larger populations and to other CRS-R items (i.e., items for the evaluation of auditory, communicative functions), as well as to the interaction of patients with their relatives. In fact, some neuroimaging studies demonstrated that an activation in some cerebral areas (such as the amygdala) was found in MCS patients after listening to a familiar voice, while is absent in VS patients. For this reasons our study could be extended to larger groups of MCS and VS patients in order to find possible differences in facial mimicry and HR, evaluating within and between-groups differences.

Conclusion

This PhD projects provides first results concerning the development of a contactless system for monitoring facial expressions and facial movements, with applications to neurology (in particular Parkinson's disease and disorders of consciousness).

Interesting results were obtained in PD patients were through a markerless and low-cost method it was demonstrated what already found in literature with more expensive techniques, that is PD patients exhibit reduced kinematic parameters of the articulatory movements during speech and reduced extents of facial expressions. Both are related to facial bradykinesia. Thus, the proposed methods could be implemented at home in order to help these patients in performing speech therapy and rehabilitation exercises, since a large percentage of them suffers from dysarthria and facial hypomimia. Moreover, this methodology allows the investigation of other disorders that affect speech production and facial movements/expressions (amyotrophic lateral sclerosis, Alzheimer's disease, stroke, etc.).

The analysis of facial movements and facial expressions was then extended to DOC patients, in order to highlight possible reactions after the administration of external stimuli. This analysis, in conjunction with HR estimation through video-based techniques, showed partial variations, with a visible and significant trend only for facial expressions. Thus these results need to be extended to larger populations differentiating between post-comatose outcomes. However, the first results obtained during these project might be a first step towards the development of an automated monitoring system for DOC patients, in order to assess patients' reactions even when nobody is near them, thus providing a reliable support to clinicians in the diagnosis rehabilitation (that today is still prone to high rates of errors).

Part I deals with the clinical and methodological background about the topics covered in this PhD thesis.

Chapter 1 is focused on motor signs of Parkinson's disease (PD), with special attention to speech and facial expressions impairments that affect high percentages of these patients.

Chapter 2 provides an overview about disorders of consciousness (DOC), focusing on the differential diagnosis between vegetative state and minimally conscious state. At date, this diagnosis is performed through neurobehavioral scales and is prone to high rates of misdiagnosis.

Chapter 3 summarizes the state of the art of the processing techniques used in this project. Through this chapter, the following topics are covered: acoustical analysis to study voice and speech problems in Parkinson's disease; kinematic analysis of articulatory movements during speech and investigations in dysarthric patients; automatic analysis of facial expressions with applications to Parkinsonian patients with hypomimia; heart rate estimation through video-based methods.

In **Part II** video and signal processing techniques used throughout the studies performed in this project are presented.

In **Chapter 4** the aims of the work and the audio/video acquisition protocols developed during this project are described. In particular, the study of speech and articulatory movements in PD patients was performed using depth sensors (i.e. Microsoft Kinect), in order to extract three-dimensional information about the observed scene.

Chapter 5 is focused on the acoustical analysis of PD speech. In this part of the work, voiced/unvoiced segmentation algorithms were applied to detect time parameters related to dysprosody during syllable and sentence repetition tasks.

The syllable repetition task is also used for the analysis of articulatory movements (in particular lips movements) through a markerless technique developed within this project and composed by a depth sensor and a face tracking algorithm.

Chapter 6 summarizes the tests performed to test the accuracy of this method against a marker-based optoelectronic system, while in **Chapter 7** this technique is implemented to study the articulatory movements during speech in two groups: PD patients and age-matched control subjects. The aim is to highlight the presence of reduced articulatory movements (in terms of displacement, speed and acceleration of the articulatory organs) through a fully contactless and low-cost method. In fact, most of the works in this field performed these investigations through marker-based expensive methods (EMA, optoelectronic systems, etc.).

Chapter 8 is focused on the study of facial expressions in PD patients. In this part of the work acted expressions (upon request of the clinician) and imitated expressions (after the visualization of a picture with the requested expression) were studied. The aims are the objectification of facial hypomimia and the investigation of emotion recognition processes that seem to be impaired in PD patients. Today a big debate in the research community is focused on the nature of hypomimia in PD patients. In fact, many works support the hypothesis that the decrease or loss of facial expressions could be due to an impairment in facial emotion recognition, while other studies state that hypomimia is the consequence of a motor impairment. Thus, the aim of this work was setting quantitative measures about facial mimicry through face tracking and classification algorithms.

Chapters 9 and 10 are focused on methodologies used for studying DOC patients. This part combines the study of facial expressions (through the methods exposed in Chapter 8) and heart rate through video-based techniques. In **Chapter 9** a recursive HR estimation method is proposed, comparing its performance against a reference ECG. In **Chapter 10**, the aforementioned methods are used to study patients' reactions after the administration of standardized stimuli, as expected by the CRS-R protocol.

Part III concerns the experimental results obtained through the different steps of the project. Each chapter is related to a chapter of Part II and reports the results obtained with the described methodologies.

Chapter 11 describes the results obtained with the acoustical analysis algorithms described in Part II - Chapter 5 and applied to PD patients.

In **Chapter 12** the results about the accuracy of the markerless method for studying speech articulators (described in Part II - Chapter 6) are shown.

Chapter 13 reports the results on the kinematic analysis of articulatory movements in PD patients obtained with the methodologies reported in Part II - Chapters 6 and 7.

Chapter 14 is focused on the results of the study of facial hypomimia in PD patients through the facial expression recognition algorithms described in Part II - Chapter 8.

Chapter 15 concerns the accuracy results of the video-based HR estimation algorithm whose processing details are reported in Part II - Chapter 9.

Chapter 16 reports the results of the study on DOC patients, where reactions to standardized stimuli are investigated through the methods described in Part II - Chapters 9 and 10.

PART I - BACKGROUND

1. Hypokinetic dysarthria and facial hypomimia in Parkinson's disease

Alterations of voluntary and involuntary movements of the human being are clinical features common to many neurodegenerative diseases. Considering the constant increase in life expectancy of the world population and the high rate of incidence of neurological disorders, the automation of the diagnostic routines and of the rehabilitation processes is highly desirable. One of the most significant cases is Parkinson's disease, a neurodegenerative illness that affects around 2% of the world population over 65 years old.

In this chapter the most important features of Parkinson's disease are explored, highlighting two motor symptoms that are the main subjects of this thesis: speech and facial expression impairments.

1.1 Parkinson's disease - background

Idiopathic Parkinson's disease (PD) is a neurodegenerative illness that involves neurons in the zona compacta of the *substantia nigra* of the midbrain and other pigmented nuclei [1,2,3]. *Substantia nigra* is a pigmented group of neurons composed by two parts: the *pars reticulata* and the *pars compacta* (Fig. 1.1). This loss of neurons leads to an insufficient synthesis and production of dopamine, an essential neurotransmitter for the control of body movements.

This pathology is associated with a wide range of motor (tremor, stiffness, bradykinesia, postural instability) and non-motor (depression, cognitive impairments, sleep and mood disorders) symptoms that significantly reduce the quality of life of patients [4,5]. PD was described for the first time as a "Shaking Palsy" by James Parkinson in 1817 [6,7]. PD is the second most common neurodegenerative disease, following the Alzheimer's disease [8]. Currently, in Italy, there are around 230000 people affected by PD, with higher incidence in male subjects (around 60% of patients are male). Due to the increase in life expectancy and the increasing aging of the world population, it is expected that in 2030 the number of cases will be double [9,10]. The onset age is around 65 years old, although there is a small percentage of patients less than 50 years old.

1.1.1 Pathogenesis

Idiopathic PD is a chronic and progressive disease of the extrapyramidal central nervous system, defined by a clinical syndrome characterized by bradykinesia, rigidity, tremor and postural instability [2,4,11,12]. This syndrome is commonly known as "Parkinsonism", and is characterized by all or some of the aforementioned motor symptoms. There are different kinds of Parkinsonism, which may differ for the underlying causes (gene mutations, metabolic disorders, toxins or drugs) [4].

PD is characterized by the degeneration of the dopaminergic neurons in the *pars compacta* of the *substantia nigra* [1,2] as shown in Fig. 1.1.

It is estimated that around 60% of the neuronal population involved is already compromised when symptoms become visible.

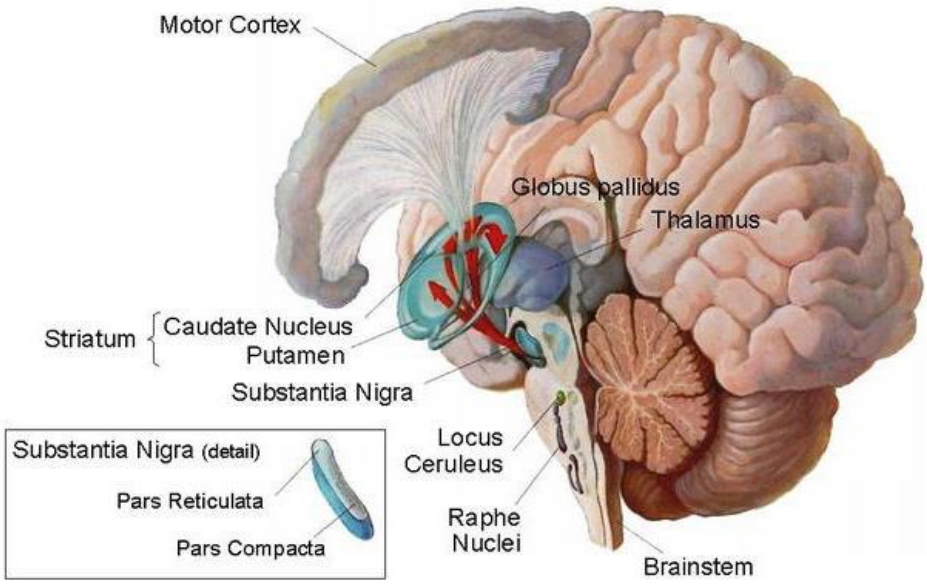

Fig. 1.1: Brain regions affected by Parkinson's disease [13].

Neurons of the *substantia nigra* normally produce dopamine, sending signals to the basal ganglia, a system whose purpose is to associate sensory information and motor commands for a proper motion planning. In basal ganglia disorders, as PD, the reduction of these signals affects the regular activation/deactivation of the individual components of a motor sequence.

At histopathological level, the neurons of the *pars compacta* may present the following characteristics: depigmentation, gliosis and appearance of the Lewy bodies, abnormal inclusions of protein masses composed by α-synuclein [13]. An example of the Lewy bodies in the *substantia nigra* is reported in Fig. 1.2.

Although a multifactorial origin is considered for PD, involving genetic and environmental factors, to date the exact causes of the onset the disease are not yet fully known [6].

1.1.2 Symptoms

As mentioned above, PD is characterized by three cardinal motor signs: tremor at rest, muscular rigidity and akynesia/bradykinesia [11,14]. During the course of the disease other non-motor symptoms may appear: depression, cognitive impairments, sleep and mood disorders. This combination of motor and non-motor symptoms significantly reduces the quality of life of patients [4,5,15].

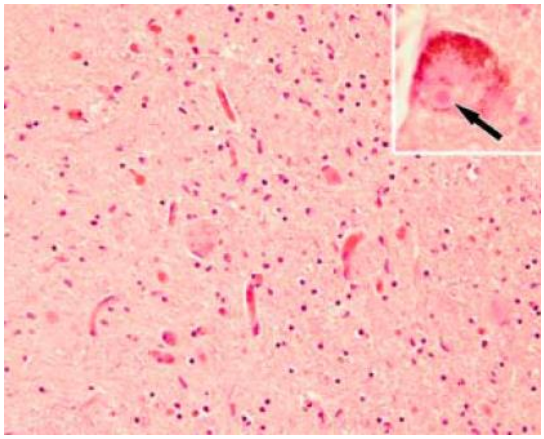

Fig. 1.2: Substantia nigra in presence of Parkinson's disease, with characteristics Lewy bodies [4].

Tremor at rest is a very common sign in PD patients, it appears often in a hand, arm, or leg, and may occur in 80% of PD patients. This sign can help in the diagnosis during the first stages of the disease. The main characteristics are: low and regular frequency (around 4-5 Hz), unilateral debut, reduction or disappearance during voluntary movements, worsening in situations of emotional stress and fatigue, complete disappearance during sleep.

Rigidity is due to an increase in muscular tone and is visible in limbs, neck and trunk. The increase in muscular stiffness is detectable by means of the passive mobilization of the body segments, especially at levels of wrists and elbows. Consequently, gait shows a deterioration, becoming rigid, in small steps and sometimes with accelerations (festination of gait) or blocks (freezing of gait), especially in changing direction.

Bradykinesia is a slowdown in speed and amplitude of the motor gesture. This symptom is responsible for the reduction and slowdown of limbs movements. Parkinsonian patients may be subjects with reduced extents of movements. Bradykinesia is manifested also in the facial muscles, and is responsible of the reduction of the facial expressions. In several cases, facial bradykinesia may give rise to a complete lack of facial mimicry, giving the impression of a facial paresis. This symptom is further described in Section 1.2.2.

During the course of the disease other motor signs, such as postural instability and “freezing” of gait, are very common. In particular, during freezing, patients may experience transient episodes of short duration with difficulties in starting the action, with the feeling of having feet glued on the ground. Other common motor sings are speech and voice impairments. During speech production, it is common to recognize the same motor signs visible in limbs: reduced extents and slowdown of movements, muscular rigidity and tremor. Speech and voice impairments are further described in Section 1.2.1.

In addition to motor signs, there are several non-motor symptoms: depression, cognitive impairments, sleep and mood disorders and sexual dysfunctions [4,5]. Among the cognitive impairments, attention

disorders, depressive syndromes and dementia (especially in the latest stages) are common in PD patients [10].

1.1.3 Diagnosis

The diagnosis of Parkinson's disease is primarily based on a careful anamnesis during the neurological examination. Nevertheless, there may be many diagnostic errors due to the subjectivity of the examiner. For these reasons some diagnostic criteria were formulated; the most used are: the criteria of the UK Parkinson's disease Brain Bank and the Gelb's criteria [4,10].

The UK Brain Bank criteria require the presence of bradykinesia along with at least one of the following symptoms: rigidity, resting tremor, postural instability. Other causes of the Parkinsonism must be excluded, i.e. stroke, head injury, neoplastic pathologies, psychotropic drugs treatment. Moreover, another fundamental requirement is the presence of at least three signs among: unilateral onset, tremor at rest, persistent asymmetry, response to levodopa, and levodopa-induced dyskinesia.

Gelb's criteria require the presence of at least two of the following symptoms: tremor at rest, bradykinesia, rigidity, unilateral onset, of which at least one among tremor and bradykinesia. Moreover, as for the UK Brain Bank criteria, other possible causes of the Parkinsonian syndrome should be excluded.

Neuroimaging techniques can be used to exclude other pathologies with similar symptoms. Among these methodologies, Single Photon Emission Computed Tomography (SPECT) and Positron Emission Tomography (PET) are used as an aid to the diagnosis. Another neuroimaging technique is the "dat-scan" a scintigraphy technique able to identify the loss of dopaminergic neurons in the *striatum* and in the *substantia nigra* in patients with uncertain PD diagnosis. This technique is performed with the injection of a radiotracer (ioflupane I-123) and allows the evaluation of the extent of damages to the dopaminergic system [4].

Despite pharmacological and neurosurgical treatments, one of the main problems of PD is its progressive nature. Thus, a constant and accurate evaluation of the motor symptoms progression is fundamental to modify and adapt the treatments during the course of the disease. For these reasons, PD patients undergo periodic and thorough assessments of symptoms [10] through the use of quantitative scales, such as the Unified Parkinson's Disease Rating Scale (UPDRS) and the Hoehn and Yahr (H&Y) staging scale [16,17].

The UDPRS was firstly introduced in 1987 and published in a new version in 2003 by the Movement Disorder Society (MDS). Its aim is to give a standardized and quantitative assessment of the disease symptoms. This scale is composed by four parts:

- Part I: evaluation of mood and behavioral aspects of patients (intellectual impairments, thought disorder, depression, motivation and initiative);

- Part II: auto-evaluation of the activities of daily living (speech performance, salivation and swallowing, handwriting, cutting food and handling utensils, dressing, hygiene, falling, walking and freezing, tremor, etc.)
- Part III: motor examination of patients (discussed in depth in section 1.2);
- Part IV: quantification of therapy-related complications (dyskinesia, clinical fluctuations, etc.).

Each section is subdivided into several items. For each of these items, clinicians give a score that ranges between 0 and 4: 0 - absence of the symptom, 1 - slight symptom, 2 - mild to moderate symptom, 3 - marked symptom, 4 - severe symptom. The main advantage of this scale is its uniform interpretability by means of a numeric score and constitutes an important means to reduce the misdiagnosis rate. However, these evaluations may be subjective since they depend on the human observation. Thus, this scale could fail in the assessment of slight progressions of the symptoms.

Another important scale is the H&Y staging scale [16]. Through this scale it is possible to divide the disease into different stages, depending on the symptoms progression evaluated through a clinical observation. The disease progression is divided into 5 severity phases, according to an increasing score from 1 to 5. Other intermediate stages were later introduced (1.5 and 2.5), in order to describe the disease progression with more precision. The H&Y disease stages are reported in Fig. 1.3.

Hoehn and Yahr Scale	Modified Hoehn and Yahr Scale
1: Only unilateral involvement, usually with minimal or no functional disability	1.0: Unilateral involvement only
	1.5: Unilateral and axial involvement
2: Bilateral or midline involvement without impairment of balance	2.0: Bilateral involvement without impairment of balance
3: Bilateral disease: mild to moderate disability with impaired postural reflexes; physically independent	2.5: Mild bilateral disease with recovery on pull test
	3.0: Mild to moderate bilateral disease; some postural instability; physically independent
4: Severely disabling disease; still able to walk or stand unassisted	4.0: Severe disability; still able to walk or stand unassisted
5: Confinement to bed or wheelchair unless aided	5.0: Wheelchair bound or bedridden unless aided

Fig.1.3: Hoehn and Yahr disease stages (left) and modified Hoehn and Yahr scale (right) with the intermediate stages 1.5 and 2.5.

1.1.4 Therapies

Pharmacological treatments

Cardinal motor signs of PD respond favorably to the dopaminergic therapy, especially to levodopa, for a variable number of years depending on the single patient. This response to the treatment is a feature so important as to be included in the diagnostic criteria of idiopathic PD exposed in the previous section [11,14]. Conversely, postural instability and freezing of gait, whose pathogenesis cannot be entirely attributed to the deficit of the nigrostriatal pathway, generally do not respond satisfactorily to

the dopaminergic therapy [5]. Moreover, some non-motor signs (as psychotic manifestations) can even be induced or unmasked by the dopaminergic therapy [18,19,20].

Levodopa (or L-Dopa) is a dopamine precursor which is able to cross the blood-brain barrier. Despite its powerful effect on PD symptoms, in most of the patients the prolonged use of this drug is accompanied by the onset of complications, like fluctuations in motor performance during the day and involuntary movements (dyskinesia and dystonia). These phenomena depend on a complex interaction between the progression of PD and the effects of levodopa and are related to the duration of the treatment as well as to the dose of the taken medication.

Dyskinesias are irregular movements (both fast and slow) that mainly occur in facial muscles, neck, jaw, tongue, trunk causing in turn abnormal movements of shoulders and hips [21]. It is widely accepted that the stable response to the L-dopa that characterizes the first stages of PD mainly depends on the ability of the remaining dopaminergic neurons to transform the exogenous L-Dopa into dopamine that is successively released according to the physiological mechanisms. With the nigro-striatal pathway degeneration, the transformation of L-Dopa in Dopamine is possible only outside the nigro-striatal neurons. Thus, the effects of the neurotransmitter become dependent on the brain and plasma levels of L-Dopa. Experimental results indicate that the pulsatile dopaminergic stimulation related to the fluctuating striatal concentration of L-Dopa plays a key role for the onset of motor complications in PD [21]. For these reasons continuous and constant dopaminergic stimulation is necessary in order to plan an efficient pharmacological treatment. Some examples of this continuous dopaminergic administration are: the use of inhibitory drugs of the key enzymes in the metabolism of Dopamine, sub-cutaneous apomorphine administrations and administration of Dopamine directly into the duodenum.

Neurosurgical treatments

In cases where drug therapy does not provide further benefits for the control of PD symptoms, some neurosurgical treatments are implemented. The most recent and widespread approach is the Deep Brain Stimulation (DBS), preferred to other ablative treatments (as pallidotomy). During DBS intervention, the stimulation electrodes are implanted in particular brain regions (as the subthalamic nucleus - STN or the *globus pallidus interna* - GPi) and positioned by means of stereotactic surgery, in order to place the electrodes as accurately as possible. Thus, the stimulation electrodes are connected to a pulse generator that is implanted in a subcutaneous region of the chest [10]. An example of a DBS system is reported in Fig.1.4 Several studies investigated and demonstrated the efficacy of DBS for the control of PD motor symptoms. DBS.

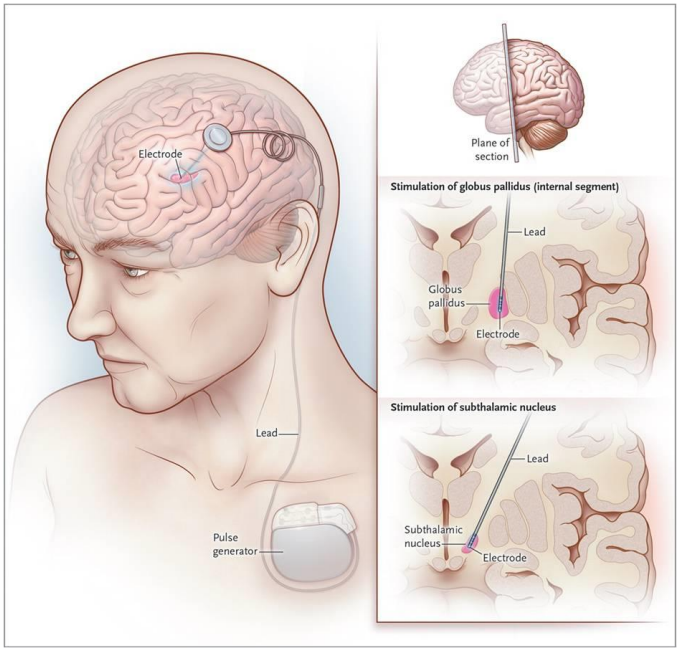

Fig.1.4: Brain regions were the stimulation electrodes are implanted

DBS is able to improve tremor, rigidity and bradykinesia in 15-30 minutes after the stimulation, both with low (60 Hz) and high frequencies (> 100 Hz). High frequency stimulations also alleviate drug-induced motor fluctuations or dyskinesia [22,23]. Other studies [22] reported slight improvements of gait and balance after a few hours of stimulation, with low frequency and high amplitude. However, the main disadvantages of this technique are:

- Currently, no standardized guidelines for programming the DBS (in terms of pulse frequency and amplitude) are available concerning balance and gait disorders [22];
- Neurosurgery should be considered with caution in patients over 70 years old; considering the high incidence of PD over 65 years old, this is a big drawback;
- In addition, not all PD patients are candidates for neurosurgical treatments. Some of the selection criteria include, in addition to the age, the presence of the disease for at least five years, presence of motor complications due to drug therapy and the absence of comorbidity [22,24].

Nevertheless, this technique is able to reduce up to the 50-70% the extent of cardinal motor signs. However, many authors demonstrated that DBS does not improve other common motor sings such speech and voice impairments [25].

1.1.5 Rehabilitation

Despite the different pharmacological and neurosurgical treatments, PD is a degenerative disease that causes a progressive and severe motor disability. Moreover, some motor symptoms are immune to the treatment or, in case of dyskinesia or dystonia, are instead the results of the treatments themselves. For these reasons all of the aforementioned therapies should be accompanied by a rehabilitation program,

in order to strengthen the residual motor skills preventing further damages due to the reduced motility. Depending on the particular needs, patients can follow different kind of rehabilitation: motor rehabilitation, occupational therapy and speech therapy.

The aims of motor rehabilitation are: the improvement of muscular strength, the increase of the amplitude of movements and the reduction of muscular rigidity, by means of physical exercises. The use of visual or acoustic cues can improve some motor aspects of PD patients, as the stride length, the balance and the freezing episodes, for example giving information about the right pace of a motor task. Rehabilitation, if undertaken promptly, is particularly useful and effective in slowing the progression of the disease and reducing the need to increase the drug dosage. However, a constant and daily exercise is fundamental in order to obtain satisfactory results. For these reasons, low-cost and markerless systems have become popular in the last years, in order to increase the number of patients that can perform rehabilitation exercises at home. Most of these systems are based on technologies originally developed for videogames. Two important examples are: the Nintendo Wii and the Microsoft Kinect (that is used in this project to study facial movements in 3D) [26,27,28]. Their use has been encouraged mainly by their low cost (around 100-200 Euros) and their easy use in the home environment. Videogames developed on these systems are useful for the rehabilitation practice, because it is easy to provide a sensory feedback (visual or auditory), thus stimulating the postural control and facilitating exercises that involve complex movements.

1.2 Motor signs evaluation - UPDRS part III

As already indicated in section 1.1.3, the prolonged use of L-Dopa is often accompanied with the onset of complications and fluctuations in motor performance during the day. A reliable evaluation of these fluctuations is of crucial importance to optimize the drug therapy during the course of the disease. The use of semi-quantitative scales facilitates a standardized and reproducible assessment of motor symptoms during the visit. For this purpose the UPDRS [17] is worldwide used, whose section III includes the necessary items for the evaluation of motor signs in PD. For each task, the practitioner provides a score (ranging between 0 and 4, sect. 1.1.3), proportional to the severity of symptoms. The motor UPDRS is divided into 18 items, one for each motor symptom:

1. Speech: patients have a spontaneous conversation with the practitioner about activities of daily livings, work, etc. Clinicians assess the volume, the modulation (prosody) and clarity of speech, the intelligibility and the presence of palilalia (repetition syllables) and tachyphemia (rapid speech, overlapping of syllables).
2. Facial expression: patients are observed at rest and seated, for at least 10 seconds, with and without a conversation. The examiner evaluates the eye-blinking rate, the loss of facial expressions during the session, the ability in displaying spontaneous smile and the opening of the lips.

3. Rigidity: patients are evaluated in passive movements of the main joints, in a relaxed position. The examiner moves patient's limbs and neck in order to evaluate the muscular stiffness.
4. Finger taps: patient has to tap the thumb with the index finger, repeating the movement for 10 times as faster and wider as possible (repeated for both hands). The examiner considers speed, amplitude, hesitations, interruptions and amplitude reduction along the task.
5. Hands movements: similar to item 4, but during an opening-closing movement of the hand (patient opens and closes hands in rapid succession).
6. Rapid alternating movements of hands: pronation-supination movements of hands, vertically and horizontally, with as large an amplitude as possible (for 10 times and for both hands). The examiner considers speed, amplitude, hesitations, interruptions and amplitude reduction along the task.
7. Feet movements: patients, in a seated position, have to keep their heels on the ground and then tap their toes for 10 times, in a rapid and wide succession (repeated for both feet). The examiner considers speed, amplitude, hesitations, interruptions and amplitude reduction along the task.
8. Leg agility: similar to item 7, but during the beating of the heel on the ground in rapid succession. Patients should picking up entire leg of at least 3 inches from the ground level.
9. Arising from chair: patient must attempt to rise from a chair, with arms folded across chest. The examiner considers hesitations and interruptions during the rising gesture. Once the patient is standing, the examiner evaluates the posture (item 13).
10. Gait: patient has to walk for a few meters and then must turn around and go back towards the examiner. The examiner considers the stride length, the speed, the amplitude of the heel lifting from the ground, the dragging of the heel during the gait. Moreover, the examiner can evaluate possible freezing events (item 11), in particular hesitations in the beginning or in the changing of direction, and the posture during gait (item 13). The examiner considers the number of backwards steps or possible falls.
11. Freezing of gait: see item 10.
12. Postural stability: response to sudden, strong posterior displacement produced by pull on shoulders while patient erect with eyes open and feet slightly apart (retropulsion test).
13. Posture: see items 9 and10.
14. Body Bradykinesia and hypokinesia: this item combines slowness, hesitancy, decreased arm swing, small amplitude, and poorness of movement observed during all the tasks.
15. Postural tremor of hands: the patient has to stretch the arms forward with palms facing down. The examiner considers the presence and the amplitude of tremors.
16. Kinematic tremor of hands: this item is evaluated during and index-nose test. The patient has to repeat a movement touching his nose and then reaching a target (i.e., the examiner's finger).

This task is repeated for both hands. The examiner considers the presence and the amplitude of tremor during the arm movements.

17. Amplitude of rest tremor: overall assessment of the tremor amplitude (at rest) during the whole examination.
18. Continuity of rest tremor: overall assessment of the continuity of tremor at rest during the whole examination.

Once the examiner has evaluated and assigned a score for each one of these items, the overall motor UDPRS score is the sum of the 18 sub-tasks. This score allows a quantitative evaluation of the disease progression. However, an objective assessment only during the visits (conducted every 3-6 months) is not sufficient to evaluate fluctuations that occur during the days. In addition, the possibility to use diaries compiled by the patient or care giver in the various phases of the day, does not guarantee the completeness and standardization of the contained information.

1.2.1 Speech and voice impairments

Impairments of voice and speech (hypokinetic dysarthria) are among the main signs of the disease. They affect about 70% of PD patients [29]. PD patients may present alterations related to all speech dimensions (i.e., the four subsystems of speech production): respiration, phonation, articulation and prosody. Therefore, they might exhibit reduced variability of pitch and loudness, hoarseness, reduced stress, imprecise consonant articulation, unorthodox speech silence and speech rate alterations [30]. These alterations result in a reduced intelligibility of these patients that may cause a tendency to a social isolation. The severity of speech and voice impairments increases with the disease progression.
Voice in PD patients has the following characteristics:

- low intensity and fading during speech: even if patients start to speak with high vocal intensities, the volume progressively decreases;
- vocal tremor: some patients may experience a vocal tremor due to the tremor of the muscles of the larynx;
- monotone pitch: pitch variations are small; this results in a loss of expressiveness during speech;
- hesitations in starting to speak: it is hard to start speaking and then to maintain a steady voice until the end of a conversation;
- difficulties in pronunciation and articulation: often the final phonemes of the words are unclear or even absent;
- speech acceleration and loss of timing control: syllables and words may pile up and flow without pauses. There may be a progressive acceleration in the pronunciation of words and syllables in the end of a sentence.
- nasalization: the soft palate cannot move properly and failing to block the passage of air to the nose.

From these features it is easy to assess many analogies with other motor impairments in PD patients: tremor, speech acceleration (like a festination of speech), hesitation in starting motor gestures and bradykinesia that in turn causes a reduction of the articulatory movements.

This set of speech and voice impairments in PD patients is usually known as "dysarthria". This term refers to a group of speech disorders caused by an alteration of the muscular control of the pneumo-phono-articulatory organs due to damage in the basic motor processes involved in speech production [31,32]. Different neurological diseases with different brain lesions or different causes, may lead to different kind of dysarthria [33]: flaccid dysarthria, spastic dysarthria, ataxic dysarthria hypokinetic dysarthria, hyperkinetic dysarthria and mixed dysarthria. In most of PD patients, dysarthria is usually "hypokinetic". This term refers to the reduced range of movements involved in speech production. In fact, hypokinetic dysarthria is characterized by reduced peak velocities and displacements of the articulators (i.e., jaw, lips and tongue) during speech movements. Moreover, as explained above PD patients exhibit a harsh breathy voice quality. This is due to a non-complete closing of the vocal folds.

All of these characteristics are the result of impairments in the muscular control of the pneumo-phono-articulatory system. Unlike other types of dysarthria and neurological disorders, hypokinetic dysarthria in PD involves all the speech dimensions, thus resulting in alterations in muscular control of lungs, larynx, lips, soft palate, tongue, jaw.

1.2.2 Facial expressions impairments

One of the most common motor signs of Parkinson's disease is facial bradykinesia, consisting in the reduction and slowness of facial movements. Facial bradykinesia may affect both the upper and the lower part of the face. PD patients exhibit a blink rate reduction and reduced smile expressions (both spontaneous and posed), as well as impaired voluntary orofacial movements (as reported in the previous section). Facial bradykinesia differs from bradykinesia of limbs for different reasons:

- lack of joints on which the facial muscles act;
- facial muscles have few or no proprioceptors;
- rigidity and tremor rarely influence facial muscles;
- limbs movements are mainly voluntary, in contrast to facial movements that are affected by the emotional state.

Face is an important means for conveying information about the emotional state of a person. Facial bradykinesia is reflected in a reduction or loss of spontaneous facial movements and emotional facial expressions called hypomimia. PD face is often a "masked" or "poker face", typical of people without any interest in the surrounding environment. External observers and in particular practitioners, may have difficulties in decoding the emotional state behind the "mask" [34]. This inaccurate impression could lead to problems with social relationships and interactions in PD patients with higher masking [35].

One of the most altered facial expressions in Parkinsonian patients is smile. When PD patients experience spontaneous happiness, their smiles are perceived as fake and insensitive, because of the lack of cheeks raising and due to a loss of tone of the orbicularis oculi muscles. This impairment in spontaneous facial expressions is consistent with the neuroanatomic evidence of damage to the extrapyramidal motor system. Moreover, PD patients may experience difficulties in performing voluntary facial expressions and facial movements [36,37,38,39,40]. Although posed expressions are believed to be originated from the cortical motor strip, a system that should be intact in PD patients, also these expressions appear to be impaired in PD patients [38,39]. The main signs of hypomimia in PD are [41]:

- Wider palpebral distances, that in conjunction with the reduced blink rate gives the impression of a staring expression;
- Flattened nasolabial folds;
- Unintentional lips separation (mouth opening).

In most of PD patients hypomimia is bilateral and symmetrical, although low percentages (around 6%) of them reported symptoms of hemihypomimia (i.e., one side of the face is more affected than the contralateral one). In particular, Ozekmekci et al., 2007 [42] reported that hemihypomimia affects the right hemiface in patients with predominantly right-sided PD symptomatology.

Hypomimia is improved by the dopaminergic therapy, while some studies [43] demonstrated that neurosurgical treatments (in particular the STN-DBS) may worsen this sign.

As mentioned above one of the most common signs in PD hypomimia is the reduced spontaneous and posed smiling. This reduction in smiling could also be related to depression [43,44]. Today a big debate in the research community is focused on the nature of the hypomimia in PD patients. In fact, many works support the hypothesis that the loss of facial expressions could be due to impairment in facial emotion recognition, with main deficits in the decoding of disgust, fear and sadness [41].

1.3 Speech therapy and Parkinson's disease

It is now accepted that physical and speech rehabilitation (in conjunction with drug therapy) can avoid a worsening of motor symptoms [45]. However, the beneficial effects of rehabilitation may be reached only with a constant exercise. It is therefore essential that patients should be able to perform speech therapy not only in rehabilitation and clinical structures, but also in the home environment, in order to ensure a higher constancy in the exercises.

One of the most important speech therapy treatment for PD patients is the Lee Silverman Voice Treatment (LSVT®) [46,47]. LSVT is an intensive treatment program whose aim is to improve vocal fold adduction, vocal loudness and the overall voice and speech production in PD patients with a simple set of tasks designed to maximize phonatory and respiratory functions. Patients are constantly instructed and stimulated to produce a loud voice with their maximum effort during various speech

tasks (as the sustained phonations). Treatment is administered in 16 sessions in a month, divided into four individual sessions of 60 minutes per week. Moreover, LSVT was successfully applied to other kind of Parkinsonism (Shy-Drager syndrome, multi-system atrophy and progressive supranuclear palsy) and to patients with stroke, multiple sclerosis, Down syndrome and cerebral palsy.

Several studies support the improvements in vocal loudness, intonation and voice quality after the administration of the LSVT, with improvements maintained up to two years after the speech therapy treatment. Moreover, other studies showed that LSVT can improve other common problems in PD dysartrhic patients, i.e. facial hypomimia and swallowing disorders [48].

Another key role in speech therapy for PD, is played by the exercises for speech articulators and facial muscles [49]. The aim of these exercises is to reinforce and improve the amplitude of the oro-facial and articulatory movements, in order to get better speech articulation performance. Some of these exercises are:

- Movements of tongue (pull out, rotations and movements inside the mouth);
- Exercises for lips and cheeks (repetitive movements like opening and closing, smiling, protrusion and rounding of lips, puffing cheeks);
- Displaying of facial expressions;
- Movements of the eyelids.

Despite the high percentages of PD patients that suffer from dysarthria, only a small part of them undergoes speech therapy. This discrepancy is due to several factors:

- during group sessions, the speech therapist has difficulty in placing the right attention to each patient, in order to evaluate the exercises and provide a feedback to the patients;
- most of the patients with hypokinetic dysarthria, due to neurodegenerative diseases, are elderly people, who could encounter difficulties in moving to specialized centers;
- in addition, these patients should perform also the exercises at home. However, they might not be stimulated without the supervision of the therapist;
- despite several studies demonstrated the effectiveness of speech therapy on the phono-articulatory performances of PD patients there is still a lack of awareness about benefits that can be achieved with speech therapy treatments.

For these reasons a system that could automatically provide a feedback about the articulatory movements and that could also be implemented in home environment would be highly beneficial. Therefore, this system should be as much as possible, contact-less and at reasonable price.

2. Differential diagnosis of disorders of consciousness

Disorders of consciousness (DOC) after coma in severe brain injury are: Vegetative State (VS), characterized by the absence of self-awareness and awareness of the environment, and Minimally Conscious State (MCS), in which certain behaviors, although functionally inconsistent, are sufficiently reproducible to be distinguished from simple reflex responses. The interest in these conditions has gradually spread in the medical and scientific community in relation to the relevant bioethical theme of end of life decision-making processes. Institutional information about incidence and prevalence provide variable data. However, it is estimated that only in the Tuscan region there are approximately 220 new cases every 1 million inhabitants per year [50]. Nationwide, there are around 2000 patients who require continuous care in specialized centers or at home with obvious economic implications. The differential diagnosis between VS and MCS is currently based on neuro-behavioral scales, instrumental examinations (Electroencephalography- EEG, evoked potentials) and functional neuroimaging techniques, with a high rate of misdiagnosis. Since only MCS patients may evolve towards a recovery, the importance of combining rating scales with objective, non-invasive, contactless and cheap methods is of increasing relevance. Another important aspect for these subjects is the assessment of pain, which still relies on the perceptual evaluation of facial expressions, respiratory rate, body language, which are difficult to interpret.

2.1 Disorders of consciousness - background

Consciousness is a multifactorial concept composed by two main elements: self-awareness and awareness of the surrounding environment ("awareness" for simplicity) and "wakefulness". According to the different levels of these two components, there can be many conditions (Fig.2.1):

- Wake state: the person is awake and has a full self-awareness and awareness of the surrounding environment;
- Drowsiness;
- Light sleep;
- Deep Sleep;
- Coma: the person has a complete failure of the excitatory system, with closed eyes, and is unable to be woken up with external stimuli.

In neurology, coma is a deep unconsciousness state, characterized by the absence of both aspects of consciousness (awareness and wakefulness), and caused by the inactivation of those neuronal circuits responsible of maintaining a waking state. These circuits are in the area of the brainstem called reticular activating system (RAS). RAS grows from the spine in a reticular formation and is connected to midbrain, myelencephalon, hypothalamus and cerebral cortex.

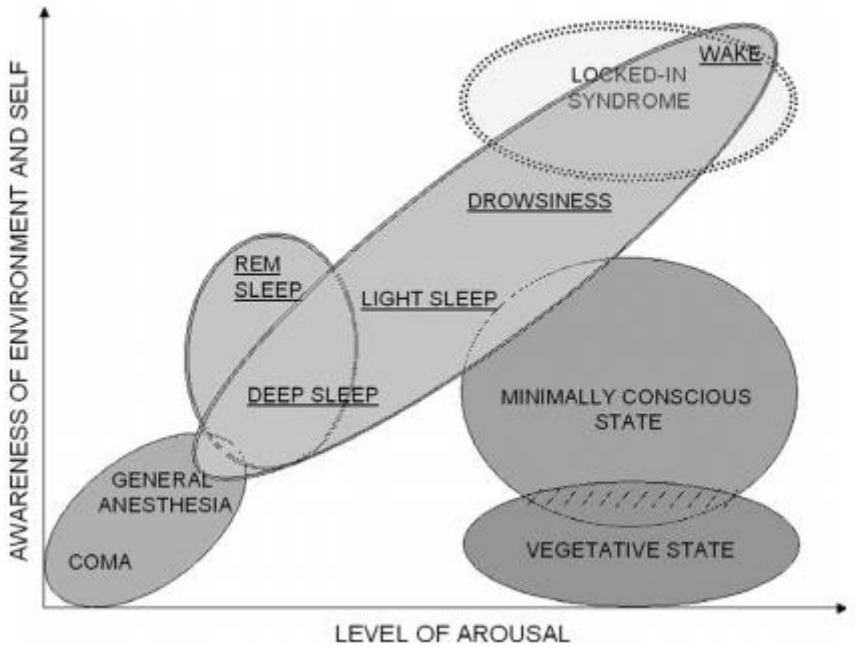

Fig. 2.1: Arousal-Awareness plane: VS and MCS are in a region with high arousal (between drowsiness and wake) and low awareness [51].

Coma is also defined as an "unarousable unresponsiveness state" and can be provoked by intoxications (drugs, alcohol, toxins), metabolic disorders (hypoglycemia, hyperglycemia, ketoacidosis) or damages and diseases of the central nervous system (stroke, head trauma, hypoxia) [52]. In most cases, coma is a transient condition that can last for a few weeks, ending when patients open the eyes, recovering the wakefulness. Although the re-opening of the eyes is sufficient to establish the end of this first acute phase, it is not enough for determining what the clinical outcome of patients is. In fact, it is necessary to understand whether patients recover the awareness and at what level. According to these criteria, the main post-comatose outcomes can be (Fig.2.2) [53]:

- Vegetative state;
- Minimally conscious state;
- Locked-in syndrome;
- Brain death
- Recovery.

Vegetative state (VS) is a syndrome characterized by the lack of awareness, and the lack of intentional and communicative responses [54]. VS is also called Unresponsive Wakeful Syndrome (UWS), since VS patients maintain a sleep-wake rhythm, they can smile, crying, grimacing and emit sounds, but without certain reasons [55].

VS patients can react to behavioral test, but they are unable to reproduce response in a continuative way. Also VS can be viewed as a transient state that, in turn, can evolve into:

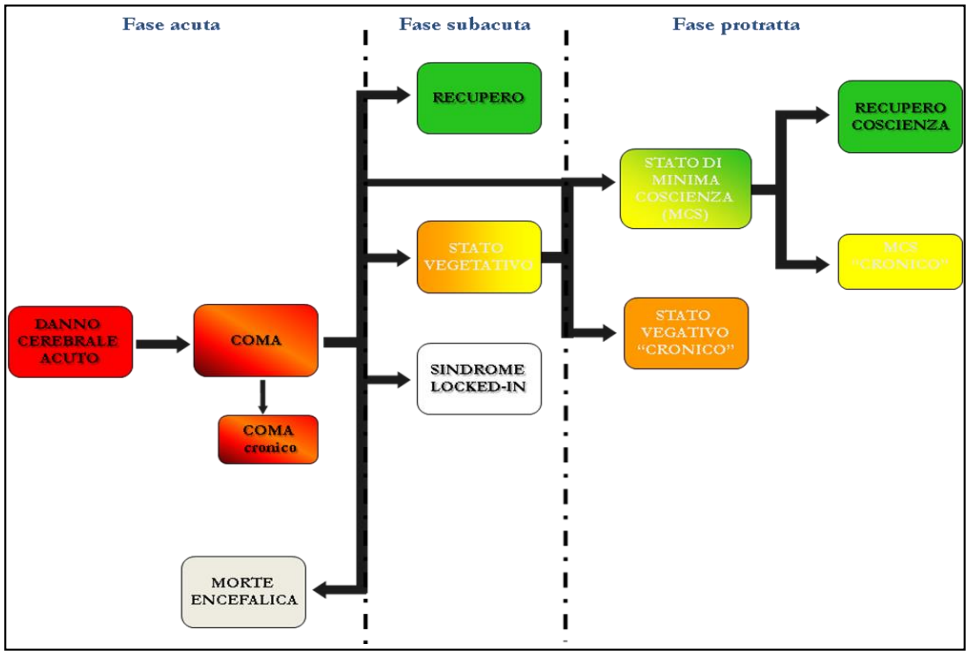

Fig. 2.2: Main post-comatose outcome [53].

- Persistent (or continuing) vegetative state (PVS);
- Minimally conscious state (MCS).

PVS concerns patients that after coma do not recover the awareness within 30 days. The etiology is variable and can be the result of: traumatic brain injury (TBJ), ischemia, hypoxia and encephalitis. The prognosis depends on the causes of the brain damage. In case of TBJ there are no certain indicators of the negative outcome before a period of 12 months after the injury. In fact, the percentage of PVS patients that recovered after this period is very low (around 1.6 % [56]). In case of hypoxic-ischemic attack, several indicators for a reliable prognosis may already exist in the first weeks after the injury. Of course, a confirmation of the persistence of VS is necessary after 3-6 months [56]. In that period, the neurologist has to identify the absence of reproducible, intentional or voluntary responses after the administration of visual, acoustic, tactile and noxious stimuli. Moreover, there should be also a lack in language understanding. PVS is characterized by an irregular but cyclic sleep-wake rhythm, in contrast to patients in coma that maintain their eyes closed [57]. For PVS patients, the probability to get out of the coma decreases with time and after a year patients who regained consciousness may present significant disabilities in proportion to the time spent in such condition.

MCS, instead, is characterized by a partial recover of the awareness. There may be some cognitive-mediated behaviors, inconsistent in functional terms, but maintained long enough to be distinguished from simple reflex responses [58]. These patients may experience smiles or other responses after the administration of external stimuli (audio-visual, emotional and linguistic stimuli), suggesting their ability to interact with the surrounding environment [55]. MCS is often considered as an intermediate state between VS and the presence of consciousness. MCS can be further divided into 2 states: Chronic MCS and Acute MCS. MCS is chronic when no further improvements occur in time, while is

acute when there are improvements over time that may also lead to a substantial recovery of consciousness.

For the diagnosis of the MCS, it is necessary to demonstrate a persistent presence of the awareness, proving that the ability to perform simple commands (i.e. to respond to verbal and emotive stimuli) is not due to simple a reflex activity.

Locked-in syndrome, describe by Plum and Posner in 1966 [52] is a particular condition characterized by the presence of awareness and wakefulness. However, patients are not able to move due to the complete paralysis of all the voluntary muscles of body. This results in a quadriplegia with inability to speak, but with preserved cognitive functions. The communication occur only with the codification of eye movements (i.e. eye-blinking), since their efferent pathways are not compromised by the brain damages. It is not uncommon for patients with locked-in syndrome to be classified as VS patients.

Nowadays, there are different diagnostic methods for DOC and for the differential diagnosis between VS and MCS. The most important are:

- The use of neurobehavioral scales;
- Neuroimaging techniques;
- Electrophysiological studies with the support of vital signs monitoring.

In the next sections, these techniques will be briefly illustrated, focusing more on the use of neurobehavioral scales. However, all of these methods present intrinsic problems and disadvantages. The diagnosis based on neurobehavioral scales is a practical and reproducible method, making it widely used in the clinical practice. However, it relies on the subjective perception of the clinicians and is mainly based on the experience of the examiner that therefore might provide biased evaluations. Neuroimaging techniques are very precise, providing very accurate information about the physiological process underlying the disorder (often in real-time). However they cannot be used regularly due to the high costs of the technologies. Finally, electrophysiological studies are reproducible, but sometimes can lead to contradictory results that lead to confusion about the basic concepts of consciousness, rather than helping the clinical diagnosis. Moreover, the presence of cumbersome devices and sensors attached to the patient's body may be a further disadvantage of these techniques.

2.2 Differential diagnosis: vegetative state and minimally conscious state

As mentioned above, it is easy to understand that the differential diagnosis between VS and MCS can be extremely hard and prone to a high rate of misdiagnosis. This differential diagnosis is mainly performed by means of neurobehavioral scale, with the support of neurophysiological studies (EEG, evoked potentials) and functional neuroimaging techniques (SPECT, PET, fMRI). However, the latter techniques are mainly used in the research field, due to the high costs.

The use of neurobehavioral scales is based on the clinical observation performed by trained observers (neurologists, therapists, etc.) and allows not only a classification of the clinical status of patients, but

also the quantification of the behavioral changes and evolutions. The most widely used scales are: the Glasgow Coma Scale (GCS) and the Coma Recovery Scaler - Revised (CRS-R) [59,60,61,62], that will be described in the next section. Other neurobehavioral scales are: the Sensory Modality Assessment and Rehabilitation Technique (SMART) [63], the Disorders of Consciousness Scale (DOCS) [64,65], the Level of Cognitive Functions (LCF), the Disability Rating Scale (DRS) and the Bartel's index [66]. However, these methods will not be considered in this chapter.

A key role to plan therapy and rehabilitation treatments for DOC patients, is played by the first diagnosis after the end of coma [67]. In fact, MCS patients who undergo a well-defined rehabilitation treatment are more prone to a better consciousness recovery than VS patients. However, as stated above, an important problem is the high rate of misdiagnosis between VS and MCS. Childs et al., 1993 [68] investigated this misdiagnosis rate in DOC patients: of 49 recruited DOC patients (with different etiology), 18 of them (around 37%) were incorrectly classified. Another study by Andrews et al., 1996 [69], showed that the error could be higher, around 43%. A more recent study, confirms these rates, with a misdiagnosis rate around 41% [70]. Of course, a correct diagnosis of DOC patients may influence both public expenditure and private health costs [58]. Despite the number patients with severe brain injury and low consciousness levels is negligible if compared to other neurological diseases (between 56 and 70 cases per million), the impact on the economy is very important: the cost per person during the course of these conditions may vary from 600000$ up to 1785000$.

2.2.1 Neurobehavioral scales

The classification of the clinical status is measured by means of different neurobehavioral scales: the simplified scales, the levels scales and the score scales. Among the score scales, the most important are the GCS and the CRS-R.

Glasgow Coma Scale

The GCS was the first score scale, introduced in 1974 by Teasdale and Jennet [59]. This scale, assigns a grade to the consciousness alteration on the basis of the patient's responses to external stimuli (noxious, verbal, etc.). The overall score ranges from 3 (deep coma) to 15 (patient is awake and conscious), as reported in Fig. 2.3.

This scale evaluates the opening of the eyes (assignig a score between 1 and 4) the motor responses (with a score between 1 and 6) and the verbal response (with a score beteween 1 and 5). Patients with scores ≤ 8 (for instance, Eyes opening = 1, Motor response = 5 and Verbal response = 2) are in coma and are unable to open the eyes after noxious and verbal stimuli [60]

Glasgow Coma Scale

BEHAVIOR	RESPONSE	SCORE
Eye opening response	Spontaneously	4
	To speech	3
	To pain	2
	No response	1
Best verbal response	Oriented to time, place, and person	5
	Confused	4
	Inappropriate words	3
	Incomprehensible sounds	2
	No response	1
Best motor response	Obeys commands	6
	Moves to localized pain	5
	Flexion withdrawal from pain	4
	Abnormal flexion (decorticate)	3
	Abnormal extension (decerebrate)	2
	No response	1
Total score:	*Best response*	15
	Comatose client	8 or less
	Totally unresponsive	3

Fig. 2.3: Glasgow Coma Scale [59].

Coma Recovery Scale - Revised

Given the high false negative and positive rates, Giacino et al., 2004 [61,62] restructured the JFK Coma Recovery Scale, firstly proposed in 1991. The aim of this scale is to provide a standardized method for the differential diagnosis of between VS and MCS, for the prognostic evaluation and for planning the therapeutic treatments in DOC patients. CRS-R seems to be the most promising tool for the differential diagnosis of VS-MCS and it is the most widely used scale in the clinical routine. The aim of the CRS-R is to quantify the auditory, visual, motor, oro-motor, communicative and arousal functions of DOC patients. This scale is divided into 29 items grouped into 6 sub-scales, each one relatives to the aforementioned functions (Fig. 2.4). For each sub-scale, the observer provides a score on the basis of the presence or absence of specific responses after the administration of standardized stimuli. The lowest score of each subscale suggests a reflex activity, while the highest scores are index of cognitive-mediated responses.

APPENDIX

JFK COMA RECOVERY SCALE-REVISED © 2004
VERSIONE ITALIANA © 2007

Scheda di registrazione

Questa scheda dovrebbe essere utilizzata solo insieme con le "Linee guida per le modalità di impiego e di determinazione del punteggio della CRS-R" che forniscono le istruzioni per la somministrazione standardizzata della scala.

Paziente:	Diagnosi:	Eziologia:
Data di insorgenza:	Data di ricovero:	

		Data															
		Settimana	Ric	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Scala per la funzione uditiva																		
4	Movimenti consistenti su ordine*																	
3	Movimenti riproducibili su ordine*																	
2	Localizzazione del suono																	
1	Reazione di sussulto uditivo																	
0	Nessuna risposta																	
Scala per la funzione visiva																		
5	Riconoscimento dell'oggetto*																	
4	Localizzazione dell'oggetto: raggiungimento*																	
3	Inseguimento visivo*																	
2	Fissazione*																	
1	Reazione di sussulto visivo																	
0	Nessuna risposta																	
Scala per la funzione motoria																		
6	Uso funzionale dell'oggetto†																	
5	Risposte motorie automatiche*																	
4	Manipolazione degli oggetti*																	
3	Localizzazione dello stimolo nocicettivo*																	
2	Allontanamento in flessione																	
1	Postura anomala																	
0	Nessuna risposta/flaccidità																	
Scala per la funzione motoria orale/verbale																		
3	Verbalizzazione comprensibile*																	
2	Vocalizzazione/movimenti orali																	
1	Movimenti orali riflessi																	
0	Nessuna risposta																	
Scala per la comunicazione																		
2	Funzionale: Appropriata†																	
1	Non funzionale: intenzionale*																	
0	Nessuna risposta																	
Scala per la vigilanza																		
3	Attenzione																	
2	Apertura degli occhi senza stimolazione																	
1	Apertura degli occhi con stimolazione																	
0	Non risvegliabile																	
PUNTEGGIO TOTALE																		

* Indica Stato di Minima Coscienza.
† Indica emergenza dallo Stato di Minima Coscienza.

Fig. 2.4: Italian version of the JFK Coma Recovery Scale - Revised [62].

The auditory scale score is assigned observing movements of limbs, head and eyes related to a specific command (for instance looking at an object) or to a specific sound (for examples, external sound or clapping hands). Concerning the visual scale, the examiner has to observe the ability of patients to discern and reach with arms some objects, as well as the ability to pursuit specific objects with gaze. The motor function is assessed administering a noxious stimulus to hands or feet. The presence or absence of flexor response determines the score. The oro-motor verbal scale evaluates the ability of patients to produce verbalizations, vocalizations or just simple reflex movements of the articulatory organs (jaw, lips and tongue). Similarly, the communicative scale observes the patient's ability to communicate in an appropriate and functional way (i.e. vocalizing or verbalizing) or by means of non-functional cues (for instance through the eye-blinking). Finally, the wakefulness scale, provides an overall score, about the patient's state during the whole examination.

The overall score ranges from 0 to 23 and the differential diagnosis between VS and MCS is performed according to the following criteria [61,62]:

- Vegetative state: auditory, motor and oro-motor verbal scores ≤ 2, or visual score = 1 and communicative score = 0;
- Minimally conscious state: auditory score between 3 and 4, or visual score between 2 and 5, or motor score between 3 and 5, or communicative score = 1;
- Minimally conscious state towards the emersion: motor score = 6, or communicative score = 2.

However, it is not always easy to classify the patient's response on the basis of these criteria. Often, there could be many misleading events (like the eye-blinking) where it is very difficult to relate the event itself to the administered stimulus and to understand its nature (reflex or cognitive-mediated).

Pain evaluation in DOC patients

Another important aspect to be considered in these patients is the perception of pain. In subjects that are able to communicate, pain can be assessed by means of self-assessment scales, like the Visual Analogic Scale (VAS). Through this method, subject reports an "amplitude" of the perceived pain on a graduated line of length equal to 10 cm [71]. However, this evaluation is not applicable to DOC patients that are unable to communicate. In these patients, as in other severe dementia states, pain can be assessed through scales that consider vital signs (heart rate, respiratory rate, etc.) as well as physiognomic parameters (facial expression, postures and body language).

The Nociception Coma Scale (NCS) was the first proposed method for the assessment of pain in MCS and VS patients [72]. It relies on 4 different sub-scales that evaluate the motor, verbal, visual and facial responses. For each one of these scales, the score ranges between 0 (null response) and 3 (maximum response), for an overall score between 0 and 12 points. The evaluation is performed during the baseline and after the administration of noxious and non-noxious stimuli.

One of the most widespread tools to assess pain in DOC patients is the Pain Assessment in Advanced Dementia Scale (PAINAD) [73]. The observer gives a score from 0 (normal) to 2 (severe) to each of the 5 subsections: respiratory rhythm, vocalization, facial expressions, body language and consolation. The overall score varies between 0 and 10 and through this score it is possible to have several pain levels:

- Slight pain: score between 1 and 3;
- Moderate pain: score between 4 and 6;
- Strong pain: score between 7 and 10.

However, the interpretation of these cues could be very subjective, leading to a difficult and biased assessment.

2.2.2 Neuroimaging techniques

As a result of what discussed above, it is easy to observe that the differential diagnosis between VS and MCS still relies on the ability and experience of the clinicians, resulting very subjective. For this reason, many studies tried to elucidate the dynamic underlying neuronal processes in DOC patients by means of neuroimaging techniques.

Owen et al., 2002 [57] investigated the cerebral regions involved in facial recognition and speech perception in three PVS patients, using PET. Areas involved in visual recognition are those of the fusiform right gyrus, and areas responsible of speech perception and recognition are those of the superior temporal gyrus, both included in the temporal lobe. In two of the three VS patients the responses were comparable with control subjects. These results led to the conclusion that the PVS diagnosis was wrong (hypothesis supported by the fact that one of these patients reemerged from PVS) or that, despite the differential diagnosis performed with the CRS-R led to similar scores, each patient reacts in a very different way after the administration of external stimuli. However, other studies performed through the use of fludeoxyglucose as PET marker [74] demonstrated that the amount of glucose used in VS patients both at regional and global level, is lower than that used in age-matched healthy control (HC) subjects.

Other studies [75,76] investigated the brain responses to speech and voice stimuli in MCS patients by means of fMRI [75] and PET [76]. In particular, an activation of the amygdala (the area responsible of the integration of superior motor processes, like emotions) was found after listening to a familiar voice (mother's voice). Boyle et al., 2004 [77], demonstrated that MCS patients and HC subjects had similar responses to acoustic stimuli, by using PET. In fact, HC subjects highlighted the activation of the Brodmann's areas 41, 42 and 22 (i.e. the fusiform right gyrus and superior temporal gyrus), similar to those encountered in MCS patients. In contrast VS patients did not show any activation of area 22. This could be due to the different level of consciousness that leads to an ability to process more detailed and complex information.

2.2.3 Electrophysiological studies and vital signs monitoring

Other diagnostic investigations in DOC patients, rely on the electrophysiological techniques. In particular, the most widely used techniques are the Electroencephalography (EEG) and the Event-Related Potentials (ERP). An ERP is a neurophysiological examination that studies the response of the nervous system to a sensory stimulus, analyzing the afferent nervous pathways. These responses can be the result of a thought or a perception and are measured through an EEG. Since the acquired response is the sum of the activation of different brain areas, the process is repeated several times and then the results are averaged, in order to delete the background brain activity. In general the type of ERP is identified with a letter (N - negative or P - positive) that expresses the polarity and is followed by the typical latency after a stimulus (for instance: P400 is a positive peak that occurs 400 ms after the stimulus).

EEG studies were performed to monitor the brain activities in DOC patients. In [78], the main differences between these patients and HC subjects were investigated studying the different EEG waves:

- Alpha waves (frequency range: 8-13 Hz, amplitude: 20-200 μV), typical of quiet conditions and wakefulness with closed eyes;
- Beta waves (frequency range: 13.5-30 Hz, mean amplitude 19 μV), intense brain activities;
- Theta waves (frequency range: 4-7.5 Hz, mean amplitude 75 μV), typical of adults with strong emotional stress, metabolic disorders and *medulla oblongata* lesions;
- Delta waves (frequency range: 0.5-4 Hz, amplitude: 1-200 μV), typical of deep sleep.

In general the EEG signals of VS patients have a slowdown of theta and/or delta waves with an attenuation of signals proportional to the extent of brain damage. It is not uncommon to find patterns typical of epilepsy.

A case study by Coleman et al., 2009 [79], shown an EEG pattern with a prevalence of delta and theta waves. In general, the EEG power spectral density of MCS patients is characterized by a decrease of the power content in particular frequency ranges (> 20 Hz) typical of brain activities during the wakefulness.

Vital signs monitoring

In recent years, many studies proposed to investigate vital signs, in particular the heart rate (HR), in DOC patients, in order to monitor possible reactions in situations of emotional stress or after the administration of exogenous stimuli. The heart rate monitoring, and in particular the study of its variations (Heart Rate Variability - HRV) is a good index of the autonomic nervous system activity.

Riganello et al., 2010 [80] studied VS patients during the listening to various music samples, selected to provoke different emotions. The HR was acquired through photopletysmography during the task, comparing HC subjects and VS patients. Moreover, HC subjects self-assessed their emotional state during the listening task (choosing between positive and negative emotion), in order to label the music

sample with a particular emotional flag. Results showed that for both groups the HR variations are similar, according to the aroused emotion.

Flotta et al., 2013 [81] exploited the idea to study HR in DOC patients, proposing an integrated system for storage, processing and analysis of data coming from VS and MCS patients. The acquisition of signals and data during the monitoring of these patients, is an important process that allowed creating a database. This dataset is important to help clinicians during the diagnosis phase, being a support for the decision of the best rehabilitation process for each patient.

3. State of the art

3.1 Acoustical analysis of patients with Parkinson's disease

Idiopathic Parkinson's disease is associated with a wide range of motor (tremor, stiffness, bradykinesia, postural instability) and non-motor (depression, cognitive impairments, sleep and mood disorders) symptoms that significantly reduce the quality of life of patients [4,5]. Impairments of voice and speech (hypokinetic dysarthria) are among the main signs of the disease and can affect about 70% of PD patients [29].

As already introduced in Chapter 1, PD patients may present alterations related to all speech dimensions (i.e. the four subsystems of speech production): respiration, phonation, articulation and prosody. Therefore, they might exhibit reduced variability of pitch and loudness, hoarseness, reduced stress, imprecise consonant articulation, unorthodox speech silence and speech rate alterations [30].

The non-invasiveness of acoustic measures, along with the debut of speech impairments in the early stage of the disease [82], has made this field very attractive towards researchers. Many authors tried to detect features of voice and speech belonging to the aforementioned speech dimensions (in particular phonation, articulation and prosody) that could discriminate PD patients from healthy subjects. Such features could provide an aid to an early diagnosis of PD, but could also help tracking the disease progression and testing the efficacy of drug therapy.

Phonation is the production of sound at the level of the larynx due to the vibration of vocal folds. The most widespread task for studying this speech dimension is based on the acoustical analysis of a sustained vowel emitted by the patient at comfortable pitch and energy. Rusz et al. [83,84] showed that significant differences exist between PD patients and healthy subjects concerning some acoustic parameters: jitter, shimmer, HNR (Harmonics to Noise Ratio) and PPE (Pitch Period Entropy, a parameter related to the inefficiency of voice frequency control [85]). Specifically HNR, a parameter related to voice roughness and breathiness, was proven useful by Tsanas et al. [86] whose results confirm that PD patients tend to have a harsher voice as compared to healthy subjects.

Articulation includes the set of movements of the articulatory organs (tongue, lips and jaw) that, by modifying the vocal tract resonant cavities, allow modelling the sound coming out from the vocal folds. Thus, most of the studies on speech articulation are based on the first two formant frequencies F1 and F2. Recently, Rusz et al. [87] showed that articulatory impairments in PD patients can be detected from F2u where FKj is the K-th formant for vowel j, tVSA (triangular Vowel Space Area, defined as: abs(F1i * (F2a – F2u) + F1a * (F2u – F2i) + F1u * (F2i – F2a))/2), F2i /F2u, and VAI (Vowel Articulation Index, (F2i + F1a)/(F1i + F1u + F2u + F2a)). In particular, differences are enhanced by specific tasks such as sentence repetition, reading a passage and monologue. Walsh et al. [88] showed a reduction of the F2 transition during the emission of a diphthong within a sentence in PD patients. This alteration reflects the undershooting of tongue and/or jaw motion. Other studies

[89,90,91] compared different acoustic parameters related to articulation, highlighting that VAI is more accurate than VSA in discriminating PD from healthy controls. Moreover, Sapir et al. [91] proposed a novel parameter FCR (Formant Centralization Ratio) and demonstrated that FCR and F2i/F2u are not gender-sensitive.

These results on speech articulation highlight the hypokinetic characteristics of speech impairments typical of PD, since the reduction of articulatory movements is reflected in the alterations of parameters derived from the formant frequencies.

One of the most common speech impairments in PD is dysprosody that includes alterations of rhythm and speed of speech, articulation, speech/pause ratio, pitch intensity and its variations [30]. Many authors carried out studies on prosodic patterns of voice in PD patients, assessing parameters related to speech rate. Skodda et al. [92] found a speech rate variation (number of syllables per second) closely related to the progression of the disease. This modification is characterized by articulatory acceleration in the early stages of the disease, followed by slowdown in advanced stages. Speech rate measures were also used to test the dopaminergic therapy effects on Parkinsonian voice. However, no significant differences were found during this pharmacological treatment [93].

Another widespread task used to evaluate speech rhythm disorders is the syllable repetition task (oral diadochokinetic test). It was shown that PD patients tend to have less control on rhythm stability during this task, with a tendency to increase the pace of repetition. These results reflect a dysfunction at the level of basal ganglia that control the temporal regulation of a motor sequence [94]. Rhythm alterations derived from the oral diadochokinetic test were also found by Rusz et al. [84], who found that PD patients have fewer syllables per second with respect to HC subjects. Furthermore, PD patients showed a reduction in fundamental frequency (F0) modulation and intensity and an increase of the number of pauses per second within a monologue or when reading a text. A reduction of F0 variability was also found in other works [95], as well as a reduction of pauses within polysyllabic words.

These results demonstrate that rhythm variations, pause alterations and the reduction of F0 variability are relevant parameters to detect dysprosody patterns in PD.

Other tests consist in the repetition of a short sentence, as reported in [87,88], although these studies considered only the articulatory dimension of speech.

Most of the works on acoustic analysis in PD patients make use of freely available software tools like PRAAT [96] or commercial tools like the Multi Dimensional Voice Program (MDVP by Kay Elemetrics Corp. [97]).

However, these tools are not provided with a pre-processing step that allows analyzing different parts of an audio recording in a completely automatic way. Specifically, as said above, the repetition of a standardized sentence is a commonly performed test in PD patients. With MDVP or PRAAT a manual selection of each sentence from the whole recording has to be made first. Then the tool is applied sentence by sentence for the estimation of the acoustic parameters. Therefore the automation of

manual segmentation would be desirable, being a time-consuming and operator-dependent step, especially in the case of long recordings.

Few attempts have been made in order to solve this technological challenge for the acoustic analysis in Parkinson's disease. Cmejla et al., 2013 [98] proposed a Bayesian autoregressive change point detector (BACD) for the evaluation of speech disfluency and articulation impairment. Tsanas et al. [99] used the AHTD (At Home Testing Device by Intel Corp. [100]) for tele-monitoring PD patients, in order to track the disease progression at home. However, the audio processing (concerning sustained vowels only) is made offline, as the voice records are sent to a server located in the medical center. Moreover, the analysis of sustained vowels, albeit providing useful parameters for the distinction between PD and healthy subjects, does not allow studying articulatory and prosodic alterations.

Therefore it is relevant to develop a robust method for a complete automation of the process. This would be essential also for the implementation of an analysis system to be used at home for monitoring and rehabilitation of PD patients. In particular, in the framework of speech therapy it is of basic importance for the patient to get an immediate feedback from the system in order to improve articulation.

3.2 Kinematic analysis of articulatory movements

3.2.1 Methods for articulatory movements analysis

In the past decades several techniques were proposed and used for studying the movements of the articulators (lips, tongue and jaw): x-ray imaging (cineradiography, x-ray microbeam system), magnetic resonance imaging (MRI), ultrasound technique, electromagnetic articulography (EMA) and optoelectronic systems (passive and active) [101]. Applications of these methods may include: the study of speech disorders in neurological illnesses (Parkinson's disease, amyotrophic lateral sclerosis, etc.) by means of optoelectronic [88,102], electromagnetic [103] and x-ray microbeam systems [104]; the use of the EMA to estimate the parameters of an articulatory model [105]; the study of tongue movements for speech therapy applications [106,107] by means of electromagnetic techniques.

However, the main disadvantage of these methods is related to the high costs that limit their use only in specialized laboratories. To our knowledge one of the few attempts to go beyond this limit is presented in the work of Feng et al., 2014 [108] where the performance of a system composed by 2 consumer-grade cameras in conjunction with a motion tracking software, are compared against an optoelectronic technique to track lips and jaw movements. However, the use of markerless systems (as well as low cost) would be desirable, in order to reduce the preparation time of the experiment and the discomfort for patients.

The spread of 3D low-cost sensors (like Microsoft Kinect, Asus Xtion, Primesense Carmine, Creative Senz3D, etc.) that enables providing 3D information of the observed scene, could help us for these

purposes, in order to extract trajectories and kinematic parameters in the 3D space, with the possibility to analyze some fundamental articulatory parameters like lip protrusion. At date, no previous work has still considered the use of 3D depth sensors for studying articulatory movements. Moreover, these devices could imply beneficial applications for speech therapy since most of the tasks consist in facial and articulatory movements and a feedback on how these movements are carried out is fundamental [109].

3.2.2 Applications to patients with Parkinson's disease and other neurological disorders

In most of PD patients, dysarthria is usually "hypokinetic". This term refers to the reduced range of movements involved in speech production. In fact, hypokinetic dysarthria is characterized by reduced peak velocities and displacements of the articulators during speech movements [110]. This articulatory undershoot may lead to alterations in some acoustical parameters that are typical of articulation, such as the reduction of the second formant transition slope or a reduction of the tVSA [89,90,91].

Several approaches and methods have been implemented to describe the kinematic characteristics of the articulators in hypokinetic dysarthria associated with PD. Most studies [88,103,104,111,112,113,114] pointed out a reduction in terms of velocity and range of movements of the articulators, although results of unimpaired articulatory movements in Parkinsonian patients were also presented [115].

One of the most widespread tasks for kinematic analysis of the articulators is the production of syllables. Through this test several authors [111,113,114] found a reduction of the displacement and of the peak velocities of lower lip and jaw during the opening and closing of the mouth. A slowdown of jaw and lower lip movements was demonstrated also by Forrest et al., 1989 [112]; on the other hand they found an increase of the closing velocity of the lower lip in PD patients. This result might reflect an alteration of the motion control due to the severity of the dysarthria. In fact, as mentioned above, although the majority of PD patients suffers from hypokinetic dysarthria, a small percentage experiences symptoms of hyperkinetic dysarthria, whose occurrence could be related to the prolonged administration of drug therapies that causes the onset of involuntary movements (dyskinesia) [110].

Other works on the articulatory kinematics in PD patients with hypokinetic dysarthria focused on tongue movements during speech. Yunusova et al., 2008 [104] studied movements of tongue, jaw and lower lip during the pronunciation of words, finding that tongue movements of PD patients could be more discriminative as compared to control subjects, although a reduction in lips and jaw kinematics still exists. Wong et al., 2012 [103] studied tongue movements during a sentence repetition task, in order to discriminate between dysartrhic and non-dysarthric PD patients. Unlike most of the previous findings, the authors demonstrated that dysarthric PD patients exhibited wider ranges of movements with an increase of peak velocities and accelerations.

Although it is well accepted that PD patients with hypokinetic dysarthria exhibit a reduced articulatory kinematic, some conflicting results were found. Walsh et al., 2012 [88] tried to elucidate this point

studying jaw and lower lip movements during the opening and closing gestures in case of bilabial consonants. The authors demonstrated that PD patients exhibited a reduced articulatory kinematic, highlighted by reduced velocities of jaw and lower lip. These results support the hypothesis that a "downscaling" in speech production occurs in PD patients with hypokinetic dysarthria.

Another important point concerns the implemented methodologies. As stated in the previous section, the kinematic analysis of the articulators was performed through several motion capture technologies. The most important are: optoelectronic systems [88,113,115], electromagnetic articulography (EMA) [103], X-ray techniques [104] and Magnetic Resonance Imaging (MRI) [101]. However, all these techniques are quite expensive and their use is limited to research within highly specialized laboratories. Moreover, some of the most widely used techniques (optoelectronic systems and EMA above all) are marker-based and need quite long preparation protocols in order to achieve good results.

Despite hypokinetic dysarthria may afflict a large number of PD patients, today only a small percentage of PD patients undergoes speech therapy with specific protocols aimed at increasing their articulatory movements. This is due to several factors:

- during group sessions, the speech therapist has big difficulties to pay the same attention to each patient in order to evaluate the performance during exercises and provide an immediate feedback;
- several patients with hypokinetic dysarthria due to neurodegenerative diseases are elderly people that often have difficulties in moving to specialized centers;
- patients should perform the exercises also at home. However, they do not because they lack the presence of the therapist.

In the last years, several results in monitoring and rehabilitation of dysarthria in home environment were achieved with the help of acoustical analysis [99,100]. Instead, few results were obtained for the automation of exercises that involve facial muscles like those commonly performed in speech therapy protocols. The main reasons are related to the high costs of the methods used to study articulatory movements.

3.3 Automatic analysis of facial expressions

3.3.1 Methodologies for facial expressions recognition

The communication of the emotional state of a person consists of two processes: the expression, where a specific set of features is produced, and the perception where an observer receives these features and can infer the emotional state of an individual. In more detail, the expression process is the externalization of the emotional state by means of physical signs (also called "expressed cues"), such as facial expressions, vocal variations and body gestures [116]. Thus, facial expressions play a fundamental role in this process.

The development of automatic systems for facial expression recognition has become one of the most popular topic in computer science and artificial intelligence, founding a new research field called "affective computing" [117]. Examples of applications are: robotics, Human-Computer Interface, video games and entertainment, psychiatry [118], neurology [119], automotive [120], security, etc. These studies are the result of decades of research in psychology that over the 20th century tried to describe in an objective manner facial expressions related to emotional states. This led to the definition of standards for the description and the recognition of facial expressions from the single facial muscles movements. One of the most popular standards is the Facial Action Coding System (FACS), developed by Ekman and Friesen in 1978 [117,121]. FACS allows decomposing a facial expression into specific Action Units (AU). An AU is a change that occurs in the face caused by one or more facial muscles; it can be viewed as the minimum change that in conjunction with other AUs causes a facial expression. FACS decomposes facial expressions into 44 AUs, plus other codes for head and eyes movements. Some of the existing AUs are:

- AU1 - Inner brow raiser, caused by Frontalis (pars medialis) muscle;
- AU2 - Outer brow raiser, caused by Frontalis (pars lateralis) muscle;
- AU4 - Brow lowerer, caused by Corrugator supercilii and depressor supercilii muscles;
- AU5 - Cheek raiser, caused by Orbicularis oculi (pars orbitalis) muscle;
- AU9 - Nose wrinkle, caused by Levator labii superioris alaeque nasi muscle;
- AU12 - Lip corner pull, caused by the zygomaticus major muscle;
- AU15 - Lip corner depressor, caused by the depressor angulis oris muscle;
- AU16 - Lower lip depressor, caused by the depressor labii inferioris muscle.

Thus, facial expressions can be obtained as a combination of several AUs. For instance, happiness is the combination of AU6 and AU12, while disgust is the combination of AU9, AU15 and AU16.

Until a couple of decades ago, facial expression recognition was performed only by the perceptual evaluation of trained raters, who based their work on standards like FACS. However, this field has significantly benefited from the latest improvements in computer vision and machine learning since the 90s to date. In general, an automatic facial expression recognizer is a system consisting of 3 parts [117,122]:

1. Face detection and tracking. In this first step, a face must be identified in the image or in a video frame. In case of video analysis the face must be tracked along the subsequent video frames, in order to keep track of facial variations in time.
2. Feature extraction. From the region of interest (ROI) detected in the previous step, some geometrical or appearance facial features are extracted. This step is essential in order to represent the face as a set of several components, thus focusing on those features that could bring information about expression variations. These features can be divided into two categories: permanent features (eyes, eyebrows, nose and lips) and transient features (facial lines and wrinkles that appear only in presence of particular expressions).

3. Expression recognition. After the face has been represented as a set of geometric and/or appearance features, these features are classified in order to predict the class they belong to (i.e., which expression or which combination of AUs). Usually, a classification algorithm is firstly trained on one or more databases of labeled expressions and then tested on facial expressions that do not belong to the training set.

Several approaches have been proposed to automatically recognize facial expressions. Usually, all of these methods follow the above scheme, but differ for the different algorithms used to address the three steps [117,122,123,124].

Concerning face detection and tracking, one of the most used algorithms is the Active Appearance Model (AAM) [122,125,126]. AAM builds statistical models of shape and appearance by means of the Principal Component Analysis (PCA) performed on a set of training faces [125,126]. Other efficient face detection and tracking algorithms used for facial expressions recognition are the Supervised Descending Method (SDM) [127], the constrained local model (CLM) [128] and the particle filtering [129].

Considering the facial features used to predict a facial expression, the proposed systems can be divided into two classes: systems that use geometrical (or shape) features and systems that use appearance features [124]. Most of the geometrical-based systems exploit the facial points provided by the face tracker, since the tracked landmarks describe the shape of permanent facial features (eyebrows, eyes, nose, lips, face contour, etc.) [117,122,123,130,131]. Appearance-based systems, instead, describe facial expressions using the information of the image texture extracted with descriptors such as local binary pattern (LBP), Gabor filters, histogram of oriented gradients (HoG), etc. [117,122,123]. The aim is to describe texture variations that occur in the face while performing an expression, trying to catch characteristic signs such wrinkle, furrows and facial lines.

Other important differences among the proposed approaches consist in the temporal representation of an expression that could be static or dynamic [117,122]. Static representation classifies facial expressions frame-by-frame; in contrast, dynamic (or spatio-temporal) representation considers the temporal dynamic of a facial expressions, working on time windows of frames. The latter approach allows identifying the usually known dynamics of facial expressions: onset-apex-offset [2015_Sariyanidi]. Of course, temporal or static representations give rise to different classification strategies. Commonly used classifier for facial expressions are: Support Vector Machine (SVM), Adaboost, Neural Networks, Hidden Markov Models (HMM), etc. In particular HMMs were often used when the temporal dynamic of the expression is required [122,123].

Another important difference among the proposed approach is the affective model. Some methods aim at classifying basic expressions [132,133], while other methods perform an AU recognition [118,131,134]. Moreover, depending on the databases used for training the classifiers, it is possible to recognize spontaneous or posed expressions. The available databases may include posed expressions [135,136,137], spontaneous expressions [138,139], or both [140,141].

For a more detailed list of different methods proposed and implemented in the past years, please refer to [117,122,123,124].

3.3.2 Applications to patients with Parkinson's disease and other neurological disorders

As described above, thanks to the perception process an observer can infer the emotional state of another person through the expressed facial features. However, if the pattern of expressed features is not unambiguous and does not vary among different emotional states, the underlying emotion cannot be reliably communicated and thus detected by the observer [116]. This is the case of Parkinsonian patients with facial hypomimia, where external observers may have difficulties in decoding the emotional state behind the "mask", with severe problems in social relationships [34,35]. Moreover, facial bradykinesia affects all the functions that involve orofacial movements with severe problems of speech, swallowing, drooling, etc.

Several works aim at detecting the most impaired expressions in PD patients. However, it is still unclear whether facial hypomimia is a pure motor disorder or is a secondary effect of an expression recognition impairment. Simons et al., 2004 [36] investigated facial expressivity in PD patients during several tasks: watching video clips, social interactions and expression posing. Results demonstrated that PD patients had a lower expressivity than HC subjects with problems in exhibiting some action units: AU4 (brow lowerer), AU9 (nose wrinkler), AU10 (upper lip raiser), AU1+2 (brow raiser) and AU6+12 (check raise plus lip corner pull). Moreover, PD patients had difficulties in the imitation of happiness, surprise and disgust. The reduction of the overall expressivity was also demonstrated by Bowers et al, 2006 [40], although the authors did not find that some expressions are more impaired than others. The analysis of AUs in PD patients was also performed by other studies [119,142]. In [119] the authors combined facial Electromyography (EMG) and classifiers for AUs detection, finding a reduced expressivity in PD patients, with an expressive reduction along with the increase of the disease severity [119]. Vinokurov et al., 2015 [142], instead, used the AUs detected by means of a 3D depth camera, in order to train an algorithm for the automatic identification of hypomimia in PD. Despite a reduction of facial expressivity was demonstrated for PD patients, no further information was provided about the most impaired expressions and/or the most impaired facial movements.

Another study [143] tried to elucidate the possible relationship between the reduced expressiveness in PD patients and the impairment in recognizing emotions (alexithymia). The authors demonstrated that PD patients had reduced facial expressiveness (both static and dynamic), with difficulties in acting particular expressions (happiness, surprise and sadness). Moreover, they reported an impairment in recognizing disgust, fear and anger. However, from this work it is still not understood if voluntary acted expressions are impaired in PD patients.

From these studies it is well accepted that PD patients had a reduced expressivity with respect to HC subjects. However, only few studies provided quantitative information about the impaired facial expressions and/or action units [36,143]. Moreover, despite the huge developments of affective

computing (as described in the previous section), the implementation of automatic methods for facial expression recognition is a fairly unexplored field in PD patients with hypomimia. Of the reported studies, only Wu et al., 2014 [119] made use of a fully automated facial expression recognizer. In particular, using appearance and geometrical facial features they were able to detect 11 AUs through a binary SVM for the recognition of each AU. Other studies [40,142] implemented semi-automatic methods. In [40] a frame differencing algorithm is used to quantify the expression change along the recorded videos; however, it required the manual selection of some facial landmarks by the user. In [142], the authors used a commercially available software called faceshift[1] [144] that in conjunction with consumer depth cameras (like Microsoft Kinect) is able to track 3D facial landmarks and drive an avatar. However, this algorithm is person-specific and requires a calibration for each subject.

Other applications in psychiatry and neurology [118] demonstrated the high potentials of the algorithms for facial expression recognition. Thus, we believe that these systems could be used to objectify facial hypomimia in PD patients, in order to extract objective measures of facial mimicry and help to settle the still open issues: is facial hypomimia a pure motor disorder? Are PD patients able to act voluntary facial expressions?

Moreover, the use of contactless video-based systems like those described in the previous section, can be an important means for analyzing facial expressions also in rehabilitation framework (in particular speech therapy), where patients would derive a definite advantage about a real-time feedback of the right facial expressions/movements to perform.

3.4 Contactless heart rate estimation

Heart rate (HR) estimation through contactless techniques has become a popular field in biomedical signal processing. The estimation of vital signs without any sensor attached to the patient's skin can bring important benefits in those cases where a prolonged monitoring is required (for instance in long-term bedridden patients or neonatal intensive care units [145]). Depending on the underlying physical principle, these methods can be classified into several classes [146,147]:

- Electromagnetic-based systems;
- Laser-based systems;
- Image-based systems, both thermal and visible imaging;

In general, an ideal contactless HR monitoring system, should be:

- fully automated;
- robust against motion artifacts;
- robust against light variations (especially in case of image-based systems);
- low cost.

[1] In 2015 Apple Inc. bought Faceshift GmbH. No updates are available on the website www.faceshift.com

In particular, the latter requirement could allow the spread of these systems in the everyday clinical practice.

On the basis of the above requirements, image-based systems gained great importance in the last 5 years. The physical principle of these methods is similar to that exploited by photoplethysmography (PPG). PPG is a non-invasive technique for sensing the cardiovascular pulse wave by means of transmitted/reflected light variations [148]. Pressure pulses coming from the heart cause volume changes also in the skin vessels. These changes could be detected by illuminating the skin with a light source and then measuring the transmitted or reflected light [148]. Light absorption in blood is mainly caused by hemoglobin, the protein responsible for the oxygen transportation in the circulatory system. PPG uses an external light source (usually red or infra-red light emitting diodes - LEDs); however, in recent years it was demonstrated that PPG measures can be extracted using ambient light as a source and digital cameras as sensors. In particular, most of the works [145,146,148,149] focused the attention on the face region. Among red (R), green (G) and blue (B) channels of a RGB color image, signals extracted from the G channel reported the highest amount of plethysmographic signal, because of the high absorption of ambient light by hemoglobin in that particular wavelength range (between 500 and 600 nm) [148,149]. Thus, by studying frame-by-frame skin color variations it is possible to estimate the fluctuations due to the skin perfusion, and thus to the heartbeat.

In order to overcome some problems such as the robustness against movements and illumination changes, several computer vision and signal processing techniques were used. In particular, most of the proposed methods used a face tracking algorithm in order to keep track of head and facial movements, reducing the analysis to a small facial ROI [146,148]. Moreover, several works used Blind Source Separation (BSS) techniques (in particular Independent Component Analysis - ICA), in order to filter out the noise components from the signals [146,148,150]. In the next sections, a brief overview of image-based systems for HR estimation is presented.

3.4.1 BSS-based methods for HR estimation

Blind Source Separation (BSS) indicates a set of techniques for recovering different source signals that mixed together gave rise to the observed signals, without any information about the mixing process. Among BSS methods, the most widely used for the image-based HR estimation is ICA [145,146,148]. In general, given a set of *m* observed signals $\mathbf{x}(t) = [x_1(t), x_2(t), \dots x_m(y)]^T$ and a set of *n* unknown source signals $\mathbf{s}(t) = [s_1(t), s_2(t), \dots s_n(t)]^T$, $\mathbf{x}(t)$ can be expressed as function of the underlying sources by the following equation:

$$\boldsymbol{x}(t) = \boldsymbol{f}\{\boldsymbol{s}(t)\} + \boldsymbol{n}(t) \tag{3.1}$$

where $\boldsymbol{f}$ can be any unknown function and $\boldsymbol{n}(t)$ is additive noise. However, without any prior assumption the problem cannot be fixed [150]. Thus, the aim of ICA is to obtain an estimation of the

mixing matrix (called unmixing matrix) representative of the underlying mixing process, in order to find an estimation of $\boldsymbol{s}$ given the observations $\boldsymbol{x}$. The assumptions of ICA are [148,150]:

- Linearity: the observed signals are obtained as linear combination of the unknown sources:

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) + \boldsymbol{n}(t) \tag{3.2}$$

  were $\boldsymbol{A}$ is the *n* x *m* mixing matrix;
- Noise term $\boldsymbol{n}(t)$ negligible: this hypothesis reduces the equation to:

$$\boldsymbol{x}(t) = \boldsymbol{A}\boldsymbol{s}(t) \tag{3.3}$$

- The number of unknown sources is less than or equal to the number of observed signals ($n \leq m$);
- Stationary mixing: the statistics of the mixing matrix $\boldsymbol{A}$ (and thus the physics of the mixing process) do not vary over the time;
- Statistical independence of the sources: the sources are assumed to be mutually independent.

Thus, the ICA problem is reduced to the estimation of an unmixing matrix $\boldsymbol{W}$ that approximates the inverse of the original mixing matrix $\boldsymbol{A}$:

$$\hat{\boldsymbol{s}}(t) = \boldsymbol{W}\boldsymbol{x}(t) \tag{3.4}$$

where $\hat{\boldsymbol{s}}(t)$ is an estimation of $\boldsymbol{s}(t)$. The estimation of the sources is performed through the central limit theorem: a sum of independent random variables is more Gaussian than the original variables. Thus, the observed signals $\boldsymbol{x}(t)$ should be more Gaussian than the estimated sources; in other words, $\boldsymbol{W}$ maximizes the non-Gaussianity of each source [150]. Different methods for the estimation of the unmixing matrix were proposed. The most important are: FastICA [151], Infomax [152] and JADE [153].

In case of image-based methods for HR estimation, the observed signals are the variations of the three color channels (R, G and B) of the RGB color video frames. Thus, considering the above assumptions, it is possible to estimate a maximum of 3 source signals [148].

In general, an ICA-based method for HR estimation consists of the following steps:

- Video acquisition;
- Face detection and ROI location;
- Pre-processing (usually composed by signal detrend, normalization and band-pass filtering);
- ICA to estimate the underlying sources starting from the pre-processed signals;
- post-processing (usually a band-pass filtering on the estimated sources);
- HR estimation from the estimated sources.

Concerning the acquisition step, most of the authors used common RGB webcams [145,146,148,154,155], although high speed cameras [156] and 5-channels color cameras (RGBCO, where C is cyan and O is orange) [157] were also implemented. In particular in the latter study, the authors demonstrated that the best results can be obtained using a combination of cyan, green and orange channels.

Once that the facial ROIs have been detected (forehead and cheeks are the most widely used ROIs), most of the works performed a spatial average within the ROI, over the image channel [145,146,148,155,156,158]. Thus, for each video frame a value for each color channel is obtained. Temporal variations of color channels are then detrended using a smoothness priors approach [159] and normalized by subtracting the mean value and dividing by the standard deviation of the signals [146,157,158]. Moreover, a band-pass filtering can be used to limit the frequency range of the signal to that typical of the heart rate (from around 0.6 Hz up to 4 Hz).

The most used ICA method is the Joint Approximate Diagonalization of Eigenmatrices (JADE) algorithm [153]. Another approach called FASTICA [151] is used by Lewandowska et al., 2011 [154]. Once the independent sources have been estimated, some authors performed another band-pass filtering, in the same frequency range of the heart beat (0.6-4 Hz, corresponding to 36-240 bpm) [146,158].

Finally, the filtered sources are used to estimate the HR. The most widely used methods are based on the highest peak of the Fast Fourier Transform (FFT) in the frequency range of interest (0.6-4 Hz) [148,156,157,158], or by computing the inter-beat interval [146], by peak-picking in the selected source.

An issue raised by most of the ICA-based works is the choice of the source to be considered for HR estimation. Poh et al., 2010 [148] stated that the second component usually contains the strongest pulse signal; this phenomenon might be related to the highest plethysmographic content in the second observed signal (i.e., the green channel of a RGB image) [148,149]. Another method proposed to identify the best source for HR estimation is to select the one with the highest spectral peak in the frequency range of interest [146]. Monkaresi et al., 2014 [160] tried to address this issue starting from the assumption that all the three components can contain useful information about the HR; thus, the selection of just one source (completely excluding the other two) should be a restricting choice. Through machine learning techniques (linear regression and k-nearest neighbor) the authors developed a method able to predict the best source for HR estimation, within a temporal window. This method is based on the selection of the source with the spectral content that minimizes the Root Mean Square Error (RMSE) between the HR estimation and HR reference value.

3.4.2 Other methods

Many other approaches that do not use BSS methods were proposed. However, for most of them the processing framework is the same of that reported in the previous section. The main goal is still the

filtering (and the amplification) of skin color variations due to the heartbeat, in order to predict the frequency at which these variations occur.

A very popular algorithm to amplify and reveal small temporal variations in videos, and thus color variations due to the skin perfusion, is the Eulerian Video Magnification (EVM) [161]. The processing steps of this algorithm are illustrated in Fig. 3.1. First, a spatial decomposition is applied to the video, in order to decompose each frame in different spatial frequency bands. These frequency bands are then filtered with a pass-band filter (the same for each band), that in case of HR estimation can be tuned to the frequency range of the heartbeat (0.4-4 Hz [161]). Then, each frequency band is amplified by a given amplification factor in order to enhance different details in different frequency bands. Finally, each filtered and amplified band is added back to the original signal, to build an output video where only events of interest are amplified.

Other image-based HR estimation methods are based on the wavelet transform, both continuous [162] and discrete [163] and on autoregressive (AR) models [164]. In particular, Tarassenko et al., 2014 [164] considered two ROIs: one on the face and another one on a fixed background (i.e., the wall). Temporal variations of the green channel of both ROIs are modelled with a 9-order AR model, within temporal windows of 15s of duration. Afterwards, poles that are common to both models are cancelled from the AR model of the facial ROI. This method is based on the assumption that the model built on the background describes illumination changes due only to the artificial light flickering and other noise factors. Thus, the "pole cancellation" procedure provides a new AR model with poles that describe only temporal variations due to the heart rate [164].

Xu et al., 2014 [165] exploited the Lambert-Beer law and the linear relationship between skin absorbance and hemoglobin absorbance to determine a pixel quotient in log space (a quantity related to the hemoglobin concentration and thus to blood concentration). Then, temporal variations of this quantity provide information about the frequency at which the blood concentration varies in time due to the heartbeat [165].

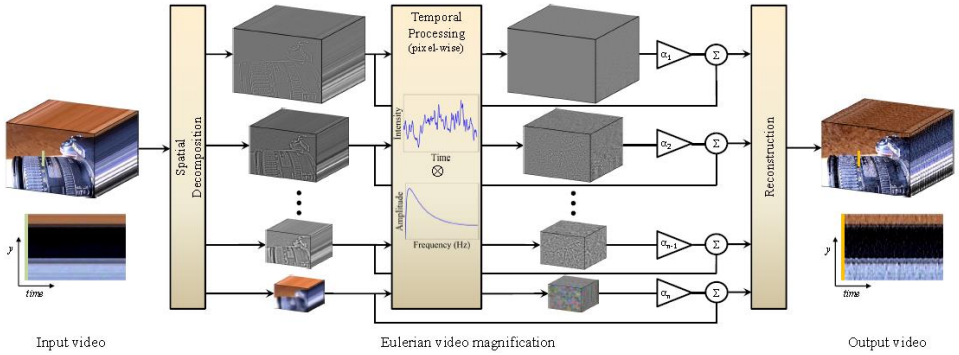

Fig. 3.1: Processing steps of the Eulerian Video Magnification algorithm [161].

PART II - MATERIALS AND METHODS

4. Acquisition Protocols

During this PhD project, the following collaborations were established:

- Unit of Neurology - Hospital "Nuovo San Giovanni di Dio" (Firenze) and *Associazione Italiana Parkinsoniani* (Firenze), for the recruitment of PD patients and HC subjects;
- "Villa delle Terme" rehabilitation center (Impruneta, Firenze) and "Don Carlo Gnocchi" Foundation IRCSS (Scandicci, Firenze) for the recruitment of DOC patients.

In case of PD patients, the aim was to extract acoustical and kinematic parameters related to speech disorders and facial expression impairments. This allows:

- Objectifying the perceptual analysis performed by clinicians;
- Developing a markerless system to monitor the disease progression, in particular the symptoms related to hypokinetic dysarthria and hypomimia;
- Developing a markerless system for speech therapy purposes that can be used in a domestic environment.

In case of DOC patients the aim was slightly different, since the project was focused on the assessment of reflex and cognitive responses to standardized stimuli, administered according to the CRS-R protocol. This allows:

- Objectifying the perceptual analysis performed by clinicians;
- Developing a markerless system for monitoring DOC patients;
- Assessing several reactions and aspects related to the perception of pain in patients unable to communicate.

Both case studies (PD and DOC patients) share the study of facial movements and facial expressions by means of video based techniques. However, the definition of two experimental protocols was necessary in order to meet the two different aims.

This chapter describes the two acquisition protocols as well as the two datasets of patients recruited throughout the project.

4.1 Audio-Video recordings of patients with Parkinson's disease

The definition of the audio-video acquisition protocol for PD patients comes from the first two item of the UPDRS part III: the speech task and the facial expressions task. During the speech task, patients have to talk as spontaneously as possible in a monologue or in an interview with the clinician. However, in order to standardize and make the acoustic and kinematic measurements reproducible, this task was revised (in accordance with expert clinicians) and divided into two parts:

- Subtask 1: repetition of a "voiced sentence" (/*Il bambino ama le aiuole della mamma*/) at least 10 times, as spontaneously as possible, at a comfortable loudness. Thanks to reliable voiced/unvoiced segmentation algorithms, it is possible to extract temporal parameters such as

sentence and pause duration, pauses inside a sentence and speech rate. Given the high prevalence of vowel sounds, the use of this sentence allows obtaining other clinically useful parameters (fundamental frequency, formant frequencies, jitter, shimmer, harmonic to noise ratio, etc.).

- Subtask 2: repetition of the syllable /pa/, at least 25 times at a rate as regular as possible. The voiced/unvoiced segmentation allows the extraction of temporal parameters to quantify rhythm impairments.

During the UPDRS facial expression task the patient is at rest in a seated position for at least 10 seconds. During that period, the following features are observed by the clinician: eye-blink frequency, loss of facial expressions, spontaneous smiling and lips opening. In order to perform experiments, I revised this task, in accordance with expert clinicians dividing it into three parts:

- Displaying of neutral expression for at least 10 seconds;
- Displaying of basic expressions (happiness, anger, disgust and sadness) upon request from the clinician;
- Displaying of basic expressions (happiness, anger, disgust and sadness) by imitating emotive faces shown on a screen (Fig. 4.1);

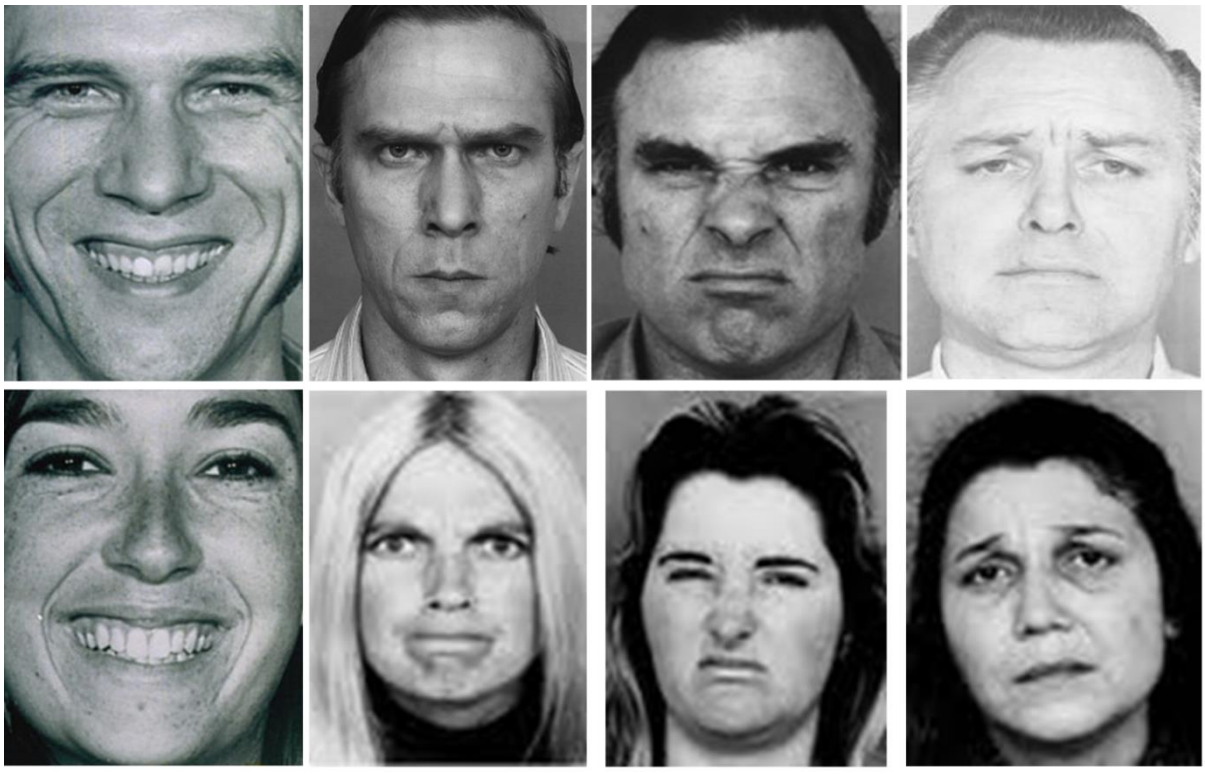

Fig. 4.1: Facial expression for the imitation task. From left to right: happy, anger, disgust and sadness, for men (upper row) and women (lower row).

Moreover, the assessment of differences between the emotions expressed by imitation an those expressed on request was performed, as many studies highlight that PD patients are impaired in recognizing emotions on other faces (Part I - sections 1.2.2 and 3.3.2).

The experiments were carried out in a quiet room of the "San Giovanni di Dio" Hospital, Firenze, Italy. Speech signals were recorded on a standard personal computer using Audacity software (version: 2.0.3) with a Shure SM58 microphone and a Tascam US-144 audio board. Signals were digitized at a sampling frequency Fs = 44.1 kHz. The microphone, fixed on a boom, was positioned at a distance of 5 cm from the subject's mouth. Participants were asked to avoid acceleration and slowdown of the articulatory velocity during the repetitions of syllables and sentences. Subjects' face was kept under constant and uniform illumination during the whole acquisition.

The subjects' face was recorded using the Microsoft Kinect for Windows sensor and was kept under constant and uniform illumination during the whole acquisition. The Microsoft Kinect is a structured light sensor that provides two video streams: a color stream (in the RGB color space), and a depth one where each pixel codes the distance of the points in the scene from the camera plane. This low-cost device allows markerless assessment of movements in the 3D space, overcoming the limits of conventional cameras. The Kinect sensor was placed in front of the subject's face at a distance between 0.5 and 0.7 m from the mouth and at a height close to that of the subject's eyes. This distance was chosen as a trade-off between the technical specifications provided by the manufacturer (in "near range" mode the minimum distance is 0.4 m [166]) and the need of having the subject's face as close as possible to the camera, in order to achieve the best accuracy in tracking the 3D facial points.

The resolution of both video streams was 640x480 pixels at 30 frames per second (fps). Color frames were recorded in 24-bit RGB images (8 bits per channel), while depth frames were recorded in 16-bit, 1 channel images. These features are the best trade-off, in terms of spatial and temporal resolution, provided by this kind of sensor to achieve good accuracies for tracking fast movements like those of the lips during the syllable repetition task. An overview of the experimental setting for audio-video acquisitions is reported in Fig. 4.2. Both streams were recorded and stored in *avi* files through the OpenNI (ver. 2.2) and OpenCV (ver. 2.4.9) libraries using a customized code written in C++ language. Despite both tasks (speech and facial expression task) were performed recording both video streams, for the analysis of facial expressions we used only the color stream, as explained in Part II, chapter 8.

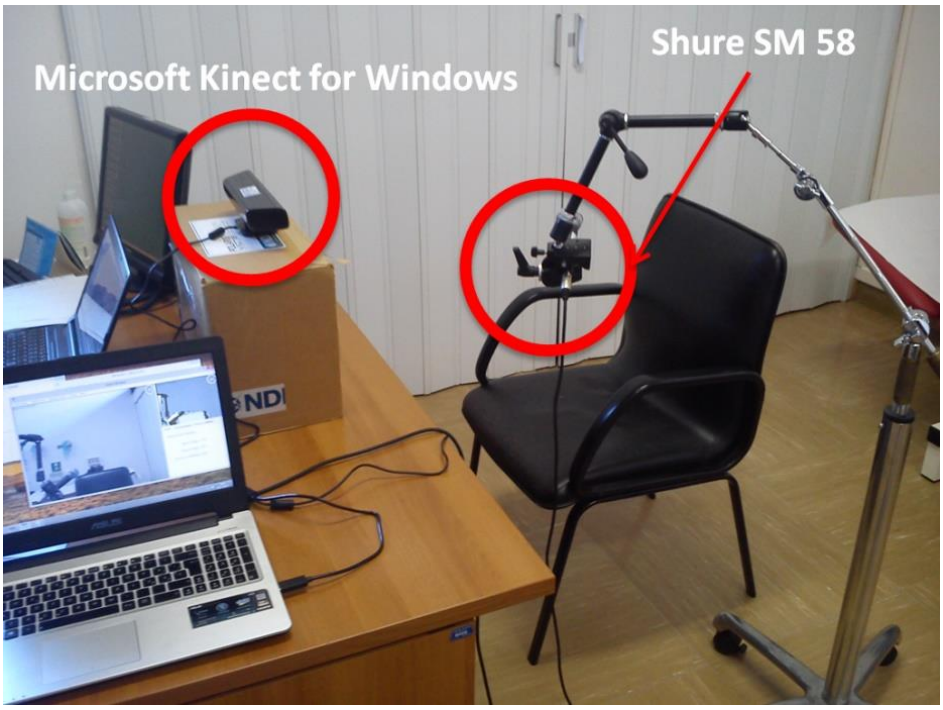

Fig. 4.2: the experimental set up for audio-video acquisitions in PD patients and HC subjects.

4.1.1 Dataset

During this PhD project the following subjects were recruited:

- 27 PD patients (18 male, 9 female), age: 71.6 ± 8.2 years, disease duration 8.5 ± 5.0 years, Hoehn & Yahr scale: 2.1 ± 0.4, UPDRS motor score: 16.6 ± 10.4. Audio recordings during the speech task were performed on 25 patients (16 male, 9 female). Video recordings (color and depth streams) during the speech task were performed on 16 patients (10 male, 6 female), while video acquisitions (color stream only) of facial expressions were performed on 18 patients (14 male, 4 female).
- 32 HC subjects (18 male, 14 female), age: 67.7 ± 7.8 years. Audio recordings during the speech task were performed on 23 subjects (11 male, 12 female). Video recordings (color and depth streams) during the speech task were performed on 19 patients (13 male, 6 female), while video acquisitions (color stream only) of facial expressions were performed on 17 subjects (6 male, 11 female).

4.2 Video recordings of patients with disorders of consciousness

The aim of is to assess facial expressions and facial movements related to reflex and cognitive behaviors, as well as vital signs (in particular the heart rate) estimated by means of video-based techniques. Starting from the CRS-R protocol, the useful items for the automatic analysis of facial features were selected, i.e. those tasks that do not involve occlusions of the patient's face by objects or movements of the clinician:

1. Auditory function scale:
    a. Score 2/4 - Localization to sound;
    b. Score 1/4 - Auditory startle;
2. Visual function scale:
    a. Score 3/5 - Visual pursuit;
    b. Score 2/5 - Fixation;
    c. Score 1/5 - Visual startle;
3. Motor function scale:
    a. Score 3-1/6 - Localization to noxious stimulation;
4. Oro-motor/verbal function scale:
    a. Score 3/3 - Intelligible verbalization;
    b. Score 2/3 - Vocalization/oral movement;
    c. Score 1/3 - Oral reflexive movement;
5. Communication scale:
    a. Score 2/2 - Functional: accurate;
    b. Score 1/2 - Non-functional: intentional;

In particular, as one of the aims was the assessment of several reactions and features related to the perception of pain in patients that are unable to communicate, we focused our attention on the Item 3 (Motor function scale - Localization to noxious stimulation): unlike the original purpose of this item, i.e. the evaluation of the motor response in the limbs, we wanted to study facial expressions and vital signs after the administration of the noxious stimulation. It is thus possible to assess facial features and vital signs related to the level of pain felt by the patient.

During the experiments, patients' faces were recorded with the Creative Senz3D camera and were kept under constant and uniform illumination during the whole acquisition. Despite this sensor can provide two video streams (like the Kinect), only the color stream was used for these experiments. In fact, unlike the kinematic analysis of speech movements performed in PD patients, in this case the purpose was not the assessment of kinematic parameters of the articulators, as DOC patients rarely exhibit some articulatory movements, vocalizations and verbalizations. Although the depth stream could provide additional information, it was preferred to work just on the color stream, thus saving computer memory for video recordings.

Before each experiment, the camera was placed in front of the patient's face at a distance around 0.5 m from the mouth and at height close to that of the subject's eyes. The camera's optical axis had to be perpendicular to the subject's face, as shown in Fig. 4.3 Thus, according to these criteria, the face can be visible inside the scene for the whole recording. During the CRS-R administration, patients were in a half-seated position in the hospital bed.

The resolution of the color stream was 640x480 pixels at 30 frames per second (fps). Color frames were recorded in 24-bit RGB images (8 bits per channel). Videos were stored in *avi* files through the OpenCV (ver. 2.4.9) libraries using a customized code written in C++ language.

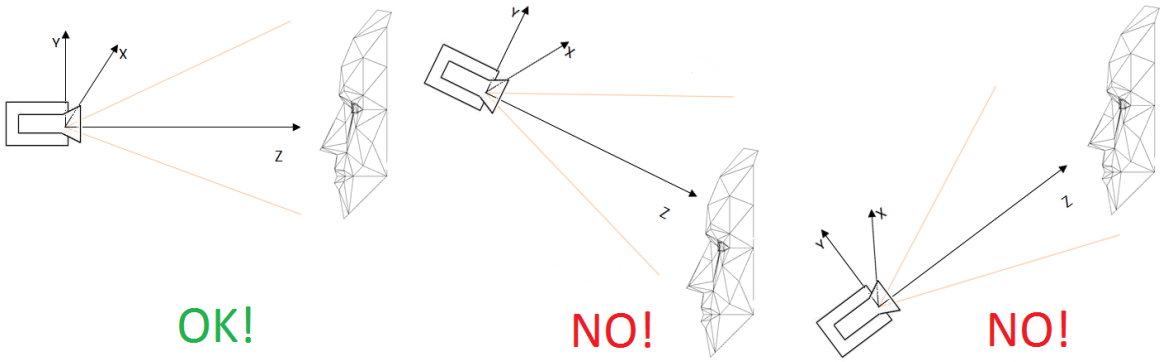

Fig. 4.3: Right orientation of the camera for the acquisitions

At the same time, the audio signal consisting of the clinician's voice was recorded with the built-in microphones of the Senz3D sensor (Fs = 44.1 kHz, recording software: Audacity, ver. 2.0.5). The audio signals where synchronized with the video recordings. The audio was recorded just for the identification of the time instants when the stimuli were administered (in particular for the localization of the noxious stimulation, where the stimulus is provided pressing the big toe or the thumb of patients, thus in a region not included in the video recordings). The time instants related to the external stimuli were identified with a short acoustic signal in the audio track.

4.2.1 Dataset

During the PhD project the following subjects were recruited: 13 DOC patients (8 male, 5 female), age: 54.9 ± 15.8 years, months after the onset: 35.3 ±27.9, 11 VS patients and 2 MCS patients, CRS-R score 6.1 ± 1.4, etiology: 8 post-anoxic, 5 traumatic brain injury. Because of the poor quality of some acquisitions, only nine (7 VS and 2 MCS) of the 13 patients were considered for the analysis of facial features and vital signs during the administration of the noxious stimulation.

5. Acoustical analysis of PD speech

As already stated in sections 1.3 and 3.1 of Part I, impairments of voice and speech (hypokinetic dysarthria) are among the main signs of the disease. PD patients may present alterations related to all speech dimensions (i.e. the four subsystems of speech production): respiration, phonation, articulation and prosody. Therefore, they might exhibit reduced variability of pitch and loudness, hoarseness, reduced stress, imprecise consonant articulation, unorthodox speech silence and speech rate alterations [30]. The non-invasiveness of acoustic measures, along with the debut of speech impairments in the early stage of the disease [82], has made this field very attractive. Many authors tried to detect features of voice and speech belonging to the aforementioned speech dimensions (in particular phonation, articulation and prosody) that could discriminate PD patients from healthy subjects. Such features could provide an aid to an early diagnosis of PD, but could also help tracking the disease progression and testing the efficacy of drug therapy.

In particular, this part of the work is focused on the study of prosody in PD patients. Dysprosody includes alterations of rhythm and speed of speech, articulation, speech/pause ratio, pitch intensity and its variations [30]. A widespread task used to evaluate speech rhythm disorders is the syllable repetition task (oral diadochokinetic test). Through this task, it was demonstrated that the alteration of the rhythm of speech is one of the most prominent features of dysprosody in patients with Parkinson's disease (PD) [94]. PD patients show a higher pace variability during a syllable repetition task, with tendencies to accelerate the rhythm [94]. However, in past studies the utterances were cut manually; this leads to long working times, incompatible with the current pace of work of the clinicians.

Moreover, most of the works on acoustic analysis in PD patients make use of freely available software tools like PRAAT [96] or commercial tools like the Multi Dimensional Voice Program (MDVP by Kay Elemetrics Corp. [97]). However, these tools are not provided with a pre-processing step that allows analyzing different parts of an audio recording in a completely automatic way. Specifically, as said above, the repetition of syllables is a commonly performed test in PD patients. With MDVP or PRAAT a manual selection of each sentence from the whole recording has to be made first. Then the tool is applied sentence by sentence for the estimation of the acoustic parameters. Therefore the automation of manual segmentation would be desirable, being a time-consuming and operator-dependent step, especially in the case of long recordings. Therefore it is relevant to develop a robust method for a complete automation of the process. This would be essential also for the implementation of an analysis system to be used at home for monitoring and rehabilitation of PD patients.

Thus, the aim of this part of the work is the development and the implementation of an automatic method for studying dysprosody in PD patients. This work is divided into two main parts:

1. To test the performance of an automatic voiced-unvoiced (AVU) segmentation algorithm, a dataset of speech signals (syllable repetition task) already analyzed in [94] was used. In this paper the temporal parameters related to dysprososdy were obtained by manual labeling the repetitions

of the syllable /pa/ uttered by PD patients and healthy control (HC) subjects. Thus, the results extracted with the AVU segmentation algorithm are compared to those provided by the authors of [94].

2. Once the accuracy of the automatic methods is tested and new temporal parameters related to dysprosody were defined, the attention was moved to the identification of dysprosodic patterns during a sentence repetition task through standard and new parameters.

Although the monologue is the most accurate task for revealing speech alterations in PD patients [83,87], the use of this task to develop and standardize an automatic analysis procedure is not straightforward. For this reason, in this part of the work we choose the repetition of a sentence to study dysprosody in idiopathic Parkinson's disease.

5.1 Experimental settings

5.1.1 Dataset

Subjects for testing the performance of the AVU algorithm

The database used for this work consists in a subset of samples already used for a previous study [94]. 32 signals from PD patients and 29 signals from HC subjects were analyzed. An exhaustive description of the groups' characteristics is reported in [94].

Subjects for studying dysprosody in PD speech

20 patients with idiopathic Parkinson's disease were recruited at the Department of Neurology of the Hospital "San Giovanni di Dio", Firenze, Italy. Patients' age ranged from 53 to 83 years (mean: 72.2 years; standard deviation SD: 8.6 years). 14 patients were male, 6 were female. At the time of the experiment, disease duration ranged from 2 to 15 years (mean: 8.3 years, SD: 4.6 years). Before the experiment each patient underwent a neurological examination. The Hoehn and Yahr disease stage [16] ranged from 1.5 to 3 (2.1 ± 0.4) and the UPDRS motor score (UPDRS part III [17]) ranged from 5 to 33 (15.7 ± 9.9), while the speech task (item 18 of the UPDRS protocol) gave results equal to 0 or 1. Thus, PD patients assessed through the perceptual evaluation did not show any (or very minimal) problem related to speech. All PD patients were under levodopa medication and were tested during their "on" state.

A group of 19 healthy subjects was tested as control group (age: 54-85 years, mean: 68.1 years, standard deviation: 8.4 years), 9 male and 10 female. A summary of subjects' characteristics is reported in Tab. 5.1.

All subjects were Italian native speakers. Signed informed consent was obtained from all the participants.

Table 5.1 - The dataset

	PD patients		Healthy subjects	
	Mean	SD	Mean	SD
Age (years)	72.2	8.6	68.1	8.4
Male	14		9	
Female	6		10	
Disease duration (years)	8.3	5.1	-	
Hoehn & Yahr stage	2.1	0.4	-	
UPDRS motor score	16.6	10.4	-	
UPDRS speech	0.6	0.5	-	

5.1.2 Experimental setup

Experimental setup for testing the performance of the AVU algorithm

An exhaustive description of the acquisition protocol can be found in [94]. The speech task consisted in the repetition of the syllable /pa/ for at least 25 times, in a comfortable steady pace without accelerating or slowing the articulatory velocity. The first 20 utterances were considered for the analysis.

Experimental setup for studying dysprosody in PD speech

Each subject was asked to repeat a standardized Italian voiced sentence ("*Il bambino ama le aiuole della mamma*") at least 10 times as spontaneously as possible at comfortable loudness without hastening the speech, therefore properly separating the consecutive sentences.

Speech signals were recorded on a standard personal computer using Audacity software (version: 2.0.3) by means of a Shure SM58 microphone and a Tascam US-144 audio board. Signals were digitized at a sampling frequency Fs= 44.1 kHz. The microphone, fixed on a boom, was positioned at a distance of 5 cm from the subject's mouth. The experiments were carried out in a quiet room of the "San Giovanni di Dio" hospital and subjects were required to stay seated during the test.

5.2 Methods

5.2.1 Testing the performance of the AVU algorithm

The aim AVU segmentation algorithm consists of finding each utterance (syllable or sentence) in the repetition task. After band-pass pre-filtering the signal (IIR Butterworth filter, 5^{th} order, cut frequencies: 50 Hz – 1000 Hz), a 20 ms sliding temporal window is applied on the whole signal. Within each window the Short-Term Energy (STE) is computed according to the following equation:

$$STE = log_{10}\left(\frac{\sum_{i=1}^{n} s(i)^2}{n} + k\right) \quad (5.1)$$

The STE values are stored in an energy vector and their histogram is computed. A suitable modification of Otsu's method [167] is then applied to the histogram to find two optimal thresholds

that split the energy histogram into two classes corresponding to voiced and unvoiced frames respectively. We point out that the method does not require defining empirical thresholds (as in other software tools) as they are adaptively computed according to the characteristics of the signal under test. The method is discussed in detail in [167,168] where its capabilities are also shown.

Figure 5.1 shows an example of the AVU algorithm on a syllable repetition task. A report with the starting and ending points of each voiced part is automatically saved as .xls file.

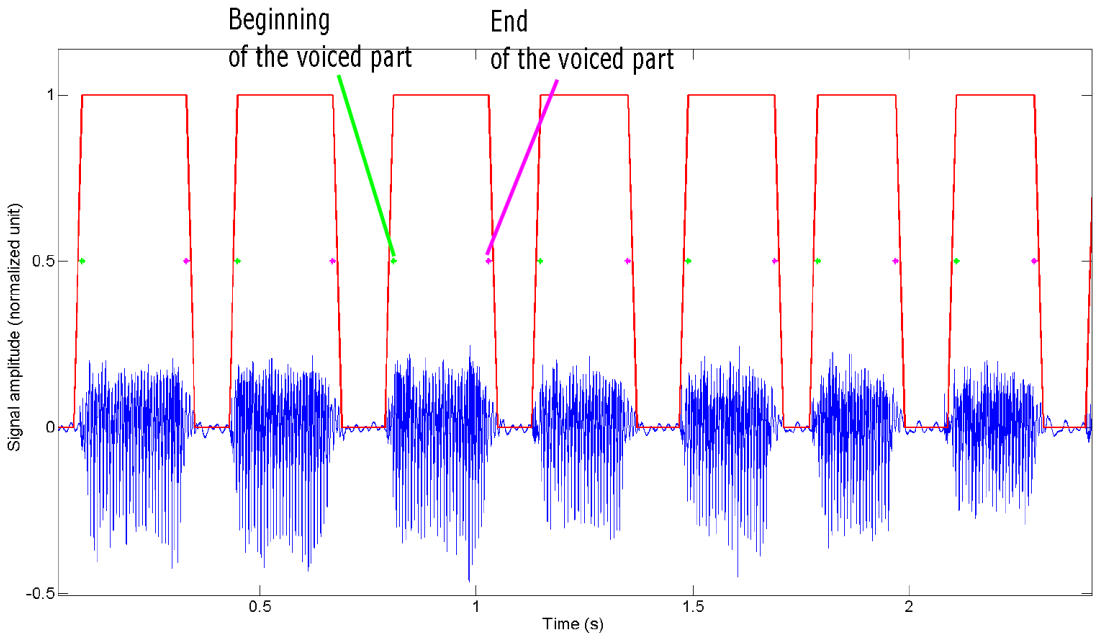

Fig. 5.1: example of the result of the AVU segmentation algorithm (red line) on a syllable repetition task.

In order to compare the results from the AVU algorithm with those extracted manually, the following parameters were computed [94]:

- Mean interval Duration (IntDur), as the average of the intervals between a vocalization and the next one;
- Standard Deviation (SD) of the intervals;
- Mean interval Duration for the first 4 repetitions ($avIntDur_{1\text{-}4}$), for the repetitions from the 5^{th} to the 12^{th} ($avIntDur_{5\text{-}12}$) and for the repetitions from the 13^{th} to the 20^{th} ($avIntDur_{13\text{-}20}$).

Moreover, on the same syllable repetitions, we compared PD patients with HC subjects, introducing new temporal parameters. We defined the Duty Cycle (D%) of the repetition, as the percent of voiced time with respect to the utterance duration. For this parameter we calculated the same measures defined in [94] for the interval duration:

- Mean value of the first 20 repetitions (D_{mean});
- Standard Deviation of the first 20 intervals (D_{SD});

- Mean value for the first 4 repetitions (avD_{1-4}), for the repetitions from the 5th to the 12th (avD_{5-12}) and for the repetitions from the 13th to the 20th (avD_{13-20});
- Coefficient of Variation (D_{COV}) on the whole task and the relative coefficient of variation ($D_{COV5-20}$) defined as:

$$D_{COV} = {}^{D_{SD}}\Big/_{\left[D_{mean}/\sqrt{20}\right]} \times 100 \tag{5.2}$$

$$D_{COV5-20} = {}^{D_{SD5-20}}\Big/_{\left[avD_{1-4}/\sqrt{16}\right]} \times 100 \tag{5.3}$$

- The ratio between avD_{5-12} and avD_{1-4} ($DRelStab_{5-12}$), the ratio between avD_{13-20} and avD_{1-4} ($DRelStab_{13-20}$) and their difference (DPA).

5.2.2 Comparison between HC subjects and PD patients on sentence repetitions

In this work we aimed at separating the repetitions of the voiced sentence: "*Il bambino ama le aiuole della mamma*". Since the sentence is mainly composed by vowel sounds, the voiced/unvoiced segmentation coincides in this case with speech/silence segmentation.

The acoustic analysis is composed by two parts: the AVU segmentation and the extraction of acoustic parameters from the voiced parts only.

In this study the LTAA (Long Term Audio Analyzer) module, proposed by our group [168] is implemented in the latest release of the software BioVoice [169] (Rel. 3.0), thus making up a single tool that performs both pre-processing and audio analysis. Moreover, the new tool allows analyzing sequentially and automatically several audio signals, thus significantly reducing the processing time as the manual user intervention to upload individual audio files to be analyzed is no longer required. Results (Excel sheets, plots, tables, etc.) are stored in separate folders, one for each signal. The software interface is shown in Fig 5.2.

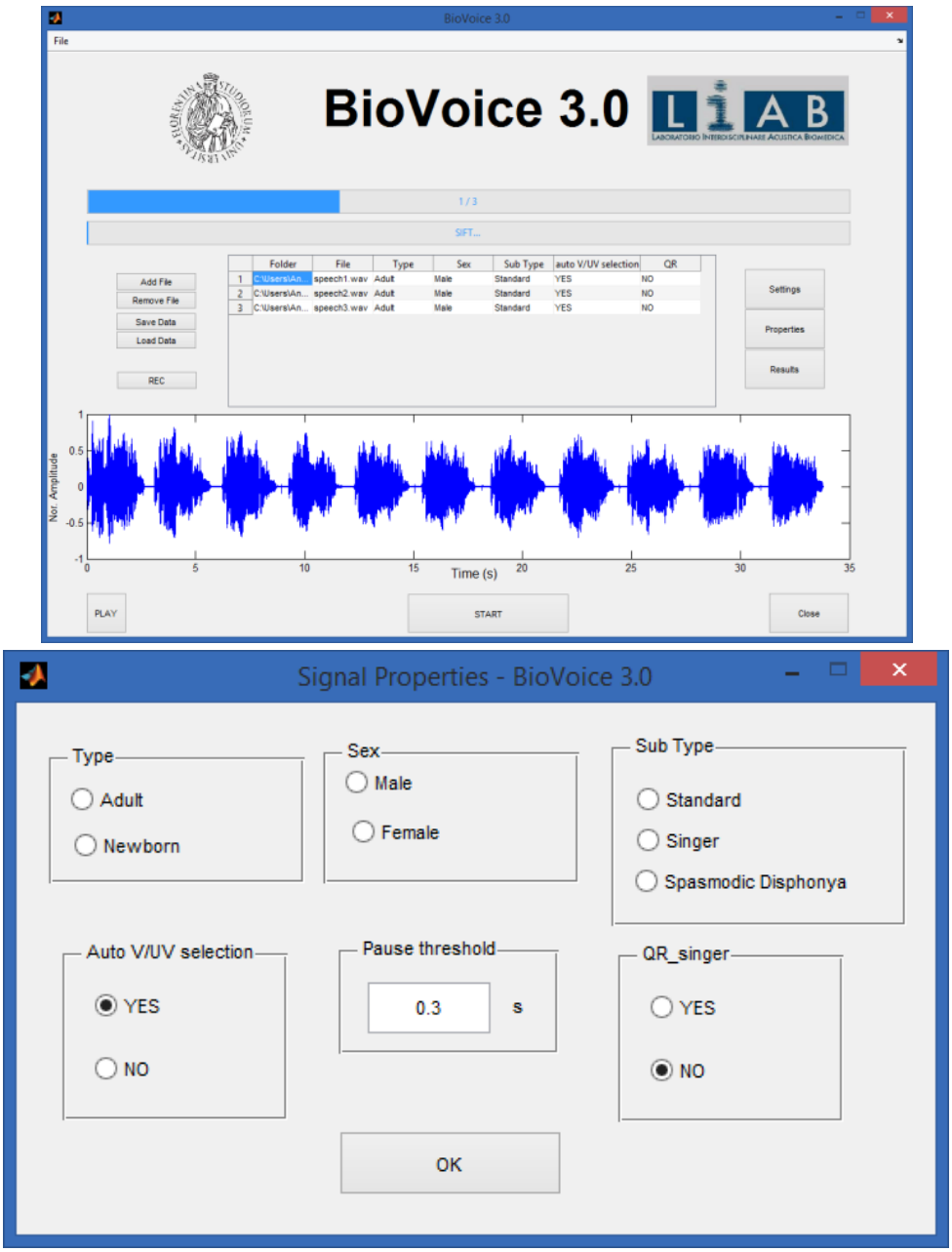

Fig. 5.2: Main software interface (upper figure) with the possibility to upload and automatically analyze several tracks. The lower figure shows the panel for choosing the signal properties. According to the selection (adult/newborn, male/female, etc.) BioVoice automatically sets some internal parameters such as pre-filtering, window's length for F0 estimation, F0 range, etc.

Figure 5.3 shows an example AVU segmentation process applied to the sentence repetition task of a healthy control subject (upper figure) and to a PD patient (lower figure).

After segmentation, some criteria are required to identify which starting and ending points of voiced parts really correspond to the beginning and to the end of each sentence. In fact, it may happen that short pauses are found within each sentence. These pauses, though significant from the clinical point of view, must be distinguished from the longer pauses between each sentence repetition. To this end we set an empirical threshold (0.3 s): an unvoiced segment (between two voiced parts) is considered a

pause within the sentence if it lasts less than this threshold. This value may be changed according to the task under consideration (Fig. 5.3, lower plot).

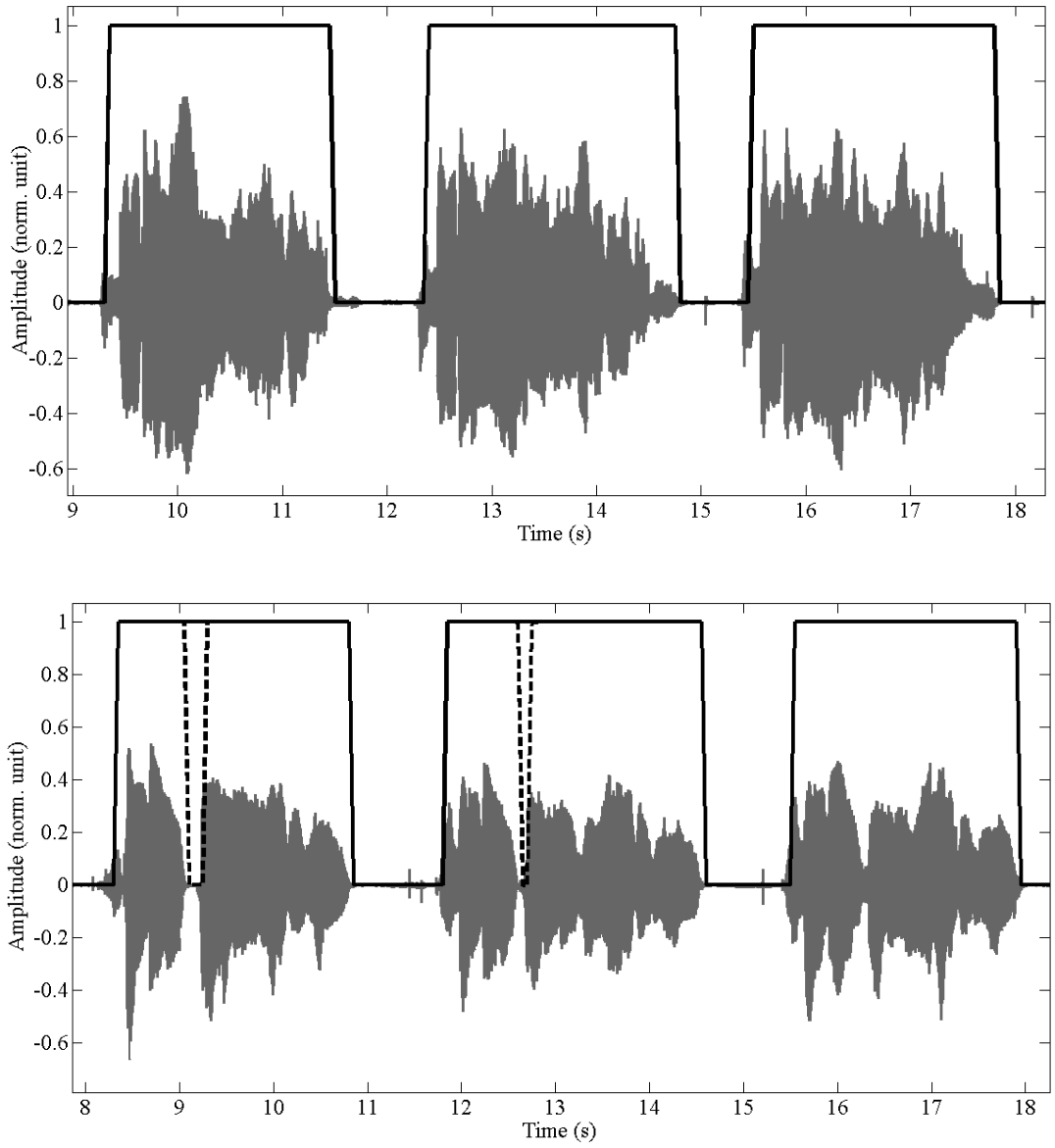

Fig. 5.3: Result of the VU segmentation process. The upper figure shows a control case where three repetitions of the vocalic sentence are correctly segmented. The lower plot concerns a PD patients, who makes pauses also within each sentence that are correctly identified (dashed black line) and removed according to the empirical threshold settled at 0.3s.

On each sentence, the following parameters are extracted:

- $T_{sentence}$: sentence duration [s], computed as the time interval between the beginning of a sentence and that of the next one. Thus $T_{sentence}$ is an index of the pace of repetition;
- T_{inter}: inter-sentence duration [s], computed as the time interval between the end of a sentence and the beginning of the next one;
- T_{pause}: pause duration [s], computed as the sum of "breaks" (short pauses) inside a sentence (less than 0.3 s);

- D%: Duty Cycle, defined as the percent of voiced time with respect to the sentence duration;

$$D = \frac{T_{sentence} - T_{inter} - T_{pause}}{T_{sentence}} \times 100 \quad (5.3)$$

- NSR: Net Speech Rate in syllables/s, defined as the number of syllables of the sentence, divided by the effective speech time ($T_{sentence} - T_{inter} - T_{pause}$) [92].

The second part of the analysis consists in the extraction of some acoustic parameters from each sentence. In this paper the following relevant parameters are considered: the fundamental frequency (F0) and the noise level. They were estimated with BioVoice and compared to PRAAT and MDVP.

With BioVoice F0 estimation is performed by means of a two-step algorithm. First, the Simple Inverse Filter Tracking (SIFT) is applied to signal time windows of short and fixed length; afterwards, F0 is adaptively estimated on signal frames of variable length through the Average Magnitude Difference Function (AMDF) within the range provided by the SIFT [169,170]. With both MDVP and PRAAT F0 is estimated through the autocorrelation of the signal calculated on short frames. For details and differences among these approaches, refer to [96,97,169].

Spectral noise is closely related to degree of perceived voice hoarseness [171,172], thus we expect a not negligible amount of noise in Parkinsonian speech samples, as also reported in [83]. With BioVoice noise is estimated by means of an adaptive version of the Normalised Noise Energy method (ANNE) [171]. This method was in fact successfully applied to pathological voices in [169,170].

Once F0 and noise are tracked within each sentence, the following parameters are computed:

- F0 Coefficient of Variation (F0CV), defined as the ratio between F0 standard deviation and F0 mean value;
- F0 Normalized Range (F0NR), defined as the difference between F0 max and F0 min, divided by the mean value of F0;
- ANNE mean value.

The normalization of the F0 values was performed in order to take into account ~~for~~ gender differences within the samples. We compared our results concerning F0CV, F0NR and noise with those obtained with PRAAT and MDVP. We remark here that each tool implements different techniques both for F0 and noise estimation. Specifically, PRAAT computes the HNR (Harmonics to Noise Ratio) while MDVP estimates the NHR (Noise to Harmonics Ratio). A comparison of the performance of BioVoice, PRAAT and MDVP is reported in [173,174,175,176].

Summarizing, in order to have a set of characteristics related to dysprosody in PD patients, for each sentence we computed parameters related to temporal features of speech ($T_{sentence}$, T_{inter}, T_{pause}, D% and NSR), to pitch variability (F0CV and F0NR) and to the degree of dysphonia (mean ANNE).

Statistical Analysis

A two-tailed t-test was applied to assess the significance of differences between the PD patients and controls. The difference was considered significant for $p < 0.05$. Given the different methods for noise estimation implemented in BioVoice, PRAAT and MDVP, the strength of the relationship among these methods was tested computing the correlation coefficient between ANNE and HNR and between ANNE and NHR. Statistical analysis was performed with Microsoft Excel 2010.

6. Markerless analysis of articulatory movements during speech (validation on healthy subjects)

The main disadvantages of the existing methods for studying articulatory movements during speech (i.e., x-ray imaging, magnetic resonance imaging - MRI, ultrasound technique, electromagnetic articulography - EMA, optoelectronic systems, etc.) are the high cost and the discomfort for participants or patients. Moreover, the above techniques needs a lengthy preparation protocol, thus their use is limited to the research field (Part I - Section 3.2.1).

In order to broaden the kinematic studies of speech articulation for speech therapy purposes, or track the disease progression, the use of a low-cost and fully contactless system would be desirable.

In the last five years the spreading of 3D video sensors (like Microsoft Kinect, Asus Xtion, Primesense Carmine, etc.), has revolutionized the world of videogames and not only, providing new possibilities to study body movements without any sensor attached to the subject. These devices could be used to get trajectories and kinematic parameters in the 3D space, and to analyze some fundamental articulatory parameters like lip protrusion. Moreover, these devices could be integrated in speech therapy applications since most of the tasks consist of tracking facial and articulatory movements to provide clinically useful feedback [109].

Speech therapy can address the slowdown of speech disorders related to neurological illnesses, such as hypokinetic dysarthria associated with Parkinson's disease or stroke [106,109,177]. Although speech disorders concern a quite large population, speech therapy is often applied to a small number of patients. This is due to the following factors:

- During group sessions, the speech therapist has difficulty to pay the same level of attention to each patient, in order to evaluate the therapy exercises and provide a valuable feedback to patients;
- Due to neurodegenerative diseases, most of the patients with hypokinetic dysarthria are elderly people, that frequently have trouble getting and attending therapy in specialized centers;
- In several cases, patients should continue the therapy exercises at home. However, they often are not enough motivated without the direct supervision of the therapist.

For these reasons, it arises the need for a system which could automatically provide a feedback about the articulatory movements and that could also be integrated in home environment. Therefore, this system should be as much as possible at reasonable price and based on a contact-less technique.

Some applications of the Kinect sensor have been developed also for speech therapy purposes, in order to study and automatically identify the therapeutic exercises that involve facial movements [178]. However, to our knowledge, the accuracy of a fully markerless technique to study speech articulation was never tested. Thus, our aim is to test the performance of a system composed by a 3D depth sensor and a face tracking algorithm in order to track lip movements during speech. In this study the accuracy is verified against an established optoelectronic method.

6.1 Experimental Settings

To test the performances of the markerless system (based on the Primesense Carmine 1.09 sensor, which is a Kinect-like device, and the *Intraface* tracking algorithm) to track articulatory movements, we used an optoelectronic technique (Vicon Motion Systems Ltd.) widely used as accurate marker-based motion capture system. Both systems (Primesense and Vicon) were used simultaneously and the different streams were acquired synchronously. In the following sections, we present the systems used in our study and we describe the method.

6.1.1 Markerless system

Video acquisitions

During the experiments the subjects' faces were recorded with the depth sensor Primesense Carmine 1.09. This device was chosen for its ability to work at short distances (0.4-1.5 m), thus appropriate for face movements. As classical structured-light sensors, it provides two video streams: the color video (like a normal webcam) and the depth stream, where the pixels of each frame code the distance of a point in the scene from the camera plane. The image resolution of both streams was set at 320 x 240 pixels. Both videos were acquired synchronously at 30 frames per second, and stored as *avi* files by means of the OpenNI (ver. 2.2) and OpenCV (ver. 2.4.9) libraries.

The device was located in front of the subject's face (at the height of the mouth) at a distance around 0.7-0.8 m from the lips, according to the specifications provided by the manufacturer.

For the automatic identification of the facial features, the tracking algorithm *Intraface*[2] was used. This algorithm fits to the video frames a face model composed of 49 points, on the basis of texture descriptors like SIFT (Scale-Invariant Feature Transform) [127,179]. This algorithm was chosen for its robustness against illumination changes, for its ability to describe asymmetrical face movements (very important in the context of speech therapy applications) and for its efficiency [127]. In particular, lips are modeled as a set of 18 points: 12 on the outer border and 6 on the inner border. In our study, only 7 points on the outer border were considered for the analysis (Fig. 6.1) to compare the performance of the system against the marker-based method.

Audio acquisitions and streams synchronization

An audio track of the speech corpora (words, sentences and syllables) uttered during the experiment was recorded using the depth sensor that is provided with two built-in microphones. These acquisitions were performed through the software Audacity (Version 2.0.5) on the same laptop used for the video streams. Stereo signals were recorded at 44100 Hz of sampling rate and coded at 16 bits. Although our aim was just the evaluation of the articulatory movements, the audio stream was of basic importance for the alignment of the different streams and for the evaluation of the errors for each

[2] *Intraface* tracking algorithm is available at http://www.humansensing.cs.cmu.edu/intraface/. The algorithms are available for research purposes only, commercial use is not allowed.

phoneme of the corpus. In order to synchronize the four streams we built and used a simple electronic circuit composed by: a normally open switch, a LED with red visible light, an infrared LED (with wavelength in the near infra-red, in order to be seen by the Vicon cameras, between 800 and 900 nm) and a buzzer. At the beginning and at the end of each acquisition the button was pressed two consecutive times, so that in the audio stream it was possible to hear the acoustic signal, to see the red light in the color stream and an extra marker (the infra-red LED) could be identified in the marker-based stream. Image registration and frame synchronization between the two Kinect streams (depth and color) were performed via dedicated software, through the same tool used for the acquisition.

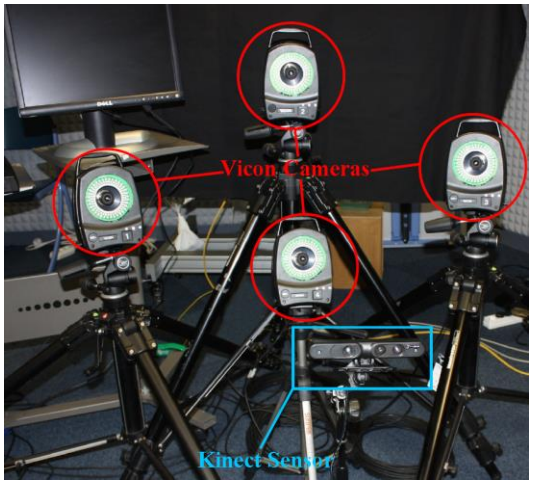

Fig. 6.1: Experimental setting: marker-based system and Kinect-sensor displacement.

6.1.2 Marker-based system

To compare the performance of the aforementioned markerless method, we used an optoelectronic system (Vicon Motion Systems Ltd., UK) as a reference. This system was composed by four cameras (MX3+ model) with special optics for near range applications. Sixteen reflective markers of 3mm diameter were glued on the faces of the subjects. This size is suitable to study facial movements without interfering with the face tracker.

Before each acquisition, the markers were accurately located in some precise facial points defined by the *Intraface* model: two for each eyebrow, three on the nose, seven on the outer border of the lips (one for each corner – L1 and L4, two on the upper lip – L2 and L3 – and three on the lower lip – L5-L7) and two on the chin (Fig. 6.2). The 3D trajectories of these markers were recorded synchronously using the infrared cameras at 100 Hz of sampling frequency and reconstructed using the Vicon Nexus software.

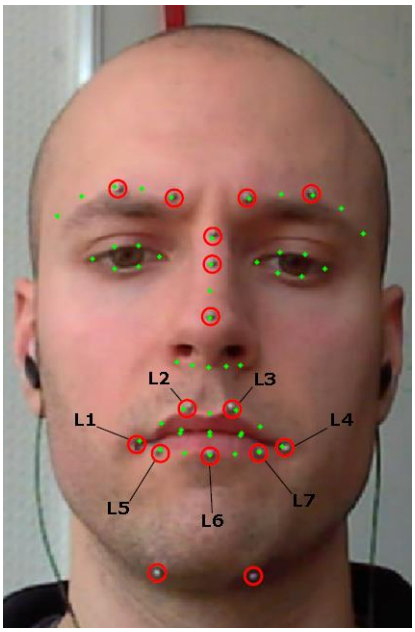

Fig. 6.2: Intraface tracker model points (green dots) and optical markers locations (red circles). The markers were located in the same position of some model points, in order to estimate the 3D rigid transformation to register the two sets of points.

6.1.3 Speech corpora and data collection

The acquisitions were performed in a quiet room with reduced environmental noise of the LORIA research center of Nancy, France. Two healthy subjects were recruited for the experiment: an Italian native speaker and a French one. Subjects were seated in front of the camera at a distance between 0.7 and 0.8 m from the Primesense sensor. This range is a tradeoff between the device characteristics and its distance from the subject's face (as close as possible) without interfering with the field of view of the Vicon cameras (Fig. 6.1 and 6.3).

As stated above, before each acquisition, 16 reflective markers were accurately glued on the subject's face in precise anatomical points chosen in correspondence to selected points of the *Intraface* model (presented in the next section), as shown in Fig. 6.2. We checked that the markers did not alter the acquisition quality of the *Intraface* tracker.

Each subject was asked to read and pronounce the corpus (displayed on screen in front of the subject, Fig. 6.3) without any excess of head movements. We chose two corpora (one for each language), both composed of 50 meaningful sentences and 100 meaningful words. The French sentences were extracted from the Comberscure corpus [180], while the words were chosen from the Lafon lists [181]. The Italian sentences and words were chosen from the corpus defined by Bocca and Pellegrini [182]. Moreover, both subjects had to repeat the syllable /pa/ for at least 30 times with a single breath. The subjects' face was kept under constant and uniform illumination during the whole recordings (Fig. 6.3).

Fig. 6.3: A healthy subject during the speech acquisitions

6.2 Methods

6.2.1 Data processing

As mentioned above, to detect some facial feature points (as the lips) a tracking algorithm capable of detecting and tracking these points, is needed. The face tracker *Intraface* used in this work fits a model of the face to the color image using texture descriptors (SIFT) [127,179]. Unlike other face trackers based on *a priori* learned face model, such as active appearance models [125,126], here each landmark position is directly optimized to the current frame based on texture descriptors. This involves a better ability to generalize situations never seen in the training set, like asymmetrical face movements, leading to a higher flexibility. Moreover, since the fitting is based on SIFT descriptors, this algorithm is robust against illumination changes [127]. The *Intraface* tracker fits to the scene a model composed of 49 points: 10 for the eyebrows, 12 for the eyes, 9 for the nose and 18 for the lips (12 on the outer contour, 8 on the inner contour) as shown in Fig.6.2. For this study we considered the points of the eyebrows, nose and of the outer lips contour.

Since the face tracker works on 2D color images, it was necessary to extract the 3D coordinates of the points of interest. Thus, for each of these points we computed the coordinates on the lateral axis (X) and on the vertical axis (Y) starting from the depth information (frontal axis – Z).

Since we registered and synchronized the depth frames with the color frames provided by the depth sensor, to extract the Z value (in mm) for each point we just needed to sample the depth image in the same pixel coordinates provided by the face tracker. According to the scheme in Fig. 6.4, we calculated the X and Y coordinates with the following formulas [183]:

$$X = Z\,\frac{(x-c_x)}{f} \qquad with\ f = \frac{W}{2}\left[\tan\left(\frac{FOV_h}{2}\right)\right]^{-1} \tag{6.1}$$

$$Y = Z\,\frac{(y-c_y)}{f} \qquad with\ f = \frac{H}{2}\left[\tan\left(\frac{FOV_v}{2}\right)\right]^{-1} \tag{6.2}$$

where x and y are the coordinates on the image plane (in pixels) of a point of coordinates $[X\ Y\ Z]^T$ in the 3D space, f is the focal length (in pixels) of the camera, (c_x,c_y) are the coordinates (in pixels) of the principal point, W and H are the dimensions of the image (width and height, respectively) in pixels, FOV_h and FOV_v are the horizontal and vertical field of view (equal to 57.5° and 45°, respectively).

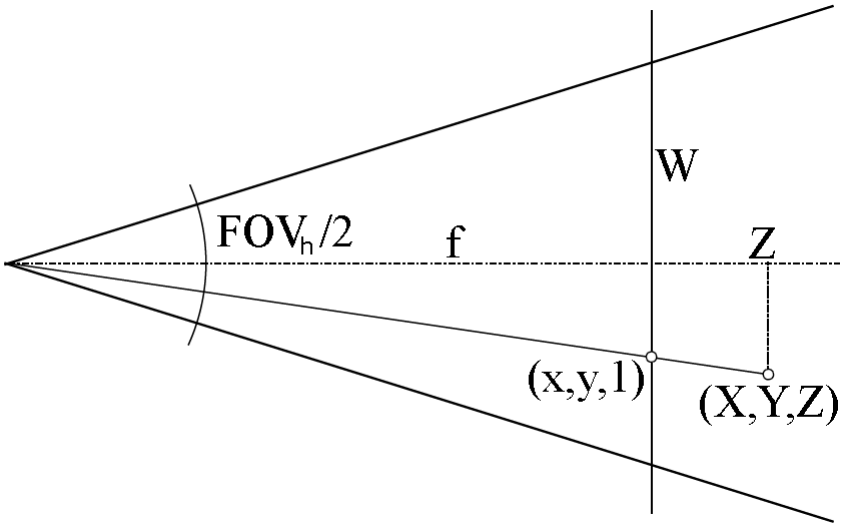

Fig. 6.4: Pinhole camera model. This model was used to retrieve the 3D coordinates of the face points (estimated with the markerless system), starting from the image coordinates plus the depth information Z (according to equations 6.1 and 6.2).

As the two reference frames are different, the two sets of points must be aligned in the space to compare the trajectories of the points of interest extracted with the markerless system with the reference ones. To do this, paying great attention to locate the markers in the same position as some *Intraface* points, the 3D rigid transformation that allows mapping the markerless points in the marker-based reference frame can be estimated. Using pairs of corresponding points provided by the two systems, the rotation matrix **R** and the translation vector **T** were estimated through a least squares solution and this transformation was applied to each point extracted from the markerless system:

$$P'_k = \boldsymbol{R}\, P_k + \boldsymbol{T} \tag{6.3}$$

where P_k is a generic point of coordinates $[X_k\ Y_k\ Z_k]^T$ in the marker-less reference frame mapped to the Vicon reference frame (P'_k) through the 3x3 rotation matrix $\boldsymbol{R}$ and the translation vector **T**. Thus, knowing couples of corresponding points in the two reference frames, it was possible to estimate the transformation parameters. Using a least squares solution and making use of more pairs of points than those required by the number of unknowns, we overestimated the system reducing the effect of noise on the estimation.

For this work we used 7 pairs of points, respectively: two for each eyebrow, two for the nose and one for the lips (midpoint on the lower lip – L6, Fig. 6.2).

Speech Alignment

In order to measure the tracking errors of the lips points for each phoneme, the audio signals were processed by means of two speech alignment software. First, the stereo speech samples were converted to mono and then resampled at 16 kHz. The French corpus was aligned using Sphinx [184] while the Italian text was aligned with the SPPAS software, based on the Julius Speech Recognition Engine [185,186]. Both tools are based on Hidden Markov Models (HMM). After the alignment was completed, we obtained the files with the phonemes included in the corpora, the starting and the ending instants for each of these phonemes. The temporal values allowed aligning the phonemes to the trajectories extracted with the two methods (marker-less and marker-based) and the subsequent error calculation for each phoneme, as shown in Fig. 6.5.

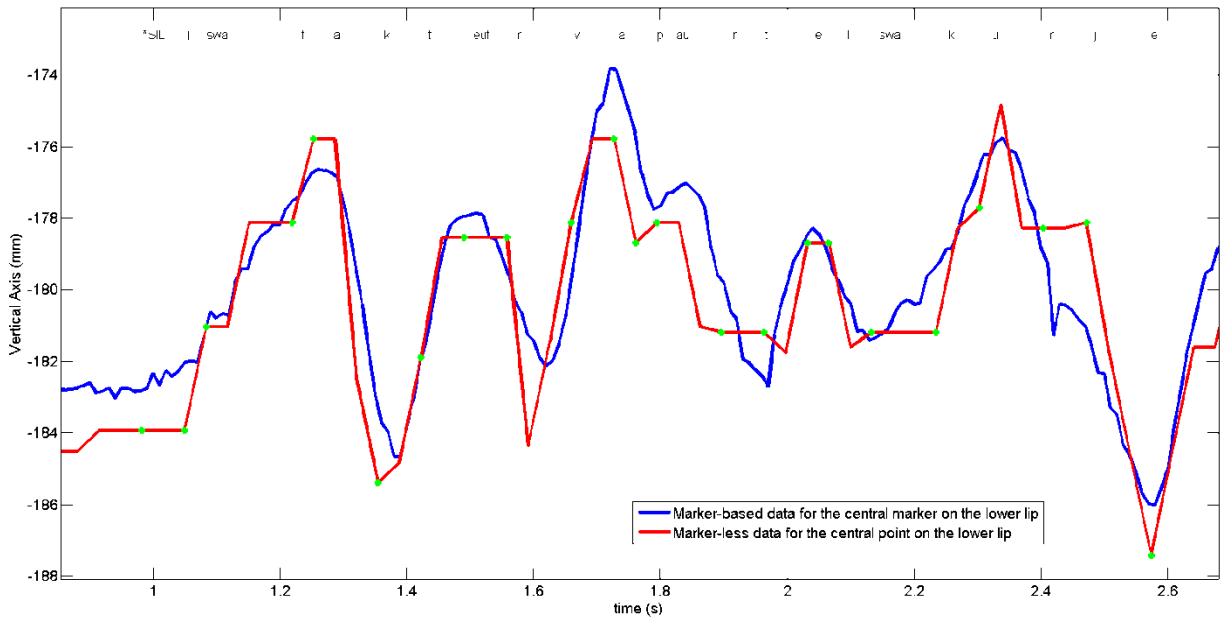

Fig. 6.5: vertical coordinate of point L6 during the pronunciation of a French sentence ("*Le facteur va porter le courrier*"), whose phonemes obtained after the alignment process are shown on top.

6.2.2 Articulatory parameters and error measure

Once the two sets of points were aligned, the trajectories extracted with the markerless system were resampled at 100 Hz using a spline interpolation technique. In this way, the comparison with the reference trajectories was obtained computing the root-mean-square error (RMSE) in mm, according to the following formula:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{6.4}$$

Where *N* is the number of samples of the trajectories during a single phoneme or syllable repetition, y_i is the *i-th* sample of the marker-based trajectory and $\hat{y}_i$ is the corresponding sample extracted from the resampled markerless trajectory.

In addition to the comparison on the 3D coordinates of the points of interest (points L1-7 in Fig. 6.2), we computed some articulatory parameters:

- Lip width: distance on the lateral axis between the two corner points (points L1 and L4);
- Lip opening: distance on the vertical axis from the midpoint between points L2 and L3, and the central lower lip point (point L6);
- Lip protrusion: distance on the frontal axis from the midpoint between points L2, L3 and L6 and a fixed reference point, in this case the nose tip.

All these parameters were normalized with respect to head rotation angles. These angles were calculated from the markers located on the eyebrows and nose.

Afterwards, for each syllable repetition the following kinematic parameters (for both systems) were computed: the maximum velocity (V_{open}) and acceleration (A_{open}) during the opening phase, the maximum velocity (V_{close}) and acceleration (A_{close}) during the closing phase. These parameters were calculated differentiating in time the trajectory on the vertical axis of the central point of the lower lip (point L6 in Fig. 6.2). V_{open} was calculated as the minimum speed value during the first half of the repetition, while V_{close} corresponds to the maximum speed value from the time instant of V_{close} up to the end of the utterance (Fig. 6.6). The same criteria were adopted to extract A_{open} and A_{close} from the acceleration values of the lip point (Fig. 6.6). Moreover, for each syllable repetition the Pearson's correlation coefficient between trajectories, velocities and accelerations extracted with both systems was computed. Correlation values close to 1 indicate that the trends of displacement, speed and acceleration calculated with the proposed method are very similar to the ground truth, as shown in Fig. 6.6.

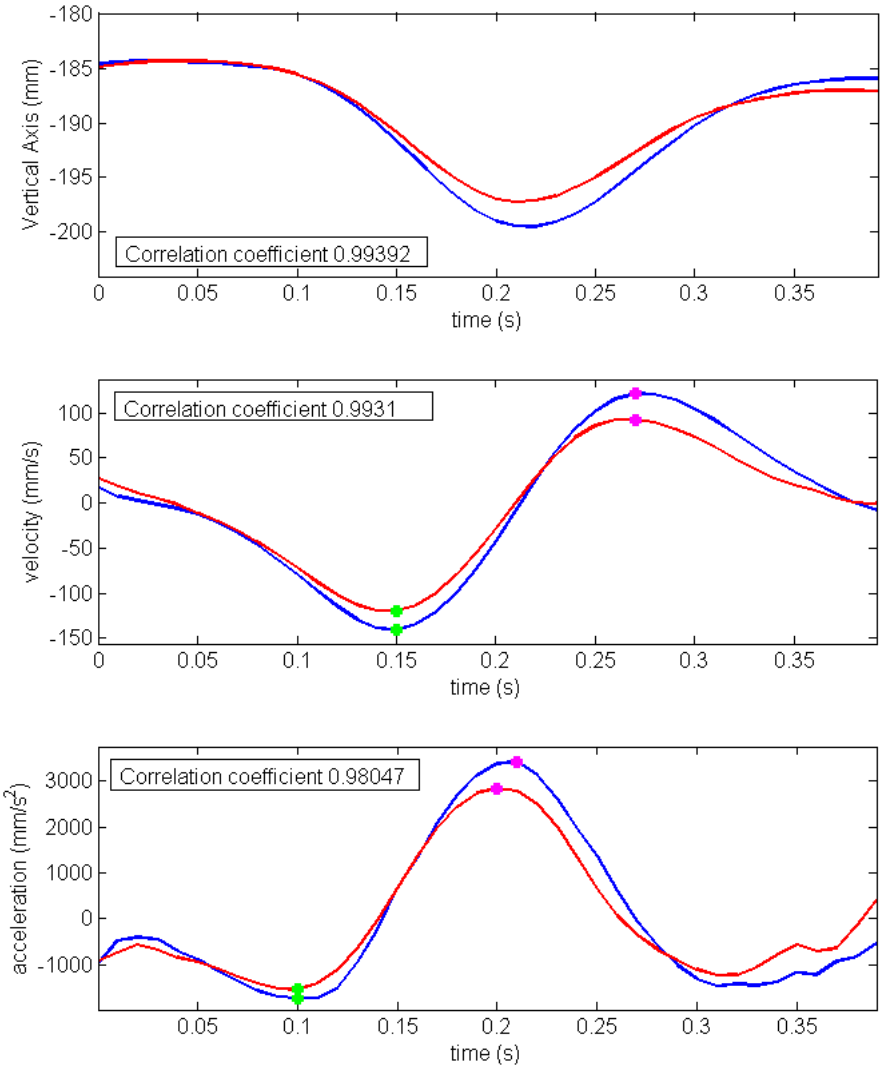

Fig. 6.6: vertical trajectory of the central point of the lower lip (upper plot) during the repetition of the syllable /pa/; speed (central plot) and acceleration (lower plot) on the vertical axis. The blue lines are relative to the reference method (marker-based), while the red lines are estimated with the markerless technique. The green points indicate the maximum velocities and accelerations during the opening phase, while the magenta points are the maximum velocities and accelerations during the closing phase.

6.2.3 Depth accuracy

The manufacturer of the Primesense sensor provided only the spatial resolution at 0.5 m from the camera, equal to 1 mm for the depth and 0.9 mm for the other two axes. Since our experiments were performed at a distance between 0.7 and 0.8 m, we expected lower resolutions. To estimate the error introduced by the Kinect sensor in the estimation of the depth value of a point in the scene (and thus for the estimation of its 3D coordinates), we used a phantom object composed by a box on which 7 reflective markers (of the same type used during experiments) were glued on small squares of yellow paper (surface = 1 cm^2) on one surface of the box (Fig. 6.7). These markers were located like a cross: the points on the horizontal axis were equally spaced of 25 mm, while those on the vertical axis were spaced of 50 mm. The Primesense sensor and the Vicon cameras were placed in front of a table on

which the box was moved with a constant speed from a distance of 900 mm to 500 mm from the depth device. This test was repeated twice.

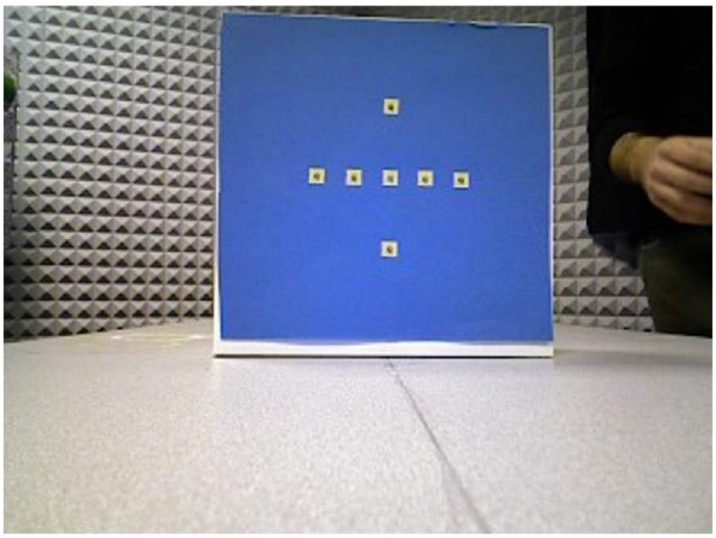

Fig. 6.7: Phantom object used to estimate the depth accuracy of the Primesense sensor.

We calculated the mean RMSE for the 7 points along the 3 coordinates for the entire range of movement covered by the box. Finally, dividing this range into intervals of 20 mm, within which the mean RMSE was computed, 20 error values were obtained.

In order to check if the errors introduced by the Primesense sensor in the estimation of the depth values was comparable to the RMSE values obtained during speech acquisitions, we computed the mean distance of the lips (in mm) from the camera.

7. Markerless analysis of articulatory movements during speech (applications to PD patients)

Although it is well accepted that PD patients with hypokinetic dysarthria exhibit a reduced articulatory kinematic, some conflicting results were found. Walsh et al., 2012 [88] tried to elucidate this point studying jaw and lower lip movements during the opening and closing gestures in the case of bilabial consonants. The authors demonstrated that PD patients exhibited a reduced articulatory kinematic, highlighted by reduced velocities of jaw and lower lip. These results support the hypothesis that a "downscaling" in speech production occurs in PD patients with hypokinetic dysarthria.

As explained in Part I - Chapter 3, an important issue concerns the implemented methodologies. In the past decades the kinematic analysis of the articulators were performed through several motion capture technologies. The most important are: optoelectronic systems [88,113,115], electromagnetic articulography (EMA) [103], X-ray techniques [104] and Magnetic Resonance Imaging (MRI) [101]. However, all these techniques are quite expensive and their use is limited to research within highly specialized laboratories. Moreover, some of the most widely used techniques (optoelectronic systems and EMA above all) are marker-based and need quite long preparation protocols in order to achieve good results.

In the previous chapter, a fully markerless and low-cost method to study the articulatory movements in 3D (in particular lips movements) during speech was proposed. We tested its accuracy against an optoelectronic marker-based method during the repetition of words, sentences and syllables. This system is able to track lips movements during speech, combining a face tracking algorithm and a 3D depth sensor. This markerless technique, whose accuracy was proven for tracking lips movements (results on the accuracy are reported in Part III - Chapter 12), is thus applied here to analyze movements of the lower lip during a syllable repetition task, both in PD patients and HC subjects. The aim is that of testing the reliability of a cheap and markerless approach for assessing signs of hypokinetic dysarthria (in particular, alterations of peak velocities and accelerations) as already demonstrated by Walsh et al., 2012 [88].

This system could be used for the analysis of the articulatory movements during speech, leading to a significant improvement in monitoring disease progression, and in speech therapy in patients with dysarthria due to neurodegenerative diseases. In fact, this system could be easily implemented in a home environment increasing the percentage of patients that undergoes speech therapy.

7.1 Experimental settings

7.1.1 Subjects

14 PD patients were recruited at the Unit of Neurology of the Florence Health Authority ("San Giovanni di Dio" Hospital, Firenze, Italy), and at the *Associazione Italiana Parkinsoniani (AIP)* –

Sezione di Firenze, Firenze, Italy. PD patients' age ranged between 62 and 80 years (mean: 71.6 years, standard deviation: 7.0 years). 9 Patients were male and 5 were female. Before carrying out the experiment, each patient underwent a neurological examination. The Hoehn and Yahr disease stage [16] ranged from 1.5 to 2.5 (2.0 ± 0.3) and the Unified Parkinson's Disease Rating Scale (UPDRS) motor score (UPDRS part III [17]) ranged from 5 to 43 (16.0 ± 12.0), while the speech task (item 18 of the UPDRS part III protocol) gave results equal to 0 or 1. Through this item, the neurologist judges the Parkinsonian continuous speech, paying attention on those signs related to dysphonia and dysprosody, but still considering global aspects as the intelligibility. In this way, ratings do not directly concern the issue addressed in this paper (the articulatory undershoot). Thus, PD patients assessed through the perceptual evaluation of the neurologist showed no speech problems (speech item = 0) or slight problems (speech item = 1) consisting in loss of modulation, diction or volume, without major alterations of speech intelligibility. All PD patients were under levodopa medication and were tested during their "on" state.

An age-matched control group composed by 14 HC subjects with no history of neurological disease was recruited. HC subjects' age ranged from 60 to 85 years (mean: 69.0 years; standard deviation: 7.4 years). 8 subjects were male and 6 were female. Table 7.1 summarizes the features of the two groups, both consisting of Italian native speakers. Signed informed consent was obtained from all the participants.

Tab. 7.1: Characteristics (mean values and standard deviations) of the two groups analyzed in this work

	PD patients		HC subjects	
	Mean	SD	Mean	SD
Age (years)	71.6	7.0	69.0	7.4
Male	9		8	
Female	5		6	
Disease duration (years)	8.4	6.1	-	
Hoehn & Yahr stage	2.0	0.3	-	
UPDRS motor score	16.0	12.0	-	
UPDRS speech	0.6	0.5	-	

7.1.2 Experimental setup

The experiments were carried out in a quiet room of the "San Giovanni di Dio" Hospital, Firenze, Italy. The speech task consisted in the repetition of the syllable /pa/ for at least 25 times within a single breath, in a comfortable steady pace. According to [187], participants were asked to avoid acceleration and slowdown of the articulatory velocity. As shown in Skodda et al., 2011 [187] a difference between PD patients and HC subjects might be present in the pace of repetition (i.e., PD patients tend to accelerate the rhythm of repetition). However, the analysis of this point goes beyond the aims of the present work and is not considered here. Subjects were seated during the experiment trying to avoid abrupt head movements during the whole task.

The subjects' face was recorded using the Microsoft Kinect for Windows sensor. The aim was that of detecting the 3D coordinates of some facial points during the speech task. The Microsoft Kinect is a structured light sensor that provides two video streams: a color stream (in the RGB color space), and a depth one where each pixel codes the distance of the points in the scene from the camera plane. The Kinect sensor was placed in front of the subject's face at a distance between 0.5 and 0.7 m from the mouth and at a height close to that of the subject's eyes. This distance was chosen as a trade-off between the technical specifications provided by the manufacturer (in "near range" mode the minimum distance is 0.4 m [166]) and the need of having the subject's face as close as possible to the camera, in order to achieve the best accuracy in tracking the 3D facial points.

7.2 Methods

7.2.1 Video specifications

The resolution of both streams was 640x480 pixels at 30 frames per second (fps). Color frames were recorded in 24-bit RGB images (8 bits per channel), while depth frames were recorded in 16-bit, 1 channel images. These features are the best trade-off, in terms of spatial and temporal resolution, provided by this kind of sensor to achieve good accuracies for tracking fast movements like those of the lips during the syllable repetition task.
Both streams were recorded and stored in *avi* files through the OpenNI (ver. 2.2) and OpenCV (ver. 2.4.9) libraries using a customized code written in C++ language.

7.2.2 Depth-color registration

Facial feature points were tracked on the color images by the face tracking algorithm that will be described in the next section. To provide the 3D location of these points, the alignment between depth and color frames is required. In fact, as the two streams come from two different and unaligned cameras (color and infrared cameras, as reported in Fig. 7.1) a stereo calibration step is firstly required. This step involves two parts:

- Intrinsic calibration for both cameras: we retrieved the internal parameters of each camera (i.e., focal length, principal point, skew coefficient and distortions). An exhaustive description of these parameters is reported in [188]. These parameters allow the estimation of the 3D coordinates of facial points, as described in the next section.
- Extrinsic calibration: the rotation matrix and the translation vector that define the position of one camera with respect to the other one are estimated. In our case we estimated the roto-translation matrix required to map the position of each depth pixel in the color reference frame.

In this work, the calibration step was performed using the Camera Calibration Toolbox for Matlab [189] by simultaneously recording with the two cameras 25 images of a checkerboard in different

positions and angles. This procedure was performed just once before making the video recordings on both groups.

Once the intrinsic parameters of each camera and the extrinsic relationship between the two reference frames were estimated, it was possible to map the pixels of the depth image into the color reference frame, as shown in Fig. 7.1.

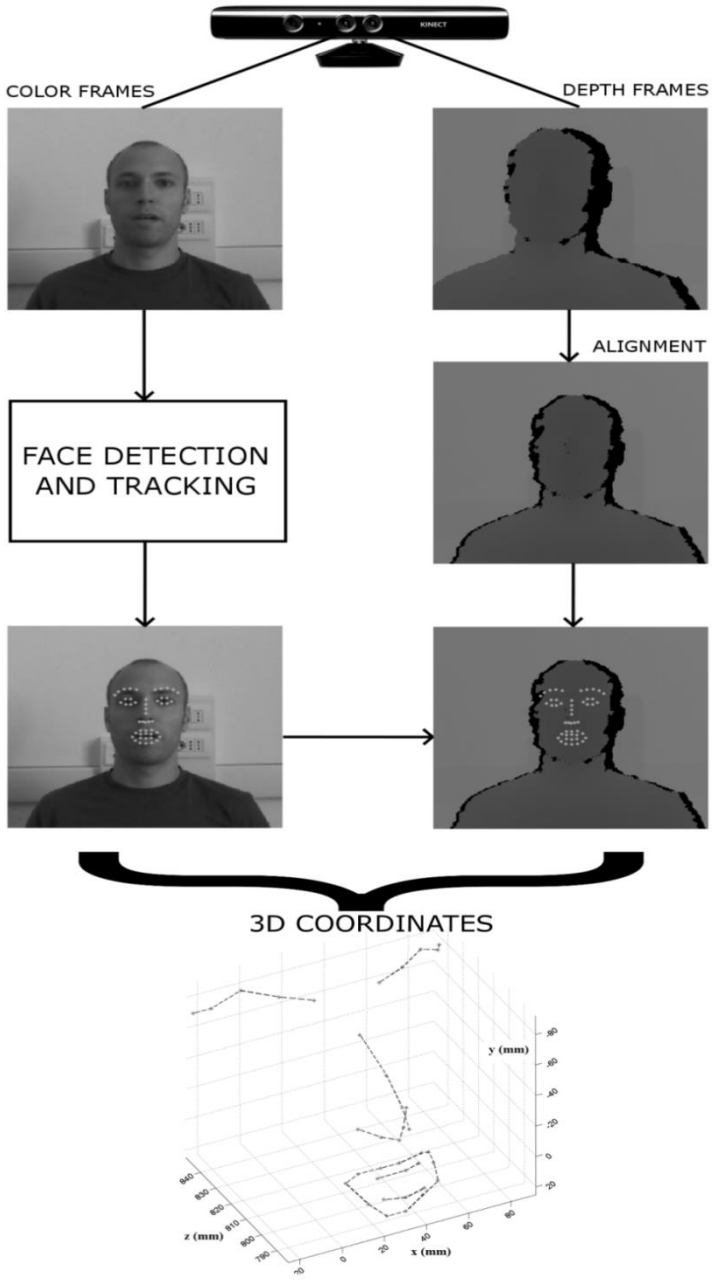

Fig. 7.1: Main video processing steps. Left - The face tracking algorithm detects the facial feature points in the colour video frames. Right - The information of the depth frames allows computing the 3D coordinates of the face points (bottom plot).

The alignment process is of great relevance in order to achieve good results in the next steps, in particular in the estimation of the 3D coordinates of facial points. In fact, the face tracker used to automatically locate and track the facial landmarks relies only on the color frames, providing 2D

points in the image plane. Therefore, to estimate the 3D position of these points, their distance from the camera plane must be known. After the alignment, these distances are given by the values of the depth image in the same pixel coordinates of the points of interest detected in the color image, sampled at the same time instant.

Although a depth to color registration was possible by means of the OpenNI libraries, a manual calibration of the sensor was carried out. In fact the intrinsic parameters on which this "factory calibration" is based might not be very accurate. Thus, we preferred to perform the calibration before carrying out the experiments.

7.2.3 Face tracking

As previously introduced, the facial feature points were located and tracked by means of a face tracking algorithm. In this work we used the *Intraface* tracking algorithm as implemented in the previous chapter, where its performance (combined with a depth sensor) in tracking lip movements during speech was tested. This tracking algorithm fits a face model made up by 49 feature points to the color images provided by the camera. These points are set as follows: 10 for the eyebrows, 12 for the eyes, 9 for the nose and 18 for the lips (12 on the outer contour, 8 on the inner contour) as shown in Fig. 7.2. To solve the optimization problem that consists in minimizing the distance between the model and the image, the algorithm uses texture descriptors (Scale Invariant Features Transform –

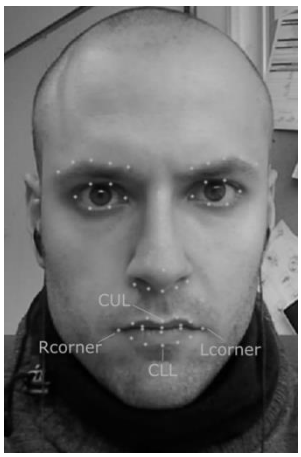

Fig. 7.2: Face model used in this work. The Intraface tracker allows detecting and tracking 49 face points (dots). For the kinematic analysis the following points are taken into account: Central Upper Lip (CUL), Central Lower Lip (CLL) and the two mouth.

SIFT). These descriptors make the tracker robust against illumination changes [127,179]. Moreover *Intraface* was chosen for its ability to generalize situations and face movements never seen in the training set, like asymmetrical movements of the mouth and eyelid movements. This could lead to a greater flexibility in view of the development of a system for speech therapy purposes, where exercises of facial muscles that involve asymmetrical movements are of great importance.

As the face tracker works on 2D color images, the facial feature points have coordinates in the image plane. Thus, process steps to get the 3D coordinates of the points of interest are the same already

discussed in Part II - Section 6.2.1. We exploited the intrinsic and extrinsic camera parameters obtained in section 7.2.2 during camera calibration. In this work the lens distortion parameters were not taken into account since the Kinect color camera uses lenses with low distortion [190].

7.2.4 Artefacts correction

The depth images are estimated by means of a structured light coding, thus in some cases (in particular during the opening phase) it was difficult to estimate the distances from the camera plane for the area inside the mouth. Frequently this area assumes zero values. Since lips movements are fast during the repetition of the syllable /pa/, sometimes the tracked points on the external contour of the lips could be located on "border" areas were the depth value is zero. Indeed, according to equations (6.1) and (6.2), if Z (frontal axis) is equal to 0 also X (lateral axis) and Y (vertical axis) are equal to zero. This problem results in the presence of artefacts in the trajectories, as shown in Fig. 7.3. The time intervals where the trajectory is equal to zero were detected and corrected through a nearest neighbor interpolation in order to remove these artefacts. This method was preferred to other techniques (such as linear or cubic interpolation) for its low memory requirements and the fast computation time that make it suitable for real-time applications. An example of the artefacts correction process is reported in Fig. 7.3.

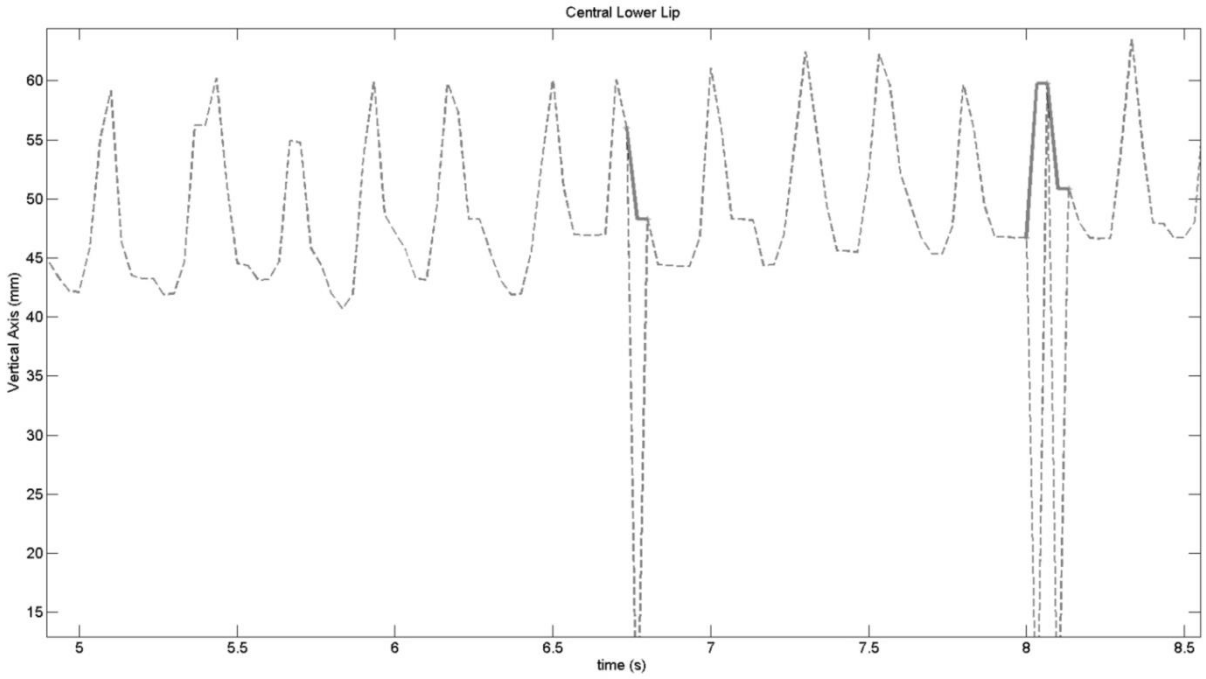

Fig. 7.3: Artefacts correction for lips trajectories. The time intervals where the trajectory (dashed line) is equal to zero are detected and corrected through a nearest neighbor interpolation (solid segments). This figure refers to the vertical trajectory of the Central Lower Lip (CLL) point during the syllable repetition task.

7.2.5 3D kinematic parameters

After the 3D coordinates of the facial feature points were computed, the velocity and acceleration on the vertical axis were calculated for the central point of the lower lip (CLL point, Fig. 7.2). We considered just the movements on the vertical axis since they are the most important ones during the repetition of the syllable /pa/. First, the trajectory of interest was smoothed with a 5-point moving average window (165 ms) to avoid large distortions of the curve especially during fast repetitions. Then the minimum values (that correspond to the closure time instants) were detected. The time interval between two consecutive minima corresponds to the period of a single syllable repetition. Thus, for each time interval, the velocity and the acceleration on the vertical axis were computed as the first and the second time derivatives, respectively. The maximum and minimum values were computed for both velocity and acceleration of each repetition. The maximum value is relative to the opening phase ($v_{opening}$ and $a_{opening}$) while the minimum value refers to the closing phase ($v_{closing}$ and $a_{closing}$) (Fig. 7.4).

For each repetition we also computed the opening of the lips as the difference between the vertical coordinates of the central points on the lower lip (CLL) and on the upper lip (CUL) (Fig. 7.2). This distance was normalized with respect to the head roll angle (rotations around the frontal axis). We did not take into account the rotations around the other 2 axes as the camera was placed in front of the subject's face at a height close to that of the eyes. Finally, for each repetition the following parameters were computed:

- normalized range of opening $\Delta Opening_{norm}$: difference between the maximum and the minimum opening values divided by its mean value;
- normalized maximum opening value $MaxOpening_{norm}$: the maximum opening value within a repetition, divided by the width of the lips. The width was calculated as the difference between the X-coordinates of the points Rcorner and Lcorner normalized with respect to the roll angle of the head.

These normalized values allowed taking into account the anatomical variations among subjects that result in different values of opening and width of the mouth.

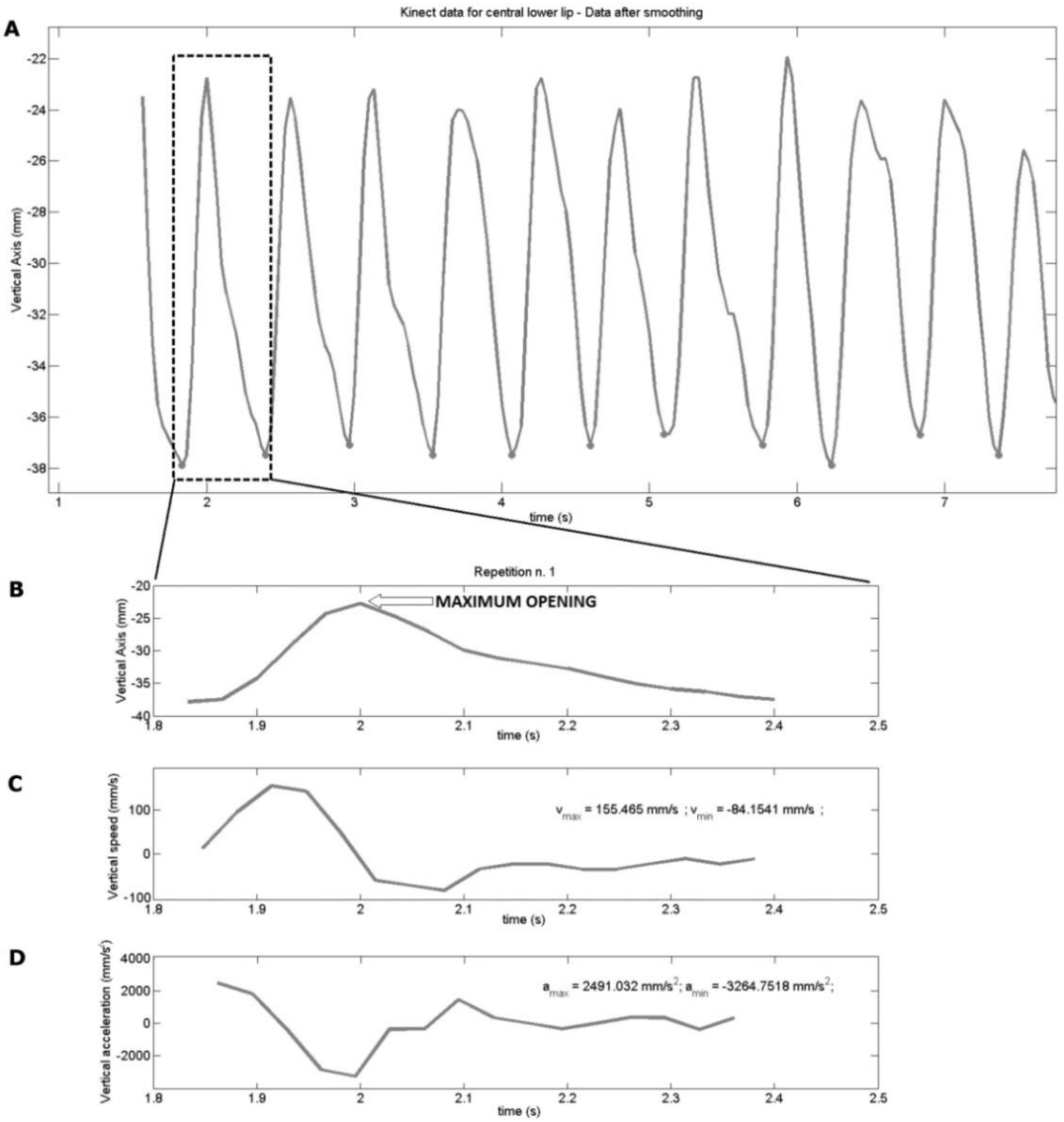

Fig. 7.4: Computation of the kinematic parameters (velocity and acceleration). A) The minima of the vertical trajectory of point CLL (solid line in the upper plot) are detected (dots) to separate each repetition. B) Vertical trajectory of point CLL of a single repetition. C) Velocity on the vertical axis calculated as the first time derivative of the trajectory. D) Acceleration on the vertical axis calculated as the second time derivative of the trajectory. The maximum velocity and acceleration values are relative to the opening phase, while the minimum values refer to the closing phase.

7.2.6 2D kinematic parameters

As the repetition of the syllable /pa/ involves main movements on the vertical axis, we computed the same kinematic parameters defined in the previous section (opening and closing velocity and acceleration, normalized range of opening and normalized maximum value of opening) from only 2D images. Indeed, if differences between groups could be detected also through 2D analysis, the advantages of this markerless system for monitoring and rehabilitation of speech diseases might be extended to the use of a simple webcam without requiring a depth stream. This would allow a further reduction of the costs making the system easier and applicable to smartphones and tablets.

Starting from the 2D face points tracked by *Intraface*, the x and y coordinates of the points in the image plane were normalized with respect to their maximum values along the whole task. The maximum y coordinate is the vertical coordinate of the point CLL in the maximum opening instant while the maximum x value is that of the external point of the left eyebrow (Fig. 7.2). Thus, the coordinates assume values between 0 and 1 for all the subjects. After the normalization the same procedure applied to the 3D kinematic parameters was used here.

7.2.7 Statistical analysis

A two-tailed t-test was performed to assess the significance of differences between PD patients and HC subjects. The degrees of freedom of the distribution are equal to 28. The difference was considered significant for $p < 0.05$.

8. Analysis of facial expressions and movements in PD patients

In this chapter the methods for facial expressions analysis and recognition in PD patients are described. As already introduced in Part I - Sections 1.2.2 and 3.3.2, some of the main motor signs of Parkinson's disease are hypomimia and facial bradykinesia. PD patients often experience serious difficulties in displaying both voluntary and spontaneous facial expressions. Thus, an objective evaluation of this sign is essential for the assessment of hypomimia (as an aid to the facial expressions item of UPDRS part III) and for rehabilitation, in particular for speech therapy.
In this project, facial expressions of PD patients were studied during the displaying of basic expressions. Some of the methods implemented here will be used also for studying facial features in DOC patients (Part II - chapter 10), for the automated analysis of reactions to standardized stimuli.

8.1 Experimental settings

As previously introduced (Part II - Chapter 4), PD patients were evaluated during audio-visual recordings for the analysis of speech and facial expressions impairments. In fact, the facial expressions task was originally proposed to study facial hypomimia and facial bradykinesia during the UPDRS motor signs evaluation. According to the UPDRS item, the clinician has to observe the patient at rest in a seated position for at least 10 seconds. During that period, the following features are observed by the clinician: eye-blink frequency, loss of facial expressions, spontaneous smiling and lips opening. In order to perform the experiments, this task was revised in accordance with expert clinicians. In the next sections, sample groups recruited for this experiment and the experimental setup will be described.

8.1.1 Subjects

17 PD patients were recruited at the Department of Neurology of the Hospital "San Giovanni di Dio", Firenze, Italy. Patients' age ranged from 53 to 83 years (mean: 71.9 years; standard deviation SD: 9.2 years). 13 patients were male, 4 were female. At the time of the experiment, disease duration ranged from 2 to 20 years (mean: 8.2 years, SD: 5.0 years). Before the experiment each patient underwent a neurological examination. The Hoehn and Yahr disease stage [16] ranged from 1.5 to 3 (2.1 ± 0.4) and the UPDRS motor score (UPDRS part III [17]) ranged from 6 to 43 (17.5 ± 10.3). All PD patients were under levodopa medication and were tested during their "on" state.
A group of 17 healthy subjects was tested as control group (age: 53-84 years, mean: 68.8 years, standard deviation: 7.5 years), 6 male and 11 female. A summary of subjects' characteristics is reported in Tab. 8.1.

All subjects were Italian native speakers. Signed informed consent was obtained from all the participants.

Table 8.1 – The dataset

	PD patients		HC subjects	
	Mean	SD	Mean	SD
Age (years)	71.9	9.2	68.4	7.5
Male	13		6	
Female	4		11	
Disease duration (years)	8.2	5.0	-	
Hoehn & Yahr stage	2.1	0.4	-	
UPDRS motor score	17.5	10.3	-	

8.1.2 Experimental setup

Each subject was asked to perform the following tasks:

- Displaying a neutral expression for at least 10 seconds;
- Displaying basic expressions (happiness, anger, disgust and sadness) upon request of the clinician;
- Displaying basic expressions (happiness, anger, disgust and sadness) by imitating emotive faces shown on a screen (Fig. 4.1 in Part I - Chapter 4);

Thus, for each subject we obtained: 1 neutral video and 8 expressive videos (4 with acted/requested expressions and 4 with imitated expressions).

The subjects' face was recorded using the Microsoft Kinect for Windows sensor as already described in Part II - Chapter 7. The Kinect sensor was placed in front of the subject's face at a distance between 0.5 and 0.7 m from the mouth and at a height close to that of the subject's eyes. However, unlike the experiments performed on speech articulatory movements, in this case we used just the color stream (like a standard webcam). Videos were acquired at 640x480 pixels at 30 frames per second (fps). Color frames were recorded in 24-bit RGB images (8 bits per channel). Color stream was recorded and stored in *avi* files through the OpenNI (ver. 2.2) and OpenCV (ver. 2.4.9) libraries using a customized code written in C++ language.

The experiments were carried out in a quiet room of the "San Giovanni di Dio" hospital and subjects were required to stay seated during the test.

8.2 Methods

On the recorded videos, two tests were performed:

- The first one consists in the analysis of facial features with respect to the neutral baseline, in order to find the most discriminative features between HC subjects and PD patients,

characterizing the deviant PD expressions, and giving an objective quantification of facial hypomimia in PD;

- The second test concerns the implementation of automatic facial expressions recognition algorithms, to study how much far the PD expressions are from the standard expressions (in terms of classification results and percentage of a targeted expression).

In the first case we want to measure the amount of facial movements in PD patients in order to find objective indices of facial mimicry in impaired expressions.

In the second case, the aim is that of assessing the intensity of facial expressions that PD patients are able to reach, with respect to healthy control subjects. This could provide useful information and real-time feedback for speech therapy, in order to produce more enhanced facial movements, reaching a higher intensity of the expression.

8.2.1 Analysis of expressive features with respect to the neutral baseline

Each video was manually labeled, detecting the acted or the imitated expression (i.e. expected expression). For each PD patient/HC subject we have 9 videos: 1 for the neutral state, 4 for the posed expressions, 4 for the imitated expressions (with the exception of a patient that was not able to display sadness and disgust requested by the clinician). The duration of these videos is variable and ranges between 3 s and 12 s.

Face tracking and facial features

The automatic identification of facial landmarks was performed through the *Intraface* tracking algorithm [127], already used to study facial and articulatory movements in PD patients and HC subjects during various speech tasks (Part II - Chapters 6 and 7). Starting from the 49 facial landmarks tracked by this algorithm (Fig. 8.1), according to Soleymani et *al*., 2012 [130] we focused our attention on the following facial features (f1-f20, Fig. 8.2):

- Eyebrows: Angles between the horizontal line connecting the inner corners of the eyes and the line that connects inner and outer eyebrow (f1, f2), the vertical distances from the outer eyebrows to the line that connects the inner corners of the eyes (f3, f4) (see Fig. 8.2a).
- Eyes: Distances between the outer eyes' corner and their upper eyelids (f5, f9), distances between the inner eyes' corner and their upper eyelid (f6, f10), distances between the outer eyes' corner and their lower eyelids (f8, f12), distances between the inner eyes' corner and their lower eyelids (f7, f11), vertical distances between the upper eyelids and the lower eyelids (f13, f14) (see Fig. 8.2b).
- Mouth: Distances between the upper lip and mouth corners (f15, f16), distances between the lower lip and mouth corners (f18, f18), distances between the mouth corners (f19), vertical distance between the upper and the lower lip (f20) (see Fig. 8.2c).

The line that connects the inner eye corners was used as a reference line since the inner eye corners are stable facial points, i.e., changes in facial expression do not induce any changes in the position of these points. For each sequence, the aforementioned 20 features were extracted for all the frames.

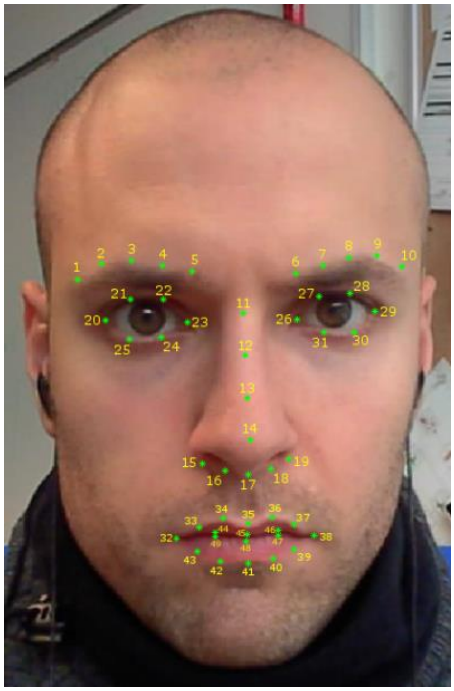

Fig. 8.1: Indexed Intraface points from which the facial features were computed

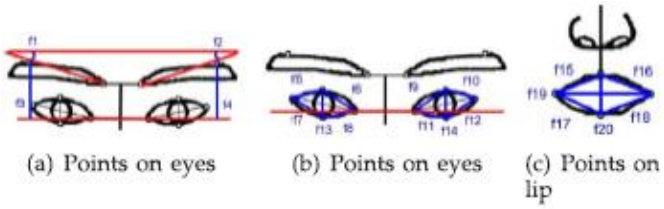

Fig. 8.2: facial features extracted from the Intraface points: a) Eyebrows features, b) eyes features and c) mouth features [130].

Baseline Building

As PD patients may exhibit impairments of facial expressions/movements, a baseline of the neutral state was built, considering an average template from the neutral videos. Thus, each PD patient and HC subject has his own neutral baseline (i.e. a vector of 20 facial features), from which we can compute the displacement of the facial features during the tasks. The neutral baseline was built as follows:

- for each video frame of the neutral video the 49 facial landmarks were detected by means of the *Intraface* tracking algorithm;
- to define geometric features, an average facial template from the facial landmarks vector was iteratively built by Procrustes analysis. For each video frame, each facial landmark was aligned to the template, updating the template by averaging the aligned landmarks [118];

- once the average facial template was computed (i.e., a vector of 49 2D points), the 20 facial features were calculated as above.

Facial features extraction from the expressive videos

After the set-up of the neutral baseline, the analysis was performed on the expressive videos (both acted and imitated expressions). Thus, for each subject and for each expressive video:

- the 49 2D facial points provided by the *Intraface* tracking algorithm were extracted from each video frame;
- for each frame the extracted facial model was aligned to the neutral template of the current subject by affine transformation (that includes rotations, translations, scaling and skewing), in order to suppress within-subject head pose variations. This affine transformation was estimated from 4 pairs of corresponding points in the facial model of the current frame and in the average neutral model: the two inner corners of the eyes, the nose tip and the point between the two eyes (points 23, 26, 14 and 11 in Fig. 8.2). These points were chosen because of their stability with respect to non-rigid facial movements [131].
- Once the face model for the current video frame was aligned to the neutral template, the 20 facial features were extracted as above.

Thus, for each expressive video, a vector of 20 features for each frame was obtained.

Facial features analysis

Given the extracted parameters (i.e. facial features of the neutral baseline and facial features from the expressive videos), for each subject and for each expressive videos, the following processing was performed. For each video frame the Euclidean distance was computed between the facial features vector and the baseline features vector built for the current subject. This distance provides global information about the displacement of facial features from the neutral expression, during the displaying of different facial expressions. On the Euclidean distance vector, the following statistics were calculated along the duration of the expressive video: mean value, median, standard deviation, skewness, kurtosis, maximum value, minimum value and range (i.e., difference between maximum and minimum values).

Thus, for each subject and for each expressive video, we got: 8 statistic features of the Euclidean distance between the 20 “expressive” facial features and the neutral baseline.

8.2.2 Automatic facial expression recognition

In this part of the study an automatic facial expression classifier was trained on different databases of posed and spontaneous expressions. However, most of the available facial expressions databases are composed by videos and images taken from healthy subjects. Starting from this consideration, the aim is to measure the intensity of facial expressions that PD patients are able to reach, with respect to

healthy control subjects that are assumed to express "standard" expressions. This measure is extracted through the prediction given by the classification algorithms.

Since supervised learning methods were implemented, this part is divided into two sub-sections:

- the training phase: where the classification algorithm is trained on labeled images;
- the test phase: where the trained algorithm is tested over the subjects recruited in this study (both PD patients and HC subjects).

Video recordings considered for the test phase are the same used for the analysis of facial features with respect to the neutral baseline.

Training phase

The main items to be considered during this phase are: the facial features to feed the classifier, the pre-processing step on facial features, the classification algorithms, the databases used for training the facial expressions classifier and the number of expressions taken into account. The choice of these items is of crucial importance and must be carefully evaluated, in order to get good performances in terms of classification accuracy.

Concerning facial features, we decided to use the same 20 geometric features described and used in the previous section. This choice is reasonable in terms of dimensionality (20 features) and characterization of the face, since we can describe the behavior of the most important components of the face (eyebrows, eyes and mouth).

Two classifiers are used: kernel ridge regression and Multi-class Support Vector Machine (SVM). In both cases, one-versus-rest classifiers were built in order to perform a multiclass classification, training a single classifier per class.

SVM is a supervised learning algorithm that divides a feature space into two classes through a linear separation (i.e., a hyperplane). This hyperplane is built exploiting the maximum separation margin between the two classes. One of the biggest advantages of SVM is the capability to deal also with problems in which features are not linearly separable. Through the so called "kernel trick" it is possible to map the original features in a higher-dimensional space in which features can be separable. Popular kernel classes are: polynomial functions and Gaussian radial basis function (as the one used in this work). The kernel trick can be used also for other classification algorithms such as ridge regression. Ridge regression differs from linear regression for the presence of a penalty term on the size of model parameters to be estimated. Kernel ridge regression is similar to SVM, especially in its non-linear version; however, the objective to optimize depends on all the training examples and not just on a subset of them (support vectors) as in case of SVM [191,192,193].

Two databases are considered: the Extended Cohn-Kanade database (CK+) and the Radboud Faces database (RaFD) [137,141]. The CK+ database contains facial expressions from 210 adults of different age (18-50 years old), cultures and races. Participants performed a series of 23 facial displays, including single action units, combination of action units and posed expressions (happy, sad, disgust,

fear, angry, surprise and contempt). Each sequence begins with a neutral expression. Of the whole database, 327 image sequences include a nominal emotional label, manually checked from certified FACS coders, thus we assume this set as reliably annotated. The number of occurrences for each stereotypical expression is reported in Tab. 8.2.

Tab. 8.2: number of occurrences for each facial expression class in the CK+ database

Emotion label	number of images
Angry	45
Contempt	18
Disgust	59
Fear	25
Happy	69
Sadness	28
Surprise	83

Afterwards, in order to enrich the database and make it more balanced, for each subject and for each expression sequence we considered the last 4 frames (i.e. the frames were the particular facial expression is more enhanced). In contrast, in order to build the neutral expression class, for each subject and for each sequence, we chose the first frame. Thus, we selected the following instances: 327 for neutral, 180 for anger, 72 for contempt, 236 for disgust, 100 for fear, 276 for happy, 112 for sad and 332 for surprise, for a total of 1635 frames.

The RaFD database is composed by images of posed expressions obtained from 57 adults (Caucasian males and females and Moroccan Dutch males) and 10 Caucasian Dutch children (both boys and girls). Each subject showed eight facial expressions (neutral, anger, sadness, fear, disgust, surprise, happiness, and contempt, the same of the CK+ database) with three gaze directions and were got from five different camera angles (in steps of 45°) [137]. Thus the total amount of images for each subject is 120, for a total number of images contained in the database of 8040. In our case, we excluded: children images and profile images, taking into account only 3 viewpoints of the acquisitions (45°, 90° - frontal view, 135°). Thus we obtained 513 samples for each expression, for a total of 4104 images.

Considering the pre-processing step, before training the classifier, a landmark template from training data was built by Procrustes analysis. For each frontal training face each facial landmark (extracted with *Intraface*) was aligned to the template, updating the template by averaging the aligned landmarks. After the template was built, each shape of the training set was aligned to the template by estimating the affine transformation. This step is essential to suppress within-subject head pose variations and inter-subject geometric differences. As already introduced in the past section, the affine transformation was estimated from 4 pairs of corresponding points: the two inner corners of the eyes, the nose tip and the point between the two eyes (on the top of the nose). These points were chosen because of their stability with respect to non-rigid facial movements.

In order to find the best combination of components (i.e., classification algorithms, databases, number of expressions) in terms of classification accuracy, we combined all the possible solutions among:

- Two possible classifiers: Kernel Ridge Regression or Multi-Class SVM;
- Three possible databases: RaFD, CK+ or both databases combined together;
- Three possible expression sets: all expressions (8 expressions - 7 and neutral), excluding contempt (6 and neutral) and considering only the expressions used during our acquisitions (neutral, anger, disgust, happy and sad);

Thus, we obtained 18 different possibilities that were evaluated with a 10-fold cross-validation to extract the accuracy of the classification. The accuracy for each test is reported in Tab. 8.3.

Tab. 8.3: different tests performed with different classifiers, databases and set of emotions.

Classifier	DB	Emotions	Accuracy
Kridge regression	RaFD	All (7 + neu)	65.28%
Kridge regression	RaFD	No contempt (6 + neu)	73.10%
Kridge regression	RaFD	5 (neu, ang, dis, hap, sad)	80.66%
Kridge regression	CK+	All (7 + neu)	70.28%
Kridge regression	CK+	No contempt (6 + neu)	73.45%
Kridge regression	CK+	5 (neu, ang, dis, hap, sad)	78.69%
Kridge regression	RaFD & CK+	All (7 + neu)	62.75%
Kridge regression	RaFD & CK+	No contempt (6 + neu)	69.91%
Kridge regression	RaFD & CK+	5 (neu, ang, dis, hap, sad)	75.57%
Multiclass - SVM	RaFD	All (7 + neu)	80.56%
Multiclass - SVM	RaFD	No contempt (6 + neu)	84.49%
Multiclass - SVM	RaFD	5 (neu, ang, dis, hap, sad)	86.08%
Multiclass - SVM	CK+	All (7 + neu)	96.51%
Multiclass - SVM	CK+	No contempt (6 + neu)	97.06%
Multiclass - SVM	CK+	5 (neu, ang, dis, hap, sad)	98.41%
Multiclass - SVM	RaFD & CK+	All (7 + neu)	82.56%
Multiclass - SVM	RaFD & CK+	No contempt (6 + neu)	86.24%
Multiclass - SVM	RaFD & CK+	5 (neu, ang, dis, hap, sad)	87.56%

From table 8.3, the following observations can be drawn:

- Multi-Class SVM gives higher accuracies than Kernel Ridge Regression (always higher than 80%), thus the first 9 attempts in Tab. 8.3 can be excluded from the selection process;
- Experiments with only the CK+ database seem to give better results. However, this database, if considered alone, is not well balanced (despite the enrichment method adopted).

For these reasons we decided to choose as classifier for the testing phase the Multi-Class SVM, trained on both databases (RaFD and CK+) with the 5 expressions used in our acquisition (last row of Tab. 8.3). Since our experiments were controlled and we knew which was the target expression that each subject had to reach during the single task, we decided to force the classifier to recognize only these expressions. Thus, the landmark template for the alignment of the training shapes was built using frontal faces from both databases.

In Tab. 8.4, confusion matrix and other measures of the classification performances (precision, recall and f-measure) are reported for the chosen case. Accuracy is defined as the ratio between the correct classifications and the overall number of classifications (sum of the diagonal elements of the confusion

matrix divided by the sum of all the elements of the matrix). The other measures are: precision, defined as the ratio between true positives and the sum of true positives and false positives; recall, the total number of true positives divided by the sum of true positives and false negatives; F-measure, defined as:

$$2\frac{(Precision*Recall)}{(Precision+Recall)} \tag{8.1}$$

and ranges between 0 (worst results) and 1 (best results).

Tab. 8.4: Confusion matrix, accuracy, precision, recall and f-measure for the selected classifier

								Instances		Accuracy	Precision	Recall	F-Measure
		Predicted											
		Neutral	Anger	Disgust	Happy	Sadness		3696					
Actual	Neutral	714	37	41	4	44		840		0,875541	0,830233	0,85	0,84
	Anger	42	608	19	0	24		693			0,882438	0,877345	0,879884
	Disgust	9	3	736	1	0		749			0,823266	0,982644	0,895922
	Happy	2	2	17	767	1		789			0,992238	0,972117	0,982074
	Sadness	93	39	81	1	411		625			0,85625	0,6576	0,743891

Test phase

Once the classifier has been trained, we use as test set our database composed by HC subjects and PD patients recruited during the experiments. Thus, for each subject and for each expressive video (8 videos with 4 imitated expression and 4 acted expressions), we got the predicted facial expression label for each frame of the video. Since the Multi-Class SVM is a one-vs-rest classifier, each classifier produces the predicted class likelihoods by giving two scores: one corresponding to the negative class and one corresponding to the positive class. For instance, if the classifier is trained to recognize between happy and non-happy instances, the first score concerns the non-happy class, while the second score the happy class. Thus, the predicted facial expression label was determined as the maximum among all the positive scores returned by the Multi-class SVM (winner-takes-all strategy).

9. Contact-less video-based tracking of heart rate

As explained in Part I - Section 3.4, several approaches were proposed to process the reflected plethysmographic signal of the skin recorded with a webcam. Most of them rely on the processing of the color channels of RGB video frames. The chromatic variations due to the cardiac rhythm are estimated by means of the Independent Component Analysis (ICA) [146]. Following this approach we propose a recursive method for the heart rate (HR) estimation during a mild physical exercise. The accuracy of the proposed method is compared to the electrocardiographic (ECG) data and to the results obtained with the continuous wavelet transform (CWT) on four healthy subjects.

9.1 Experimental settings

To elicit a HR variation four healthy control subjects sat on a cycling ergometer during the experiment. The experiment duration was 450s, of which: 180s at rest, 90s during mild exercise (setting a power of 65 W and pedaling at 60-70 cycles per minute) and 180 s at rest. The ECG signal was recorded with a video-EEG system (EBNeuro Mizar-Sirius, 128 Hz sampling frequency) and the subject's face was recorded with a Creative Senz3D webcam. Video resolution was 1280x720 pixels at 30 frames per second (fps) and color frames were stored in 24-bit RGB images (8 bits per channel). The ECG signal was extracted from the precordial leads V1 and V6. Subjects had to maintain the forehead free from hair or other occlusions during the whole task. This requirement allows extracting color variations from the skin's surface without any other contribution due to external factors. Moreover, subjects had to avoid abrupt head movements, for a stable video recording of the face.
Color videos were recorded and stored in *avi* files through the OpenCV (ver. 2.4.9) libraries using a customized code written in C++ language.

9.2 Methods

The HR estimation process is made of different steps: face detection and tracking, identification of the Regions of Interest (ROIs) in the face, extraction and pre-processing of the color variations within the selected ROIs, ICA on the pre-processed signals to detect the independent components that generated those variations and HR estimation on the processed components.

9.2.1 Face tracking and ROIs extraction

The first step was the automated identification and location facial landmarks (eyebrows, eyes, nose and mouth) in the video frames. To perform this task, the *Intraface* tracker algorithm [127] was used, as already done for the other experiments described in the previous chapters (Part II - Chapters 6-8). Once the face and its main features (eyebrows, eyes, nose and mouth) were detected within the image

plane, two ROIs were defined, in order to extract the R, G and B channel variations. Starting from the 49 tracked points, the following ROIs were automatically located: one on the forehead and one in the central face region (including cheeks and nose), as reported in Fig. 9.1. These regions were chosen for their higher skin blood perfusion, for their higher visibility and uniform illumination. These regions were already used in literature [154,194,195] for video-based HR estimation.

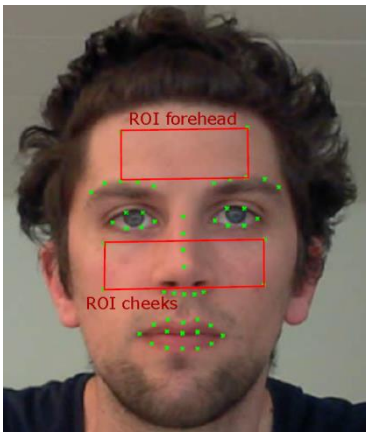

Fig. 9.1: ROIs considered for this study, selected from the *Intraface* landmarks.

According to the landmarks distribution provided by the *Intraface* tracking algorithm (Fig. 8.1, Part II - Section 8.2.1), these ROIs were selected as follows:

- Forehead ROI: the two lower vertices of the rectangular ROI were selected in correspondence of points 3 and 8 (central points of the eyebrows). Their horizontal distance defines the ROI width. Starting from these points, the upper vertices of the rectangle were located at a vertical distance of size equal to 1/3 of the width. This value was selected in order to fit the ROI within the forehead, without considering hair regions.
- Cheeks ROI: the two upper points of the rectangular ROI were selected at the horizontal coordinate of points 2 and 8 (the second external eyebrows points) and at a vertical coordinates between points 12 and 13 (the two central nose points). Then, the lower vertices of this ROI were located at a vertical coordinate between points 14 (nose tip) and 17 (central nose base).

ROIs were normalized with respect to the head rotation angles, in order to make their positioning integral with head movements.

9.2.2 RGB components extraction and pre-processing

Starting from the two ROIs the R, G and B signals were extracted by averaging pixels values (on the different 3 channels) along the whole ROI. Thus, for each frame an average value of R, G and B was obtained for both ROIs. This process was repeated for the whole video duration, in order to extract 3 signals for each ROI.

Afterwards, to remove low frequencies a detrend based on a smoothness priors approach [159] was applied. This detrend, already used in [146,148,154], operates as a time-varying high pass FIR filter and the frequency response can be adjusted by setting a single smoothing parameter λ, that in our case was set equal to 100 (corresponding to a cutoff frequency of the filter of 0.66 Hz). Thus, events with frequency below 39 beats per minute (bpm) were filtered out.

The detrended signals were then normalized in amplitude subtracting their mean value and dividing by the standard deviation. In Fig. 9.2 the result of detrend and normalization on R, G and B signals is reported.

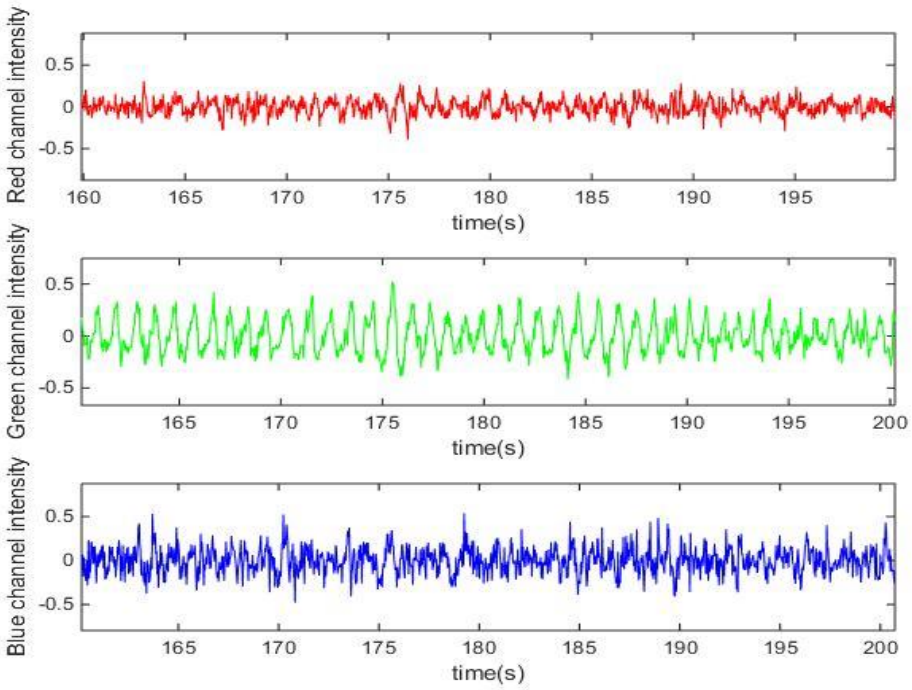

Fig. 9.2: mean R, G and B channel intensities after detrend and normalization. The highest oscillations consistent to the blood pulsations in the pre-processed green channel [148,149].

9.2.3 Estimation of the independent components

As introduced in Part I - Section 3.4, the idea behind video-based methods to estimate HR is to exploit the volumetric variations due to the cardiac cycle that cause changes in blood vessels of the facial skin. These variations modify the path length of the incident light, so that the reflected light can contain information about the cardiac cycle [148]. The hemoglobin absorptivity varies across the visible and near-infrared spectral range, with the highest peak in the green/yellow wavelength range (around 510-590 nm) [148,149]. In fact, Fig. 9.2, highlights that the highest oscillations consistent to the blood pulsations are in the pre-processed green channel.

However, video recordings performed with RGB color cameras contain a mixture of the reflected plethysmographic signal with other underlying sources. In particular, the pre-processed signals considered until now are a mixture of different sources: skin color variations due to the blood

perfusion, light variations due to the sunlight or to the artificial illumination flickering, noise, shadows due to the head movements, etc. Thus, the purpose of ICA is to unmix these sources, identifying the variations due to the cardiac cycle. In this part of the work the decomposition into three source signals was performed through the ICA based on the Joint Approximate Diagonalization of the Eigenmatrices (JADE) algorithm, already used for the same purpose in [146,148,157,160].

On the estimated sources a smoothing with a 5-point moving average window and a bandpass filtering between 0.6 and 3.5 Hz with a 128-points Hamming window were applied. In particular, the band-pass filtering allows restricting the frequency range of the heart rate between 36 and 210 bpm. The three filtered sources are reported in Fig. 9.3.

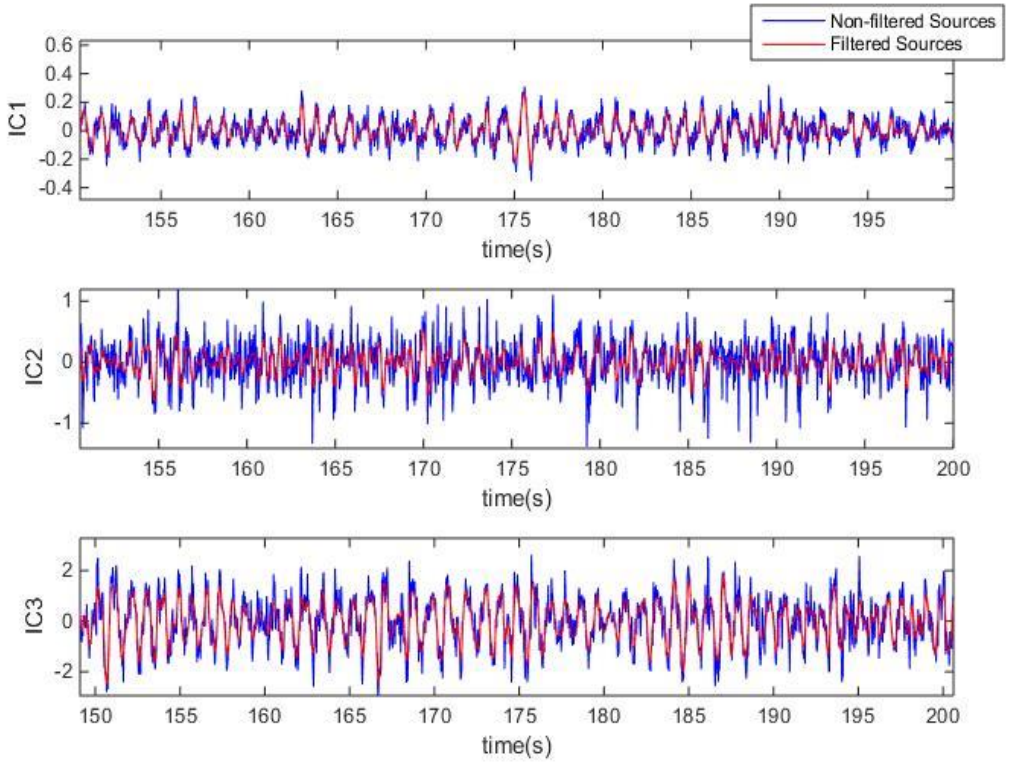

Fig. 9.3: Filtered sources obtained after the ICA with JADE algorithm

9.2.4 HR estimation

Autoregressive method

To track the HR variations an autoregressive (AR) model with recursive least squares (RLS) estimation was applied. Through an AR model (of order p), a signal y(t) can be expressed as:

$$y(t) = -\sum_{k=1}^{p} \alpha_k y(t-k) + u(t) \tag{9.1}$$

where α_k (k = 1…p) are the p+1 unknown model parameters and u(t) is a zero-mean white noise process with variance σ_u^2. Through the estimation of α_k and σ_u^2, the power spectral density can be computed as follows:

$$P_{AR} = \frac{T\hat{\sigma}_u^2}{\left|1+\sum_{k=1}^{p} a_k e^{-j2\pi fkT}\right|} \tag{9.2}$$

with T equal to the sampling period of y(t). The basic RLS algorithm assigns the same weight to all the past data for the estimation of the actual one. A more refined version is provided with the so-called "forgetting factor" λ that assigns less weight to older data. Specifically data older than T_0 samples are considered with a weight around 36% of the most recent one. T_0 is defined as:

$$T_0 = 1/(1-\lambda) \tag{9.3}$$

According to formula 9.3, a forgetting factor λ=0.9889 allowed taking into account measurements up to the previous 90 samples (3 s). At each time instant the first peak of the AR power spectral density (PSD) was estimated and converted in bpm. A low model order p=4 was found capable of smoothing the PSD peaks giving a single maximum in the frequency range of interest (Fig. 9.4) [196].

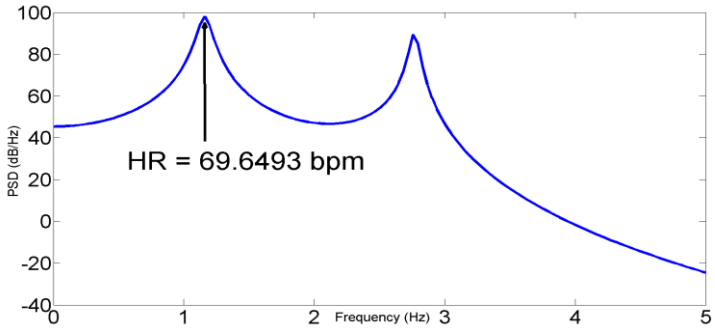

Fig. 9.4: AR Power Spectral Density (PSD) obtained by a RLS estimation of the parameters. In a given instant the HR is the frequency value of the first peak, converted in bpm.

Complex Wavelet Tansform

For each one of the filtered sources the CWT was computed using a complex Morlet as mother wavelet. The complex Morlet function is obtained from the product between a Gaussian function and a complex exponential, according to the following formula:

$$\psi(t) = \frac{1}{\pi^{1/4}} e^{j2\pi f_0 t} e^{-t^2/2\sigma^2} \tag{9.4}$$

where the first term is a normalization factor, f_0 is the central frequency of the mother wavelet and σ is the standard deviation of the Gaussian function. The Gaussian standard deviation defines the scale of the wavelet; in fact, since the 99.7% of the area under a Gaussian function is within an interval of 6σ,

it is possible to extract the central frequency of the wavelet by setting the number of significant oscillations of the wavelet:

$$f_0 = \frac{n}{6\sigma} \tag{9.5}$$

This relationship derives from the assumption that n oscillations are included in an interval of 6σ. The higher σ the lower the central frequency f_0. In this work, we set n = 7.

Afterwards, the peak frequency was computed as the weighted average of the peaks within the frequency range of interest (0.6-3.5 Hz). The peaks amplitudes were considered as the weights in this calculation. An example of the spectrogram obtained from the second independent component is reported in Fig. 9.5

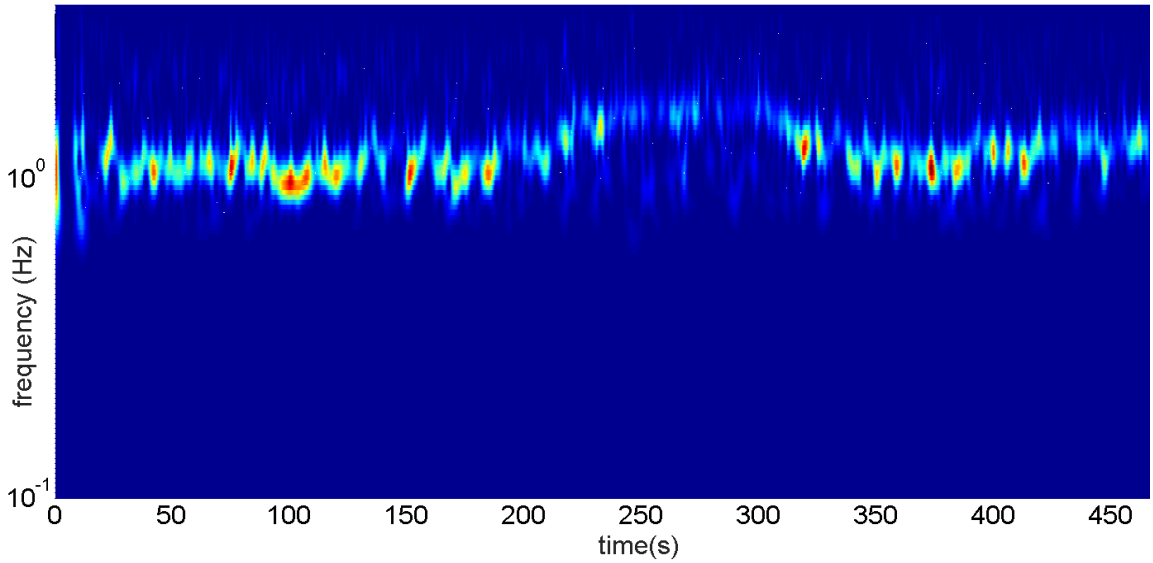

Fig. 9.5: Spectrogram obtained for the second independent component. It is possible to see a high energy band that lies in the frequency range of interest (0.6-3.5 Hz) and varies in correspondence to the HR variations due to the physical exercise (between 200 and 300 s).

9.2.5 Error calculation

Once the HR estimations were extracted for both ROIs (forehead and cheeks) and for both estimation methods (AR models and CWT), they were resampled at 128 Hz (sampling frequency of the video EEG) using a spline interpolation technique. Thus, the comparison with the ground truth HR values was made according to the root-mean-square error (RMSE) in bpm, given by:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{9.6}$$

Where N is the number of samples of the video recording, y_i is the *i-th* sample of the reference HR and $\hat{y}_i$ is the corresponding HR estimation with the video-based technique.

10. Video Analysis of DOC patients: facial movements and vital signs

In this chapter we propose to study facial expressions and heart rate in patients with disorders of consciousness (DOCs) by means of low cost and contactless techniques. The aim of this part of the work is the implementation of a markerless method for monitoring and studying DOC patients, in particular after the administration of standardized stimuli (such as those included in the CRS-R protocol - Part I - Chapter 2). The proposed system merges some of the methods already described in the previous chapters, in particular: the analysis of facial expressions (Part II - Chapter 8) and the contactless estimation of the heart rate (Part II - Chapter 9).

The aim is to extract quantitative and objective information to help clinicians in diagnostic assessment through the analysis of possible reactions after standardized stimuli in DOC patients. Moreover, this system could be used as an aid to the monitoring of these patients, even when there are no caregivers or family members near them.

This will lead to the development of a tool for monitoring of patients in vegetative/minimally conscious state, capable of automatically assess reflex and cognitive-mediated behaviors during the whole day, relating them to possible external stimuli. Moreover, it will bring more information both to the clinicians and relatives about the level of pain of the patient.

10.1 Experimental settings

DOC patients studied in this chapter were evaluated during the administration of the CRS-R. Although all the CRS-R items were video recorded, only the motor evaluation function was considered for this study. During this item a noxious stimulus is administered to hands or feet. The presence or absence of flexor response determines the score (Part I - Chapter 2). However, unlike the original purpose of this item, our aim was the study of facial expressions and vital signs after the administration of the noxious stimulation. It is thus possible to assess facial features and vital signs related to the level of pain felt by the patient. Signed informed consent was obtained from family member or caregivers of the participants.

10.1.1 Subjects

Nine DOC patients were video recorded at the “Villa delle Terme” rehabilitation center (Impruneta, Firenze). Seven patients were male and two were female. At the time of the experiments the patients’ age ranged from 35 to 80 years (mean value: 53.4 years, standard deviation: 14.2 years) and the months after the brain injury ranged from 4 and 93 (mean value: 31.7 months, standard deviation: 27.6 months). For five patients, the etiology was post-anoxic/ischemic, four patients had a cerebral

hemorrhage and one patient a traumatic brain injury. The total CRS-R score assessed at the moment of the experiment ranged between 4 and 9 (mean value: 6.2, standard deviation: 1.5).

10.1.2 Experimental setup

During the CRS-R administration, patients' faces were recorded with the Creative Senz3D camera, under a constant and uniform illumination. Despite this sensor can provide two video streams (like the Kinect), only the color stream was used for these experiments. In fact, unlike the kinematic analysis of speech movements performed in PD patients (Part II - Chapter 7 and Part III - Chapter 13), in this case the purpose was not the assessment of kinematic parameters of the articulators, as DOC patients rarely exhibit some articulatory movements, vocalizations and verbalizations. Although the depth stream could provide additional information, it was preferred to work just on the color stream, thus saving computer memory for video recordings.

Before each experiment, the camera was placed in front of the patient's face at a distance around 0.5 m from the mouth and at height close to that of the subject's eyes. The camera's optical axis was set perpendicular to the subject's face, as shown in Fig. 4.3 of Part II - Section 4.2. Thus, according to these criteria, the face can be visible inside the scene for the whole recording. During the CRS-R administration, patients were in a half-seated position in the hospital bed.

The resolution of the color stream was 640x480 pixels at 30 frames per second (fps). Color frames were recorded in 24-bit RGB images (8 bits per channel). Videos were stored in *avi* files through the OpenCV (ver. 2.4.9) libraries using a customized code written in C++ language.

At the same time, the audio signal consisting of the clinician's voice was recorded with the built-in microphones of the Senz3D sensor (Fs = 44.1 kHz, recording software: Audacity, ver. 2.0.5). The audio signals were synchronized with the video recordings. The audio was recorded just for detecting the time instants when the stimuli were administered. In particular the noxious stimulation administrations were identified in the recorded track (and in turn in the video frames) by an acoustic signal (with a similar procedure described in Part II - Section 6.1.1 "*Audio acquisitions and streams synchronization*").

Videos had different durations that range from 38 s up to 220 s. For some patients, more than one stimulus was administered.

10.2 Methods

As mentioned above, most of the processing steps are the same described in the previous chapters. In particular, for this work we merged the analysis of facial features with respect to the neutral baseline (Part II - Section 8.2.1) and the contactless video-based estimation of the heart rate (Part II - Chapter 9).

10.2.1 Facial features extraction

The analysis of facial features comes from the consideration that facial expressions could bring important information about patient's reactions to external stimuli and pain as they are unable to communicate this condition. In fact, most of the perceptual ratings performed by clinicians are based on the assessment of the facial mimicry (see PAINAD scale, Part I - Section 2.2.1). However, a standard facial expression recognition approach (i.e., training of a classifier on basic expressions and then testing on video frames, like the approach proposed in Part II - Section 8.2.2) is not well-suited for these patients. In fact, they do not exhibit standard expressions, even if sometimes they can display smiles and grimaces. In this case, a more "patient-fitted" approach that allows extracting relative information about the evolution of the studied features during tests, is more appropriate. For these reasons we decided to follow the same steps already described in Part II - Section 8.2.2. Thus, for each video recording, the following processing steps were performed:

- Neutral template building on the first second (30 frames). These frames are required and used for the neutral template, because the noxious stimuli were not yet administered. As reported in Part II - Section 8.2.2, facial landmarks were assessed by means of the *Intraface* tracker and then the baseline template was built by Procrustes analysis. On the average facial template, the 20 facial features reported in Fig. 8.3 (Part II - Chapter 8) were computed.
- On the whole duration of the video the 49 facial landmarks provided by the *Intraface* algorithm were extracted and then aligned to the neutral baseline by affine transformation. Afterwards the 20 facial features were extracted for each frame.
- For each frame the Euclidean distance was computed between the facial features vector and the neutral template built for the current subject. This distance provides a global information about the displacement of facial features from the baseline state (i.e., before the administration of the external stimulus).

10.2.2 Contactless HR estimation

The HR could provide important information to evaluate patient's reactions and the level of pain. However, a continuous and constant monitoring of HR could be expensive and uncomfortable for the presence of sensors attached to the patient's body. Thus, we estimated HR from the same videos processed to extract facial features, in the same way as already explained in Chapter 9, namely:

- Face tracking and ROIs detection. Two ROIs were automatically located with the *Intraface* tracking algorithm: one on the forehead and one in the central facial region (including cheeks and nose).
- For both ROIs the following preprocessing steps were applied to the temporal trends of the mean values of the R, G and B channels: detrend based on a smoothness priors approach

(smoothing parameter = 100, cutoff frequency = 0.66 Hz); mean value subtraction and normalization with respect to the standard deviation.

- Estimation of the three independent sources by means of ICA based on the joint approximate diagonalization of the eigenmatrices algorithm (JADE). Afterwards, a smoothing with a 5-point moving average window and a bandpass filtering between 0.6 and 3.5 Hz with a 128-points Hamming window were applied to the three sources.
- Heart rate estimation. To compute the HR values an autoregressive (AR) model with recursive least squares (RLS) estimation was applied. At each time instant the first peak of the AR power spectral density (PSD) was estimated and converted in beats per minute (bpm). A low model order p=4 was found capable of smoothing the PSD peaks giving a single maximum in the frequency range of interest. A forgetting factor f=0.9889 allowed taking into account measurements up to the previous 3s. For each one of the filtered sources the CWT was computed using a complex Morlet as mother wavelet: the peak frequency was obtained as the weighted average of the peaks within the frequency range of interest and weights are the peaks amplitudes.

10.2.3 Evaluation of patients reactions

For each patient we got: a facial features vector (20 facial features for each video frame), a distance vector (Euclidean distance between the facial features and the baseline, for each frame) and a HR estimation. Thus, each one of these vectors was divided into different parts:

- From the beginning of the recording up to the administration of the first stimulus. This phase is called “baseline”;
- After the administration of a noxious stimulus up to the next administration or to the end of the video. These phases are called “stimulus1”, “stimulus2”, etc., according to the number of noxious stimuli administered for each patient (in this study this number ranges from 1 to 4).

This decomposition allow to evaluate the behavior of heart rate, facial features and its distance from the baseline during the different phases, in order to assess possible variations due to the external stimulation.

PART III - RESULTS

11. Acoustical analysis of PD speech[3]

In this section, results obtained from the acoustical analysis of patients with Parkinson's disease, are reported. As already stated in Part II - Chapter 5, the study is divided into two main parts:

- The first part concerns performance evaluation of an AVU segmentation algorithm on a dataset of speech signals (syllable repetition task) already analyzed in [94]. Temporal parameters extracted with the AVU segmentation algorithm are compared with those provided by the authors of [94]. Moreover, a comparison between HC subjects and PD patients was performed.
- Once the accuracy of the automatic methods is tested and new temporal parameters related to dysprosody were defined, the attention was moved to the identification of dysprosodic patterns during a sentence repetition task through standard and new parameters.

11.1 Results

11.1.1 Performance of the AVU algorithm on syllable repetitions

The parameters used to compare the two methods are reported in Tab. 11.1. Concerning the comparison between PD patients and HC subjects (Tab. 11.2), both groups show similar values of D_{mean} (around 0.50 – this indicates that on average the 50% of the interval duration is made by the vocalization). The only parameter that shows significant differences between two groups is $DRelStab_{13\text{-}20}$, with HC subjects that exhibit higher values (92.97 ± 8.60 vs. 87.33 ± 10.36, p = 0.02).

Tab. 11.1: mean values, standard deviations and correlation coefficients for the parameters used to compare the AVU with the reference results extracted with the manual labeling.

Parameter	Group	Manual labeling	AVU	Correlation
IntDur (ms)	PD	545.15 ± 182.08	550.06 ± 186.59	
	HC	505.56 ± 234.33	507.28 ± 236.09	0.997
SD (ms)	PD	39.56 ± 20.15	41.56 ± 20.29	
	HC	28.50 ± 16.51	36.03 ± 19.28	0.905
$avIntDur_{1\text{-}4}$ (ms)	PD	553.77 ± 181.95	554.22 ± 184.44	
	HC	501.38 ± 219.38	509.48 ± 230.49	0.995
$avIntDur_{5\text{-}12}$ (ms)	PD	554.13 ± 189.01	557.03 ± 190.48	
	HC	512.85 ± 240.00	514.14 ± 239.36	0.998
$avIntDur_{13\text{-}20}$ (ms)	PD	536.30 ± 188.41	541.02 ± 188.16	
	HC	500.36 ± 237.52	499.31 ± 237.47	0.997

[3] These results, as well as methods discussed in Part II - Chapter 5 can also be found in [200,201,202,203,204]

11.1.2 Comparison between HC subjects and PD patients on sentence repetitions

Mean values and standard deviations (SD) of parameters computed with BioVoice, PRAAT and MDVP on sentence repetition signals are reported in Tab. 11.3. As not all participants show breaks within single sentences, the parameter T_{pause} was considered only for D% and NSR calculation. For this reason, this parameter was not reported in Tab. 11.3.

Tab. 11.2: mean values, standard deviations and correlation coefficients for the parameters used to compare PD patients and HC subjects during the syllable repetition task (data from [94])

Parameters	HC subjects	PD patients	t-test
D_{mean}	0.50 ± 0.11	0.50 ± 0.12	p = 0.82
D_{SD}	0.05 ± 0.02	0.06 ± 0.03	p = 0.08
avD_{1-4}	0.53 ± 0.11	0.55 ± 0.13	p = 0.62
avD_{5-12}	0.50 ± 0.11	0.50 ± 0.13	p = 0.82
avD_{13-20}	0.49 ± 0.12	0.48 ± 0.12	p = 0.56
D_{COV}	2.31 ± 1.16	2.82 ± 1.47	p = 0.11
$D_{COV5-20}$	2.33 ± 1.27	2.58 ± 1.21	p = 0.34
$DRelStab_{5-12}$	94.97 ± 7.67	90.53 ± 9.91	p = 0.08
$DRelStab_{13-20}$	92.93 ± 8.60	87.33 ± 10.36	p = 0.02
DPA	2.04 ± 7.70	3.21 ± 6.51	p = 0.29

These results show that significant differences exist between the two groups for parameters T_{inter}, D% and NSR. In particular, there is a marginal significance of the difference in the means as far as T_{inter} is concerned ($t(39) = 2.03$, $p = .049$), with PD patients have higher values than healthy controls. Significant differences exist as far as D% ($t(39) = 2.68$, $p = .011$) and NSR ($t(39) = 2.68$, $p = .011$) are concerned. In particular, D% is larger in the control group while NSR is larger in PD patients (NSR being inversely proportional to D%). No significant differences were found in the parameters related to F0 and noise. However, the ratio between F0CV of the last sentence and F0CV of the first sentence is larger in PD patients ($t(39) = 1.91$, $p = .063$) with respect to controls. Similar result were found with PRAAT ($t(39) = 1.99$, $p = .054$) and MDVP ($t(39) = 1.79$, $p = .080$) (Tab. 11.3). A slight difference, though not significant was also found for ANNE: the healthy control group exhibits larger negative values (that is, less noise) than PD patients. Again similar results were found with PRAAT and MDVP: no significant differences were found in the parameters related to F0 and noise. Moreover, with our data no relevant correlation nor any significant trend or pattern (increasing or decreasing) during the sentence repetition task was found between the following parameters: disease duration and UPDRS motor score.

Tab. 11.3: - Mean value, standard deviation and t-test result for the acoustical parameters extracted from the sentences repetition task

Software tool	Parameters	PD patients		Healthy subjects		t-test
		Mean	SD	Mean	SD	
BioVoice	$T_{sentence}$ (s)	3.05	0.67	3.14	0.56	$t(39) = 0.42$, $p = .67$
	T_{inter} (s)	0.76	0.33	0.57	0.23	$t(39) = 2.03$, $p = .049$
	D%	73.77	7.37	79.77	6.52	$t(39) = 2.68$, $p = .011$
	NSR (syll./s)	6.54	0.97	5.79	0.77	$t(39) = 2.68$, $p = .011$
	F0CV	0.13	0.04	0.12	0.04	$t(39) = 0.83$, $p = .41$
	F0CV last/first	1.06	0.34	0.89	0.19	$t(39) = 1.91$, $p = .063$
	F0NR	0.56	1.14	0.51	0.15	$t(39) = 1.13$, $p = .27$
	F0NR last/first	1.15	0.65	0.98	0.37	$t(39) = 0.97$, $p = .34$
	ANNE (dB)	-20.00	3.78	-21.73	3.89	$t(39) = 1.39$, $p = .17$
Praat	F0CV	0.14	0.04	0.14	0.05	$t(39) = 0.29$, $p = .78$
	F0CV last/first	1.14	0.56	0.87	0.17	$t(39) = 1.99$, $p = .054$
	F0NR	0.65	0.14	0.69	0.17	$t(39) = 0.77$, $p = .44$
	F0NR last/first	1.05	0.41	0.96	0.32	$t(39) = 0.74$, $p = .47$
	HNR (dB)	12.74	4.41	15.28	3.97	$t(39) = 1.86$, $p = .071$
MDVP	F0CV	0.15	0.04	0.16	0.08	$t(39) = 0.38$, $p = .71$
	F0CV last/first	1.11	0.44	0.88	0.33	$t(39) = 1.79$, $p = .080$
	F0NR	0.85	0.34	0.78	0.19	$t(39) = 0.79$, $p = .43$
	F0NR last/first	1.06	0.31	1.04	0.36	$t(39) = 0.16$, $p = .87$
	NHR	0.23	0.10	0.20	0.05	$t(39) = 1.31$, $p = .20$

11.2 Discussion

11.2.1 Performance of the AVU algorithm on syllable repetitions

Comparing the AVU results with the manual labeling, the automatic segmentation provides results very close to the reference. These results are very promising in order to implement an automatic and computationally fast method to assess dysprosody in PD patients, since the processing time is on average below 2 s for signals with duration of 20 s. However, for the SD it was not possible to highlight the same significant differences of the manual results (Tab. 11.1), although a higher variability in PD patients is still visible. For these reasons, the comparison on a larger dataset is needed.

Concerning the comparison between PD patients and HC subjects, since $DRelStab_{13\text{-}20}$ is defined as the ratio between $avD_{13\text{-}20}$ and $avD_{1\text{-}4}$, this result indicates a decreasing of D during the repetitions. In fact, since D is the percentage of voiced time related to the whole utterance duration, and that a decreasing of Interval Duration along the repetitions is present in PD patients but not significant (Tab. 11.1), it could mean that there is a reduction of the vowel duration during the task. This is true for both groups, but more enhanced in PD patients.

11.2.2 Comparison between HC subjects and PD patients on sentence repetitions

The results of this study show that PD patients exhibit an alteration of prosodic patterns of speech during a sentence repetition task. With respect to control subjects, PD patients have longer pauses between each sentence repetition (T_{inter}) and a lower percentage of "voiced time" during the entire repetition period (duty cycle D%). On the other hand, a decrease of duty cycle leads to an increase of NSR. Therefore PD patients tend to have a shorter time period occupied by speech than healthy controls, but at the expense of a longer recovery time, since no significant difference was found in $T_{sentence}$ (the time from the start of a sentence to the start of the next one).

At a first glance our results on T_{inter} (time interval between two adjacent sentences) are in contrast with other findings [92,93,95], where PD patients (of both genders) showed a less percentage of pauses than healthy controls. However, it should be noted that in these studies the task is not the repetition of a sentence but the reading of a passage composed by four sentences.

The main difference might be due to the fact that we have considered only inter-sentence pauses using intra-sentence pauses for the computation of the other parameters (D% and NSR).

Our results concerning the increase of NSR in PD patients are consistent with other findings [92,95] where an increase of the number of syllables per second was found during the reading of a text. However, other studies that used the reading of a passage [84,93] did not show any significant difference in this parameter. Thus, it is likely that speech rate alterations could be more easily raised from the repetition of a sentence rather than by the reading of a text.

It's worth noting that in our study we choose the sentence repetition task for two main reasons: first, we aimed at studying temporal parameters related to the pace of the repetition; second, some patients may have problems with regarding a passage due to poor eyesight and to age (most PD patients are elderly) and this may lead to an altered emotional state with obvious consequences on the quality and characteristics of the voice. Other studies, as well as results exposed in the previous section, show that alterations in speech rhythm could be pointed out by the widespread syllable repetition task (i.e., the repetition of syllables /pa/ or /pa/-/ta/-/ka/). It was shown that this test could reveal vocal pace variations; in particular, rhythm acceleration was found in PD patients [94].

Temporal parameters (T_{inter}, $T_{sentence}$, D% and NSR) are computed with the AVU module of BioVoice. A comparison with the other two software tools (PRAAT and MDVP) was not carried out because they both provide only global information (relating to the entire recording) about the voiced and the unvoiced parts. More details on the comparison among the three software as far as VU segmentation performance is concerned are reported in [168].

No significant differences were found in the F0-related parameters (F0CV and F0NR). The only observed change between control and PD patients, although not significant, is the ratio between F0CV of the last sentence and F0CV of the first sentence ($t(39) = 1.91$, $p = .063$) that is higher in PD patients than in healthy controls (1.06 vs 0.89). Similar results were obtained with PRAAT ($t(39) =$

1.99, p = .054) and MDVP ($t(39) = 1.79$, p = .080), with higher values for the control groups with respect to PD patients (1.14 vs 0.87 with PRAAT, 1.11 vs 0.88 with MDVP).

This suggests that the potential non-significant trend observed may be worth exploring in a larger sample. Other studies concerning F0 variations showed that PD patients exhibit reduced F0 standard deviation and variation range during a passage reading and a monolog [84,92,93,95]. This reflects the reduced variability of F0 which is one of the most common speech alterations in Parkinson's disease. The different result obtained with our study could be due to the fact that here the sentence repetition was not performed with any particular intonation, unlike e.g. in Rusz et al. [84] where 10 sentences were pronounced according to different emotional context. From the comparison with these findings, it seems that the sentence repetition task is less meaningful with respect to the monolog or the text reading in enhancing reduced variability of F0 in PD.

Concerning noise, no significant differences were found although some variation is detected (Tab. 11.3). As expected healthy subjects exhibit larger negative values of ANNE than PD patients (-21.73 dB vs -20.00 dB) that reflects a slightly higher noise in PD patients. However both groups are in the range of good quality of the voice. This is confirmed by similar results with PRAAT (lower HNR in PD patients) and with MDVP (higher NHR in PD patients). As the three tools apply different methods for noise estimation, the correlation coefficients between ANNE and HNR ($r(37) = - 0.84$, $p < .001$) and between ANNE and NHR ($r(37) = 0.65$, $p < .001$) were computed, showing moderate correlations between them.

In conclusion, our results show that in PD patients with no or minimal speech problems (UPDRS item 18 score between 0 and 1) the speech rate and temporal alterations are the most noticeable features of dysprosody during a sentence repetition task.

The advantage of our method is that the chosen sentence (which consists mainly of vowel sounds) not only allows an automatic segmentation within the whole signal, but also the estimation of other acoustic parameters (fundamental frequency, noise, etc.).

Skodda et al. [197] showed that tVSA (total Vowel Space Area) and VAI (Voice Articulation Index) computed from the first and the second formant frequencies of the vowels /a/, /i/ and /u/, are predictive of PD progression. Since our sentence ("*Il bambino ama le aiuole della mamma*") contains all these vowels in a triphthong (/aiu/ inside /aiuole/), it would be easy to evaluate also these parameters. Other possibilities could be: the analysis of the three corner vowels where each vowel (/a/, /i/ and /u/) is extracted separately from the same sentence and the analysis of "staccato" vowels uttered in isolation. This problem will be addressed in future work.

A disadvantage of our method consists in the empirical threshold to identify inter-sentence pauses. Despite most of subjects showed pauses/breaks inside a sentence shorter than the inter-sentence ones (and lower than the empirical threshold), in two cases (1 PD patient and 1 healthy control the length of some inter-sentence pause was comparable to the length of some intra-sentence pause. In these two cases the selection of the inter-sentence pauses was made by hand.

Thus, further developments will concern the implementation of a speech recognition module in order to detect the words that delimit each sentence. In this way we would be able to consider all the pauses between such words as well as the inter-sentence ones. Moreover, a speech recognition module would allow detecting a set of words within a "less controlled" framework, such as a monolog. Indeed some studies [83,87] showed that the monolog is the most predictive task for speech impairment assessment in PD. Through this module it would thus be possible to automatically perform the acoustic analysis of some features extracted from the monolog, while meeting requirements in order to standardize different and heterogeneous acquisition (i.e., to extract some particular vowel from the same words in monolog, etc.), as suggested in [87].

12. Markerless analysis of articulatory movements during speech (validation on healthy subjects)[4]

In this chapter, results on the performance of the proposed markerless method for studying the articulatory movements (in particular lips movements) are reported (methods are shown in Part II - Chapter 6). In particular, the RMSE of the tracked lip points is calculated exploiting words and sentences of the chosen speech corpora, while syllable repetitions (/pa/ for at least 25 times) were used to compare our markerless method with the marker-based reference in the estimation of some kinematic parameters (peak velocities and accelerations of the lower lip during the opening and closing phases).

12.1 Results

12.1.1 Error assessment

The acquisition relative to the French subject was carried out at a mean distance of the mouth from the Primesense sensor of (737.99 ± 41.88) mm, while those relative to the Italian subject were performed at a mean distance of (770.20 ± 14.14) mm. The mean values and the standard deviations of the RMSE for the 3D trajectories of the 7 points of interest as well as the RMSE for the 3 articulatory parameters were reported in Tab. 12.1. These results were computed on the whole corpus for a total of 3413 phonemes (1743 coming from the French corpus and 1670 from the Italian corpus). Since the points extracted with the markerless method were mapped to the Vicon reference frame, the notation used in this table refers to Vicon coordinates system: x is the lateral axis, y is the frontal axis and z is the vertical axis.

Considering the mean distances at which the experiments were performed, the mean errors introduced by the Primesense sensor, for the French corpus were 0.99 mm on the lateral axis, 0.83 mm on the vertical axis and 1.13 mm on the frontal axis (Fig. 12.1). For the Italian corpus, the mean errors were 1.17 mm on the lateral axis, 0.81 mm on the vertical axis and 1.21 mm on the frontal axis (Fig. 12.1).

[4] Part of these results, as well as methods discussed in Part II - Chapter 6 can also be found in [205,206]

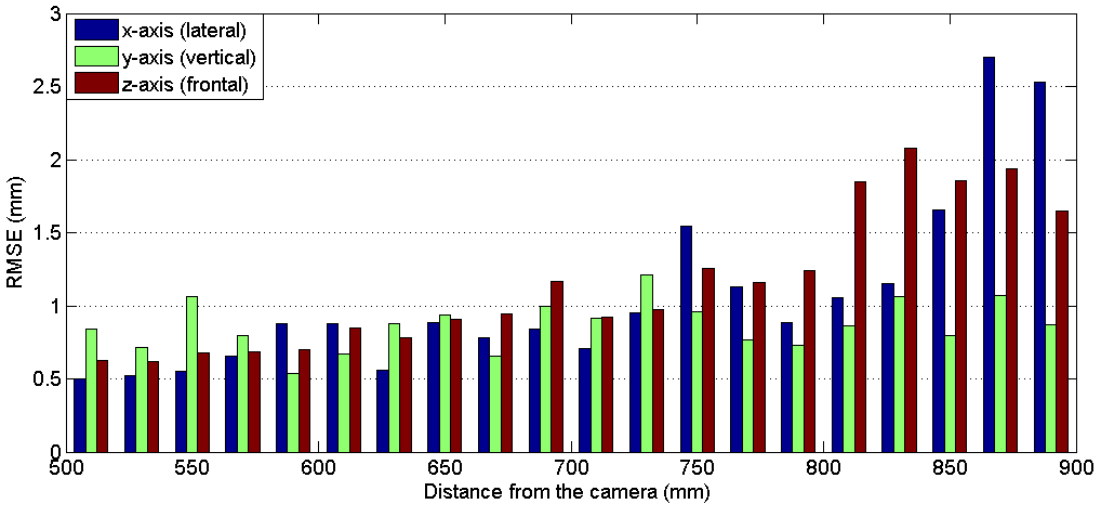

Fig. 12.1: mean errors introduced by the Primesense sensor with respect to the frontal distance from the camera

As shown in tab. 12.1, the mean values of RMSE relative to the French corpus ranged between 1 and 3 mm (except for the coordinate y of the points L5 and L7 and the coordinate x of the point L7). Even in the Italian corpus we noticed mean values less than 3 mm, although errors higher than 3 and 4 mm are more frequent. In particular, for the coordinate y of the points L2, L3, L5 and L7 the RMSEs ranged between 4 and 5 mm. Considering the depth accuracy of the device, the error for the depth values (z-axis) was approximately of 0.7 mm for z < 600 mm, ranged between 0.8 and 1.2 mm for z between 600 mm and 800 mm and is about 2 mm for z > 800 mm. Similar values were present on the x-axis, although the increase of the RMSE values over 2 mm occurs after 850-870 mm of distance. On the y-axis (vertical), the error seems to be constant along the entire range covered by the object. In fact, we found that the RMSE ranged between 0.7 mm and 1.2 mm in the considered distance range.

Tab: 12.1. RMSE (mean values and standard deviations) for the 7 points of interest and for the 3 articulatory parameters

Points and	French RMSE (mm)			Italian RMSE (mm)		
Parameters	x	y	z	x	y	z
L1	1.63 ± 0.99	2.15 ± 0.96	2.57 ± 1.53	2.91 ± 2.30	2.35 ± 1.54	1.61 ± 1.01
L2	1.39 ± 0.82	2.98 ± 1.44	1.40 ± 0.96	2.79 ± 1.85	4.17 ± 2.17	1.50 ± 0.99
L3	1.28 ± 0.95	2.99 ± 1.49	1.50 ± 1.04	5.62 ± 3.06	4.14 ± 1.68	1.57 ± 1.39
L4	2.10 ± 1.14	1.53 ± 0.85	2.23 ± 1.51	2.54 ± 2.08	1.85 ± 1.27	1.92 ± 1.21
L5	1.30 ± 0.80	3.43 ± 1.40	2.46 ± 1.52	2.93 ± 2.30	5.50 ± 2.81	2.36 ± 1.27
L6	1.96 ± 1.13	2.47 ± 1.39	2.49 ± 1.61	2.85 ± 2.49	2.92 ± 1.85	2.47 ± 1.31
L7	3.15 ± 1.54	6.50 ± 2.08	2.44 ± 1.59	3.29 ± 2.73	5.08 ± 2.62	2.63 ± 1.51
Width		2.07 ± 1.22			1.86 ± 1.07	
Opening		2.72 ± 1.66			3.81 ± 2.39	
Protrusion		1.36 ± 0.81			4.45 ± 2.08	

12.1.2 Kinematic analysis on syllable repetitions

The analysis on the syllable repetition task was conducted on a total of 80 utterances. The kinematic parameters (mean values and standard deviations) for speed and acceleration during opening and closing phases relative to the same point are reported in Tab. 12.2.

The correlation coefficient for the trajectory on the vertical axis of the point L6 was (0.96 ± 0.03), while those for speed and acceleration were respectively (0.95 ± 0.05) and (0.88 ± 0.10).

Tab. 12.2: Mean values and standard deviations of the kinematic parameters during the opening and closing phases of the syllable repetition

Kinematic parameters	Marker-based	Markerless
V_{open} (mm/s)	-114.37 ± 30.55	-96.39 ± 23.48
V_{close} (mm/s)	100.01 ± 45.82	79.77 ± 26.97
A_{open} (mm/s^2)	-1689.63 ± 559.86	-1759.10 ± 665.73
A_{close} (mm/s^2)	2619.02 ± 1068.82	2141.91 ± 910.86

12.2 Discussion

For most of the points, the RMSE mean values ranged between 1 and 3 mm (Tab. 12.1). Considering the low image resolution used for the experiment this is a very promising result. Further information about acceptable error ranges could be provided by future works that could be devoted to the calculation of error measures with other algorithms, devices and configurations.

Concerning the articulatory parameters (Tab. 12.1), we obtained good results for the width in both corpora (mean RMSE around 2 mm) and for opening and protrusion in the French corpus (mean values: 2.72 mm and 1.36 mm, respectively). Instead, the errors for opening and protrusion in the Italian corpus were higher, with values over 4 mm for protrusion. Since protrusion is computed from the y-coordinate of points L2, L3 and L6, bigger errors for these points could lead to a bigger protrusion error; in fact, the error on the y-coordinate for these 3 points is always bigger in the Italian corpus with respect to the French one (Tab. 12.1).

These differences could be due to several factors. First of all, the markers were accurately positioned to match the *Intraface* points, but this positioning presents an intrinsic error due to the manual settings. For the opening, the higher RMSE could be due to the higher distance of the acquisition and different orientation of the Primesense with respect to the face that, along with the low resolution of the images, could lead to bigger errors.

The kinematic parameters reported in Tab. 12.2 show a tendency to underestimate the module of the maximum and the minimum speed values (closing and opening phases) with differences around 20 mm/s. An underestimation is visible also for the closing acceleration, while during the opening phase the two estimates seem to be closer.

Although the results on kinematic parameters seem to be inconsistent, from the plot in Fig. 6.6 of Part II - Section 6.2.2, and from the correlation values between the two systems, it is possible to observe

that the trajectories, the velocities and the accelerations extracted with the markerless technique were very similar when compared with the reference. This suggests that a bias is present in the estimation of the kinematic parameters.

This bias might be due to the distance from the face at which the device was located (about 0.8 m), or to the different framerate of the systems (30 Hz for the depth sensor, 100 Hz for the marker-based method). This distance was a trade-off between the need to move the sensor as close as possible to the subject's face and its characteristic (range of work: 0.4-1.5 m), without interfering with the field of view of the Vicon cameras. The distance, in conjunction with the low image resolution (320 x 240 pixels) probably explains these differences. However, further experiments with structured light sensors should consider an experimental design with higher frame resolutions (at least 640 x 480 pixels) and smaller distances from the subject's face (i.e., 0.5-0.6 m, according to the specification provided by the manufacturer).

Considering the depth accuracy of the device we found that the errors on the x (lateral) and z (frontal) axes exhibited similar behaviors, with higher values with increasing distance from the camera. Considering the distances at which the speech acquisitions were performed (0.7-0.8 m) the errors in the depth estimation reflected what has already been observed in the results of table 12.1, namely for most of points the biggest error is on the frontal axis (that in the Vicon reference frame is the y-axis). However, this error is always smaller than those reported in Tab. 12.1 (only in the case of the point L4 might be similar, since it is between 1.5 mm and 1.8 mm). Even for the other two coordinates the device errors were lower than those computed during speech experiments. We believe that this is due to the low resolution of the video frames (320 x 240 pixels) and the distance of the Primesense to the subject. This means that the pixels of the color images (those on which the face tracker works) correspond to an area of the face larger than that which would be using a higher resolution and/or decreasing the distance between the face and the Primesense camera. Thus, some depth variations of the face (in particular the cavities in the corner of the mouth due to the lip anatomy) might be indistinguishable.

For these reasons for further experiments that will involve structured light cameras (Microsoft Kinect, Primesense sensors, Asus Xtion, etc.) we strongly recommend to acquire images of at least 640 x 480 pixels of resolution for both streams.

Although the working range used for this experiment led to reasonable errors in the estimation of the 3D coordinates another requirement to adopt for future experiments is to perform the acquisitions at distances lower than 0.7 m. In fact, according to our results and considering the range of working specified by the manufacturer (0.4-1.5 m), the resolution on the 3 axes should be better at distances between 0.4 and 0.6 m from the camera. This recommendation allows increasing indirectly the resolution. However, due to the technical design of this experiment, we could not bring the camera closer to the subject's face.

13. Markerless Analysis of articulatory movements during speech (application to PD patients)[5]

13.1 Results

The parameters obtained with the kinematic analysis ($v_{opening}$, $a_{opening}$, $v_{closing}$, $a_{closing}$, $\Delta Opening_{norm}$, $MaxOpening_{norm}$, already defined in Part II - Chapter 7) are reported in Table 13.1 for both 2D and 3D analysis. Specifically: $v_{opening}$ and $a_{opening}$ refer to the maximum speed and acceleration of point CLL during the opening phase, $v_{closing}$ and $a_{closing}$ refer to the maximum speed and acceleration of point CLL during the closing phase, $\Delta Opening_{norm}$ and $MaxOpening_{norm}$ are the normalized range of opening and the normalized maximum opening value, respectively. For simplicity, values related to the closing phase ($v_{closing}$ and $a_{closing}$, Table 13.1) are reported in absolute value since they should be negative. Indeed in this experiment the values on the vertical axis increase downward. This is due to the fact that the 3D coordinates are computed starting from the 2D image coordinates where the plane origin is located in the upper left corner and the y axis is positive downward. The closing movement direction is opposite to that axis leading to negative values of speed and acceleration.

Concerning the 3D analysis, the results show significant differences in $v_{opening}$ ($t(28) = 2.49$, $p = .019$), $v_{closing}$ ($t(28) = 2.32$, $p = .028$) and $a_{opening}$ ($t(28) = 2.13$, $p = .043$). All these values are lower in PD patients. In particular, the opening and closing velocities are reduced by more than 25 mm/s (94.94 ± 33.40 mm/s vs 64.45 ± 30.94 mm/s for $v_{opening}$, 87.85 ± 31.28 mm/s vs 61.54 ± 28.49 mm/s for $v_{closing}$). Lower values were found also for $a_{opening}$, although not significant. Concerning the opening parameters ($\Delta Opening_{norm}$, $MaxOpening_{norm}$) the normalized range of opening is lower in PD patients (0.65 ± 0.36 vs 0.46 ± 0.23) although not significant, while the normalized maximum value is comparable with values around 0.4 in both groups (i.e., the maximum opening is about the 40% of the mouth width). These values are normalized to take into account the anatomical variations among subjects that result in different values of opening and width of the mouth, as explained in Part II - Chapter 7.

As far as the 2D kinematic analysis is concerned, similar considerations can be drawn. Velocities and accelerations are reduced in PD patients, although no significant differences were found.

[5] These results, as well as methods discussed in Part II - Chapter 7 can also be found in [207]

Figure 13.1 shows the trend of the six 3D parameters described above during the whole repetition task. The plots show a decrease of all parameters for both groups, more pronounced for velocity in PD patients. To provide a quantitative evaluation of this trend a linear regression was applied. The slope of the regression line is reported in Table 13.2 for each parameter and both PD and HC subjects. Indeed Table 13.2 shows that the decrease of the velocities in HC subject is slower than in PD patients (-0.41 vs -0.72 for $v_{opening}$ and -0.38 vs -0.77 for $v_{closing}$) while for the other parameters the trend is comparable.

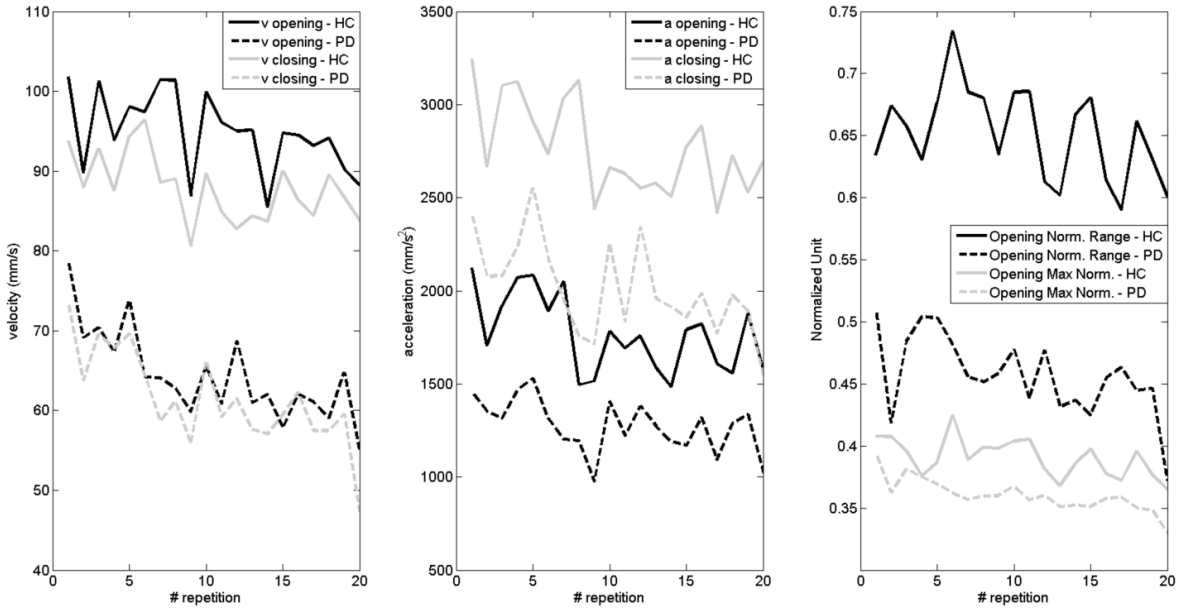

Fig. 13.1: Kinematic parameters along the whole repetition task. Solid lines: HC subjects; Dashed lines: PD patients. Left plot: trend of peak velocities. Middle plot: trend of peak acceleration values. Right plot: trend of normalized opening values. All the parameters show a decrease in both groups, however it seems to be more pronounced in PD patients especially for velocities (left plot and Tab. 13.2).

Tab. 13.2: Slope of the regression line of the kinematic parameters along the entire repetition task.

	PD	HC
$v_{opening}$	-0.72	-0.41
$v_{closing}$	-0.77	-0.38
$a_{opening}$	-12	-19
$a_{closing}$	-26	-25
$\Delta Opening_{norm}$	-0.0034	-0.0027
$MaxOpening_{norm}$	-0.0018	-0.0015

13.2 Discussion

Our results confirm the findings reported by Walsh et al., 2012 [88] where PD patients exhibited reduced peak velocities of lower lip, both for opening and closing phases. In addition, the peak values of acceleration are reduced in PD patients with significant differences only for the opening phase. Thus, our work supports most of the literature on the kinematic analysis of the articulators in PD

patients with hypokinetic dysarthria which states that most of these patients exhibit a downscaling of the articulatory movements [88,104,111,112,113,114].

The novel contribution of this work concerns the assessment of this downscaling by means of a fully markerless and low-cost method. In Part II - Chapters 6 and 7, we described in detail the video-processing framework: camera calibration, face tracking, 3D coordinates estimation and calculation of the kinematic parameters. All of these steps can be easily automated, providing an easy-to-use method for speech therapy and disease progression monitoring. By exploiting an existing face tracking algorithm and computer vision methods, the application of contact sensors or markers to the subject's face is no longer required and no other manual setting is needed with obvious advantages for the patient.

The accuracy of this technique was already demonstrated in the previous chapter although those experiments rely only on healthy subjects. Thus, a further development should be focused on testing this markerless method against a marker-based reference also for dysarthric patients.

Because of the limited number of PD patients, we did not make any distinction among different levels of severity of dysarthria. Thus, future studies will be devoted to increasing the dataset, recruiting PD patients at different stages of the disease and with different levels of speech impairment, in order to investigate the evolution of the kinematic parameters. We expect a reduction of the kinematic parameters proportional to the severity of dysarthria, although further investigations are strictly necessary. As far as the normalized range of opening ($\Delta Opening_{norm}$) is concerned, we found lower values in PD patients with respect to HC subjects although this difference is not significant. This result supports previous findings [88] where the displacement of the lower lip during the utterance of bilabial sounds was lower in PD patients with respect to HC subjects. In our case PD patients have lower values of $\Delta Opening_{norm}$. This could be due to the fact that the maximum/minimum interval of lower lip displacement is reduced in PD patients, since this value is defined as the difference between the maximum and the minimum opening values divided by its mean value. Differently from [88] here we normalized the opening values. This allows taking into account the anatomical variations of different subjects. Conversely, the results of $MaxOpening_{norm}$ are similar between the two groups. As the maximum/minimum interval is reduced in PD patients, but the maximum value of opening is similar between groups, the decrease of the normalized range of opening in PD patients might be due to an increase of the minimum value of opening. This means that during the pronunciation of the bilabial plosive /p/ the occlusion phase of the lips before the "burst" (releasing phase of the airflow) is less pronounced in PD patients. This might be due to a weakness in tightening the lips.

Concerning the 2D analysis, we found a decrease in the kinematic parameters of PD patients with respect to HC subjects as well as for the 3D analysis. However, the comparison among the 2D parameters did not show any significant difference between groups. For this reason, we recommend the use of both video streams (colour and depth) to evaluate the articulatory kinematic to get highly accurate results. New experiments should be performed on larger patients groups to better assess if

these differences could be detected with a single colour video stream. Thanks to the proposed methodology, the simple 2D parameters proposed here could be easily implemented in a user friendly smartphone app for rehabilitation purposes. This would increase the number of PD patients undergoing monitoring and provide further clinical data to the clinicians. As the only significant differences between the groups were found for 3D parameters, we evaluate the trends of these measures during the task. Figure 13.1 shows a decrease of all the parameters along the syllable repetition, although from Table 13.2 a clear decrease of the regression line slope is present for $v_{opening}$ and $v_{closing}$ only. This could be due to speech-related fatigue that mainly occurs in PD patients. However, a recent study [198] shows that fatigue manifestations would not be so noticeable in PD patients. In our case the performed speech task is too short to draw conclusions about speech-related fatigue, while in [199] a speech task of 1 hour of duration was used. Nevertheless we could apply our method to assess those findings from a kinematic point of view.

In contrast to Walsh et al., 2012 [88] we use a syllable repetition task, a widely used speech task already implemented in many studies on the kinematic analysis of the articulators in PD patients [111,113,114]. However, the analysed bilabial movements are similar to those proposed there (/pa/ in our study, /paIp/ in [88]). Further developments will be devoted to the implementation of this technique to other speech tasks in particular those that involve spontaneous speech (passage reading or monologue).

Concerning the methodology some considerations should be drawn. Two main differences exist between the current work and the previous chapter where we tested the accuracy of the markerless system:

1. During the accuracy tests we used the depth sensor Primesense Carmine 1.09 while in the current study we carried out the experiments by means of the Microsoft Kinect for Windows. However, both devices have very similar hardware and also the Kinect sensor used in our experiments can work in a near range mode [199].
2. To assess the accuracy of the proposed method we used an image resolution of 320x240 pixels (Part II - Chapter 6). Though this is a quite low resolution for our purposes, promising results were found in terms of tracking errors of the lips during speech production. We concluded that higher resolutions should be adopted for the experiments that involve the study of the articulatory movements with 3D depth sensors. A recommendation was to use at least 640x480 pixels for both streams (colour and depth) according to the device features.

As already demonstrated in the previous chapter the error introduced in the depth estimation by the sensor is fairly constant (around 1 mm) between 0.5 and 0.7 m far away from the camera. Considering these values and the working range provided by the manufacturer (greater than 0.4m) we performed these experiments within these distances. Moreover, subjects were seated during the experiments, therefore these boundaries were reasonably kept even in the case of PD patients with involuntary movements.

In the previous chapter we demonstrated that with low image resolutions our method was able to accurately track the trajectories and the trends of velocity and acceleration of the lower lip. Despite these good results, with correlation coefficients over 0.95 for trajectories and velocities when compared to the marker-based reference, we noted that the markerless method underestimates the peak velocities of about 20 mm/s. In the present work we cannot state if the estimation of the peak velocities and accelerations were underestimated as we have not yet compared our markerless method against a marker-based technology with higher image resolutions. Thus, further experiments will concern also this topic.

With this simple markerless system we are able to track lips movements with good accuracy and detect significant differences in the kinematic parameters of the lower lip between PD patients and HC subjects. However, the main limitation of this system is its capability of tracking only the external articulators being based on colour and infrared cameras. Thus, as speech involves a complex coordination of several articulators, complete and exhaustive results could be obtained only combining markerless and marker-based techniques (for instance, EMA for studying tongue movements).

14. Analysis of facial expressions and movements in PD patients

14.1 Results - Analysis of expressive features with respect to the neutral baseline

For each video frame of the expressed videos the Euclidean distance from the current face model and the neutral template for each subject was computed. As already reported in Part II - Chapter 8, for each video, the following measures related to the distance were computed: mean value, standard deviation, skewness, kurtosis, maximum value, minimum value and range. On average, HC subjects reported higher distances than PD patients along the whole tasks (12.68 ± 5.05 for HC subjects vs 9.35 ± 3.85 for PD patients, $p < .00001$). Then, for each one of the 4 expressions, the following comparison on the aforementioned measures was performed: PD patients during the acted expressions vs PD patients during the imitated expressions (PD_act vs PD_im), HC subjects during the acted expressions vs HC subjects during the imitated expressions (HC_act vs HC_im), PD patients vs HC subjects during the acted expressions (PD_act vs HC_act) and PD patients vs HC subjects during the imitated expressions (PD_im vs HC_im). Results are reported in Tables 14.1-14.4. Within group differences (PD_act vs PD_im and HC_act vs HC_im) are reported in bold, while between group differences (PD_act vs HC_act and PD_im vs HC_im) are highlighted in grey.

Tab. 14.1: Distance measures during the anger expression for PD patients and HC subjects. Results are reported for both tasks (acted and imitated expression)

	Distance measures	HC subjects	PD patients
Acted	Mean	**10.13 ± 3.44**	8.05 ± 3.38
	SD	1.93 ± 0.96	2.11 ± 1.43
	Skewness	-0.14 ± 1.63	0.64 ± 0.95
	Kurtosis	6.74 ± 8.05	4.42 ± 2.65
	Max	16.11 ± 6.20	14.27 ± 5.75
	Min	**4.94 ± 1.98**	4.03 ± 1.51
	Range	11.17 ± 6.08	10.24 ± 4.99
Imitated	Mean	**12.50 ± 4.95**	8.44 ± 2.80
	SD	1.65 ± 0.59	1.73 ± 1.05
	Skewness	0.06 ± 0.97	0.59 ± 1.35
	Kurtosis	4.40 ± 2.18	5.82 ± 5.87
	Max	17.40 ± 5.96	13.47 ± 4.08
	Min	**8.12 ± 4.71**	4.96 ± 2.20
	Range	9.28 ± 3.38	8.51 ± 4.38

Tab. 14.2: Distance measures during the disgust expression for PD patients and HC subjects. Results are reported for both tasks (acted and imitated expression)

	Distance measures	HC subjects	PD patients
Acted	Mean	**12.77 ± 4.98**	9.38 ± 3.25
	SD	2.50 ± 1.15	2.31 ± 1.10
	Skewness	-0.29 ± 1.23	0.36 ± 0.93
	Kurtosis	4.57 ± 3.27	3.53 ± 1.55
	Max	18.73 ± 6.84	15.62 ± 5.28
	Min	**6.78 ± 4.74**	4.96 ± 1.73
	Range	11.95 ± 5.61	10.65 ± 4.60
Imitated	Mean	**14.88 ± 5.22**	10.38 ± 4.79
	SD	2.11 ± 1.10	2.69 ± 3.19
	Skewness	-0.10 ± 0.93	0.08 ± 0.86
	Kurtosis	4.39 ± 3.63	3.33 ± 1.58
	Max	20.24 ± 6.12	18.89 ± 8.17
	Min	**9.75 ± 5.28**	5.34 ± 2.31
	Range	10.49 ± 4.58	10.54 ± 8.25

Tab. 14.3: Distance measures during the happiness expression for PD patients and HC subjects. Results are reported for both tasks (acted and imitated expression)

	Distance measures	HC subjects	PD patients
Acted	Mean	13.86 ± 5.86	10.94 ± 4.84
	SD	3.68 ± 2.34	2.81 ± 0.93
	Skewness	0.34 ± 1.26	0.32 ± 0.86
	Kurtosis	5.22 ± 6.63	3.18 ± 1.05
	Max	24.75 ± 9.95	17.80 ± 5.84
	Min	**6.17 ± 3.10**	5.45 ± 3.44
	Range	**18.57 ± 9.43**	12.35 ± 3.69
Imitated	Mean	14.17 ± 6.62	10.21 ±4.52
	SD	2.49 ± 1.29	2.42 ± 1.11
	Skewness	0.16 ± 1.25	0.31 ± 0.99
	Kurtosis	5.16 ± 3.71	3.68 ± 2.84
	Max	21.22 ± 6.85	16.01 ± 5.57
	Min	**8.54 ± 5.14**	5.26 ± 2.56
	Range	**12.69 ± 5.23**	10.75 ± 4.79

Tab. 14.4: Distance measures during the sadness expression for PD patients and HC subjects. Results are reported for both tasks (acted and imitated expression)

	Distance measures	HC subjects	PD patients
Acted	Mean	11.14 ± 3.47	8.76 ± 2.48
	SD	2.37 ±1.39	1.91 ± 0.97
	Skewness	0.04 ±1.38	0.41 ± 0.84
	Kurtosis	5.48 ± 4.13	3.77 ± 2.49
	Max	17.55 ± 6.28	13.89 ± 4.02
	Min	**5.54 ± 2.12**	4.92 ± 1.68
	Range	12.01 ±5.99	8.97 ± 3.87
Imitated	Mean	11.97 ± 4.13	8.63 ± 2.94
	SD	1.71 ± 0.65	1.98 ± 0.93
	Skewness	0.35 ± 1.24	0.46 ± 0.84
	Kurtosis	5.26 ± 4.31	3.39 ± 1.82
	Max	16.91 ± 4.52	13.73 ± 3.99
	Min	**7.45 ± 3.11**	4.99 ± 2.15
	Range	9.46 ± 3.36	8.74 ± 3.31

14.1.1 Comparison between groups (PD_act vs HC_act and PD_im vs HC_im)

Significant differences were found in the imitation of the anger expression. PD patients showed lower values of mean (8.44 ± 2.80 vs 12.50 ± 4.95, p = 0.007), maximum (13.47 ± 4.08 vs 17.40 ± 5.96, p = 0.033) and minimum distance (4.96 ± 2.20 vs 8.12 ± 4.71, p = 0.020), as reported in Tab. 14.4.

During the disgust, significant differences were found in both tasks (acted and imitated expressions) for mean distance, with lower values in PD patients (Tab 14.2). During the acted expression mean distance was 9.38 ± 3.25 for PD patients and 12.77 ± 4.98 for HC subjects (p = 0.027), while during the imitated expression mean distance was 10.38 ± 4.79 for PD patients and 14.88 ± 5.22 for HC subjects (p = 0.014).

Concerning the happiness, significant differences were found in both tasks (Tab. 14.3). During the acted expression PD patients showed reduced maximum (17.80 ± 5.84 for PD patients, 24.75 ± 9.95 for HC subjects, p = 0.020) and range values of distance (12.35 ± 3.69 for PD patients, 18.57 ± 9.43 for HC subjects, p = 0.019). During the imitation task, PD patients showed reduced maximum (16.01 ± 5.57 for PD patients, 21.22 ± 6.85 for HC subjects, p = 0.021) and minimum values of distance (5.26 ± 2.56 for PD patients, 8.54 ± 5.14 for HC subjects, p = 0.028).

Significant differences were also found during the display of sadness (Tab 14.4). Concerning the acted expression, PD patients showed lower mean values of distance (8.76 ± 2.48 for PD patients, 11.14 ± 3.47 for HC subjects, p = 0.031). During the imitation task, PD patients showed lower mean (8.63 ± 2.94 for PD patients, 11.97 ± 4.13 for HC subjects, p = 0.011), maximum (13.73 ± 3.99 for PD patients, 16.91 ± 4.52 for HC subjects, p = 0.037) and minimum values of distance (4.99 ± 2.15 for PD patients, 7.45 ± 3.11 for HC subjects, p = 0.012).

14.1.2 Comparison within groups (PD_act vs PD_im and HC_act vs HC_im)

During the anger expression HC subjects showed an increase of the following parameters from the acted expression to the imitation task (Tab. 14.1): mean (10.13 ± 3.44 during the acted anger, 12.50 ± 4.95 during the imitated anger, p = 0.049) and minimum values of distance (4.94 ± 1.98 during the acted anger, 8.12 ± 4.71 during the imitated anger, p = 0.011). The same differences were also found during the disgust (Tab. 14.2): HC subjects showed an increase (from the acted task to the imitated task) of mean (12.77 ± 4.98 during the acted disgust, 14.88 ± 5.22 during the imitated disgust, p = 0.0056) and minimum values of distance (6.78 ± 4.74 during the acted disgust, 9.75 ± 5.28 during the imitated disgust, p = 0.0015).

Concerning happiness, HC subjects showed an increase from the acted expression to the imitated expression in minimum value of distance (6.17 ± 3.10 during the acted happiness, 8.54 ± 5.14 during the imitated happiness, p= 0.039) and a decrease of the range value of distance (18.57 ± 9.43 during the acted happiness, 12.69 ± 5.23, p = 0.028), as reported in Tab. 14.3.

During sadness (Tab. 14.4), only a significant increase for HC subjects in the minimum value of distance from the acted to the imitated expression was found (5.54 ± 2.12 during the acted sadness, 7.45 ± 3.11 during the imitated sadness, p = 0.005).

No significant differences were found in the PD group between the acted expressions and the imitated expressions.

14.2 Results - Automated facial expression recognition

For each expressive video the SVM-scores were extracted. The multiclass SVM provides 5 scores (1 for each class), thus for each video frame the maximum score was computed, retrieving the predicted class among: neutral, anger, disgust, happiness and sadness. This procedure provided the total number of frames assigned to a particular class for an expressive video. In order to test whether predictions coincide with the target expressions of a particular video, the total number of occurrences of each class was computed for both groups (PD patients and HC subjects) and for the acquisition tasks (acted and posed expressions). Afterwards, these values were normalized with respect to the overall number of occurrences (i.e., classified video frames), in order to get a percentage of a particular expression.

For each one of the 4 target expressions (both acted and imitated) the pie plots of these percentages are shown in Figures 14.1-14.4. For each target expression they provide a visual representation of the average expressions recognized by the classifier in a particular task.

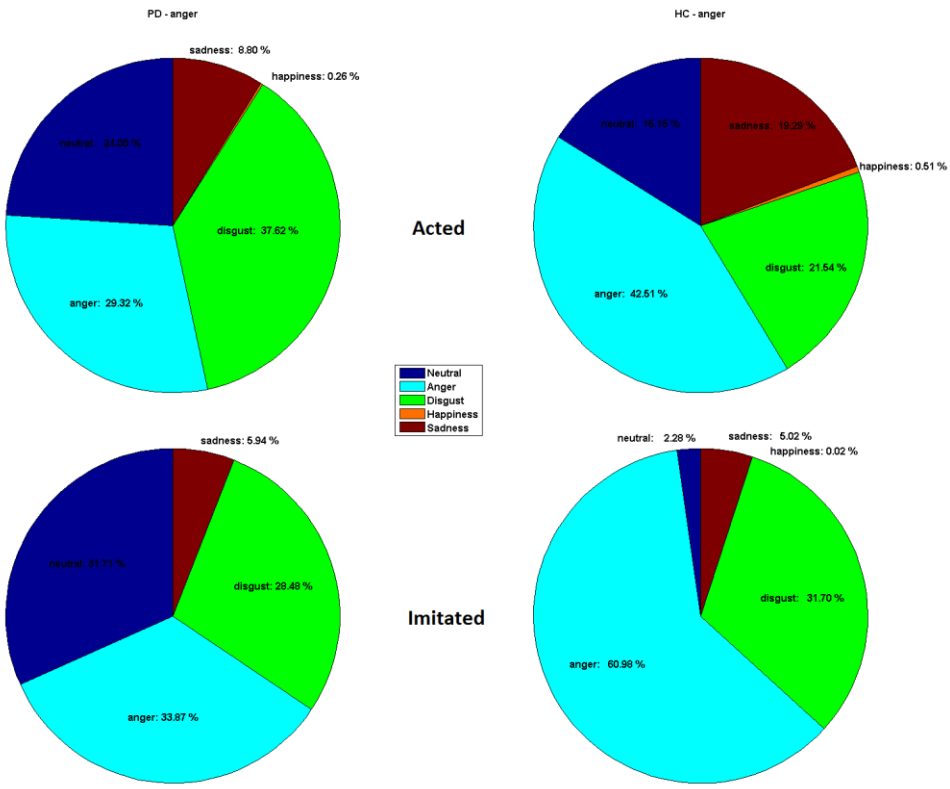

Fig. 14.1: Pie charts for anger: acted expressions (upper plots) and imitated expressions (lower plots). The two plots on the left concern PD patients, while those on the right concern HC subjects.

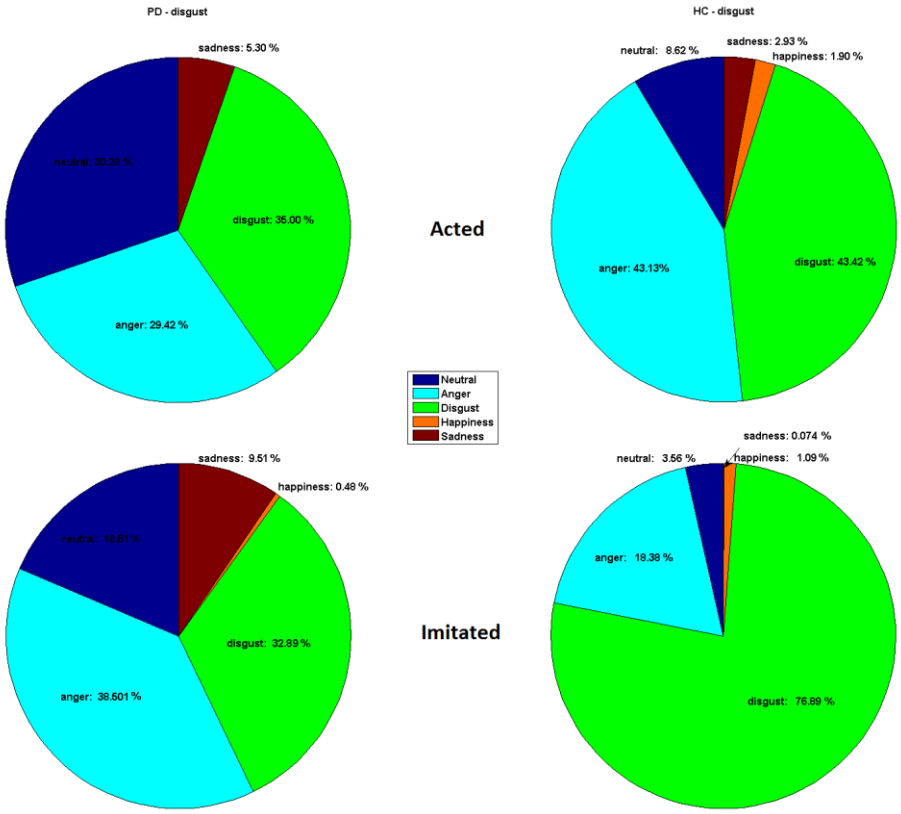

Fig. 14.2: Pie charts for disgust: acted expressions (upper plots) and imitated expressions (lower plots). The two plots on the left concern PD patients, while those on the right concern HC subjects.

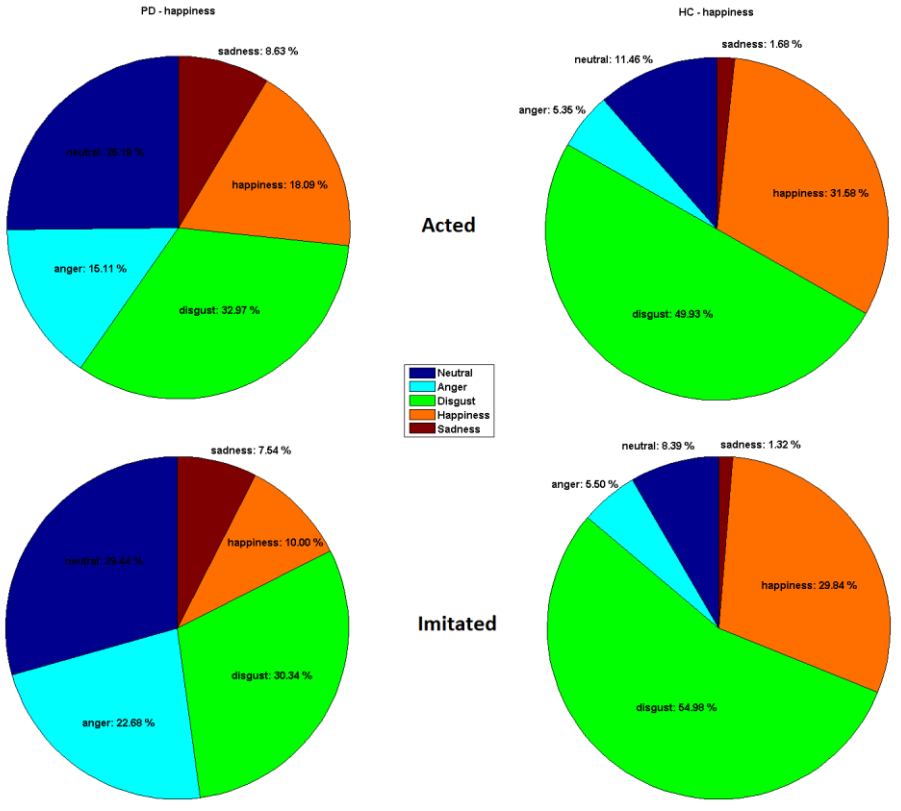

Fig. 14.3: Pie charts for happiness: acted expression (upper plots) and imitated expressions (lower plots). The two plots on the left concern PD patients, while those on the right concern HC subjects.

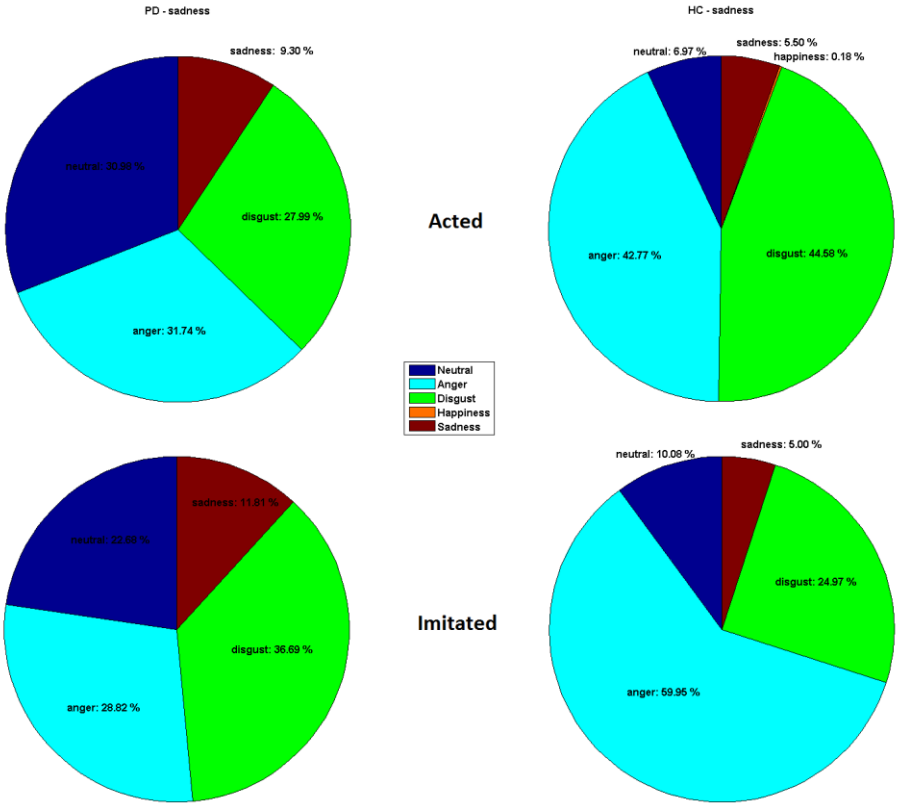

Fig. 14.4: Pie charts for sadness: acted expression (upper plots) and imitated expressions (lower plots). The two plots on the left concern PD patients, while those on the right concern HC subjects.

Moreover, the SVM output scores were considered as an index of the similarity of a particular expression to the target one, since we supposed that the learned expressions were the standard ones (as we trained the classifier on databases of expressions of healthy subjects). This provides further information about the capability of one group in exhibiting an expression better than the other group (i.e. closer to the standard expression). Results are reported in Table 14.5.

Tab. 14.5: SVM average scores during the anger expressions for PD patients and HC subjects. The target expression of this task is reported in red, while the highest average scores are reported in bold for both tasks (acted and imitated)

		Acted	Imitated
HC subjects	Neutral	-0.63 ± 0.17	-0.75 ± 0.12
	Anger	**-0.49 ± 0.41**	**-0.37 ± 0.38**
	Disgust	-0.69 ±0.16	-0.64 ± 0.18
	Happiness	-0.75 ± 0.11	-0.76 ± 0.12
	Sadness	-0.69 ± 0.16	-0.74 ± 0.15
PD patients	Neutral	-0.57 ± 0.26	-0.57 ± 0.34
	Anger	**-0.48 ± 0.44**	**-0.42 ± 0.59**
	Disgust	-0.73 ± 0.18	-0.74 ± 0.20
	Happiness	-0.76 ± 0.12	-0.79 ± 0.13
	Sadness	-0.71 ± 0.21	-0.75 ± 0.18

Tab. 14.6: SVM average scores during the disgust expressions for PD patients and HC subjects. The target expression of this task is reported in red, while the highest average scores are reported in bold for both tasks (acted and imitated)

		Acted	Imitated
HC subjects	Neutral	-0.69 ± 0.11	-0.70 ± 0.10
	Anger	**-0.51 ± 0.30**	-0.59 ± 0.34
	Disgust	-0.61 ± 0.13	**-0.54 ± 0.14**
	Happiness	-0.71 ± 0.11	-0.67 ± 0.10
	Sadness	-0.73 ±0.07	-0.74 ± 0.07
PD patients	Neutral	-0.56 ± 0.33	-0.68 ± 0.20
	Anger	**-0.44 ± 0.50**	**-0.38 ± 0.56**
	Disgust	-0.69 ± 0.19	-0.70 ± 0.19
	Happiness	-0.77 ± 0.12	-0.77 ± 0.13
	Sadness	-0.80 ± 0.17	-0.73 ± 0.19

Tab. 14.7: SVM average scores during the happiness expressions for PD patients and HC subjects. The target expression of this task is reported in red, while the highest average scores are reported in bold for both tasks (acted and imitated)

		Acted	Imitated
HC subjects	Neutral	-0.64 ± 0.08	-0.65 ± 0.15
	Anger	-0.78 ± 0.10	-0.77 ± 0.09
	Disgust	-0.58 ± 0.12	-0.59 ± 0.13
	Happiness	**-0.50 ±0.22**	**-0.49 ± 0.34**
	Sadness	-0.74 ±0.06	-0.74 ± 0.08
PD patients	Neutral	**-0.57 ±0.17**	**-0.53 ± 0.26**
	Anger	-0.69 ± 0.29	-0.60 ± 0.47
	Disgust	-0.72 ± 0.19	-0.68 ± 0.20
	Happiness	-0.59 ± 0.28	-0.66 ± 0.19
	Sadness	-0.70 ± 0.15	-0.77 ± 0.15

Tab. 14.8: SVM average scores during the sadness expressions for PD patients and HC subjects. The target expression of this task is reported in red, while the highest average scores are reported in bold for both tasks (acted and imitated)

		Acted	Imitated
HC subjects	Neutral	-0.70 ± 0.09	-0.69 ± 0.21
	Anger	**-0.51 ± 0.30**	**-0.46 ± 0.33**
	Disgust	-0.64 ± 0.12	-0.66 ± 0.12
	Happiness	-0.72 ± 0.10	-0.73 ± 0.10
	Sadness	-0.71 ± 0.11	-0.71 ± 0.12
PD patients	Neutral	-0.51 ± 0.32	-0.64 ± 0.26
	Anger	**-0.50 ± 0.47**	**-0.45 ± 0.52**
	Disgust	-0.71 ± 0.19	-0.71 ± 0.19
	Happiness	-0.78 ± 0.11	-0.76 ± 0.12
	Sadness	-0.76 ± 0.24	-0.71 ± 0.23

14.2.1 Anger

Fig. 14.1 shows that HC subjects perform the expression of anger better than PD patients during the acted task (42.51% vs 29.32%). In both groups, the other most frequent recognized expression is disgust (21.54% in HC subjects and 37.62% in PD patients). Neutral expression is higher in PD

patients than in HC subjects (24.00% vs 16.15%), the percentage of sadness exceeds 10% only in HC subjects (19.29%).

During the imitation task there is a noticeable increase of the target expression only in HC subjects (from 42.51% to 60.98%), while the increase of the anger percentage in PD patients is around 4% (from 29.32% to 33.87%). On the other hand, PD patients show an increase of the neutral expression (from 24.00% to 31.71%), while this expression almost disappears in HC subjects.

Results Tab. 14.5 show that in both groups the highest score is relative to the target expression (anger).

14.2.2 Disgust

Fig. 14.2 shows that HC subjects perform the expression of disgust better than PD patients during the acted task (43.42% vs 35.00%). In both groups, a large percentage of anger is present (43.13% in HC subjects and 29.42% in PD patients), with PD patients also showing higher percentage of neutral (30.28%). During the imitation task there is a large increase of the target expression only in HC subjects (from 43.42% to 76.89%), while in PD patients there is even a decrease of the disgust (from 35.00% to 32.89%) at the expense of an increase of the anger expression (from 29.42% to 38.51%). During the imitation PD patients still show a higher percentage of neutral than HC subjects (18.62% vs 3.56%).

Results in Tab. 14.6 show that in both groups the highest score is related to the anger class during the acted expression. This is compatible with the high anger percentages showed in the upper plots of Fig. 14.2. However, during the imitation task, the average highest score in PD patients is still related to the anger class, in contrast to HC subjects, where the highest score coincides with the target expression.

14.2.3 Happiness

Fig. 14.3 shows that HC subjects perform the expression of happiness better than PD patients during the acted task (31.58% vs 18.09%). However, in both groups, this is not the highest percentage, since disgust is 49.93% in HC subjects and 32.97% in PD patients. During the imitation, the percentage of happiness decreases in both groups (from 31.58% to 29.84% in HC subjects, from 18.09% to 10.00% in PD patients). Also in this case disgust has the highest percentage (54.98% for HC subjects and 30.34% for PD patients), with an increase of the neutral expression in PD patients from 25.19% to 29.44%.

However, results in Tab. 14.7 show that HC subjects have the highest score in correspondence to the target expression during both tasks (acted and imitated), while the highest score for PD patients corresponds to the neutral class in both tasks.

14.2.4 Sadness

Fig. 14.4 shows that both groups have low percentage of sadness. This is the only case when the percentage of the target expression is higher in PD patients than in HC subjects (9.30% vs 5.50% during the acted expression, 11.81% vs 5.00% during the imitating task). In both groups, during the acted expression there are high percentages of disgust (44.58% in HC subjects and 27.99% in PD patients) and anger (42.77% in HC subjects and 31.74% in PD patients). During the acted expression there is a large increase of the anger percentage in HC subjects (from 42.77% to 59.95%), while PD patients show an increase of disgust (from 27.99% to 36.69%), with a consequent decrease of anger (from 31.74% to 28.82%) and neutral (from 30.98% to 22.68%). However, neutral still remains higher in PD patients than in HC subjects for both tasks.
Tab. 14.8 shows that both groups have the highest average score in correspondence to the anger class for both tasks.

14.3 Discussion

Considering PD patients, the percentage of target expression did not undergo to big changes between acted and imitated tasks (Fig. 14.1-14.4). The percentage of neutral expression was always higher than HC subjects, during both tasks. Moreover, PD patients never showed an expression that clearly exceeds the others, as would be expected especially during the imitation. This suggests that PD patients had poor ability to show acted expressions both upon request and after that a visual aid has been shown for the imitation.
In contrast, HC subjects showed higher variations (between acted and imitated tasks) of the target expression than PD patients especially for anger and disgust. Moreover, during the imitation of anger and disgust (Fig. 14.1-14.2) the target expression greatly exceeded the other expressions. This suggests that HC subjects were able to show a certain expression unambiguously upon request, but they clearly improved this ability when the visual aid for the imitation is shown. However, this result was not found during happiness and sadness (Fig. 14.3-14.4). Concerning happiness, the target expression remained stable between acted and imitated tasks and it was not the highest expression in terms of percentage. In fact, a high percentage of disgust was visible during acted and imitated tasks (Fig. 14.3). Watching the acquired video clips, we noticed that many HC subjects exaggerated this expression showing an excessive nose wrinkler, characteristic of disgust. Unlike PD patients, HC subjects did not show high percentages of neutral expression (Fig. 14.3); this means that they tried to exhibit happiness, but sometimes they failed in performing the right expression.
Sadness was the expression with the worst results in terms of target expression percentage. Both groups showed very low percentage of the target expression. PD patients always showed high percentages of neutral, disgust and anger. These three expressions fill almost the entire pie plot in Fig. 14.4. HC subjects, showed similar results but with a clear reduction of the neutral expression. Thus, as

in case of happiness, HC subjects tried to perform the expression but failed to do the right facial movements. In fact, after the experiment, the opinion of both groups was that it is very difficult to express sadness even when there is an image for the imitation. In this task both groups tend to assume different expressions, in particular anger and disgust.

The fact that HC subjects tried to perform happiness and sadness without succeed is proven by the results in Tab. 14.3 and 14.4. In fact, PD patients showed lower maximum and range values of distance during the acted happiness, lower maximum and minimum values of distance during the imitated happiness, a lower mean value of distance during the acted sadness and lower mean, maximum and minimum values of distance during the imitated sadness.

Considering the Euclidean distance from the neutral baseline, PD patients never showed significant differences between acted and imitated expressions (Tab. 14.1-14.4). This confirms the results of pie plots in Fig. 14.1-14.4, that is, the performances of PD patients in exhibiting a certain expression remain stable even after that a visual aid has been shown for the imitation. In contrast, HC subjects showed an increase in mean distance between acted and imitated expressions, for both anger and disgust. This confirms that a higher expressivity is present during the imitation for HC subjects. Other significant variations were found in the minimum value of distance for all the expressions, although the decrease of the range value of distance during happiness indicates that HC subjects did not show big variations (between acted and imitated tasks) of the overall expressivity during this expression.

Considering only the acted expressions we can state that PD patients always showed a higher percentage of neutral than HC subjects (Fig. 14.1-14.4). Distance results showed that no significant differences exist in the overall expressivity between HC subjects and PD patients during anger (Tab. 14.1), although the target expression is higher in HC subjects (Fig. 14.1). During acted disgust and acted sadness, PD patients showed a reduction of the mean distance with respect to HC subjects. Thus, the overall expressivity is reduced for PD patients. PD patients showed a lower expressivity also during the acted happiness, as suggested by a significant reduction of maximum and range values of distance.

The higher percentages of neutral expression in PD patients were confirmed also during the imitated task. The differences between PD patients and HC subjects are greatly enhanced during the imitation, since a significant reduction was detected in: mean distance during anger, disgust and sadness; maximum and minimum distance during anger, happiness and sadness. Thus, the reduction of the overall expressivity of PD patients is more enhanced during the imitated than the acted expressions.

These results showed that anger and disgust are the two expressions in which HC subjects showed a higher increase in the target expression during the imitation task. This means that HC subjects benefited from the displaying of the target expression in order to perform the imitation of those facial expressions. This was not true for PD patients, where the percentage of the target expressions (anger and disgust) remained stable or decreased from the acted task to the imitated task.

In order to quantify the amount of facial movements (and thus the overall expressivity) of both groups, the Euclidean distance of the facial model from the neutral baseline was computed. On average, HC subjects reported higher distances than PD patients along the whole tasks; this confirms that HC subjects showed larger movements during both posed and imitated facial expressions.

From Tables 14.1-14.4 we note that PD patients did not show any significant difference between acted and imitated expressions in all the 4 tasks. This confirms what suggested by the pie plots in Figures 14.1-14.4, namely PD patients did not improve their facial mimicry even when they were asked to imitate another expression. The reduced Euclidean distance from the neutral baseline means that PD patients showed a global reduction of expressivity during the tasks. These results agree with other findings reported in literature [36,40,119,143].

In particular, Simons et al., 2004 [37] found a reduction of posed smiles and difficulties in imitating happiness, surprise and disgust in PD patients (facial movements from participants were coded by a certified FACS coder). The authors found that Parkinsonian patients had difficulties in exhibiting the following AUs during posed expressions: AU4 (brow lowerer), AU9 (nose wrinkler), AU10 (upper lip raiser), AU1+2 (brow raiser) and AU6+12 (cheek raise plus lip corner pull). In particular, AU4 is involved in sadness, fear and anger, AU9 is involved in disgust and AU6+12 are characteristic of happiness. In our case, we did not detected AUs, but the reduced percentage of disgust and anger (Fig. 14.1-14.2) in PD patients, could be due to the impairment in exhibiting these movements: brow lowerer and nose wrinkler.

In all the acquisitions PD patients exhibited higher percentages of neutral with respect to HC subjects. This confirms the presence of a hypomimia that reduced the displaying of other facial expressions. In fact, even when HC subjects did not succeed in reaching the target expression (happiness - Fig.14.3, sadness - Fig. 14.4), the neutral percentage is always lower if compared to PD patients.

As reported in Ricciardi et al., 2015 [143], it is still unclear whether PD patients are unable to show posed expressions, since conflicting results are present in literature. Our results suggest that PD patients have difficulties in exhibiting posed expressions, if compared with HC subjects. However, the most evident result is that PD patients did not improve the ability to reach a requested expression even when a visual aid is proposed (in this case the imitation of an expression); in contrast, HC subjects showed this improvement, especially for anger and disgust.

These results do not provide a solution to the issue about hypomimia in PD (i.e., hypomimia is a pure motor disorder or is the consequence of a failed recognizing process). Actually, our experiments suggest that both the hypotheses may be valid, and that a joint action of two phenomena may be the most reasonable scenario: PD patients had lower performance in reaching a target expression upon request, with lower overall expressivity and facial movements (thus, confirming a motor disorder), but they did not improve these performance when they had to imitate a provided expression.

Limitations of this study are the heterogeneous PD sample and the exclusion of fear and surprise. More accurate results could be provided by analyzing patients with different stages of the disease, in

order to study facial hypomimia along the course of the disease, and maybe drawing more useful information about impairments in expression recognition processes. More information could be extracted by studying AUs, as already done in [36,119]. In particular, the automatic detection of AUs could be an important aid for therapeutic exercises that involve facial mimicry, in order to develop a system for speech therapy and rehabilitation focusing mainly on those movements that seem to be more impaired in PD patients (as already demonstrated by [37]).

Another point that could be elucidated in future works is the cultural dependency of these methods. A still open debate in literature (in particular in psychology) is the cultural specificity in the expression and perception of emotions. Previous works [116] demonstrated that it is difficult to draw firm conclusions about the expression of emotions across cultures, although some differences can be present. This issue could be focused also in PD patients of different cultures, in order to see if the disease can modify inter-cultural aspects and in which way.

A clear advantage of this study is the objectification of facial hypomimia in PD patients by means of a contactless technology. Although Wu et al., 2014 [119] already performed an analysis of facial mimicry in PD patients through AUs classification techniques, in that study the authors recorded a group of PD patients monitored with facial EMG and ECG. The automatic analysis of facial expressions is a continuously evolving field of research that finds several applications also in medicine [118]. In particular, we believe that this system can be used to extract objective measures of facial mimicry constituting an important means for analyzing facial expressions also in rehabilitation framework (in particular speech therapy), whereby patients may obtain definite advantages about a real-time feedback of the right facial expressions/movements to perform.

15. Contact-less video-based tracking of heart rate[6]

15.1 Results

Mean values and standard deviations of the RMSE for the two estimation methods (AR-RLS and Wavelet-based methods) are reported in Tab. 15.1. Both methods were compared with the reference ECG measures for the 3 independent sources (Source 1, Source 2 and Source 3 in Tab. 15.1) extracted from the two facial ROIs: forehead and cheeks.

Tab. 15.1: Mean value and standard deviation of RMSE for both estimation methods (AR-RLS and Wavelet-based methods) computed on the three independent sources extracted from the two facial ROIs (forehead and cheeks). These methods were compared with reference HR values extracted from an ECG. Values are expressed in bpm.

ROI	Independent Source	AR-RLS	Wavelet
Cheeks	Source 1	14.42 ± 4.25	14.75 ± 3.59
	Source 2	16.44 ± 2.00	16.81 ± 1.84
	Source 3	11.10 ± 3.50	12.65 ± 2.64
Forehead	Source 1	12.48 ± 2.59	12.90 ± 2.60
	Source 2	14.79 ± 0.99	16.70 ± 2.34
	Source 3	9.58 ± 2.82	11.67 ± 2.25

Among the sources of both ROIs the one who gives the best results was Source 3 extracted from the forehead (Tab. 15.1). Considering this signal for the estimation, the AR-RLS methods gave better results than the wavelet-based one, with mean RMSE around 9 bpm.

Fig. 15.1 shows the results of both estimation methods compared to the reference values (blue line). The HR estimation was performed on the third independent source extracted from the forehead.

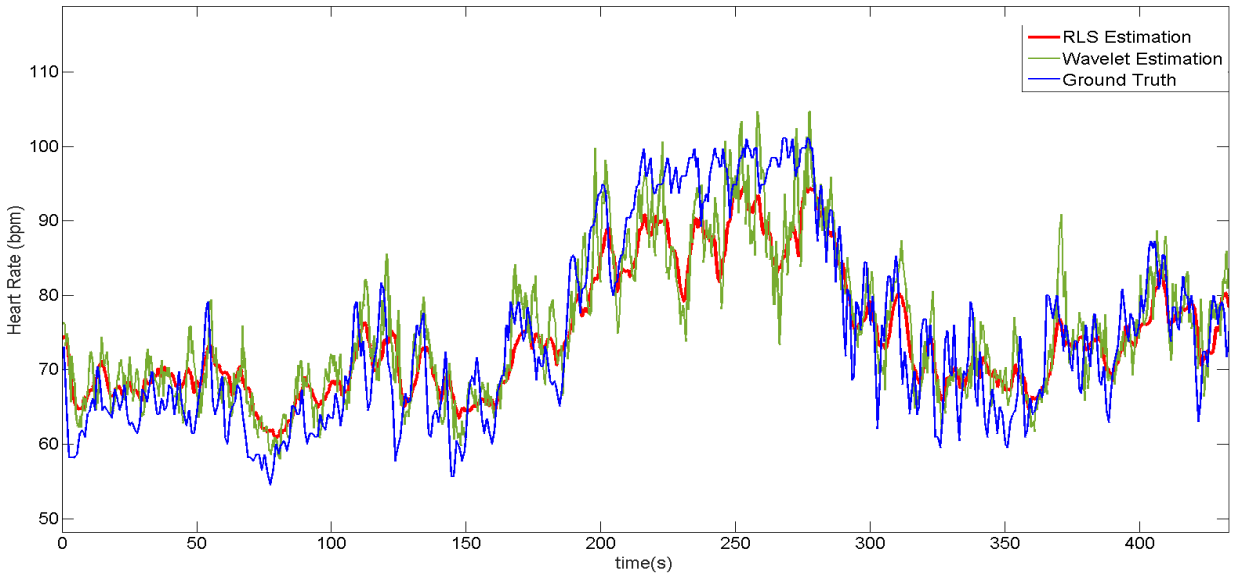

Fig. 15.1: Comparison between the AR-RLS estimation (red) wavelet estimation (green) and reference data (blue).

[6] Part of these results, as well as methods discussed in Part II - Chapter 9 can also be found in [208]

15.2 Discussion

A recursive HR estimation method from video signals was proposed. This method was compared against a reference ECG and against a wavelet-based method. Results in Tab. 15.1 highlight that the best results can be achieved using the third source signal extracted from the forehead. The use of a recursive estimation method allowed tracking heart rate variations under mild physical exercise (Fig. 15.1).

Our results seem to be in contrast with other proposed methods [146,148], where the best source for HR estimation was the second one. Although one of the problems of ICA is the lack of an order between the estimated sources (this leads to difficulties on how to interpret the sources [150]), in those works the authors stated that the second one could be related to the green channel of color images, where the highest plethysmographic content is present. The best result in terms of accuracy was 4.63 bpm in presence of motion artifacts [148].

On the other hand, Monkaresi et al., 2014 [160], demonstrated that useful information for HR estimation could be contained in all the estimated ICA sources. Performing a similar task to that used in our work (indoor cycling), the authors used machine learning techniques (linear regression and k-nearest neighbor) to predict the best source for HR estimation within a temporal window. However, when they considered the acquisition as a whole, the third source gave the best results. The estimation error reached values around 4 bpm when the authors combined ICA and k-NN [160].

In our work, the errors against the reference HR are higher than those proposed in [148,160]. Nevertheless our method is able to track HR variations during physical exercises and motion artifacts. However, since our aim is to apply this method to study DOC patients and their reactions after the administration of external stimuli, this accuracy could be acceptable for our purposes. In fact, this kind of patients does not perform large movements of the head (especially patients in persistent vegetative state). Indeed the higher error values could be due to the head motion that occurs when subjects are pedaling. Because of its recursive nature, an improvement could be the implementation of a variable forgetting factor, in order to decrease the memory of the system during rapid HR changes and increase it during the stable phases. This would allow a better tracking of the varying dynamics of the system.

16. Video analysis of DOC patients: facial movements and HR

16.1 Results

Mean values of Euclidean distance and HR for each subject are reported in Tab. 16.1. As not all patients had the same number of stimuli, two analyses were performed. The first one considered those patients who had at least 3 stimuli after the baseline (N = 7) and the second one those patients who had at least 2 stimuli after the baseline interval (N = 8).

16.1.1 Analysis on 3 stimuli

A one-way repeated-measures analysis of variance (ANOVA) was performed to compare the effect of the noxious stimuli over the 4 time instants (before the first stimulus - baseline - and after the first 3 stimuli) on the Euclidean distance of facial features from the neutral template and on the HR estimation. This analysis was performed on 7 patients.

The repeated-measures ANOVA showed that the stimulus administration produced a significant increase in the Euclidean distance from the neutral template over the 4 time instants, $F(3,18) = 7.1994$, $p = .002$. No significant differences were found in HR values over the 4 time instants, $F(3,18) = 2.6155$, $p = .08$. Boxplots of Euclidean distance and HR over the considered 4 time intervals are reported in Figures 16.1 and 16.2.

For both measures (distance and HR), three paired t-tests were used to make post hoc comparisons between the following conditions: baseline vs 1st stimulus, baseline vs 2nd stimulus and baseline vs 3rd stimulus. The paired t-tests performed on the distance reported the following results: no significant differences exist between the baseline (mean: 10.99, SD: 6.97) and after the administration of the first stimulus (mean: 13.79, SD: 9.18), $t(6) = 1.73$, $p = .13$; there was a significant difference between the baseline and the interval after the second stimulus (mean: 19.28, SD: 10.09), $t(6) = 3.05$, $p = .02$, and between the baseline and the interval after the third stimulus (mean: 24.49, SD: 13.48), $t(6) = 4.21$, $p = .006$. Concerning HR, no significant differences were found during the comparison between the baseline and the three intervals after the stimuli. Mean values and standard deviations of distance and HR for the 4 time intervals are reported in Tab. 16.2.

No significant correlations were found between distance and CRS-R score, age and months after the onset and between HR, CRS-R score, age and months after the onset.

16.1.2 Analysis on 2 stimuli

A one-way repeated-measures ANOVA was performed to compare the effect of the noxious stimuli over the 2 time instants (before the first stimulus - baseline - and after the first 2 stimuli) on the

Euclidean distance of facial features from the neutral baseline and on the HR estimation. This analysis was performed on 8 patients.

The repeated-measures ANOVA showed that no significant differences exist in the Euclidean distance from the neutral template over the 3 time instants, $F(2,14) = 2.4275$, $p = .12$, while a significant difference exists in the HR estimation, $F(2,14) = 6.1699$, $p = .012$. Boxplots of Euclidean distance and HR over the 3 time intervals are reported in Figures 16.3 and 16.4.

For both measures (distance and HR), two paired t-tests were applied to make post hoc comparisons between the following conditions: baseline vs 1st stimulus and baseline vs 2nd stimulus. The paired t-tests performed on the distance reported the following results: no significant differences exist between the baseline (mean: 10.87, SD: 6.46) and after the administration of the first stimulus (mean: 13.20, SD: 8.67), $t(7) = 1.56$, $p = .16$; there was a significant difference between the baseline and the time interval after the second stimulus (mean: 17.77, SD: 10.28), $t(7) = 2.52$, $p = .04$. Concerning HR, there was a significant difference between the baseline and the time interval between the first stimulus (mean:, SD:), $t(7) = -2.48$, $p = .04$, while no significant differences were found during the comparison between the baseline and the interval after the second stimulus. Mean values and standard deviations of distance and HR for the 3 time intervals are reported in Tab. 16.2.

No significant correlations were found between distance and CRS, age and months after the onset and between HR, CRS, age and months after the onset.

Tab. 16.1: Mean values of Euclidean distance and HR for each patient considered in the study. Seven patients (1,2,3,4,5,7 and 8) were considered for the analysis on the first 3 stimuli, while 8 patients (1 to 8) were considered for the analysis on the first 2 stimuli.

Patient	Gender	Age (years)	Onset (months)	CRS-R score	Baseline	Stimulus1	Stimulus2	Stimulus3	Stimulus4	
P1	F	41	29	7	3.65	6.29	20.39	20.31	-	Distance
					82.60	77.84	70.99	84.02	-	HR (bpm)
P2	M	56	22	6	22.88	28.92	36.80	44.63	-	Distance
					78.55	75.69	72.39	76.98	-	HR (bpm)
P3	F	35	93	7	4.16	4.01	11.11	10.48	-	Distance
					83.45	81.62	85.45	87.17	-	HR (bpm)
P4	M	80	15	6	12.23	14.00	20.76	35.34	14.93	Distance
					88.28	82.57	85.52	92.48	89.16	HR (bpm)
P5	M	25	14	7	16.29	14.98	14.92	26.90	27.59	Distance
					73.36	75.50	76.37	82.69	86.85	HR (bpm)
P6	M	50	20	9	10.08	8.97	7.18	-	-	Distance
					87.61	84.59	86.72	-	-	HR (bpm)
P7	M	48	4	5	6.51	6.34	5.85	6.12	6.72	Distance
					81.43	82.39	79.15	81.83	88.06	HR (bpm)
P8	F	51	29	4	11.18	22.04	25.16	27.67	28.91	Distance
					81.74	74.83	74.10	77.07	77.58	HR (bpm)
P9	M	66	59	5	17.43	16.15	-	-	-	Distance
					110.80	95.08	-	-	-	HR (bpm)

Tab. 16.2: Mean values and standard deviations for Euclidean distance and HR during the two analyses (N = number of patients; * $p < .05$, ** $p < .01$)

Analysis	Measure	Baseline	Stim1	Stim2	Stim3
3 stimuli	Distance	10.99 ± 6.97	13.79 ± 9.18	19.28 ± 10.09 *	24.49 ± 13.48 **
(N = 7)	HR	81.34 ± 4.58	78.63 ± 3.47	77.71 ± 5.94	83.18 ± 5.49
2 stimuli	Distance	10.87 ± 6.46	13.20 ± 8.67	17.77 ± 10.28 *	-
(N = 8)	HR	82.13 ± 4.78	79.38 ± 3.84 *	78.84 ± 6.35	-

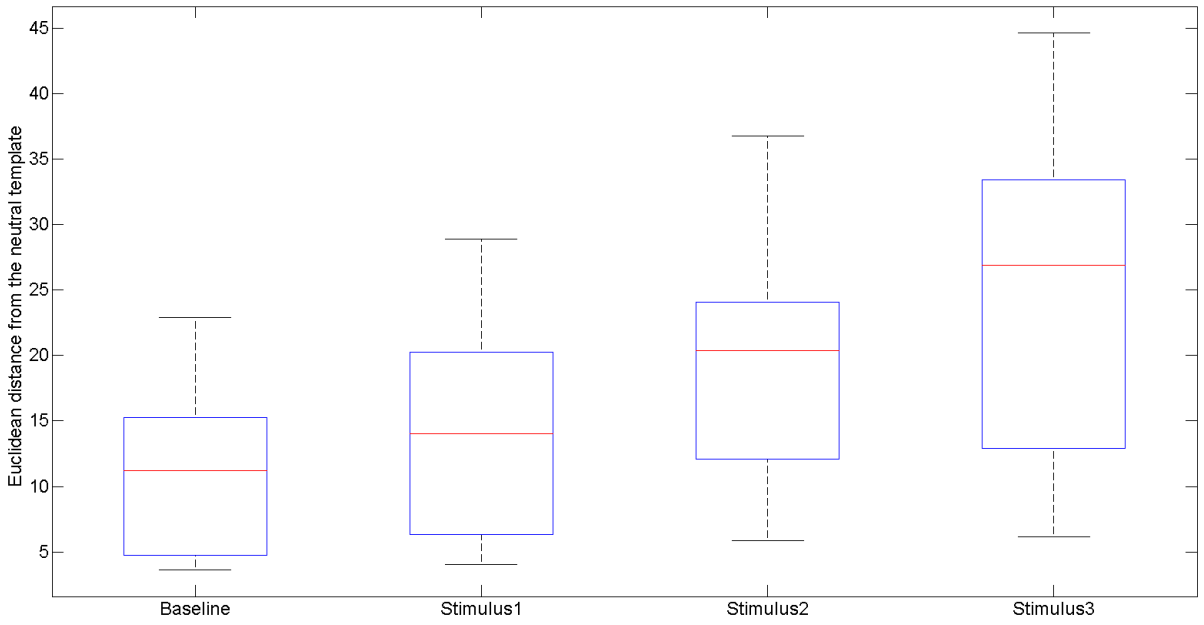

Fig. 16.1: Boxplots of the Euclidean distance of facial features from the neutral template over the 4 time interval considered for the study (baseline - before the administration of the first stimulus, and after the administration of the first 3 stimuli).

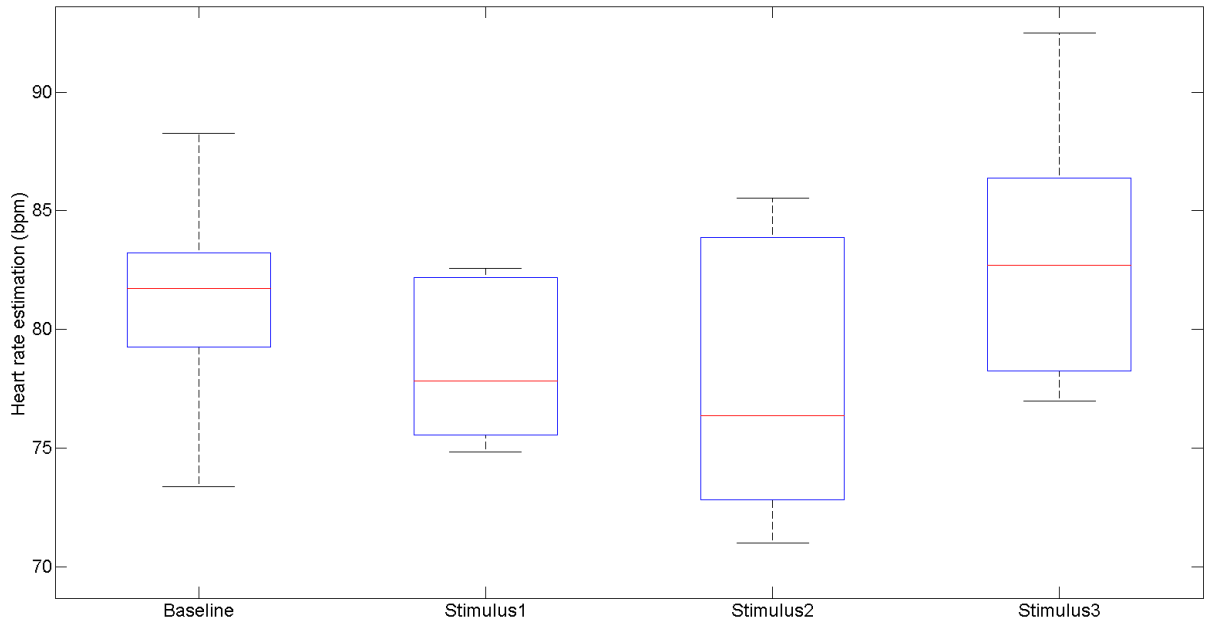

Fig. 16.2: Boxplots of the HR over the 4 time interval considered for the study (baseline - before the administration of the first stimulus, and after the administration of the first 3 stimuli).

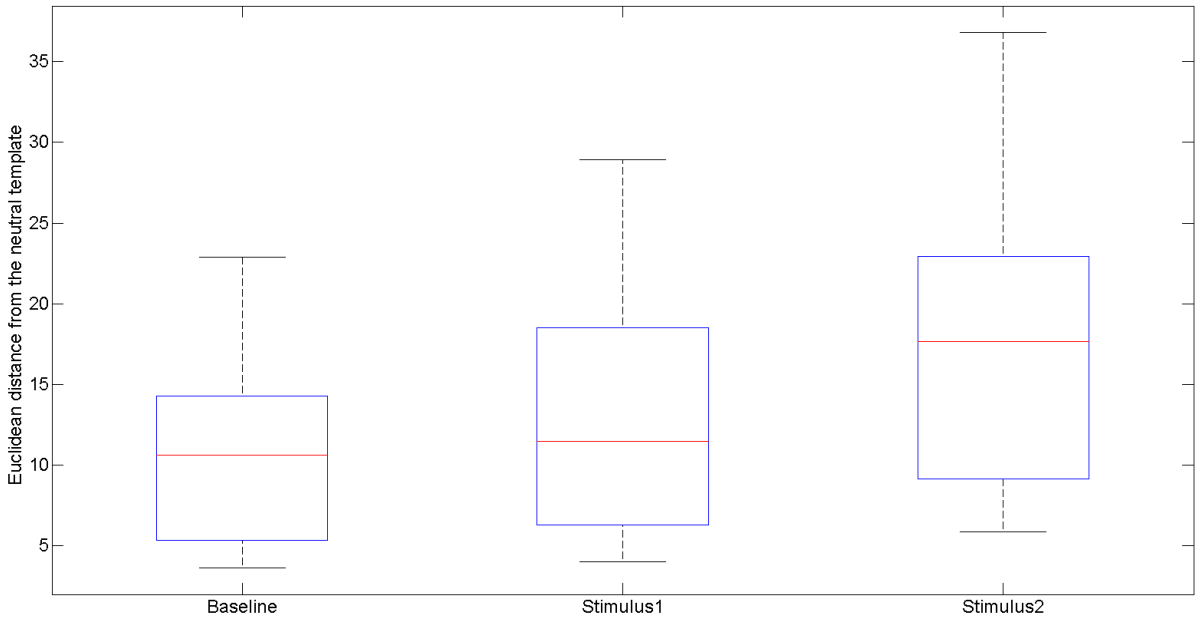

Fig. 16.3: Boxplots of the Euclidean distance of facial features from the neutral template over the 3 time interval considered for the study (baseline - before the administration of the first stimulus, and after the administration of the first 2 stimuli).

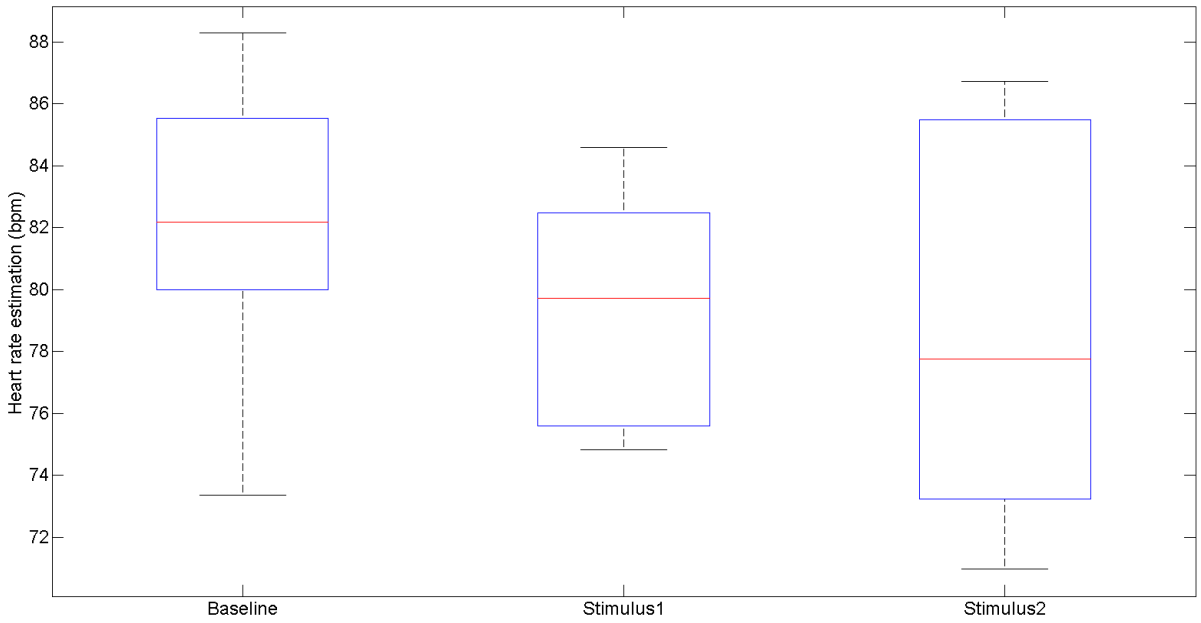

Fig. 16.4: Boxplots of the HR over the 3 time interval considered for the study (baseline - before the administration of the first stimulus, and after the administration of the first 2 stimuli).

16.2 Discussion

As reported in Fig. 16.1 and 16.3 patients showed a significant increase of the Euclidean distance of facial features from the neutral template during the course of the experiment (in particular after the second stimulus). This is consistent with an increase of facial mimicry and facial movements that might indicate a patient's reaction. This distance has lower values during the baseline time interval

(i.e. before the administration of the first stimulus), reflecting a lower activity in facial mimicry with a facial expression similar to the neutral template. However, the same visible trend is not present for HR (figures 16.2 and 16.4); in fact, results in Tab 16.2 confirm that the mean HR remains stable during the first 3 stimuli, around 80 bpm. Moreover, the variations along the administered stimuli are very small (around 5 bpm on average) and thus comparable with the accuracy of this estimation method that in the previous chapter was demonstrated to be around 9 bpm. Thus, it is not possible to draw any conclusion on HR variations, although it is reasonable to suppose that no large variations occurred in DOC patients, at least considering the HR.

This could depend on the small number of patients considered for this study (only 9 patients). These patients had different etiology, different age and different post-comatose outcome (in this case we had 7 VS patients and only 2 MCS patients). More accurate results could be obtained with a larger data set, differentiating between VS and MCS patients. However, this is a pilot study and a first attempt to perform a contactless automated monitoring of DOC patients, that could be easily extended to larger populations and to other CRS-R items (i.e., items for the evaluation of auditory, communicative functions), as well as to the interaction of patients with their relatives. In fact, some neuroimaging studies [75,76] demonstrated that an activation in some cerebral areas (such as the amygdala) was found in MCS patients after listening to a familiar voice, but is absent in VS patients. Thus, our study could be extended in order to find possible differences in facial mimicry and HR in larger groups of MCS and VS patients, evaluating both within and between-groups differences.

Conclusion

This PhD project provides first results concerning the development of a markerless system for monitoring facial expressions, facial movements and speech with applications to neurology (in particular Parkinson's disease and disorders of consciousness).

A markerless method for tracking articulatory movements during speech was tested against a marker-based optoelectronic reference, obtaining good results in terms of accuracy. Interesting results were obtained in PD patients were through this markerless and low-cost approach it was demonstrated what already found in literature with more expensive techniques, that is PD patients exhibit reduced kinematic parameters of the articulatory movements (speed and acceleration of lips) during speech and reduced extents of facial expressions. Both are related to facial bradykinesia. Thus, the proposed methods could be implemented at home in order to help these patients in performing speech therapy and rehabilitation exercises, since a large percentage of them suffers from dysarthria and facial hypomimia.

Interesting results were also found as far as the acoustical analysis of speech is concerned, with PD patients that show speech rate alterations during syllable and sentence repetition tasks. Thus, the analysis of facial expressions and articulatory movements, in conjunction with the acoustical analysis of speech signals can provide an important means for tracking the disease progression in different diseases that involve impairments in the speech production and facial movements. Moreover, the methodologies proposed in this thesis can also be extended for the investigation of other disorders, such as amyotrophic lateral sclerosis, Alzheimer's disease, stroke, etc.

The analysis of facial movements and facial expressions was also applied to DOC patients, in order to highlight possible reactions after the administration of external stimuli. This analysis, in conjunction with HR estimation through video-based techniques, highlighted some changes with a clear and significant trend only for facial expressions. Thus, these results need to be extended to larger populations differentiating between post-comatose outcomes. However, the first results obtained during this project might be a first step towards the development of an automated monitoring system for DOC patients, in order to assess patients' reactions even when nobody is near them. This could provide a reliable support to clinicians in the differential diagnosis between vegetative state and minimally conscious state that today is still prone to high rates of error.

References

[1] Damier P, Hirsch EC, Agid Y, Graybiel AM. The substantia nigra of the human brain II. Patterns of loss of dopamine-containing neurons in Parkinson's disease, Brain. 1999; 199:1437-1448.

[2] Braak H, Del Tredici K, Rüb U, de Vos RAI, Jansen Steur ENH, Braak E. Staging of brain pathology related to sporadic Parkinson's disease. Neurobiol Aging. 2003; 24: 197–211.

[3] Dickson JM, Grünewald RA. Somatic symptom progression in idiopathic Parkinson's disease. Parkinsonism Relat Disord. 2004; 10: 487–492.

[4] Dickson DW. Parkinson's disease and parkinsonism: neuropathology. Cold Spring Harb Perspect Med. 2012; 2(8): 1–15.

[5] Chaudhuri KR, Healy DG, Schapira AHV. Non-motor symptoms of Parkinson's disease: diagnosis and management. Lancet Neurol. 2006; 5: 235–245.

[6] www.parkinson.it. Accessed: March 14, 2016.

[7] Kraus PH, Lemke MR, Reichmann H. Kinetic tremor in Parkinson's disease - an underrated symptom. J Neural Transm. 2006; 113: 845-853.

[8] Giuberti M, Ferrari G, Contin L, et al. On the characterization of leg agility in patiens with Parkinson's disease. In: 10th International Conference on Wearable and Implantable Body Sensor Networks (BSN) of IEEE. Cambridge, MA, USA; May 2013; 1–6.

[9] Giuberti M, Ferrari G, Contin L, et al. Linking UPDRS scores and kinematic variables in the leg agility task of Parkinsonians. In: 11th International Conference on Wearable and Implantable Body Sensor Networks (BSN) of IEEE. Zurich; June 2014; 115-120.

[10] Italian Ministry of Health. Linea Guida 24. Diagnosi e terapia della malattia di Parkinson; 2013: http://www.parkinson-italia.it/upl/cms/attach/20130530/150328056_1050.pdf

[11] Gelb DJ, Oliver E, Gilman S. Diagnostic criteria for Parkinson's disease. Arch Neurol. January 1999; 56(1):33-9.

[12] Klinke R, Pape H, Kurtz A, Silbernagl S. Fisiologia. 3rd ed. Napoli, Italy: Edises; 2012.

[13] http://nawrot.psych.ndsu.nodak.edu/Courses/465Projects15/Parkinsons/PDtermproject.htm. Accessed: January 25, 2015.

[14] Hughes AJ, Daniel SE, Kilford L, Lees AJ. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinic-pathological study of 100 cases. J Neurol Neurosur Ps. 1992; 55: 181-184.

[15] Schrag A, Jahanshahi M, Quinn N. What contributes to quality of life in patients with Parkinson's disease?. J Neurol Neurosur Ps. 2000; 69: 308-312.

[16] Goetz CG, Poewe W, Rascol O, et al. Movement disorder society task force report on the Hoehn and Yahr staging scale: status and recommendations. Movement Disord. 2004;19(9): 1020–1028.

[17] Fahn S, Elton R, Members of the UPDRS Development Committee. In: Fahn S, Marsden CD, Calne DB, Goldstein M, Eds. Recent Development in Parkinson's Disease. vol. 2. Florham Park, NJ: Macmillan Health Care Information; 1987: 153-163, 293-304.

[18] Moskovitz C, Moses H, Klawans HL. Levodopa-induced psychosis: a kindling phenomena. Am J Psychiatry. 1978. 135: 669-675.

[19] Goldman JG, Goetz CG, Berry-Kravis E, Leurgans S, Zhou L. Genetic polymorphisms in Parkinson disease subjects with and without hallucinations: an analysis of the cholecystokinin system. Arch Neurol. 2004; 23: 2398-2403.

[20] Pezzella FR, Colosimo C, Vanacore N, et al. Prevalence of clinical features of hedonistic homeostatic dysregulation in Parkinson's disease. Movement Disord. 2005; 20: 77-81.

[21] Calabresi P, Di Filippo M, Ghiglieri V, Tambasco N, Picconi B. Levodopa-induced dyskinesias in patients with Parkinson's disease: filling the bench-to-bedside gap. Lancet Neurol. September 2010; 9:1106-17.

[22] Mera TO, Filipkowskia DE, Riley DE, et al. Quantitative analysis of gait and balance response to deep brain stimulation in Parkinson's disease. Gait Posture. 2013; 38(1): 109–114.

[23] DeLong MR. Discovery of High-Frequency Deep Brain Stimulation for Treatment of Parkinson Disease: 2014 Lasker Award. JAMA-J Am Med Assoc. 2014; 312(11):1093-1094.

[24] Shukla AW, Okun MS. Surgical Treatment of Parkinson's Disease: Patients, Targets, Devices, and Approaches. Neurotherapeutics. 2014; 11:47–59.

[25] Skodda S. Effect of deep brain stimulation on speech performance in Parkinson's disease. Parkinsons Dis. 2012. 2012: 850596.

[26] Galna B, Jackson D, Schofield G, et al. Retrainig function in people with Parkinson's disease using Microsoft Kinect: game design and pilot testing. J Neuroeng Rehabil. 2014; 11:60.

[27] Pompeu JE, dos Santos Mendes FA, da Silva KG, et al. Effect of Nintendo Wii™-based motor and cognitive training on activities of daily living in patients with Parkinson's disease: a randomized clinical trial. Physiotherapy. 2012; 98: 196-204.

[28] Galna B, Barry G, Jackson D, Mhiripiri D, Oliver P. Accuracy of the Microsoft Kinect sensor for measuring movement in people with Parkinson's disease. Gait Posture. 2014; 39:1062-1068.

[29] Hartelius L, Svensson P. Speech and swallowing symptoms associated with Parkinson's disease and multiple sclerosis: a survey. Folia Phoniatr Logop. 1994; 46: 9–17.

[30] Darley FL, Aronson AE, Brown JR. Motor Speech Disorders. Philadelphia, PA: Saunders; 1975.

[31] Romualdi P. Diagnosi differenziale rispetto alle afasie e alle aprassie nel linguaggio. Omega: 1993.

[32] Ruoppolo G, Schindler A, Amitrano A, Genovese E. Manuale di foniatria e logopedia. Roma, Italy: SEU Società Editrice Universo; 2012.

[33] http://www.d.umn.edu/~mmizuko/2230/msd.htm. Accessed: March 14, 2016.

[34] Tickle-Degnen L, Doyle Lyons K. Practitioners' impressions of patients with Parkinson's disease: the social ecology of the expressive mask. So Sci Med. 2004; 58: 603-614.

[35] Hemmesch AR, Tickle-Degnen L, Zebrowitz LA. The influence of facial masking and sex on older adults' impressions of individuals with Parkinson's disease. Psychol Aging. September 2009; 24(3): 542-549.

[36] Simons G, Smith Pasqualini MC, Reddy V, Wood J. Emotional and nonemotional facial expressions in people with Parkinson's disease. J Int Neuropsych Soc. 2004; 10: 521-535.

[37] Simons G, Ellgring H, Smith Pasqualini MC. Disturbance of spontaneous and posed facial expressions in Parkinson's disease. Cognition Emotion. 2003; 17: 759–778.

[38] Jacobs DH, Shuren J, Bowers D, Heilman KM. Emotional facial imagery, perception, and expression in Parkinson's disease. Neurology. 1994; 45: 1696–1702.

[39] Madeley P, Ellis AW, Mindham, RHS. Facial expressions and Parkinson's disease. Behav Neurol. 1995; 8: 115–119.

[40] Bowers D, Miller K, Bosch W, et al. Faces of emotion in Parkinsons disease: micro-expressivity and bradykinesia during voluntary facial expressions. J Int Neuropsych Soc. 2006; 12: 765-773.

[41] Bologna M, Fabbrini G, Marsili L, et al. Facial Bradykinesia. J Neurol Neurosur Ps. 2013; 84: 681-685.

[42] Özekmekçi S, Benbir G, Özdogan FY, Ertan S, Kiziltan ME. Hemihypomimia, a rare persistent sign in Parkinson's disease. J Neurol. 2007. 254: 347-350.

[43] Weiss D, Wächter T, Breit S, et al. Involuntary eyelid closure after STN-DBS: evidence for different pathophysiological entities. J Neurol Neurosurg Ps. 2010; 81: 1002-1007.

[44] Mergl R, Mavrogiorgou P, Hegerl U, Juckel G. Kinematical analysis of emotionally induced facial expressions: a novel tool to investigate hypomimia in patients suffering from depression. J Neurol Neurosurg Ps. 2005; 76: 138-140.

[45] Schulz GM, Grant M. Effects of speech therapy and pharmacologic and surgical treatments in voice and speech in Parkinson's disease: a review of the literature. J Commun Disord. 2000; 33: 59-88.

[46] Ramig LO, Fox C, Sapir S. Parkinson's disease: speech and voice disorders and their treatment with the Lee Silverman Voice Treatment. Semin Speech Lang. 2004; 25(2): 169-180.

[47] S, Spielman JL, Ramig LO, Story BH, Fox C. Effects of intensive voice treatment (the Lee Silverman Voice Treatment [LSVT]) on vowel articulation in dysartrhic individuals with idiopathic Parkinson disease: acoustic and perceptual findings. J Speech Lang Hear R. 2007; 50: 899-912.

[48] https://www.lsvtglobal.com/patient-resources/what-is-lsvt-loud. Accessed: March 14, 2016.

[49] Burattin M, Caligari M, Deambrogio R, et al. Manuale di autoriabilitazione a domicilio del paziente affetto da malattia di Parkinson. 2012. http://www.parkinson-italia.it/upl/cms/attach/20120611/121001778_9088.pdf. Accessed: March 14, 2016.

[50] https://www.ars.toscana.it/it/aree-dintervento/problemi-di-salute/gravi-cerebrolesioni-acquisite.html. Accessed: March 14, 2016.

[51] Laureys S, Perrin F, Brédart S. Self-consciousness in non-communicative patients. Conscious Cogn. 2007; 16: 722–741.

[52] Plum F, Posner JB. The diagnosis of stupor and coma. 3rd ed. New York, NY: Wiley; 1983.

[53] Amantini A, Ragazzoni A. Documento sulle indagini neurofisiologiche nei pazienti con DOC. http://www.sinc-italia.it/FCKFiles/allegati_articoli/34_allegato_Documento_Amantini-RagazzoniinviatoCDSINC.doc. Accessed: March 14, 2016.

[54] Jennett B, Plum F. Persistent vegetative state after brain damage: A syndrome in search of a name. Lancet. 1972; 1: 734-737.

[55] Chatelle C, Laureys S, Majerus S, Schnakers C. Eye gaze and conscious processing in severely brain-injured patients. Behav Brain Sci. 2010; 33(6): 432-433.

[56] The Multi-Society Task Force on PVS. Medical aspects of the persistent vegetative state-first of two parts. N. Engl. J. Med. 1994; 330:1499–1508.

[57] Owen AM, Menon DK, Johnsrude IS, et al. Detecting Residual Cognitive Function in Persistent Vegetative State. Oxford University, Neurocase. 2002; 8: 394–403.

[58] Giacino JT, Ashwal S, Childs N, et al. The minimally conscious state: Definition and diagnostic criteria. Neurology. 2002; 58: 349-353.

[59] Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet. 1974; 2(7872): 81-4.

[60] Gabbe BJ, Cameron PA, Finch CF. The status of the Glasgow Coma Scale. Emerg Med. 2003; 15: 353-360.

[61] Giacino JT, Kalmar K, Whyte J. The JFK Coma Recovery Scale-Revised: measurement characteristics and diagnostic utility. Arch Phys Med Rehabil. 2004; 85: 2020–2029.

[62] Lombardi F, Gatta G, Sacco S, Muratori A, Carolei A. The Italian version of the Coma Recovery Scale-Revised (CRS-R). Funct Neurol. 2007; 22(1): 47-61.

[63] Gill-Thwaites H, Munday R. The sensory modality assessment and rehabilitation technique (SMART): a valid and reliable assessment for vegetative state and minimally conscious state patients. Brain Injury. 2004; 18:1255– 1269.

[64] Pape TL, Heinemann AW, Kelly JP. A measure of neurobehavioral functioning after coma. Part I: Theory, reliability, and validity of disorders of consciousness scale. J Rehabil Res Dev. 2005; 42:1–18.

[65] Pape TL, Senno RG, Guernon A. A measure of neurobehavioral functioning after coma. Part II: Clinical and scientific implementation. J Rehabil Res Dev. 2005; 42:19–28.

[66] Riganello F, Dolce G, Cortese MD, Sannita WG. Responsiveness And Prognosis In The Severe Disorder Of Consciousness. In: Schaffer A, Muller J, Eds. Brain Damage: Causes, Management and Prognosis. Hauppage, NY: Nova Publisher, Inc; 2012: 117-136.

[67] Giacino JT, Kalmar K. The Vegetative and Minimally Conscious State: a Comparison of Clinical Features and Functional Outcome. J Head Trauma Rehabil. 1997; 12(4): 36-51.

[68] Childs NL, Mercer WN, Childs HW. Accuracy of diagnosis of persistent vegetative state. Neurology. 1993; 43(8):1465-7.

[69] Andrews K, Murphy L, Munday R, Littlewood C. Misdiagnosis of the vegetative state: retrospective study in a rehabilitation unit. Brit Med J. 1996; 313(7048): 13-16.

[70] Schnakers C, Vanhaudenhuyse A, Giacino J, et al. Diagnostic accuracy of the vegetative and minimally conscious state: Clinical consensus versus standardized neurobehavioral assessment. BMC Neurol. 2009; 9:35.

[71] Hawker GA, Mian S, Kendzerska T, French M. Measures of adult pain. Arthritis Care Res. 2011; 63(11): 240-252.

[72] Schnakers C, Chatelle C, Vanhaudenhuyse A, et al. The Nociception Coma Scale: A new tool to assess nociception in disorders of consciousness. Pain. 2010: 148(2): 215-219.

[73] Warden V, Hurley AC, Volicer L. Development and psychometric evaluation of the Pain Assessment in Advanced Dementia (PAINAD) scale. J Am Med Dir Assoc. 2003; 4(1): 9-15.

[74] Kobylarz EJ, Schiff ND. Neurophysiological correlates of persistent vegetative and minimally conscious states. Neuropsychol Rehabil. 2005;15(3-4): 323–332.

[75] Bekinschtein T, Niklison J, Sigman L, et al. Emotion processing in the minimally conscious state. J Neurol Neurosur Ps. 2004; 75: 788–793.

[76] Laureys S, Perrin F, Faymonville M-E, et al. Cerebral processing in the minimally conscious state. Neurology. 2004; 63:916–918.

[77] Boly M, Faymonville M, Damas P, Lambermont B, Del Fiore G, Degueldre C. Auditory processing in severely brain injured patients: Differences between the minimally conscious state and the vegetative state. Arch Neurol. 2004; 61: 233–238.

[78] Kulkarni VP, Lin K, Benbadis SR. EEG findings in the persistent vegetative state. J Clin Neurophysiol. 2007; 24: 433-437.

[79] Coleman MR, Bekinschtein T, Monti MM, Owen AM, Pickard JD. A multimodal approach to the assessment of patients with disorders of consciousness. Prog Brain Res. 2009; 177: 231-248.

[80] Riganello F, Candelieri A, Quintieri M, Conforti D, Dolce G. Heart rate variability: an index of brain processing in vegetative state? An artificial intelligence, data mining study. J Clin Neurophysiol. 2010; 121: 2024-2034.

[81] Flotta L, Riganello F, Sannita WG. Intelligent monitoring of subject with severe disorder of consciousness. In: Proceedings of the 4th International Conference on Sensor Device

Technologies and Applications (SENSORDEVICES 2013); Aug 25-31, 2013; Barcelona, Spain.

[82] Harel BL, Cannizzaro MS, Cohen H, Reilly N, Snyder PJ. Acoustic characteristics of Parkinsonian speech: a potential biomarker of early disease progression and treatment. J Neurolinguist. 2004;17: 439–453.

[83] Rusz J, Cmejla R, Ruzickova H, et al. Evaluation of speech impairment in early stages of Parkinson's disease: a prospective study with the role of pharmacotherapy. J Neural Transm. 2013; 120: 319–329.

[84] Rusz J, Cmejla R, Ruzickova H, Ruzicka E. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease J Acoust Soc Am. 2011; 129(1): 350-367.

[85] Little MA, McSharry PE, Hunter EJ, Spielman J, Ramig LO. Suitability of dysphonia measurements for telemonitoring of Parkinson's disease. IEEE Trans Biomed Eng. 2009; 56(4): 1015-1022.

[86] Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng. 2012; 59 (5): 1264-1271.

[87] Rusz J, Cmejla R, Tykalova T. Imprecise vowel articulation as a potential early marker of Parkinson's disease: effect of speaking task. J Acoust Soc Am. 2013; 134(3): 2171-2181.

[88] Walsh B, Smith A. Basic parameters of articulatory movements and acoustics in individuals with Parkinson's disease. Movement Disord. 2012: 27(7): 843–850.

[89] Skodda S, Visser W, Schlegel U. Vowel articulation in Parkinson's disease. J Voice. 2011; 25(4): 467–472.

[90] Skodda S. Erratum: Vowel articulation in Parkinson's disease. 2011; 25: 467–472.

[91] Sapir S, Ramig LO, Spielman JL, Fox C. Formant Centralization Ratio (FCR): a proposal for a new acoustic measure of dysarthric speech. J Speech Lang Hear Res. 2010; 53(1): 1-20.

[92] Skodda S, Rinsche H, Schlegel U. Progression of dysprosody in Parkinson's disease over time - a longitudinal study. Movement Disord. 2009; 24: 716–722.

[93] Skodda S, Visser W, Schlegel U. Short- and long-term dopaminergic effects on dysarthria in early Parkinson's disease. J Neural Transm. 2010; 117: 197-205.

[94] Skodda S, Flasskamp A, Schlegel U. Instability of syllable repetition as a model for impaired motor processing: is Parkinson's disease a rhythm disorder? J Neural Transm. 2010; 117: 605–612.

[95] Skodda S, Visser W, Schlegel U. Gender-related patterns of dysprosody in Parkinson's disease and correlation between speech variables and motor symptoms. J Voice. 2011; 25(1): 76–82.

[96] Boersma P, Weenink D. Praat, a system for doing phonetics by computer. Glot Int. 2001; 5: 341–345.

[97] Kay Elemetrics Corp. Multi-Dimensional Voice Program (MDVP): Software Instruction Manual, Kay Elemetrics, Lincon Park, 2003.

[98] Cmejla R, Rusz J, Bergl P, Vokral J. Bayesian changepoint detection for automatic assessment of fluency and articulatory disorders. Speech Commun. 2013; 55: 178–189.

[99] Tsanas A, Little MA, McSharry PE, Ramig LO. Accurate telemonitoring of Parkinson's disease progression by a non-invasive speech test. IEEE Trans Biomed Eng. 2010; 57: 884–893.

[100] Goetz CG, Stebbins GT, Wolff D, et al. Testing objective measures of motor impairment in early Parkinson's disease: feasibility study of an at-home testing device. Movement Disord. 2008; 24(4): 551–556.

[101] Earnest MM, Max L. En route to the three-dimensional registration and analysis of speech movements: Instrumental techniques for the study of articulatory kinematics. Contemp Issues Commun Sci Disord. 2003; 30: 5-25.

[102] Shellikeri S, Yunusova Y, Thomas D, Green JR, Zinman L. Compensatory articulation in amyotrophic lateral sclerosis: Tongue and jaw interactions. J Acoust Soc Am. 2013; 5: 133.

[103] Wong MN, Murdoch BE, Whelan B. Lingual kinematics during rapid syllable repetition in Parkinson's disease. Int J Lang Comm Dis. 2012; 47: 578-588.

[104] Yunusova Y, Weismer G, Westbury JR, Lindstrom MJ. Articulatory movements during vowels in speakers with dysarthria and healthy controls. J Speech Lang Hear R. 2008; 51: 596-611.

[105] Toutios A, Ouni S, Laprie Y. Estimating the control parameters of an articulatory model from electromagnetic articulograph data. J Acoust Soc Am. 2011; 129(5): 3245-3257.

[106] Shtern M, Haworth MB, Yunusova Y, Baljko M, Faloustsos P. A game system for speech rehabilitation. In: Proceedings of the 5th International Conferences on Motion in Games (MiG); Nov 15-17, 2012; Rennes, France. 43-54.

[107] Katz W, Campbell T, Wang J, et al. Opti-speech: a real-time, 3D visual feedback system for speech training. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014); Sep 14-18, 2014; Singapore. 1174-1178.

[108] Feng Y, Max L. Accuracy and precision of a custom camera-based system for 2-D and 3-D motion tracking during speech and nonspeech motor tasks. J Speech Lang Hear R. 2014; 57: 426-438.

[109] Lanz C, Denzler J, Gross HM. Facial movement dysfunctions: Conceptual design of a therapy-accompanying training system. In: AAl-Kongress 2013; Jan 22-23, 2013; Berlin, Germany.

[110] Tjiaden K. Speech and swallowing in Parkinson's disease. Top Geriatr Rehabil. 2008; 24:115–126.

[111] Caligiuri MP. Labial kinematics during speech in patients with Parkinsonian rigidity. Brain. 1987; 110:1033–1044.

[112] Forrest K, Weismer G, Turner GS. Kinematic, acoustic, and perceptual analyses of connected speech produced by Parkinsonian and normal geriatric speakers. J Acoust Soc Am. 1989; 85:2608–2622.

[113] Svensson P, Henningson C, Karlsson S. Speech motor control in Parkinson's disease: a comparison between a clinical assessment protocol and a quantitative analysis of mandibular movements. Folia Phoniatr (Basel). 1993; 45:157–164.

[114] Forrest K, Weismer G. Dynamic aspects of lower lip movement in Parkinsonian and neurologically normal geriatric speakers' production of stress. J Speech Hear Res. 1995; 38:260–272.

[115] Ackermann H, Hertrich I, Daum I, et al. Kinematic analysis of articulatory movements in central motor disorders. Movement Disord. 1997; 12:1019–1027.

[116] Scherer KR, Clark-Polner E, Mortillaro M. In the eye of the beholder? Universality and cultural specificity in the expression and perception of emotion. Int J Psychol. 2011: 46(6): 401-435.

[117] Bettadapura V. Face expression recognition and analysis: the state of the art. arXiv preprint arXiv:1203.6722 2012.

[118] Hamm J, Kohler CG, Gur RC, Verma R. Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders. J Neurosci Meth. 2011; 200: 237-256.

[119] Wu P, Gonzalez I, Patsis G, et al. Objectifying facial expressivity assessment of Parkinson's patients: preliminary study. Comput Math Methods Med. 2014; 1-12.

[120] Jabon M, Bailenson J, Pontikakis E, Takayama L, Nass C. Facial expression analysis for predicting unsafe driving behavior. IEEE Pervasive Comput. 2010; 10(4): 84-95.

[121] Ekman P, Friesen WV. Manual for the facial action coding system. Palo Alto, CA: Consulting Psychologists Press; 1977.

[122] Sariyanidi E, Gunes H, Cavallaro A. Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans Pattern Anal Mach Intell. 2015; 37(6): 1113-1133.

[123] Zeng Z, Pantic M, Roisman GI, Huang TS. A survey of affect recognition methods: audio, visual and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell. 2009; 31(1): 39-58

[124] Valstar MF, Mehu M, Jiang B, Pantic M, Scherer K. Meta-analysis of the first facial expression recognition challenge. IEEE Trans Syst Man Cybern. 2012; 42(4): 966-979.

[125] Cootes TF, Edwards GJ, Taylor CJ. Active appearance models. IEEE Trans Pattern Anal Mach Intell. 2001; 23(6): 681-685.

[126] Matthews I, Baker S. Active appearance models revisited. Int J Comput Vision. 2004; 60(2):135-164.

[127] Xiong X, De la Torre F. Supervised descent method and its applications to face alignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); June 23–28, 2013; Portland, OR, USA. 532–539.

[128] Cristinacce D, Cootes T. Feature detection and tracking with constrained local models. In: Proceedings of the British Machine Vision Conference (BMVC); Sep 4-6, 2006; Edinburgh, UK. vol. 3: 929-938.

[129] Patras I, Pantic M. Particle filtering with factorized likelihoods for tracking facial features. In: Proceedings of the 6th IEEE International Conference on Automatic Face and Gesture Recognition; May 19, 2004; Seoul, South Korea.

[130] Soleymany M, Lichtenauer J, Pun T, Pantic M. A multimodal database for affect recognition and implicit tagging. IEEE Trans Affect Comput. 2012; 3(1):1-14.

[131] Pantic M, Patras I. Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In: Proceedings of the 2005 IEEE International Conference on Systems, Man and Cybernetics (SMC); Oct 10-12, 2005; Waikoloa, HI, USA.

[132] Zhao G, Pietikäinen M. Boosted multi-resolution spatiotemporal descriptors for facial expression recognition. Pattern Recogn Lett. 2009; 30(12): 1117-1127.

[133] Jeni LA, Girard JM, Cohn JF, De La Torre F. Continuous au intensity estimation using localized, sparse facial feature space. In: Proceedings of the 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG); Apr 22-26, 2013; Shanghai, China.

[134] Tian Y-L, Kanade T, Cohn J. Recognizing action units for facial expression analysis. IEEE Trans Pattern Anal Mach Intell. 2001; 23(2): 1-19.

[135] Kanade T, Cohn JF, Tian Y. Comprehensive database for facial expression analysis. In: Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition; Mar 28-30, 2000; Grenoble, France.

[136] Gross R, Matthews I, Cohn J, Kanade T, Baker S. Multipie. Image Vision Comput. 2010; 28(5): 807-813.

[137] Langner O, Dotsch R, Bijlstra G, Wigboldus DHJ, Hawk ST, van Knippenberg A. Presentation and validation of the Radboud Faces Database. Cognition Emotion. 2010; 24(8), 1377-1388.

[138] McKeown G, Valstar M, Cowie R, Pantic M, Schroder M. The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. IEEE Trans Affect Comput. 2012; 3(1): 5-17.

[139] Bartlett M, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J. Automatic recognition of facial actions in spontaneous expressions. J Multimed. 2006; 1(6): 22-35.

[140] Pantic M, Valstar M, Rademaker R, Maat L. Web-based database for facial expression analysis. In: Proceedings of the 2005 IEEE International Conference on Multimedia and Expo (ICME 2005); Jul 6-8, 2005; Amsterdam, The Netherlands.

[141] Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The extended Cohn-Kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW); Jun 13-18, 2010; San Francisco, CA, USA.

[142] Vinokurov N, Weinshall D, Arkadir D, Bergman H, Linetsky E. Quantifying Hypomimia in Parkinson Patients Using a Depth Camera. In: the 5th EAI International Symposium on Pervasive Computing Paradigms for Mental Health. Sep 24-25, 2015; Milan, Italy.

[143] Ricciardi L, Bologna M, Morgante F, et al. Reduced facial expressiveness in Parkinson's disease: a pure motor disorder? J Neurol Sci. 2015; 358(1-2): 125-130

[144] http://www.faceshift.com. Accessed: March 14, 2016.

[145] Scalise L, Bernacchia N, Ercoli I, Marchionni P. Heart rate measurement in neonatal patients using a webcamera. In: Proceedings of the 2012 IEEE International Symposium on Medical Measurements and Applications (MeMeA); May 18-19, 2012; Budapest, Hungary.

[146] Poh M-Z, McDuff DJ, Picard RW. Advancements in noncontact, multiparameter physiological measurements using a webcam. IEEE Trans Biomed Eng. 2011; 58(1): 7-11.

[147] Scalise L (2012). Non Contact Heart Monitoring. In: Richard Millis, Ed. Advances in Electrocardiograms - Methods and Analysis. Rijeka, Croatia: InTech; 2012: 82-106.

[148] Poh M-Z, McDuff DJ, Picard RW. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. Opt Express. 2010; 18(10): 10762-10774.

[149] Verkruysse W, Svaasand LO, Nelson JS. Remote plethysmographic imaging using ambient light. Opt Express. 2008; 16(26): 21434-21445.

[150] James CJ, Hesse CW. Independent component analysis for biomedical signals. Physiol Meas. 2005; 26: R15-R39.

[151] Hyvärinen A, Oja E. A fast fixed-point algorithm for independent component analysis. Neural Comput. 1997; 9: 483–1492

[152] Bell AJ, Sejnowski TJ. An information-maximization approach to blind separation and blind deconvolution. Neural Comput. 1995. 7: 1129-1159.

[153] Cardoso JF, Souloumiac A. Blind beamforming for non-Gaussian signals. Radar and Signal Processing, IEE Proceedings F. 1993; 140(6): 362-370.

[154] Lewandowska M, Rumiński J, Kocejko T, Nowak J. Measuring pulse rate with a webcam - A non-contact method for evaluating cardiac activity. In: Proceedings of the 2011 Federated Conference on Computer Science and Information Systems (FedCSIS 2011); Sep 18-21, 2011; Szczecin, Poland.

[155] Yu Y-P, Kwan B-H, Lim C-L, Wong S-L, Raveendran P. Video-based heart rate measurement using short-time Fourier transform. In: Proceedings of the International Symposium on Intelligent Signal Processing and Communications Systems (ISPACS 2013); Nov 12-15, 2013; Naha, Okinawa, Japan.

[156] Dactu D, Cidota M, Lukosch S, Rothkrantz L. Noncontact automatic heart rate analysis in visible spectrum by specific face regions. In: Proceedings of the 14th International Conference on Computer Systems and Technologies; Jun 28-29, 2013; Ruse, Bulgaria.

[157] McDuff D, Gontarek S, Picard RW. Improvements in remote cardiopulmonary measurement using a five band digital camera. IEEE Trans Biomed Eng. 2014; 61(10): 2593-2601.

[158] Zhang Q, Xu G-Q, Wang M, Zhou Y, Feng W. Webcam based non-contact real-time monitoring for the physiological parameters of drivers. In: Proceedings of the 4th Annual International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (IEEE-CYBER 2014); Jun 4-7, 2014; Hong Kong, China.

[159] Tarvainen MP, Ranta-Aho PO, Karjalainen PA. An advanced detrending method with application to HRV analysis. IEEE Trans Biomed Eng. 2002; 49(2): 172-175.

[160] Monkaresi H, Calvo RA, Yan H. A machine learning approach to improve contactless heart rate monitoring using a webcam. IEEE J Biomed Health Inform. 2014; 18(4): 1153-1160.

[161] Wu H-Y, Rubinstein M, Shih E, Guttag J, Durand F, Freeman W. Eulerian Video Magnification for Revealing Subtle Changes in the World. ACM Trans Graphic (Proceedings SIGGRAPH 2012). 2012; 31(4).

[162] Bousefsaf F, Maaoui C, Pruski A. Remote detection of mental workload changes using cardiac parameters assessed with a low-cost webcam. Comput Biol Med. 2014; 53: 154-163.

[163] Bal U. Non-contact estimation of heart rate and oxygen saturation using ambient light. Biomed Opt Express. 2015; 6(1): 86-97.

[164] Tarassenko L, Villaroel M, Guazzi A, Jorge J, Clifton DA, Pugh C. Non-contact video-based vital sign monitoring using ambient light and auto-regressive models. Physiol Meas. 2014; 35: 807-831.

[165] Xu S, Sun L, Rohde GK. Robust efficient estimation of heart rate pulse from video. Biomed Opt Express. 2014; 5(4): 1124-1135.

[166] https://msdn.microsoft.com. Accessed August 20, 2015.

[167] Otsu N. A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybern. 1979; 9: 62–66.

[168] Orlandi S, Dejonckere PH, Schoentgen J, Lebacq J, Ruqja N, Manfredi C. Effective pre-processing of long term noisy audio recordings: an aid to clinical monitoring. Biomed Signal Process Control. 2013; 8: 799–810.

[169] Manfredi C, Bocchi L, Cantarella G. A multipurpose user-friendly tool for voice analysis: application to pathological adult voices. Biomed Signal Process Control. 2009; 4: 212–220.

[170] Manfredi C, Peretti G. A new insight into post-surgical objective voice quality evaluation. Application to thyroplastic medialization. IEEE Trans Biomed Eng. 2006; 53(3): 442–451.

[171] Kasuya H, Ogawa S, Mashima K, Ebihara S. Normalised noise energy as an acoustic measure to evaluate pathologic voice. J Acoust Soc Am. 1986; 80(5): 1329–1334.

[172] Yumoto E, Gould WJ. Harmonics-to-noise ratio as an index of the degree of hoarseness. J Acoust Soc Am. 1982: 71(6): 1544–1550.

[173] Dejonckere PH, Schoentgen J, Giordano A, Fraj S, Bocchi L, Manfredi C. Validity of jitter measures in non-quasi-periodic voices. Part I: Perceptual and computer performances in cycle pattern recognition. Logoped Phoniatr Vocol. 2011; 36(2): 70–77.

[174] Manfredi C, Giordano A, Schoentgen J, Fraj S, Bocchi L, Dejonckere PH. Validity of jitter measures in non-quasi-periodic voices. Part II: The effect of noise. Logoped Phoniatr Vocol. 2011; 36(2): 78–89.

[175] Dejonckere PH, Giordano A, Schoentgen J, Fraj S, Bocchi L, Manfredi C. To what degree of voice perturbation are jitter measurements valid? A novel approach with synthesized vowels and visuo-perceptual pattern recognition. Biomed Signal Process Control. 2012; 7: 37–42.

[176] Manfredi C, Giordano A, Schoentgen J, Fraj S, Dejonckere PH. Perturbation measurements in highly irregular voice signals: performance/validity of analysis software tools. Biomed Signal Process Control. 2012; 7(4): 409–416.

[177] Russell JA, Ciucci MR, Connor NP, Schallert T. Targeted exercise therapy for voice and swallow in persons with Parkinson's disease. Brain Res. 2010; 1341: 3-11.

[178] Lanz C, Olgay BS, Denzler J, Gross H-M. Automated classification of therapeutic face exercises using the Kinect. In: Proceedings of the 8th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2013); Feb 21-24, 2013; Barcelona, Spain.

[179] Lowe DG. Distinctive image features from scale-invariant keypoints. Int J Comput Vision. 2004; 60(2): 91-110.

[180] Combescure P. Vingt listes de dix phrases phonéiquement équilibrées. Revue d'acoustique. 1981; 14(56).

[181] Lafon JCH. Le test phonétique et la mesure de l'audition. Eindhoven, The Netherlands: Dunod; 1964.

[182] Bocca E, Pellegrini A. Studio statistico sulla composizione fonetica della lingua italiana e sua applicazione pratica all'audiometria con la parola. Archivio Italiano di Otologia, Rinologia e Laringologia 1950; 56(5): 116-141.

[183] Szeliski R. Computer vision: algorithms and applications. London, UK: Springer; 2010.

[184] Lee K-F, Hon H-W, Reddy R. An overview of the SPHINX speech recognition system. IEEE Trans Acoust Speech. 1990; 38(1): 35-45.

[185] Bigi B. SPPAS: a tool for the phonetic segmentations of Speech. In: Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012); May 21-27, 2012; Istanbul, Turkey.

[186] Bigi B. SPPAS - Multi-lingual Approaches to the Automatic Annotation of Speech. "The Phonetician", International Society of Phonetic Sciences. 2015; 111-112: 55-69.

[187] Skodda S, Flasskamp A, Schlegel U. Instability of syllable repetition as a marker of disease progression in Parkinson's disease: a longitudinal study. Movement Disord. 2011; 1: 59-64.

[188] Heikkilä J, Silven O. A four-step camera calibration procedure with implicit image correction. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition; Jun 17-19, 1997; San Juan, Puerto Rico. 1106–1112.

[189] http://www.vision.caltech.edu/bouguetj/calib_doc/. Accessed August 20, 2015.

[190] Khoshelham K, Elberink SO. Accuracy and resolution of Kinect depth data for indoor mapping and applications. Sensors (Basel). 2012; 12: 1437-1454.

[191] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. 2nd Ed. New York, NY: Springer; 2009.

[192] http://clopinet.com/isabelle/Projects/modelselect/MFAQ.html. Accessed: March 13, 2016.

[193] http://scikit-learn.org/stable/supervised_learning.html. Accessed: March 13, 2016.

[194] Balakrishnan G, Durand F, Guttag J. Detecting pulse from head motions in video. In: Proceedings of the 2013 IEEE Computer Vision and Pattern Recognition Conference (CVPR 2013); Jun 23-28, 2013; Portland, OR, USA.

[195] Shan L, Yu M. Video-based heart rate measurement using head motion tracking and ICA. In Proceedings of the 2013 6th International Congress on Image and Signal Processing (CISP); Dec 16-18, 2013; Hangzhou, China.

[196] Fort A, Manfredi C, Rocchi S. Recursive autoregressive spectral maps for ocular pathology detection. Ultrasound Med Biol. 1997; 23(3): 391-403.

[197] Skodda S, Grönheit W, Schlegel U. Impairment of vowel articulation as a possible marker of disease progression in Parkinson's disease. PLOS ONE. 2012; 7(2): e32132.

[198] Makashay MJ, Cannard KR, Solomon NP. Speech-related fatigue and fatigability in Parkinson's disease. Clinical Linguist Phonet. 2015; 29: 27-45.

[199] Di Filippo NM, Jouaneh MK. Characterization of different Microsoft Kinect sensor models. Sensors (Basel). 2015; 15: 4554-4564.

[200] Bandini A, Giovannelli F, Orlandi S, et al. Automatic identification of dysprosody in idiopathic Parkinson's disease. Biomed Signal Process Control. 2015; 17: 47-54.

[201] Bandini A, Giovannelli F, Cincotta M, et al. Abnormal rhythms of speech inpatients iwth idiopathic Parkinson's disease. In: Proceedings of the 8th International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications (MAVEBA 2013); Dec 16-18; 2013; Florence, Italy. 67-70.

[202] Bandini A, Giovannelli F, Cincotta M, et al. Automatic detection of prosody patterns in patients with idiopathic Parkinson's disease. Clin Neurophysiol (Proceedings of the 59th National Congress of SINC - *Società Italiana di Neurofisiologia Clinical*; May 14-17; Milan, Italy). 2015; 126(1): e17-e18.

[203] Bandini A, Giovannelli F, Zaccara G, et al. Acoustic and kinematic measure of speech in idiopathic Parkinson's disease by means of contact-less techniques. In: Proceedings of the 4th National Congress of Bioengineering (GNB 2014); Jun 25-27, 2014; Pavia, Italy.

[204] Bandini A, Skodda S, Orlandi S, Manfredi C. Manual vs automatic segmentation of syllable repetition: application to dysprosody in idiopathic Parkinson's disease. In: Proceedings of the 11th Pan-European Voice Conference (PEVOC 2015); Aug 31 - Sep 2, 2015; Florence, Italy.

[205] Bandini A, Ouni S, Orlandi S, Manfredi C. Evaluating a markerless method for studying articulatory movements: applications to a syllable repetition task. In: Proceedings of the 9th International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications (MAVEBA 2015); Sep 2-4, 2015; Florence, Italy.

[206] Bandini A, Ouni S, Cosi P, Orlandi S, Manfredi C. Accuracy of a markerless acquisition technique for studying speech articulators. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015); Sep 6-10, 2015; Dresden, Germany.

[207] Bandini A, Orlandi S, Giovannelli F, et al. Markerless analysis of articulatory movements in patients with Parkinson's disease. J Voice. 2015; *in press.*

[208] Bandini A, Orlandi S, Capo A, Vannetti F, Pasquini G, Manfredi C. Contact-less video-based tracking of heart rate. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC 2015). Aug 25-29, 2015; Milan, Italy.

List of Publications

The following is a list of works related to the PhD project and published in the last three years.

International peer-reviewed journals:

- Bandini A, Orlandi S, Giovannelli F, et al. Markerless analysis of articulatory movements in patients with Parkinson's disease. J Voice. 2015; in press. doi:10.1016/j.jvoice.2015.10.014
- Bandini A, Giovannelli F, Orlandi S, et al. Automatic identification of dysprosody in idiopathic Parkinson's disease. Biomed Signal Process Control. 2015; 17: 47-54. doi:10.1016/j.bspc.2014.07.006

International conferences proceedings:

- Bandini A, Ouni S, Cosi P, Orlandi S, Manfredi C. Accuracy of a markerless acquisition technique for studying speech articulators. In: Proceedings of the 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015); Sep 6-10, 2015; Dresden, Germany.
- Bandini A, Ouni S, Orlandi S, Manfredi C. Evaluating a markerless method for studying articulatory movements: applications to a syllable repetition task. In: Proceedings of the 9th International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications (MAVEBA 2015); Sep 2-4, 2015; Florence, Italy.
- Bandini A, Skodda S, Orlandi S, Manfredi C. Manual vs automatic segmentation of syllable repetition: application to dysprosody in idiopathic Parkinson's disease. In: Proceedings of the 11th Pan-European Voice Conference (PEVOC 2015); Aug 31 - Sep 2, 2015; Florence, Italy.
- Bandini A, Orlandi S, Capo A, Vannetti F, Pasquini G, Manfredi C. Contact-less video-based tracking of heart rate. In: 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (IEEE EMBC 2015). Aug 25-29, 2015; Milan, Italy.
- Bandini A, Orlandi S, Giovannelli F, et al. Acoustical and video analysis for the assessment of speech impairments in idiopathic Parkinson's disease. In: 11th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research (AQL 2015). Apr 8-9, 2015; London, UK.
- Bandini A, Giovannelli F, Orlandi S, et al. Acoustic and kinematic analysis of speech in idiopathic Parkinson's disease. In: XXII Annual Pacific Voice Conference. Apr 11-13, 2014; Krakow, Poland.
- Bandini A, Giovannelli F, Cincotta M, et al. Abnormal rhythms of speech inpatients iwth idiopathic Parkinson's disease. In: Proceedings of the 8th International Workshop on Models and Analysis of Vocal Emission for Biomedical Applications (MAVEBA 2013); Dec 16-18; 2013; Florence, Italy. 67-70.

National Conferences Proceedings

- Bandini A, Giovannelli F, Zaccara G, et al. Acoustic and kinematic measure of speech in idiopathic Parkinson's disease by means of contact-less techniques. In: Proceedings of the 4th National Congress of Bioengineering (GNB 2014); Jun 25-27, 2014; Pavia, Italy.
- Bandini A, Giovannelli F, Cincotta M, et al. Automatic detection of prosody patterns in patients with idiopathic Parkinson's disease. Clin Neurophysiol (Proceedings of the 59th National Congress of SINC - *Società Italiana di Neurofisiologia Clinica*; May 14-17; Milan, Italy). 2015; 126(1): e17-e18.